From Gene Trees to Species Trees II: Species Tree Inference by Minimizing Deep Coalescence Events

Louxin Zhang

Abstract—When gene copies are sampled from various species, the resulting gene tree might disagree with the containing species tree. The primary causes of gene tree and species tree discord include incomplete lineage sorting, horizontal gene transfer, and gene duplication and loss. Each of these events yields a different parsimony criterion for inferring the (containing) species tree from gene trees. With incomplete lineage sorting, species tree inference is to find the tree minimizing extra gene lineages that had to coexist along species lineages; with gene duplication, it becomes to find the tree minimizing gene duplications and/or losses. In this paper, we present the following results:

(i) The deep coalescence cost is equal to the number of gene losses minus two times the gene duplication cost in the reconciliation of a uniquely leaf labeled gene tree and a species tree. The deep coalescence cost can be computed in linear time for any arbitrary gene tree and species tree.

(ii) The deep coalescence cost is always no less than the gene duplication cost in the reconciliation of an arbitrary gene tree and a species tree.

(iii) Species tree inference by minimizing deep coalescence events is NP-hard.

Index Terms—Gene tree and species tree reconciliation, deep coalescence, gene duplication and loss, the Parsimony principle, NP-hardness.

1 Introduction

Gene trees are fundamental to molecular systematics. Traditionally, a gene tree is reconstructed from DNA sequence variation at individual genetic loci in a group of species and is taken as the phylogenetic tree of the species due to sequencing technology limitations. However, when gene copies are sampled from various species, the resulting gene tree might disagree with the species tree. As such, the relationship between gene trees and species trees has been the focus of many studies (see for example [5], [11], [19], [24], [26], [30], [32]). It has long been recognized that gene trees can be used to estimate species divergence time, ancestral population sizes and even the containing species tree although they may not accurately reflect the species tree [7], [14], [20].

The discord of gene trees and the containing species tree can arise from horizontal gene transfer, incomplete lineage sorting, and gene duplication and loss. The importance of these causes depends on the considered genes and species. Hence, inferring the species tree from gene trees has been investigated under various parsimony criteria. With incomplete lineage sorting (also called deep coalescence), the problem is to find the tree minimizing extra gene lineages that had to coexist along species lineages [19]; with gene duplication, it becomes to find the tree minimizing gene duplications and/or losses [11], [24], [12], [27].

Inferring the species tree from a set of gene trees has often been studied under the gene duplication cost [1], [2], [3], [6], [8], [13], [15], [17], [29], [33] until very recently. In a seminal work [19], Maddison addressed incomplete lineage sorting in the framework of coalescence theory. Coalescence theory is an active branch of population genetics concerned with tracing the genealogical history of a present-day gene copy. For a gene sampled from two individuals, one may ask: How deep in time do these two lineages coalesce? Hence, the depth of this coalescence is a measure of the relationship between two sampled gene copies. The more deep in time coalescence occurs, the more distantly related they are. Maddison proposed to use the total number of "extra" gene lineages that fail to coalesce on a species tree to measure the inconsistence of a gene tree and the species tree, called the *deep coalescence cost*. For the gene tree and species tree shown in Fig. 1, there are three gene lineages on a branch and two gene lineages on another branch that fail to coalesce, giving the deep coalescence cost of 3. Since coalescence theory provides the probability that a gene tree would exist in a species tree, it allows the inference problem to be studied in an explicitly statistical framework [4], [28]. This seems to give the deep coalescence model an advantage over the other models.

This paper is a sequel of [18] that studies the complexity and algorithmic issues of inferring the species

- *Louxin Zhang is with the Department of Mathematics, National University of Singapore, 10 Lower Kent Ridge Road, Singapore 119076. Email: matzlx@nus.edu.sg*

tree from a set of gene trees with the gene duplication/loss cost in the reconciliation of a gene tree and the containing species tree. Here, we present an equation of the deep coalescence cost, the duplication cost, and the number of gene losses. We also show that the deep coalescence cost is no less than the gene duplication cost. Although deep coalescence and gene duplication are two different mechanisms responsible for the discord of gene trees and species trees, this relationship suggests that the deep coalescence cost and the duplication cost are closely related to each other as a similarity measure of trees. We further show that inferring species tree from gene trees by minimizing the deep coalescence cost is also NP-hard.

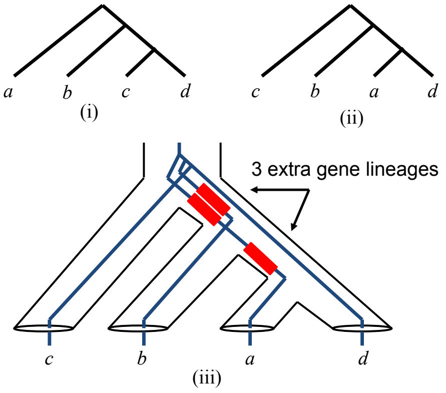


Fig. 1. (i) A gene tree. (ii) A species tree. (iii) The reconciliation of the gene tree in (i) and the species tree in (ii) has deep coalescence cost 3.

2 Basic Definitions and Notations

In this section, we shall introduce basic definitions and notations on gene duplication, gene loss and deep coalescence that are used in the rest of the paper.

2.1 Species Trees and Gene Trees

For a set of n taxa, their evolutionary history is modeled as a rooted, full binary tree with n leaves in which leaves are uniquely labeled with taxa, representing the labeling taxa, and internal nodes are unlabeled. Here, the 'fullness' means that each internal node has exactly two children. Such a tree is called a *species tree*. In a species tree, each unlabeled internal node is considered as a *taxon family* which include as its members the subordinate species represented by the leaves below it. Thus, the evolutionary relation "m is a descendant of n" is expressed using the set-theoretic notation as "$m \subsetneq n$". We also call an internal node an *ancestor* of the species below it.

The model for gene evolutionary relationship is also a rooted, full binary tree with leaves representing genes, called a *gene tree*. Usually, a gene tree is reconstructed from a collection of gene family members sampled from the considered species. We label the gene copies by the species from which they are sampled. In a gene tree, leaf labels may not be unique as two or more gene copies might be sampled in a species. An internal node g corresponds to a multiset of leaf labels.

Finally, for a species or gene tree T, we use $L(T)$ to denote the set of leaf labels of it. We write $t \in T$ to denote that t is an internal node of T. For any $t \in T$, $a(t)$ and $b(t)$ are used to denote its two children.

2.2 Gene Duplication

Let G be a gene tree and S a species tree such that $L(G) \subseteq L(S)$. For any nodes s', s'' in S, the *least common ancestor* of s' and s'' is defined to be the internal node $s \in S$ satisfying that $s', s'' \subseteq s$, but the children of s do not have this containment property, which is denoted by $\text{lca}(s', s'')$. To reconcile G and S, each node g of G is mapped to a unique node $M(g)$ in S as

$$M(g) = \begin{cases} \text{the leaf of } S \text{ with the same label}, & g \text{ is a leaf}, \\ \text{lca}\left(M(a(g)), M(b(g))\right), & \text{otherwise}. \end{cases}$$

This mapping M was first considered in [11] and then formulated in [24]. We call M the lca mapping or reconciliation of G with S. Obviously, if $g' \subseteq g$, $M(g') \subseteq M(g)$.

Definition 2.1: Let g be an internal node of G. If $M(c(g)) = M(g)$ for some child $c(g)$ of g, then we say that a duplication occurs at $M(g)$ (or more exactly in the lineage entering $M(g)$) in S.

The total number of duplications arising in the lca reconciliation of G in S is proposed to measure the discord of the gene tree and species tree and is called the *duplication cost*. We use $c_{dup}(G, S)$ to denote the duplication cost for G and S. Note that the duplication cost is not symmetric.

2.3 Gene Loss

A subset A of nodes of a species tree S is incomparable if, for any $x, y \in A$, one is not an ancestor of the other. For an incomparable subset A in S, the restriction of S on A is the smallest subtree of S containing A as its leaf set, denoted by $R_S(A)$. It is easy to see that the root of $R_S(A)$ is the least common ancestor of the nodes from A. The homomorphic subtree $S|_A$ of S induced by A is a tree obtained from $R_S(A)$ by contracting all degree-2 nodes except for the root of $R_S(A)$.

Let G be a gene tree such that $L(G) \subseteq L(S)$. $S|_{L(G)}$ is well defined. To reconcile G and S in this general case, we consider the lca mapping M from G to $S|_{L(G)}$. For any two nodes s and s' of $S|_{L(G)}$ such that $s \subsetneq s'$, we define

$$d(s, s') = |\{h \in S|_{L(G)} \ : \ s \subsetneq h \subsetneq s'\}|.$$

That is, $d(s, s')$ is the number of nodes on the path from s' to s excluding s and s'.

Recall that $a(g)$ and $b(g)$ denote the children of g. The number of losses l_g associated with g is defined as

$$l_g = \begin{cases} 0, & \text{if } M(g) = M(a(g)) = M(b(g)), \\ f(a(g)) + 1, & \text{if } M(a(g)) \subsetneq M(g) = M(b(g)), \\ \sum_{h=a(g),b(g)} f(h), & \text{if } M(a(g)), M(b(g)) \subsetneq M(g), \end{cases}$$

where, for a non-root node x in G, $f(x) = d\left(M(x), M\left(p(x)\right)\right)$ in which $p(x)$ denotes the parent of x. This definition of l_g is a generalization of the loss cost given in [12]. When $L(G) = L(S)$, our definition is then identical to the one given in [12].

The *gene loss cost* of the reconciliation of G in S is defined as the total number of losses $\sum_{g\in G} l_g$. We denote this gene loss cost for G and S by $c_{loss}(G, S)$.

2.4 Deep Coalescence

Let G be a gene tree and S a species tree such that $L(G) = L(S)$. Under the lca reconciliation $M : G \to S$, if a branch e of S is on the k paths from $M(g_i)$ to $M(c(g_i))$, $g_i \in G$ ($1 \le i \le k$), then we say that there are $k-1$ 'extra' lineages failing to coalesce on e. The *deep coalescence* (DC) *cost* is defined as the total number of the 'extra' lineages on all branches of S in the reconciliation M of G with S (see [19]), which is denoted by $c_{dc}(G, S)$. Note that the concept of deep coalescence is meaningful only if S has 2 or more leaves. We assume this fact throughout the paper.

In general, if $L(G) \subseteq L(S)$, the deep coalescence cost $c_{dc}(G, S)$ is defined as $c_{dc}(G, S|_{L(G)})$, where $S|_{L(G)}$ is the homomorphic subtree of S induced by $L(G)$. Such a generalization will be used in the study of inferring the species tree from a set of gene trees

3 An Equation of the Duplication, Loss and DC Costs

We have seen that deep coalescences, gene losses and duplications are inferred through gene tree and species tree reconciliation. In fact, the number of these events are indeed closely related through a simple equation.

Definition 3.1: Let G be a gene tree and S a species tree such that $L(G) \subseteq L(S)$. Under the lca reconciliation $M : G \to S$, an internal node $g \in G$ is of

- type-1 if $M(g') \subsetneq M(g)$ for every child g' of g;
- type-2 if there exists a unique child g' such that $M(g') = M(g)$;
- type-3 if $M(g') = M(g)$ for every child g' of g.

Note that type-2 or type-3 internal nodes correspond one-to-one with duplication events.

Theorem 3.1: Let G be a uniquely leaf-labeled gene tree and S a species tree such that $L(G) = L(S)$. Then,

$$c_{dc}(G, S) = c_{loss}(G, S) - 2c_{dup}(G, S).$$

Proof: Let G and S have n leaves. Assume that there are k_1 type-1 internal nodes

$$g_{11}, g_{12}, \dots, g_{1k_1},$$

k_2 type-2 internal nodes

$$g_{21}, g_{22}, \dots, g_{2k_2},$$

and k_3 type-3 internal nodes

$$g_{31}, g_{32}, \dots, g_{3k_3}$$

in G under the lca reconciliation $M : G \to S$, respectively. Since G is a full binary tree with n leaves, G has $n-1$ internal nodes and hence

$$k_1 + k_2 + k_3 = n - 1. \tag{1}$$

Additionally, type-2 and type-3 nodes correspond one-to-one with duplication events,

$$c_{dup}(G, S) = k_2 + k_3. \tag{2}$$

For simplicity, we assume that g' and g'' are the children of g for each type-1 internal node g; we also assume that $a(g)$ is the unique child such that $M\left(a(g)\right) \subsetneq M(g)$ for each type-2 node g. Since we use $d\left(M(h), M(g)\right)$ to denote the number of nodes on the path from $M(g)$ to $M(h)$ for a node g and its child h, the number of lineages contained in the path is $d\left(M(h), M(g)\right) + 1$. Therefore, by Eqn. (1) and (2) and the fact that $|E(S)| = 2n - 2$,

$$\begin{aligned} c_{dc}(G, S) &= \sum_{j=1}^{k_1} \left\{ \left[d\left(M(g'_{1j}), M(g_{1j})\right) + 1 \right] \right. \\ &\qquad \left. + \left[d\left(M(g''_{1j}), M(g_{1j})\right) + 1 \right] \right\} \\ &\quad + \sum_{j=1}^{k_2} \left[d\left(M(a(g_{2j})), M(g_{2j})\right) + 1 \right] - |E(S)| \\ &= c_{loss}(G, S) + 2k_1 - (2n - 2) \\ &= c_{loss}(G, S) - 2(k_2 + k_3) \\ &= c_{loss}(G, S) - 2c_{dup}(G, S). \end{aligned}$$

This concludes the proof. □

Remarks. (i) Following the proof of the equation in the above theorem, one can easily see that for an arbitrary gene tree G in which there may be two or more gene copies from a species and a species tree S such that $L(G) \subseteq L(S)$,

$$\begin{aligned} & c_{dc}(G, S) \\ = \; & c_{loss}(G, S) - 2c_{dup}(G, S) + (|G| - |R_S(L(G))|) \end{aligned} \tag{3}$$

where $|T|$ denotes the number of the nodes of T for $T = G, R_S(L(G))$. Note that when $L(G) \neq L(S)$, G is mapped onto $R_S(L(G))$, which is the restriction of S on the set of leaves whose labels are in $L(G)$ and may not be a fully binary tree.

(ii) With the presence of multiple gene copies, the last term in the right-hand side of Eqn (3) is the size difference of the gene tree and species tree. Since ancient gene duplication produces more gene copies than recent gene duplication, the deep coalescence cost penalizes ancient gene duplication more than recent one if gene loss is rare. This fact suggests that the deep coalescence cost might be more suitable for inferring recent duplication events than the gene duplication cost.

(iii) Since the number of gene duplications and losses can be calculated in linear time [33], [18], the first remark implies that the deep coalescence cost can also be computed in linear time.

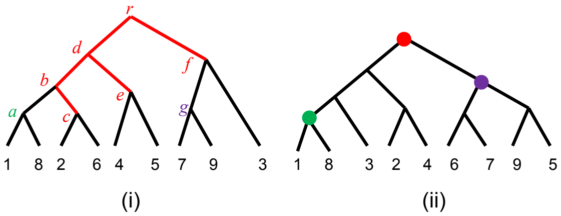


Fig. 2. (i) A gene tree. (ii) A species tree. In the lca reconciliation M of the gene tree with the species tree, a is mapped to the left highlighted node, b, c, d, e, f and r to the root, and g to the right highlighted node. The nodes b, c, d, e, f, r form a subtree of the gene tree.

By Theorem 3.1, $c_{dc}(G, S) \leq c_{loss}(G, S)$ for a species tree S and a uniquely leaf labeled gene tree G. Now we show that the DC cost is also bounded below by the duplication cost for any arbitrary gene trees.

Theorem 3.2: Let G be a uniquely leaf-labeled gene tree and S a species tree such that $L(G) = L(S)$. Then, $c_{dc}(G, S) \geq c_{dup}(G, S)$.

Proof: Denote the image node set of the lca mapping M by $M(G)$, which is a subset of nodes in the species tree S. For any internal node $s \in M(G)$, we use $M^{-1}(s)$ to denote all internal nodes g of the gene tree that are mapped to s under M. For any nodes x and a descendant y of x in the gene tree G, if $M(x) = M(y) = s$, then $M(g) = s$ for each node in the path from x to y. Since G is uniquely leaf labeled, all internal nodes in $M^{-1}(s)$ form a rooted subtree of G, denoted by $T^{-1}(s)$, as illustrated in Fig. 2.

$T^{-1}(s)$ is not a full binary tree in general. In particular, its root might have degree 1. Let n'_s, n''_s, n'''_s denote the number of non-root degree-1, degree-2 and degree-3 nodes in the subtree $T^{-1}(s)$, respectively. Assume that $T^{-1}(s)$ has two or more nodes. Then, by definition, the root of $T^{-1}(s)$ corresponds with a gene duplication in the reconciliation of G and S; each degree-2 or degree-3 node of $T^{-1}(s)$ also corresponds with a gene duplication. Therefore, there are $n''_s + n'''_s + 1$ duplication events at s. We now consider two cases.

Case 1. The root of $T^{-1}(s)$ has degree 1. Then $T^{-1}(s)$ has $n'''_s + 1$ leaves, that is $n'_s = n'''_s + 1$. For each leaf of $T^{-1}(s)$, it has two children that are mapped to a node below s in the species tree S; each non-root degree-2 node has exactly one child that is mapped to a node below s and so is the root since it has degree 1. Thus, there are $2\,(n'''_s + 1) + n''_s + 1$ image paths that contain one of the two lineages from s to one of its children.

Case 2. The root has degree 2. In this case, $T^{-1}(s)$ has $n'''_s + 2$ leaves and there are $2\,(n'''_s + 2) + n''_s$ image paths that contain one of the two lineages from s to one of its children.

By distributing the DC and duplication costs to each image node s in $M(G)$, we obtain that

$$\begin{aligned} c_{dc}(G, S) &\geq \sum_{s \in M(G): |T^{-1}(x)|>1} (\text{the no. of extra gene} \\ &\qquad\qquad \text{lineages in the branches leaving } s) \\ &\geq \sum_{s \in M(G): |T^{-1}(x)|>1} (2n'''_s + n''_s + 1) \\ &\geq \sum_{s \in M(G): |T^{-1}(x)|>1} (n'''_s + n''_s + 1) \\ &= c_{dup}(G, S). \qquad (4) \end{aligned}$$

This finishes the proof. □

Remark The fact $c_{dc}(G, S) \geq c_{dup}(G, S)$ holds even for any arbitrary gene tree in which 2 or more leaves with the same label, which represent genes sampled from the same species, and any species tree such that the lca reconciliation does not map any internal node to a leaf. In the general case, T^{-s} might be a forest – a union of rooted trees. However, the estimation (4) in the proof is still valid if the sum is over all the subtrees that are mapped to a node in the species tree, i.e. T^{-s} is replaced by a subtree of each resulting forest.

4 NP-hardness of the Species Tree Problem

The parsimony criterion is often used for inference in biology. Hence, inferring a species tree from a set of gene trees is usually formulated as the following algorithmic problem.

Species Tree Problem
INPUT: A set of gene trees G_i, $1 \leq i \leq n$.
SOLUTION: A species tree S that minimizes the total cost $\sum_i c(G_i, S)$, where $c(,)$ is a reconciliation cost function.

It is proved that the species tree problem is NP-hard for the duplication and loss costs in [18], which can also be generalized to the duplication plus loss cost. In this section, we prove the following theorem.

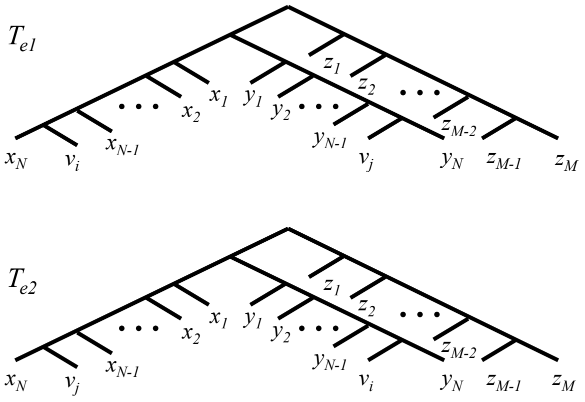


Fig. 3. Gene trees defined for each edge $e = (v_i, v_j)$.

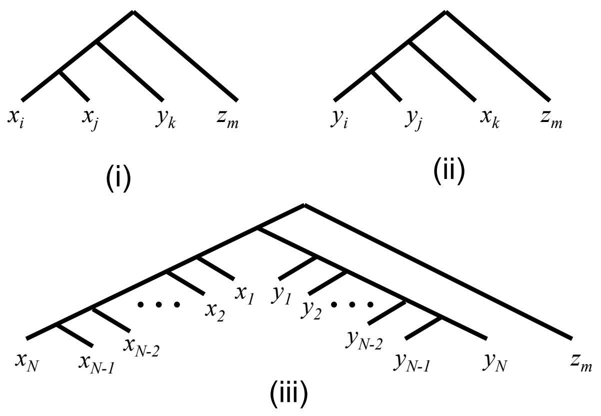


Fig. 4. ‘Structural’ gene trees with parameters i, j, k, m.

Theorem 4.1: The species tree problem is NP-hard under the DC cost.

Proof. Given a gene tree G and a species tree S, the DC cost $c_{dc}(G, S)$ can be computed in polynomial time since gene duplications and losses can be counted in linear time [33]. Therefore, the species tree problem is in NP.

To prove its NP-hardness, we reduce the Maximum Cut problem to the decision version of the species tree problem. Given an instance graph $\mathcal{G} = (V, E)$ and a positive integer I, the Maximum Cut problem is to partition the node set V into two disjoint subsets V_1 and V_2 such that there are at least I edges from E that have one endpoint in V_1 and one endpoint in V_2. Assume that $V = \{v_1, v_2, \cdots, v_n\}$ and $|E|$ denotes the number of edges in E, where $n > 3$. We construct a set $\mathcal{A}$ of gene trees to obtain a corresponding instance of the species tree problem.

Choose $N > n^2$ and $M \geq n^2N(N+1) + |E|$. For each node v_i ($1 \leq i \leq n$), we introduce a label with the same name v_i. We also introduce $2N + M$ extra labels x_i, y_i, $1 \leq i \leq N$ and z_j, $1 \leq j \leq M$. For each edge $e = (v_i, v_j) \in E$, we add to $\mathcal{A}$ two gene trees T_{e1} and T_{e2} as shown in Fig. 3. These two trees are same except that the leaf labels v_i and v_j are swapped.

Let the trees shown in Fig. 4 (i)-(iii) be written as $G_{(i,j,k,m)}$, $G'_{(i,j,k,m)}$ and $F[\{x_i\}, \{y_i\}, z_m]$, respectively. Besides the ‘edge’ gene trees T_{e1} and T_{e2} ($e \in E$), the set $\mathcal{A}$ of gene trees also contains

$$G_{(i,j,k,m)},\ 1 \leq i, j, k \leq N,\ i < j,\ 1 \leq m \leq M,$$
$$G'_{(i,j,k,m)},\ 1 \leq i, j, k \leq N,\ i < j,\ 1 \leq m \leq M,$$
$$G''_m = F[\{x_i\}, \{y_i\}, z_m],\ 1 \leq m \leq M.$$

These three classes of gene trees are introduced to restrict the topology of the optimal species tree for the defined instance of the problem. Hence, we call them ‘structural’ gene trees. The NP-completeness of the decision version of the species tree problem follows from the following two lemmas. Although the proof is long, the idea is quite simple. The parameter M is set so large that the structural gene trees will force the species trees with the minimum DC cost to be three line subtrees joined together as shown in Fig. 5, one of which contains x_is and some v_js, giving a cut of the graph $\mathcal{G}$.

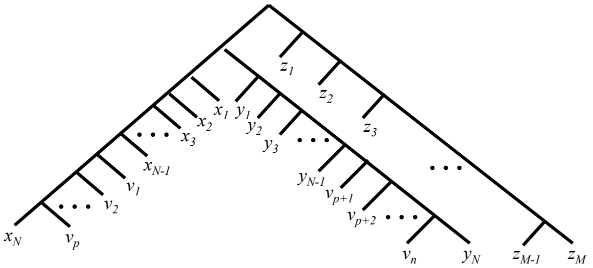


Fig. 5. Species tree $S_{\mathcal{G}}$ defined from a cut (V_1, V_2) of $\mathcal{G}$ in Lemma 4.1.

Lemma 4.1: If the graph $\mathcal{G}$ has a cut of d edges, there is a species tree $S_{\mathcal{G}}$ having the DC cost

$$c_{dc}(\mathcal{A}, S_{\mathcal{G}}) = N(N+1)|E| + |E| - d.$$

Proof: Assume that the node set V of the graph $\mathcal{G}$ is divided into $V_1 = \{v_1, v_2, \cdots, v_p\}$ and $V_2 = \{v_{p+1}, v_{p+2}, \cdots, v_n\}$ such that there are exactly d edges having one endpoint in V_1 and one endpoint in V_2. We define a species tree $S_{\mathcal{G}}$ as shown in Fig. 5.

First, we observe that

$$c_{dc}(G_{(i,j,k,m)}, S_{\mathcal{G}}) = 0,$$
$$c_{dc}(G'_{(i,j,k,m)}, S_{\mathcal{G}}) = 0,$$
$$c_{dc}(G''_m, S_{\mathcal{G}}) = 0,$$

for each possible i, j, k, m.

Consider a non-cut edge $e = (v_i, v_j)$ ($i < j$). Since $L(T_{e1}) = L(T_{e2}) \subset L(S_{\mathcal{G}})$,

$$c_{dc}\left(T_{e1}, S_{\mathcal{G}}\right) = c_{dc}\left(T_{e1}, S_{\mathcal{G}}|_{L(T_{e1})}\right)$$

and

$$c_{dc}\left(T_{e2}, S_{\mathcal{G}}\right) = c_{dc}\left(T_{e2}, S_{\mathcal{G}}|_{L(T_{e2})}\right).$$

To determine these DC costs, we consider $S_{\mathcal{G}}|_{L(T_{e1})}$. Without loss of generality, we assume $v_i, v_j \in V_1$ and hence $S_{\mathcal{G}}|_{L(T_{e1})}$ becomes the one shown in Fig. 6. In the reconciliation of T_{e1} and $S_{\mathcal{G}}|_{L(T_{e1})}$,, all extra lineages occur in the line subtrees with x’s and y’s; there are no deep coalescence events in the branch $(p(u), u)$ or on branches on the paths from the root to z_M.

The left child of the root of T_{e1} is mapped to u; $p(x_N) = p(v_i)$ is mapped to v and $p(x_k)$ is mapped to the corresponding $p(x_k)$ for each $1 \leq k \leq N-1$; $p(y_1), p(y_2), \cdots, p(y_N)$ are all mapped to u since v_j belongs to the left subtree and y_N belongs to the right subtree of u. Therefore, there is exactly one extra lineage in each of the $N+1$ branches on the path from u to $p(v_j)$. In addition, since, for each $1 \leq k \leq N$, the branch $(p(y_k), y_k)$ of T_{e1} is mapped the path from u to y_k, there are $N-1$ extra lineages on the branch $(u, p(y_1))$ and $N-k$

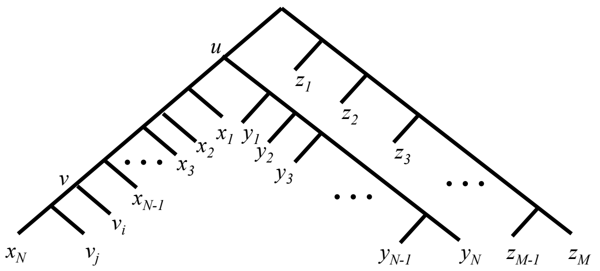


Fig. 6. The tree $S_{\mathcal{G}}|_{L(T_{e1})}$ defined in Lemma 4.1 when $v_i, v_j \in V_1$ and $i < j$.

extra lineages on the branch $(p(y_{k-1}), p(y_k))$ for $k \geq 2$. In total, we have

$$c_{dc}(T_{e1}, S_{\mathcal{G}}) = \frac{1}{2}N(N-1) + N + 1$$

Similarly,

$$c_{dc}(T_{e2}, S_{\mathcal{G}}) = \frac{1}{2}N(N-1) + N.$$

For each cut edge $e = (v_i, v_j)$ $(i < j)$ with one endpoint in V_1, say $v_i \in V_1$, and another in V_2, we have that

$$c_{dc}(T_{e1}, S_{\mathcal{G}}) = 0, \quad c_{dc}(T_{e2}, S_{\mathcal{G}}) = N(N+1). \qquad (5)$$

Therefore, we have

$$c_{dc}(\mathcal{A}, S_{\mathcal{G}}) = N(N+1)|E| + |E| - d.$$

This finishes the proof of the lemma. □

Lemma 4.2: If there is a species tree S having the DC cost $c_{dc}(\mathcal{A}, S) = N(N+1)|E| + t$, then the graph $\mathcal{G}$ has a cut of at least $|E| - t$ edges.

Proof: If $t > |E|$, the fact is trivial. Hence, without loss of generality, we may assume that $t \leq |E|$. Here, we use LT$[a, \ldots, b, c]$ to denote the line tree with leaves labeled by a, b, …, c, respectively, as shown in Fig. 7 (i). Note that the leaf a is a child of the root in LT$[a, \ldots, b, c]$. For a set of trees T', T'', …, T''', we use

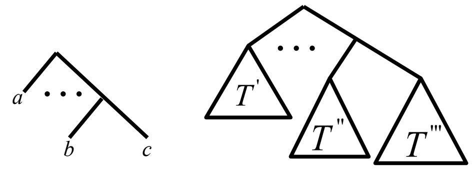
$$\text{LT}[T', \ldots, T'', T''']$$

to denote the tree obtained by replacing each leaf by a corresponding subtree in LT$[a, \ldots, b, c]$ as shown in Fig. 7 (ii).

Let B be a subset of leaves in the species tree S and the least common ancestor of the leaves from B be r_B in S. Recall that the homomorphic subtree $S|_B$ of S induced by B is the tree obtained from S by removing all the nodes and edges that are not on a path from r_B to a leaf from B and then contracting all the degree-2 node except for the root r_B. For example, for $S_{\mathcal{G}}$ defined in Lemma4.1, $S_{\mathcal{G}}|_{\{x_1, x_2, y_1\}} = \text{LT}[y_1, x_1, x_2]$.

Set

$$U = \{x_i, y_i : 1 \leq i \leq N\} \cup \{v_1, v_2, \cdots, v_n\},$$
$$Z = \{z_1, z_2, \cdots, z_M\}. \qquad (6)$$



(i) (ii)

Fig. 7. (i) Line tree LT$[a, \ldots, b, c]$. (ii) The resulting tree LT$[T', \ldots, T'', T''']$ after replacing each leaf with a tree in a line tree.

By replacing the children of a two-leaf rooted tree with $S|_U$ and $S|_Z$, we obtain a species tree $S' = \text{LT}[S|_U, S|_Z]$ from S. First, S' has the following property.

Fact 1 $c_{dc}(\mathcal{A}, S') \leq c_{dc}(\mathcal{A}, S) = N(N+1)|E| + t$.

Proof. For each gene tree $T = T_{e1}$ or T_{e2}, we use f and f' to denote the lca mappings from T to S and S', respectively. Let r be the root of T. Assume that $a(r)$ is the left child of r, the least common ancestor of x_is and y_is, and $b(r)$ the right child of r. For each edge $e = (u_1, u_2)$ on a path from $b(r)$ to some z_i, by the definition of $S|_Z$,

$$d(f'(u_1), f'(u_2)) \leq d(f(u_1), f(u_2)),$$

and, furthermore, $f(u_1) = f(u_2)$ if and only if $f'(u_1) = f'(u_2)$. For each edge below $a(r)$, the same property holds. However, the edges incident to the root of T may not satisfy the property discussed above. It is possible that $f(r) = f(a(r))$ and/or $f(r) = f(b(r))$. However, $f'(r) = r'$, $f'(a(r)) = a(r')$ and $f'(b(r)) = b(r')$, where r' is the root of S', $a(r')$ and $b(r')$ the root of $S|_U$ and $S|_Z$ respectively. Since no other lineages fail to coalesce with $(r, a(r))$ on $(r', a(r'))$ and with $(r, b(r))$ on $(r', b(r'))$ respectively, these two edges does not contribute the deep coalescence cost. Thus, $c_{dc}(T, S') \leq c_{dc}(T, S)$.

Similarly, we also have the following three inequalities

$$\begin{aligned} c_{dc}\left(G(i,j,k,m), S'\right) &\leq c_{dc}\left(G(i,j,k,m), S\right) \\ c_{dc}\left(G'(i,j,k,m), S'\right) &\leq c_{dc}\left(G'(i,j,k,m), S\right) \\ c_{dc}\left(G''(m), S'\right) &\leq c_{dc}\left(G''(m), S\right) \end{aligned}$$

for any i, j, k, m. Thus, the fact holds. □

Fact 2. In $S|_U$, all the leaves x_i must be below one child of the root and all the leaves y_i must be below the other child of the root. In other words, $S|_U = \text{LT}[T_1, T_2]$, where T_1 is a tree over x_i and some v_is and T_2 is a tree over y_is and some v_js.

Proof. Assume that the fact is false. There are x_i, x_j and y_k such that $S|_{\{x_i, x_j, y_k\}} = (S|_U)|_{\{x_i, x_j, y_k\}} = \text{LT}[x_i, x_j, y_k]$, or there are y_i, y_j and x_k such that $S|_{\{y_i, y_j, x_k\}} = (S|_U)|_{\{y_i, y_j, x_k\}} = \text{LT}[y_i, y_j, x_k]$. If the former is true, then,

$$c_{dc}\left(G_{(i,j,k,m)}, S'\right) \geq 1, \ 1 \leq m \leq M.$$

This implies that

$$N(N+1)|E|+t \geq c_{dc}(\mathcal{A}, S') \geq \sum_{m=1}^{M} c_{dc}\left(G_{(i,j,k,m)}, S'\right) = M,$$

contradicting to the fact that $M \geq N(N+1)n^2$.

If the latter is true, for any $1 \leq m \leq M$,

$$c_{dc}\left(G'_{(i,j,k,m)}, S'\right) \geq 1.$$

Again, we have that $c_{dc}(\mathcal{A}, S') \geq M$, leading to a contradiction. □

Let $X = \{x_1, x_2, \ldots, x_N\}$ and $Y = \{y_1, y_2, \ldots, y_N\}$. Then $S'|_X = (S|_U)|_X$ and $S'|_Y = (S|_U)|_Y$.

Fact 3. $S'|_X = \text{LT}[x_1, x_2, \ldots, x_N]$ and $S'|_Y = \text{LT}[y_1, y_2, \ldots, y_N]$.

Proof. Note that $G''_m|_X = \text{LT}[x_1, x_2, \ldots, x_N]$ and $G''_m|_Y = \text{LT}[y_1, y_2, \ldots, y_N]$ for any $1 \leq m \leq M$. If the claim is false, then, $c_{dc}(G''_m, S') \geq 1$ for any m and hence

$$N(N+1)|E| + t \geq c_{dc}(\mathcal{A}, S') \geq \sum_{m=1}^{M} c_{dc}(G''_m, S') = M,$$

a contradiction as in the proof of Fact 2. □

Let the least common ancestor of x_is and y_is be w in S'. We have shown that x_is are below one child of w, say w_1, and y_i's are below the other child of r, say w_2. Recall that $S'|_X$ and S'_Y are two line trees.

Fact 4. For each edge $e = (v_i, v_j)$ $(i < j)$ such that v_i and v_j are in the same subtree as x_is or as y_is, then

$$c_{dc}(T_{e1}, S') + c_{dc}(T_{e2}, S') \geq N(N+1) + 1.$$

Proof. Without loss of generality, we may assume that v_i and v_j are below w_1 in the same subtree as x_is. We consider the following two cases.

If

$$\begin{aligned} & S|_{X \cup \{v_i, v_j\}} \\ = \;& \text{LT}[x_1, x_2, \cdots, x_k, v_i, x_{k+1}, \cdots, x_m, v_j, v_{m+1}, \cdots, x_N] \end{aligned}$$

for some $0 \leq k \leq m \leq N$, we have that

$$c_{dc}(T_{e1}, S') = \frac{1}{2}N(N-1) + N + 1 + \frac{1}{2}(N-k)(N-k-1)$$

and

$$c_{dc}(T_{e2}, S') = \frac{1}{2}N(N-1) + k + 1 + \frac{1}{2}(N-m)(N-m-1).$$

Hence,

$$\begin{aligned} & c_{dc}(T_{e1}, S') + c_{dc}(T_{e2}, S') \\ \geq \;& N(N-1) + N + 1 + \frac{1}{2}[(N-k)(N-k-1) + 2k + 2] \\ \geq \;& N(N+1) + 1 \end{aligned}$$

as the minimum value of $(N-k)(N-k-1) + 2k + 2$ is N (reaching at $k = N-2, N-1$).

If

$$\begin{aligned} & S|_{X \cup \{v_i, v_j\}} \\ = \;& \text{LT}[x_1, x_2, \cdots, x_k, \text{LT}[v_i, v_j], x_{k+1}, \cdots, x_{N-1}, x_N] \end{aligned}$$

for some $0 \leq k \leq N$, we have that

$$c_{dc}(T_{e1}, S') = \frac{1}{2}N(N-1) + k + 2 + \frac{1}{2}(N-k)(N-k-1)$$

and

$$c_{dc}(T_{e2}, S') = \frac{1}{2}N(N-1) + k + 2 + \frac{1}{2}(N-k)(N-k-1).$$

Therefore,

$$\begin{aligned} & c_{dc}(T_{e1}, S') + c_{dc}(T_{e2}, S') \\ \geq \;& N(N-1) + 2k + 4 + (N-k)(N-k-1) \\ \geq \;& N(N+1) + 2 \end{aligned}$$

as the minimum value of $2k + (N-k)(N-k-1)$ is 2N-2 (reaching at $k = N-1, N-2$). The fact is proved. □

Fact 5. For each edge $e = (v_i, v_j)$ such that v_i is below w_1 in the same subtree as x_i and v_j is below w_2 in the subtree as y_is. Then,

$$c_{dc}(T_{e1}, S') + c_{dc}(T_{e2}, S') \geq N(N+1).$$

Proof. Let

$$S|_{X \cup \{v_i\}} = \text{LT}[x_1, x_2, \cdots, x_k, v_i, x_{k+1}, \cdots, x_{N-1}, x_N]$$

and

$$S|_{Y \cup \{v_j\}} = \text{LT}[y_1, y_2, \cdots, y_m, v_j, y_{m+1}, \cdots, y_{N-1}, y_N].$$

We have that all the internal nodes in T_{e2} are mapped onto the least common ancestor w of x_is and y_js and thus

$$c_{dc}(T_{e2}, S') = N(N+1).$$

Since $c_{dc}(T_{e1}, S') \geq 0$, the fact is proved. □

Let V_1 denote the subset of leaves v_i below w_1 in the same subtree as x_is and V_2 the subset of leaves v_j below w_2 in the same subtree as y_is. Then (V_1, V_2) is a cut of the graph $\mathcal{G}$. Assume there are p cut edges. Since there are $|E| - p$ non-cut edges,

$$\begin{aligned} & N(N+1)|E| + t \\ = \;& c_{dc}(\mathcal{A}, S') \\ \geq \;& (|E| - p)N(N+1) + pN(N+1) + (|E| - p) \\ = \;& N(N+1)|E| + |E| - p, \end{aligned}$$

which implies that $p \geq |E| - t$. This finishes the proof of Lemma 4.2. □

5 Conclusion

We conclude this paper by posing three related research problems. In this paper, we have proved that species tree inference by minimizing deep coalescences is NP-hard. This justifies the effort from different groups in seeking efficient heuristic methods for the inference problem [20], [31]. We have also discussed the relationship of the deep coalescence cost and the gene duplication cost. There are two research problems arising from this study. Does the species tree inference problem remain NP-hard if every gene tree has the same set of taxon labels? Is there any polynomial-time algorithm with constant approximation ratio for the species tree problem in the deep coalescence model? Note that the heuristic method developed by Than and Nakhleh in [31] seems to be effective.

The parametric complexity of the species tree inference by minimizing gene duplications has been studied in the past several years. Is it possible to develop efficient algorithms for parametric species tree inference under the deep coalescence model?

Acknowledgements

The author would like to thank four anonymous reviewers for their useful commentary. LX Zhang was financially supported by AcRF R146-000-109-112. A preliminary version of this work was presented in the poster session of the RECOMB'2000.

References

[1] M.S. Bansal and O. Eulenstein, "The multiple gene duplication problem revisited", *Bioinformatics*, vol. 24, pp. 132-138, 2008.

[2] C. Chauve, J.P. Doyon, and N. El-Mabrouk, "Gene family evolution by duplication, speciation, and loss", *J. Comput. Biol.*, vol. 15, pp. 1043-1062, 2008.

[3] K. Chen, D. Durand, and M. Farach-Colton, "Notung: A program for dating gene duplications and optimizing gene family trees", *J. Comput. Biol.*, vol. 7, pp. 429-447, 2000.

[4] J.H. Degnan and L.A. Salter, "Gene tree distribution under the coalescence process", *Evolution*, vol. 59, pp. 24-37, 2005.

[5] J.J. Doyle, "Gene trees and species trees: molecular systematics as one-character taxonomy", *Syst. Bot.*, vol. 17, pp. 144-163, 1992.

[6] D. Durand, B.V. Halldorsson, and B. Vernot, "A hybrid micro-macroevolutionary approach to gene tree reconstruction", *J. Comput. Biol.*, vol. 13, pp. 320-335, 2006.

[7] S.V. Edwards and P. Beerli, "Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeography studies", *Evolution*, vol. 54, pp. 1839-1854, 2000.

[8] O. Eulenstein, B. Mirkin, and M. Vingron, "Duplication-based measures of difference between gene and species trees", *J. Comput. Biol.*, vol. 5, pp. 135-148, 1998.

[9] W. Fitch, "Distinguishing homologous from analogous proteins", *Syst. Zool.*, vol. 19, pp. 99-113, 1970.

[10] M. Garey and D. Johnson, *Computers and Intractability: A Guide to the Theory of NP-completeness*, W. H. Freeman, New York, 1979.

[11] M. Goodman, J. Czelusniak, G.W. Moore, A.E. Romero-Herrera, and G. Matsuda, "Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences", *Syst. Zool.*, vol. 28, pp. 132-163, 1979.

[12] R. Guigó, I. Muchnik, and T. Smith, "Reconstruction of ancient molecular phylogeny", *Mol. Phylogenet. Evol.*, vol. 6, pp. 189-213, 1996.

[13] M.T. Hallett and J. Lagergren, "New algorithms for the duplication-loss model", In *Proceedings of RECOMB'2000*, pp. 138-146, 2000.

[14] J. Hey and R. Nielsen, "Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*", *Genetics*, vol. 167, pp. 747-760, 2004.

[15] R. Libeskind-Hadas and M.A. Charleston, "On the computational complexity of the reticulate cophylogeny reconstruction problem", *J. Comput. Biol.*, vol. 16, pp. 105-117, 2009.

[16] L. Liu, L.L. Yu, L. Kubatko, D.K. Pearl, S.V. Edwards, "Coalescent methods for estimating phylogenetic trees", *Mol. Phylogenet. Evol.*, vol. 53, pp. 320-328, 2009.

[17] C.W. Luo, M.C. Chen, Y.C. Chen, W.L. Yang, H.F. Liu, and K.-M. Chao, "Linear-time algorithms for the multiple gene duplication problems", *IEEE Trans. Comput Biol. and Bioinform.* (in press), 2010.

[18] B. Ma, M. Li and L.X. Zhang, "From gene trees to species trees", *SIAM J. Comput.*, vol. 30, pp. 729-752, 2001.

[19] W.P. Maddison, "Gene trees in species trees", *Syst. Biol.*, vol. 46, pp. 523-536, 1997.

[20] W.P. Maddison and L. Knowles, "Inferring phylogeny despite incomplete lineage sorting", *Syst. Biol.*, vol. 55, pp. 21-30, 2006.

[21] B. Mirkin, I. Muchnik and T. Smith, "A biologically meaningful model for comparing molecular phylogenies", *J. Comput. Biol.*, vol. 2, pp. 493-507, 1995.

[22] M.M. Miyamoto and W.T. Fitch, "Testing species phylogenies and phylogenetic methods with congruence", *Syst. Biol.*, vol. 44, pp. 64-76, 1995.

[23] M. Nei, *Molecular Evolutionary Genetics*, Columbia University Press, New York, 1987.

[24] R. Page, "Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas", *Syst. Bio.*, vol. 43, pp. 58-77, 1994.

[25] R. Page and M. Charleston, "From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem", *Mol. Phylogenet. Evol.*, vol. 7, pp. 231-240, 1997.

[26] P. Pamilo and M. Nei, "Relationship between gene trees and species trees", *Mol. Biol. Evol.*, vol. 5, pp. 568-583, 1988.

[27] F. Ronquist, "Phylogenetic approaches in coevolution and biogeography", *Zool. Scripta*, vol. 26, pp. 313-322, 1997.

[28] N.A. Rosenberg, "The probability of topological concordance of gene trees and species trees", *Theor. Popul. Biol.*, vol. 61, pp. 225-247, 2002.

[29] C. Roth, A. Rastogi, L. Arvestad, K. Dittmar, S. Light, D. Ekman, A. David, and D.A. Liberles, "Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms", *J Exp. Zool. Part B*, vol. 308, pp. 58-73, 2007.

[30] N. Takahata, "Gene genealogy in three related population: Consistency probability between gene and population trees", *Genetics*, vol. 122, pp. 957-966, 1989.

[31] C. Than and L. Nakhlen, "Species tree inference by minimizing deep coalescences", *PLoS Comput. Biol.*, vol. 5, e1000501.doi:10.1371/journal.pcbi.1000501, 2009.

[32] C.-I. Wu, "Inference of species phylogeny in relation to segregation of ancient polymorphisms", *Genetics*, vol. 127, pp. 429-435, 1991.

[33] L.X. Zhang, "On a Mirkin-Muchnik-Smith conjecture for comparing molecular phylogenies", *J. Comput. Biol.*, vol. 4, pp. 177-188, 1997.

Figure and table captions

Figure 1: (i) A gene tree. (ii) A species tree. (iii) The reconciliation of the gene tree in (i) and the species tree in (ii) has deep coalescence cost 3.

Figure 2: (i) A gene tree. (ii) A species tree. In the lca reconciliation M of the gene tree with the species tree, a is mapped to the left highlighted node, b, c, d, e, f and r to the root, and g to the right highlighted node. The nodes b, c, d, e, f, r form a subtree of the gene tree.

Figure 3: Gene trees defined for each edge $e = (v_i, v_j)$.

Figure 4: 'Structural' gene trees with parameters i, j, k, m.

Figure 5: Species tree $S_{\mathcal{G}}$ defined from a cut (V_1, V_2) of $\mathcal{G}$ in Lemma 4.1.

Figure 6: The tree $S_{\mathcal{G}}|_{L(T_{e1})}$ defined in Lemma 4.1 when $v_i, v_j \in V_1$ and $i < j$.

Figure 7: (i) Line tree LT$[a, \dots, b, c]$. (ii) The resulting tree LT$[T', \dots, T'', T''']$ after replacing each leaf with a tree in a line tree.