On Reconstructing Species Trees From Gene Trees In Term Of Duplications And Losses

Bin Ma[*], Ming Li[†], and Louxin Zhang[‡]

Abstract

This paper studies various properties of the least common ancestors mapping, the duplication and mutation costs, and the complexity of finding a species tree from gene trees.

1 Introduction

Since DNA sequences have become easier to obtain, emphasis has been placed on constructing gene trees and from these, reconstructing evolutionary trees for species (called *species trees*) in the evolutionary biology([8, 18, 6]). The current strategy for reconstructing species trees is based on the separate consideration of distinct gene families represented by homologous sequences. The homologous sequences are assumed to evolve in the same way as species. However, because of the presence of paralogy, sorting of ancestral polymorphism and horizontal transfer, gene trees and species trees are often inconsistent([19, 23, 26, 28]). Hence, a major problem that arises is how to reconcile different, sometimes contradictory, gene trees into a species tree([7]). This problem has been studied extensively for the last two decades. Several similarity/dissimilarity measures for gene trees and species trees have been proposed and efficient comparison methods have been investigated([24, 27, 11, 14, 13, 16, 1, 5, 12, 15].)

This paper studies the problem of combining different gene trees into a species tree under two duplication-based similarity/dissimilarity measures. These measures are proposed by Goodman et al.([11]), Page([21]), and Guigó *et al.* ([12]). Gene divergence causes all inconsistency among different gene trees and can be the results of either speciation or duplication([20]). If the common ancestry of two genes can be tracked back to a speciation event, then they are said to be related by *orthology*; if it is tracked back to a duplication event, then they are related by *paralogy*([7]). Taking account of orthology and paralogy evolutions, Goodman *et al.* proposed a similarity/dissimilarity measure for annotating species tree with duplications, gene losses and the nucleotide replacements([11]). Later, Page developed a method based on duplications for interpreting inconsistency between vertebrate globin gene trees and the species tree based on morphological data([21]); Guigó *et al.* elaborated the idea for identifying and locating the gene duplications in eukaryotic history([12]).

The duplication cost introduced by Page and the mutation cost by Guigó *et al.* are based on a mapping from gene trees to a species tree.Assuming that only genes from each contemporary species are presented in gene trees, we may denote a contemporary species and the genes from that species by a same symbol. In a gene tree, an ancestral gene is uniquely defined by the set of contemporary genes descending it. Similarly, in a species tree, an ancient species is defined by the contemporary species descending it. The mapping M from a gene tree to a species tree maps a contemporary gene to the corresponding species, and an ancestral one to the most recent one which contains that gene. Hence, we call it the *least common ancestor*(*l.c.a.*) *mapping* in this paper. When the gene and species trees are inconsistent, it may map an ancestral gene, say g, and its child $c(g)$ to the same ancient species. In this case, we say a *duplication* happens at g. Furthermore, roughly speaking, the number of losses associated with g is defined as the total number of interspecies between $M(g)$ and $M(c(g))$ for all children $c(g)$. To measure the similarity/dissimilarity between a gene and species trees, Page defined the duplication cost as the number of duplications, and Guigó *et al.* defined the mutation cost as the sum of the numbers of duplications and of gene losses (under the l.c.a. mapping). The mutation cost is not only biological meaningful([15]), but also efficiently computable, as proved by Eulenstein and Vingron([3]) and Zhang([29]) independently (see also [4]). Recon-

[*]Department of Mathematics, Beijing University, Beijing 100871, People's Republic of China. Email: bma@sxx11.math.pku.edu.cn.

[†]The work was done at City University of Hong Kong. Supported in part by the NSERC Operating Grant OGP0046506, ITRC, and a CGAT grant. Address: Department of Computer Science, University of Waterloo, Waterloo, Ont. N2L 3G1, Canada. E-mail: mli@math.uwaterloo.ca

[‡]BioInformatics Center & Institute of Systems Science, National University of Singapore, Singapore 119597. Email: lxzhang@iss.nus.sg

structing a global species tree is based on the parsimonious criterion of minimizing the duplication or mutation cost between the gene trees and the species tree. Such a problem was investigated by Guigó *et al.*([12]) under the duplication cost. In their paper, they developed a heuristic method for the problem using a nearest neighbor interchange searching algorithm and applied it to infer a most likely phylogenetic relationship among 16 major higher eukaryotic taxa from the sequences of 53 different genes.

Our main contribution has three aspects. First, we study the properties of the l.c.a. mapping as well as the duplication and mutation costs. In particular, we prove a less obvious fact that the duplication cost satisfies the triangle inequality(Lemma 4.1). Second, the complexity of reconstructing an optimal species tree from gene trees is investigated. We prove that the problem is NP-complete under both the duplication and mutation costs. The concept of a reconciled tree was introduced by Goodman *et al.*([11]) and formalized by Page([21]) as a means of describing historical associations including genes and species. We also prove that finding the best reconciled tree for gene trees is NP-complete. These results may justify the necessary of developing heuristic methods for reconstructing species trees such as one proposed by Guigó *et al.*([12]) and the necessary of experimental research conducted by Page and Charleston([22]). Third, we define a new metric for measuring the similarity/dissimilarity between two trees with same uniquely labelled leaves. A disadvantage of the duplication cost is its asymmetric property. Because of this, a new metric satisfying the metric axioms is proposed. Like the mutation cost, the new metric is efficiently computable. Furthermore, under this new metric, we prove that the problem of reconstructing a species tree from gene trees can be approximated within constant factor 2 in polynomial time.

2 Comparing gene and species trees - duplications and losses

In this section we briefly define gene and species trees, and introduce two duplication-based measures for comparing gene and species trees. For their biological meaning, we refer the reader to [11], [21], and [15]. We also refer the reader to Garey and Johnson's book [10] for NP-completeness and approximation algorithms.

2.1 Species trees and gene trees. For a set I of N biological taxa, the model for their evolutionary history is a full, rooted binary tree T with N leaves each labeled by a distinct taxon in I, in which each internal node has exactly two children. Such a tree is usually called a *species tree*. In a species tree, any internal node denotes an ancestor of its subordinate species represented by leaves below it and are considered as a subset (called *cluster*) of the taxa set I. Thus, the evolutionary relation "m is a descendant of n" is expressed, in set-theoretic setting, just as "$m \subset n$", where we use the strict inclusion, in contrast to notation $m \subseteq n$, which allows the equality of m and n.

The model for gene relationship is a full, rooted binary tree with labelled leaves. Usually, a gene tree is constructed from a selection of genes each appearing in the studied species. For example, the gene family of hemoglobin genes in vertebrates contains α-hemoglobin and β-hemoglobin. A gene tree based on these two genes is illustrated in Figure 1 for human, chimp and horse([3]). Note that the labels in a gene tree may not be unique. Hence, an internal node g corresponds to a multiset $M_g = \{x_1^{i_1}, x_2^{i_2}, \cdots, x_m^{i_m}\}$, where i_j is the number of its subordinate leaves labelled with x_j. The cluster of g is just the set

$$S_g = \{x_1, x_2, \cdots, x_m\}.$$

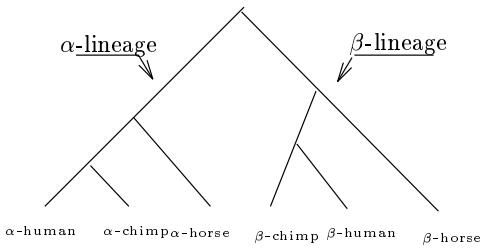


Figure 1: *A gene tree based on α-hemoglobin and β-hemoglobin.*

2.2 Gene duplications and losses. Given a gene tree G and a species tree S such that $L(G) \subseteq L(S)$. For any node $g \in G$, we define $M(g)$ be the node of S being its least common ancestors, that is, the smallest cluster containing the cluster of g, S_g. This correspondence M, first considered by Goodman *et al.* ([11]), is referred as a mapping of G into S by Page ([21]). We call M the *l.c.a. mapping* from G to S. Obviously, if $g' \subset g$, then $M(g') \subseteq M(g)$, and any leaf is mapped onto a leaf with the same label. For an internal node g, we use $c(g)$ to denote a child of g. Note that each internal node g has exactly two children.

DEFINITION 2.1. *Let g be an internal node of G. $G(g)$ and $S(M(g))$ are root-inconsistent if $M(c(g)) = M(g)$ for some child $c(g)$ of g.*

If $G(g)$ and $S(M(g))$ is root-inconsistent, a *duplication* is said to happen at g. The total number

$t_{dup}(G,S)$ of duplications happening in G under the l.c.a. mapping M is proposed as a measure of the similarity/dissimilarity of the gene tree G and the species tree S ([11, 21]). We call such a measure the *duplication cost*. Now we list two properties of this measure, which will be used later. Their proofs are easy and so are omitted.

PROPOSITION 2.1. *Let G be a gene tree and S a species tree. Then, $t_{dup}(G,S) = 0$ if and only if G is identical to $S|_{L(G)}$.*

PROPOSITION 2.2. *Let g be the root of G with children $a(g)$ and $b(g)$ and let s the root of S with children $a(s)$ and $b(s)$. Then, if a duplication happens at g under the l.c.a. mapping from G to S, then, $t_{dup}(G,S) = 1 + t_{dup}(a(G),S) + t_{dup}(b(G),S)$.*

Furthermore, the duplication cost also satisfies the triangle inequality, which is proved in Lemma 4.1 in Section 4. Under the duplication cost, the problem of finding the 'best' species tree from a set of known gene trees can be formulated as the following minimization problem.

Optimal Species Tree I(OST I)
INSTANCE: Given n gene trees $G_1, G_2, \cdots, G_n$.
QUESTION: Find a species tree S with the minimum duplication cost $\sum_{i=1}^{n} t_{dup}(G_i,S)$.

A subset L of nodes in a species tree S is *disjoint* if $x \cap y = \phi$ for any $x, y \in L$. For a disjoint subset L in S, we denote by S' the smallest subtree of S containing L as its leaf set. The *homomorphic subtree* $S|_L$ of S induced by L is a tree obtained from S' by contracting all degree 2 nodes except for its root.

Now, we define the *number of gene losses* associated with the l.c.a. mapping M from G to S. Since $L(G) \subseteq L(S)$, $S|_{L(G)}$ is well defined and M induces a l.c.a. mapping M' from G to $S|L(G)$. Let g and g' be two nodes in $S|_{L(G)}$ such that $g \subseteq g'$. Define

$$d(g,g') = |\{h \in S|_{L(G)} \mid g \subset h \subset g'\}.$$

Let $a(g)$ and $b(g)$ denote the two children of g. The *number of losses* l_g associated to g is

$$l_g = \begin{cases} 0 & \text{if } M'(g) = M'(a(g)) = M'(b(g)); \\ d(a(g),g) + 1 & \text{if } M'(g) \subset M'(a(g)) \ \&\ M'(g) = M'(b(g)); \\ d(a(g),g) + d(b(g),g) & \text{if } M'(g) \subset M'(a(g)) \ \&\ M'(g) \subset M'(b(g)). \end{cases}$$

Note that our definition of $l(g)$ is a variant of one defined by Guigó, Muchnik and Smith([12]). The *mutation cost* is defined as the sum of t_{dup} and the total number of losses, $l(G,S) = \sum_{g \in G} l_g$. This measure turns out to be identical to a biological meaningful measure defined in Mirkin *et al.*([15]) when G have the same number of uniquely labelled leaves as S, which was proved in [3] and [29] independently (see also [4]). The problem of finding the 'best' species tree from a set of known gene trees under this measure is formulated as:

Optimal Species Tree II(OST II)
INSTANCE: Given n gene trees $G_1, G_2, \cdots, G_n$.
QUESTION: Find a species tree S with the minimum mutation cost $\sum_{i=1}^{n}(t_{dup}(G_i,S) + l(G,S))$.

2.3 Reconciled Trees. Let G be a gene tree and S a species tree. The *reconciled tree* $T_r(G,S)$ of G with respect to S is the smallest tree with labelled leaves such that

(1) It contains the only clusters of S,

(2) It contains G as a subtree, i.e. $T_r(G,S)|_{L(G)} = G$, and

(3) For two children $a(g)$ and $b(g)$ of $g \in T_r$, $C_{a(g)} \cap C_{b(g)} = \phi$, or $S_{a(g)} = S_{b(g)} = S_g$.

An efficient algorithm for computing a reconciled tree given a gene and species trees was presented in Page([21]). Reconstructing a species tree from a gene tree can be formulated as:

Optimal Species Tree III(OST III)
INSTANCE: Given a gene tree G.
QUESTION: Find a species tree S with the minimum duplication cost $t_{dup}(T_r(G,S),S)$.

3 NP-completeness of finding optimal species trees

3.1 Optimal Species Tree I. Given n trees $T_1, T_2, \cdots, T_n$, we use $L[T_1, T_2, \cdots, T_n]$ to denote the tree T shown in Figure 2. When T_i is a single labelled node, the resulting tree is obviously a line tree, in which each internal node has a leaf as one of its children.

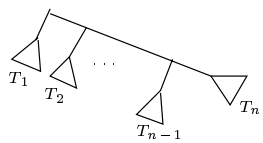


Figure 2: The tree $L[T_1, T_2, \cdots, T_n]$.

THEOREM 3.1. *The problem OST I is NP-complete.*

Proof. The problem is in NP. This is because there are exponential many species trees with leaves labelled with a given set of species and for each tree, the total

number of duplications can be easily calculated in linear time([29]).

To prove ite NP-completeness, we reduce the independent set problem to OST I. Assume that an instance $G = (V, E)$ of the independent set problem is given, where $V = \{v_1, v_2, \cdots, v_n\}$. We construct the corresponding instance of the problem OST I as follows.

Let $N = 7n^3$. For each v_i, we introduce N labels l_{ip}, $1 \leq p \leq N$ and a line tree $T_i = L[l_{i1}, l_{i2}, \cdots, l_{iN}]$. For each pair (i, j) $(1 \leq i \neq j \leq n)$ such that $(v_i, v_j) \in E$, we define two trees G_{1ij} and G_{2ij} with leaves labelled by $A = \{l_{ip} \mid 0 \leq i \leq n, 1 \leq p \leq N\}$ as shown in Figure 3 (a) and (b). Note that $G_{kij} \neq G_{kji}$ for $k = 1, 2$. Finally, for each i, we define a tree G_i with leaves labelled by A as shown in Figure 3 (c). Obviously, such a construction can be carried out in polynomial time. Hence, the NP-completeness of OST I derives from the following fact.

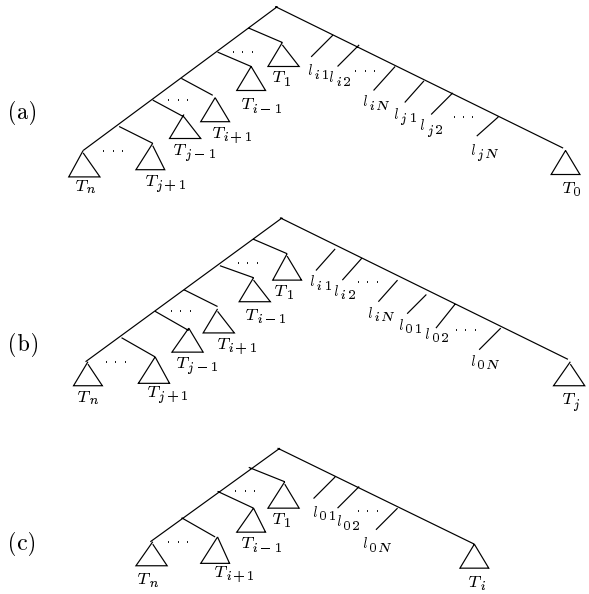


Figure 3: Gene trees defined in terms of edges and notes in the graph.

Claim The graph G contains an independent set of size C if and only if there is a special tree S for all the gene trees G_{1ij}, G_{2ij} and G_i, $1 \leq i, j \leq n$ with duplication cost $c < (4|E| + n - C + \frac{1}{2})N$.

Proof. $(\Rightarrow)$ Assume G contains an independent set K of size C. Without loss of generality, we assume $V(K) = \{v_1, v_2, \cdots, v_C\}$. Then, we define a species tree S as

$$S = L[l_{n1}, \cdots, l_n, \cdots, l_{(C+1)1}, \cdots, l_{(C+1)N}, l_{01}, \cdots, l_{0N}, l_{C1}, \cdots, l_{CN}, \cdots, l_{11}, \cdots, l_{1N}].$$

For each $i \leq C$, $c(G_i, S) = n - 1$. For each $i > C$, $c(G_i, S) = N + n - 1$. Further, we can verify that for any $i < j$ such that $(v_i, v_j) \in E$ and such that $j > C$,

$$\begin{aligned} &4(N + C) \leq \textstyle\sum_{k=1}^{2}(c(G_{kij}, S) + c(G_{kji}, S)) \\ &\leq 4(N + C + 1). \end{aligned} \tag{3.1}$$

Thus, the duplication cost c is

$$\begin{aligned} &\textstyle\sum_{(v_i,v_j)\in E}(\sum_{k=1}^{2}(c(G_{kij}, S) + c(G_{kji}, S))) \\ &+ \textstyle\sum_{1\leq i\leq n} c(G_i, S) \\ &\leq (4|E| + n - C)N + 3n^3 \\ &< (4|E| + n - C + \tfrac{1}{2})N. \end{aligned}$$

$(\Leftarrow)$ We prove it by contradiction. Suppose that the optimal duplication cost is c for G_{1ij}, G_{2ij} and G_i, $1 \leq i, j \leq n$. Let $A_i = \{l_{ip} \mid 1 \leq p \leq N\}$.

Fact 1. There is an optimal species tree S such that $S|_{A_i} = T_i$ for every $i \in [0, n]$.

Proof. Assume that S is any optimal species tree. For any i, we use $lca(A_i)$ to denote the least common ancestor of the leaves of A_i in the tree S. Let $lca(A_i) = p \in S$. If $S|_{A_i} = S(p)|_{A_i} \neq T_i$, then there is a subtree, say T, in $S(p)$ such that each of two subtrees $a(r(T))$ and $b(r(T))$ contains at least two labelled leaves in A_i. Let subtree $a(r(T))$ contain k such labelled leaves $l_{ij_1}, l_{ij_2}, \cdots, l_{ij_k}$, where $2 \leq k \leq N - 2$. Without loss of generality, we may assume that for any k' and k'' such that $k' < k''$, either $p(l_{ij_{k'}})$ and $p(l_{ij_{k''}})$ are disjoint, or $p(l_{ij_{k'}}) \subseteq p(l_{ij_{k''}})$. We construct a tree S'' from S by replacing the subtree T with $L[a(r(T))|_{A-A_i}, l_{ij_k}, \cdots, l_{ij_1}, b(r(T))]$. By definition of gene trees, we can verify that the duplication cost c'' of S'' is at most c. Since S is optimal, we have that $c'' = c$ and so S'' is also optimal. Applying above procedure repeatedly, we will finally obtain a desired optimal species tree. This concludes the proof of the fact.

Let S be an optimal species tree satisfying Fact 1. Then the inclusion relationship $\subseteq$ among $lca(A_i)$ in S can be extended into a total order $\prec$ such that for any i and j, $lca(A_i) \prec lca(A_j)$ if $lca(A_i) \subseteq lca(A_j)$ in S. Let $lca(A_{i_n}) \prec lca(A_{i_{n-1}}) \prec \cdots \prec lca(A_{i_0})$. Then, we define a line tree S' as

$$S' = L[l_{i_0 1}, \cdots, l_{i_0 N}, l_{i_1 1}, \cdots, l_{i_1 N}, \cdots, l_{i_n 1}, \cdots, l_{i_n N}].$$

Let S' have duplication cost c'. We have the following two facts.

Fact 2. $c' \leq 3n^3 + c$

Proof. Since $S'|_{A_i} = S|_{A_i} = T_i$, no duplications happen at all subtrees T_i $(0 \leq i \leq n)$ in each gene tree $G_{1i'j'}$, $G_{2i'j'}$ and $G_{i'}$. On the other hand, since S' and S have the same inclusion relationship $\subset$ among all

$lca(A_i)(0 \le i \le n)$, the duplication cost on all the right subtrees of gene trees are same. Note that there are at most $n' = n^2 + 2n(n-1)(n-2)$ other vertices that have not been considered above. We have that $c' \le n' \le 3n^3$. This finishes the proof of Fact 2.

Fact 3. $c' > (4|E| + n - C + 1)N$.

Proof. Let $E_< = \{(v_i, v_j) \in E \mid p_{S'}(l_{i1}), p_{S'}(l_{j1}) \subset p_{S'}(l_{01})\}$ and $V_< = \{v_i \in V | p_{S'}(l_{i1}) \subset p_{S'}(l_{01})\}$. If $G = (V, E)$ does not contain an independent set of size C. Then, $|E_<| + C - |V_<| \ge 1$. In fact, this is trivial if $|V_<| < C$. Otherwise, let the restriction subgraph $G|_{V_<}$ have a largest independent set K'. Then, $|K'| \le C - 1$. Since K' is largest, for any node $v \in V - K'$, $(v, v') \in E$ for some $v' \in K'$. This implies that $|E_<| \ge |V_<| - |K'| \ge |V_<| - C + 1$, i.e., $|E_<| + C - |V_<| \ge 1$ when $|V_<| \ge C$.

It is easy to verify that, for any i, j such that $(v_i, v_j) \in E$ and such that $lca(\{l_{i1}\}), lca(\{l_{j1}\}) \subset lca(\{l_{01}\})$ in S',

$$\sum_{k=1}^{2} c(G_{kij}, S') = 6N + 4(|V_<| - 1). \tag{3.2}$$

Hence, by Formula (3.1) and (3.2), we have

$$\begin{aligned} c' &= \textstyle\sum_{v_i \in V - V_<} c(G_i, S) + \sum_{v_i \in V_<} c(G_i, S) \\ &+ \textstyle\sum_{(v_i, v_j) \notin V} (\sum_k^2 c(G_{kij}, S') + \sum_k^2 c(G_{kji}, S')) \\ &\ge (4|E| + n - C + 1)N. \end{aligned}$$

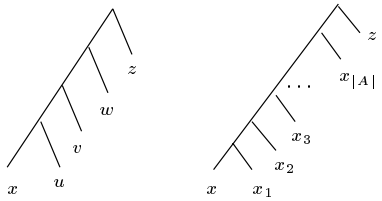
Then, Fact 3 is proved.

Combining Fact 2 and Fact 3, we have that $c > (4|E| + n - C + 1/2)N$, a contradiction. Thus, we finish the proof of Claim and so Theorem 3.1.

Remark. We have actually proved that OST I is NP-complete even for all gene trees with the same uniquely labelled leaves. Such a stronger conclusion will be used to prove that OST III is NP-complete in Section 2.3.

3.2 Optimal Species Tree II. Let C be a set of full binary trees G with leaves uniquely labelled by $L(G)$, and let T be a full binary tree with leaves uniquely labelled by $\sum_{G \in S} L(G)$. We say that C is *compatible* with T if for every $G \in S$, the homomorphic subtree $T|_{L(G)|}$ of T induced by $L(G)$ is G. It is *compatible* if it is compatible with some tree with leaves labelled by $\sum_{G \in S} L(G)$. Finally, recall that $L[z, w, v, u, x]$ denotes a rooted line tree with 5 leaves z, w, v, u, x as shown in Figure 4 (a).

LEMMA 3.1. *If a collection C of 5-leave rooted line trees $L[y, w_i, v_i, u_i, x]$ is compatible, then it is compatible with a rooted line tree $L[y, x_n, x_{n-1}, \cdots x_1, x]$, where $\{x_1, x_2, x_3, \cdots, x_n\} = \cup\{u_i, v_i, w_i\}$.*



(a) (b)

Figure 4: Rooted line trees.

Proof. Choose a label z not in $\{x, y\}$ and $\cup_i\{u_i, v_i, w_i\}$. For each $t = L[y, w_i, v_i, u_i, x]$, we add an edge between z and the root so that the resulting tree t^z is an unrooted, full binary tree in which each internal node has degree-3. It is not difficult to see that t^z is defined by the following set of quartets(see [25]):

$$Q(t^z) = \{xu_i|v_iz, xv_i|w_iz, xu_i|yz, xv_i|yz, xw_i|yz\}.$$

Suppose C is compatible with a rooted, full binary tree T, then, $C^z = \{t^z \mid t \in C\}$ is compatible with T^z, and thus quartet set $\cup_{t \in C} Q(t^z)$ is compatible with T^z. By a lemma in [25], $\cup_{t \in C} Q(t^z)$ is compatible with an xz-caterpillar $x|u_1u_2 \cdots u_{|A|}y|z$. This implies that C is compatible with the binary tree rooted at the internal node that is jointed with z (after the removal of z), which has the form shown in Figure 4 (b).

THEOREM 3.2. *The problem OST II is NP-complete.*

Proof. The problem is obviously in NP as the problem OST I. To prove its NP-completeness, we now describe a transformation from the cyclic ordering problem([10]):

> **Instance**: A finite set A, and a collection C of ordered triples (a, b, c) of distinct elements from A.
>
> **Question**: Is there a one-to-one function $f : A \to \{1, 2, \cdots, |A|\}$ such that, for each $(a, b, c) \in C$, we have either $f(a) < f(b) < f(c)$ or $f(b) < f(c) < f(a)$ or $f(c) < f(a) < f(b)$?

which is proved to be NP-complete by Galil and Megiddo in [9].

Suppose an instance of the cyclic ordering problem is given. We construct for each ordered triple $\pi = (a, b, c) \in C$ three gene trees $G_1^\pi = L[y, c, b, a, x]$, $G_2^\pi = L[y, a, c, b, x]$ and $G_3^\pi = L[y, b, a, c, x]$ as shown in Figure 5, where x and y are two new labels fixed for all triples in C. Now, we consider a collection $G(C) = \{G_i^\pi \mid 1 \le i \le 3, \pi \in C\}$ of $3|C|$ gene trees. Obviously, such a construction can be carried out in polynomial time.

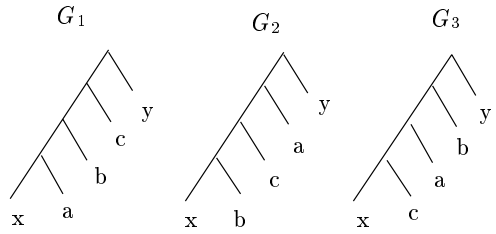


Figure 5: Three trees correspond to an ordered triple (a, b, c).

We claim that there is a species tree with leaves $A \cup \{x, y\}$ having the mutation cost at most $14|C|$ if and only if A has a cyclic ordering.

Suppose a cyclic ordering f exists. Let $f(i)$ denote the ith smallest element in A and let $S = L[y, f(|A|), \cdots, f(2), f(1), x]$(see Figure 6).

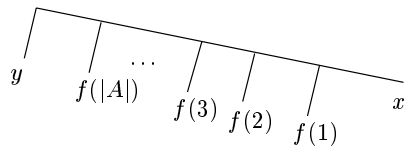


Figure 6: The species tree constructed from a cyclic ordering f.

For a triple $\pi = (a, b, c) \in C$, without loss of generality, we may assume that $f(a) < f(b) < f(c)$. Then G_1^π is the homomorphic subtree of S on $\{x, a, b, c, y\}$. Thus, $c(G_1, S) = 0$, $c(G_2, S) = 5$ and $c(G_3, S) = 9$. Hence, the total mutation cost over all $3|C|$ gene trees is $14|C|$.

Conversely, suppose that T is a species tree with leaves $A \cup \{x, y\}$ having the mutation cost at most $14|C|$. Then we have

Claim For any $\pi = (a, b, c) \in C$, the homomorphic subtree of T on $\{x, a, b, c, y\}$ is G_1, G_2 or G_3 as shown in Figure 5.

Proof The homomorphic subtree T' of T on $\{x, a, b, c, y\}$ is a full, binary tree with five labeled leaves. Assume it is not any of G_1^π, G_2^π and G_3^π. All possible homomorphic subtrees are illustrated in Figure 7 and Figure 8 and the case-by-case analysis of the mutation cost of G_1, G_2 and G_3 with T is shown in Table 1.

Hence, T has the mutation cost at least $14|C| + 1$. This is a contradiction. This finishes the proof of Claim.

By Lemma 3.1, there exists a line tree such that for each triple $\pi = (a, b, c)$, the homomorphic subtree on $\{x, y, a, b, c\}$ is one of the gene trees $G_1^\pi, G_2^\pi, G_3^\pi$. It is not difficult to see that such a line tree inducing a cyclic ordering. This concludes the proof of Theorem 3.2.

Cases	T_i	(a)	(b)	(c)	(d)
Cost	17	18	27	45	29,32
Cases	(e)	(f)	(g)	(h)	(i)
Cost	31,34	32,35	37	34	37
Cases	(j)	(k)	(l)	(m)	
Cost	26,29	20	33	29,32	

Table 1: Case-by-case analysis of duplications.

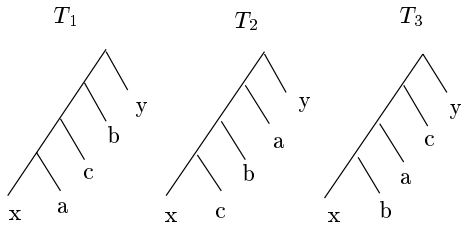


Figure 7: Three trees in the first column in Table 1.

3.3 Optimal Species Tree III. First, we have the following property, which is derived from the definition of reconciled trees.

LEMMA 3.2. *Given a gene tree G and a species tree S. Let T_r be the reconciled tree of G with respect to S and g be an internal node in G. If g is mapped to $t \in T_r$ under the l.c.a mapping. Then, $T_r(t)$ is the reconciled tree of $G(g)$ with respect to $S(t)$.*

LEMMA 3.3. *Let T_r be the reconciled tree of G with respect to S. Then, $t_{dup}(T_r, S) = t_{dup}(G, S)$.*

Proof. We prove this by induction on the number of leaves in S. It is obviously true for a species tree having only three leaves. Now assume that S has at least 4 leaves. Let t be the root of T_r with children $a(t)$ and $b(t)$, let g be the root of G with children $a(g)$ and $b(g)$ and let s be the root of S with children $a(s)$ and $b(s)$. We consider the following cases.

Case 1. $a(t) \cap b(t) = \phi$.

Since G is identical to $T_r|_{L(G)}$, under the l.c.a. mapping from G to T_r, $a(g)$ is mapped to a node $t_1 \subseteq a(t)$, and $b(g)$ to a node $t_2 \subseteq b(t)$. Note that t_1 and t_2 are also two clusters in S. For simplicity, we still use t_1 and t_2 to denote such two corresponding nodes. By Lemma 3.2, $T_r(t_1) = T_r(G(a(g)), S(t_1))$ and $T_r(t_2) = T_r(G(b(g)), S(t_2))$. By induction, $t_{dup}(T_r(t_1), S(t_1)) = t_{dup}(G(a(g)), S(t_1))$ and $t_{dup}(T_r(t_2), S(t_2)) = t_{dup}(G(b(g)), S(t_2))$. Since

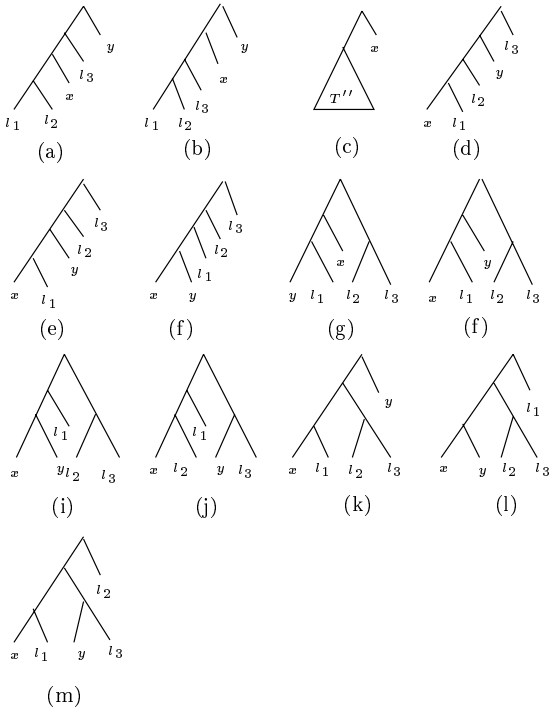


Figure 8: Cases 2-14 in the proof of Claim 1.

$t_1 \subseteq a(t)$ and $t_2 \subseteq b(t)$, g is not a duplication node under both l.c.a. mappings from G to T_r and to S respectively. Thus,

$$\begin{aligned} & t_{dup}(T_r(G,S),S) \\ &= t_{dup}(T_r(a(t)),S(a(s))) + t_{dup}(T_r(b(t)),S(b(s))) \\ &= t_{dup}(G(a(g)),S(a(s))) + t_{dup}(G(b(g)),S(b(s))) \\ &= t_{dup}(G,S). \end{aligned}$$

Case 2. $a(t) = b(t)$.

Then, by definition, $a(t) = b(t) = t$. Furthermore, either $a(g)$ is mapped to $a(t)$ or $b(g)$ is mapped to $b(t)$. Without loss of generality, we may assume that the former is true. Let $b(g)$ be mapped to t'. Note that $t' \subseteq b(t), s$. Under the l.c.a. mapping from G to S, $a(g)$ is mapped to s, the root of S. Thus, by induction,

$$\begin{aligned} & t_{dup}(T_r,S) \\ &= 1 + t_{dup}(T_r(a(t)),S) + t_{dup}(T_r(b(t)),S) \\ &= 1 + t_{dup}(a(g),S) + t_{dup}(G(b(t)),S(t')) \\ &= t_{dup}(G,S). \end{aligned}$$

This proves Lemma 3.3.

Therefore, the problem OST III is a special case of the problem OST I in which each instance has only one gene tree. Unfortunately, such a problem is still NP-complete.

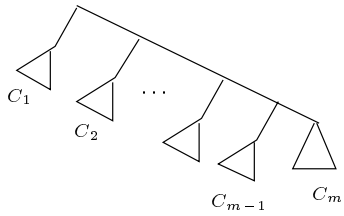


Figure 9: *Connection of m gene trees in a right line tree.*

THEOREM 3.3. *The problem OST III is NP-complete.*

Proof. Obviously, such a problem is in NP. Now we ptove its NP-completeness. By Lemma 3.3, we need only to prove the following problem to be NP-complete:

> Given a gene tree, find a species tree S with the minimum duplication cost $t_{dup}(G,S)$.

Given a class C of m gene trees with the same n labelled leaves, we construct a gene G by connecting all the gene trees in C through a right line trees as shown in Figure 9. Since all the gene trees in C have the same labelled leaves, we have that for any species tree S,

$$t_{dup}(G,S) = m - 1 + \sum_{1 \le i \le m} t_{dup}(G_i,S).$$

This finishes the reduction from an NP-complete problem to the problem given above(see the remark after Theorem 3.1).

4 A New Metric

In this section, we introduce a new metric for arbitrary full trees based on the concept of duplications. Note that in a full tree, each internal node has degree ≥ 3.

4.1 Definition. Given two full trees T_1 and T_2, we define the l.c.a. mapping M from T_1 and T_2 as before and we say a duplication happens at $n \in T_1$ under M if and only if for some child $c(n)$ of n such that $M(c(n)) = M(n)$. We still use $t_{dup}(T_1,T_2)$ to denote the number of duplications between T_1 and T_2.

Let T be a full tree. For any internal edge $e = (u,v)$, the *contraction tree* of T at e is the resulting tree after the removal of e and combining u and v into a new node p such that p is adjacent to all the adjacencies of both u and v.

LEMMA 4.1. *The duplication cost satisfies the triangle inequality, i.e., $t_{dup}(T_1,T_3) \leq t_{dup}(T_1,T_2) + t_{dup}(T_2,T_3)$. for any three full trees T_1, T_2 and T_3 with same uniquely labeled leaves.*

Proof. Let M_{ij} denote the l.c.a. mapping from T_i to T_j. Now let T_1' be the resulting tree from T_1 by contracting all edges (u, v) such that $M_{12}(u) = M_{12}(v)$. Then, there is no duplications between T_1' and T_2. Furthermore, let M_{12}' be the mapping from T_1' to T_2, we have the following claim.

Claim 1. For any $m \in T_1'$, $M_{12}'(m) = m$. Thus, $t_{dup}(T_1', T_2) = 0$.

Proof. This can be proved by induction.

Claim 2. $t_{dup}(T_1, T_3) \leq t_{dup}(T_1, T_2) + t_{dup}(T_2, T_3)$ if $t_{dup}(T_1', T_3) \leq t_{dup}(T_2, T_3)$.

Proof. Under the mapping M_{13}, a duplication happens at a node $n \in T_1$ if and only if $M_{13}(n) = M_{13}(c(n))$ for some child $c(n)$ of n. Let D denote the set of such duplication nodes in T_1 under M_{13}. We divide D into two disjoint subsets:

$$D_1 = \{n \in D \mid M_{12}(n) = M_{12}(c(n))\},$$

and

$$D_2 = \{n \in D \mid M_{12}(n) \neq M_{12}(c(n))\}.$$

Obviously, $|D_1| \leq t_{dup}(T_1, T_2)$. By definition, $t_{dup}(T_1, T_3) = |D_1| + |D_2| \leq t_{dup}(T_1, T_2) + t_{dup}(T_1', T_3) \leq t_{dup}(T_1, T_2) + t_{dup}(T_2, T_3)$ if $t_{dup}(T_1', T_3) \leq t_{dup}(T_2, T_3)$.

Let $M_{12}'(n) = p$ and $M_{12}'(c(n)) = q$. Then, by Claim 1, $n = p$ and $c(n) = q$. If $M_{13}(n) = M_{13}(c(n))$, then all nodes in the path from $M_{23}(p)$ and $M_{23}(q)$ is mapped to the same node in M_3. This implies that $t_{dup}(T_1', T_3) \leq t_{dup}(T_2, T_3)$. This finishes the proof of Lemm 4.1.

Now we define a new similarity/dissimilarity measure between two full trees as

$$d(T_1, T_2) = \frac{t_{dup}(T_1, T_2) + t_{dup}(T_2, T_1)}{2}.$$

Since the duplication cost is computable in linear time, the measure $d(.,.)$ is also efficiently computable. Further, it satisfies the three metric axioms.

PROPOSITION 4.1. *For any three full trees T_1, T_2 and T_3, $d(.,.)$ satisfies the following properties:*

(1) $d(T_1, T_2) = 0$ if and only if $T_1 = T_2$;

(2) $d(T_1, T_3 \leq d(T_1, T_2) + d(T_2, T_3)$;

(3) $d(T_1, T_2) = d(T_2, T_1)$.

In what follows, we call $d(.,.)$ the *symmetric duplication cost*. Interestingly enough, the symmetric duplication cost is closely related to the *nearest neighbor interchange*(nni) distance, which was introduced independently in [17] and [24]. An nni operation swaps two subtrees that are separated by an internal edge (u, v) as show in Figure 10. The *nni distance*, $D_{nni}(T_1, T_2)$, between two full trees T_1 and T_2 is defined as the minimum number of nni operations required to transform one tree into the other.

PROPOSITION 4.2. *For any species trees T_1 and T_2, $d(T_1, T_2) \leq D_{nni}(T_1, T_2)$.*

Proof. Suppose T_1 is converted into T_2 by one nni operation. Then, we can easily verify that $d(T_1, T_2) = 1$. Thus, $d(T_1, T_2) \leq D_{nni}(T_1, T_2)$. Since $d(.,.)$ satisfies the triangle inequality, the result hold in general also.

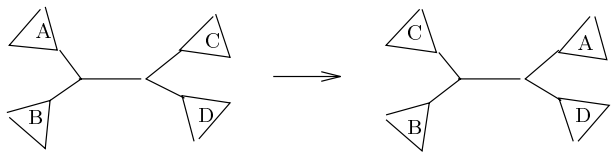


Figure 10: A possible nni operation on an internal edge (u, v): exchange A and C.

Although it is unknown whether the problem of finding an optimal species tree from gene trees is NP-complete with the cost $d(.,.)$ or not, we have the following approximation result.

THEOREM 4.1. *There is a polynomial-time approximation of ratio 2 to the problem of finding an optimal species from gene trees with the symmetric duplication cost $d(.,.)$.*

Proof. Given an input of n gene trees $G_1, G_2, \cdots, G_n$, we compute $\sum_{i \neq j}^{n} d(T_i, T_j)$ for each $j \leq n$ and output G_j with the minimum cost $\sum_{i \neq j}^{n} d(T_i, T_j)$ as the species tree. We now prove that the output species tree has at most two times the optimal cost. Assume that G_1 is the output and S is an optimal species tree. Then,

$$\begin{aligned} &\textstyle\sum_{i \leq n} d(T_i, T_1) \leq (\sum_{i \leq n} \sum_{j \leq n} d(T_i, T_j))/n \\ &\leq (\textstyle\sum_{i \leq n} \sum_{j \leq n} (d(T_i, S) + d(T_j, S)))/n \\ &\leq 2 \textstyle\sum_{i \leq n} d(T_i, S). \end{aligned}$$

This proves Theorem 4.1.

5 A general problem

We have studied the properties of the duplication and mutation costs, and the computational complexity of reconstructing a global species tree from gene trees. We have proved that various versions of the problem are NP-complete. As a consequence it is unlikely that there is an efficient algorithm for these problems. However, our complexity results are the start point for the development of good approximation, heuristic algorithms and methods specific to the type of given data. This is an area we are currently investigating.

Furthermore, a general problem may be more interesting. There are a large family of genes each having several, distinct copies in the studied species. In order

to derive a gene tree that truly reflects the evolution of species, one needs knowledge about which copies of the gene are comparable. This is usually impossible untill careful study of the species. However, one may have confidence to a certain degree in different gene trees. Hence, it is natural to propose the following problem. We use I^+ to denote the set of integer numbers and let m be any similarity/dissimilarity measure between gene and species trees.

General Optimal Species Tree(GOST)
INSTANCE: A set of n gene trees $G_1, G_2, \cdots, G_n$, to each tree a confidence value $c_i \in I^+$ is associated.
QUESTION: Find a species tree S with the minimum cost $\sum c_i m(G_i, S)$.

Clearly, GOST is NP-complete under the duplication cost and the mutation cost. To the nni distance, the conclusion is also true.

THEOREM 5.1. *The problem GOST is NP-complete for the NNI distance.*

Sketch of Proof. We reduce the problem of computing nni distance between two trees(see [2] for its NP-completeness) to GOST. Given two binary trees T_1 and T_2 with n leaves. By applying an nni operation to T_1, we may obtain as many as $2n-2$ different resulting trees. Let T_3 be such a tree, i.e., $d_{nni}(T_3, T_1) = 1$. We consider the following instance I of GOST:

$$I = \{T_1, T_2, T_3, c_1 = 2, c_2 = 2, c_3 = 1\}.$$

Let S be an optimal species tree for I. Then one can easily verify that $S = T_3$ if and only if $d_{nni}(T_1, T_2) = d_{nni}(T_1, T_3) + d_{nni}(T_3, T_2)$. Note that the nni distance $d_{nni}(T_1, T_2)$ is at most $n \log n$. If GOST is solved in polynomial time, we can compute $d_{nni}(T_1, T_2)$ using an efficient search. For each T_3 such that $d_{nni}(T_1, T_3) = 1$, compute the optimal species tree S for the instance I defined above. If $S = T_3$, then compute $d_{nni}(T_3, T_2)$ recursively and output $1 + d_{nni}(T_3, T_2)$. This finishes the reduction and so the proof.

Note that Theorem 4.1 can not be generalized to the problem GOST. Therefore, it is challenging to develop polynomial algorithms with constant approximation factor for GOST for the various measures studied here.

References

[1] E.N. III Adams, N-trees as nestings: complexity, similarity and consensus, *J. Classification* 3(1986), 299-317.

[2] B. DasGupta, X. He, T. Jiang, M. Li, J. Tromp and L. Zhang. On distance between phylogenetic trees. In *Proc. of the 8th SODA*, 427-436, 1997.

[3] O. Eulenstein and M. Vingron, On the equivalence of two tree mapping measures, *Arbeitspapiere der GMD*, 936, Bonn, Germany.

[4] Eulenstein, Mirkin and Vingron, Duplication-based measures of difference between gene and species trees, *Submitted for publication.*

[5] M. Farach and M. Thorup, Fast comparison of evolutionary trees, In *Proc. of the 5th SODA*, 481-488, 1994.

[6] J. Felsenstein, Phylogenies from molecular sequence: Inference and reliability, *Ann. Review Genet.* 22(1988), 521-561.

[7] W. Fitch, Distinguishing homologous and analogous proteins, *Syst. Zool.* 19(1970), 99-113.

[8] W. Fitch and E. Margoliash, Construction of phylogenetic trees, *Science* 155(1967), 279-284.

[9] Z. Galil and N. Megiddo, Cyclic ordering is NP-Complete, *Theoret. Comput. Sci.* 5(1977), 179-182.

[10] M. Garey and D. Johnson, *Computers and Intractability: A Guide to the Theory of NP-completeness*, W. H. Freeman, 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.* 28(1979), 132-163.

[12] R. Guigó, I. Muchnik and T. Smith, Reconstruction of ancient molecular phylogeny, *Molecular Phylogenetics and Evolution* 6(1996), No. 2, 189-213.

[13] M.D. Hendy, C.H.C. Little and D. Penny, Comparing trees with pendant vertices labeled, *SIAM J. Appl. Math.* 44(1984), 1054-1067.

[14] T. Margush and F. R. McMorris, Consensus n-Trees, *Bull. of Math. Biol.* 43(1981), 239-244.

[15] B. Mirkin, I. Muchnik and T. Smith, A biologically meaningful model for comparing molecular phylogenies, *J. Comput. Biology* 2(1995), 493-507.

[16] B. Mirkin and S. N. Rodin, *Graphs and Genes*, Springer-Verlag, Bonn, Germany, 1984.

[17] G. W. Moore, M. Goodman and J. Barnabas, An iterative approach from the standpoint of the additive hypothesis to the dendrogram problem posed by molecular data sets, *J. Theoret. Biol.* 38(1973), 423-457.

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

[19] J. E. Neigel and J. C. Avise, Phylogenetic relationship of mitochondrial DNA under various demographic models of speciation, *Evolutionary Processes and Theory*, 515-534, Academic Press, New York, 1986.

[20] S. Ohno, *Evolution by gene duplication.* Springer-Verlag, Berlin, 1970.

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

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

[23] P. Pamilo and M. Nei, Relationship between gene trees

and species trees. *Mol. Bio. Evol.* 5 (1988), 568-583.

[24] D. F. Robinson, Comparison of labeled trees with valency trees, *J. Combin. Theory*, Series B, 11(1971), 105-119.

[25] M. Steel, The complexity of reconstructing trees from qualitative characters and subtrees, *J. Classifications* 9(1992), 91-116.

[26] N. Takahata, Gene genealogy in three related population: Consistency probability between gene and population trees, *Genetics* 122(1989), 957-966.

[27] M. Waterman and T. Smith, On the similarity of dendrograms, *J. Theoret. Biol.* 73(1978), 789-800.

[28] C.-I. Wu, Inference of species phylogeny in relation to segregation of ancient polymorphisms, *Genetics* 127(1991), 429-435.

[29] L. Zhang, On a Mirkin-Muchnik-Smith conjecture for comparing molecular phylogenies, *J. of Comput. Biology* 4(1997), 177-188.