On a Mirkin-Muchnik-Smith Conjecture for Comparing Molecular Phylogenies

Louxin Zhang
lxzhang@iss.nus.sg
BioInformatics Center
Institute of Systems Science
Heng Mui Keng Terrace
Singapore 119597

Abstract

A conjecture of Mirkin, Muchnik and Smith is answered affirmatively which connects the inconsistency function, a biologically meaningful dissimilarity measure for a gene and species tree, to the mutation cost function, a combinatorial measure based on mapping of trees. A linear-time algorithm for computing the inconsistency function is also derived from the conjecture.

1 Introduction

As DNA sequences have become easier to obtain, interesting emphasis has been placed on constructing gene trees and from these, reconstructing evolutionary trees for species. Because of the presence of paralogy and sorting of ancestral polymorphism, gene trees and species trees are often inconsistent (e.g., Neigel and Avise, 1986; Pamilo and Nei, 1988; Takahata, 1989; Wu, 1991). Therefore, a major concern that arises is how to combine different, sometimes contradictory, gene trees into an evolutionary tree called species tree (Fitch, 1970; Goodman *et al.*, 1979; Nei, 1987). Several ideas have been suggested for the last twenty years (see, for example, Robinson 1971; Waterman and Smith, 1978; Margush and McMorris, 1981; Hendy *et al.*, 1984; Adams, 1986; Barthélemy, 1986). A common characterization of these ideas is that they consider phylogenetic trees as formal mathematical objects and proposed the similarity/dissimilarity measures based only on combinatorial consideration. The weakness of these measures is lack of biological meaning (Mirkin *et al.*, 1995) and computable intractability in general (see, for example, DasGupta *et al.*, 1996).

Biologically meaningful similarity/dissimilarity measures are also addressed. Since all contradictions among different gene trees are resulted from the gene divergence, the divergence should be presented and explained in the combined species tree (Fitch, 1970; Goodman *et al.*, 1979). The gene divergence can be the results of either speciation or duplication (Ohno, 1970). The speciation happens between species. If the gene divergence occurs with only speciation, the gene and species trees are identical. But, the duplication event happens within species. When duplications occur, the gene and species trees might be inconsistent. If the common ancestry of two genes can be tracted back to a speciation event, then they are said to be related by *orthology* (Fitch, 1970); if it is tracked back to a duplication event, then they are related by *paralogy.* Taking into account orthology and paralogy evolutions, Goodman *et al.* (1979) proposed a new similarity/dissimilarity measure for annotating species tree with duplications, gene losses and the nucleotide replacements. Later, Guigó *et al.* (1994) elaborated the idea for identifying and locating the gene duplications in eukariotic history.

Guigó *et al.* (1994) introduced the mutation cost functions for measuring the dissimilarity between gene and species trees using so called mapping of the trees. The mapping of trees was considered implicitly in Goodman *et al.* (1979) and explicitly in Page (1994). As we shall see later, the concept of mapping and thus the mutation cost function is rather formal and technical. Therefore, the definitions of duplications and subsequent events based on mapping are not substantiated with any biologically meaningful model. This leads Mirkin *et al.* (1995) to propose a new, biological meaningful model for explaining and measuring the dissimilarity between a single gene and species tree. They first formalized the concepts of gene duplication, loss, information gap in graph-theoretic terms, and then proposed a procedure for measuring the dissimilarity between a gene and species tree by comparing every subtree of the gene tree with the corresponding subtree in the species tree. Whenever an inconsistency occurs, a duplication event is assumed to explain the inconsistency. This duplication is reflected in the species tree with the event's history leading to the current situation in species. The history along the tree involves certain gene copy losses which accompany the duplication during the evolutionary history. The total number of duplications, loss and information gap events involved in all the inconsistency is defined to be the inconsistency measure between the gene and species tree. In the same paper, after numerous testing, Mirkin *et al.* conjecture that the inconsistency measure coincides with the mutation cost function.

The major goal of our work is to prove that the conjecture is true (in Section 3.2). This may justify the use of the inconsistency measure in evolutionary tree reconstruction. Our work also provides an efficient algorithm for computing the inconsistency function (in Section 3.3).

2 A model based on gene duplications

In this section we briefly introduce the Mirkin-Muchnik-Smith model for comparing a gene species tree. For its biological meaning, the reader is referred to Mirkin *et al.* (1995).

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 element of I. Any internal nodes denote ancestors of the taxa in I and are considered as a subset (also called cluster) of its subordinate leaves. Thus, the evolutionary relation "m is a descendant of n" is expressed, in set-theoretic setting, just as "$m \subset n$", where we use strict inclusion, in contrast to notation $m \subseteq n$ allowing the equality of m and n.

Each internal node has two children, which are denoted by $a(n)$ and $b(n)$. If $n_1, n_2, \cdots n_k$

is a path connecting node n_1 and its descendant n_l, then $n_k \subset n_{k-1} \subset \cdots \subset n_1$, and any node m belongs to the path between n_1 and n_l if and only if $n_l \subset m \subset n_1$, and it is called an *intermediate* between n_1 and n_l.

A subset S of nodes of T is incompatible if $x \cap y = \phi$ for any $x, y \in S$. For an incompatible subset S in T, we denote by $T|S$ the smallest subtree T containing S as its leaf set. The *homomorphic subtree* of T induced by S is the subgraph obtained from $T|S$ by contracting all degree 2 nodes except for its root. These concepts are illustrated in Figure 1.

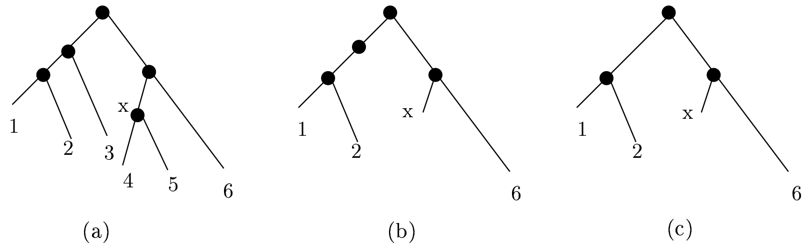

Figure 1: *(a) A phyolgenetic tree T; (b) The subtree $T|S$ for $S = \{1, 2, x, 6\}$; (c) The homomorphic subtree induced by S.*

In order to annotate duplication history into a phylogenetic tree, the concept of duplication is introduced in Mirkin *et al.* (1995).

Definition 2.1 *Let T be a phylogenetic tree with leaf set I representing a set of N species and let $L = \{\phi, +, -, +-\}$. A mapping $\delta : T \to L$ is called a gene duplication if (1) it is monotone, that is, $\delta(m) \subseteq \delta(n)$ when $m \subseteq n$, and, (2) it is saturated, that is, $\delta(n) = +-$ for the root n of T.*

A node $m \in T$ is called *mixed* if $\delta(m) = +-$, *speciated* if $\delta(m) = +$ or $-$, and *gapped* if $\delta(m) = \phi$. The maximal mixed, speciated/gapped nodes have particular evolutionary meaning: the maximal mixed node is the root, which represents the duplication event itself, the maximal speciated/gapped nodes correspond to the gene losses. The total number of these maximal nodes will be called the *complexity* of duplication δ, which counts for the total number of the evolutionary distinct events associated with the duplication[1].

2.1 Comparing a gene tree with a species tree

Let T be a species tree rooted at t with leaf set I and let G be a single gene tree rooted at g with leaf set J such that $J \subseteq I$. We also assume that in the gene tree G there is exactly a gene for each species. Therefore, there is an injection from the leaves of G to those of T. T and G will be said to be *root-consistent* if each of the sets of the leaf descendants, $a(g), b(g)$, of the children of the root g is contained in a child-set $a(t)$ or $b(t)$ of t, and *root-inconsistent* otherwise.

[1]In Mirkin *et al.* (1995), the complexity is differently defined as the total number of maximal duplications, maximal speciated and maximal gapped nodes. For so-called operational duplications, their definition and ours become identical. Main reason for modifying the definition is that under our definition, the conjecture is always true, not just for operational duplications.

Obviously, the root-consistent means that the 'root branches' of the tree G consistent with the 'root branches' of T; an event causing the divergence of the root descendants occurring in the species tree T is also reflected in the divergence of the specific gene family represented by the gene tree G.

If G and T are root-inconsistent, we use a duplication event in the root of T to express the inconsistency. Thus, we pose the following postulate.

> **Duplication/Speciation Principle**(Mirkin *et al.*, 1995). *Root-inconsistency of the trees G and T means that a duplication event in the gene corresponding to tree G occurs at the root of the species tree T and evolves in T in such a way that contemporary organisms corresponding to the leaves in $a(g)$ have one of the divergent gene copies and the leaves in $b(g)$ the other.*

Duplication/Speciation Principle induces the following simplest duplication assignment $\delta_g : T \to \{+-, +, -, \phi\}$:

1. for any $i \in I$, $\delta_g(i)$ is defined by

$$\delta_g(i) = \begin{cases} + & \text{if } \ i \in a(g), \\ - & \text{if } \ i \in b(g), \\ \phi & \text{if } \ i \in I - J; \end{cases}$$

2. for any internal node $n \in T$,

$$\delta_g(n) = \begin{cases} + & \text{if } n \cap a(g) \neq \phi \text{ and } n \cap b(g) = \phi, \\ - & \text{if } n \cap b(g) \neq \phi \text{ and } n \cap a(g) = \phi, \\ +- & \text{if } n \cap a(g) \neq \phi \text{ and } n \cap b(g) \neq \phi, \\ \phi & \text{if } n \subseteq I - J. \end{cases}$$

The mapping δ_g can be computed just in one bottom-up run through tree T using at most N steps, each involving comparing two sets of at most N elements. The complexity $c(\delta_g)$ demonstrates the extent of the biologically meaningful difference between G and T, and will be denoted as $c(g, T) = c(\delta_g)$.

To compare the entire gene tree G with species tree T, we need comparing all the subtrees of G with T. For simplicity, we assume that these two trees have an identical leaf set, that is, $I = J$. Recall that $T(n)$ denotes the subtree of T rooted at the node n. The comparing procedure consists of sequential comparisons of all the gene subtrees $G(m)$ with those subtrees, $T(n)$, of T such that $m \subseteq n$. If $G(m)$ is root-consistent with a subtree $T(n)$ of T, we proceed to subtrees of $T(n)$. When $G(m)$ is root-inconsistent with $T(n)$, the minimum duplication assignment $\delta_{nm} : T(n) \to \{+-, +, -, \phi\}$ is defined as a duplication of gene G in the node n ($\in T$) to explain the inconsistency. Figure 2 illustrates the comparison of a gene tree G (in (b)) with a species tree T (in (a)). There are three subtrees rooted at A, B, and C which are root-inconsistent with the corresponding subtrees. The corresponding duplications are shown in Figure 2 (c), (d), and (e) respectively, where maximal speciated/gapped nodes are marked with square boxes. The costs of duplications in (c), (d), and (e) are 8, 5 and 4 respectively.

The total inconsistency, $c(T/G)$, between T and G is defined through all the duplication events:

$$c(T/G) = \sum_{g' \in G} c(g', T).$$

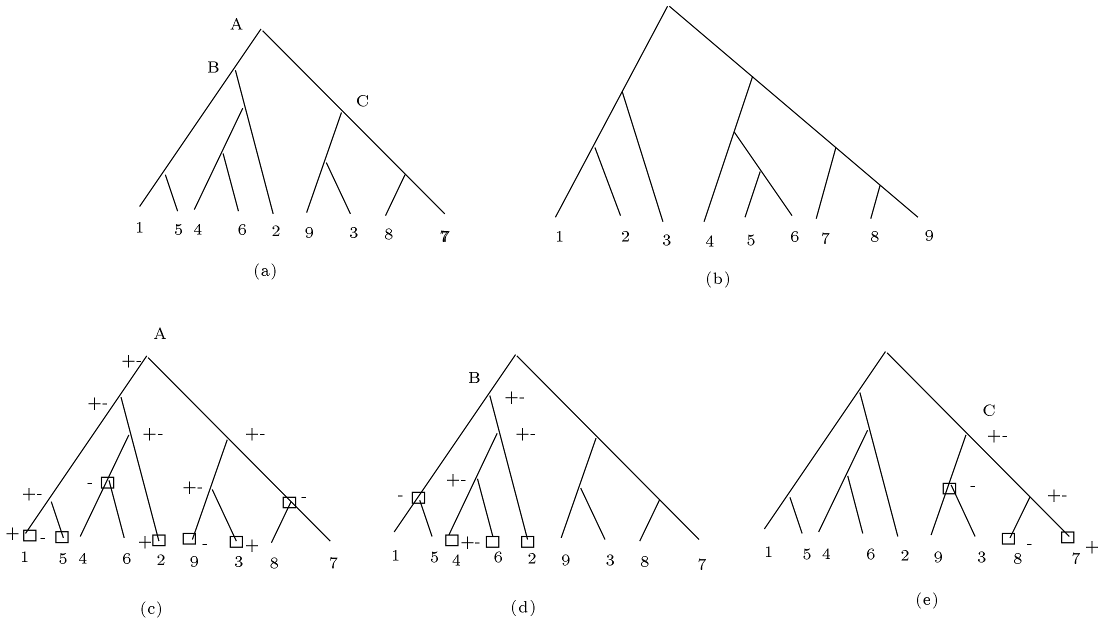

Figure 2: *Duplications between a species tree (a) and a gene tree (b)*

3 Computing the inconsistency

3.1 Mirkin-Muchnik-Smith conjecture

Given a species tree T and a gene tree G with the same leaf set of N taxa. By definition, computing the total inconsistency requires $O(N^3)$ steps: any of $(N-1)$ non-singleton subtrees of G is compared with at most $(N-1)$ subtrees of T, and each comparison involves checking for the root-inconsistency and then defining the duplication if necessary, which takes at most N^2 steps. Actually, computing the total inconsistency of T and G is much easier based on a conjecture posed by Mirkin *et al.* (1995)which relates $c(T/G)$ to the combinatorial properties of mapping G into T. Before stating the conjecture, we introducing some necessary concepts and facts regarding to mapping G into T.

For any node $g \in G$, we use $M(g)$ to denote the node of T being its least common ancestor, that is, the smallest cluster satisfying $g \subseteq M(g)$. This correspondence M, first considered by Goodman *et al.* (1979), is referred to as mapping of G into T by Page (1994). We call $M(g)$ the *destination* of g. Recall that for an internal node g, $a(g)$ and $b(g)$ denotes its two children.

Definition 3.1 *Let g be an internal node of G. It is said to be type-1 under the mapping if $M(a(g)) \subset M(g)$ and $M(b(g)) \subset M(g)$, where $a(g)$ and $b(g)$ are the two children of g; it is type-2 if $M(a(g)) \subset M(g)$ and $M(b(g)) = M(g)$ or vice verse; it is type-3 if $M(a(g)) = M(b(g)) = M(g)$.*

Obviously, the following fact is true.

Proposition 3.1 *(***Mirkin et al., 1995***) A node $g \in G$ is a type-2 or type-3 node if and only if subtrees $G(g)$ and $T(M(g))$ are root-inconsistent.*

The mutation cost function associated with the mapping itself is defined as follows. We use G_i to denote the set of all type-i nodes in G for $i = 1, 2, 3$.

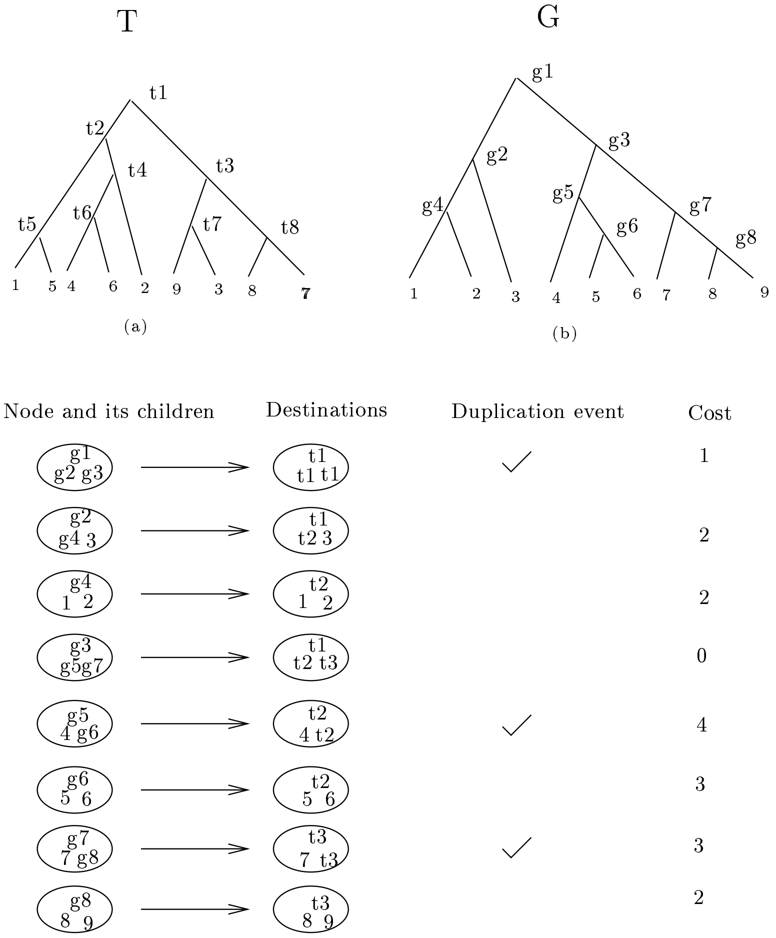

Figure 3: *Mapping G onto T.*

Definition 3.2 (**Guigó** *et al., 1994) The cost $L(g)$ associated with $g \in G$ is defined as:*

$$L(g) = \begin{cases} |M(g)M(a(g))| + |M(g)M(b(g))| & \text{if } g \in G_1 \\ |M(g)M(a(g))| + 2 & \text{if } g \in G_2 \text{ and } M(a(g)) \subset M(g), \\ 1 & \text{if } g \in G_3. \end{cases}$$

where $|M(g)M(a(g))|$ denotes the number of intermediate nodes between $M(g)$ and its descendant $M(a(g))$ in T. The mutation cost function $c(G,T)$ associated with the mapping of G into T is the sum of all $L(g)$ of the internal nodes $g \in G$.

Figure 3 presents the mapping of the gene tree G onto the species tree T which are illustrated in Figure 2. In their paper, Mirkin *et al.* (1995) conjecture that $c(T/G) = c(G,T)$ for any two trees with the same labeled leaf set.

3.2 The proof of the conjecture

In this section, we shall prove that their conjecture holds. Here, we assume that T is a species tree and G a gene tree and both have the same leave set I.

Proposition 3.2 *If T and G are root-inconsistent and δ denotes the duplication assignment from G to T induced by the Duplication/Speciation Principle, then the number of the maximal speciation/gapped nodes is equal to the number of the mixed nodes plus 1.*

Sketch of Proof. It follows from the fact that any intermediate node between the root of T and a maximal speciated/gapped node is a mixed node. □

Now we consider the duplication mapping M from G to T. Suppose that there are k_1 type-1 nodes, g_{1i} $(1 \le i \le k_1)$, k_2 type-2 nodes, g_{2i} $(1 \le i \le k_2)$ and k_3 type-3 nodes, g_{3i} $(1 \le i \le k_3)$. Since G is a full binary tree with N leaves, G has $N-1$ internal nodes. Therefore,

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

By Proposition 3.1, we also have

Proposition 3.3 *There are $k_2 + k_3$ duplications between G and T.*

For a type-1 node g_{1i}, $1 \le i \le k_1$, $M(a(g_{1i}))$ and $M(b(g_{1i}))$ are distinct from $M(g_{1i})$. The unique path from $M(a(g_{1i}))$ to $M(b(g_{1i}))$ through $M(g_{1i})$ is called a *path* in the mapping M from G to T. For our purpose, we say that such path *starts* at $M(g_{1i})$. We also say that such a path *passes* through any intermediate between $M(a(g_{1i}))$ and $M(g_{1i})$ or between $M(b(g_{1i}))$ and $M(g_{1i})$. For a type-2 node g_{2i}, $1 \le i \le k_2$, let $M(a(g_{2i})) \subset M(g_{2i})$ and let $M(b(g_{2i})) = M(g_{2i})$. The unique path from $M(g_{2i})$ to its descendant $M(a(g_{2i}))$ is called a *path* in the mapping M from G to T, starting at $M(g_{2i})$. Such a path *passes* through all intermediates between $M(a(g_{1i}))$ and $M(g_{1i})$.

Proposition 3.4 *For any non-duplication (i.e., type-1) node x, the total number of duplications in which x is mixed is exactly one less than the number of paths passing through x in the mapping M of G in T.*

For any duplication node t_i, the total number of duplications in which t_i is mixed is one less than the sum of the numbers of paths passing through t_i and of the paths starting at t_i.

For understanding Proposition 3.4 and the following proof, one would better study the example illustrated in Figure 2 and Figure 3 again.

Proof. Consider a non-duplication node $x \in T$. Set

$$G'_x = \{g' \in G \mid M(g') \subset x \subset M(p(g'))\},$$

$$G''_x = \{g'' \in G \mid x \cap g'' \neq \phi \;\&\; x \subset M(g'')\},$$

$$\overline{G}'_x = \{g' \in G \mid M(g') \cap x = \phi \;\&\; p(g') \in G''_x, \;\&\; a(p(g')) \in G'_x\},$$

where $a(p(g'))$ is the other child of $p(g')$. It is not difficult to see that G''_x contains all intermediate nodes between the nodes in G'_x and the root g of G, and $\overline{G}'_x$ consists of all siblings of nodes in G'_x. Furthermore, the parents of all nodes in G'_x are either type-1 or type-2 nodes and the corresponding paths starting at their destinations pass through x. In fact, as we shall see later, for a node $g \in G'_x$, if its parent is a type-1 node, then its sibling is in $\overline{G}'_x$; if its parent is a type-2 node, then its sibling is in G''_x. Therefore, $G'_x \cup \overline{G}'_x \cup G''_x$ is a subtree rooted at g, the root of G. Consider the homomorphic subtree, G_x, of G induced by nodes in $G'_x \cup \overline{G}'_x$. Recall that G_x is obtained from $G'_x \cup \overline{G}'_x \cup G''_x$ after the contraction of all degree-2 nodes except for the root.

Claim 1. Let $y \in G$. If y induces a duplication at $M(y) \in T$ in which x is mixed. Then, y is a node in the subtree G_x.

Proof. Since duplication occurs in $M(y)$, then $G(y)$ and $T(M(y))$ are root-inconsistent. Therefore, $M(a(y)) = M(y)$ or/and $M(b(y)) = M(y)$. Since x is mixed in this duplication, we have $x \subseteq M(y)$, which implies that $y \in G''_x$. Since $a(y) \cap x \neq \phi$ and $b(y) \cap x \neq \phi$, y is not a degree-2 node in $G'_x \cup \overline{G}'_x \cup G''_x$, and thus in G_x. □

Conversely, we have

Claim 2. Let $y \in G_x$. If $a(y) \in G''_x$ and $b(y) \in G''_x$, then, y is a type-3 or type-2 node and x is mixed in the duplication induced by y.

Proof. Since $x \subset M(a(y))$ and $x \subset M(b(y))$, both $M(a(y))$ and $M(b(y))$ are intermediate nodes in the path from the root t to x. If $M(b(y)) \subset M(a(y))$, then $M(y) = M(a(y))$, and $M(y) = M(b(y))$ otherwise. By the definition of G''_x, $x \cap a(y)$ and $x \cap b(y)$ are non-empty, and so x is mixed in the duplication induced by y. □

Claim 1 and Claim 2 reflect that all duplications in which x is mixed are induced by the corresponding internal nodes of G_x. On the other hand, there are also one-to-one corresponding between all paths passing through x and all type-1 or type-2 nodes in G_x. Given a node $y \in G$. Suppose y is a type-1 or type-2 node and the corresponding path passes through x. By the construction of G_x, $y \in G_x$. Furthermore, we have the following facts.

Claim 3. Let $y \in G_x$. If $a(y) \in G'_x$ and $b(y) \in \overline{G}'_x$, then y is a type-1 node and the corresponding path starting at $M(y)$ passes through x.

Proof. By definition, $M(a(y)) \subset x \subset M(y)$ and $M(b(y)) \subset M(y)$. Thus, y is a type-1 node. Obviously, the corresponding path passes through x. □

Claim 4. Let $y \in G_x$. If $a(y) \in G'_x$ and $b(y) \in G''_x$, then y is a type-2 node and the corresponding path starting at $M(y)$ passes through x.

Proof. Since $x \subset M(b(y))$ and $M(a(y)) \subset x$, then $y = a(y) \cup b(y) \subset x \cup M(b(y)) = M(b(y))$. By the definition, $M(y) = M(b(y))$. Thus, y is a type-2 node. Since $M(a(y)) \subset x \subset M(y)$, the corresponding path passes through x. □

Let the numbers of type-1, type-2, type-3 nodes in G_x be n_1, n_2, n_3 respectively. By Claim 1 and Claim 2, the number of duplications in which x is mixed is $n_2 + n_3$. By Claim 3 and Claim 4, the number of paths passing x is $n_1 + n_2$. Since the subtree G_x is a binary, full, then $n_1 = n_3 + 1$. Thus, the fact is true for a non-duplication node.

Similarly, the fact can be proved for duplication nodes. The only difference is that we have to include all node $y \in G$ such that $M(y) = x$ in the set G''_x. This concludes the proof of Proposition 3.4. □

Let duplication occur at p nodes of T, $t_1, t_2, \cdots, t_p$, and let there be t_i duplications at node t_i: $D_{i1}, D_{i2}, \cdots, D_{it_i}$. Then, by Proposition 3.3,

$$\sum_{i=1}^{p} t_i = k_2 + k_3. \tag{2}$$

By Propositions 3.2, 3.4 and Formula (1), (2),

$$
\begin{aligned}
c(T/G) &= \sum_{i=1}^{p}\sum_{j=1}^{t_i} c(D_{ij}) \\
&= \sum_{i=1}^{p} t_i + \sum_{1\le i\le k_1}(|M(g_{1i})M(a(g_{1i}))| + |M(g_{1i})M(b(g_{1i}))|) + k_1 \\
&\quad + \sum_{1\le i\le k_2}|M(g_{2i})M(a(g_{2i}))| + k_2 - (N-1) + k_2 + k_3 \\
&= k_2 + k_3 + \sum_{1\le i\le k_1}(|M(g_{1i})M(a(g_{1i}))| + |M(g_{1i})M(b(g_{1i}))|) \\
&\quad + \sum_{1\le i\le k_2}|M(g_{2i})M(a(g_{2i}))| + k_2 \\
&= \sum_{1\le i\le k_1}(|M(g_{1i})M(a(g_{1i}))| + |M(g_{1i})M(b(g_{1i}))|) \\
&\quad + \sum_{1\le i\le k_2}(2 + |M(g_{2i})M(a(g_{2i}))|) + k_3 \\
&= c(G,T)
\end{aligned}
$$

Hence, we have proved the following conjecture.

Theorem 3.1 (**Mirkin-Muchnik-Smith Conjecture**) *For any T and G with the same set of leaves, $c(T/G) = c(G,T)$).*

3.3 A linear-time algorithm

In this subsection, we shall present a linear-time algorithm for computing the cost function $c(G,T)$ for a gene tree G and a species tree T.

First, we compute the mapping M from G to T in linear time. Given a node $u \in G$, by the definition of mapping, its destination $M(u)$ is the lowest common ancestor of $M(a(u))$ and $M(b(u))$. This simple observation leads the following algorithm for computing the mapping from a gene tree to a species tree. In our algorithm, we define an auxiliary binary tree MT with information attached to various nodes. MT has the same structure as G. At each node $u \in MT$, we associated a pointer $m(u)$, which points to the destination $M(u)$ of u after it is computed. For $u, v \in G$, we use $lca(u,v)$ to denote the lowest common ancestor of u and v and $LCA(u,v)$ to denote the instruction for finding $lca(u,v)$. Finally, recall that in postorder, $i < j$ if and only if i is to the left of j or a descendant of j(see Aho *et al.*, 1974).

Algorithm 1

1. Generate a sequence, S, of LCA instructions by processing each node of T in postorder. For each internal node $u \in G$, generate an instruction: $LCA(M(a(u)), M(b(u)))$;
2. Execute the instruction sequence S. After finishing an instruction corresponding to $u \in G$, save the result to $MT(u)$.
3. Output $MT(u)$.

Theorem 3.2 *The mapping MT can be computed in $O(n)$ time and $O(n)$ space on a RAM using Algorithm 1.*

Proof. Obviously, Step 1 takes $O(n)$ times and the instruction set S can be saved in $O(n)$ spaces.

In the instruction sequence S, the nodes involving in an instruction depend on the results of two previous instructions, and thus each instruction must be answered before processing the next. Therefore, we use the on-line algorithms for finding lowest common ancestors presented in Harel and Tarjan(1984) or Schieber and Vishkin (1988). Both algorithms takes $O(n)$ time and $O(n)$ space for executing the sequence S of $(n-1)$ LCA instructions. $\Box$

Recall that for a node $u \in T$, the depth of u is the length of the simple path from the root to itself. The depth of the root is obviously 0. For computing the mapping cost function $c(G,T)$, we need to preprocess the tree T to get the depth $D(u)$ of each node $u \in T$ and then calculate the cost using information arrays D and MT.

Theorem 3.3 *The mapping cost can be computed in $O(n)$ time and $O(n)$ space on a RAM using Algorithm 2.*

Proof. The correctness follows from the definition of the mapping cost. Observe that Both Step 1 and Step 3 take $O(n)$ time and $O(n)$ space. By Theorem 3.2, Step 2 takes $O(n)$ time and $O(n)$ space. Thus, Algorithm has the required time and space complexity. $\Box$.

Algorithm 2

1. Process each node of T in preorder. For each node u, calculate its depth $D(u)$.
2. Use Algorithm 1 to compute the mapping MT from G into T.
3. Compute the mapping cost c by process each internal node of MT. Initially, $c = 0$. For each internal node $u \in MT(u)$, update the cost by
   $c = c + D(m(a(u))) + D(m(b(u))) - 2D(m(u)) - 2$
   if $m(a(u)) \neq m(u)$ and $m(b(u)) \neq m(u)$,
   $c = c + D(m(a(u))) - D(m(u)) + 1$
   if only $m(a(u)) \neq m(u)$, or
   $c = c + 1$
   otherwise.

By Theorem 3.1, the inconsistency cost is also computable in linear time and space on a RAM.

Acknowledgement

This work was supported by a CGAT grant when author was in University of Waterloo. The author would like to thank Ilya Muchnik for helpful discussions regarding to the topic in this paper through e-mail.

References

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

[2] A.V. Aho, J.E. Hopcroft and J.D. Ullman, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, Mass., 1974.

[3] J.P. Barthélemy, B. Leclerc, and B. Monjardet, On the use of ordered sets in problems of comparison and consensus of classifications, *Journal of Classification* 3(1986), 187-224.

[4] B. DasGupta, X. He, T. Jiang, M. Li, J. Tromp and L. Zhang. On distance between phylogenetic trees. To appear on *the ACM-SIAM Symposium on Discrete Algorithms*, 1997.

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

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

[7] R. Guigó, I. Muchnik and and T. Smith, Reconstruction of ancient molecular phylogeny, *Molecular Phylogenetics and Evolution*, in press.

[8] D. Harel and R.E. Tarjan, Fast algorithms for finding nearest common ancestors, *SIAM J. on Comput.* 13(1984), No. 2, 338-355.

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

[10] T. Margush and F. R. McMorris, Consensus n-Trees, *Bulletin of Mathematical Biology* 43(1981), 239-244.

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

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

[13] 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.

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

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

[16] P. Pamilo and M. Nei, Relationship between gene trees and species trees. *Mol. Bio. Evol.* 5 (1988), 568-583.

[17] D. F. Robinson, Comparison of labeled trees with valency three, *Journal of Combinatorial Theory*, Series B, 11(1971), 105-119.

[18] B. Schieber and U. Vishkin, On finding lowest common ancestors: simplification and parallelization, *SIAM J. on Comput.* 17(1988), No. 6, 1253-1262.

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

[20] M. S. Waterman and T. F. Smith, On the similarity of dendrograms, *Journal of Theoretical Biology* 73(1978), 789-800.

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