Duplication Cost Diameters

Paweł Górecki[1], Jarosław Paszek[1], and Oliver Eulenstein[2]

[1] Department of Mathematics, Informatics and Mechanics, University of Warsaw, Poland
{gorecki,j.paszek}@mimuw.edu.pl
[2] Department of Computer Science, Iowa State University, USA
oeulenst@cs.iastate.edu

Abstract. The gene duplication problem seeks a species tree that reconciles given gene trees with the minimum number of gene duplication events, called gene duplication cost. To better assess species trees inferred by the gene duplication problem we study diameters of the gene duplication cost, which describe fundamental mathematical properties of this cost. The gene duplication cost is defined for a gene tree, a species tree, and a leaf labeling function that maps the leaf-genes of the gene tree to the leaf-species. The diameters of this cost are its maximal values when one topology or both topologies of the trees involved are fixed under all possible leaf labelings, and are fundamental in understanding how gene trees and species trees relate. We describe the properties and formulas for these diameters for bijective and general leaf labelings, and present efficient algorithms to compute the diameters and their corresponding leaf labelings. Moreover, we provide experimental evaluations demonstrating applications of diameters for the gene duplication problem.

1 Introduction

A basic tenet of all biological disciplines is the common evolutionary history of all life forms, including all extant species. The evolutionary relationships among such entities are usually represented as species trees and are a key tool in understanding evolution and its complex events that have engineered the enormous species and phenotypic diversity to date. Species trees are fundamental to evolutionary biology, but are also essential tools for an array of other disciplines such as agronomy, biochemistry, conservation biology, epidemiology, and medical sciences. For example, evolutionary trees are increasingly used to study the dynamic range of patients' cancer progressions, and to tailor corresponding treatments [22]. Species trees were also used to develop pesticides [17], to control invasive species [16], and to predict outbreaks of infectious diseases [14]. Invariably common to such studies is that large-sale species trees need to be accurately inferred.

One approach to construct large-scale trees is to utilize the rapidly growing availability of *gene trees*, i.e. the evolutionary history of genes. Gene trees describe how parts of the species genomes have evolved, and thus can be assembled into larger evolutionary trees of species. Unfortunately, evolutionary mechanism can cause conflicting evolutionary relationships between gene trees and the topology of the species tree along whose branches they have evolved [26]. Resolving such conflicts has become a grand

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, pp. 212–223, 2014.

challenge in the field of phylogenetic tree inference. There has been considerable interest in inference approaches that account for conflict involving gene duplication [5], which is a major and frequently occurring evolutionary mechanism [25].

One such approach is solving the gene duplication problem, which is well studied [5]. Given a collection of gene trees, this problem seeks a species tree that is a median tree of the given trees under the gene duplication cost, i.e. the fewest number of duplication events that can explain the conflict between a gene tree and a species tree [12,23]. Despite the NP-hardness of the gene duplication problem [19], effective local search heuristics [1,30] have produced credible estimates for this problem [4,20,21,24]. Recently, exact dynamic programming solutions have been developed for the gene duplication problem that were able to solve instances of up to 22 taxa within a few hours on a standard workstation [3]. Furthermore, there are modifications of the gene duplication problem that handle various types of input trees, such as erroneous trees [8], unrooted trees [11], and non-binary trees [18].

Unfortunately, despite ongoing work on the gene duplication problem, little is known about the mathematical properties of the gene duplication cost that is at the heart of the gene duplication problem. Here we investigate into diameters of this cost, that is, the maximal values of this cost when one shape or both shapes of the input trees are fixed. Diameters of the duplication cost are fundamental in understanding how gene trees and species trees relate under this cost [12].

Related Work. Goodman et al.'s pioneering work [7] introduced the gene duplication cost between a gene tree and a species tree that are both rooted and full binary. It is also assumed that the leaves of the gene tree are related to the leaves of the species tree by a function that is called *leaf labeling*. The leaf labeling is thought to relate the leaves of the species tree by a leaf of the gene tree from which it was sampled. An extension of the leaf labeling, called *least common ancestor (LCA) mapping*, relates every node of the gene tree to the most recent species in the species tree that could have contained this gene. A node in the gene tree is a *gene duplication* if it has the same lca mapping as one of its children. The *duplication cost* between a gene tree and a species tree under a given leaf labeling is the number of gene duplications. While diameter for other cost functions used for tree inference are well understood [10,28], diameters for the gene duplication cost have not been investigated yet.

Our Contribution. Under the gene duplication cost we study diameters of trees when leaf labelings are constrained to be bijective and when they are unconstrained. In particular, we study under the assumption that only the topologies of the trees involved are given (i) the diameter of a gene tree and a species tree, (ii) the diameter of a gene tree, and (iii) the diameter of a species tree. For example, the bijective diameter of a gene tree is the maximal duplication cost among all gene and species tree pairs such that the gene tree is fixed and its leaf labeling is bijective.

For bijective leaf labelings we show that the diameter of a gene tree and a species tree is equal to the number of non-cherry nodes (a node that is not the parent of two leaves) in the gene tree. While it follows that this diameter is linear time computable, we show that a leaf labeling that induces this diameter is also linear time computable. We also provide the exact conditions for a gene tree and a species tree to establish this diameter. For other types of diameters we describe their formulas. Moreover, we study the

properties of the expected values of the bijective diameters and provide formulas for them. Based on this theoretical results, we evaluate computationally the expected values of bijective diameters. We also provide two experiments showing applications of unconstrained diameters in the gene duplication problem for two empirical datasets [13,27].

2 Preliminaries

2.1 Basic Definitions

We begin by recalling some basic definitions from phylogenetic theory. A *(phylogenetic) tree* is a connected acyclic graph such that exactly one of its nodes has a degree of two (root), and all its remaining nodes have a degree one (leaves) and three. The nodes with degree at least two are called internal. By $\preceq$ we denote the partial order in a tree T such that $a \preceq b$ if b is a vertex on the shortest path between a and the root of T. Note that $a \prec b$ is equivalent to $a \preceq b$ and $a \neq b$. The least common ancestor of a and b in T is denoted by $a \oplus b$. We write that a and b are *comparable* if $a \preceq b$ or $b \preceq a$. If a is not the root of T, then the *parent* of a, denoted by πa, is the least node v such that $a \prec v$. If two nodes a and b have the same parent, then a is called *a sibling* of b. A sibling of a will be denoted by σa. $T(a)$ is the maximal subtree of T rooted at a. $|T|$ denotes the number of leaves in T. A *cherry* in T is a subtree of T that has exactly two leaves. A leaf that is not an element of a cherry is called *free leaf*. By L_T we denote the set of all leaves in T.

A *species tree* is a tree whose leaves are called *species*. A *gene tree* over a set of species X, is a triple $\langle V_G, E_G, \Lambda_G \rangle$ such that $\langle V_G, E_G \rangle$ is a tree and Λ_G is the leaf labeling function $\Lambda_G \colon L_G \to X$, called *labeling*. For simplicity, if the species tree S is known, we write that the gene tree is over S instead of L_S. Traditionally, gene and species trees are defined by nested parenthesis notation. For a species tree S and a gene tree G over S, let $M \colon V_G \to V_S$ be the *least common ancestor (LCA) mapping* between G and S that preserves the labeling of the leaves. In other words, $M|_{L_G} = \Lambda_G$, and for any non-root node a we have $M(\pi a) = M(a) \oplus M(\sigma a)$.

Parenthesis may be omitted in formulas for more clarity. For instance, instead of writing $M(\pi x)$ to denote the LCA mapping of πx, we write $M\pi x$. If $P = \langle p_1, p_2, \ldots, p_k \rangle$ is a sequence of nodes of G, then by MP, we denote the sequence $\langle Mp_1, Mp_2, \ldots, Mp_k \rangle$. If Q is a set of nodes, we define MQ to be $\{Mq \colon q \in Q\}$.

A *path* P in a tree T is a non-empty sequence of nodes without repetitions such that for every adjacent v and w in P, $\{v, w\} \in E_T$. Note that every path in a tree T has a unique $\preceq$-maximal element. We denote it by $\max P$. A path P is called *simple* (in T) if its vertexes are comparable and its unique $\preceq$-minimal element will be denoted by $\min P$. A *path partition* Π of T is a set of pairwise disjoint paths in T such that $\bigcup \Pi = V_T$.

A tree is called *caterpillar* if it contains exactly one cherry. A tree T is called *trivial* if T is a single noded tree, i.e., $|V_T| = 1$. By $\mathcal{C}(T)$ we denote the set of all cherry roots from a tree T. By χ_T we denote the number of all cherries in a tree T.

We call an internal node g from G *a S-duplication (node), or duplication*, if $Mg = Mc$ for some child c of g. The *duplication cost* (D) from G to S, denoted by $\mathrm{D}(G, S)$, is defined as the total number of duplication nodes present in G.

2.2 Duplication Diameters

For a species tree S, we denote by $\mathbb{G}(S)$ the set of all gene trees over S. By $\mathbb{B}(S)$ we denote the set of all gene trees over S with bijective labeling.

Let G be a gene tree. By $\hat{G}$ we denote the unlabeled tree obtained from G by forgetting the labeling. We define several types of *bijective diameters*. For trees T and S with the same number of leaves we define *the bijective duplication diameter for fixed shapes* as:

$$\mathrm{b^D}(T,S) = \max\{\mathrm{D}(G,S)\colon G \in \mathbb{B}(S), \hat{G} = T\}$$

We define *the bijective duplication diameter for a fixed species tree*:

$$\mathrm{b^D}(\star,S) = \max_{G\in\mathbb{B}(S)} \mathrm{b^D}(G,S).$$

Next, we define *the bijective duplication diameter for a fixed shape of a gene tree*:

$$\mathrm{b^D}(T,\star) = \max_{S} \mathrm{b^D}(T,S).$$

Similarly to $\mathrm{b^D}(T,S)$, $\mathrm{b^D}(T,\star)$ and $\mathrm{b^D}(\star,S)$ we introduce $\mathrm{u^D}(T,S)$, $\mathrm{u^D}(T,\star)$ and $\mathrm{u^D}(\star,S)$, respectively, to denote the *unconstrained duplication diameters* by replacing $\mathbb{B}(S)$ in the above definitions with $\mathbb{G}(S)$, $\mathrm{b^D}$ with $\mathrm{u^D}$ and relaxing the assumption that T and S have the same size. We omit straightforward definitions.

3 Results

3.1 Bijective Duplication Diameters

In this section all labelings are bijections. First we define problems related to bijective diameters.

Problem 1. Given a tree T and a species tree S with the same number of leaves. Find a gene tree G such that $\hat{G} = T$ and $\mathrm{D}(G,S) = \mathrm{b^D}(T,S)$.

Problem 2. Given a tree T. Find a species tree S and a gene tree G such that $\hat{G} = T$ and $\mathrm{D}(G,S) = \mathrm{b^D}(T,\star)$.

Problem 3. Given a species tree S. Find a gene tree $G \in \mathbb{B}(S)$ such that $\mathrm{D}(G,S) = \mathrm{b^D}(\star,S)$.

First, we show the upper bound for the diameter:

Lemma 1. *For every species tree S and every tree G both with n leaves:* $\mathrm{b^D}(G,S) \leqslant n - 1 - \chi_G$.

Proof. It is obvious that a gene tree can have at most $n-1$ duplication nodes. Additionally, a cherry root in G cannot be a duplication node. Thus, the upper bound for $\mathrm{b^D}(\hat{G},S)$ is $n - 1 - \chi_G$.

We now show that the upper bound is reached by showing a simple procedure that induces the maximal number of gene duplications. First we introduce a notion of a *trunk* [10], that will be used to assign mappings to a cherry leaves.

A *trunk* of a species tree S is a non-empty sequence $\Upsilon = \langle \Upsilon_1, \Upsilon_2, \ldots, \Upsilon_k \rangle$, of internal nodes from S that starts in the root of S and satisfies (I) both children of Υ_k have the same number of descendant leaves, and (II) for $i > 1$, Υ_i is the child of Υ_{i-1} having more descendant leaves than its sibling. A subtree T of S is called a *limb*, if the root of T is not a trunk node, while its parent is a trunk node. S can be represented in *limb format* by using the standard nested parenthesis notation: $S = (S_1, (S_2, \ldots, (S_{|\Upsilon|}, S_{|\Upsilon|+1})))$, where S_i is a limb of S, for each i. For a leaf a in S let Υa be the lowest trunk node, whose subtree contains a.

Algorithm 1. Inference of a gene tree that induces the diameter b^{D}

1. **Input:** A species tree S and a tree T both with $n > 1$ leaves. **Output:** A gene tree G such that $T = \hat{G}$ and $\mathrm{D}(G, S) = \mathrm{b}^{\mathrm{D}}(T, S)$.
2. **Comment** We define the labeling function Λ for T.
3. **Let** $\langle \xi_1, \xi_2, \ldots, \xi_{\chi_G} \rangle$ be a sequence of all cherry roots from T. **Let** $(S_1, (S_2, \ldots, (S_k, S_{k+1})))$ be a limb representation of S. **Let** $v_1, v_2, \ldots, v_n$ be the sequence of all leaves of S, such that, for every $i < j$, if $v_i \in S_k$ and $v_j \in S_l$, then $k \leqslant l$.
4. **For** $i = 1, 2, \ldots, \chi_G$ **Do** if x and y are the leaves of $G(\xi_i)$, set $\Lambda(x) := v_i$ and $\Lambda(y) := v_{n-\chi_G+i}$,
5. **Let** $v_{\chi_G+1}, v_{\chi_G+2}, \ldots, v_{n-\chi_G}$ be the labels of the remaining $n - 2 * \chi_G$ free leaves of G.
6. **Return** $\langle V_T, E_T, \Lambda \rangle$.

Theorem 1. *For every species tree S and every tree T both with n leaves: $\mathrm{b}^{\mathrm{D}}(T, S)$ equals the number of non-cherry nodes from T, that is, $n - 1 - \chi_T$.*

Proof. We show that Algorithm 1 infers a labeling that induces the maximal number of gene duplications. Let Λ be the function inferred by the algorithm. It is easy to see that Λ is a well defined labeling for T. Then, for every $1 \leqslant i < j \leqslant |L_S|$, we have $M\xi_i \geq M\xi_j$, thus all cherry mappings are comparable. We show that every non-cherry node is a S-duplication in a gene tree $\langle V_T, E_T, \Lambda \rangle$. Let g be a non-cherry node and let i be the minimal index of a cherry root from ξ that is present in $G(g)$. All mappings of internal nodes are comparable, thus $M\xi_i = Mg$. Moreover, at least one child of g is mapped to $M\xi$, thus g is a duplication. This completes the proof.

Now, we introduce a notion of a cherry partition, that is crucial for the bijective diameters of the duplication cost. A path partition Π of a tree G is called *cherry partition (of G)* if every path that contains an internal node is a simple path whose minimal element is a cherry root. We say that the LCA mapping M between a species tree S and a gene tree G over S *induces a cherry partition* if there is a cherry partition Π of G such that for every path P in Π, MP consists of a single element. Let $A \subseteq V_G$, then by $G \triangle A$ we denote the set of all maximal subtrees of G that do not contain any nodes from A.

First we classify cherry partitions.

Lemma 2. *Let S be a species tree, G be a gene tree over S and M be the LCA mapping between G and S. Then, M induces a cherry partition of G if and only if for every internal node g of G there is a cherry K in $G(g)$ such that $MK = Mg$.*

Proof. ($\Rightarrow$). Assume that M induces a cherry partition Π. Let g be an internal node of G. The proof is by induction on the structure of G. We consider two cases. (1) If g is a cherry root, then the property is trivially satisfied. (2) Otherwise, g has a child c such that g and c belongs the the same path from Π. Thus, $Mg = Mc \preceq M\sigma c$. By the inductive hypothesis there is a cherry K in $G(c)$ such that $Mc = MK$. By the previous observation, we have $MK = Mg$. This completes the first part of the proof. ($\Leftarrow$). We show that there exists a cherry partition of G induced by M. Consider the following procedure, where for each i, X_i is a set of subtrees of G: (I) Let $X_0 = \{G\}$. (II) For $0 < i \leqslant \chi_G$, let $X_i = (T \triangle r) \cup (X_{i-1} \backslash \{T\})$ and t_i is the root of T, where T is some non-trivial subtree of G from X_{i-1} and r is a cherry root from T such that $Mr = Mt_i$ (in G). It is easy to proof that for each $i \geqslant 0$, X_i is a set of disjoint subtrees of G satisfying $\bigcup_{T \in X_i} \mathcal{C}(T) = \mathcal{C}(G) \backslash \{r_1, r_2, \ldots, r_i\}$. Thus, $X_{\chi_G} = \bigcup_{g \in L_G} G(g)$, i.e., it is composed of all trivial subtrees of G. For $i = \{1, 2, \ldots, \chi_G\}$, let P_i be the simple and the shortest path connecting r_i and t_i. Note, that MP_i is a one-element set. Now, it should be clear that $\bigcup P_i$ is the set of all internal nodes of G, and the following family of simple paths $\bigcup_{g \in L_G} \langle g \rangle \cup \{P_1, P_2, \ldots, P_{\chi_G}\}$ is a cherry partition of G induced by M.

Theorem 2. *Let S be a species tree, G be a gene tree over S and M be the LCA mapping between G and S. Then, $\mathrm{b^D}(\hat{G}, S) = \mathrm{D}(G, S)$ if and only if M induces a cherry partition of G.*

Proof. ($\Rightarrow$). Assume that $\mathrm{b^D}(\hat{G}, S) = \mathrm{D}(G, S)$. Then by Theorem 1 every non-cherry node is a duplication. By Lemma 2 it is sufficient to show that for every internal g, there is a cherry K in $G(g)$ such that $Mg = MK$. The proof is by induction on the structure of G. If g is a cherry root the condition is trivially satisfied. Otherwise, g is a duplication and has a non-leaf child c such that $Mg = Mc$. Then by the inductive hypothesis, there is a cherry K in $G(c)$ such that $MK = Mc$. Thus, $MK = Mg$ which completes the inductive proof. ($\Leftarrow$). Assume that there is a cherry partition induced by M. Let g be a non-cherry node of G. Then, by Lemma 2 there is a cherry root r in $G(g)$ such that $Mg = Mr$. Thus, for some child c of g, $Mc = Mg$. We conclude that g is a duplication. We proved that every non-cherry node is a duplication. Thus, by Theorem 1, $\mathrm{D}(G, S) = \mathrm{b^D}(\hat{G}, S)$.

Theorem 3. *Given a species tree S and a tree T both with n leaves, Algorithm 1 infers a gene tree G such that $\hat{G} = T$ and $D(G, S) = \mathrm{b^D}(T, S)$. The time complexity of Algorithm 1 is $O(n)$.*

Proof. Algorithm 1 is adopted from [10]. Correctness of Algorithm 1 follows from Theorem 1. It should be clear that every step of the algorithm can be completed in $O(n)$ number of steps. For more details please refer to [10].

In the remaining part of this section, we study other bijective diameters.

Theorem 4. *For every tree G and every species tree S with the same number of leaves, $\mathrm{b^D}(G, S) = \mathrm{b^D}(G, \star)$.*

Proof. The proof follows immediately from Theorem 1.

We conclude that Problem 2 can be solved by choosing any species tree with $|G|$ leaves and applying Algorithm 1.

Theorem 5. *For a species tree S with $n > 1$ leaves* $\mathrm{b}^{\mathrm{D}}(\star, S) = n - 2$.

Proof. It follows from Theorem 1, that the gene tree G that maximizes duplication cost should have the minimal number of cherries (which is 1 in this case). Thus, such a tree is a caterpillar tree. Moreover, by Lemma 2 and Theorem 2 the root of the only cherry present in G has to be mapped to the root of S.

We conclude that Problem 3 has a simple solution.

Theorem 6. *For a species tree S and a gene tree $G \in \mathbb{B}(S)$, $\mathrm{D}(G, S) = \mathrm{b}^{\mathrm{D}}(\star, S)$ if and only if G is a caterpillar tree such that the only cherry of G is mapped to the root of S.*

Proof. ($\Leftarrow$). It easy to see that $\mathrm{D}(G, S) = n - 2$. Thus by Theorem 5 $\mathrm{D}(G, S) = \mathrm{b}^{\mathrm{D}}(\star, S)$. ($\Rightarrow$). See the proof of Theorem 5.

3.2 Unconstrained Duplication Diameters

In this section, we show similar results for the unconstrained diameters. To avoid repetitions we skip formal definitions of unconstrained problems.

Theorem 7. *For trees T and S, we have* $\mathrm{u}^{\mathrm{D}}(T, S) = \mathrm{u}^{\mathrm{D}}(T, \star) = |T| - 1$.

Proof. For the first diameter, choose the labeling for T that is a constant function. Then, every internal node is a duplication. Thus, we have $|T| - 1$ gene duplications, which is the maximal possible duplication cost. The same holds true for $\mathrm{u}^{\mathrm{D}}(T, \star)$.

Theorem 8. *For a species tree S,* $\mathrm{u}^{\mathrm{D}}(\star, S) = +\infty$.

Proof. Assume that we have a sequence of trees $T_1, T_2, \ldots$ over S such that T_n has n leaves. Then, by Theorem 7 $\lim_{n \to +\infty} \mathrm{u}^{\mathrm{D}}(T_n, S) = \lim_{n \to +\infty} n - 1 = +\infty$. We conclude, $\mathrm{u}^{\mathrm{D}}(\star, S) = +\infty$.

3.3 Expected Number of Gene Duplications

In this section we show the formulas for the expected values of diameters considered in this paper.

In this section we assume $n!! = 1$ for $n \leqslant 2$. For a set X consisting of $n > 0$ species, by $\mathbb{S}(X, c)$, we denote the set of all pairwise non-isomorphic bijectively labeled gene trees over X (i.e., having n leaves uniquely labeled by elements from X) and c cherries. By $t_{n,c}$ we denote the size of $\mathbb{S}(X, c)$.

Lemma 3. *For $n > 1$ and $c \geqslant 0$, we have*

$$t_{n,c} = \begin{cases} (2c-1)!!(2c-3)!! & \text{if } n = 2c \geqslant 2, \\ t_{n-1,c-1}(n-2c+1) + t_{n-1,c}(n+2c-2) & \text{if } n > 2c > 0, \\ 1 & \text{if } n = 1, c = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Proof. We omit the proof for brevity.

Based on the recurrence for the number of rooted binary leaf-labeled trees with n leaves [6], we can derive the following recurrence for $t_{n,c}$

Lemma 4. *If $n \geqslant 3$ and $c \in \{1, 2, \ldots, \lfloor n/2 \rfloor\}$, then*

$$2t_{n,c} = \sum_{k=1}^{n-1} \binom{n}{k} \sum_{i=0}^{c} t_{k,i} t_{n-k,c-i}.$$

Proof. Every species tree $S \in \mathbb{S}(X, c)$ has the following form $S = (S', S'')$, where $S' \in \mathbb{S}(A, i)$ and $S'' \in \mathbb{S}(X \backslash A, c - i)$ for some $\varnothing \neq A \subsetneq X$, $k = |A|$ and $i \in \{0, 1, \ldots, c\}$. Thus, for every $k \in \{1, 2, \ldots, n-1\}$ we distribute k species from X in the left subtree of S. This can be done in $\binom{n}{k}$ ways. Then, for the left subtree we choose shapes with i cherries, while for the right subtree we choose shapes with $c - i$ cherries. Additional, every possibility is counted twice due to the symmetry.

Lemma 5. *For every $n \geqslant 1$, $\sum_c t_{n,c} = (2n-3)!!$.*

Proof. It follows easily from the fact that the number of rooted binary leaf-labeled trees with n leaves equals $(2n-3)!!$ [6].

Theorem 9. *Under the assumption of a uniform distribution of gene trees with n leaves, the expected value of the bijective diameter for fixed shape of gene tree equals*

$$\frac{1}{(2n-3)!!} \sum_{c} t_{n,c}(n - c - 1), \tag{1}$$

where n is the number of species.

Proof. It follows easily from Theorem 1 and Lemma 5.

The same result can be obtained for b^{D} diameter (under a uniform distribution of gene and species tree pairs). Finally, it is straightforward to proof that under the assumption of a uniform distribution of species trees, the expected value of the bijective diameter of a species tree equals $n - 1$. We omit details.

4 Experimental Evaluation and Discussion

We performed two types of experimental evaluations related to the diameters studied in this article. The first experiment analyzes the expected value of the bijective diameter of the gene tree, while the second experiment studies the effect of inferring species trees by solving the GTP problem when the cost involved is normalized by the diameter.

4.1 Expected Value of the Bijective Diameter of a Gene Tree

We computed values of the expected value of $\mathrm{b}^{\mathrm{D}}(T, \star)$ diameter according the the formula 1 from Theorem 9. By using a dynamic programming for efficient computation of the recurrence from Lemma 4, implemented in a python script, we computed the expected value of this diameter for $n = 10, 20, \ldots, 990$ (99 values in total). The overall time of computation was approximately 3 hours of a standard PC workstation (a single core, AMD processor, 1400MHz). The result is depicted in Figure 1. Based on Theorem 1 and Theorem 4, we present minimal (i.e., $\lceil n/2 \rceil - 1$) and maximal (i.e., $n - 1$) values of this diameter for fixed n. The main conclusion from this experiment is that the expected value can be well approximated by the average between these two values.

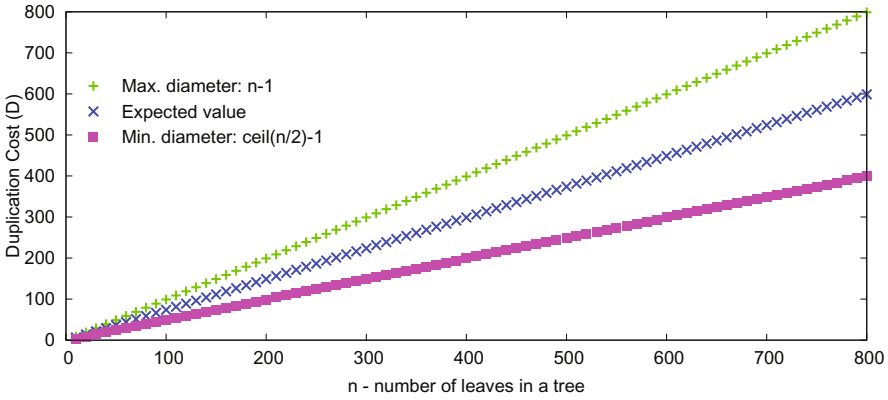

Fig. 1. The expected value of the bijective diameter for a fixed shape of a gene tree (middle line) with values of maximal and minimal values of this diameter. Here ceil(x) denote the ceil function, i.e., the smallest following integer function.

4.2 GTP Evaluation

We studied the gene tree parsimony problems (GTP) [3,9,19,29] under duplication cost and its normalized variant. The problems are defined as follows:

Problem 4 (GTP-DUP). Given a collection of gene trees Q. Find a species tree S that minimizes the total cost $\sum_{G \in Q} D(G, S)$.

Problem 5 (GTP-DUP-NORM). Given a collection of gene trees Q. Find a species tree S that minimizes the total normalized cost $\sum_{G \in Q} D(G, S) / \mathrm{b}^{\mathrm{D}}(\hat{G}, \star)$.

Since Problem 4 is known to be NP-hard, we implemented in Java a classical hill climbing heuristic based on the nearest neighborhood interchange (NNI) local search algorithm [2,9]. We used our computer program to perform four computational experiments on two publicly available datasets under the standard duplication cost (GTP-DUP) and the normalized cost (GTP-DUP-NORM).

Guigó dataset. The first dataset consists of 53 gene trees from 16 eukaryotes from [13] (median size of a gene tree is 4.66). It is known from [3], that there are exactly 71 optimal species trees having the minimal total duplication cost equal to 36. Our heuristic

was able to infer all these optimal trees. The heuristic run for the normalized variant yielded the same set of species trees and the normalized cost 8.9809.

TreeFam dataset. The second dataset consists of 1274 curated gene family trees from TreeFam v7.0 [27] spanning 25 mostly animal species. The gene trees were rooted by using FastUrec [9] with the species tree based on the NCBI taxonomy (see Figure 2b). Median size of a gene tree in this dataset is 31.80. Multiple runs of our program inferred a single optimal species tree with 7451 gene duplications in total (see Figure 2a). The same tree was inferred for the normalized variant of the duplication cost with 177.3693.

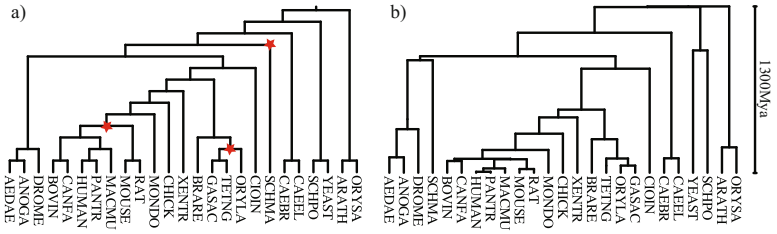

Fig. 2. Species trees for TreeFam dataset. Left: an optimal tree inferred by our heuristic. Right: Species tree based on NCBI taxonomy with branch lengths obtained from diversification dates of the TreeTime database [15]. Stars in the left tree denote the differences between both trees. Note that lengths of the branches of the left tree are not proportional to the time.

The total runtime for these computational experiments was approximately 3 hours on a server with 64 cores (8 Opteron AMD processors, 1400 MHz). The prototype computer programs are available on request.

5 Conclusions and Future Outlook

In this article we investigated into diameters of the duplication cost under several variants and two types of leaf labelings. We proved mathematical properties describing these diameters. Based on these properties we proposed simple formulas for the diameters and efficient algorithms to compute the diameters and their corresponding leaf labelings. In particular we presented an optimal, linear time, algorithm for the bijective case when the shapes of both input trees are fixed.

This is a continuation of our previous research on the deep coalescence diameters [10], that lays foundations for further study on the diameters of other reconciliation based cost functions, e.g., duplication-loss or loss costs [11,23,28].

Our GTP experiments for the duplication cost show no difference between optimal species trees inferred for the standard and normalized variants of the cost. This is likely due to the low resolution of the duplication cost function. However, our experiments are only based on two examples, and future studies will include more complex analyses. Furthermore, we will also investigate in expected values of diameters under various phylogenetic models.

Acknowledgements. The authors wish to thank the three anonymous referees whose constructive suggestions and criticisms have helped considerably to improve the quality of this paper. We also would like to thank Prof. J. Tiuryn for helpful comments. This work was conducted as a part of the Gene Tree Reconciliation Working Group at the National Institute for Mathematical and Biological Synthesis, sponsored by the U.S. National Science Foundation, the U.S. Department of Homeland Security, and the U.S. Department of Agriculture through NSF Award #EF-0832858, with additional support from The University of Tennessee, Knoxville. Support was provided by the grant of NCN (2011/01/B/ST6/02777), and to O. Eulenstein by the U.S. National Science Foundation Award #CCF-1017189.

References

1. Bansal, M.S., Eulenstein, O.: Algorithms for genome-scale phylogenetics using gene tree parsimony. IEEE/ACM Trans. Comput. Biol. Bioinform. 10(4), 939–956 (2013)
2. Bansal, M.S., Eulenstein, O., Wehe, A.: The gene-duplication problem: near-linear time algorithms for nni-based local searches. IEEE/ACM Trans. Comput. Biol. Bioinform. 6(2), 221–231 (2009)
3. Chang, W.-C., Górecki, P., Eulenstein, O.: Exact solutions for species tree inference from discordant gene trees. J. Bioinform. Comput. Biol. 11(05), 1342005 (2013)
4. Cotton, J.A., Page, R.D.M.: Going nuclear: gene family evolution and vertebrate phylogeny reconciled. Proc. Biol. Sci. 269(1500), 1555–1561 (2002)
5. Eulenstein, O., Huzurbazar, S., Liberles, D.A.: Reconciling Phylogenetic Trees. In: Evolution after Gene Duplication. John Wiley, Hoboken (2010)
6. Felsenstein, J.: The number of evolutionary trees. Syst. Zool. 27, 27–33 (1978)
7. Goodman, M., Czelusniak, J., Moore, G.W., Romero-Herrera, A.E., Matsuda, G.: Fitting the gene lineage into its species lineage. A parsimony strategy illustrated by cladograms constructed from globin sequences. Systematic Zoology 28, 132–163 (1979)
8. Górecki, P., Eulenstein, O.: Algorithms: simultaneous error-correction and rooting for gene tree reconciliation and the gene duplication problem. BMC Bioinformatics 13(suppl. 10), S14 (2012)
9. Górecki, P., Burleigh, J.G., Eulenstein, O.: GTP supertrees from unrooted gene trees: Linear time algorithms for NNI based local searches. In: Bleris, L., Măndoiu, I., Schwartz, R., Wang, J. (eds.) ISBRA 2012. LNCS, vol. 7292, pp. 102–114. Springer, Heidelberg (2012)
10. Górecki, P., Eulenstein, O.: Maximizing deep coalescence cost. In: Accepted to IEEE/ACM Trans. Comput. Biol. Bioinform. (2014), preprint is available at `http://dx.doi.org/10.1109/TCBB.2013.144`
11. Górecki, P., Eulenstein, O., Tiuryn, J.: Unrooted tree reconciliation: A unified approach. IEEE/ACM Trans. Comput. Biol. Bioinform. 10(2), 522–536 (2013)
12. Górecki, P., Tiuryn, J.: DLS-trees: A model of evolutionary scenarios. Theor. Comput. Sci. 359(1-3), 378–399 (2006)
13. Guigó, R., Muchnik, I.B., Smith, T.F.: Reconstruction of ancient molecular phylogeny. Mol. Phylogenet. Evol. 6(2), 189–213 (1996)
14. Harris, S.R., Cartwright, E.J.P., Török, M.E., Holden, M.T.G., Brown, N.M., Ogilvy-Stuart, A.L., Ellington, M.J., Quail, M.A., Bentley, S.D., Parkhill, J., Peacock, S.J.: Whole-genome sequencing for analysis of an outbreak of meticillin-resistant staphylococcus aureus: a descriptive study. Lancet Infect. Dis. 13(2), 130–136 (2013)
15. Hedges, S.B., Dudley, J., Kumar, S.: Timetree: a public knowledge-base of divergence times among organisms. Bioinformatics 22(23), 2971–2972 (2006)

16. Hufbauer, R.A., et al.: Population structure, ploidy levels and allelopathy of Centaurea maculosa (spotted knapweed) and C. diffusa (diffuse knapweed) in North America and Eurasia. In: Proceedings of the International Symposium on Biological Control of Weeds, pp. 121–126 (2003)
17. Jackson, A.P.: A reconciliation analysis of host switching in plant-fungal symbioses. Evolution 58(9), 1909–1923 (2004)
18. Lafond, M., Swenson, K.M., El-Mabrouk, N.: An optimal reconciliation algorithm for gene trees with polytomies. In: Raphael, B., Tang, J. (eds.) WABI 2012. LNCS (LNBI), vol. 7534, pp. 106–122. Springer, Heidelberg (2012)
19. Ma, B., Li, M., Zhang, L.: From gene trees to species trees. SIAM Journal on Computing 30(3), 729–752 (2000)
20. Martin, A.P., Burg, T.M.: Perils of paralogy: using hsp70 genes for inferring organismal phylogenies. Syst. Biol. 51(4), 570–587 (2002)
21. McGowen, M.R., Clark, C., Gatesy, J.: The vestigial olfactory receptor subgenome of odontocete whales: phylogenetic congruence between gene-tree reconciliation and supermatrix methods. Syst. Biol. 57(4), 574–590 (2008)
22. Nik-Zainal, S., et al.: The life history of 21 breast cancers. Cell 149(5), 994–1007 (2012)
23. Page, R.D.M.: Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. Systematic Biology 43(1), 58–77 (1994)
24. Page, R.D.M.: Extracting species trees from complex gene trees: reconciled trees and vertebrate phylogeny. Mol. Phylogenet. Evol. 14, 89–106 (2000)
25. Page, R.D.M., Holmes, E.C.: Molecular evolution: a phylogenetic approach. Blackwell Science (1998)
26. Rokas, A., Williams, B.L., King, N., Carroll, S.B.: Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425, 798–804 (2003)
27. Ruan, J., et al.: TreeFam: 2008 Update. Nucleic Acids Res. 36, D735–D740 (2008)
28. Than, C.V., Rosenberg, N.A.: Mathematical properties of the deep coalescence cost. IEEE/ACM Trans. Comput. Biol. Bioinform. 10(1), 61–72 (2013)
29. Wehe, A., Bansal, M.S., Burleigh, G.J., Eulenstein, O.: DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. Bioinformatics 24(13), 1540–1541 (2008)
30. Wehe, A., Burleigh, J.G., Eulenstein, O.: Efficient algorithms for knowledge-enhanced supertree and supermatrix phylogenetic problems. IEEE/ACM Trans. Comput. Biol. Bioinform. 10(6), 1432–1441 (2013)