

Phylogenetic Tree Reconciliation: Mean Values for Fixed Gene Trees

Paweł Górecki¹, Alexey Markin², Agnieszka Mykowiecka¹, Jarosław Paszek¹,
and Oliver Eulenstein²

¹ Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, PL
{gorecki|agnieszka.mykowiecka|j.paszek}@mimuw.edu.pl

² Department of Computer Science, Iowa State University, USA
{amarkin|oeulenstein}@iastate.edu

Abstract. Phylogenetic tree reconciliation is a widely used approach for analyzing the inconsistencies between the evolutionary histories of genes, and the species through which they have evolved. An important aspect of tree reconciliation are the cost functions involved that are the minimum number of evolutionary events explaining such inconsistencies. Mean values for these functions are fundamental when analyzing tree reconciliations. Here we describe mean value formulas when a history of genes is fixed for the cost functions for the events gene duplication, gene loss and gene duplication-loss, under the uniform model of species trees. We show that these formulas can be efficiently computed, and finally analyze the mean values using empirical and simulated data.

Keywords: tree reconciliation, duplication-loss model, deep coalescence, speciation, gene duplication, gene loss, bijectively labelled tree, uniform model of trees, mean value

1 Introduction

Phylogenetic tree reconciliation is a powerful tool for analyzing the inconsistencies between the evolutionary histories of genes, and the species through which they have evolved. Through algorithmic advances in tree reconciliation such analyses have become common practice in various biological research areas, such as molecular biology and microbiology [21]. For example tree reconciliation is used to illuminate the dynamics of gene family evolution in terms of complex evolutionary processes [5,20]. Reconciling trees is also one of the most reliable approaches for identifying truly orthologous genes [1,2], which is a fundamental task in understanding the evolution of genetic function [19].

Tree reconciliation is a process that takes two trees as input, a *gene tree* that is the evolutionary history of genes, and a *species tree* that is the evolutionary history of the species hosting the genes. It seeks an embedding of the gene tree into the species tree (i.e., the evolution of the gene tree along the branches of the species tree) that explains possible inconsistencies between the two trees by

inferring the minimum number of evolutionary events, such as gene duplication, gene loss, the combination of gene duplication-loss, and deep coalescence.

An important aspect of tree reconciliation is its associated cost that is the (minimum) number of evolutionary events inferred by the process. This, for example, allows the comparative analysis of gene trees in the context of their corresponding species trees [25,26], which is a standard approach for synthesizing large-scale species trees from collections of discordant gene trees [3,6].

The widespread usage of tree reconciliation in practice has led to a growing interest in analyzing reconciliation cost functions. This includes analyzing the *diameters* of such functions that are the maximum costs when one or both tree topologies are given [11,12,14,13]. More recently, the mean values of reconciliation cost functions have been studied when either a gene tree or a species tree is given. The *mean value for a gene tree* for a reconciliation cost function is the mean of the costs between the gene tree and all of its corresponding species trees. The *mean value of a species tree* is defined similarly. These mean values have been studied under two classic probability models for phylogenetic trees that are the uniform model and the Yule-Harding model [18,24,28].

Here we study the mean values for a gene tree under the uniform distribution for the the reconciliation functions for each of the events, gene duplication and loss, gene duplication, and gene loss.

Previous Work. The pioneering work of Goodman et al. [9] introduced the approach for reconciling a gene tree with a corresponding species tree, where both of these trees are rooted and full binary. This approach is embedding the gene tree into the species tree using a *mapping* that relates every gene in the gene tree to its *host species* that is the most recent species that could have contained the gene. Consequently, the mapping is relating every leaf-gene of the gene tree to the species from which it has been sampled. When restricted to the leaf-genes, the mapping is referred to as *leaf-labeling*. Based on this mapping the evolutionary events, gene duplication, gene loss, and the combination of gene duplication and subsequent loss (in short, duplication-loss) are identified. A gene is a *gene duplication* when it has a child with the same host species, and a *gene loss* is accounted for by a maximum subtree in the species tree that has no host species (i.e., no mapping from the gene tree). While other embeddings are possible [15] the mapping describes the most parsimonious embedding in terms of the number of gene duplication and loss events [7,4,15]. The reconciliation cost function associated with each of these events counts the number of their occurrences in terms of gene duplications, gene losses, and gene duplications plus losses, and are termed *duplication*, *loss*, and *duplication-loss* cost functions respectively. The deep coalescence cost function, introduced by Maddison [22], is also based on the reconciliation approach. Edges in the species tree may have embedded edges from the gene tree, which are called *lineages*. The *deep coalescence cost function* counts for every edge in the species tree the number of lineages minus one, which are thought to be caused by deep coalescence events. From the mathematical point of view, the gene loss cost function is a linear combination of gene duplication and deep coalescence cost functions [16,31], and therefore, any prop-

erty derived for these two functions can naturally be translated into gene loss and gene duplication-loss cost functions. All of the described reconciliation functions have been defined for general leaf-labelings and for bijective leaf-labelings.

The focus of this work are the mean values of the described reconciliation cost functions for bijective leaf-labelings under the uniform distribution of phylogenetic trees. Mean value formulas have been described for a given species tree for the deep coalescence cost function [29]. More recently such formulas have also been described for the gene duplication, gene loss, and gene duplication-loss cost functions [17]. For the computation time to obtain these mean values let n be the size of the given species tree. The mean values for a given species tree under the uniform model can be computed in $O(n)$ time for the deep coalescence cost function, and in time $O(n^3)$ for the gene duplication, gene loss, and gene duplication-loss cost functions [17]. Mean value formulas for a given gene tree have only been described for the deep coalescence cost function [29], and this value is computable in $O(n)$ time, where n is the size of the given gene tree.

Our Contributions. In this article we develop the formulas to compute the mean values for the reconciliation cost using gene duplication and loss, gene duplication, and gene loss events when the gene tree is given under a uniform distribution for the species trees. We show that these formulas can be computed in time $O(n^3)$ for a given gene tree of size n . Finally, we conducted comparative studies for fixed gene and fixed species tree means for our reconciliation costs and performed an analysis of an empirical dataset consisting of thousands of gene family trees.

2 Basic definitions

We follow the basic definitions and notation from [16,31]. Let X be a non-empty set of n species (taxa). The set of all full binary and rooted trees whose leaves are bijectively labeled by the species in X is denoted by $R(X)$. Trees in $R(X)$ are denoted by using the standard nested parenthesis notation. Given a tree $T \in R(X)$, we denote its node and edge sets by V_T and E_T respectively. The root of T is denoted by $\text{root}(T)$ and the parent of a non-root node v is denoted by $\text{par}(v)$. We denote the least common ancestor of nodes $v, w \in V_T$ in tree T by $\text{lca}_T(v, w)$. A *cluster* (or also called *clade*) of a node $v \in V_T$ is the set of all leaf labels of the subtree of T rooted at v .

In phylogenetic tree reconciliation a gene tree is embedded into its corresponding species tree. In this work we assume that both types of trees have the same bijective labelling of leaves. Therefore, we assume that every gene tree and every species tree is an element of $R(X)$. For a (gene) tree $G \in R(X)$ and a (species) tree $S \in R(X)$ the *least common ancestor mapping between G and S* , or *lca-mapping*, $M: V_G \rightarrow V_S$, is defined as $M(g) = s$ if g and s are leaves with the same label, and $M(g) = \text{lca}_S(M(g'), M(g''))$ if g has two children g' and g'' . An internal node g is called a *duplication*, or an *S-duplication*, if $M(g) = M(a)$ for a child a of g . Every internal non-duplication node is called a *speciation*. The *duplication cost*, denoted by $D(G, S)$, is the total number of *S-duplications*

in G [25]. The deep coalescence cost function [22,23,31] can be expressed by $\text{DC}(G, S) := \sum_{g \in V_G \setminus \{\text{root}(G)\}} (\|\mathbf{M}(g), \mathbf{M}(\text{par}(g))\| - 1)$, where $\|a, b\|$ is the number of edges on the simple path connecting nodes $a, b \in S_V$. The reader is referred to [29] for alternative definitions of DC . Finally, we can provide formulas for the loss and duplication-loss cost functions [31]: $\text{L}(G, S) := 2 \text{D}(G, S) + \text{DC}(G, S)$ and $\text{DL}(G, S) := \text{D}(G, S) + \text{L}(G, S)$. For a more detailed introduction to the model please refer to [15,22,25].

3 Results

In the uniform model of binary trees an equal probability is assigned to each possible leaf labeled binary tree with n leaves. In this model rooted trees can be generated by uniform and random insertions of one edge to any edge including the rooting edge at each step. For example, given a rooted tree $(a, (b, c))$, the following five four-labelled trees can be created by inserting a new edge with a leaf d : $((a, d), b), c)$, $((a, (b, d)), c)$, $((a, b), d), c)$, $((a, b), (c, d))$, and $((a, b), c), d)$.

We analyse the mean of the duplication cost in the uniform model of rooted leaf-labeled trees. Let $R(X)$ denote the set of all bijectively labeled rooted trees over a non-empty set X . Then, the mean of duplication cost for a fixed gene tree $G \in R(X)$ under a probabilistic model of species trees is:

$$\bar{\text{D}}_u(G) = \sum_{S \in R(X)} \mathbb{P}(S) \text{D}(G, S). \quad (1)$$

Recall that size of $R(X)$ is $b(n) = (2n - 3)!!$, where $k!!$ is the double factorial, i.e., $k!! = k \cdot (k - 2)!!$ and $0!! = (-1)!! = 1$. Hence, in the uniform model for every tree $T \in R(X)$ has probability $\mathbb{P}(T) = \frac{1}{b(n)}$.

Now we introduce a notion of a (rooted) split. Every non-leaf node $v \in V_T$, induces a *split* $A|B$, where A and B are the clusters of children of v . The set of all splits in T is denoted by $\text{Spl}(T)$. As an example, $\text{Spl}(((a, b), (c, d))) = \{\{\{a, b\}, \{c, d\}\}, \{\{a\}, \{b\}\}, \{\{c\}, \{d\}\}\}$, which we describe by using the simplified split notation: $\{ab|cd, a|b, c|d\}$.

For a split $A|B$ induced by a node v from a fixed gene tree $G \in R(X)$, by $\xi_n^{\text{Dup}}(A, B)$ we denote the number of species trees S from $R(X)$ such that v is an S -duplication node. Similarly, we define $\xi_n^{\text{Spec}}(A, B)$ for speciation nodes.

Lemma 1. *For a gene tree G with n leaves,*

$$\sum_{A|B \in \text{Spl}(G)} \xi_n^{\text{Dup}}(A, B) + \xi_n^{\text{Spec}}(A, B) = b(n) \cdot (n - 1).$$

Now, the mean (1) is equivalent to

$$\bar{\text{D}}_u(G) = \frac{1}{b(n)} \sum_{A|B \in \text{Spl}(G)} \xi_n^{\text{Dup}}(A, B) = n - 1 - \frac{1}{b(n)} \sum_{A|B \in \text{Spl}(G)} \xi_n^{\text{Spec}}(A, B). \quad (2)$$

Similarly to [17], it is more convenient to count directly the number of speciation nodes rather than duplications.

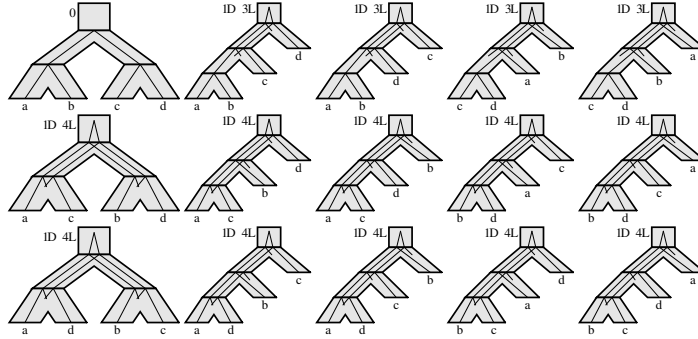


Fig. 1. Embeddings (scenarios) of $G = ((a, b), (c, d))$ into every species four-leaf species tree [15]. Each scenario is summarized with two numbers denoting the number of gene duplications (D) and the number of gene losses (L). We have 14 gene duplications, 31 speciation nodes and 52 gene losses in total. In this example, $\overline{D}_u(G) = 14/15$, $\overline{L}(G) = 52/15$ and $\overline{DL}_u(G) = 66/15$.

Lemma 2. For a species tree G with n leaves and a split $A|B$ present in G

$$\xi_n^{\text{Spec}}(A, B) = \sum_{i=0}^m \sum_{j=0}^{m-i} \binom{m}{i} \binom{m-i}{j} b(|A|+i)b(|B|+j)b(m-i-j+1).$$

where $m = n - |A| - |B|$.

Proof. Let $v \in G$ has the split $A|B$. A species tree S that induces a speciation node v mapped into a node s from S can be constructed as follows. Let z be an element not in X . Let A' and B' be two disjoint supersets of A and B , respectively. Then, a species tree $S \in R(X)$ such that s has split $A'|B'$ can be constructed by replacing the leaf z in a tree $R((X \setminus (A' \cup B')) \cup \{z\})$ by a tree (S_A, S_B) such that $S_A \in R(A')$ and $S_B \in R(B')$. Then, v is a speciation node mapped to the root of (S_A, S_B) in S . On the other hand note that every S such that v from G is a speciation node mapped to a node in S , is inferred exactly once in the above procedure. \square

Now, we can state the main result that follows from Lemma 2 and Eq. 2.

Theorem 1 (Fixed gene tree mean of D under the uniform model). For a given gene tree G with n leaves

$$\overline{D}_u(G) = n-1 - \frac{1}{b(n)} \sum_{\substack{A|B \in \text{Sp}(G) \\ m=n-|A|-|B|}} \sum_{i=0}^m \sum_{j=0}^{m-i} \binom{m}{i} \binom{m-i}{j} b(|A|+i)b(|B|+j)b(m-i-j+1).$$

To obtain the mean formula for DL cost we recall the result from [29] (see Cor. 13) on the deep coalescence cost. For a gene tree G with n leaves:

$$\overline{DC}_u(G) = -(2n-1) + 2n \frac{(2n-2)!!}{b(n)} - \frac{(2n-2)!!}{b(n)} \sum_{v \in V_G} \frac{(2|C_v|-3)!!}{(2|C_v|-2)!!},$$

where C_v denotes the cluster of a node v .

Finally, we have the result for DL and L (see also similar results for fixed species tree from [17]).

Theorem 2 (Fixed gene tree mean of DL and L). *For a gene tree G we have $\overline{DL}_u(G) = 3 \cdot \overline{D}_u(G) + \overline{DC}_u(G)$ and $\overline{L}_u(G) = 2 \cdot \overline{D}_u(G) + \overline{DC}_u(G)$.*

Proof. It follows from the definition of gene loss and duplication-loss functions and the properties of mean values. \square

Given the mean formulas for DC and D it is now straightforward to obtain the exact formulas for the means of DL and L. We omit these details for brevity. See an example of mean values depicted in Fig. 1.

Computing the mean of deep coalescence for a fixed gene tree can be completed in $O(n)$ steps under assumption that double factorials are memorized and the required size of clusters is stored with the nodes of the standard pointer-like implementation of trees. For the mean of the remaining cost functions, however, we need two additional loops. Therefore, the time complexity of computing $\overline{D}_u(G)$, $\overline{L}_u(G)$ and $\overline{DL}_u(G)$ is $O(n^3)$.

4 Experimental evaluation

4.1 Mean values for tree shapes

Here we analyze the mean values of our analyzed reconciliation cost functions for all tree shapes with 3, 4, \dots 9 leaves ordered by their Furnas rank [8], which are depicted in Table 1. We observe that tree shapes with the same number of splits induce the same mean values (e.g., the two red colored tree shapes) which follows directly from the mean value formulas for deep coalescence and duplication cost functions. This property also holds for the mean values when a species tree is fixed [17]. Note, while in [17] the mean value of the duplication cost function for a fixed species tree was conjectured to grow monotonically with the Furnas rank, this is not the case for the corresponding mean values when a gene tree is fixed as indicated in Table 1. Moreover, we can observe that the mean value of the duplication cost function is maximum for caterpillar trees while it is minimum for the most balanced once.

Moreover, we compared the mean values for fixed species tree shapes from [17] with their corresponding values when the gene tree is fixed. Therefore, we computed the mean values for all gene tree shapes with up to 20 leaves, e.g., for $n = 20$ there are 293547 trees. Fig. 2 depicts two diagrams which represent the means for a fixed species tree shapes [17] and the corresponding means for a fixed gene tree shapes, respectively. While we are expecting that the blue ovoids and the red ovoids will increasingly overlap with an increasing number of taxa, we observe that this occurs earlier (i.e., for smaller sizes of taxa) for species tree shapes. For the duplication cost function we observe a broader range of means in the upper diagram, while the range for the other cost functions appears to

be broader for the species tree shapes. In combination with our previous observations from Table 1, we conclude that the properties of the duplication cost function differs significantly when comparing the two types of fixed tree means.

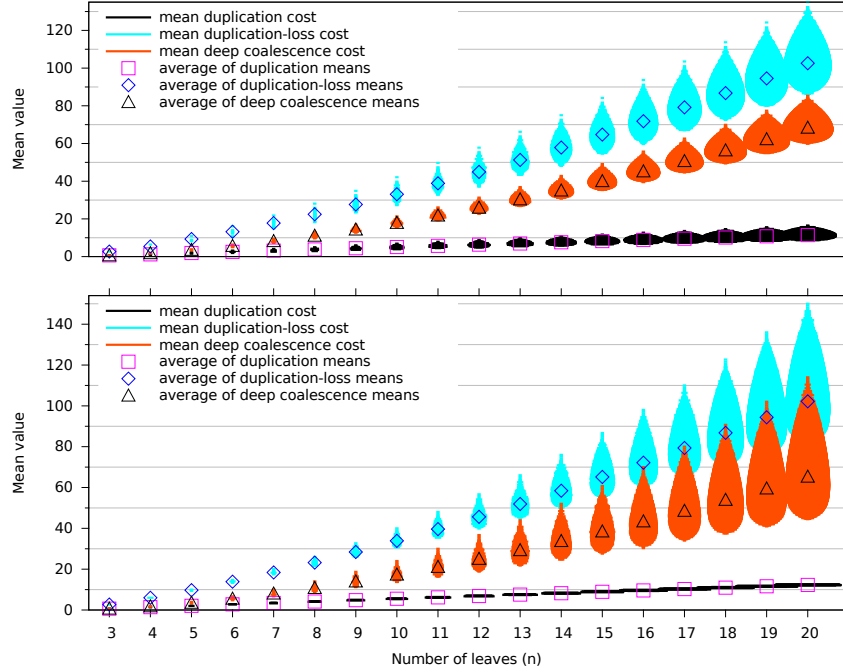


Fig. 2. *Top:* Frequency diagram of mean values of duplication, duplication-loss and deep coalescence costs for all fixed gene tree shapes for $n = 3, 4, \dots, 20$ under the uniform model of species trees. For each n , mean values for every cost were grouped into bins of size 0.01. The width of each bin is proportional to $\log_2 K$, where K is the number of gene tree shapes having the mean value in this bin. *Bottom:* The same type of diagram for means of fixed species tree taken from [17].

4.2 Empirical study

In this section we study the distribution of mean values for the duplication and duplication-loss cost functions for gene trees obtained from a baseline empirical dataset. Additionally, we evaluate how the duplication and duplication-loss costs compare to the respective mean values

Empirical dataset To evaluate the distributions of mean values on empirical phylogenetic datasets we analyzed the classic *TreeFam ver.9* dataset [27] consisting of gene family trees of 109 mostly animal species (with 71 taxa in the gene family trees on average). Among around 15 thousand rooted gene trees in the dataset, the 4070 bijectively labeled and strictly bifurcated trees were selected.

We further filtered the trees based on their size; that is, we removed all trees with less than 10 leaves in order to eliminate otherwise arising outliers due to insufficient tree size.

Given that the best-known species tree for the TreeFam dataset is not completely refined (contains many large multifurcations), we estimated the species tree using a popular supertree tool, *duptree2* [30]; the tool approximates a species tree that minimizes the duplication cost for the given set of gene trees.

Experimental setting. In order to compare mean values for gene trees of different sizes and topologies we need to bring them up to the same scale. We achieve this by normalizing the mean values by respective diameters. Note that diameters under fixed gene tree topologies can be computed exactly, both for the duplication and duplication-loss cost functions [13,10].

To assess the mean value distributions for trees taken from the empirical dataset, we compare them to *complete* distributions for trees of fixed size. That is, for a fixed number of leaves, t , we compute mean values for all possible tree topologies with t leaves. This is repeated for $t = 10, 12, 14$, and 16. Apart from serving as a complete distribution reference, these data also allows us to empirically observe how the mean-value distributions progress with the increase of taxa.

Results and discussion. Figure 3 illustrates that the mean values under the duplication cost function for the TreeFam gene trees are concentrated around the value 0.9. That is, the mean values are very close to respective cost diameters, which implies that for all the trees under consideration, most of possible species trees have a very high (close to the maximum) duplication cost. It also suggests that the proximity of a duplication cost (normalized by the diameter) to 0 indicates a high confidence in the species tree.

Further, the complete distributions of duplication means for all possible tree topologies over varying taxa size are shown on Figure 4 (left hand side) closely resemble the distribution on empirical datasets. The figure also demonstrates that the duplication-mean distribution does not seem to change much with the increase of taxa.

The empirical distribution for the duplication-loss means on Figure 3 (left-hand side, red histogram) is rather spread on the interval from approximately 0.25 to 0.7 with multiple picks. Figure 4 (right hand side) additionally shows that duplication-loss mean values (normalized by the respective diameters) gradually decrease with increasing taxa number. Given that the TreeFam dataset contains trees of varying size, the shifts in mean values for gene trees of larger size, explain the wide range of duplication-loss means on Figure 4.

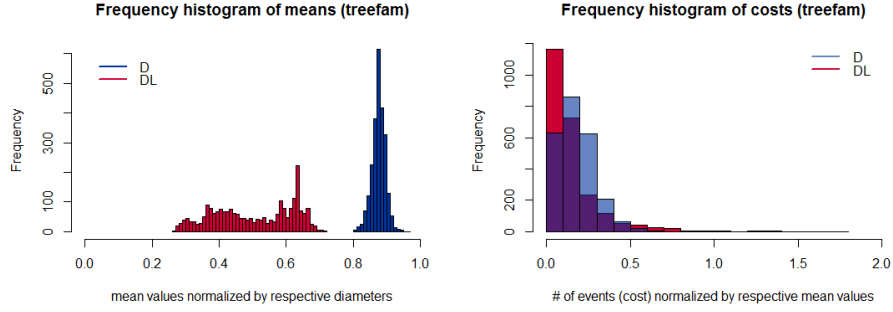


Fig. 3. Comparison of (i) mean values normalized by diameters and (ii) costs normalized by mean values for the duplication (D) and the duplication-loss (DL) cost functions (TreeFam dataset).

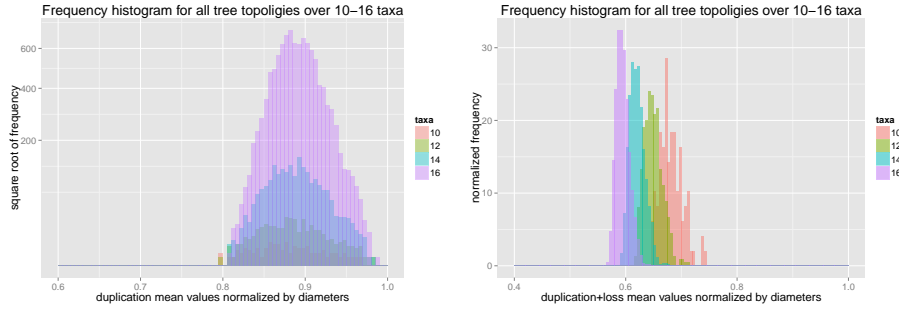


Fig. 4. *Left:* Distribution of *duplication*-mean values normalized by respective diameters. The frequencies of the histogram were scaled by a square root to achieve a more comprehensive visualization. *Right:* Distribution of *duplication-loss*-mean values normalized by respective diameters. Distributions are shown for all possible tree topologies over 10, 12, 14, and 16 taxa respectively.

Further, the mean values play an important role in the normalization of reconciliation costs, since it allows us to relate reconciliation costs that are otherwise significantly affected by topologies of the gene trees. The histogram on the right-hand side of Figure 3 shows duplication and duplication-loss costs normalized by respective mean values. While the majority of trees are concentrated below the value 0.5 (i.e., the cost is significantly smaller than the respective mean), there are some outliers for which the cost is close to the mean or even exceeds it. Such trees can be thought of as not strongly correlating with the corresponding species tree (or even correlating negatively), and they can represent gene families of interest for a researcher. Alternatively, when the reconciliation cost between

a gene tree and a species tree exceeds the mean value, it might indicate possible errors in the gene tree.

5 Conclusions

In this work we have developed the mean value formulas for a fixed gene tree for the gene duplication, gene loss and gene duplication-loss cost functions under the uniform model of species trees. We have also shown that these mean values can be efficiently computed. Our comparative experiments demonstrate that there can be fundamental differences between fixed species tree and fixed gene tree means. This motivates further analyzes that may establish deeper mathematical insights into mean values and the relations between them. Our future research in mean values of tree shapes will dovetail with these ideas.

Acknowledgements

This material is based upon work supported by the grants of the National Science Foundation under Grant No. 1617626 and the NCN #2015/19/B/ST6/00726.

References

1. Akerborg, O., Sennblad, B., Arvestad, L., Lagergren, J.: Simultaneous bayesian gene tree reconstruction and reconciliation analysis. *Proc Natl Acad Sci U S A* 106(14), 5714–9 (Apr 2009)
2. Altenhoff, A.M., Dessimoz, C.: Inferring orthology and paralogy. *Methods Mol Biol* 855, 259–79 (2012)
3. Bininda-Emonds, O.R. (ed.): *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, Computational Biology, vol. 4. Springer Verlag (2004)
4. Bonizzoni, P., Della Vedova, G., Dondi, R.: Reconciling a gene tree to a species tree under the duplication cost model. *Theoretical Computer Science* 347(1-2), 36–53 (2005)
5. David, L.A., Alm, E.J.: Rapid evolutionary innovation during an archaean genetic expansion. *Nature* 469(7328), 93–6 (Jan 2011)
6. Eulenstein, O., Huzurbazar, S., Liberles, D.: Evolution after Gene Duplication, chap. Reconciling Phylogenetic Trees, pp. 185–206. John Wiley & Sons, Inc. (2010)
7. Eulenstein, O.: Vorhersage von Genduplikationen und deren Entwicklung in der Evolution. Ph.D. thesis, Rheinische Friedrich-Wilhelms-Universität Bonn, 53012 Bonn, Germany (1998)
8. Furnas, G.W.: The generation of random, binary unordered trees. *Journal of Classification* 1(1), 187–233 (1984)
9. Goodman, M., et al.: Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Zoology* 28(2), 132–163 (1979)
10. Górecki, P., Eulenstein, O.: Bijective diameters of gene tree parsimony costs. Submitted

11. Górecki, P., Eulenstein, O.: Deep coalescence reconciliation with unrooted gene trees: Linear time algorithms. LNCS 7434, 531–542 (2012)
12. Górecki, P., Eulenstein, O.: Maximizing deep coalescence cost. IEEE-ACM Transactions on Computational Biology and Bioinformatics 11(1), 231–242 (2014)
13. Górecki, P., Eulenstein, O.: Gene tree diameter for deep coalescence. IEEE-ACM Transactions on Computational Biology and Bioinformatics 12(1), 155–165 (2015)
14. Górecki, P., Paszek, J., Eulenstein, O.: Unconstrained gene tree diameters for deep coalescence. In: Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics. pp. 114–121. BCB '14, ACM, New York, NY, USA (2014)
15. Górecki, P., Tiuryn, J.: DLS-trees: A model of evolutionary scenarios. Theoretical Computer Science 359(1-3), 378–399 (2006)
16. Górecki, P., Eulenstein, O., Tiuryn, J.: Unrooted tree reconciliation: A unified approach. IEEE-ACM Transactions on Computational Biology and Bioinformatics 10(2), 522–536 (2013)
17. Górecki, P., Paszek, J., Mykowiecka, A.: Mean values of gene duplication and loss cost functions. Lecture Notes in Computer Science 9683, 189–199 (2016)
18. Harding, E.F.: The probabilities of rooted tree-shapes generated by random bifurcation. Advances in Applied Probability 3(1), pp. 44–77 (1971)
19. Ihara, K., Umemura, T., Katagiri, I., Kitajima-Ihara, T., Sugiyama, Y., Kimura, Y., Mukohata, Y.: Evolution of the archaeal rhodopsins: evolution rate changes by gene duplication and functional differentiation. J Mol Biol 285(1), 163–74 (1999)
20. Kamneva, O.K., Knight, S.J., Liberles, D.A., Ward, N.L.: Analysis of genome content evolution in pvc bacterial super-phylum: assessment of candidate genes associated with cellular organization and lifestyle. Genome Biol Evol 4(12), 1375–90 (2012)
21. Kamneva, O.K., Ward, N.L.: Chapter 9 - reconciliation approaches to determining hgt, duplications, and losses in gene trees. In: Michael Goodfellow, I.S., Chun, J. (eds.) New Approaches to Prokaryotic Systematics, Methods in Microbiology, vol. 41, pp. 183 – 199. Academic Press (2014)
22. Maddison, W.P.: Gene trees in species trees. Systematic Biology 46, 523–536 (1997)
23. Maddison, W.P., Knowles, L.L.: Inferring phylogeny despite incomplete lineage sorting. Systematic Biology 55(1), 21–30 (2006)
24. McKenzie, A., Steel, M.: Distributions of cherries for two models of trees. Mathematical Biosciences 164(1), 81 – 92 (2000)
25. Page, R.: From Gene to Organismal Phylogeny: Reconciled Trees and the Gene Tree/Species Tree Problem. Molecular Phylogenetics and Evolution 7(2), 231–240 (1997)
26. 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)
27. Ruan, J., et al.: TreeFam: 2008 Update. Nucleic Acids Research 36, D735–40 (2008)
28. Steel, M.A., Penny, D.: Distributions of tree comparison metrics — some new results. Systematic Biology 42(2), 126–141 (1993)
29. Than, C.V., Rosenberg, N.A.: Mean deep coalescence cost under exchangeable probability distributions. Discrete Applied Mathematics 174, 11–26 (2014)
30. Wehe, A., Burleigh, J.G.: Scaling the gene duplication problem towards the tree of life: accelerating the rspr heuristic search (2010)
31. Zhang, L.: From gene trees to species trees ii: Species tree inference by minimizing deep coalescence events. IEEE-ACM Transactions on Computational Biology and Bioinformatics 8, 1685–1691 (2011)










$n \leq 6$	 0.67 2.00 2.67 0.67	 1.47 4.93 6.40 2.00	 0.93 3.47 4.40 1.60	 2.32 8.59 10.91 3.94	 1.79 7.07 8.86 3.49	 1.64 6.53 8.17 3.26	 3.21 12.87 16.08 6.44	 2.68 11.30 13.97 5.94	 2.53 10.74 13.26 5.68	 2.45 10.43 12.88 5.52	 1.92 8.85 10.77 5.02	 2.32 9.92 12.24 5.27
$n = 7$	 4.12 17.71 21.83 9.46	 3.59 16.09 19.67 8.91	 3.44 15.50 18.94 8.63	 3.36 15.18 18.54 8.46	 2.83 13.56 16.38 7.90	 3.23 14.65 17.88 8.18	 3.31 14.97 18.28 8.34	 2.78 13.34 16.12 7.78	 2.63 12.76 15.39 7.51	 3.13 14.15 17.27 7.89	 2.60 12.52 15.12 7.33	
$n = 8$	 5.04 23.05 28.10 12.97	 4.51 21.39 25.90 12.37	 4.36 20.79 25.15 12.07	 4.28 20.45 24.73 11.88	 3.75 18.79 22.54 11.29	 4.16 19.90 24.05 11.59	 4.24 20.23 24.46 11.75	 3.70 18.56 22.27 11.16	 3.55 17.96 21.51 10.86	 4.05 19.37 23.42 11.27	 3.52 17.71 21.23 10.67	 4.21 20.07 24.27 11.66
$n = 9$	 5.98 28.88 34.86 16.92	 5.44 27.17 32.62 16.29	 5.29 26.55 31.84 15.97	 5.22 26.20 31.42 15.77	 4.68 24.50 29.18 15.13	 5.09 25.63 30.72 15.45	 5.17 25.97 31.14 15.63	 4.64 24.27 28.90 14.99	 4.48 23.64 28.13 14.67	 4.99 25.08 30.07 15.11	 4.45 23.38 27.83 14.47	 5.14 25.80 30.94 15.52
	 4.61 24.10 28.70 14.89	 4.45 23.48 27.93 14.57	 4.38 23.12 27.50 14.37	 3.84 21.42 25.27 13.73	 4.25 22.55 26.80 14.05	 4.92 24.71 29.63 14.87	 4.39 23.01 27.40 14.23	 4.24 22.38 26.62 13.91	 4.87 24.40 29.26 14.67	 4.33 22.70 27.03 14.03	 3.80 20.99 24.79 13.40	 5.12 25.68 30.79 15.44
	 4.58 23.97 28.56 14.81	 4.43 23.35 27.78 14.49	 4.35 23.00 27.35 14.29	 3.82 21.29 25.12 13.65	 4.23 22.43 26.65 13.97	 4.31 22.77 27.07 14.15	 3.78 21.06 24.84 13.51	 3.62 20.44 24.06 13.19	 4.12 21.88 26.00 13.63	 3.59 20.18 23.77 13.00	 4.88 24.44 29.32 14.68	 4.34 22.73 27.08 14.04
	 4.19 22.11 26.30 13.73	 4.12 21.76 25.88 13.53	 3.58 20.06 23.64 12.89	 3.99 21.19 25.18 13.21	 4.79 23.92 28.71 14.34	 <i>4.26 22.22</i> <i>26.48 13.71</i>	 4.10 21.60 25.70 13.39	 <i>4.26 22.22</i> <i>26.48 13.71</i>	 3.72 20.52 24.24 13.07	 3.57 19.89 23.46 12.75		
												Key: \bar{D}_u \bar{L}_u $\bar{D}C_u$

Table 1. Mean values for all gene tree shapes with $n \in \{3, 4, \dots, 9\}$ leaves. The shapes are shown ordered by their Furnas rank [8]. The table is patterned after [29,17]. The two red shapes for $n = 9$ have the same number of splits, which implies equal values of the corresponding mean values.