Algorithms for Unrooted Gene Trees with Polytomies*

Paweł Górecki[1] and Oliver Eulenstein[2]

[1] Institute of Informatics, University of Warsaw, Poland, gorecki@mimuw.edu.pl
[2] Dept. of Computer Science, Iowa State University, USA, oeulenst@cs.iastate.edu

Abstract. Gene tree reconciliation is a method to reconcile gene trees that are confounded by complex histories of gene duplications with a provided species tree. The trees involved are required to be rooted and full binary. Reconciling gene trees allows not only to identify and study such histories for gene families, but is also the base for several higher level applications including the estimation of species trees from gene trees when when duplications are involved. However, gene tree reconciliation can not handle *common gene trees* that are unrooted and non-binary trees. This limitation severely limits the applicability of gene tree reconciliation, since common trees are frequently inferred by phylogenetic methods in practice. Here, we describe a linear time reconciliation of a given common gene tree with a given species tree that can be non-binary. Then, we extend this reconciliation by seeking an optimal reconciliation under all rooted and binary refinements of the common gene tree with the given species tree. Finally, we describe a polynomial time algorithm that computes such an optimal refinement for the case when the species tree is full binary.

1 Introduction

Phylogenetic tree inference from the available range of genetic sequence information is a central task in studying evolution. The number of these sequences as well as their evolutionary complexity has expanded on an unprecedented scale in recent years [14] which is increasingly challenging current phylogenetic inference methods.

Gene trees that represent the evolutionary histories of gene families can be inferred from multiple sequence alignments of these families. Such trees provide valuable information to study the evolution of biochemical function in gene families. Often gene trees are assumed to reflect the evolution of species from which the sequences were sampled, which presents a common approach of species tree inference. However, phylogenetic tree inference from sequence alignments can be largely mislead when species have complex levels of gene duplications in their genomes. Avoiding such misleading cases can severely limit the exploitation of the species' available sequence ranges. Complicating matters further, gene duplication is one of the most fundamental and frequently occurring macroevolutionary processes [15], and therefore, different phylogenetic inference techniques are required.

* Support was provided to OE by the NSF (#0830012 and #106029), and to PG and OE by NCN #2011/01/B/ST6/02777 and the NIMBioS Working Group: Gene Tree Reconciliation through NSF #EF-0832858.

Gene tree reconciliation is an optimization problem that given a gene tree and a species tree, both rooted and full binary, seeks a reconciliation of the gene tree with the species tree that requires the minimum number of gene duplications, which is termed *reconciliation cost*. The gene tree reconciliation problem can be solved in linear time and space [17]. Reconciled gene trees are valuable for studying complex histories of gene duplication and loss through which function in gene families has evolved [2].

Since the advent of Goodman et al.'s [4] pioneering work that introduced gene tree reconciliation, ongoing research has provided many different variants of gene tree reconciliation [1,6,7,12], of which some have become the basis for several higher level applications. These applications include *gene tree parsimony* that is an NP-hard problem to infer a median (species) tree from a given collection of gene trees under the reconciliation cost [13]. For a detailed overview about gene tree reconciliation the interested reader is referred to [3].

While gene tree reconciliation provides credible results in practice [5,16], it is unable to handle *common gene trees* that are un-rooted and non-binary gene trees. Unfortunately, such common trees are frequently inferred in practice when insufficient sequence information does not permit to reliably estimate the root or bifurcations in gene trees.

Here, we overcome this limitation by introducing the duplication cost for common gene trees and describe a linear time algorithm to compute this cost based on concepts of unrooted reconciliation [8,9]. Then, we extend the gene tree reconciliation problem by seeking an optimal reconciliation across all rooted and binary refinements of the common gene tree with the given species tree. Finally, we describe a polynomial time algorithm that computes such an optimal refinement for the case when the species tree is binary, by showing a linear time reduction of our problem to the problem where the gene trees are rooted [12].

2 Basic Definitions and Preliminaries

An *unrooted tree* T is an acyclic, connected, and undirected graph that has no degree-two nodes, and every degree-one node is labeled with a species name. The degree-one nodes are called *leaves*; and the remaining nodes are called *internal* nodes. A tree is binary if every internal node has degree three. A *rooted* tree is defined similar to an unrooted tree, with the difference that it has a distinguished node, called *root*. *A contraction* of an edge e of an (un)rooted tree T removes e from T and merges both ends of e into a single node. *A binary refinement* of T is a binary tree that can be transformed into T by contractions. By $L(T)$ we denote the set of all leaf labels in T.

A rooted tree S with a unique leaf labeling is called *a species tree*. For two nodes a, b of S, $a \oplus b$ is the least common ancestor of a and b in S. Let T and be a rooted tree (called rooted gene tree) such that $L(T) \subseteq L(S)$. By $M\colon T \to S$ we denote the *least common ancestor (lca) mapping* between the nodes of T and S that preserves the labeling of the leaves. The *duplication cost* between T and S, is defined by: $D(T,S) := |\{M(g) = M(c)\colon c \text{ is a child of an internal node } g \in G\}|$.

Let $G = \langle V_G, E_G \rangle$ be an unrooted tree (called unrooted gene tree). A rooting of G can be defined by choosing an edge e on which the root is to be placed. Such a tree

will be denoted by G_e. The unrooted *duplication (urD)* cost between an unrooted gene tree G and a species tree S is defined as $urD(G,S) := \min\{D(G_e,S)\colon e \in E_G\}$. The edges $e \in E_G$, such that $D(G_e,S) = urDC(G,S)$ are called *optimal*. In the remainder of this work we show first how to compute $urDC$ in linear time and space, and then solve the following problem.

Problem 1. For a given unrooted gene tree G and a binary species tree S, find the binary refinement of a rooting of G that minimizes the duplication cost.

A similar problem for rooted gene trees was solved in [12]. In the remaining section we show how to reduce Prob. 1 into the rooted problem in linear time and space.

2.1 Unrooted reconciliation.

We start with definitions introducing the basics of unrooted reconciliation. This approach is based on our previous papers [8,9,10,11]. However, for the first time, we prove properties of urD for trees with multifurcations. We assume that G is an unrooted gene tree and S is a species tree. We transform G into a directed graph $\widehat{G}$, by replacing each edge $\{v,w\}$ by a pair of directed edges $\langle v,w\rangle$ and $\langle w,v\rangle$. We label the edges of $\widehat{G}$ by the nodes of S as follows. If $v \in G$ is a leaf labeled by a, then the edge $\langle v,w\rangle$ in $\widehat{G}$ is labeled by the node in S whose label is a. Let $v \in G$ have exactly k siblings $w_1, w_2, \ldots, w_k$. If a_i and b_i are the labels of $\langle v,w_i\rangle$ and $\langle w_i,v\rangle$, respectively, then $a_i = \bigoplus_{j=1,j\neq i}^{j=k} b_j$. Let $\top$ be the root of S. Each internal node $v \in G$ defines a *star* with the center v as indicated in Fig. 1a. We refer to the undirected edge $\{v,w_i\}$ as e_i, for all $i = 1,2,\ldots k$. There is a limited number of star types in gene trees [9]. Let K

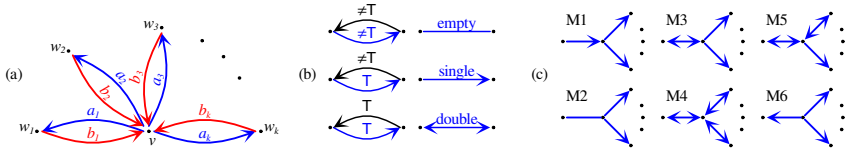

Fig. 1. (a) A star with the center v in $\widehat{G}$ and $k \geq 3$ edges. Here $e_i = \{v,w_i\}$ for $i = 1,2,\ldots,k$. (b) A simplified representation of edges (empty, single and double) that will be used through the rest of this work. The notation $\neq \top$ denotes that the label is a non-root node from S. (c) Star topologies that can be present in gene trees.

be a star with center v and k siblings as indicated in Fig. 1a. Let α denote the number of edges satisfying $a_i = \top$. Similarly, we define β for b_i's. Then, K has type: **M1** if $\alpha = 1$ and $\beta = k-1$ and all edges labeled by $\top$ are connected to the k siblings of v, **M2** if $\alpha = 0$ and $\beta = k-1$, **M3** if $\alpha = 1$ and $\beta = k$, **M4** if $\alpha = \beta = k$, **M5** if $1 < \alpha < \beta = k$ and **M6** if $\alpha = 0$ and $\beta = k$. The next proposition follows from the properties of stars [9].

Proposition 1. *G can have any number of stars M1. For the remaining stars we have three mutually exclusive cases: (i) G has a single edge (one or two stars M2) (ii) G has a double edge (stars M3-M5) or (iii) G has only single edges (exactly one star M6).*

3 Results

The next proposition show the cost changes when we move a position of the root in G.

Proposition 2. *Under the notation from Fig. 1 let $\lambda_i = D(G_{e_i}, S)$ for $i = 1, 2, \ldots, k$. Depending on the type of a star we have: (i) $\lambda_i = 1$ and $\lambda_j = 2$ for type M1 or M3 where $b_i = \top$ and $i \neq j \in \{1, 2, \ldots, k\}$, (ii) $\lambda_i \leq 1 \leq \lambda_j$ for type M2 where $a_i \neq \top \neq b_i$ and $i \neq j \in \{1, 2, \ldots, k\}$ or (iii) $\lambda_i = \lambda_j$ for M3-M6 and all i and j.*

We conclude from Prop. 1 and Prop. 2 that the optimal edges are single or double if they are present or all edges of star M6, otherwise. This observation leads to a linear time and space algorithm for $urDC$ computation similar to algorithms from [8,9].

Now we reduce Problem 1, to the problem where gene trees are rooted. We refer to the algorithm for refining rooted gene trees from [12] by $Bin(T, S)$, where T is a rooted tree and S is a binary species tree. It is known that $Bin(T, S)$ runs in $O(|T||S|)$ time [12]. For a node v by G_v we denote the tree rooted at v. Observe, that if v is a center of star M6, then $urD(G, S) = D(G_v, S) + 1$.

Theorem 1. *Alg. 1 infers a rooted binary gene tree G^* such that $D(G^*, S) = \min\{urDC(G', S) \colon G' \text{ is a binary refinement of } G\}$. Alg. 1 requires $O(|G||S|)$ time, while the reduction (steps 1-6) can be completed in $O(|G| + |S|)$ time and space.*

Algorithm 1 Resolving polytomies in unrooted gene trees

1: **Input** A binary species tree S, an unrooted gene tree G with at least three leaves $L(G) \subseteq L(S)$.
2: **Output** Optimal rooted binary refinement of G under duplication cost.
3: **Let** $m_{x,y}$ be the label (a node from S) of $\langle x, y\rangle$ in $\widehat{G}$. *// can be computed in $O(|G|)$ steps [9].*
4: **Let** v be a node from V_G and let $\top := M(L(G))$.
5: **While** there exists a node w adjacent with v such that $m_{w,v} = \top \neq m_{v,w}$ **do:** set $v := w$ (star M1).
6: **If** v is incident with a single/double edge $\langle v, w\rangle$, that is, $m_{v,w} = \top = m_{w,v}$ or $m_{v,w} \neq \top \neq m_{w,v}$
7: **then return** $Bin(G_{\langle v,w\rangle}, S)$ (optimal edge found in star M2-M5)
8: **else return** $Bin(G_v, S)$ (v is the center of star M6).

4 Conclusion

Gene tree reconciliation is elementary to phylogenetic inference as it reconciles otherwise misleading gene duplications, which is imperative for several higher level phylogenetic applications. In this work we modify gene tree reconciliation such that it can handle common gene trees, and in turn uses reconciliation to root the common gene trees and resolve their multifurcations. Our described algorithm computes this reconciliation and the rooted and refined gene tree. The presented work allows, for the first time, to explore the full range of available gene trees using gene tree reconciliation. Finally, our modified reconciliation allows us to simultaneously root and refine gene trees that are confounded by gene duplications.

Acknowledgements

We would like to thank Nadia El-Mabrouk for helpful discussions.

References

1. R. Chaudhary, J. B. Gordon, and O. Eulenstein. Efficient error correction algorithms for gene tree reconciliation based on duplication, duplication and loss, and deep coalescence. *BMC Bioinformatics*, 13(S-10):S11, 2012.
2. J. A. Cotton and R. D. M. Page. Going nuclear: gene family evolution and vertebrate phylogeny reconciled. *P. Roy. Soc. Lond. B Biol.*, 269:1555–1561, 2002.
3. O. Eulenstein, S. Huzurbazar, and D. A. Liberles. *Evolution after Gene Duplication*, chapter Reconciling Phylogenetic Trees. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2010.
4. M. Goodman, J. Czelusniak, G. W. Moore, A. E. Romero-Herrera, and G. Matsuda. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Zoology*, 28(2):132–163, 1979.
5. J. B. Gordon, M.S. Bansal, O. Eulenstein, and T.J. Vision. Inferring species trees from gene duplication episodes. In Aidong Zhang, Mark Borodovsky, Gultekin Özsoyoglu, and Armin R. Mikler, editors, *BCB*, pages 198–203. ACM, 2010.
6. P. Górecki and O. Eulenstein. A linear time algorithm for error-corrected reconciliation of unrooted gene trees. *LNCS*, 6674:148–159, 2011.
7. P. Górecki and O. Eulenstein. Algorithms: simultaneous error-correction and rooting for gene tree reconciliation and the gene duplication problem. *BMC Bioinformatics*, 13(Suppl 10):S14, 2012.
8. P. Górecki and O. Eulenstein. Deep coalescence reconciliation with unrooted gene trees: Linear time algorithms. *LNCS*, 7434:531–542, 2012.
9. P. Górecki and O. Eulenstein. A Robinson-Foulds measure to compare unrooted trees with rooted trees. *LNCS*, 7292:102–114, 2012.
10. P. Górecki, O. Eulenstein, and J. Tiuryn. Unrooted Tree Reconciliation: A Unified Approach. *TCBB*, accepted, preprint available, to appear in 2013.
11. P. Górecki and J. Tiuryn. Inferring phylogeny from whole genomes. *Bioinformatics*, 23(2):e116–e122, 2007.
12. M. Lafond, K.M. Swenson, and N. El-Mabrouk. An optimal reconciliation algorithm for gene trees with polytomies. *WABI 2012, LNCS/LNBI*, 7534:106–122, 2012.
13. B. Ma, M. Li, and L. Zhang. From gene trees to species trees. *SIAM J. Comput.*, 30(3):729–752, 2000.
14. J. E. McCormack, S. M. Hird, A. J. Zellmer, B. C. Carstens, and R. T. Brumfield. Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol Phylogenet Evol*, 66(2):526–38, Feb 2013.
15. S. Ohno. *Evolution by gene duplication*. Springer-Verlag, Berlin, 1970.
16. M. J. Sanderson and M. M. McMahon. Inferring angiosperm phylogeny from EST data with widespread gene duplication. *BMC Evolutionary Biology*, 7 (Suppl 1): S3, 2007.
17. L. Zhang. On a Mirkin-Muchnik-Smith conjecture for comparing molecular phylogenies. *Journal of Computational Biology*, 4(2):177–187, 1997.