On the Impact of Uncertain Gene Tree Rooting on Duplication-Transfer-Loss Reconciliation

Soumya Kundu[1] and Mukul S. Bansal[1,2]

[1] Department of Computer Science and Engineering,
University of Connecticut, Storrs, USA
soumya.kundu@uconn.edu
[2] Institute for Systems Genomics, University of Connecticut, Storrs, USA
mukul.bansal@uconn.edu

Duplication-Transfer-Loss (DTL) reconciliation is one of the most effective techniques for studying the evolution of gene families and inferring evolutionary events. Given the evolutionary tree for a gene family, i.e., a *gene tree*, and the evolutionary tree for the corresponding species, i.e., a *species tree*, DTL reconciliation compares the gene tree with the species tree and reconciles any differences between the two by proposing gene duplication, horizontal gene transfer, and gene loss events. DTL reconciliations are generally computed using a parsimony framework where each evolutionary event is assigned a cost and the goal is to find a reconciliation with minimum total cost [1–3]. The resulting optimization problem is called the *DTL-reconciliation problem*.

The standard formulation of the DTL-reconciliation problem requires the gene tree and the species tree to be rooted. However, while species trees can generally be confidently rooted (using outgroups, for example), gene trees are often difficult to root. As a result, the gene trees used for DTL reconciliation are often unrooted. When provided with an unrooted gene tree, existing DTL-reconciliation algorithms and software first find a root for the unrooted gene tree that yields the minimum reconciliation cost and then use the resulting rooted gene tree for the reconciliation. However, there is a critical flaw in this approach: Many gene trees have multiple optimal roots, and yet, only a single optimal root is randomly chosen to create the rooted gene tree and perform the reconciliation. Here, we perform the first in-depth analysis of the impact of uncertain gene tree rooting on DTL reconciliation and provide the first computational tools to quantify and negate the impact of gene tree rooting uncertainty.

To properly account for rooting uncertainty, we define a *consensus reconciliation*, which summarizes the different reconciliations across all optimal rootings of an unrooted gene tree and makes it possible to identify those aspects of the reconciliation that are conserved across all optimal rootings. We study basic structural properties of consensus reconciliations and analyze a large biological data set of over 4500 gene families from a broadly sampled set of 100 predominantly prokaryotic species [4]. Our analysis focuses on several fundamental aspects of DTL reconciliation with unrooted gene trees including prevalence of multiple optimal rootings, structure of optimal roots in multiply rooted gene trees, impact of gene tree error and evolutionary event costs, information content of consensus reconciliations, and conservation of event and mapping assignments in consensus reconciliations.

Our experimental results show that a large fraction of gene trees have multiple optimal rootings and that gene tree error significantly increases the fraction of multiply rooted gene trees. The prevalence of multiple optimal rootings is also heavily influenced by gene tree size, with smaller gene trees more likely to have multiple optimal roots. An analysis of the placement of optimal roots shows that multiple roots often, but not always, appear clustered together in the same region of the gene tree. This a highly desirable property since it maximizes the information content, or size, of consensus reconciliations and also makes it easier to estimate the "true" root position. A detailed study of the computed consensus reconciliations reveals that most aspects of the reconciliation, i.e., event and mapping assignments, remain conserved across the multiple rootings, showing that unrooted gene trees can be meaningfully reconciled even after accounting for multiple optimal roots. Our analysis also uncovers several interesting patterns in the reconciliations of singly rooted and multiply rooted gene trees.

The results of our experimental analysis have important implications for the application of DTL reconciliation in evolutionary studies, and the techniques introduced in this work make it possible to systematically avoid incorrect evolutionary inferences caused by incorrect or uncertain gene tree rooting. Our tools for computing consensus reconciliations have been implemented into the phylogenetic reconciliation software package RANGER-DTL, freely available from http://compbio.engr.uconn.edu/software/RANGER-DTL/.

References

1. Tofigh, A., Hallett, M.T., Lagergren, J.: Simultaneous identification of duplications and lateral gene transfers. IEEE/ACM Trans. Comput. Biol. Bioinform. **8**(2), 517–535 (2011)
2. Doyon, J.P., Scornavacca, C., Gorbunov, K.Y., Szöllosi, G.J., Ranwez, V., Berry, V.: An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In: Tannier, E. (ed.) RECOMB-CG. LNCS, vol. 6398, pp. 93–108. Springer, Berlin (2010)
3. Bansal, M.S., Alm, E.J., Kellis, M.: Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. Bioinformatics **28**(12), 283–291(2012)
4. David, L.A., Alm, E.J.: Rapid evolutionary innovation during an archaean genetic expansion. Nature **469**, 93–96 (2011)