OXFORD

Phylogenetics

RANGER-DTL 2.0: rigorous reconstruction of gene-family evolution by duplication, transfer and loss

Mukul S. Bansal[1,]*, Manolis Kellis[2,3], Misagh Kordi[1] and Soumya Kundu[1]

[1]Department of Computer Science and Engineering, University of Connecticut, Storrs, CT 06269, USA, [2]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA and [3]Broad Institute, Cambridge, MA 02142, USA

*To whom correspondence should be addressed.
Associate Editor: Janet Kelso

Abstract

Summary: RANGER-DTL 2.0 is a software program for inferring gene family evolution using Duplication-Transfer-Loss reconciliation. This new software is highly scalable and easy to use, and offers many new features not currently available in any other reconciliation program. RANGER-DTL 2.0 has a particular focus on reconciliation accuracy and can account for many sources of reconciliation uncertainty including uncertain gene tree rooting, gene tree topological uncertainty, multiple optimal reconciliations and alternative event cost assignments. RANGER-DTL 2.0 is open-source and written in C++ and Python.
Availability and implementation: Pre-compiled executables, source code (open-source under GNU GPL) and a detailed manual are freely available from http://compbio.engr.uconn.edu/software/RANGER-DTL/.
Contact: mukul.bansal@uconn.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Duplication-Transfer-Loss (DTL) reconciliation is widely recognized as one of the most powerful computational techniques for understanding the evolution of microbial gene families (Kamneva and Ward, 2014). DTL reconciliation works by comparing a given gene tree (for the gene family of interest) against the corresponding species tree and postulating gene duplication, horizontal gene transfer and gene loss events to explain the evolution of that gene tree inside the species tree. The result of DTL reconciliation is a mapping of the nodes of the gene tree to nodes (or edges) of the species tree, showing the embedding of the gene tree inside the species tree, as well as a labeling of each internal node of the gene tree as either a speciation, duplication, or transfer event. Such detailed knowledge of gene family evolution has many important biological applications, and the DTL reconciliation problem has therefore been extensively studied, e.g. (Bansal *et al.*, 2012, 2013; David and Alm, 2011;

Doyon *et al.*, 2010; Jacox *et al.*, 2016; Kordi and Bansal, 2016; Sjostrand *et al.*, 2014; Stolzer *et al.*, 2012; Szollosi *et al.*, 2012; Tofigh *et al.*, 2011).

While probabilistic models of DTL evolution also exist (Sjostrand *et al.*, 2014; Szollosi *et al.*, 2012), we focus here on parsimony-based models of DTL reconciliation which are much more scalable and require fewer parameters. Parsimony-based DTL reconciliation is also known to be highly accurate in practice; see Section S3 in the Supplementary Material for a detailed discussion on accuracy.

A preliminary version of RANGER-DTL (short for Rapid ANalysis of Gene family Evolution using Reconciliation-DTL) was released in 2012 with a paper on the algorithmics of DTL reconciliation (Bansal *et al.*, 2012), providing only rudimentary functionality. Despite its limited functionality, the preliminary version of RANGER-DTL has been frequently used for biological data analysis

(Dupont and Cox, 2017; Heitlinger *et al.*, 2014; Heshiki *et al.*, 2017; Jeong *et al.*, 2016; Koczyk *et al.*, 2015; Ricci *et al.*, 2015). Here, we release the first full version of RANGER-DTL with greatly extended and improved functionality, and featuring the new algorithms and techniques developed in Bansal *et al.* (2013); Kordi and Bansal (2016); Kundu and Bansal (2018).

2 Features

RANGER-DTL 2.0 is designed to enable fast and rigorous analysis of gene families and provides several advanced features not available in any other reconciliation software. The software takes as input a gene tree (rooted or unrooted) and a rooted species tree and reconciles the two by postulating speciation, duplication, transfer and loss events. Advanced capabilities of RANGER-DTL 2.0 include (i) principled handling of unrooted gene trees by considering all possible optimal rootings, (ii) uniformly random sampling of the space of all optimal reconciliations, making it possible to compute multiple optimal reconciliations and account for the variability in optimal reconciliation scenarios, (iii) use of distance-dependent transfer costs to better model transfer dynamics, (iv) handling gene tree uncertainty by collapsing weakly supported gene tree edges and computing and considering all optimal resolutions of the gene tree and (v) computing support values for individual DTL event inferences and species mapping assignments while accounting for multiple optimal reconciliations, uncertainty in gene tree rooting, alternative event cost assignments and even gene tree topological uncertainty. Furthermore, RANGER-DTL 2.0 can efficiently analyze trees with thousands of taxa.

While it can handle both undated and fully-dated species trees, the focus of RANGER-DTL 2.0 is on undated species trees, for which it offers the most options and functionality. The reason for focusing on undated species trees is explained in Section S1 in the Supplementary Material.

Several features of RANGER-DTL 2.0, including consideration of all optimal gene tree roots, all possible optimal resolutions of unresolved gene trees and distance-dependent transfer costs, are not available in any comparable software package. A detailed comparison of RANGER-DTL 2.0 with existing DTL reconciliation software appears in Section S2 of the Supplementary Material.

3 Availability and requirements

The software package consists of 10 related programs designed to work together to support various reconciliation analyses. These ten programs are organized into (i) three *core programs*, which define the core functionality of RANGER-DTL 2.0, designed to be applied sequentially, (ii) five *Supplementary programs* that provide additional functionality and (iii) two *summary scripts*. Further details on the implementation of RANGER-DTL 2.0 are given in Section S4 of the Supplementary Material.

RANGER-DTL 2.0 is available open-source under GNU General Public Licence v3. Pre-compiled executables for Linux, Mac, and Windows, source code and a detailed manual are freely available online. The eight core and Supplementary programs are written in C++ and can be compiled on any operating system with a C++ compiler supporting the ANSI C++ standard. These C++ programs use standard C++ libraries along with the freely available and widely used Boost C++ libraries (http://www.boost.org/). The two summary scripts are written in Python and can be run on any operating system with the Python interpreter. RANGER-DTL is designed to be efficient in both time complexity and memory requirements, and all programs, except for the two that consider unresolved gene trees, are scalable to hundreds or thousands of genes and taxa on commodity hardware. For instance, computing an optimal reconciliation using the core *Ranger-DTL* program for species trees and gene trees with 200 leaves and 1000 leaves each requires approximately 5 s and 9 min, respectively, on a desktop computer with a 3.1 GHz Intel i5 processor and both instances require less than 1 GB of RAM. In fact, with the supplementary program *Ranger-DTL-Fast*, reconciling the 1000-leaf trees takes less than a second.

4 Conclusion

Accurate and efficient DTL reconciliation of gene trees and species trees is crucial to understanding microbial gene and species evolution and to inferring horizontal gene transfer and other evolutionary events. RANGER-DTL 2.0 makes it possible to perform fast and rigorous analysis of gene family evolution through DTL reconciliation and offers many important features, such as consideration of all optimal gene tree roots, all possible optimal resolutions of unresolved gene trees, and distance-dependent transfer costs, that are not available in any comparable reconciliation software. RANGER-DTL is also designed to be easy to use, with easily interpretable results.

There are several additional features that we intend to add to RANGER-DTL to further improve its functionality and accuracy. These include fast heuristics for handling gene tree uncertainty and estimating its impact on the reconciliation, and consideration of transfers from unsampled or extinct lineages, e.g. (Jacox *et al.*, 2016). These and other new features will be extensively tested to assess their impact on DTL reconciliation accuracy, and those that result in an improvement will be added to RANGER-DTL.

References

Bansal,M.S. *et al.* (2012) Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics*, **28**, i283–i291.

Bansal,M.S. *et al.* (2013) Reconciliation revisited: handling multiple optima when reconciling with duplication, transfer, and loss. *J. Comput. Biol.*, **20**, 738–754.

David,L.A. and Alm,E.J. (2011) Rapid evolutionary innovation during an archaean genetic expansion. *Nature*, **469**, 93–96.

Doyon,J.-P. *et al.* (2010) An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In: Tannier, E. (ed.) *RECOMB-CG, Lecture Notes in Computer Science*, Vol. **6398**. Springer, Berlin, Heidelberg, pp. 93–108.

Dupont,P.-Y. and Cox,M.P. (2017) Genomic data quality impacts automated detection of lateral gene transfer in fungi. *G3: Genes Genomes Genet.*, **7**, 1301–1314.

Heitlinger,E. *et al.* (2014) The genome of Eimeria falciformis–reduction and specialization in a single host apicomplexan parasite. *BMC Genomics*, **15**, 696.

Heshiki,Y. *et al.* (2017) Toward a metagenomic understanding on the bacterial composition and resistome in hong kong banknotes. *Front. Microbiol.*, **8**, 632.

Jacox,E. *et al.* (2016) Eccetera: comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics*, **32**, 2056.

Jeong,H. *et al.* (2016) HGTree: database of horizontally transferred genes determined by tree reconciliation. *Nucleic Acids Res.*, **44**, D610.

Kamneva,O.K. and Ward,N.L. (2014) Reconciliation approaches to determining HGT, duplications, and losses in gene trees. In: Goodfellow, M., Sutcliffe, I. and Chun, J. (eds.) *New Approaches to Prokaryotic Systematics, Vol. 41 of Methods in Microbiology.* Academic Press, Amsterdam, Netherlands, pp. 183–199.

Koczyk,G. *et al.* (2015) The distant siblings–a phylogenomic roadmap illuminates the origins of extant diversity in fungal aromatic polyketide biosynthesis. *Genome Biol. Evol.*, **7**, 3132.

Kordi,M. and Bansal,M.S. (2016) Exact algorithms for duplication-transfer-loss reconciliation with non-binary gene trees. In: *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB 2016.* ACM, New York, USA, pp. 297–306.

Kundu,S. and Bansal,M.S. (2018) On the impact of uncertain gene tree rooting on duplication-transfer-loss reconciliation. *BMC Bioinformatics*, in press.

Ricci,J.N. *et al.* (2015) Phylogenetic analysis of HpnP reveals the origin of 2-methylhopanoid production in Alphaproteobacteria. *Geobiology*, **13**, 267–277.

Sjostrand,J. *et al.* (2014) A Bayesian method for analyzing lateral gene transfer. *Syst. Biol.*, **63**, 409–420.

Stolzer,M. *et al.* (2012) Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics*, **28**, i409–i415.

Szollosi,G.J. *et al.* (2012) Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc. Natl. Acad. Sci. USA*, **109**, 17513–17518.

Tofigh,A. *et al.* (2011) Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **8**, 517–535.