BIOINFORMATICS

Vol. 00 no. 00 2017
Pages 1–2

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], Soumya Kundu[1]

[1] Department of Computer Science & Engineering, University of Connecticut, Storrs, USA.
[2] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, USA. [3] Broad Institute of MIT and Harvard, Cambridge, USA.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

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: 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

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., (Doyon *et al.*, 2010; Tofigh *et al.*, 2011; David and Alm, 2011; Bansal *et al.*, 2012; Stolzer *et al.*, 2012; Szollosi *et al.*, 2012; Bansal *et al.*, 2013; Sjostrand *et al.*, 2014; Kordi and Bansal, 2016; Jacox *et al.*, 2016).

While probabilistic models of DTL evolution also exist (Szollosi *et al.*, 2012; Sjostrand *et al.*, 2014), 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 supplement 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 (Heitlinger *et al.*, 2014; Ricci *et al.*, 2015; Koczyk *et al.*, 2015; Jeong *et al.*, 2016; Dupont and Cox, 2017; Heshiki *et al.*, 2017). 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 supplement.

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 supplement.

3 AVAILABILITY AND REQUIREMENTS

The software package consists of ten 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 supplement.

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 $1,000$ leaves each requires approximately 5 seconds and 9 minutes, 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 $1,000$-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.

ACKNOWLEDGEMENTS

Funding: This work was supported in part by NSF CAREER award IIS 1553421 and by NSF awards MCB 1616514 and IES 1615573 to MSB, and by a University of Connecticut Summer Undergraduate Research Fund (SURF) award to SK.

Conflict of Interest: None declared.

REFERENCES

Bansal, M. S., Alm, E. J., and Kellis, M. (2012). Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics*, **28**(12), 283–291.

Bansal, M. S., Alm, E. J., and Kellis, M. (2013). Reconciliation revisited: Handling multiple optima when reconciling with duplication, transfer, and loss. *J. Comput. Biol.*, **20**(10), 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., Scornavacca, C., Gorbunov, K. Y., Szöllosi, G. J., Ranwez, V., and Berry, V. (2010). An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In E. Tannier, editor, *RECOMB-CG*, volume 6398 of *Lecture Notes in Computer Science*, pages 93–108. Springer.

Dupont, P.-Y. and Cox, M. P. (2017). Genomic Data Quality Impacts Automated Detection of Lateral Gene Transfer in Fungi. *G3: Genes, Genomes, Genetics*, **7**(4), 1301–1314.

Heitlinger, E., Spork, S., Lucius, R., and Dieterich, C. (2014). The genome of Eimeria falciformis - reduction and specialization in a single host apicomplexan parasite. *BMC Genomics*, **15**(1), 696.

Heshiki, Y., Dissanayake, T., Zheng, T., Kang, K., Yueqiong, N., Xu, Z., Sarkar, C., Woo, P. C. Y., Chow, B. K. C., Baker, D., Yan, A., Webster, C. J., Panagiotou, G., and Li, J. (2017). Toward a metagenomic understanding on the bacterial composition and resistome in hong kong banknotes. *Frontiers in Microbiology*, **8**, 632.

Jacox, E., Chauve, C., Szollosi, G. J., Ponty, Y., and Scornavacca, C. (2016). eccetera: comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics*, **32**(13), 2056.

Jeong, H., Sung, S., Kwon, T., Seo, M., Caetano-Anolls, K., Choi, S. H., Cho, S., Nasir, A., and Kim, H. (2016). HGTree: database of horizontally transferred genes determined by tree reconciliation. *Nucleic Acids Research*, **44**(D1), D610.

Kamneva, O. K. and Ward, N. L. (2014). Reconciliation approaches to determining hgt, duplications, and losses in gene trees. In M. Goodfellow, I. Sutcliffe, and J. Chun, editors, *New Approaches to Prokaryotic Systematics*, volume 41 of *Methods in Microbiology*, pages 183 – 199. Academic Press.

Koczyk, G., Dawidziuk, A., and Popiel, D. (2015). The Distant Siblings – A Phylogenomic Roadmap Illuminates the Origins of Extant Diversity in Fungal Aromatic Polyketide Biosynthesis. *Genome Biology and Evolution*, **7**(11), 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*, pages 297–306. ACM.

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

Ricci, J. N., Michel, A. J., and Newman, D. K. (2015). Phylogenetic analysis of HpnP reveals the origin of 2-methylhopanoid production in Alphaproteobacteria. *Geobiology*, **13**(3), 267–277.

Sjostrand, J., Tofigh, A., Daubin, V., Arvestad, L., Sennblad, B., and Lagergren, J. (2014). A bayesian method for analyzing lateral gene transfer. *Systematic Biology*, **63**(3), 409–420.

Stolzer, M., Lai, H., Xu, M., Sathaye, D., Vernot, B., and Durand, D. (2012). Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics*, **28**(18), 409–415.

Szollosi, G. J., Boussau, B., Abby, S. S., Tannier, E., and Daubin, V. (2012). Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc Natl Acad Sci USA*, **109**(43), 17513–17518.

Tofigh, A., Hallett, M. T., and Lagergren, J. (2011). Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Trans. Comput. Biology Bioinform.*, **8**(2), 517–535.