Bioinformatics, 2020, 1–2
doi: 10.1093/bioinformatics/btaa978
Advance Access Publication Date: 20 November 2020
Applications Note

Phylogenetics

eMPRess: a systematic cophylogeny reconciliation tool

Santi Santichaivekin[1], Qing Yang[1], Jingyi Liu[1], Ross Mawhorter[2], Justin Jiang[1], Trenton Wesley[1], Yi-Chieh Wu[1] and Ran Libeskind-Hadas [1,*]

[1]Department of Computer Science, Harvey Mudd College, Claremont, CA 91711, USA and and [2]Department of Computer Science and Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA

*To whom correspondence should be addressed.

Associate Editor: Schwartz Russell
Received on August 20, 2020; revised on November 5, 2020; accepted on November 10, 2020; editorial decision on November 8, 2020;

Abstract

Summary: We describe eMPRess, a software program for phylogenetic tree reconciliation under the duplication-transfer-loss model that systematically addresses the problems of choosing event costs and selecting representative solutions, enabling users to make more robust inferences.
Availability and implementation: eMPRess is freely available at http://www.cs.hmc.edu/empress.
Contact: hadas@cs.hmc.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The problem of reconciling pairs of phylogenetic trees arises when studying pairs of entities that have evolved in tandem, including hosts and parasites, pairs of symbionts and species and their ecological niches. Given two phylogenetic trees and an association of their tips (extant taxa), a reconciliation is a mapping of one tree (e.g. the parasite tree) into the other (e.g. the host tree) that explains their evolutionary histories subject to cospeciation, duplication, host transfer and loss events, known as the *DTL* model. Reconciliation in the DTL model is generally performed using a maximum parsimony formulation in which each event type has an associated non-negative cost (except for cospeciation which is assumed to have cost zero) and the objective is to find a mapping, known as a *maximum parsimony reconciliation (MPR)*, that minimizes the weighted cost of the constituent events. (See Supplementary Section S1.)

Several software packages have been developed for inferring MPRs under the DTL model. While all of these tools have important and unique features, none of them address the fundamental challenges of systematically selecting multiple representative event costs from the infinite number of possible combinations. With one exception [Capybara (Wang *et al.*, 2020)] which uses a fundamentally different approach than the one proposed here, none of these tools address the problem of identifying representative MPRs from the potentially exponentially large space of equally optimal solutions. (See Supplementary Section S6.) The choice of event costs and the resulting MPRs can substantially impact the conclusions reached from the data. Making inferences from a single set of event costs or a single MPR may lead to conclusions that are not supported, or are even contradicted, by other MPRs. We present the eMPRess package to address these challenges.

2 Empress software

eMPRess replaces our Jane tool (Conow *et al.*, 2010), which has over 1600 registered downloads and has been used in more than 200 cophylogenetic studies. The design of eMPRess was informed by a survey of over 100 Jane users and the advent of new methods and algorithms, including several from our research group (Libeskind-Hadas *et al.*, 2014; Mawhorter and Libeskind-Hadas, 2019; Santichaivekin *et al.*, 2019). eMPRess integrates these algorithms into a single application with an intuitive workflow for use by biologists. (See Supplementary Section S5.) The initial release of eMPRess lacks some features found in Jane including handling of multi-host parasites, non-binary trees and time zone ranges that allow the user to specify partial dating information.

To demonstrate the features and utility of eMPress (Fig. 1a), we use a host-parasite dataset of *Vidua* parasitic brood finches and their estrildid finch hosts (Sorenson *et al.*, 2004) as our running example. (See also the video on the eMPRess website for a demonstration of the eMPRess workflow using this dataset.)

Visualizing the impact of event costs—In a parsimony framework, there exist an infinite number of combinations of event costs. However, most published studies using DTL reconciliation choose either a single combination of event costs or a small sample of event costs. To understand the impact of the event costs on the solution space of MPRs, eMPRess partitions the space of event costs into a finite number of equivalence classes or 'regions' using the Costscape Algorithm (Libeskind-Hadas *et al.*, 2014). Since event costs are unitless, the loss cost is fixed to 1.0, and regions depend on duplication (x-axis) and host transfer (y-axis) costs relative to the cost of loss. Each region comprises event costs that give rise to the same set of MPRs. (Fig. 1b; see also Supplementary Section S2.)

Clustering MPR space—In general, even for a single set of event costs, there can exist an exponentially large and diverse set of MPRs

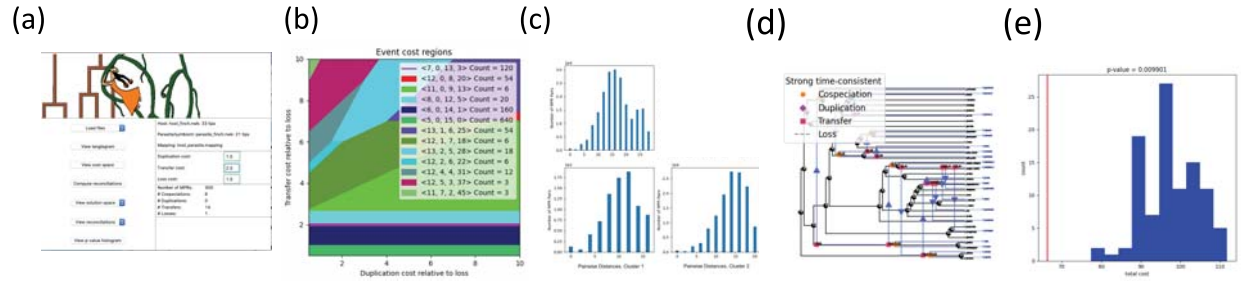

Fig. 1. Overview of eMPRess functionality using the *Vidua* dataset (Sorenson *et al.*, 2004). (**a**) eMPRess main GUI interface; (**b**) event cost landscape (Costscape) plot; (**c**) clustering of MPRs, first row shows the pairwise distances for the entire space of 640 MPRs, second row shows the pairwise distances in each of two clusters; (**d**) a median reconciliation from one of the two clusters; (**e**) *P*-value histogram showing cost of the MPR for the original data (red) and costs of MPRs for the two trees with randomized tip associations (blue)

(Haack *et al.*, 2019). To help biologists address this challenge, eMPRess computes and displays the distances between every pair of MPRs using the Pairwise Distance Algorithm (Santichaivekin *et al.*, 2019) (Fig. 1c top), where the distance between two MPRs is the number of events that are found in one MPR or the other but not both (Haack *et al.*, 2019; Huber *et al.*, 2018; Nguyen *et al.*, 2013).

The distribution of distances provides important insight into the structure of the solution space. For example, the presence of multiple modes in the distribution could indicate the need for multiple representative reconciliations. The user may then choose to cluster the solution space using our hierarchical clustering algorithm (Mawhorter and Libeskind-Hadas, 2019). In our *Vidua* example, the distribution of distances reveals two modes (Fig. 1c, top), which were disaggregated into two clusters (Fig. 1c, bottom). (See also Supplementary Section S3.)

Viewing MPRs—In addition to summarizing MPR space, eMPRess provides a tool for selecting and viewing individual MPRs (Fig. 1d) The user may choose to select one representative median MPR for the entire space (Nguyen *et al.*, 2013) or a median MPR from each cluster. eMPRess also reports whether an MPR is time-consistent, distinguishing between two types of time-consistency. (See Supplementary Section S4.)

Additional features—eMPRess has a number of additional features that are used in cophylogenetic analyses, including computing event support values, statistical tests for tree congruence (Fig. 1e) and visualizations that support zooming and saving images to files.

3 Software

eMPRess is written in Python 3. The eMPRess website is at www.cs.hmc.edu/empress, which includes tutorials, software, documentation and sample input files.

4 Conclusion

The eMPRess software offers unique features that address the problem of systematically selecting event costs and best representative MPRs from the large solution space. This tool allows biologists to make better-informed and more robust conclusions than is possible with existing tools.

Acknowledgements

The authors thank P. Andrews, A.E. Garcia, A. Garcia, D. Makhervaks, C. Ngo, S. Sehra and Z. Witzel and the anonymous reviewers for their valuable input.

Funding

This work was supported by the National Science Foundation [IIS-1751399 to Y.-C.W. and IIS-1905885 to R.L.-H.].

Conflict of Interest: none declared.

References

Conow,C. *et al.* (2010) Jane: a new tool for cophylogeny reconstruction problem. *Algorithms Mol. Biol.*, **5**, 16.

Haack,J. *et al.* (2019) Computing the diameter of the space of maximum parsimony reconciliations in the duplication-transfer-loss model. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **16**, 14–22.

Huber,K.T. *et al.* (2018) Geometric medians in reconciliation spaces of phylogenetic trees. *Inf. Process. Lett.*, **136**, 96–101.

Libeskind-Hadas,R. *et al.* (2014) Pareto-optimal phylogenetic tree reconciliation. *Bioinformatics*, **30**, i87–i95.

Mawhorter,R. and Libeskind-Hadas,R. (2019) Hierarchical clustering of maximum parsimony reconciliations. *BMC Bioinformatics*, **20**, 612.

Nguyen,T.-H. *et al.* (2013) Support measures to estimate the reliability of evolutionary events predicted by reconciliation methods. *PLoS One*, **8**, e73667.

Santichaivekin,S. *et al.* (2019) An efficient exact algorithm for computing all pairwise distances between reconciliations in the duplication-transfer-loss model. *BMC Bioinformatics*, **20**, 636.

Sorenson,M.D. *et al.* (2004) Clade-limited colonization in brood parasitic finches (*Vidua* spp.). *Syst. Biol.*, **53**, 140–153.

Wang,Y. *et al.* (2020) Capybara: equivalence ClAss enumeration of coPhylogenY event-BAsed ReconciliAtions. *Bioinformatics*, **36**, 4197–4199.