Intertwining phylogenetic trees and networks

Klaus Schliep, University of Massachusetts Boston, Boston MA, USA
Alastair J. Potts, Nelson Mandela Metropolitan University, Port Elizabeth, South Africa
David A. Morrison, Uppsala University, Uppsala, Sweden
Guido W. Grimm, University of Vienna, Vienna, Austria

Corresponding author: Alastair J. Potts, potts.a@gmail.com

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2054v1 | CC-BY 4.0 Open Access | rec: 18 May 2016, publ: 18 May 2016

Abstract

The fields of phylogenetic tree and network inference have dramatically advanced in the last decade, but independently with few attempts to bridge them. Here we provide a framework, implemented in the PHANGORN library in R, to transfer information between trees and networks. This includes: 1) identifying and labelling equivalent tree branches and network edges, 2) transferring branch support to network edges, and 3) mapping bipartition support from a sample of trees (e.g. from bootstrapping or Bayesian inference) onto network edges. The ability to readily combine tree and network information should lead to more comprehensive evolutionary comparisons and conclusions.

Keywords

Exploratory Data Analysis; Networks; PHANGORN; R; Trees;

Text

Traditional phylogenetic inference has almost exclusively relied on the assumption that evolution is successfully captured by a bifurcating tree (Mindell 2013). However, tree-based methods usually perform poorly when this assumption is violated, and phylogenetic networks should be used instead (Bapteste et al. 2013). Despite advances in both fields (e.g. Balvočiūtė et al. 2014; Salichos et al. 2014; Yang et al. 2013), the interface between trees and networks has rarely been bridged (Holland & Moulton 2003; Holland et al. 2008; Huber et al. 2016). The decision to use trees or networks is usually not dependent on any arguments over the superiority of one approach over the other (but see Morrison 2014), but rather the evolutionary complexity of the group under investigation and the resulting dataset. Nonetheless, tree-based methods remain the prime analytical choice. When the levels of conflict are great, however, researchers may resort to networks – often as a last option after all other tree-based methods have failed – to have some way of making sense of the patterns within a dataset. The only alternative is filtering the 'rogue' taxa that are causing topological conflict or a decrease in branch support (Aberer et al. 2013). The wide range of available network methods (Huson & Bryant 2006) have remained underutilised, likely because of the difficulties that arise when comparing trees and networks (such as matching tree branches to network edges).

The advances in tree and network inferences call for an integration of both methodologies.

However, a framework enabling automated integration has been lacking. Here we provide an R-based framework, implemented in the PHANGORN library (Schliep 2011), to intertwine trees and networks. Using this framework we can:

1) Compare trees and networks by identifying shared or exclusive branches or edges between trees and networks constructed for the same dataset (Fig. 1A). We hope this will help researchers bridge the psychological gap between tree- and network-thinking (Morrison 2010; Morrison 2014).
2) Map branch support (e.g. nonparametric bootstrap support: Felsenstein 1985; Bayesian posterior probabilities: Rannala & Yang 1996), incongruence values (internode certainty: Salichos et al. 2014), or any other value that can be linked to a tree branch, onto a phylogenetic network (Fig. 1B). This will help researchers e.g. to investigate non-ambiguous support (any value < 1.0/100) of tree branches, and to determine whether this is due to incompatible or insufficient signals in the underlying data (e.g. Draper et al. 2007).
3) Map bipartition frequencies from a sample of trees (e.g. from non-parametric bootstrapping or Bayesian inference) onto network edges (Fig. 1C; Grimm et al. 2006). This will help provide much-needed confidence in networks, and facilitate investigation of topological alternatives that are not captured by the tree-inference itself.

This open-source R-based tree-network framework (scripts and vignettes can be found in the Supplement Material) provides a meeting point for the output of tree and network inference software (e.g. SplitsTree: Huson & Bryant 2006; MrBayes: Ronquist et al. 2012; RAxML: Stamatakis 2014) and results can either be visualised within R or exported to other visualisation software (e.g. SplitsTree; FigTree: Rambaut 2014).

We envisage that this framework will have a multitude of uses, such as investigating specific phylogenetic signals, identifying competing evolutionary scenarios and pinpointing methodological shortcomings. For example, tree branches that are not present in the edges of a network can be identified, or vice versa; this often highlights significant and identifiable discrepancies between trees, which may arise from specific processes (e.g. rapid ancient radiations) or method-inherent biases (branching artefacts, model-induced differences). Networks can also be used to improve phylogenetic tree inference (Morrison 2010). For example, ‘lost branches’ can be identified, i.e. alternative phylogenetic splits that receive relatively high support but are not represented in the inferred tree. In addition, researchers may be interested in the differential support of topological alternatives that are masked by polytomies

in the commonly used majority rule trees, strict consensus trees, or single 'representative' trees with mapped support values (e.g. Mardulyn 2012).

This framework will help phylogenetic practitioners to readily transfer information between tree-based and network-based analyses, and thereby visualise and investigate similarities and differences between them. We believe that phylogenetic networks with edges supported by tree-based algorithms (e.g. maximum likelihood or Bayesian inference) offer the most comprehensive representation of evolutionary signal in a phylogenetic dataset irrespective of its complexity (Fig. 1C; e.g. Potts et al. 2014).

Acknowledgements

This work was supported in part by a grant from the National Science Foundation (DEB 1350474 to KS); and the Austrian Science Fund FWF (M1751-B16 to GWG).

References

Aberer AJ, Krompass D, and Stamatakis A. 2013. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. *Systematic Biology* 62:162–166.

Balvočiūtė M, Spillner A, and Moulton V. 2014. FlatNJ: A novel network-based approach to visualize evolutionary and biogeographical relationships. *Systematic Biology* 63:383–396. 10.1093/sysbio/syu001

Bapteste E, van Iersel L, Janke A, Kelchner S, Kelk S, McInerney JO, Morrison DA, Nakhleh L, Steel M, Stougie L, and Whitfield J. 2013. Networks: expanding evolutionary thinking. *Trends in Genetics* 29:439–441.

Draper I, Hedenäs L, and Grimm GW. 2007. Molecular and morphological incongruence in European species of *Isothecium* (Bryophyta). *Molecular Phylogenetics and Evolution* 42:700-716. 10.1016/j.ympev.2006.09.021

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783-791.

Grimm GW, Renner SS, Stamatakis A, and Hemleben V. 2006. A nuclear ribosomal DNA phylogeny of *Acer* inferred with maximum likelihood, splits graphs, and motif analyses of 606 sequences. *Evolutionary Bioinformatics* 2:279-294.

Holland B, and Moulton V. 2003. Consensus networks: A method for visualising incompatibilities in collections of trees. In: Benson G, and Page R, eds. *Algorithms in Bioinformatics: Third International Workshop, WABI, Budapest, Hungary*. Berlin, Heidelberg, Stuttgart: Springer Verlag, 165-176.

Holland BR, Benthin S, Lockhart PJ, Moulton V, and Huber KT. 2008. Using supernetworks to distinguish hybridization from lineage-sorting. *BMC Evolutionary Biology* 8:202.

Huber KT, Moulton V, Steel M, and Wu T. 2016. Folding and unfolding phylogenetic trees and networks. *Journal of Mathematical Biology*. 10.1007/s00285-016-0993-5

Huson DH, and Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23:254-267.

Mardulyn P. 2012. Trees and/or networks to display intraspecific DNA sequence variation? *Molecular Ecology* 21:3385–3390.

Mindell DP. 2013. The Tree of Life: metaphor, model, and heuristic device. *Systematic Biology* 62:479–489.

Morrison DA. 2010. Using data-display networks for exploratory data analysis in phylogenetic studies. *Molecular Biology and Evolution* 27:1044–1057.

Morrison DA. 2014. Is the Tree of Life the best metaphor, model, or heuristic for phylogenetics? *Systematic Biology* 63:628–638.

Potts AJ, Hedderson TA, and Grimm GW. 2014. Constructing phylogenies in the presence of intra-individual site polymorphisms (2ISPs) with a focus on the nuclear ribosomal cistron. *Systematic Biology* 63:1–16.

Rambaut A. 2014. Figtree, a graphical viewer of phylogenetic trees. 1.4.2 ed. Edinburgh: The author, Institute of Evolutionary Biology, University of Edinburgh.

Rannala B, and Yang Z. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *Journal of Molecular Evolution* 43:304-311.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, and Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* 61:539–542. 10.1093/sysbio/sys029

Salichos L, Stamatakis A, and Rokas A. 2014. Novel information theory-based measures for quantifying incongruence amoung phylogenetic trees. *Molecular Biology and Evolution* 31:1261–1271.

Schliep KP. 2011. Phangorn: phylogenetic analysis in R. *Bioinformatics* 27:592–593.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.

Yang J, Grünewald S, and Wan X-F. 2013. Quartet-Net: A quartet-based method to reconstruct phylogenetic networks. *Molecular Biology and Evolution* 30:1206–1217.

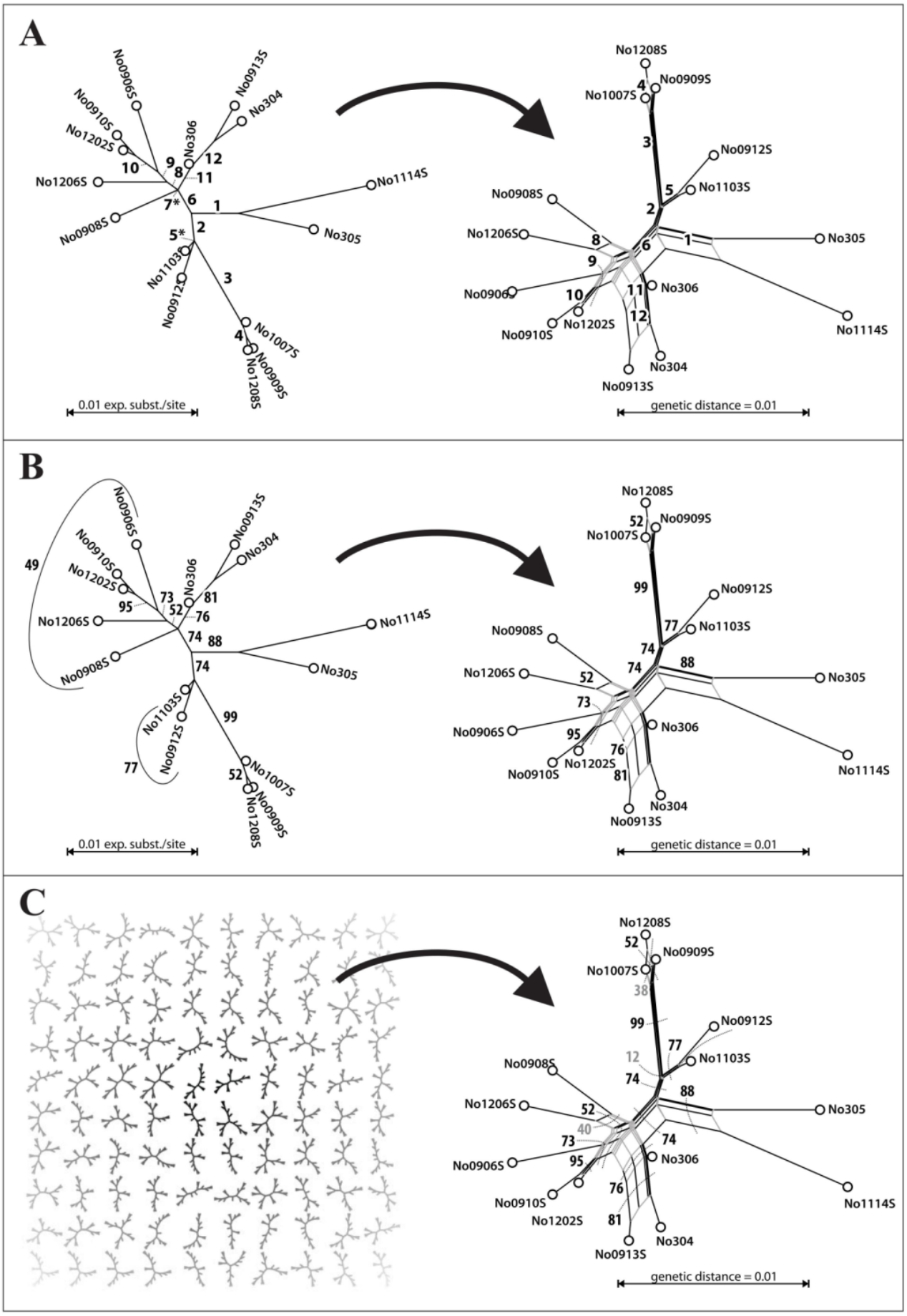

Figure 1 Mapping tree information onto a network using a mitochondrial gene (*cytB*) woodmouse (*Apodemus sylvaticus*) dataset (the standard test set from the APE library).
A. Identification of edge bundles (in black) in a neighbour-net (NN) network based on uncorrected *p*-distances that correspond to branches (labelled 1-12) in a maximum likelihood (ML) tree. Asterisks refer to zero-length tree branches (soft polytomies), of which one (branch 7) has no corresponding edge bundle in the NN network.
B. Nonparametric ML bootstrap (ML-BS) support for all branches (branch labels) defining the ML tree mapped on the corresponding edge bundles of the NN network.
C. Frequencies of bipartitions found in the ML-BS pseudoreplicates mapped on the corresponding edge bundles of the NN network using a threshold of 10% (i.e. any edge is labeled that occurs in at least 100 of the 1000 ML-BS pseudoreplicates). Edge bundles not found in the ML tree are labelled using grey numbers.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2054v1 | CC-BY 4.0 Open Access | rec: 18 May 2016, publ: 18 May 2016

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2054v1 | CC-BY 4.0 Open Access | rec: 18 May 2016, publ: 18 May 2016