Community detection on networks of topologies and bipartitions identifies conflicting phylogenetic signal

Jeremy Ash[1], Wen Huang[2], Guifang Zhou[2], James Wilgenbusch[3], Kyle Gallivan[2], Melissa Marchand[2], Jeremy M. Brown[1]

1 Deptartment of Biological Sciences, Louisiana State University, Baton Rouge, LA

2 Department of Mathematics, 3 Department of Scientific Computing, Florida State University, Tallahassee, FL

Introduction

In the phylogenomic era, there exists an ever greater need to fully characterize the information in sets of phylogenies. One potentially rich and underexploited avenue is to use networks to characterize this information [1, 2, 3], either of phylogenies themselves or their component bipartitions. Once formed, community detection methods [4, 5, 6, 7] allow researchers to explore relationships between competing phylogenetic signals in these networks. These competing signals may indicate heterogeneity in evolutionary history underlying the data or systematic error. We have implemented tools for network construction and community detection in the software TreeScaper [8]. Here, we perform an initial simulation-based benchmarking of these community detection approaches. Across a broad range of simulation scenarios, we find that when there is little conflicting signal in a multiple sequence alignment (MSA), TreeScaper recovers little evidence for community structure. However, when the MSA contains strong support for a few distinct phylogenies, TreeScaper recovers clear evidence for community structure. Network-based approaches provide a new, quantitative approach for exploring conflicting signal in sets of phylogenies.

Types of Phylogenetic Networks

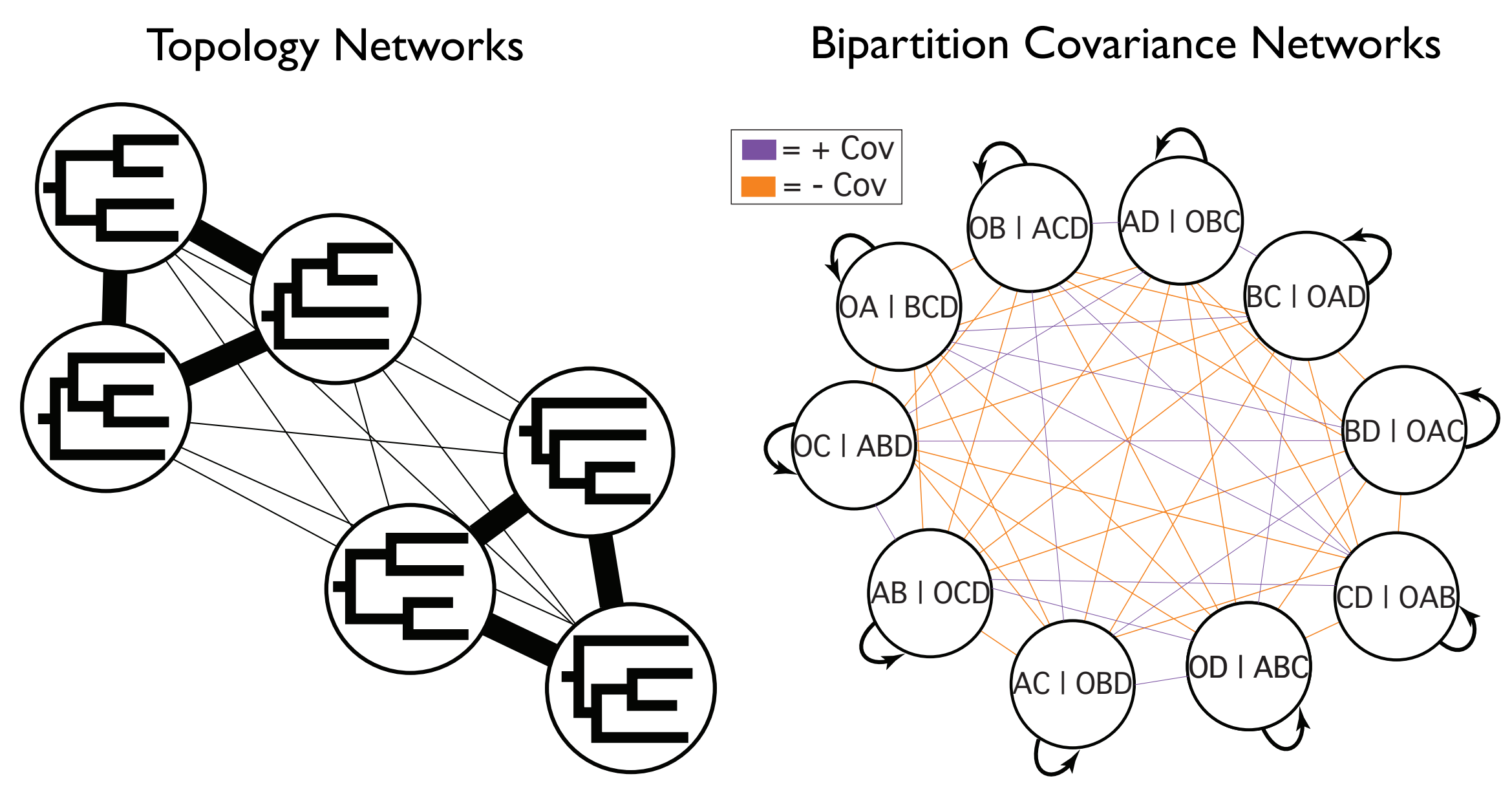

Figure 1. Networks used for analyses. (a) Topologies are nodes and their affinities, the similarities between them, are edge weights. (b) Bipartitions are nodes and their positive or negative covariances are edge weights. In these examples, (a) has community structure and clear evidence for conflicting signal, and (b) has no distinct community structure and little evidence for conflicting signal.

Methods

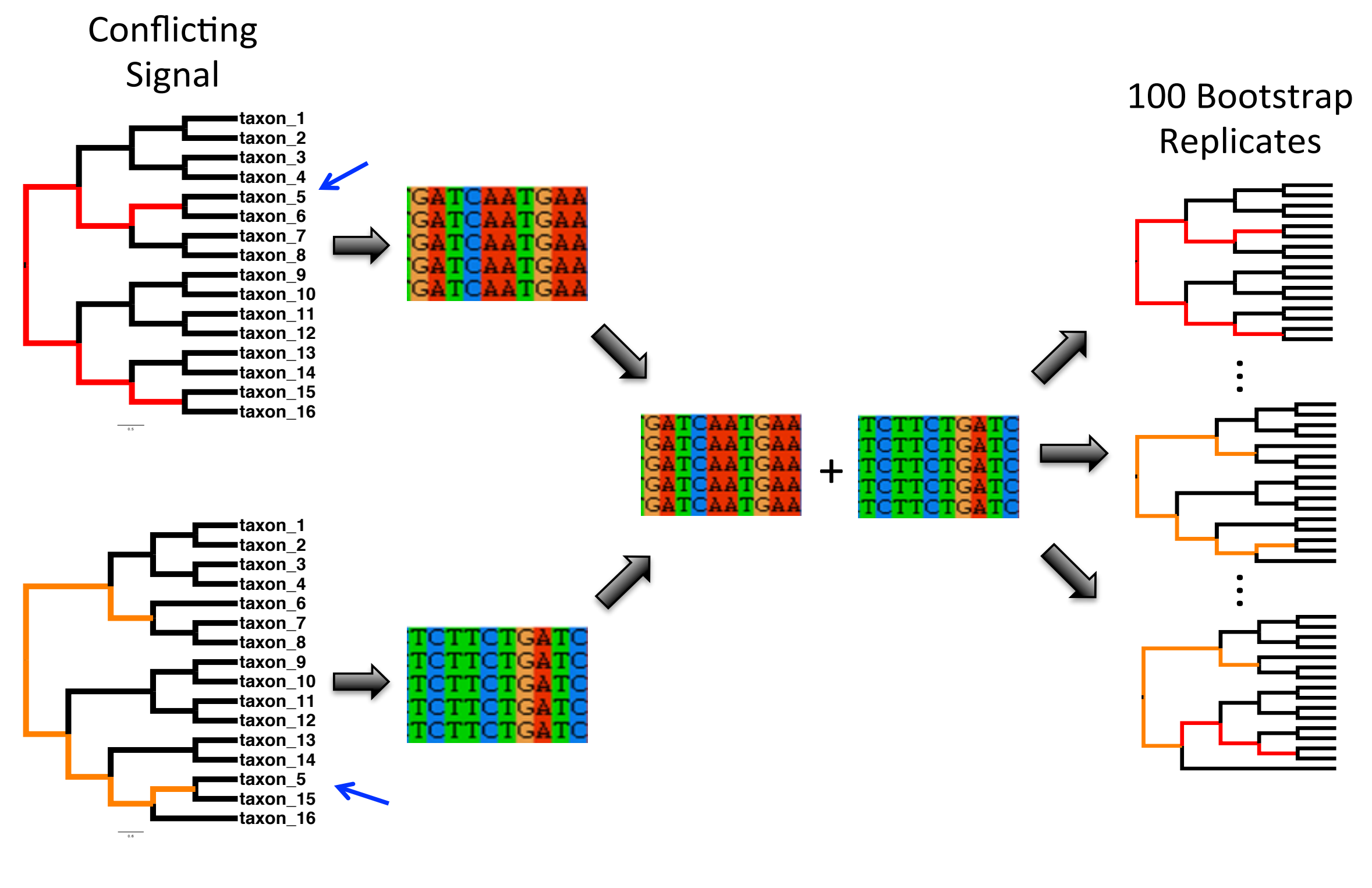

Figure 2. Simulation of tree sets with conflicting signals. Two guide trees that only differed in their placement of taxon 5 (the rogue taxon) were used to simulate in Seq-Gen [9] two equally sized MSAs. The MSAs were concatenated together. Bootstrap analyses were performed in Garli [10] on the concatenated MSAs.

Results

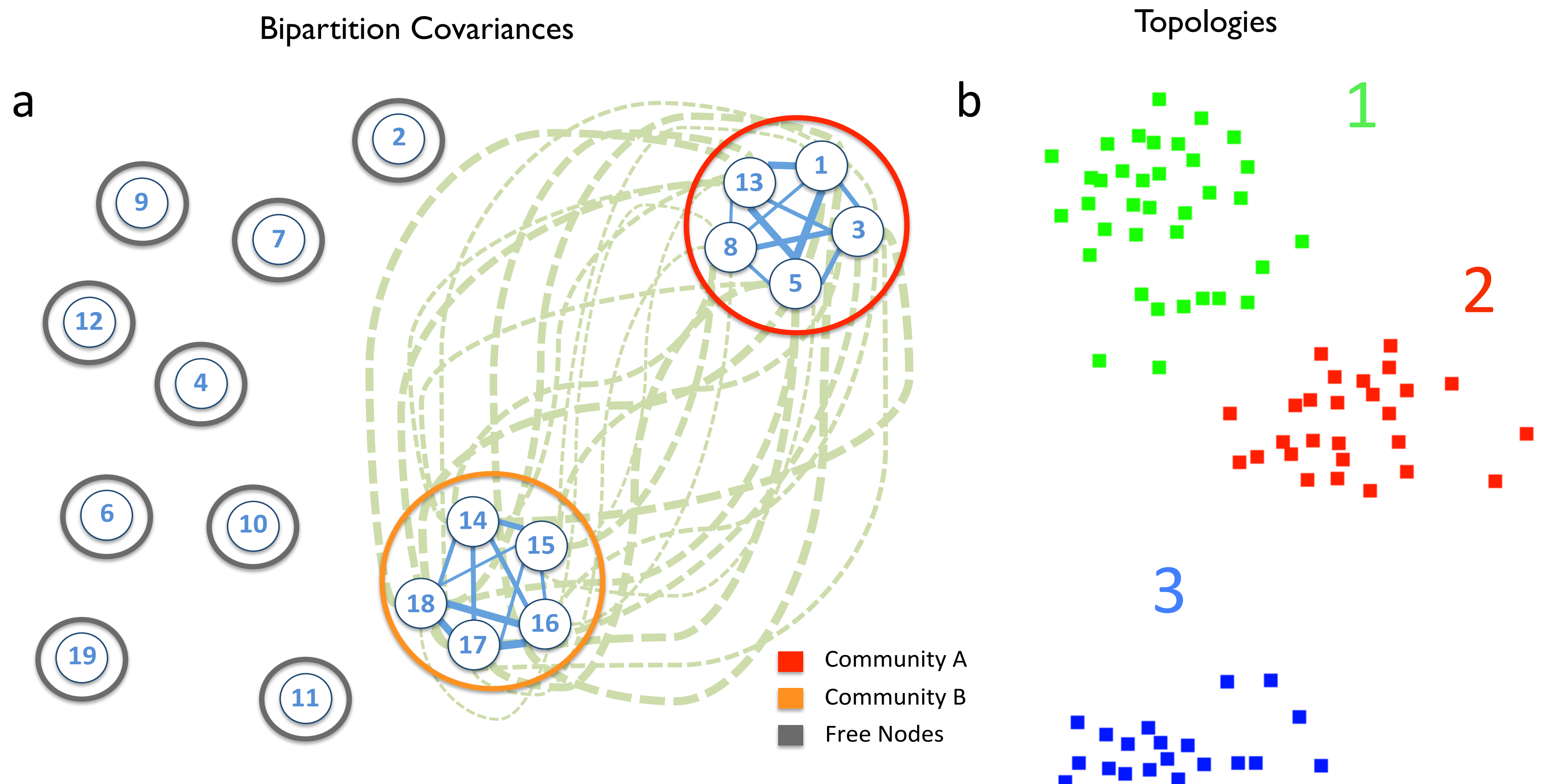

Figure 3. Network analyses of a simulated tree set in TreeScaper. (a) Community structure of bipartition covariance network. Blue and green edges indicate positive and negative covariances respectively. Colored communities contain conflicting bipartitions. (b) Community structure of topology network. Colors indicate the three affinity communities found by TreeScaper. Non-linear dimensionality reduction projection [11] is for display purposes only.

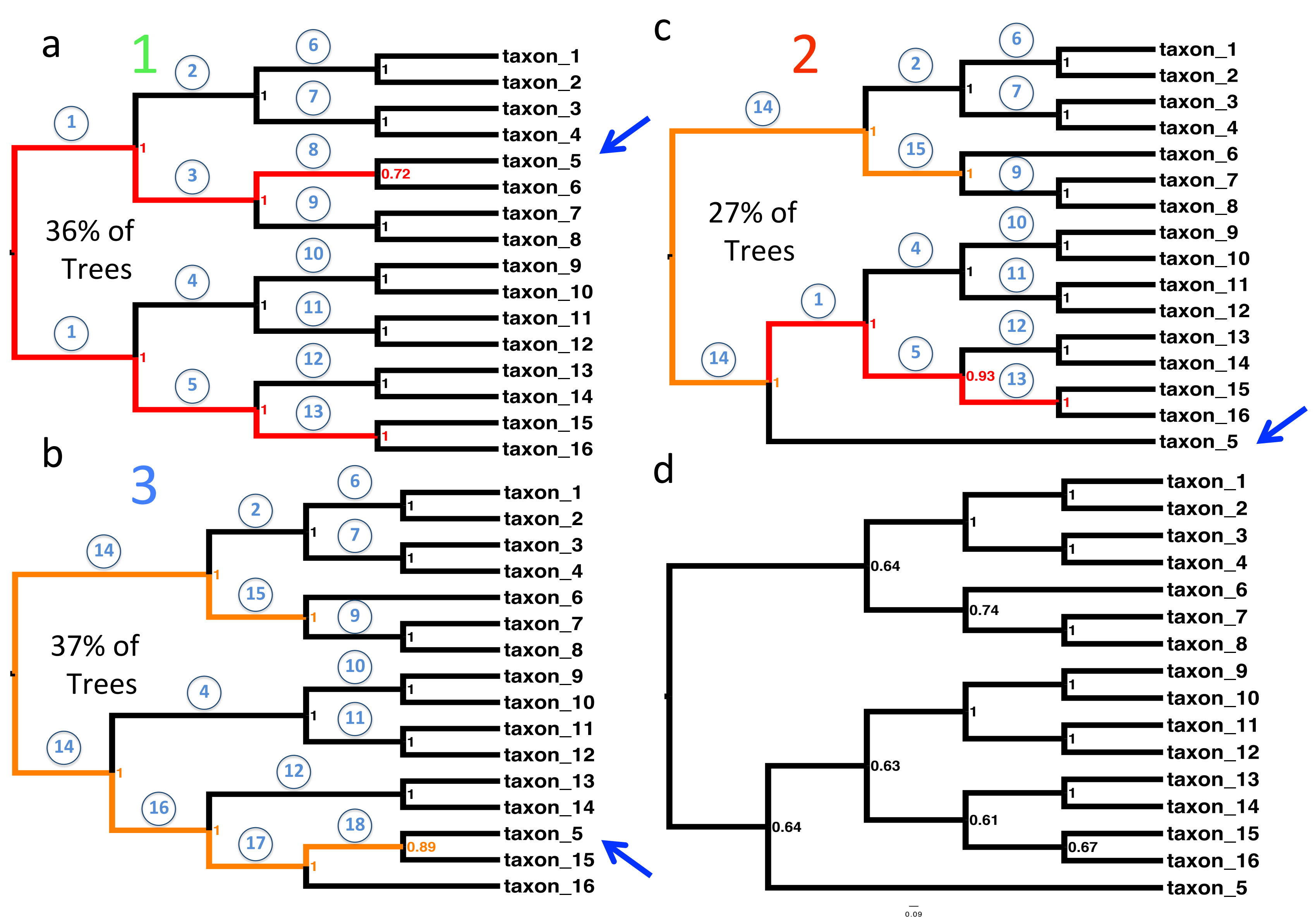

Figure 4. Trees representative of the simulated treeset. (a-c) Consensus trees of the three affinity communities found by TreeScaper. Conflicting bipartitions are in color. (d) Consensus tree of the entire tree set.

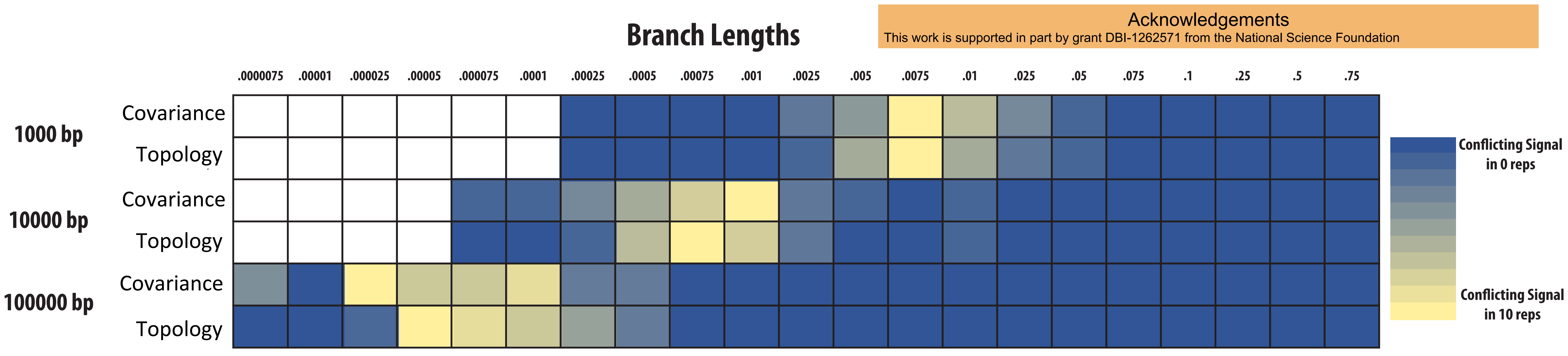

Figure 5. Window of guide tree branch lengths in which conflicting signal is detected. At very short or long branches, there is a lack of phylogenetic signal and many topologies in the tree set. Between these extremes, there is strong support for a few topologies. Heat maps show the number of replicates in which the known conflict was detected by TreeScaper. Constant-Potts Model (CPM) [5] was used for community detection.

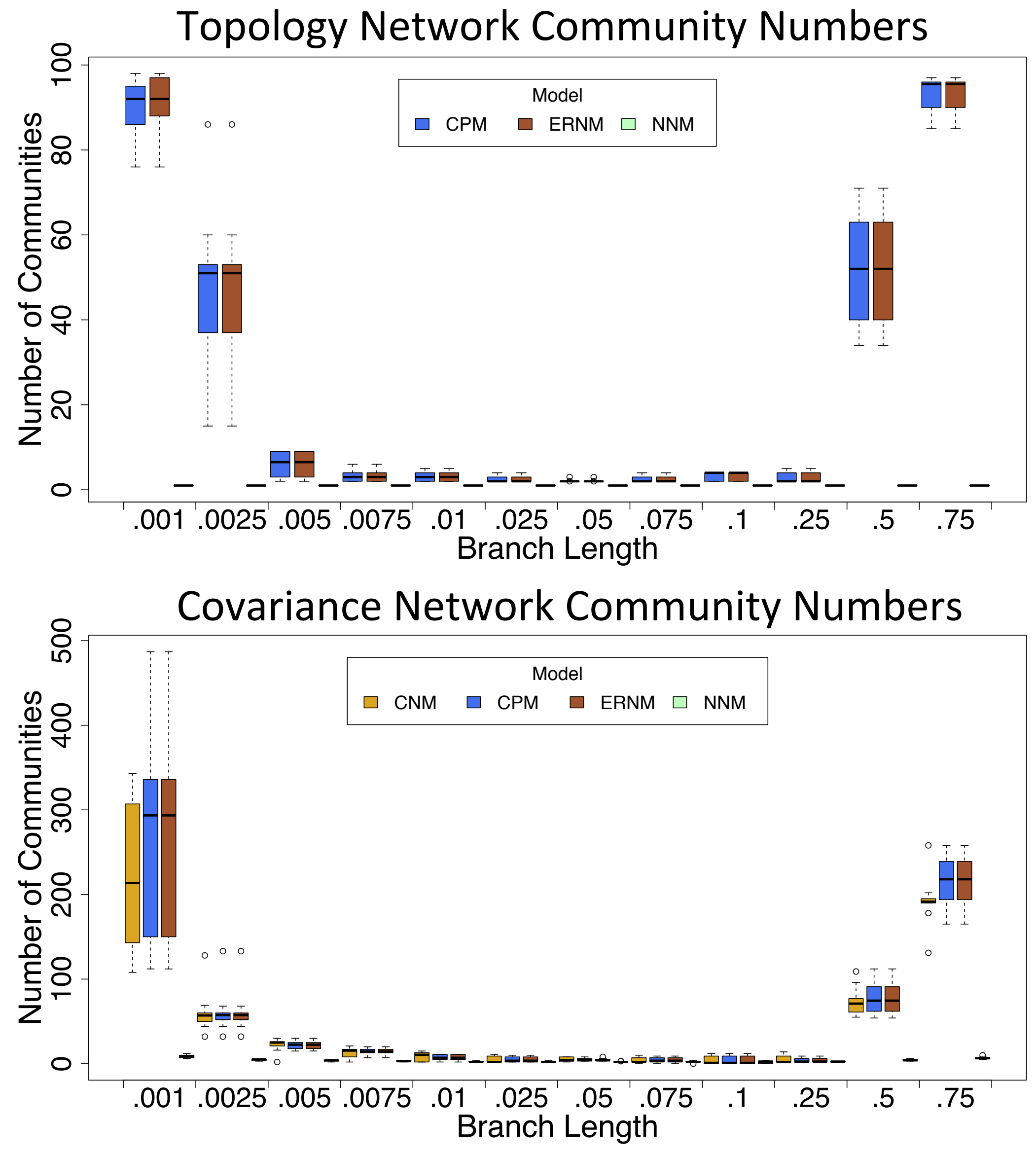

Figure 6. Box plots of the number of communities found using several community detection models: Configuration Null Model (CNM) [4], Constant-Potts Model (CPM) [5], Erdos-Renyii Model (ERNM) [6], No Null Model (NNM) [7].

Conclusions

- When there is little conflicting signal in a data set, community detection methods find little evidence for community structure.
- When there is strongly supported conflicting signal, community detection methods identify a few well supported topologies and their conflicting bipartitions.

Future Directions

- Comparison of community detection methods to alternate rogue taxon identification methods
- Use of community detection in posterior prediction and parametric bootstrapping

References

1. D.R. Maddison. (1991). The discovery and importance of multiple islands of most- parsimonious trees. Systematic Zoology. 40:315–328.
2. R.D.M. Page. (1993). On islands of trees and the efficacy of different methods of branch swapping in finding most-parsimonious trees. Systematic Biology. 42:200–210.
3. L.A. Salter. (2001). Complexity of the likelihood surface for a large DNA dataset. Systematic Biology. 50:970–978.
4. M.E.J. Newman (2006). Modularity and community structure in networks. PNAS. 103 (23) 8573-8574.
5. V.A. Traag & P. Van Dooren. (2011). Narrow scope for resolution-limit-free community detection. Physical Review E. 84:016114.
6. J. Reichardt & S. Bornholdt. (2006). Statistical Mechanics of Community detection. Phys. Rev. E 74, 016110.
7. M.E.J. Newman & M.Girvan (2004). Finding and evaluating community structure in networks. Phys. Rev. E 69, 026113.
8. Huang, W., et al. (2010). TreeScaper: software to visualize tree landscapes. http://bpd.sc.fsu.edu/index.php/diagnostic-software.
9. Rambaut A & Grassly NC. (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci. 13(3):235-8.
10. Zwickl, D. J., 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. dissertation, UT Austin.
11. Hillis, D. et al. (2005). Analysis and visualization of tree space. Sys. Bio. 54, 471-482.

Acknowledgements

This work is supported in part by grant DBI-1262571 from the National Science Foundation