Generating Synthetic Coevolutionary Histories

Ben Drinkwater
Associate Professor Michael Charleston
School of Information Technologies
FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

Introduction

One of the most important applications of cophylogenetic research is the study of *<u>zoonoses</u>*, diseases that switch hosts from one species to another. In humans it is currently estimated that 75% of emergent diseases are zoonoses [1].

- A known problem in this field is the lack of tools to generate synthetic data sets which has led to biased evaluations of algorithms to solve the cophylogeny reconstruction problem [2,3].
- An example of such an evaluation is Dowling's comparison of Brookes Parsimony Analysis and TreeMap which used manually produced histories rather than synthetic data [4].

Creating Synthetic Histories

There exists an exponential number of topologies for a single bifurcating tree [5] and therefore the growth rate is even greater when considering a pair of trees and associations between them.

- To overcome this problem and create a even distribution of solutions my framework utilizes both rooted and unrooted trees, along with the ability to clone part or full portions of the host tree to create all possible histories.
- Further, this framework generates a set of unique topologies ensuring that the solutions generated have a fair distribution of problem sets and are not clustered as can be the case for rooted trees which generally produce balanced trees when random growth is enforced.

Possible Data Sets

- All four known types of coevolutionary instances as defined by Hommola *et al* [6] can be generated using this framework.

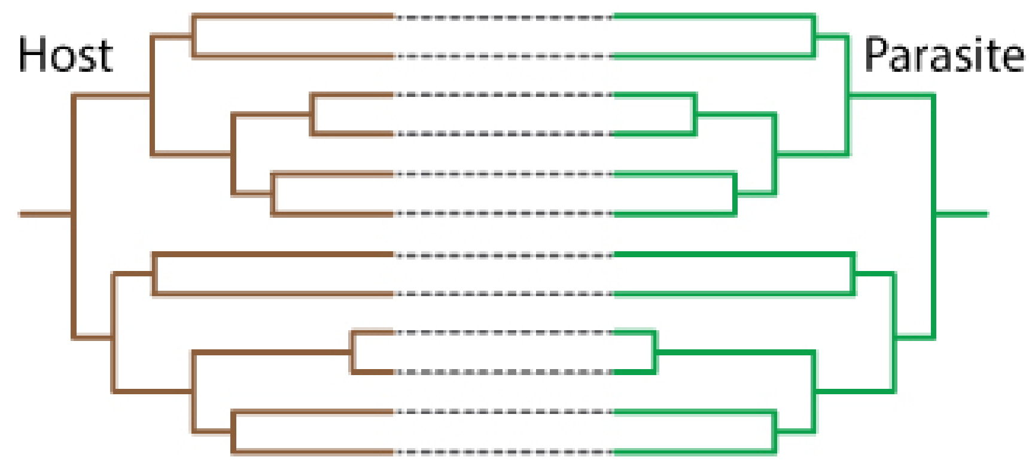

Figure 1: Identical Histories

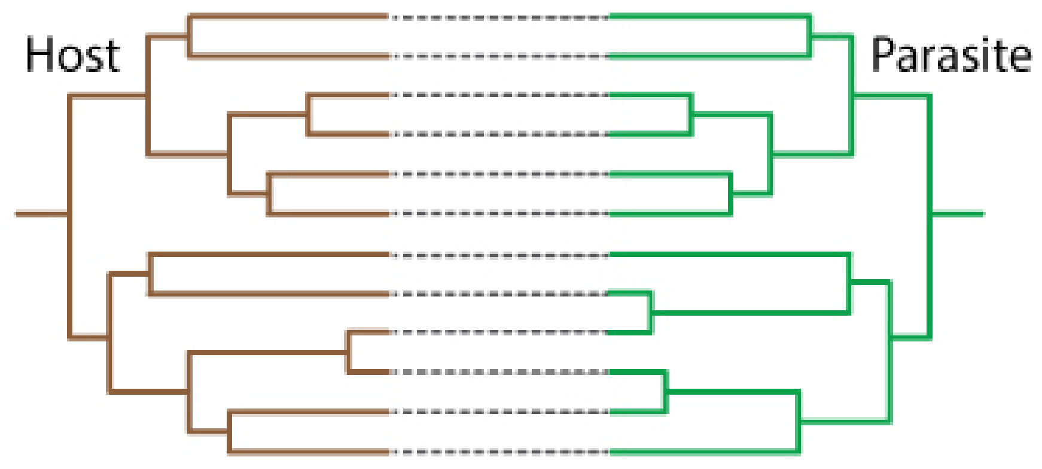

Figure 2: Topologically Similar Histories

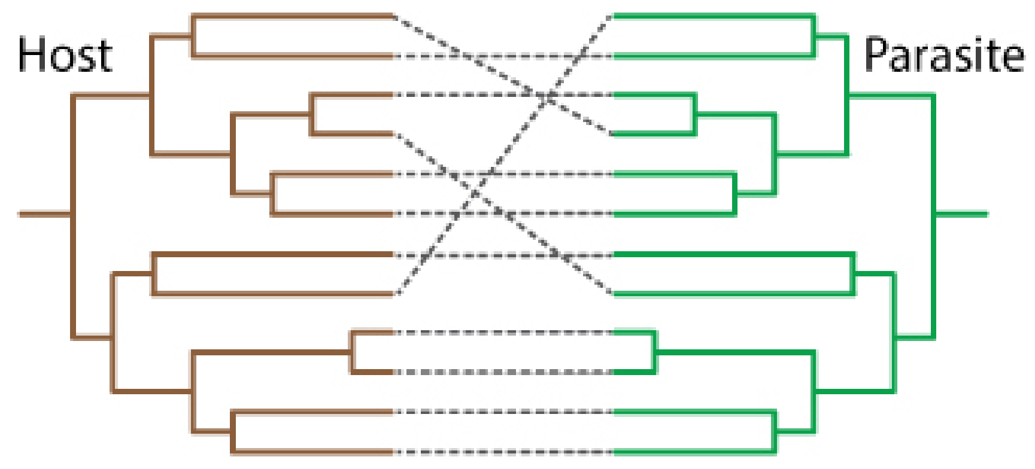

Figure 3: Randomized Associations

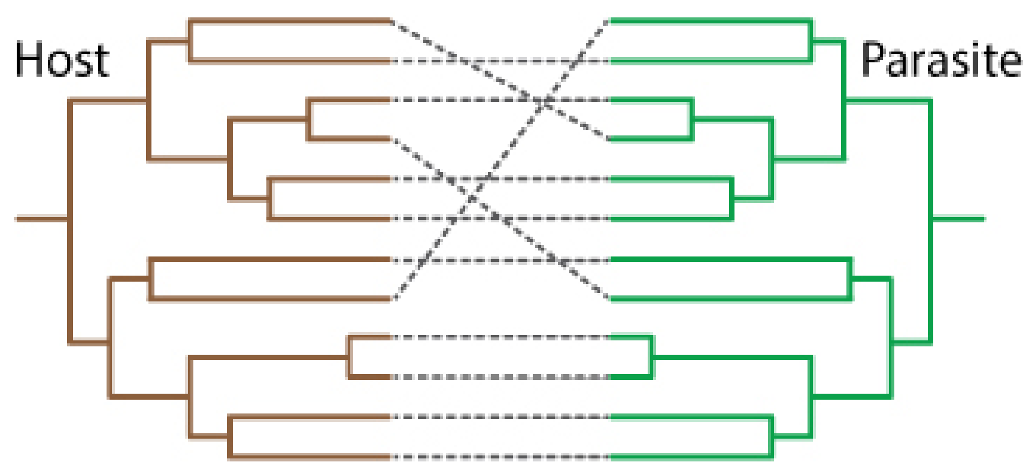

Figure 4: Independent Histories

Using this framework to evaluate existing solutions to this problem

To evaluate the histories produced by this framework it was applied to comparison of two existing techniques for solving the cophylogeny reconstruction problem.

- The tree metric and cophylogeny mapping techniques are well established techniques for solving this problem. It is however an open question in the field as to which of these techniques provides the best approximation for the statistical significance of coevolution for a particular coevolutionary instance.
- Two well known techniques from each of these respective fields are **Parafit** [7] and **Jane** [8].
- For this experiment the recent R implementation of Parafit [9] was selected and the third iteration of Jane.
- To compare these two applications a set of topologically different coevolutionary instances were created ranging from identical through to purely independent histories using five different trees sizes including 10, 15, 20, 25 and 30 leaf nodes. Overall there were **27500 test cases** utilized for this comparison.

Results

The similarity of the solutions produced

- There is no statistically significant difference between Jane and Parafit when comparing over all sample histories with both algorithms holding an inverse relationship between the likelihood for coevolution and the topological similarity.

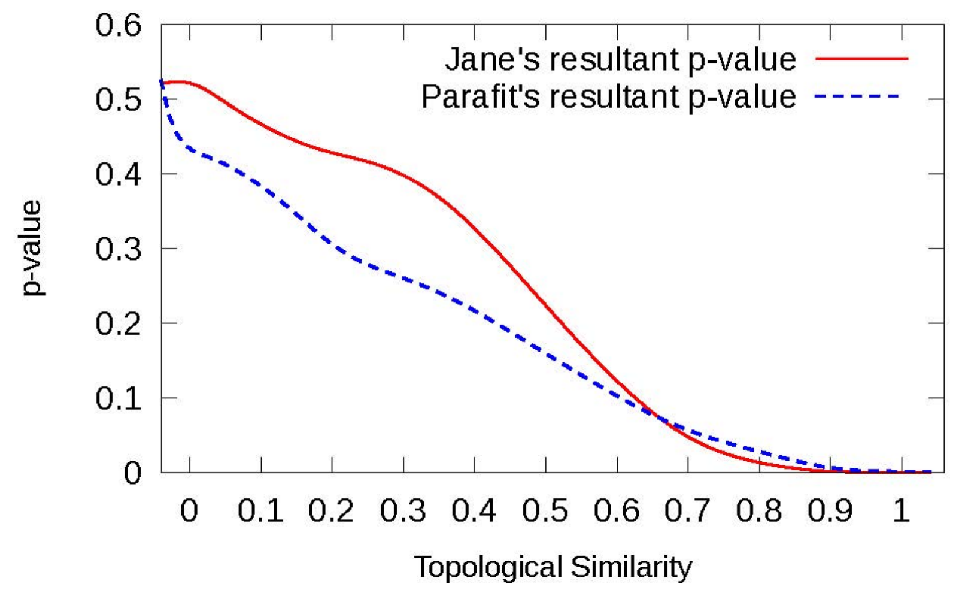

Figure 5: The mean for each Test Case size per topologically similar history.

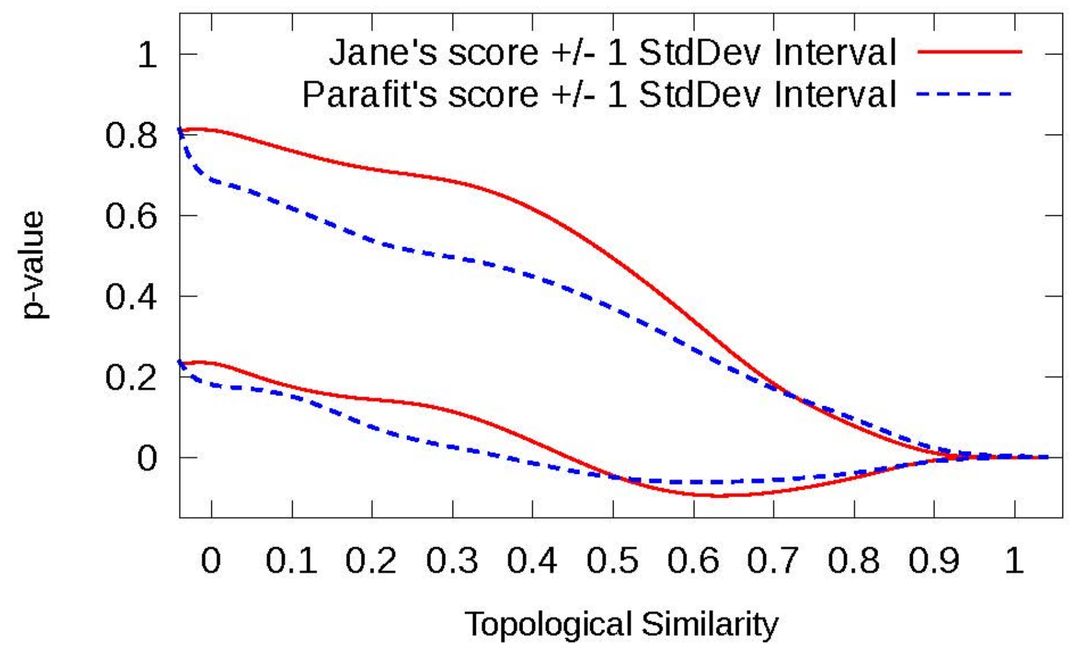

Figure 6: One Standard Deviation interval either side for the *p*-values for each approach.

The number of disagreements

- Although the comparison between Parafit and Jane across the full set of data presents a strong congruence between each technique, there are many cases where Jane and Parafit disagree. For a *p*-value of 0.05 there are up to 11% of cases in which they disagree. This number is reduced to less than 5% when the p-value is reduced to 0.01. This result can be seen in Figure 7.

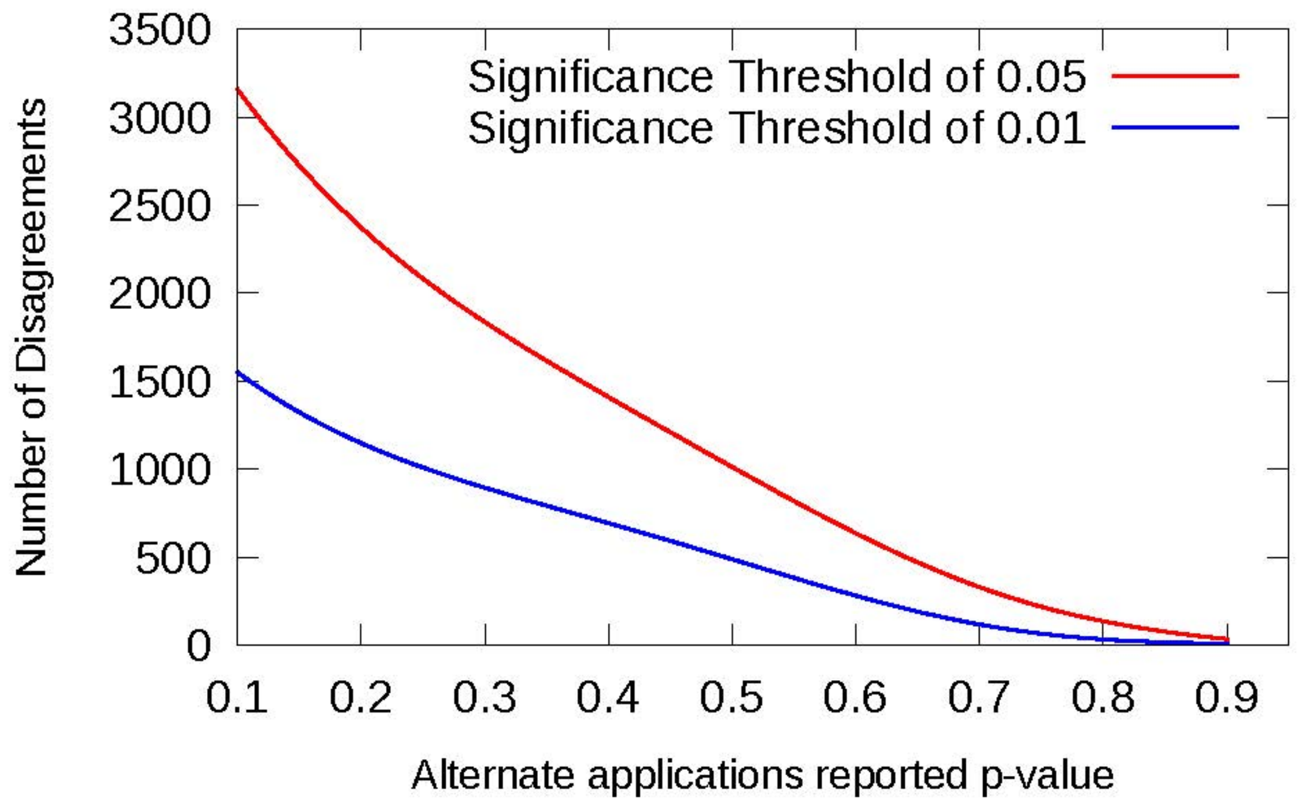

Figure 7: The number of instances where Jane and Parafit disagree on the significance of a solution when the statistical significance threshold is set at either 0.05 or 0.01.

Conclusion

- This research has identified the benefits of using a lower *p*-value when using either Jane or Parafit. Further it was observed that there is a potential of up to a 50% reduction in the number of disagreements between Jane and Parafit when reducing the *p*-value from 0.05 to 0.01.

Further Work

Expand this framework

- This framework only generates Tanglegrams which have a single association between each node between the host and parasite tree. An extension to this framework would be to include the generation of coevolving histories with multiple associations between host and parasite trees, along with the potential of leaf nodes with no associations linking branches.
- Further, this framework only allows for the host and parasite trees to have the same number of leaf nodes with the addition of multiple associations. Different tree sizes will also be possible allowing for more biologically accurate synthetic histories to be generated.

Apply this framework to other open questions in this field

- A new area for cophylogenetic research is the application of approximation algorithms to this problem. This framework has the potential to generate comprehensive training and validation data sets to assist in the creation and validation of new metaheuristic solutions to this problem.

References

[1] F. M. Tomley and M. W. Shirley, “Livestock infectious diseases and zoonoses introduction”, Philosophical Transactions of the Royal Society B-Biological Sciences, vol.364, no.1530, pp.2637-2642, 2009.

[2] M. E. Siddall, “Bracing for another decade of deception: the promise of secondary brooks parsimony analysis,” Cladistics, vol.21, no.1, pp.90--99, 2005.

[3] M. E. Siddall and S. L. Perkins, ``Brooks parsimony analysis: a valiant failure,” Cladistics-the International Journal of the Willi Hennig Society, vol.19, no.6, pp.554-564, 2003.

[4] A. Dowling, “Testing the accuracy of TreeMap and Brooks parsimony analyses of coevolutionary patterns using artificial associations,” Cladistics, vol.18, no.4, pp.416-435, 2002.

[5] J. Felsenstein, “Inferring phylogenies”, Sinauer Associates Sunderland, vol 2, p. 30, 2004

[6] K. Hommola, J. Smith, Y. Qiu, and W. Gilks, “A Permutation Test of Host--Parasite Cospeciation,”' Molecular biology and evolution, vol. 26, no. 7, p. 1457, 2009.

[7] P. Legendre, Y. Desdevises, and E. Bazin, “A statistical test for host–parasite coevolution,” Systematic Biology, vol. 51, no. 2, pp. 217–234, 2002.

[8] C.Conow, D.Fielder, Y.Ovadia, and R.Libeskind-Hadas, “Jane: A new tool for the cophylogeny reconstruction problem,” Algorithms for Molecular Biology, vol.5, no.1, p.16, 2010.

[9] E. Paradis, B. Bolker, J. Claude, H. Cuong, R. Desper, J. Legendre, Y. Noel, J. Nylander, R. Opgen-Rhein, A. Popescu et al., “Package ape,” 2012.