Performance study of supertree methods

Patrick Niehaus and Mark Holder, PhD

[niehaupr@email.sc.edu](mailto:niehaupr@email.sc.edu), [mtholder@ku.edu](mailto:mtholder@ku.edu)

**Introduction:**

Evolutionary biologists use phylogenetic trees to conceptualize, visualize, and analyze the relationships among varying species [1]. A supertree consists of several smaller phylogenetic tree estimates merged together [2-4]. The Open Tree of Life project is currently attempting to characterize the accuracy of supertrees when dealing with a large number of taxa. The results of this study may help characterize the efficiency and accuracy of a known supertree method called Matrix Representation with Parsimony (MRP) to an unpublished method known as the Tree Alignment Graph (TAG). Since TAG is a new method, it needs to be evaluated with respect to accuracy and running time.

Background:

MRP works by translating each branch of each individual tree into a matrix column, this process is repeated until a character matrix is formed [2]. All taxa that descend from the branch are represented in the matrix with a ‘1’, other taxa that are found in the tree are represented with a ‘0’, and taxa that aren’t found in the tree are represented with the missing data symbol ‘?’. Once complete, the character matrix includes all of the phylogenetic relationships contained in the source trees. Using parsimony, estimating a tree from this matrix then provides a supertree (Refer to figure 2).

TAG exploits the fact that trees are actually part of a graph. Phylogenetic trees are directed and require that each node as at most one parent. However, TAG relaxes these requirements, and allows multiple trees to be combined into one a common graph [1] (Refer to figure 2). Through this process, computational time and resources are believed to be greatly minimized.

It is currently believed that MRP is more effective and/or accurate for smaller taxa sets. When faced with a larger taxa set, MRP loses accuracy. TAG, however, is designed to work best with larger taxa inputs and is believed to be more accurate and efficient when working with a larger taxonomy input. Through this study, the two methods are being compared in an attempt to gain a better understanding of the two known methods.

A consensus tree, summarized as CON later in the results, is a tree that has all of the taxa found in the input trees. This is very helpful in the comparison of trees. False Positives occur when an input tree has an extra branch not found in the true tree. False Negatives occur when the true tree has a branch not in one of the input trees. When false negatives and false positives are added together, the total number of errors in the tree being compared is known.

Methods:

This study first requires an input taxonomy. From here, a true tree and several input trees are generated. Each input tree is then run individually through MRP and TAG as an estimate and are then compared to the true tree (Refer to figure 3). Whichever estimate is closest to the true tree (e.g. the least amount of false negatives and positives) and takes the least amount of time is believed to be the more accurate and efficient algorithm.

Results and Implications:

At this point of the study, MRP and TAG have only been compared with the primate’s taxonomy (around 1400 input lines). With the smaller taxonomy input, there are many differences in the outputs produced by MRP and TAG. **At 10 input trees, TAG appeared to be more effective then MRP especially in the number of false negatives (refer to figure 6). Surprisingly, TAG’s accuracy didn’t increase when the number of input trees were increased but the accuracy of MRP increased significantly (refer to figure 7).** While MRP seems to be the more effective algorithm in this part of the study, it is important to remember that TAG is designed to be more effective with a larger taxonomy input. Another implication is that TAG is still in development and could possibly become more effective with smaller taxonomies when completed. In the next part of this study, MRP and TAG will be compared with a larger taxonomy (e.g. the tree of life which is around 2.7 million input lines).

**Figures 4 and 5 show how the number of input trees affect the outputs of MRP and TAG in terms of false negatives and positives. MRP appears to become more effective in terms of false negatives and positives as the number of input trees increases. TAG, however, appears to become less effective as the number of trees increases, and even appears to have about the same number of false negatives/positives after around 60 input trees. Further investigation is necessary to explain this.**

\*\* I think only one of these is necessary to explain the results in this journal, which figures do you think are more effective? (e.g. figures 4 and 5 or figures 7 and 8).

Bibliography:

1. Smith SA, Brown JW, Hinchliff CE (2013) Analyzing and Synthesizing Phylogenies Using Tree Alignment Graphs. PLoS Comput Biol 9(9): e1003223. doi:10.1371/journal.pcbi.1003223

2. Bininda-Emonds OR, editor (2004) Phylogenetic supertrees: Combining information to reveal the Tree of Life (Computational Biology). Springer, 564 pp. URL http://www.amazon.com/Phylogenetic-supertrees-Combining-information-Computational/dp/1402023294.

3. Sanderson MJ, Purvis A, Henze C (1998) Phylogenetic supertrees: Assembling the trees of life. Trends in Ecology & Evolution 13: 105–109.

4. Davis KE, Hill J (2010) The Supertree Tool Kit. BMC Research Notes 3: 1–6.

5.Von Haeseler, A. "Do we still need supertrees?." BMC biology 10.1 (2012): 13.

6. Bininda-Emonds, O, R. Beck, and A. Purvis. "Getting to the Roots of Matrix Representation." Systematic Biology 54.4 (2005): 668-72.