

# Searching for Optimal Trees Under Maximum Parsimony<sup>1</sup>

Tiffani L. Williams

*Department of Computer Science, University of New Mexico, Albuquerque, NM 87131*

*{tlw@cs.unm.edu}*

With the recent explosion in the amount of genomic data available, biologists are able to consider the construction of evolutionary trees consisting of hundreds and possibly thousands of taxa. Phylogeny reconstruction is a difficult computational problem since most techniques are based upon heuristics for NP-hard optimization problems. Here, we are interested in phylogenetic trees reconstructed based on the maximum parsimony (MP) criterion, which typically returns a number of optimal-scoring trees. In the absence of knowing the “true” tree, researchers rely on substitute criteria such as MP scores. Yet, is there a strong correlation between topological accuracy and parsimony scores? Our conjecture is that once an MP search reaches its top trees, there is little phylogenetic signal to select among the trees under consideration. At this point, the trees are topologically indistinguishable. Conducting an MP search based on the observed tree topologies in conjunction with their MP scores may present a fruitful approach to reconstructing phylogenies. Such an approach may allow a search to terminate faster without any corresponding loss in accuracy.

**Experimental Methodology:** Our objective is to understand the relationship between tree topologies and MP scores. For each biological dataset, we store all of the trees found by the Parsimony Ratchet, the most successful MP search strategy to date. Afterwards, we categorize the trees found for a particular dataset by placing them into scoring bins. Let  $s$  represent the optimal score found for a dataset.  $\text{OPT}_i$  is the set of trees with score  $s + i$ . All optimally-scoring trees appear in the set  $\text{OPT}_0$ . Furthermore, let  $\text{AOPT}_i$  represent the set of trees with a score of at most  $s + i$ . More formally,  $\text{AOPT}_i = \sum_{j=0}^i \text{OPT}_j$ . (Note that  $\text{OPT}_0 \equiv \text{AOPT}_0$ .) For each set, we compute its strict- and majority-consensus tree.

Besides organizing trees into OPT and AOPT bins and comparing them, there is an additional purpose for collecting all of trees encountered during a ratchet search. We intend to establish that the relationship between tree topologies and MP scores is not as strong as previously thought. As a result, new algorithms can be developed to stop the search as soon as possible. Since we stored all of the phylogenetic trees found by parsimony ratchet, we can test our termination heuristics on the saved collection of trees and compare the results with those of the original parsimony search.

**Observations:** Our preliminary observations on biological datasets (such as the well-studied 500 rbcL dataset) suggest that strictly searching for trees in  $\text{OPT}_0$  may not be worth the computational effort. Our results show that the majority consensus tree for  $\text{OPT}_0$  is quite similar (based on RF distance) to the majority tree found for each  $\text{OPT}_i$  and  $\text{AOPT}_j$  set, where  $i, j > 0$ . Another interesting trend occurs when comparing the strict consensus tree formed from  $\text{OPT}_0$  with the strict consensus trees of  $\text{OPT}_i$  and  $\text{AOPT}_j$ , where  $i, j > 0$ . Here, the average number

---

<sup>1</sup>Joint work with Tanya Berger-Wolf, Bernard Moret, Usman Roshan, and Tandy Warnow. This work was supported by the National Science Foundation under ACI00-81404, DEB 01-20709, EIA 01-13095, EIA 01-21377, and EIA 02-03584, and by an Alfred P. Sloan Postdoctoral Fellowship in Computational Molecular Biology.

of false positives is 0, which implies that suboptimal trees do not contain edges that are not in an optimal tree. Simply stated, the suboptimal trees are proper subsets of the optimal trees. The above observations suggest that prolonging a search in order to find optimally-scoring trees is a misguided effort if a strict- or majority-consensus tree is desired. Additional observations include small distances between the trees in a set, the low number of iterations required to find trees in  $OPT_1$ , and the large amount of computing time required to obtain trees in  $OPT_0$ .

Currently, our preliminary observations are based strictly on biological datasets, which are not designed to test specific aspects of reconstruction algorithms. Furthermore, we cannot judge the outcomes on the basis of accuracy (since we do not know the “true” tree) nor can we use them to predict behavior on other datasets. On the other hand, biological datasets offer data with all of the peculiarities that are difficult to produce in simulations. We realize that a complete, experimental study of the relationship between topological accuracy and MP scores requires carefully designed experiments based on simulated and biological data. Therefore, we are creating simulation data using a variety of evolutionary substitution models to study thoroughly our preliminary observations based on “real-world” datasets.