Partial Tree Mixing – a Novel Approach to Phylogenetic Search

Kenneth Sundberg, Mark Clement, Quinn Snell, Dan Ventura, Michael F. Whiting and Keith A. Crandall

September 16, 2011

1 Introduction

Phylogenetic search is an NP-Hard [3] problem. It is however important to the analysis of biological sequences and the testing of evolutionary hypothesis [6]. As such it is necessary to employ heuristic methods.

A phylogenetic search begins by using a greedy heuristic to build an initial tree. This initial tree is then improved by the full search. Unfortunately, the greedy nature of the starting trees limits the effectiveness of the full search.

Partial Tree Mixing (PTM) addresses this issue through the use of a global representation of partition based tree space [14]. Using this representation PTM is able to quickly begin exploring this space with a global search strategy.

1.1 Related Work

The most common heuristic method for phylogenetic search is a form of hill climbing based on Tree Bisection and Reconnection (TBR) [1]. This method requires an initial tree which is commonly built using methods such as UPGMA (Unweighted Pair Group Method with Arithmetic Mean) [9] and neighbor joining [13], as well as stepwise maximum parsimony. Both distance methods and stepwise maximum parsimony are $O(n^2)$ algorithms.

Distance methods begin by computing an all-to-all distance matrix between the taxa. This is typically the hamming distance between the DNA character sequences for each taxa though some other metrics have been used [8]. The taxa are then clustered using this distance information.

Stepwise maximum parsimony begins by shuffling the taxa into a random order. Each taxon is inserted in the most parsimonious position resulting in a complete tree.

1.1.1 Cartographic Projections

Trees can be considered as collections of bipartitions of taxa as every branch in a tree divides the taxa into two sets. A tree can be represented as a vector whose components all have the value 1 or 0, where these values indicate the presence or absence of the associated bipartition.

The dimensionality of tree space is $O(n!)$, with respect to the number of taxa. Directly representing trees in this space quickly becomes prohibitive. One method of mitigating this explosive dimensionality is through cartographic projections [14] which represents trees as a small set of inner products which can be computed in $O(n)$ time.

2 Approach

PTM is able to begin global searching after only $O(n \log n)$ steps using a global search space representation [14].

The PTM method is based on the idea that an unresolved tree is an approximation of all the resolutions of that tree.

During tree mixing, unresolved trees are chosen which cover new portions of tree space. As the partial trees are kept small, many of these exploratory searches can be accomplished in a small amount of time.
Figure 1: A brief overview of the PTM algorithm. In the first phase small disjoint maximum parsimony trees are built. In the final phase no division occurs so the trees grow in size until a tree containing all of the taxa is produced.

Figure 1 shows a graphical overview of the PTM process. First the taxa are divided into disjoint sets and the initial partial trees are built. Then the partial trees mix together, exploring the global tree space. Finally the partial trees are joined to build a fully resolved tree, which can then be passed on to the usual TBR-based search.


Results

In this section two types of results are considered. First, the work examines the effects of the parameters available to the user on the time taken and on the quality of the trees found. Second, using default settings for these parameters the method is compared with other phylogenetic search programs. PTM followed by a standard TBR search is shown to find better trees than competing methods.

2.1 The Effects of Partial Tree Size

During tree mixing two parameters, a minimum and maximum, determine the range of allowable partial tree sizes. Figure 2 shows the effects of these parameters on search times. After the TBR refinement there is very little variation in the score of the final tree.

2.2 Comparison with Existing Phylogenetic Search Programs

Two comparisons are made with PTM and other existing search programs. The first is a comparison of the quality of the initial starting tree generated. The results of comparing PTM with stepwise maximum parsimony (PAUP*[15]) using both a single
starting tree and multiple starting trees is shown in Figure 3. PTM takes significantly more time than stepwise maximum parsimony. However, PTM also yields higher quality trees.

![Figure 3: A comparison of search results between PTM and stepwise maximum parsimony on several datasets. Note that in every case PTM found more parsimonious trees, but in much more time. When stepwise maximum parsimony was used to find multiple starting trees (300), PTM still found more parsimonious trees.]

The second is a comparison of the quality of the final tree found. PTM is again compared with PAUP* but is also compared with TNT[5]. As the final step of PTM is a standard TBR search, the PTM starting tree was refined using the Parsimony Ratchet [10] and a sectorial search [4], both of these techniques are also available in TNT. Figure 4 contains this comparison. The final results from PTM for all of the data sets are superior to the final results found using a stepwise tree. Furthermore, with the exception of the smallest data set, these superior trees are found in less time. A trace of a typical result is shown in Figure 5, the figure shows a search through a set of 6722 taxa.

![Figure 4: A comparison of search results between PTM and PAUP*, and TNT on several datasets. Note that in every case PTM found a more parsimonious tree than PAUP*. In all but the smallest case this tree was found in less time. TNT finishes much faster than PTM, but finds less parsimonious trees.]

### References


Figure 5: A comparison of scores found over time between Partial Tree Mixing (PTM) and PAUP*\cite{15} (Stepwise Maximum Parsimony followed by TBR). Although PAUP* achieves better scores during the early phases of the search, PTM achieves significantly better results after 30000 seconds.


