Evolutionary Bioinformatics

ORIGINAL RESEARCH

A New Support Measure to Quantify the Impact of Local Optima in Phylogenetic Analyses

Grant Brammer[1], Seung-Jin Sul[2] and Tiffani L. Williams[1]

[1]Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843, USA.
[2]J. Craig Venter Institute, Rockville, MD 20850, USA. Corresponding author email: grb@cse.tamu.edu

Abstract: Phylogentic analyses are often incorrectly assumed to have stabilized to a single optimum. However, a set of trees from a phylogenetic analysis may contain multiple distinct local optima with each optimum providing different levels of support for each clade. For situations with multiple local optima, we propose *p*-support which is a clade support measure that shows the impact optima have on a final consensus tree. Our *p*-support measure is implemented in our PeakMapper software package. We study our approach on two published, large-scale biological tree collections. PeakMapper shows that each data set contains multiple local optima. *p*-support shows that both datasets contain clades in the majority consensus tree that are only supported by a subset of the local optima. Clades with low *p*-support are most likely to benefit from further investigation. These tools provide researchers with new information regarding phylogenetic analyses beyond what is provided by other support measures alone.

Keywords: phylogenetic, support measure, visualization

Evolutionary Bioinformatics 2011:7 159–170

doi: 10.4137/EBO.S7182

This article is available from http://www.la-press.com.

Introduction

Consensus trees are one of the most popular methods for summarizing a phylogenetic analysis. Oftentimes, these trees are annotated with values (eg, bootstrap replicate percentages) to show the support for each consensus branch (or clade). Support values, in general, indicate the level of corroboration for each region of the tree. Less corroborated regions are more likely to be overturned by subsequent data whereas highly corroborated clades are more robust to consideration of further data.[1] Since we cannot ensure that a phylogenetic analysis has converged to a single global optimum without exhaustively exploring tree space, tree sets from a phylogenetic analysis may contain trees from multiple regions of tree space. We define a peak as a set of good-scoring trees with similar tree topologies. In essence, the trees in each peak are similar to one another but contain significant difference to trees in other peaks. In this way, the peaks are an estimation of the local optima that were found by the phylogenetic search.

If trees from multiple peaks are represented in a set of trees, each clade in the consensus tree may not be highly supported across the peaks. Using clade frequency as the sole basis for a support measure can result in misleading conclusions regarding the stability of evolutionary relationships found by a phylogenetic analysis. Thus, we develop a new support measure called *p*-support that incorporates information about these distinct sets of trees in order to estimate the confidence levels of inferred relationships more robustly than traditional approaches.

Definition of *p*-support

We define the *p*-support of a clade as the percentage of *p* peaks with majority support for that clade. A *p*-support value of 100% means that a clade was supported by each peak whereas 0% implies that the clade was not strongly supported by any of the *p* peaks. *p*-support can be viewed as a measure of precision at the peak level much the same way that bootstrap and jackknife support are measures of precision at the character and taxa level. High *p*-support values signal that a clade is in high agreement across the peaks and therefore is less likely to be overturned by additional analysis. Similarly to other common support measures, *p*-support can be useful in identifying the areas of a tree that may benefit the most from additional data and analysis. In this way, support measures are a useful tool in illuminating new problems and hypotheses.[2]

The most critical feature to the *p*-support measure is the identification of the *p* peaks which are the input to the *p*-support calculation. We have developed the PeakMapper algorithm to determine how many distinct sets of trees are contained in a data set as well as which trees are contained in peak. While our technique uses clustering to identify the peaks among the trees, *p*-support is independent of our PeakMapper algorithm. Any method that identifies distinct sets of trees can be used with our *p*-support measure. For instance, if tree islands[3] were detected and labeled in a data set that information could be used to compute *p*-support. Our PeakMapper software identifies peaks in a tree collection and annotates majority and strict consensus trees with *p*-support values that can be viewed in standard tree viewing packages such as FigTree. Furthermore, our PeakMapper software is designed for analyzing large-scale tree collections (eg, tens of thousands of trees).

Comparison to common support measures

Bremer support, also known as the decay index, support index, or simply SI[4] measures how many steps from the most parsimonious trees it takes to lose a branch in the consensus of the near-most-parsimonious trees. A branch in one of the most parsimonious trees is strongly supported if it is also contained in the near-most-parsimonious trees. There are similarities in the intuition behind Bremer support and *p*-support. Both methods consider the prevalence of a clade across sets of trees. Where Bremer support creates subsets of trees by iteratively relaxing the threshold for near-most-parsimonious trees to be included into the consensus, *p*-support considers distinct sets of trees. Where Bremer support seeks to compute the the point at which the clade stops appearing in the consensus of the subset of trees, *p*-support computes the prevalence of each clade in each peak. In both cases highly supported clades are highly corroborated in the data set. While Bremer support is effective only in parsimony analyses, *p*-support is not limited by the method which the trees are computed. *p*-support can be used in any analysis which contains distinct sets of trees.

Bootstrap support[5] is computed by running a set of analyses with the input sequence alignment resampled such that some characters are include twice or more and others are not included at all. This simulates the effect that reweighting or revising the data might have on the output trees. Hence, the rate that a clade appears in the resulting trees is a measure of how robust a clade is to changes in the sequence alignment. Bootstrap support is similar to taxa jackknifing[6,7] which samples the taxa set to generate input data for a set of phylogenetic analyses. The resulting trees are used to compute a consensus tree to identify areas of disagreement among the trees. This is a measure of stability in regards to the deletion of taxa.

Each of these measures defines the support of a clade based on its stability under different methods of perturbation of the input sequence data. These methods are very complementary to *p*-support which is the only support measure of corroboration across peaks in the data set. In fact, there is no strict guarantee that the trees generated from a bootstrap or jackknife analysis fall into a peak themselves. Thus, measuring the *p*-support of the trees resulting from a bootstrap or jackknife analysis may provide further information.

Summary of experimental results

Using our PeakMapper approach, we analyze two published Bayesian studies on 150 taxa of desert algae and green plants[8] and 567 taxa of angiosperms[9] data sets. The 150 taxa data set consists of 20,000 trees from two runs of the MrBayes phylogenetic heuristic. The 567 taxa data set contained 33,306 trees from 12 Bayesian runs. Both of these tree collections have high majority consensus resolution rates. Our approach shows that both tree sets contain multiple peaks—there are two and six peaks found for the 150 and 567 taxa data sets, respectively. Hence, high consensus resolution rates do not exclude the possibility of a tree set containing multiple peaks. These data sets present two interesting cases: the number of trees in the peak in the 150 taxa data set are of equal size while they are disproportional in the 567 taxa data set. These cases show how the distribution of trees across peaks can impact the resulting majority consensus tree and also show how *p*-support can provide previously unavailable information about the distribution of the clades. We show that the 150 taxa data set contains three clades that appear in the majority consensus tree but are only supported by one of the two peaks. The 567 taxa data set contains seven clades in the majority consensus tree with supported by only three of the six peaks and a clade in the majority consensus tree supported by only two of the six peaks.

Overall, our work presents systematists with a new measure called *p*-support for quantifying the robustness of inferred relationships in an evolutionary tree. We hope that *p*-support can provide researchers and the community at large with more information about the results of phylogenetic analyses—especially in regards to which regions of the tree may benefit most from further investigation.

Material and Methods

Tree collections

The biological trees used in this study were obtained from two recent Bayesian analysis, which we describe below. All trees in our collections are unique. Table 1 provides statistics concerning our tree collections.

- *Data set #1:* 20,000 trees obtained from a Bayesian analysis of an alignment of 150 taxa (23 desert taxa and 127 others from freshwater, marine, ands oil habitats) with 1,651 aligned sites.[8] Two independent runs consisting of 25 million generations (trees were

Table 1. Detailed information for the two biological datasets studied in this paper. For each data set, we list the number of trees, resolution rates of the majority and strict consensus trees, and run labels of the trees.

	Size		Resolution rate		Runs (%)
	Taxa	Trees	Majority	Strict	
Data set #1	567	33,306	92.6%	51.8%	*R*0 (9.4%) *R*1 (9.4%) *R*2 (9.4%) *R*3 (9.4%) *R*4 (8.9%) *R*5 (8.9%) *R*6 (8.9%) *R*7 (8.9%) *R*8 (6.5%) *R*9 (6.5%) *R*10 (6.5%) *R*11 (6.5%)
Data set #2	150	20,000	85.7%	34.0%	R0 (50%), R1 (50%)

sampled every 1,000 generations) were performed using the GTR+I+Γ model in MrBayes with four independent chains. The authors constructed a majority consensus tree in their study using the 20,000 trees from the last 10 million generations from each of the two runs. The resolution rates of the majority and strict consensus trees are 85.7% and 34.0%, respectively. The total number of clades across the 20,000 trees is 2,940,000, where 1,168 of them are unique.

- *Data set #2:* 33,306 trees obtained from an analysis of a three-gene, 567 taxa (560 angiosperms, seven outgroups) dataset with 4,621 aligned characters, which is one of the largest Bayesian analysis done to date.[9] Twelve runs, with four chains each, using the GTR+I+Γ model in MrBayes ran for at least 10 million generations. Trees were sampled every 1,000 generations. The authors discuss the difficulties with combining trees from multiple runs. To obtain our collection of 33,306 trees, we discard the trees from the first 8 million generations. The resolution rates of the majority and strict consensus trees are 92.6% and 51.8%, respectively. The total number of clades across the 33,306 trees is 18,784,584, where 2,444 of them are unique.

Our PeakMapper approach

In our PeakMapper approach, we use clustering techniques to determine the p peaks found in a set of t trees, which can come from a Bayesian or bootstrap analysis for example. These p peaks are then used to compute and visualize p-support across a collection of trees. The presence of p peaks is based on placing the trees into k different distinct clusters (or partitions). For a particular cluster C_i, the trees within that cluster represent a peak. Clustering has been long used as part of phylogenetic analysis but most commonly as a method to build trees. Distance-based methods such as UPGMA[10] are based on hierarchical clustering of the data. In addition to us, Stockham et al[11] have used clustering as a post-processing tool. The authors suggest presenting multiple consensus trees which more accurately reflect the distribution of clades in a set of trees as compared to a single consensus tree. They use clustering to select the optimum grouping of trees to best fit the distribution of clades in the original data tree set. However, our use of clustering differs from their approach in that we are interested in computing the number of peaks in the dataset in order to compute p-support. Furthermore, we focus on developing methods suitable for fast analysis of tens of thousands of trees.

Step 1: Creating the clade matrix.—Initially, the Newick-formatted trees are represented by a feature matrix for further processing. The feature matrix is a $t \times c$ binary matrix, where each of the t trees in the data set is represented as a row in the matrix and each column represents a unique feature (or clade). In other words, there are c unique clades for a set of t trees. Hence, the two tree collections can be represented by a $20,000 \times 1,168$ and a $33,306 \times 2,444$ clade matrix, respectively. The state of clade i for tree j is contained in cell (i, j). We mark the state of each clade for each tree as either present or absent, which we represent as '1' and '0' respectively. For fully resolved trees of n taxa, there are n–3 clades present in each tree. Clade matrices are created using a variation of the HashCS[12] algorithm, which is based on using a hash table to identify the c unique clades across the t trees over n taxa.

Step 2: Clustering the clade matrix.—Given a $t \times c$ clade matrix, we use CLUTO,[13] a freely-available, high-performance software package for clustering large high-dimensional data. CLUTO was chosen for it's ability to cluster very large data sets efficiently. CLUTO has been successfully used to cluster multiple types of data including text documents[14] and gene expression data.[15] As input, CLUTO takes either a distance matrix or a set of vectors and the number of desired clusters. The user can also set CLUTO to use different methods of clustering such as agglomerative clustering or by repeated bisections and optimize on different criteria to maximize internal similarity or minimize external differences. We have chosen to use the default settings which clusters our clade matrix (represented as vectors) by repeated bisection, computes the distance between the vectors as the cosine, and maximized the internal similarity as the fitness function. For the large data sets we analysis in this paper, CLUTO has a ten-fold increase in performance over R. However, in the future for smaller data sets, we plan to incorporate clustering analyses from these packages into our PeakMapper software.

Since the true number of clusters represented by the data is unknown, we tested a range of clusters, k, varying from 2 to 24 and evaluated the results. The best k that fits the data (ie, clade matrix) represents the number of peaks p found by the phylogenetic analysis. For clustering approaches such as CLUTO, we cannot generate a fitness score for situations where $k = 1$, which represents a single cluster. This is because fitness is measured as a ratio of internal and external similarity between clusters. With a single cluster, there is no external similarity that can be computed. Instead, we check whether any of the clusterings fit the data. If not, we reject the hypothesis that there are multiple peaks since a single peak best represents the data.

We have chosen to examine k values of 2 through 24. Assuming each of the twelve runs of our largest data set converged to completely different peaks we would require a k value of 12 to handle this case. We have tested k values up to 24 which is double the maximum number of runs in our largest data set and used k values of 2 through 24 for each data set we studied as a measure of consistency. This range was shown to be more than sufficient for our data set.

We have also applied clustering methods to the all-to-all Robinson-Foulds (RF) distance matrices,[16] a popular distance measure to compare phylogenetic trees,[11,17] to each of our two data sets. Instead of a $t \times c$ feature matrix, using distance matrices require a $t \times t$ matrix. Our two data sets required a $20{,}000 \times 20{,}000$ and $33{,}306 \times 33{,}306$ RF matrices, which we computed using HashRF.[18] Both methods of clustering produced similar results, however, the clade matrix is a smaller representation of the data than the corresponding distance matrix and results in a significantly shorter clustering time in CLUTO. By using the clade matrix representation instead of the RF matrix representation we decreased our running time by over an order of magnitude.

Step 3: Determining the number of peaks, p.—Once we have computed the clusterings of our data with k ranging from from 2 to 24, we select the number of clusters that maximizes the similarity among the trees of interest while minimizing the total number of clusters. The optimal clustering represents the number of peaks, p, found in the data. CLUTO's internal similarity measures (*ISim* and *ESim*) are our primary measure for determining the appropriate k value. $ISim_{ave}$ (intra-cluster similarity) is the average similarity between trees of each cluster a value which we would like maximized. On the other hand, $ESim_{ave}$ (inter-cluster similarity) is the average similarity of the trees of each cluster to the tress outside the cluster a value which we would like minimized. We use the $ISim_{ave}/ESim_{ave}$ ratio as the basic measure of quality for a clustering of a clade matrix. In selecting a P value, we use the elbow criteria which advocates choosing P so that adding additional peaks does not add sufficient information. In other words, we choose a P value such that increasing it does not increase the $ISim_{ave}/ESim_{ave}$ ratio significantly.

Step 4: Visualizing the P peaks.—We use multi-dimensional-scaling (MDS) plots to see the relative positions of the trees within their peaks. We used a freely available software package called High-Throughput Multi-dimensional Scaling (HiT-MDS-2)[19] for reducing the dimensionality of large data sets. HiT-MDS-2 allows the user to input data in high dimensional space and map it to a low dimensional space. HiT-MDS-2 took our $t \times c$ clade matrix and reduced it to a $t \times 2$ matrix in order to visualize the result as a scatter plot. Remember that the c clades represent features of the t trees. So, in MDS, the c features of each tree are reduced to 2 features (f_1 and f_2) by Hit-MDS-2. In our scatter plots, we plot f_1 on the x-axis and f_2 on the y-axis. Given that our data sets consist of 20,000 and 33,306 trees, plotting such a large number of points would result in an unreadable scatter plot. To enable useful inferences from our visualizations of the clustered data, we visualize the data based on a 10% sample (without replacement) of the collection of t trees. As a result, in our MDS visualizations, we plot a $s \times 2$ matrix, where $s = 0.1 \cdot t$. Hence, for our MDS plots, we have two tree samples of 2,000 and 3,331 trees consisting of 150 taxa and 567 taxa, respectively. Multiple samples and MDS runs have shown the plots are good representations of the data.

Aside from this study, other systematists have used MDS in phylogenetics.[5] However, they do not use MDS it in the context of analyzing the distinct set of tree in tree space as it relates to clade support.

Step 5: Computing and visualizing p-support. Once the p optima have been identified, we can compute the p-support value for each clade in our data set. In order to visualize the p-support values

for all c unique clades in the data set, we developed the p-map. Our visualization technique compares the range of p-support values on the y-axis to the percentage of total support on the x-axis. To increase readability, jitter is applied along the y-axis only, allowing overlapping points to form into lanes. The position on the x-axis is absolute meaning that any point on the right side of the line marking 50% support is a clade that would appear in the majority consensus trees. Each point in a p-map is to represent the standard deviation in support for a clade between clusters. Points shaded in blue are equally supported across clusters were as points shaded in red are highly supported in some clusters and barely supported in others. These plots are used later in our analysis as shown in Figures 3 and 5.

In addition to p-maps, our PeakMapper package takes as input a tree and annotates its branches with their respective p-support values. Since the Bayesian analyses studied in this paper were summarized as consensus trees, we take their strict and majority representations and annotate with p-support.

Results and Discussion

Data set #1: 20,000 trees over 150 taxa

Figure 1 suggests that the optimal number of clusters is at $k = 2$. Thus, $P = 2$ since the data set contains two peaks, $P0$ and $P1$. What is the composition of the two peaks in this data set? Table 2 and Figure 2 show that both peaks are composed of half of the 20,000 total trees, and each peak is composed of trees from a single run. Peak $P0$ is composed of trees from run $R1$ from the Bayesian analysis. Peak $P1$ consists of trees from run $R0$. There is no overlap of trees between the two peaks as each run is contained within a single peak.

Summarizing this collection of 20,000 trees as a consensus tree without acknowledging the presence of the two peaks ignores the distinct competing hypotheses that exist in this data set. We represent the influence of these hypotheses by annotating the resulting consensus tree with the p-support values of each branch in the consensus. For example, in a majority consensus tree, our annotation illuminates majority clades that are supported by a subset of the peaks. Figure 3 shows that there are three clades that would appear in the majority consensus tree but are only supported by one of the two peaks. These clades have a high standard deviation of p-support meaning that they are supported by one peak much more than the other. Since there is disagreement among the peaks, these clades more likely to be overturned by further runs of the search heuristic than those agreed on by all of the peaks.

Data set #2: 33,306 trees over 567 taxa

Even though the 567 taxa data set is more demanding in terms of having more trees and runs than the 150 taxa data set, we approach it with the same process to compute its p-support values. Examining Figure 6(a) with the elbow criteria in mind suggests a range

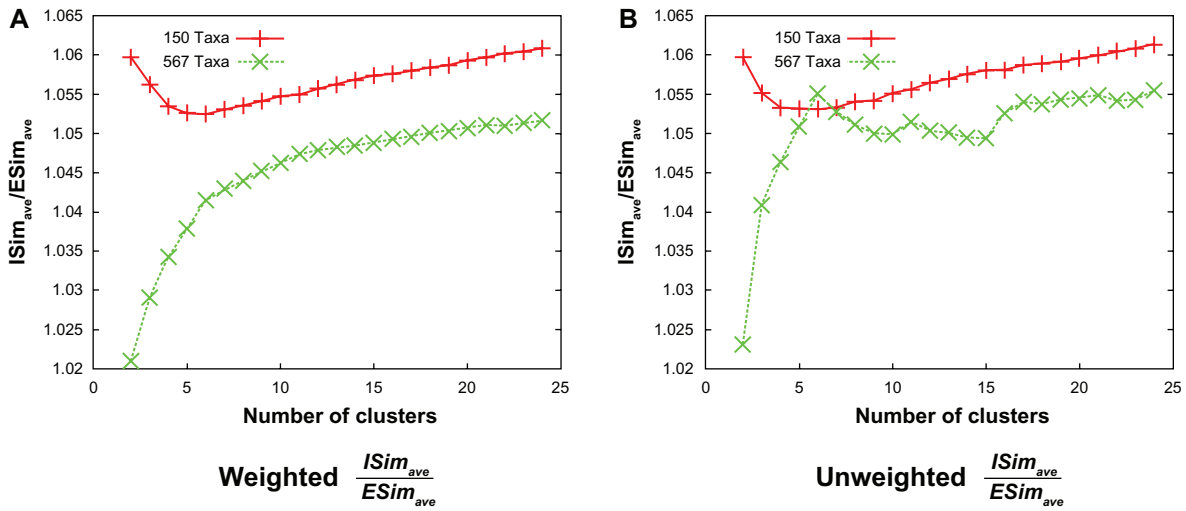


Figure 1. Selecting the appropriate number of clusters, k. Larger $ISim_{ave}/ESim_{ave}$ values are preferred since they indicate a better clustering of the data. The resolution rate reported in this table is a measure of how resolved the consensus tree is. If the consensus tree was completely binary we would have a value of 100% whereas if it was a star phylogeny it would have a value of 0%.

Table 2. Detailed information for the two peaks found for the 150 taxa trees. For each peak, we list the number of trees, resolution rates of the majority and strict consensus trees, and run labels of the trees.

Peak	Size		Resolution rate		Runs (%)
	Trees	%	Majority	Strict	
*P*0	10,000	50%	90.5%	34.7%	R1 (100%)
*P*1	10,000	50%	89.1%	37.4%	R0 (100%)

of potentially useful k values from $k = 5$ to $k = 8$. However Figure 1(b) suggests that $k = 6$ is the optimal clustering for the data set. Since $k = 6$ is potential solution in both of the analysis of the $ISim_{ave}/ESim_{ave}$ ratio and the clear peak in unweighted analysis we will select the $P = 6$ as the number of peaks for this phylogenetic analysis.

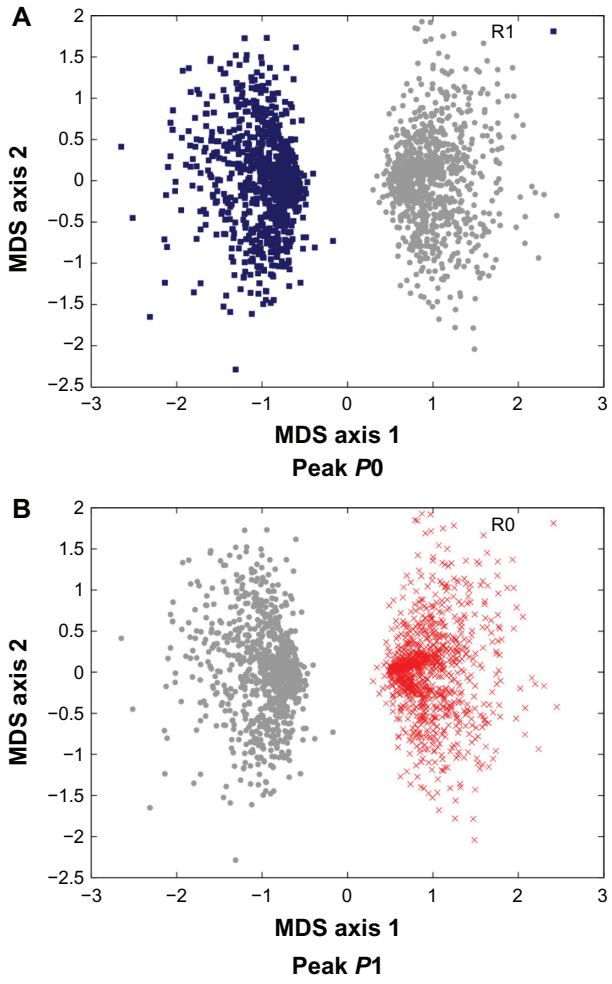


Figure 2. *150 taxa, P* = 2: Each plot shows a single peak with the trees in the peak colored to represent the Bayesian runs they came from. The MDS values are computed as a Euclidean embedding of the data points. The *r* value for this MDS analysis is 0.79, where *r* = *correlation* (*original, reconstruction*).

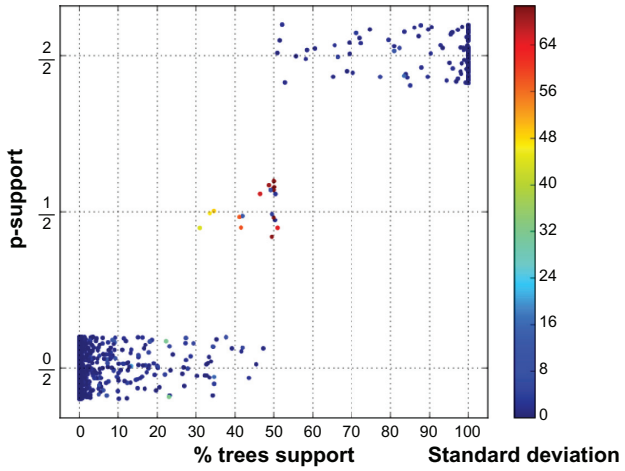


Figure 3. *p*-map. *p*-support values plotted against the percent of trees containing the clade for the 150 taxa data set with a *P* value of 2. Points greater than 50% on the *x*-axis would appear in a majority consensus tree. The points are shaded to reflect the standard deviation in the support values from the different peaks. Dark blue points mean the clade was equally supported among the peaks whereas red points mean there was a large difference in the amount of support for a clade among the peaks.

We now explore the composition of the peak and ask "What are the traits of the six peaks in the data set?" Table 3 provides statistics regarding the composition of the six peaks for the 567 taxa trees. The information in Table 3 is represented visually in Figure 4. The trees in runs *R*0, *R*6, and *R*11 are each fully contained in a single peak, *P*3, *P*1, and *P*2 respectively. These trees are placed in these peaks alone with no other trees from other runs. Hence we can say that the trees from each run are contained in their own peaks with no mixing or overlap with trees from other runs. Hence, these runs settled into distinct areas of tree space.

Alternatively, peaks *P*4 and *P*5 contain trees from multiple runs. Peak *P*4 contains trees from runs *R*1, *R*5, *R*7, and *R*8. These runs appear wholly in *P*4, and in no other peak. We can say that these four runs have stabilized to the same peak but do not overlap with the runs in any other peaks. Peak *P*5 is very similar to peak *P*4 in that it is mainly composed of runs that wholly are contained with in the peak. Runs *R*2, *R*4, *R*9 and *R*10 are all contained in this peak. The one exception is the placement of run *R*3. This run is split between two peaks. About half the run appears in peak *P*4 with four other runs and the other half of run *R*3 appears in peak *P*0 by itself. This shows that run *R*3 is split between two different peaks. It is the only run in either of our data sets to exhibit this behavior. This

Table 3. Detailed information for the six peak found by our PeakMapper approach on the 567 taxa trees. For each peak, we list the number of trees, resolution rates of the majority and strict consensus trees, and run labels of the trees.

Peak	Size		Resolution rate		Runs (%)
	Trees	%	Majority	Strict	
*P*0	1,537	4.6%	94.2%	68.4%	*R*3 (100.0%)
*P*1	2,986	9.0%	95.6%	64.7%	*R*6 (100.0%)
*P*2	2,164	6.5%	95.6%	66.8%	*R*11 (100.0%)
*P*3	3,177	9.5%	95.4%	65.1%	*R*0 (100.0%)
*P*4	11,312	34.0%	94.0%	61.0%	*R*1 (28.1%) *R*5 (26.4%) *R*7 (26.4%) *R*8 (19.1%)
*P*5	12,130	36.4%	92.4%	58.3%	*R*2 (26.2%) *R*3 (13.5%) *R*4 (24.6%) *R*9 (17.8%) *R*10 (17.8%)

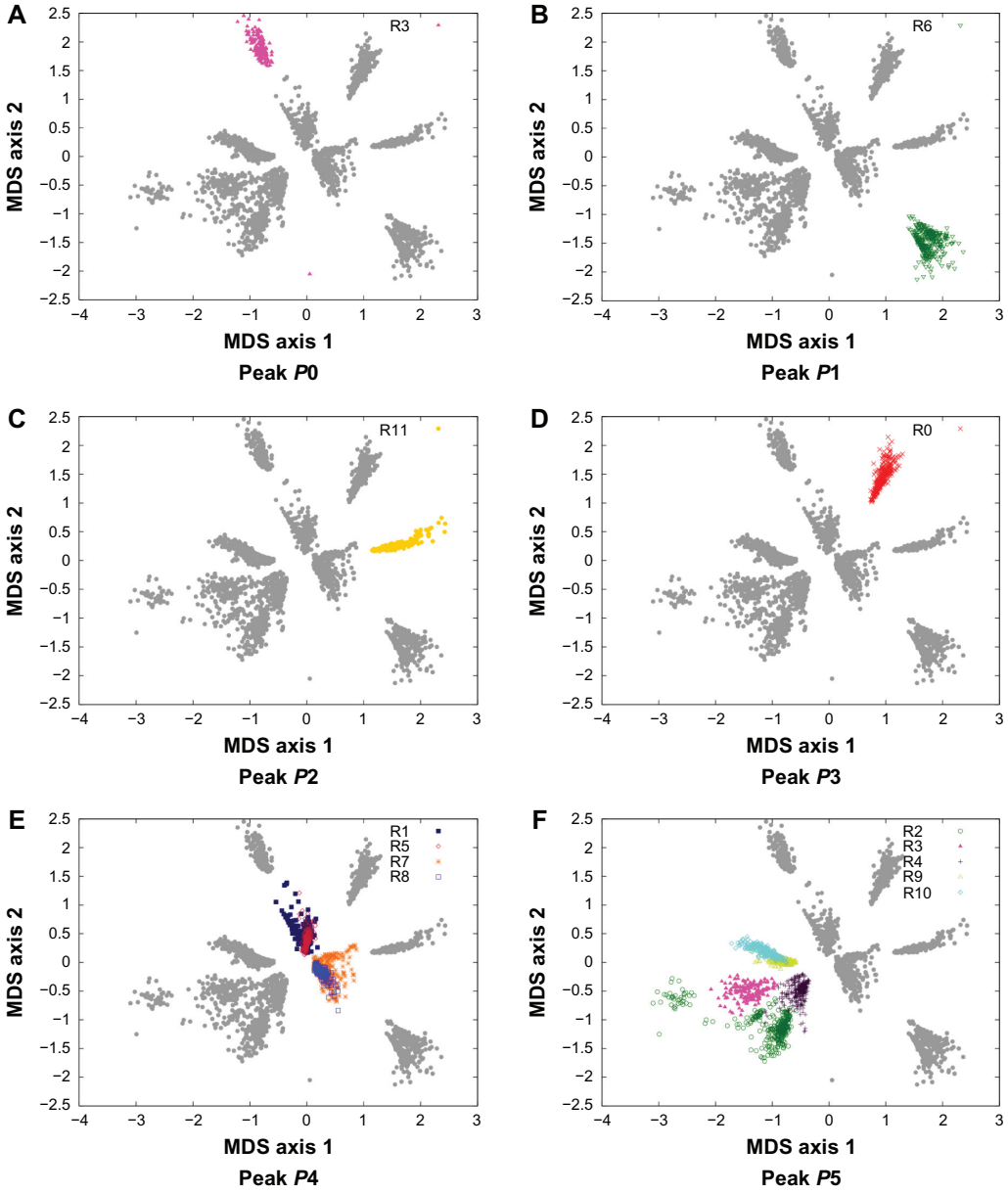


Figure 4. *567 taxa, P* = 6: Each plot shows a single peak with the trees in the peak colored to represent the Bayesian runs they came from. The MDS values are computed as a Euclidean embedding of the data points. The *r* value for this MDS analysis is 0.84, where *r* = *correlation* (*original, reconstruction*).

behavior may be the result of the phylogenetic search settling into one peak for the first part of the search only to find a better peak as the search progressed.

In this data set our analysis shows six peaks each distinct from one another. The only mixing in terms of runs appearing in multiple peaks occurs with *R*3 appearing in two peaks. Knowledge of these trends in the data has the potential to inform the process of summarizing of the trees For instance since the behavior exhibited by run *R*3 could be explained that it found one peak early in the search only to abandon it later for another peak, it may be useful to remove the section of *R*3 contained in the lesser scoring of the two peaks from the summarization. This same idea could be applied to whole sets of peaks. Given that we have six peaks and know which trees make up each peak it becomes possible to select the peak with the best overall likelihood scores and set aside the trees representing the less likely peaks.

Figure 5 shows the *p*-support values plotted against the percentage of trees containing each clade. Notice that there are a number of clades which would appear in a majority consensus tree but are supported by only a subset of the peaks found in the search. For instance, there are seven clades which appear in the majority consensus tree but, they are only supported by 50% of the peaks. There is even a clade which is only supported by 2 of the 6 peaks yet it has over 50% majority support and therefore appears on the majority consensus tree.

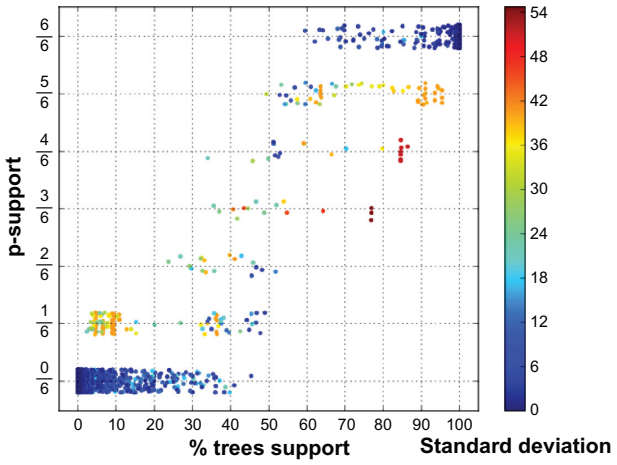


Figure 5. *p*-support values plotted against the percent of trees containing the clade for the 567 taxa data set with a *P* value of 6. Points greater than 50% on the *x*-axis would appear in a majority consensus tree. The points are shaded to reflect the standard deviation in the support values from the different peaks. Dark blue points mean the clade was equally supported among the peaks where as red points mean there was a large difference in the amount of support for a clade among the peaks.

Due to size disparity between peaks there are some clades supported by only half the peaks but are still able to achieve over 70% majority support. These clades are also interesting due to the high standard deviation in *p*-support. Some peaks heavily support those clades while other peaks barely support them if at all. The clades with low *p*-support are the ones most likely to be effected by further analysis or new data. Clades that are common to all six peaks are likely to appear in future runs of the heuristic search algorithm while clades supported by a subset of peaks are less likely to be present in a future run. To this end it may be appropriate to collapse clades with low *p*-support.

Detecting single peak tree collections

We have shown the effectiveness of our algorithm in detecting when a data set has stabilized to multiple peaks instead of a single peak. Since neither data set we analyzed contained only a single peak, we chose to create such a data set in order show how our algorithm would perform given this case. To represent this case we used a single run from our 150 taxa data set to create a 10,000 tree data set with a single peak.

We begin with by clustering our data set using $k = 2$ through $k = 24$. We then examine the $ISim_{ave}/ESim_{ave}$ ratio in Figure 6. Applying the elbow criteria can be a little tricky in this case. There is no obvious place where the gains in the $ISim_{ave}/ESim_{ave}$ ratio taper off. There is a fairly steady increase in the $ISim_{ave}/ESim_{ave}$ ratio as the k values increase. This is a strong sign that the most appropriate clustering is actually $k = 1$. As we increase k to its maximum value of n (number of trees) we expect the $ISim_{ave}/ESim_{ave}$ ratio to rise. We are looking for a peak in these values before it it becomes a continual rise. Since this isn't found, we assume a k value of one and therefore a P value of one.

Computational time

Since the two data sets studied here were obtained by using MrBayes, we compare the time of our PeakMapper algorithm to analyze the trees in the data sets with the time required by MrBayes to produce consensus trees. Our purpose is to show that for the time a systematist is willing to spend obtaining a consensus tree, our PeakMapper algorithm can produce a richer and more detailed perspective of the quality of the underlying data set by computing and visualizing *p*-support.

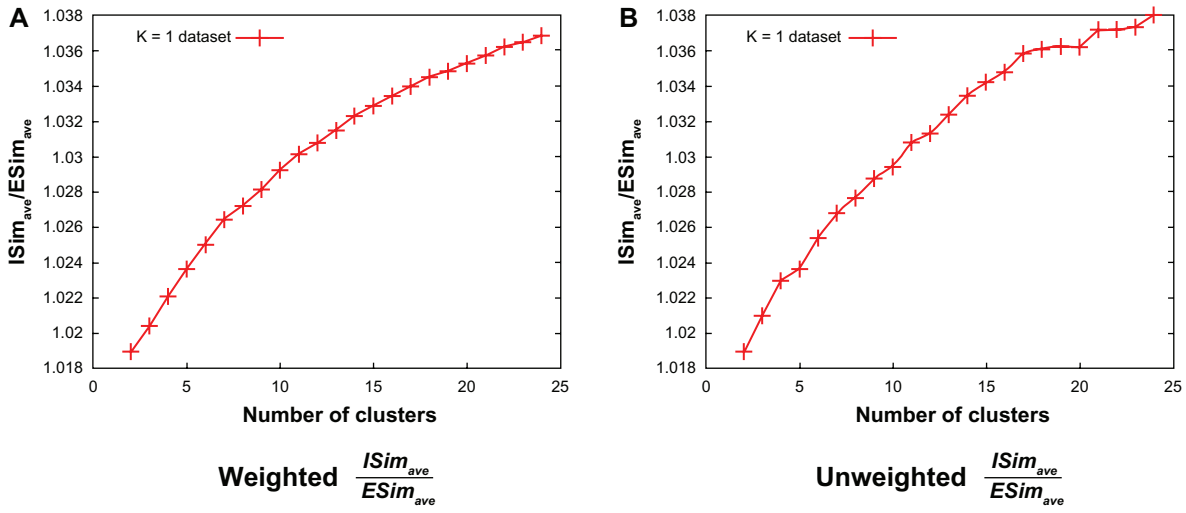
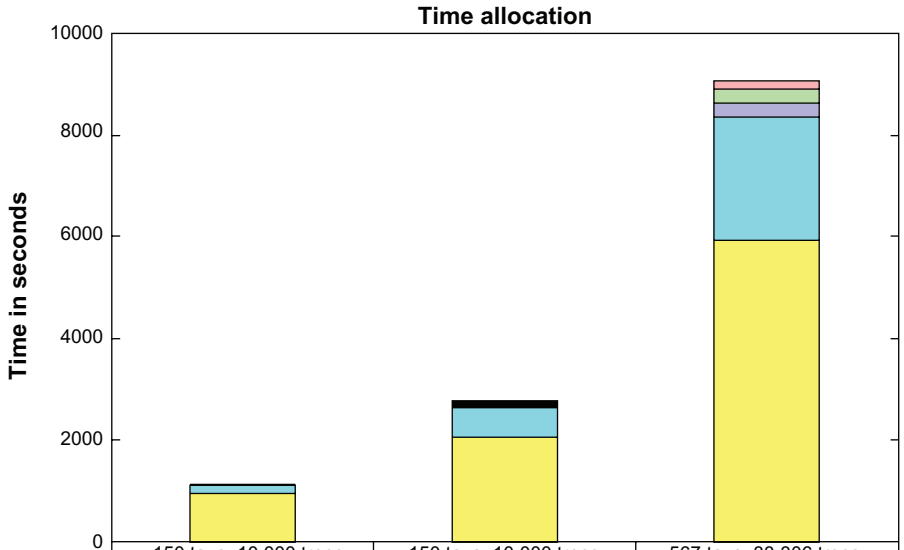


Figure 6. Selecting the appropriate number of clusters, *k*. Larger $ISim_{ave}/ESim_{ave}$ values are preferred since they indicate a better clustering of the data.

Figure 7 shows that our 567 taxa tree collection required 2 hours 22 minutes and 20 seconds and a maximum of 2.7 GB of memory to complete the creation of the clade matrix, MDS analysis, clusterings for the range of k values where $2 \leq k \leq 24$, and plotting the results. We also computed the strict and majority consensus trees for each of the 6 clusters and for the data set as a whole for a total of 14 consensus trees. This computation took a total of 5 minutes and 42 seconds. Creating a tree annotated with *p*-support took an additional 30 seconds. Table 4 shows the total computational time for the analysis presented in the paper is 2 hours, 28 minutes, and 32 seconds. This is compared to computing the strict and majority consensus tree using MrBayes which took 4 hours 14 minutes and 31 seconds. Our analysis produced much more information including, an analysis of the peaks in the data set, in less time than it took to create two consensus trees with MrBayes.



	150 taxa, 10,000 trees	150 taxa, 10,000 trees	567 taxa, 33,306 trees
Clade matrix	11	24	145
Misc	8	27	281
Consensus trees	12	50	282
MDS	151	595	2432
Clustering	952	2062	5922

Figure 7. This plot shows the time in seconds that PeakMapper spent to compute the different components of the *p*-support analysis.Clustering the data required the most time, followed by MDS, etc. The Misc category captures the running time of operations that cannot be categorized by creating a clade matrix, computing a consensus tree, computing MDS, or clustering the trees.

Table 4. Time to compute the a single consensus tree of the whole data set using the `sumt` command in MrBayes. This time only includes the creation of the consensus tree and not the phylogenetic search or any other MrBayes functionality.

Data set		MrBayes `sumt` command		PeakMapper
Taxa	Trees	Strict consensus	Majority consensus	Total time
150	10,000	2 m:42 s	2 m:43 s	18 m:54 s
150	20,000	5 m:32 s	5 m:30 s	46 m:5 s
567	33,306	2 h:07 m:17 s	2 h:07 m:14 s	2 h:28 m:32 s

For our single peak data set with 150 taxa and 10,000 trees, we completed our analysis in under 19 minutes. This included the creation of the clade matrix, clustering, MDS and computing consensus trees. Since there was only a single peak we halted our analysis after the selection of $k = 1$ as the best k. Hence no time was spent annotating a consensus tree with p-support values. The whole process including generation of the clade matrix, MDS including clustering and plotting for each k value took just over 45 minutes and under 2GB of memory. We then selected $k = 2$ as the best k. To explore the those two peaks we computed the strict and majority consensus trees for each cluster as well as for the whole data set took an additional 50 seconds. Creating a tree annotated with p-support took an additional 7 seconds. Table 4 shows the total computational time for the whole analysis which is 46 minutes and 5 seconds. As a comparison, summarizing the same data set using MrBayes' `sumt` command for both the strict and majority consensus takes a total of 11 minutes and 2 seconds. PeakMapper computes both of these consensus trees but also the strict and majority consensus trees for each of the 2 clusters, for a total of 6 consensus trees in only 50 seconds.

Figure 7 shows that our 567 taxa tree collection required 2 hours 22 minutes and 20 seconds and a maximum of 2.7 GB of memory to complete the creation of the clade matrix, MDS analysis, clusterings for the range of k values where $2 \leq k \leq 24$, and plotting the results. We also computed the strict and majority consensus trees for each of the 6 clusters and for the data set as a whole for a total of 14 consensus trees. This computation took a total of 5 minutes and 42 seconds. Creating a tree annotated with p-support took an additional 30 seconds. Table 4 shows the total computational time for the analysis presented in the paper is 2 hours 28 minutes 32 seconds. This is compared to computing the strict and majority consensus tree using MrBayes which took 4 hours 14 minutes and 31 seconds. Our analysis produced a much more information including an analysis of the peaks in the data set in less time than it took to create two consensus trees with MrBayes.

Conclusions

While p-support introduces new ideas to produce a novel clade support measure, it is intended to be compatible with the methods current used in phylogenetic analyses. Our measure neither changes nor replaces any of the ways that trees are currently produced or summarized. Furthermore, p-support is indifferent to the method on which the trees were gathered allowing it to be used with parsimony, likelihood, Bayesian or even to examine a combination of methods. p-support is intended to be an additional tool to be used along side existing measures. The peak detection methods described in this paper may be used to verify that there is only a single peak present. In other cases where the peaks are already known or computed by a different process (tree islands), p-support may be of value while the peak detection phase of the algorithm can be skipped.

It has been assumed that when multiple Bayesian analyses converge that there is a single peak present in the data set. We have shown that this is an assumption that should be investigated further. Not only is it possible for multiple peaks to be present it is possible for a single run of a search algorithm to contain multiple peaks. For data sets that contain multiple peaks, we provide methods to visualize and report their presence and impact on the final tree at a clade level. p-support allows the impact of peaks to be quantified on the reported tree and acts a measure of clade stability across multiple heuristic searches. p-support identifies clades which may most benefit from further analysis. Areas with low p-support could be improved potentially through further analysis with increased sequence data or by tuning the parameters

of the search algorithm. In sum, we hope that our work shows the importance of developing new and novel data analysis tools for understanding phylogenetic tree sets and the analyses that produced them.

Funding

This work was supported by the National Science Foundation under grants DEB-0629849, IIS-0713618, and IIS-1018785.

Acknowledgements

We would like to thank Matthew Gitzendanner, Paul Lewis, and David Soltis for providing us with the tree collections used in this paper. Moreover, this publication is based in part on work supported by Award No. KUS-C1-016-04, made by King Abdullah University of Science and Technology (KAUST).

Disclosures

Author(s) have provided signed confirmations to the publisher of their compliance with all applicable legal and ethical obligations in respect to declaration of conflicts of interest, funding, authorship and contributorship, and compliance with ethical requirements in respect to treatment of human and animal test subjects. If this article contains identifiable human subject(s) author(s) were required to supply signed patient consent prior to publication. Author(s) have confirmed that the published article is unique and not under consideration nor published by any other publication and that they have consent to reproduce any copyrighted material. The peer reviewers declared no conflicts of interest.

References

1. Egan M. Support versus corroboration. *Journal of Biomedical Informatics*. 2006;39:72–85.
2. Grant T, Kluge AG. Clade support measures and their adequacy. *Cladistics*. 2008;24:1051–64.
3. Maddison D. The discovery and importance of multiple islands of most parsimonous trees. *Syst Bio*. 1991;42:200–10.
4. Bremer K. Branch support and tree stability. *Cladistics*. 1994;10:295–304.
5. Felsenstein J. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*. 1985;39:783–91.
6. Lapointe FJ, Kirsch JAW, Bleiweiss R. Jackknifing of weighted trees: Validation of phylogenies reconstructed from distance matrices. *Molecular Phylogenetics and Evolution*. 1994;3:256–67.
7. Quenouille MH. Notes on bias in estimation. *Biometrika*. 1956;43:353–60.
8. Lewis LA, Lewis PO. Unearthing the molecular phylodiversity of desert soil green algae (chlorophyta). *Syst Bio*. 2005;54:936–47.
9. Soltis DE, Gitzendanner MA, Soltis PS. A 567-taxon data set for angiosperms: The challenges posed by bayesian analyses of large data sets. *Int J Plant Sci*. 2007;168:137–57.
10. Sokal RR, Michener CD. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*. 1958;38:1409–38.
11. Stockham C, Wang LS, Warnow T. Statistically based postprocessing of phylogenetic analysis by clustering. Pages 285–293 *in* Proceedings of 10th Int'l Conf. on Intelligent Systems for Molecular Biology (ISMB'02). 2002.
12. Sul SJ, Williams TL. An experimental analysis of consensus tree algorithms for large-scale tree collections. Pages 100–111 *in* ISBRA '09: Proceedings of the 5th International Symposium on Bioinformatics Research and Applications Springer-Verlag, Berlin, Heidelberg. 2009.
13. Karypis G. CLUTO–software for clustering high-dimensional datasets. Internet Website, last accessed, December 2010. Available from http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview. 2010.
14. Tagarelli A, Karypis G. 2008. A segment-based approach to clustering multi-topic documents. *in* Text Mining Workshop, SIAM Datamining Conference, 2008.
15. Zhao Y, Karypis G. Clustering in life sciences. Pages 183–218 *in* Functional Genomics (Brownstein MJ, Khodursky AB, eds.) vol. 224 of *Methods in Molecular Biology*. Humana Press 10.1385/1-59259-364-X:183. 2003.
16. Robinson DF, Foulds LR. Comparison of phylogenetic trees. *Mathematical Biosciences*. 1981;53:131–47.
17. Hillis DM, Heath TA, John KS. Analysis and visualization of tree space. *Syst. Biol*. 2005;54:471–82.
18. Sul SJ, Williams TL. An experimental analysis of robinson-foulds distance matrix algorithms. Pages 793–804 *in* European Symposium of Algorithms (ESA'08) vol. 5193 of *Lecture Notes in Computer Science* Springer-Verlag. 2008.
19. Strickert M, Teichmann S, Sreenivasulu N, Seiffert U. High-throughput multi-dimensional scaling (Hit-MDS) for cDNA-array expression data. Pages 625–634 *in* Artificial Neural Networks: Biological Inspirations – ICANN 2005 (W. Duch, J. Kacprzyk, E. Oja, and S. Zadrozny, eds.) vol. 3696 of *Lecture Notes in Computer Science*. Springer-Verlag, Heidelberg. 2005.

Supplementary Material

Our PeakMapper software can be found at http://faculty.cse.tamu.edu/tlw/evobio11.