This is a preprint of an article published in Journal of Computational Biology

MASTtreedist: Visualization of Tree Space based on Maximum Agreement Subtree

Hong Huang[*1] and Yongji Li[2]

[1]School of Information, University of South Florida, Tampa, FL, 33620

[2]Department of Computer Science, Sun Yetsen University, Guangzhou, China, 510275

Email: Hong Huang[*] - honghuang@usf.edu; Yongji Li- dragonlyj@gmail.com;

[*]Correspondent author.

Abstract

Phylogenetic tree construction process might produce many candidate trees as the "best estimates". As the number of the constructed phylogenetic trees grows, the need for efficiently compare their topological or physical structures arises. One of the tree comparisons software tools, the Mesquite's Tree Set Viz module allows the rapid and efficient visualization of the tree comparison distances using Multidimensional Scaling (MDS). Tree-distance measures such as Robinson-Foulds (RF) for the topological distance among different trees have been implemented in Tree Set Viz. New and sophisticate measures such as Maximum Agreement Subtree (MAST) could be continuously built upon Tree Set Viz. It could detect the common substructures among trees, and provide more precise information on the similarity of the trees, but it is NP-hard, and difficult to implement. In this paper, we present a practical tree-distance metric: MASTtreedist, a Maximum Agreement SubTree (MAST) based comparison metric in Mesquite's Tree Set Viz Module. In this metric, the efficient optimizations for the Maximum Weight Clique problem are applied. The result suggests that the proposed method can efficiently compute the MAST distances among trees, and such tree topological differences can be translated as a scatter of points in two-dimensional space. We also provide statistical evaluation of provided measure with respect to RF using experimental data sets. This new comparison module provides a new tree-tree pairwise comparison metric based on the differences of the number of MAST leaves among constructed phylogenetic trees. Such a new phylogenetic tree comparison metric improves the visualization of taxa differences by discriminating small divergences of subtree structures for phylogenetic tree reconstruction.

Availability: http://www.rc.usf.edu/MASTtree

Contact: honghuang@usf.edu

Introduction

Researchers may collect the data (such as DNA sequences) for each of the different taxa (genes, species, etc.), then construct phylogenetic trees. Many tree reconstruction methods could produce more than one candidate tree for the input dataset. Very often the number of trees can be in the hundreds or thousands (Ayre et al., 2012; Than et al., 2008; Matthews et al., 2010). These candidate trees are computed so as to resolve the conflict, summarize the information, and reduce the large number of possible solutions to select the appropriate ones for further analysis. Multiple tree visualization and construction solutions have been proposed for the management and annotations of large and single trees (Ulitsky et al., 2006; Letunic and Bork, 2007; Jordan and Piel, 2008; Santamaria and Theron, 2009), as well as comparisons of trees (Trooskens et al., 2005). Approaches such as Tree Set Viz (Amenta and Klingner, 2002; Hillis et al., 2005) were reported for visualizing sets of trees according to their similarity computed by tree-to-tree distance metrics (e.g., Robinson-Foulds) using multi-dimensional scaling (MDS). It is a module within Mesquite program, which is an open source Java-based platform that allows building and loading new analysis modules (Maddison and Maddison, 2012). The results of MDS analyses provide a tree-like visual comparison plotted as the MDS ordination plots in 2D representations of multidimensional space. The program could also generate a consensus tree by selecting the "islands" / "clusters" of the candidate trees shown in the two dimensional "tree space".

The Robinson-Foulds (RF) distance metric had been implemented in the Tree Set Viz module by summing the number of internal edges (branches) that must be collapsed or removed from one tree from another (Robinson and Foulds, 1981). Since RF computed the number of

edges in disagreement, one may wish, however, to conduct tree-to-tree comparison based on the number of leaves causing the disagreement by detecting the subtle differences reflecting their least common ancestors or their maximum similarity of the subtrees. These two methods differ conceptually, and may differ greatly in practice. However, MAST is NP-hard, and computationally sophisticated. In this research, we have provided a computationally proficient MAST tree-tree comparison metric solution for tree comparisons in Tree Set Viz Module.

Implementation

MAST can be used either in rooted or unrooted trees (Bryant 1997; Farach et al., 1995; Patric and Ostergard, 2001). It has been implemented in PAUP (Swofford, 2012). The MAST method focuses on extracting the common structure in multiple trees (Finden and Goron, 1985), whereas the Robinson-Foulds method determines how different the given trees are according to the internal edges' dissimilarity (Robinson and Foulds, 1981), therefore having a computational advantage for certain cases of phylogenetic tree reconstruction (Bryant, 1997).

The proposed MASTtreedist could compare with two trees, T1 and T2, with an agreement subtree with the largest possible number of leaves, which is given by: MAST (T1, T2) equals the number of leaves in common for the maximum subtree in T1 and T2 (Bryant, 1997). The MASTtreedist tree to tree pairwise distance metric was implemented in the Tree Set Viz module version 2.1 (Hillis et al., 2005) in the Mesquite software package (version 1.01) (Maddison and Maddison, 2012). To implement the MASTtreedist module, a parallel folder named "MAST" was built along with "Robinson-Foulds" containing the class files for the MASTtreedist module. The "MAST" class inherited "NumberFor2Trees" class from the Mesquite library. "MAST" class used "NumberFor2Trees" class method: "calculateNumber" to

compute the MAST number of leaves for two trees. At the end, MASTtreedist metric returns the value of the number of pruning leaves (total taxa minus the MAST leaves for two trees) to represent the tree "distance" (indicating similarity of trees), and such "distances" were visualized under multi-dimensional-scaling. However, MASTtreedist implementation could be run-time slow since it is NP-hard.

In this application, several techniques were used to optimize the program. MASTtreedist has many set operations implemented with a boolean array. The program also takes advantage of symmetries for rooted triples and fan triples (see their definitions in Bryant 1997, P 180). A Java build-in class "Hashtable" was used to save all possible repeats of the calculation, including the least common ancestor (Bryant 1997, p 175). Dynamic programming technique was also applied to traverse each pair of tree leaves and determines if the pair had been computed and memorizes the result for every pair of leaves in a two-dimensional array.

The Maximum Weight Clique calculation is NP-hard, and the most time-consuming step. In particular, the algorithm conducts combinatorial backward-searching from the end of the array containing all the possible subgraphs. If the nodes in the Maximum Weight Clique were stored in the front of the array, then the algorithm has to take more time to enumerate almost all possible combinations that can be found in the weighted graph. If the nodes in the Maximum Weight Clique can be put toward the end of the array, the search can end earlier without considering all possible node combinations.

A key value (integer) was assigned to each node, so that the nodes can be rearranged from low to high in accordance with the correspondent key value. The key value is related to the weight of the current node, and the degree or weight of adjacent nodes. It can be defined as the following:

$$key(k) = \sum_{i \in adj(k)} (d(i) \times w(i)) + w(k)$$

where *k* represents the current node, *adj(k)* is the set of adjacent nodes of the current node *k*, *d(i)* and *w(i)* are the degree and weight of the node *i* respectively. In this equation, the run-time complexity is $O(n \times d)$, where *n* is the number of leaves, *d* is the maximum degree of the tree. It is much faster compared to the run-time of Maximum Weight Clique. It also does not affect the overall time complexity, then can improve the computing speed for Maximum Weight Clique.

The run time complexity of the MASTtreedist metric is $O(k \times n^3 + n^d)$, where *k* is the total trees for comparison, *d* is the maximum degree of the internal nodes, and *n* is the number of total leaves of taxon.

Results

Two datasets: Camp (Cosner et al., 2000; Moret et al., 2001; Stockham et al., 2002), and PEVCCA (Stockham et al., 2002; Van de Peer et al., 2000) were obtained to test the MAST performance and compare the results with the ones from RF. The Camp dataset is for the phylogenetic trees breakpoint reconstruction for the *Campanulaceae* family (Moret et al., 2001), which contains 216 trees on 13 leaves. The PEVCCA datasets in this research contains 168 trees on 129 leaves obtained by maximum parsimony searches of the small subunit ribosomal RNA sequences (Stockham et al., 2002; Van de Peer et al., 2000).

The MASTtreedist tree comparison metric can be deployed along with other metrics (e.g., Robinson-Foulds) in the Tree Set Viz module within Mesquite. Tree Set Viz module used multi-dimensional scaling to represent the relationships between topologies (in this case, the topologies as a scatter of points in two-dimensional space. The software arranges the points such that they group according to the distance between the trees (distance between trees was

calculated using MAST or RF). The tree distances were indicated by their similar levels (similar trees are close to each other). The default step size which suggested the speed of positional change for each tree during the MDS process in Tree Set Viz was used in all analyses. The MDS was allowed to proceed until the stress function stopped changing. To avoid being trapped in local optima, this procedure was conducted multiple restarts to make sure that similar results were being achieved. We collected the tree-tree distance output generated by MAST and RF methods for further analysis.

For Camps dataset, the MAST and RF arranged the trees based on their distances differently. Results from RF shows three clusters of the trees, while the display generated by MAST can categorize the trees based on their different number of similar leaves into eight groups (Figure 1). Thus the MAST metric can detect detailed topological differences (identify trees based on their similar maximum subtrees). Figure 2 showed that the histogram of RF only had four unique tree-tree distance values, but MAST had five. This leads to the conclusion that MAST is more discriminative than RF in this dataset. MAST is also more reminiscent of the normal distribution than RF in this dataset (Figure 2).

As for the PEVCCA dataset, the scatter of dots in the two-dimensional space for MAST demonstrates a dense aggregation of the trees clusters, while the RF shows a sparse distribution of the tree clusters (Figure3A). Researchers could use Mesquite to construct a strict consensus tree by highlighting the selected tree groups. Figure 3B showed the consensus tree created by RF is less informative than the MAST one (multiple splits from a single tree node). In addition, the Figure 4 histogram indicated that RF distance values distribution has a similar shape with MAST's, however, MAST results have more unique values (40 different values for MAST, but only 29 for RF). The sparse distribution of trees in RF 2D MDS space is due to the wider range

of the tree-tree distance values, since the incremental interval is value of two in RF but one in MAST. MAST can distinguish the subgroups while RF could not detect due to having more tree comparison values. RR distance provided less discrimination than MAST distance, also lacking robustness in the face of very small changes in certain cases.

Conclusion

In this paper, we have proposed the MASTtreedist tree comparison metric for measuring distance between trees, and provided its evaluations with other measures such as RF. The RF distance metric is based on the split decompositions of the two tree topologies and is the number of edges that have no conflicts in the other tree structure (Robinson and Foulds, 1981). The new measure is the MAST distance metric complement with RF distance method, to enrich the tree comparison analysis. MAST describes the number of leaves on the largest subtree that both tree structures have in common (Moret et al., 2001). The MAST metric scores includes a maximal subset of taxa for which the subtrees calculated by the input trees in agreement, gives more meaningful results in some cases than other tree comparison matrices. This approach is particularly useful when a only a few taxa that causing the topological differences among trees, thereby providing a means of identifying these small sets of conflicted taxa.

It is practical for large tree analysis by optimizing the MAST program running time. Our testing data indicated that MASTtreedist metric, though harder to compute, and has theoretical advantages to make large-scale tree comparisons to visualize the subtle tree differences (when only a few taxa are responsible for the incongruity among trees) in the “tree space” using MDS. The program allows the user navigating a set of trees based on their topological similarity, and select subsets of trees for further analysis (such as constructing a consensus tree). This approach

could be extended to accommodate trees, where users would select trees for new tree reconstruction.

Acknowledgements

The authors would like to express their gratitude to David Swofford, and James Wilgenbusch for insightful suggestions, and comments on this research.

Author Disclosure Statement

No competing financial interests exist.

Figure 1. The phylogenetic trees distances (computed by MASTtreedist) were displayed as white dots in the 2-D spaces using the Camps dataset. MASTtreedist (left) metric showed more discriminative (identify more clusters) than Robinson-Foulds (right). Input data ("Camp.nex") can be found in website http://rc.usf.edu/MASTtree.

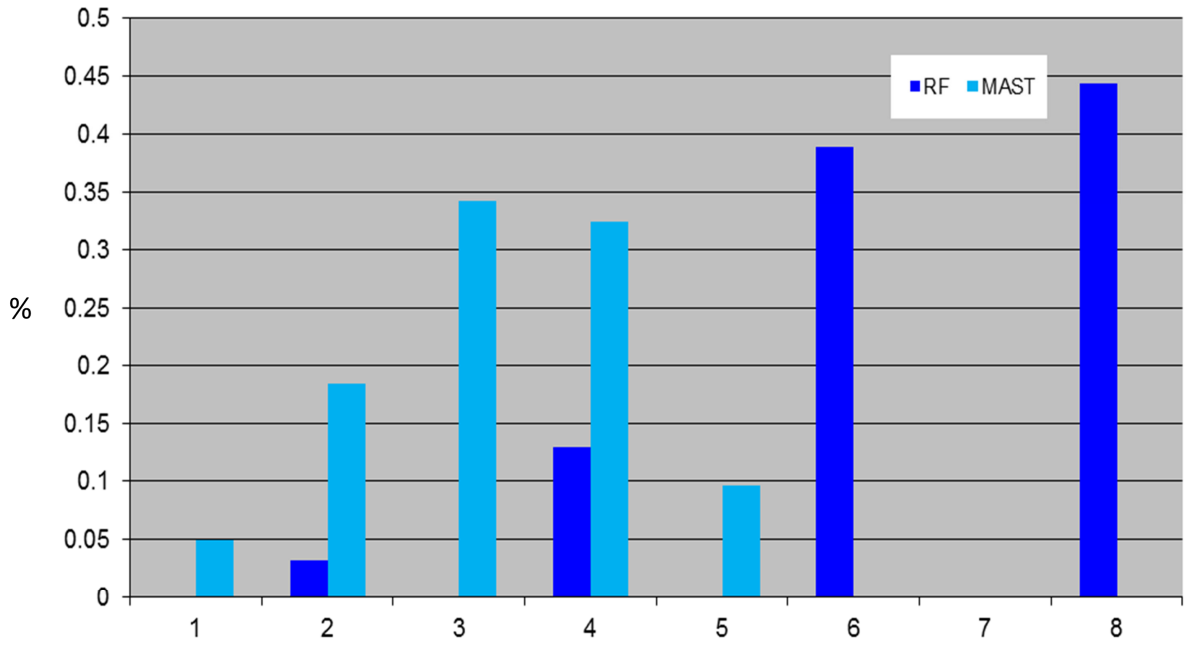

Figure 2. Comparison of distribution of MAST and RF. Comparison of RF dissimilarity measure with MAST, showing that the percentages for number of obtained pairs of trees (y axis) with certain distance values (x axis) using Camp dataset.

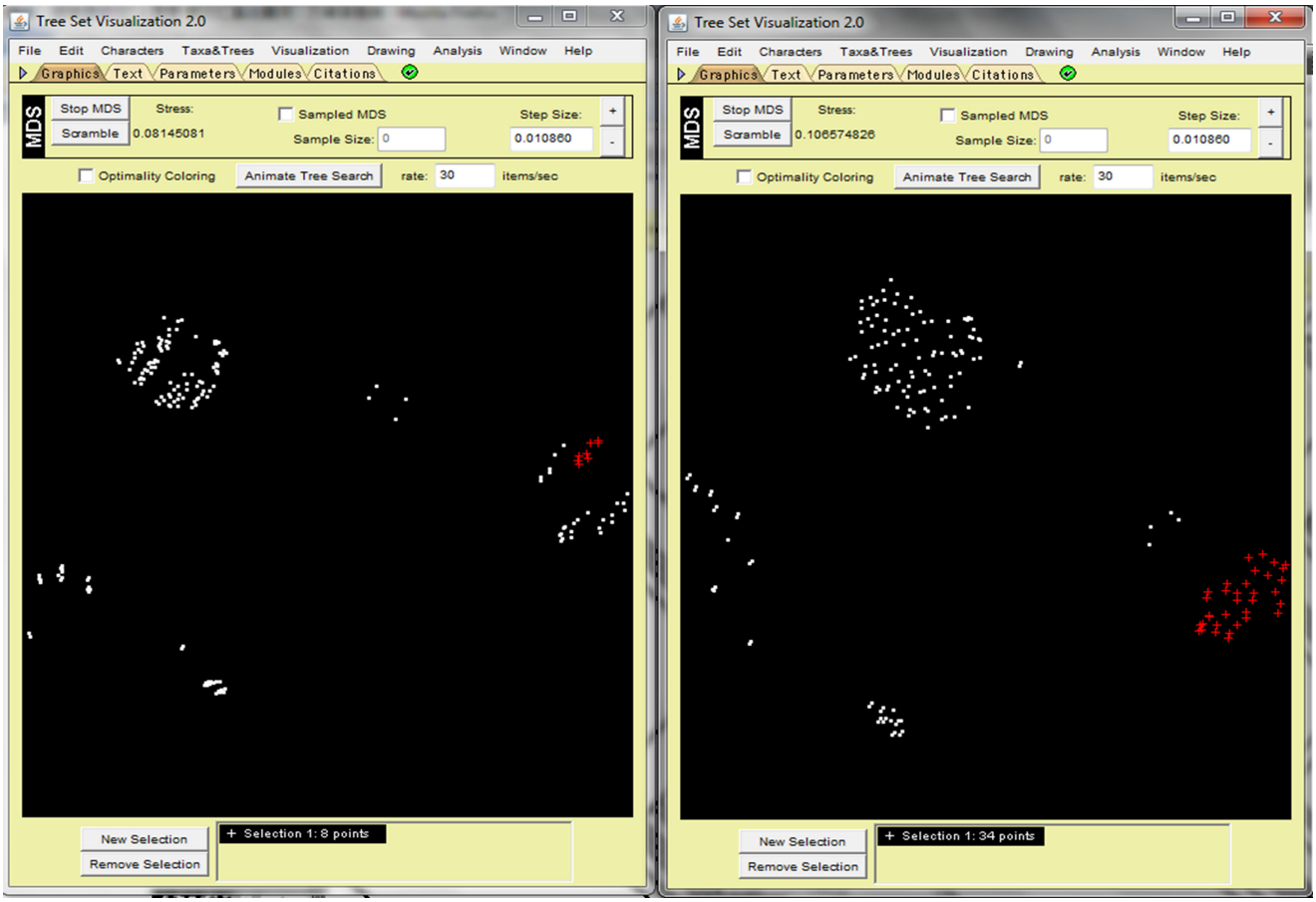

Figure 3A. The phylogenetic trees distances (computed by MASTtreedist) were displayed as white dots in the 2-D spaces using the Camps dataset. MASTtreedist (left) metric showed more discriminative (identify more clusters) than Robinson-Foulds (right). Input data ("PEVCCA.nex") can be found in website http://rc.usf.edu/MASTtree.

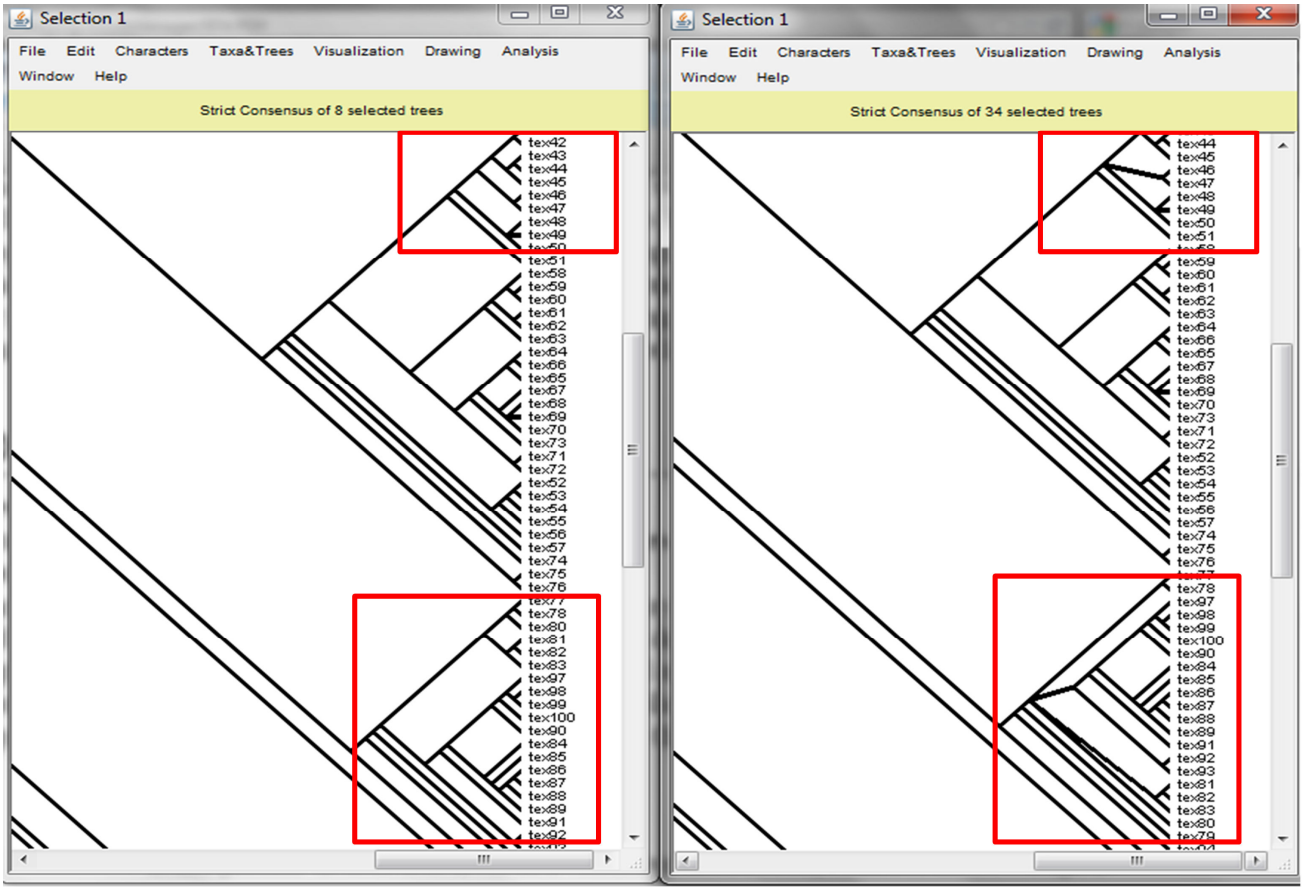

Figure 3B. Strict consensus trees were created in the MDS space computed by MAST (left) metric or RF (right). MAST could further break a tree group into multiple subgroups but RF cannot. These subgroups are belonged to a sparse, indistinguishable tree group in RF (right). The MAST consensus tree demonstrate less ambiguous information in compared to the RF one. The RF consensus tree was produced from 34 selected trees from an indiscernible group. The MAST consensus tree, however was created from a distinct subgroup (8 of 34 trees). The ambiguous relationships among tree leaves were highlighted as red.

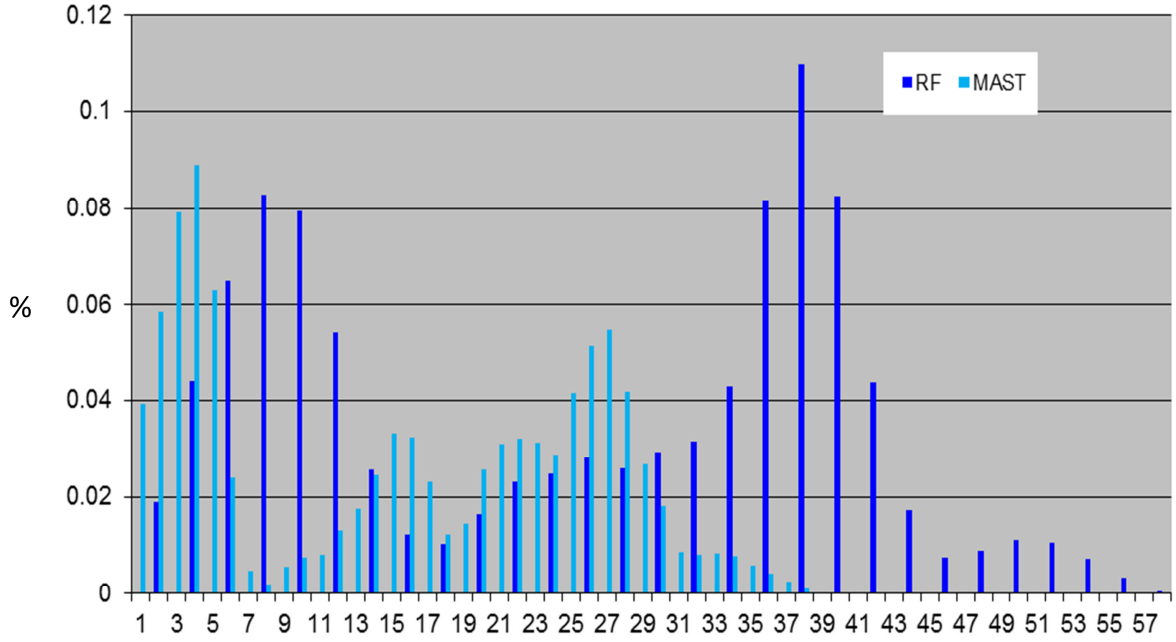

Figure 4. Comparison of distribution of MAST and RF. Comparison of RF dissimilarity measure with MAST, showing that the percentages for number of obtained pairs of trees (y axis) with certain distance values (x axis) using PEVCCA dataset.

References

Amenta N, Klingner J. 2002. Case study: Visualizing sets of evolutionary trees. 8th *IEEE Symposium on Information Visualization*, 71–74.

Ayres DL, Darling A, Zwickl DJ, Beerli P, Holder MT, Lewis PO, Huelsenbeck JP, Ronquist F, Swofford DL, Cummings MP. 2012. BEAGLE: an Application Programming Interface and High-Performance Computing Library for Statistical Phylogenetics. *Syst Biol*, 61(1):170-173.

Bryant D. 1997. Building trees, hunting for trees, and comparing trees, theory and methods in phylogentic analysis. *PhD dissertation,* University of Canterbury.

Cosner ME, Jansen RK, Moret BME, et al. 2000. A new fast heuristic for computing the breakpoint phylogeny and experimental analyses of real and synthetic data. In *In*

Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology: August 19–23, 2000; La Jolla, CA. Edited by Bourne P, Gribskov M, Altman R, Jensen N, Hope D, Lengauer T, Mitchell J, Scheeff E, Smith C, Strande S, Weissig W. Cambridge, MA: AAAI Press; 2000:104-115.

Farach M, Przytycka T, Thorup M. 1995. On the agreement of many trees. *Inform Process Lett*, 55(6):297–301.

Finden CR, Gordon AD. 1985. Obtaining common pruned trees. *J Classif* , 2:255-276.

Hillis DM, Heath TA, John KS. 2005. Analysis and visualization of tree space. *Syst Biol*, 54(3):471 - 482.

Jordan GE, Piel, WH. 2008. PhyloWidget web-based visualizations for the tree of life. *Bioinformatics,* 24,1641-1642.

Letunic I, Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23, 127–128.

Maddison WP, Maddison DR. 2012. Mesquite: a modular system for evolutionary analysis, http://mesquiteproject.org.

Matthews SJ, Williams TL. 2010. MrsRF: an efficient MapReduce algorithm for analyzing large collections of evolutionary trees. *BMC Bioinformatics*, 11(Suppl 1):S15.

Moret BME, Wang L-S, Warnow T, Wyman S. 2001. New approaches for reconstructing phylogenies from gene order data. *Bioinformatics*, 17:S165-S173

Moret BME, Wyman S, Bader DA, Warnow T, Yan M. 2001. A new implementation and detailed study of breakpoint analysis. In *Proceedings of Sixth Pacific Symposium on Biocomputing: 3–7 January 2001; Hawaii*. World Scientific Press; 583-594.

Patric R, Ostergard J. 2001. A New Algorithm for the Maximum-Weight clique problem. *Nordic Journal of Comput*, 1-13.

Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. *Mathematical Biosci*, 53:131–147.

Santamaria R, Theron R. 2009. Treevolution: visual analysis of phylogenetic trees. *Bioinformatics*, 25, 1970-1971.

Steel M, Warnow T. 1993. Kaikoura Tree Theorems-Computing the Maximum Agreement Subtree. *Inform Process Lett*, 48: 77–82.

Stockham C, Wang LS, Warnow T. 2002. Statistically Based Postprocessing of Phylogenetic Analysis by Clustering. *Proceedings of 10th Int'l Conf. on Intelligent Systems for Molecular Biology (ISMB'02)*, 285-293.

Swofford DL. 2012. PAUP 4.0* Phylogenetic Analysis Using Parsimony. Sinauer Associates, Sunderland, Masschusetts.4.0 2012.

Than C, Ruths D, Nakhleh L.2008. PhyloNet: A software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*, 9:322.

Trooskens G, Beule DD, Decouttere F, Criekinge WV. 2005. Phylogenetic trees: visualizing, customizing and detecting incongruence. *Bioinformatics*, 21, 3801–3802

Ulitsky I, Burstein D, Tuller T, Chor BT. 2006. The average common substring approach to phylogenomic reconstruction. *J Compu Biol*. 13, 336-350.

Van de Peer Y, De Rijk P, Wuyts J, Winkelmans T, De Wachter R. 2002. The European small subunit ribosomal RNA database. *Nucl Acids Res*, 28:175-176.