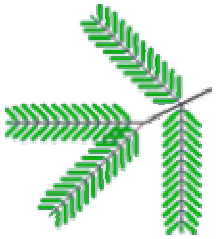
TREE SET VISUALIZATION MODULE FOR MESQUITE
VISUALIZING SETS OF PHYLOGENETIC TREES



Nina Amenta[1], Katherine St. John[2], and Jeff Klingner[3]

Tracy A. Heath[4]

Frederick Clarke[2], Denise Edwards[2], Silvio Neris[2], Ruchi Mahindru[2], and Nicolay Postarnakevich[1]

[1]University of California, Davis, Department of Computer Science
[2]City University of New York, Lehman College, Department of Computer Science
[3]Stanford University, Department of Computer Science
[4]University of Texas, Austin, Department of Biological Sciences

This research project is supported by a grant from the National Science Foundation, NSF-ITR 0121651/0121682: "Collaborative Research: Exploring the Tree of Life."

1 INTRODUCTION

Phylogenetic trees provide valuable information about evolutionary relationships and are powerful tools used in many areas of biology. For example, comprehensive phylogenies of pathogens allow researchers to predict the evolution of a virus or bacteria, aiding in the development of vaccines. Additionally, phylogenies can be effectively applied to analyses of character evolution, gene expression, and many other areas of biology, such as conservation, ecology, and developmental biology.

There are many methods available for inferring phylogenies, and such analyses can be conducted using several different types of data. Systematists can collect molecular data, such as nucleotide or amino acid sequences, or they can use morphological or behavioral information to construct an evolutionary tree. Some methods, however, can yield a large number of trees that the systematist must then summarize. For example, a set of trees may include several optimal or near-optimal parsimony trees or the set of trees may be the trees sampled by a Bayesian analysis.

Tree Set Visualization is a program capable of summarizing sets of phylogenetic trees. This program depicts a "tree space" using multi-dimensional scaling based on tree-to-tree distance metrics. Visualizing sets of trees in such a way can yield more information about the trees that may otherwise be lost using more conventional methods for summarizing large sets of trees.

1.1 Tree distance metrics

The Tree Set Visualization module calculates a matrix of pairwise tree distances. The trees are then displayed using multi-dimensional scaling (MDS) to best represent the observed distances.

Tree metrics compare topological differences among trees. While there are several different tree difference metrics, some are too computationally difficult to include in this module. For a description of the available distance measures see section 3.4.1.

1.2 Multi dimensional scaling

Multi-dimensional scaling (MDS) is a method for plotting points in a redefined space based on a matrix of distances. The Tree Set Visualization module plots the individual trees in two dimensions. Because tree distances define a hyper-dimensional space, MDS must plot the trees in such a way to minimize the amount of distortion between the observed distances (the calculated pairwise tree distances) and the represented distances (the distances depicted in two dimensions). MDS incorporates a stress function, which is used to minimize the amount of distortion.

2 GETTING STARTED

2.1 Installation and Requirements

Tree Set Visualization is a module for the Mesquite software package. Mesquite is an open source program that can be used for a number of analyses. Some of the applications available include comparative analyses, ancestral state reconstruction, coalescence, simulation, and hypothesis testing. Documentation and other information about Mesquite and its applications can be found at http://www.mesquiteproject.org/mesquite/mesquite.html.

In order to run Mesquite, a JAVA virtual machine (JAVA 1.1 or higher) must be installed. The Mesquite software is available at http://www.mesquiteproject.org/mesquite/download/download.html and is distributed according to the terms of the GNU lesser general public license (http://www.gnu.org/copyleft/lesser.html). Mesquite must be installed prior to installation of the Tree Set Visualization module.

Once Mesquite is installed, the Tree Set Visualization module can be obtained online at http://comet.lehman.cuny.edu/treeviz/software.html. Once the file is downloaded and decompressed, move the "*treecomp*" folder to "*Mesquite_Folder/mesquite*". When Mesquite is reopened and a file is loaded, the Tree Set Visualization module should be accessible in the "*Taxa&Trees*" menu.

3 VISUALIZING TREES

There are many ways in which this software can be useful. Trees can either be colored according to their optimality score or by a color designated by the user. This capability allows for several promising applications such as illustrating the increasing likelihood scores of the trees sampled by the Markov chain in a Bayesian analysis (coloring trees according to optimality score), or the user can directly compare the trees obtained from separate analyses of the same dataset (coloring sets of trees according to which run or method generated them). We expect that there may be several other ways in which this program will be useful to the scientific community.

3.1 Trees

The Tree Set Visualization module allows for either rooted or unrooted trees. However, both types of trees cannot be combined in a single analysis. Trees containing polytomies are also acceptable.

It is important to note that in order to load a very large number of trees; a significant amount of memory must be allotted to the Mesquite program. Additionally, the performance of Tree Set Visualization, in terms of analyzing many trees, is dependent on the machine running the program.

3.2 Tree Files

Trees can be stored in the program using a few approaches. However the trees must be in NEXUS format. A tree block can be included in the NEXUS file containing your data matrix (example3.2.1.nex). This method requires no translation table in the tree block. Second, a data file is opened in Mesquite containing only a data matrix (no trees block), you can load trees from a separate tree file by going to the "*Taxa&Trees*" menu and selecting "*Get trees from file*". Choose one of the 3 options and select the appropriate tree file. This approach also does not require a translation table in the trees block. Finally, you can load a tree file into mesquite even if you have not loaded a data matrix simply by selecting "*File→Open→File*" (or by typing ctrl-O for Windows/UNIX or command-O for MacOS) and choose the appropriate tree file (example3.2.2.tre). If the file contains only a trees block, then, there must be a translation table for the file to load properly. If there is no translation table in the trees block, however, the file can still be loaded if it also contains a taxa block with a taxon list.

3.3 Starting Tree Set Visualization

Once a file is loaded and trees are stored, the Tree Set Visualization module can be started. Go to "*Taxa&Trees*" in the menu and select "*Tree Set Visualization 2.0*". A window will appear prompting you to select the trees for visualization (figure 3.1). Select "*Stored Trees*" to load the trees from file.

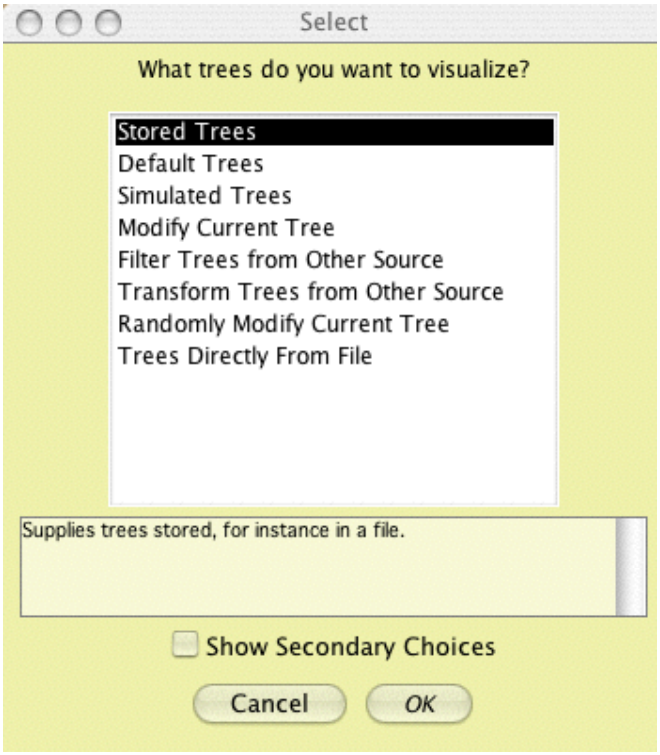


Figure 3.1: Indicate the source of the trees for visualization

Once the source of the trees has been indicated, another window will appear prompting you to choose a tree distance metric (figure 3.2).

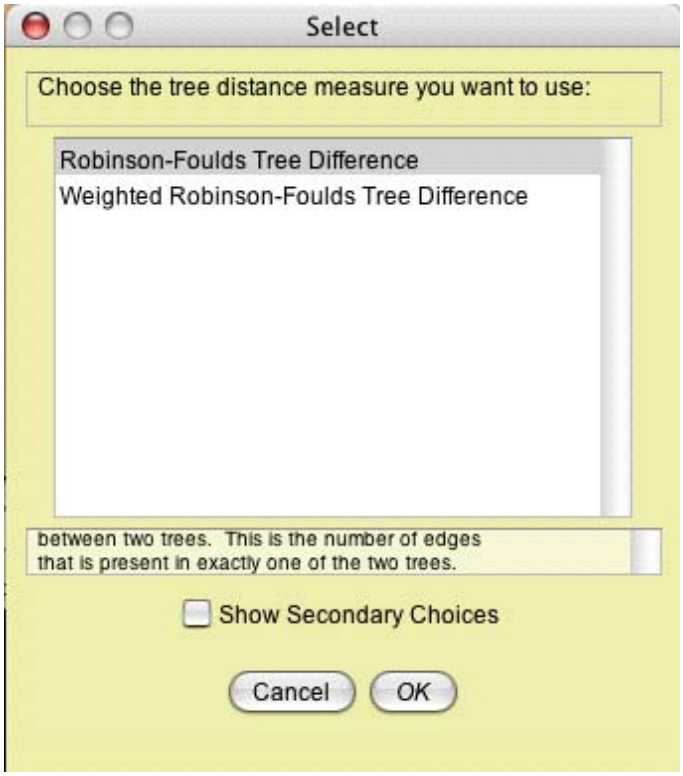


Figure 3.2: Indicate the tree distance metric to be calculated

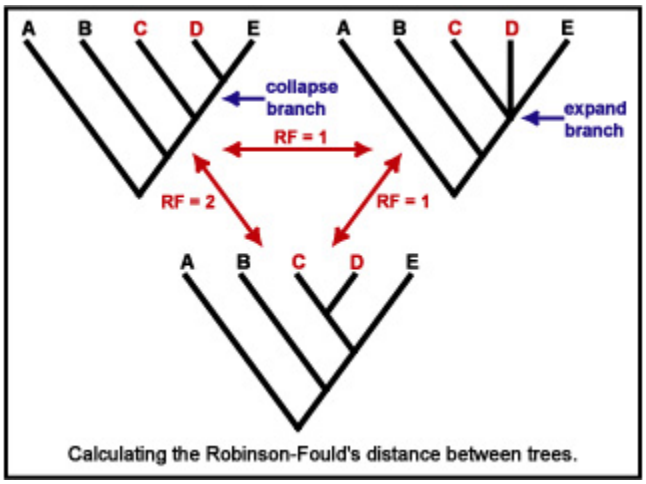
3.4 Choosing a tree distance metric

The Tree Set Visualization module can calculate the pairwise tree distances using a few of the available metrics. Alternatively, the user can use tree distances computed by a different program by uploading the matrix from file.

3.4.1 Available Metrics in TSV

Below are the currently available options for tree distance metrics. Each metric calculates tree distances differently and each leads to different visual representations of the tree differences. Also, the user should be aware that as the number of trees and the number of taxa in the trees increase, the computational time required for calculating the distance metrics also increases.

Robinson-Fould's tree difference – This tree distance metric is perhaps the most commonly used measure of tree difference. It sums the number of internal branches that must either be collapsed or expanded to move between two topologies. This metric is also called "symmetric distance" (Penny and Hendy, 1985) and is the default metric used for calculating pairwise tree distances in PAUP* 4.0b1 (Swofford, 2001). R-F distance is proportional to the number of nearest neighbor interchanges between two trees (e.g. one nearest neighbor interchange is equivalent to an R-F distance of two).



Calculating the Robinson-Fould's distance between trees.

Weighted Robinson-Fould's distance – This variant of R-F distance incorporates branch length information to calculate tree distances. Therefore, the weighted R-F distance between two trees increases as the differences in their branch lengths increase. For example, two trees that have identical topologies can have a weighted R-F distance greater than zero if they do not have identical branch lengths. If this metric is selected for visualization, branch length information must be included in all of the tree descriptions.

3.5 Starting MDS

After the tree distances are selected and the calculation of the matrix is completed, the Tree Set Visualization window will open. Above the field displaying the trees are several options (figure 3.3). To begin the multi-dimensional scaling of the tree distances matrix, click on the "*Start MDS*" button in the upper left corner.

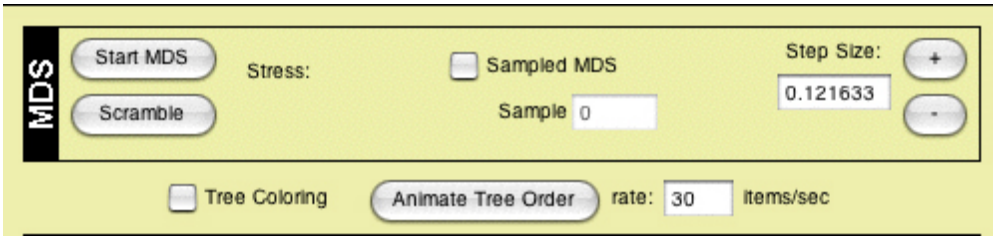


Figure 3.3: The Tree Set Visualization window command buttons

The points in the display field should begin repositioning and the stress value should start to decrease.

3.6 Stress and when to stop MDS

MDS repositions the individual points to most accurately depict the trees in two-dimensional space and minimize the stress value. The stress value is a measure of the distortion between the displayed distances and the observed distance matrix. As the visualization proceeds, this value should steadily decrease as the two-dimensional representation of the distances approximates the actual tree-to-tree distances. It is important to understand that the tree space displayed is redefined for every set of trees. Therefore, stress values and MDS plots from one set of trees cannot be compared with those from another set containing different trees.

With very few trees, the process of visualization is very rapid. After starting multi-dimensional scaling, the points in the display field should immediately reposition and the stress level will begin to stabilize. When the stress value stops decreasing (or is continually oscillating between two minutely different values), the MDS can be discontinued.

It is possible for the MDS to get trapped in local optima. This can be avoided by conducting multiple restarts. After the MDS has been stopped, it can be re-started by clicking the "*Scramble*" button and then starting the MDS again.

3.7 Step size

When the Tree Set Visualization window is opened, the step size is displayed next to buttons that increase or decrease the step size value. The default value is an optimal value which is calculated based on the number of trees and the range of pairwise distances. The step size indicates the degree of positional change for each tree during the MDS process. If this value is too high, the points will appear to move erratically over tree space and may never optimize the stress value. Should the step size be set too low, however, the trees will be repositioned at very small increments and the MDS will take a very long time.

It may be beneficial to initially increase the step size at the beginning of the MDS analysis, and then return it to the default value. Once it appears as though the stress

value has reached an optimum, decrease the step size and continue the run until the stress ceases to change. This practice may prevent the analysis from getting trapped in a local optimum.

3.8 Sampled MDS

This feature is useful when analyzing a very large set of trees. When "Sampled MDS" is selected, a percentage (default is 10%) of random trees are circled. Then, when the MDS is initiated, the points are repositioned with respect to the circled points instead of all points.

Sampled MDS is effective when applied at the beginning of the MDS run. Once the stress value appears to stabilize, turn off "Sampled MDS" and continue to run the analysis. Visualizing a large set of trees this way may reduce the amount of time it takes to obtain the optimal visual representation of tree space.

3.9 Coloring trees

This option is very useful for visualizing structure in a set of trees based on either optimality score or a user-designated category for each tree. The values for the tree scores are uploaded from file.

3.9.1 Color trees based on optimality score

A set of trees can be colored based on optimality scores by loading the scores output files from either PAUP or MrBayes. Alternatively, the user can also create a scores file. Once the tree distances have been calculated and the Tree Set Visualization window is open, select "*Tree Coloring*". A window will open to allow you to select the appropriate file. Then, the visualization should change such that the individual points are colored according to their likelihood score. Scores files can be created in PAUP for any set of trees and can be directly loaded to the TSV module (see example3.9.1.tre and example3.9.1.scores). The output files generated from a MrBayes run include both a tree file (with the ".t" file extension) and a parameters file (ending with a ".p" file extension). The MrBayes parameters file can be loaded into the TSV module to color the trees from the ".t" file based on their individual likelihood score. However, often the tree file obtained from a MrBayes run contains such a large number of trees that it may not be possible to view them all. (See example3.9.1.tre and example3.9.1.p)

3.9.2 Coloring groups of trees

Alternatively, a scores file can be created for any set of trees. For example, the user can combine sets of trees from different analyses of the same data set. Colors can be assigned to specific groups of trees from each of the analyses, so that tree space can be easily visualized. The file can be created using a spreadsheet program such as MS Excel. See the example files (*example3.9.2.tre & example3.9.2.scores*) for the proper file format for this type of visualization.

3.10 Animate tree order

This feature allows the user to illustrate where each of the trees falls in tree space. After the MDS is stopped, when "*Animate Tree Order*" is initiated, the points will appear according to their order in the tree file. If the trees are from a tree search, this feature can depict the process of that search.

The user can also increase or decrease the rate at which trees are displayed during the animation. The default value is 30 trees/second. When this value is increased, the animation will proceed at a much faster rate.

3.11 Consensus trees

Consensus tree methods are commonly used to summarize sets of phylogenetic trees. These methods combine tree topologies such that if two trees have any identical bipartitions, these overlaps are represented in the resulting consensus tree.

3.11.1 Types of consensus trees

Strict Consensus – This consensus tree method results in a summary tree containing only the bipartitions found in *every* source tree. Otherwise, polytomies are used to represent areas of the tree that have multiple resolutions in the source trees.

Majority Rule Consensus – The consensus tree produced by this method contains more resolution than the strict consensus because bipartitions found in most of the source trees are included in the summary tree. Typically, if more than half of the trees have a particular bipartition it is included in the consensus.

3.11.2 Viewing consensus trees

In the Tree Set Visualization window, the user can select an individual tree or a group of trees to be displayed in a separate window. To select a single tree, simply click on a point in the tree space. A window will appear displaying that tree and its name (the name can be incorporated in the tree description in the tree file). If the user selects a group of trees, highlighting a number of points with the mouse or by clicking on individual points while holding the Ctrl key (Windows) or command key (MacOS), a consensus tree will be displayed in a separate window.

3.12 Write to file

The Tree Set Visualization module is capable of producing a postscript file of the visualization. This feature allows the user to save the tree visualization (see example3.12.ps).

3.12 Changing the tree distance metric

Once the MDS visualization is completed, the user can change the tree distance metric in order to depict the set of trees in a different way. To change the distance measure, go to the "*Visualization*" menu and select "*Tree Difference Metric → <new metric>*". The distance matrix will be recalculated, when this is completed, click on the "*Scramble*" button and restart the MDS.

3.13 Increasing Mesquite's Memory

In order to analyze several thousand trees using the Tree Set Visualization module, the user must allocate more heap memory to Mesquite. Below are some ways of doing this.

3.13.1 Windows

One way to increase the memory used by Mesquite is to create a Java shortcut. First, go to the *C:\WINDOWS\system32* folder and create a shortcut for the Java executable (java.exe). Once this is done, it may be helpful to move the shortcut to a more accessible folder. Then, open the shortcut properties (right click on icon and go to properties). Under the tab labeled "Shortcut", change the target to: *C:\WINDOWS\system32\java.exe -Xmx500mb -cp "C:\Program Files\Mesquite_Folder" mesquite.Mesquite*. The value in *-Xmx500mb* indicates the amount of memory and can be increased or decreased based on how much system memory is installed on your computer. Then, under the tab labeled "General", you can change the name of the shortcut from "*Shortcut to java.exe*" to "*Mesquite*" or whatever you may prefer. Once you have applied these changes, Mesquite will open when the new shortcut is opened.

3.13.2 Mac OSX

In OS X, one way to run Mesquite with a bigger heap is to execute the program from the terminal window. In the terminal window go to the "Mesquite_folder" directory. Once in this folder you can execute the program using the following command: `java -Xmx500m mesquite/Mesquite`. The value in *-Xmx500mb* indicates the amount of memory and can be increased or decreased based on how much system memory is installed on your computer. When the command is executed, mesquite should open on your desktop.

3 USEFUL REFERENCES

Amenta, N., F. Clarke, and K. S. John. 2003. A linear-time majority tree algorithm. *Algorithms in Bioinformatics, Proceedings* 2812:216-227.

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

Borg, I., and P. Groenen. 1997. *Modern Multidimensional Scaling*. Springer-Verlag, Heidelberg.

Buneman, P. 1971. The recovery of trees from measures of dissimilarity. Pages 387–395 in *Mathematics in Archaeological and Historical Sciences*. (F. R. Hodson, D. G. Kendall, and P. Tautu, eds.). Edinburgh University Press, Edinburgh.

Day, W. H. E. 1985. Optimal algorithms for comparing trees with labeled leaves. *Journal of Classification* 2:7–28.

Hillis, D.M., T.A. Heath, and K. St. John. 2004. Analysis and Visualization of Tree Space. *Systematic Biology*, in press.

Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogeny. Bioinformatics 17:754–755.

Huelsenbeck, J. P., F. Ronquist, R. Nielson, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294:2310–2314.

Lingoes, J. C., E. E. Roskam, and I. Borg. 1979. *Geometric Representations of Relational Data*, 2nd edition. Mathesis Press, Ann Arbor, Michigan.

Maddison, D. R. 1991. The discovery and importance of multiple islands of most-parsimonious trees. *Systematic Zoology* 40:315–328.

Maddsion, W. P., and D. R. Maddison. 1992. *MacClade: Analysis of Phylogeny and Character Evolution*. Sinauer, Sunderland, Massachusetts (see also http://www.macclade.org).

Maddison, D. R., D. L. Swofford, and W. P. Maddison. 1997. NEXUS: an extensible file format for systematic information. *Systematic Biology* 46:590-621.

Maddison, W.P., and D.R. Maddison. 2004. *Mesquite: A Modular System for Evolutionary Analysis*. Version 1.02 http://mesquiteproject.org.

Penny, D., and M. D. Hendy. 1985. The use of tree comparison metrics. *Systematic Zoology* 34:75–82.

Robinson, D. F., and L. R. Foulds, 1979. Comparison of weighted labeled trees. *Lecture Notes in Mathematics* 748:119–126.

Robinson, D. F., and L. R. Foulds. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences* 53:131–147.

Stockham, C., L.-S. Wang, and T. Warnow. 2002. Statistically based postprocessing of phylogenetic analysis by clustering. *Bioinformatics* 18:S285–S293.

Swofford, D. L. 2000. *PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods)*. Sinauer, Sunderland, Massachusetts.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pages 407-514 *in Molecular Systematics*, 2nd ed. (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer Associates, Sunderland, Massachusetts.

Young, F. W. and R. M. Hamer. 1987. *Multidimensional Scaling: History, Theory and Applications*. Erlbaum, New York.