# Visualizing Restricted Landscapes of Phylogenetic Trees

Ingrid Montealegre and Katherine St. John
The Graduate Center, CUNY
Department of Computer Science
365 Fifth Avenue, New York City, NY 10016
ingrid_nyc@excite.com & stjohn@lehman.cuny.edu

[**Keywords:** phylogeny, visualization.]

## Abstract

*We are designing tools to visualize very large sets of phylogenetic trees. Our tools give a three dimesional representation of treespace, with two dimensions representing the clustering of trees under multidimensional scaling, and the third dimension (the "height") the score of the tree (i.e. parsimony or maximum likelihood score). The user can rotate the resulting distribution to get a sense of the three-dimensional structure. This is implemented as part of the Mesquite system for phylogenetic analysis.*

## 1. Introduction

Evolutionary trees, or phylogenies, are an essential tool in biology, used in all kinds of processes such as understanding evolution, designing new drugs, predicting gene expression, and determining the origin of a virus strain. The most popular optimization methods for reconstructing evolutionary trees are intractable [7]. As a result, phylogenetic analysis usually produce a large set of best possible trees found during the search, not a single optimal tree. We are designing visualization tools to efficiently view and analyze thousands of such trees at a time, and tools that allow side-by-side comparisons of very large trees.

We have developed software to analyze, cluster, and visualize large datasets of phylogenetic trees [1]. Our software, `treecomp` is built as a module in the Mesquite System [5] and is freely available. The initial version of our software uses multidimensional scaling (MDS) to embed trees in two dimensional space (under the Robinson-Foulds distance). The user can select individual trees to view and compare, or multiple trees to view their consensus tree. See Figure 1.

The success of this led us to extend the visualization to three dimensions, increasing the amount of viewable information. The third dimension holds another metric, such as an optimality score. Now the user can explore the relationship between the two metrics and gain further understanding of the tree space by viewing the resulting landscape. Since the number of trees is prohibitively large for even small number of taxa (for 20 taxa, there are more than $10^{22}$ trees), we focus on restricted landscapes, that is, landscapes restricted to a set of input trees.

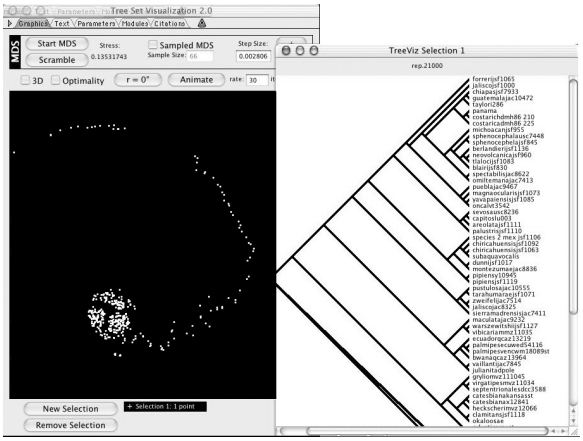



**Figure 1. Two dimensional representation of 650 trees of new world frogs trees, with a selected tree. (Data from Hillis Lab, UT Austin.)**

## 2. Prior Work

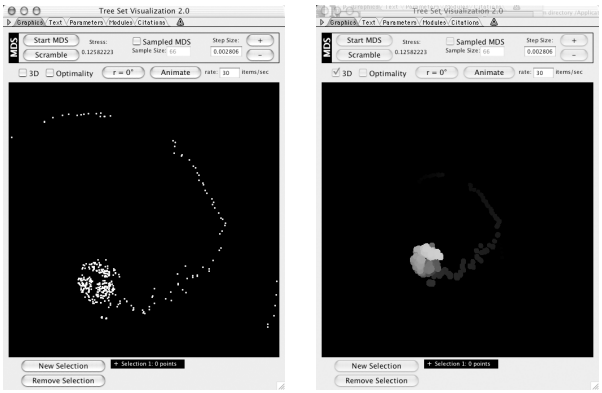
Previous work has focused on defining the question and computing the characteristics of the space. Mike Charleston [3] outlines a useful view of the space of phylogenetic trees, where the *landscape* of an optimisation problem is the solution space (the trees) and an optimality criterion (such as parsimony or maximum likelihood score). He focuses on measuring the landscapes via methods such as random sampling. The measuring of the fitness of a landscape was also

addressed recently [2] from a different viewpoint. These works measure analytically the landscape, while we focus on a complementary question of *viewing* the landscapes.

## 3. Our Contribution

Our tool focuses on visualizing large data sets of trees and their associated score (i.e. a parsimony or maximum likelihood score [6, 4]). Previous versions of our tool clustered sets of trees in two dimensions, using multidimensional scaling (MDS) (see [1] for more details). Our current version represents the solution set to some optimality criteria in three dimensions, using the third dimension for the score. In Figure 2, we show the dataset of 650 trees, clustered under MDS in two dimesions and in three dimensions, where the shading represents the maximum likelihood score of the trees (trees and scoring by a run of [4]).

By viewing the landscape restricted to the input set we can quickly find the distribution of values of the selected optimality criteria and characteristics of optimal solutions. Similarly, we can use the landscapes representing different optimality criteria to find the similarities and differences among the set of trees. When we view the two dimensional image produced by running MDS we lose information on the relative position of the cluster to each other. When we switch to a 3D projection we see that many of the clusters in the tail are further away. In Figure 3 we show two more rotations of the projection, showing more information about the relationship of the clusters.




2D MDS 3D projection

**Figure 2. Comparison of 2D MDS to equivalent 3D projections. Darker discs are further away than lighter discs.**

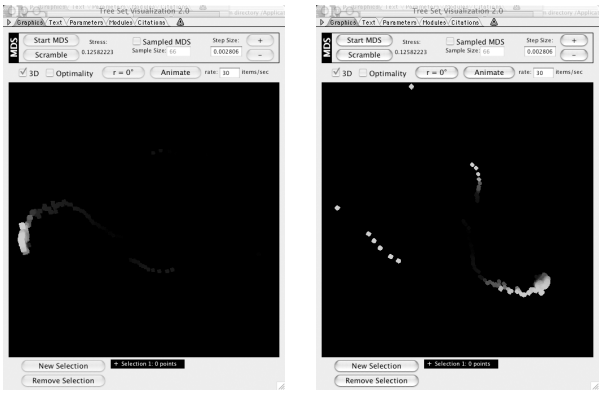



**Figure 3. Rotated views of the 3D projections.**

## 4. Future Work & Acknowledgments

We are extending the visualization to show the "surface" of the restricted landscape. Intriguing questions include how to do you sample and display small regions of treespace efficiently to allow the user to zoom-in and explore regions of interest.

We thank Jeff Klingner for the tree set visualization module, Wayne and David Maddison for Mesquite, Ken Perlin for very useful Java code, David Hillis and his laboratory for their encouragement and sample datasets, and Nina Amenta for helpful discussions and insights. This project was supported by NSF-ITR 0121651/0121682.

## References

[1] N. Amenta and J. Klingner. Case study: Visualizing sets of evolutionary trees. In *8th IEEE Symposium on Information Visualization (InfoVis 2002)*, pages 71–74, 2002. Software available at `comet.lehman.cuny.edu/treeviz/`.

[2] O. Bastert, D. Rockmore, P. Stadler, and G. Tinhofer. Landscapes on spaces of trees. Appl. Math. Comput., in press.

[3] M. Charleston. Landscape characteristics of tree space, 1996. `http://taxonomy.zoology.gla.ac.uk/ mac/`.

[4] J. P. Huelsenbeck and F. Ronquist. Mrbayes: Bayesian inference of phylogeny, 2001.

[5] W. Maddison and D. Maddison. Mesquite: a modular system for evolutionary analysis. version 0.992, 2002. Available from `http://mesquiteproject.org`.

[6] D. Swofford. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4*. Sinauer Associates, Sunderland, Massachusetts, 2002.

[7] D. Swofford and G. Olsen. Phylogeny reconstruction. In D. M. Hillis and E. C. Moritz, editors, *Molecular Systematics*, pages 411–501. Sinauer Associates, Sunderland, 1990.