Visualizing Sets of Evolutionary Trees

the undergraduate thesis of

Jeff Klingner

working under the supervision of

Dr. Nina Amenta

May 2001

The Department of Computer Science
The University of Texas at Austin

Abstract

One of the problems with current methods for phylogenetic reconstruction is the large number of equally parsimonious trees that are often found during a tree search; understanding these large sets of trees is a challenge for biologists. I explored the utility of a data visualization technique in creating 2D and 3D images of tree sets in order to improve researchers' understanding of the sets. I used multidimensional scaling based on Robinson-Foulds inter-tree distances to construct the visualizations. Direct visualization of taxa differences was also explored. I found that structure in the tree sets was reflected in the visualizations. Visual clustering in our pictures corresponds to islands of phylogenetic trees in the set and also reveals additional structure. Direct visualization of taxa differences provides a good alternative to displaying divergence with branch lengths and may be useful in a divide-and-conquer approach to phylogenetic reconstruction. I integrated all of the computational steps needed for building the visualizations into a module for the phylogeny software package Mesquite.

Table of Contents

1 Background

1.1 What is a Phylogenetic Tree?

A phylogenetic tree is a depiction of the evolutionary relationships of a set of organisms. A tree gives a natural representation of any hierarchical organization, and ever since Linneaus introduced the idea of a hierarchical classification of biological diversity, trees have been used to represent the organization of life on earth. With Darwin and his theory of evolution by natural selection came the notion that the classification of species should be based on true relatedness. It was realized that the branching structure of trees did more than simply portray a hierarchy; it could be seen as a direct representation of the branching diversification of life through history. Darwin realized this, and an illustration in his *On the Origin of Species* (Fig. 1) is the first phylogenetic tree. In Darwin's words, (pp. 129-130)

> The affinities of all the beings of the same class have sometimes been represented by a great tree. I believe this simile largely speaks the truth. The green and budding twigs may represent existing species; and those produced during each former year may represent the long succession of extinct species … The limbs divided into great branches, and these into lesser and lesser branches, were themselves once, when the tree was small, budding twigs; and this connexion of the former and present buds by ramifying branches may well represent the classification of all extinct and living species in groups subordinate to groups … As buds give rise by growth to fresh buds, and these, if vigorous, branch out and overtop on all sides many a feeble branch, so by generation I believe it has been with the great Tree of Life, which fills with its dead and broken branches the crust of the earth and covers the surface with its ever branching and beautiful ramifications.

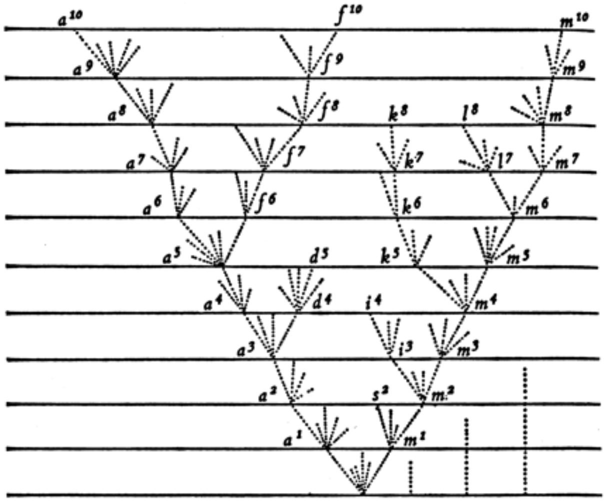


Figure 1 The first phylogenetic tree, from Darwin's *On the Origin of Species*

Darwin's theory brought about the creation of the branch of science called systematics, the study of biological diversity in an evolutionary context. Phylogenetic trees are very important to systematists, and a huge amount of current research throughout the biological sciences is

1.2 How are phylogenetic trees calculated?

There are many methods currently in use for estimating phylogenies. Maximum parsimony (shortest tree) is the most widespread, but maximum likelihood is also used frequently. Maximum parsimony and maximum likelihood are each based on different assumptions about the underlying evolutionary process, but all follow the same basic algorithmic pattern:

1. Choose a tree.
2. Calculate the statistic of interest (parsimony or likelihood) for that tree.
3. Repeat steps 1-2 until all trees have been checked or the user says stop.
4. Return the trees(s) with the best score.

1.3 Computational obstacles

The basic difficulty faced with this paradigm is the astronomical number of trees to check. For n taxa, there are (2n-5)!! distinct trees that can connect them. In practice, the user always stops the algorithm early; checking all trees for any reasonable number of taxa is intractable.

Computational research so far has focused on two parts of the algorithm: choosing the next tree well, and the fast calculation of scores for each tree. Choosing the next tree well amounts to a smart sampling of tree-space. There is just not enough time to check every possible tree, but we can try to check those trees that we suspect have promise. The basic approach is to look at the trees that are similar to the best we have found so far. Work on the fast evaluation of trees has also helped but does not address the problem's enormous computational complexity.

Two practical troubles have surfaced as a result of the computational difficulty of phylogenetic tree inference. The first is a lack of confidence in results. The algorithms discussed above return the best tree(s) they have seen so far, but without checking all of the trees, it can't be known that better trees weren't passed over. There is no way to guarantee an optimal solution without performing a complete search. The second difficulty lies not in the size of the problems but in the size of the solutions. Current techniques often return a very large number of trees with the same (optimal so far) score. For example, in a parsimony search for a phylogenetic tree to connect 28 genera of sunflowers, Kim and Jansen (1995) found 8,235 equally parsimonious trees after a search of several weeks. Furthermore, the true tree may not in fact be one of the shortest. A researcher may ask that some sub-optimal trees be returned as well, further exacerbating this problem.

1.4 Dealing with a large number of returned trees

The trouble with getting so many trees back is in understanding them. One cannot simply flip through a thousand trees one at a time and get any idea about the true tree that they approximate. The most common method used for understanding a large number of trees is to calculate a consensus tree for the group. A consensus tree is a sort of average or lowest common denominator of trees. The most conservative is called the strict consensus tree; it contains all bipartitions that are present in every tree of the set under consideration. (See Figure 2.) A divides the leaf nodes of the tree into two distinct sets. A tree can be completely and uniquely specified by a list of the bipartitions it contains.

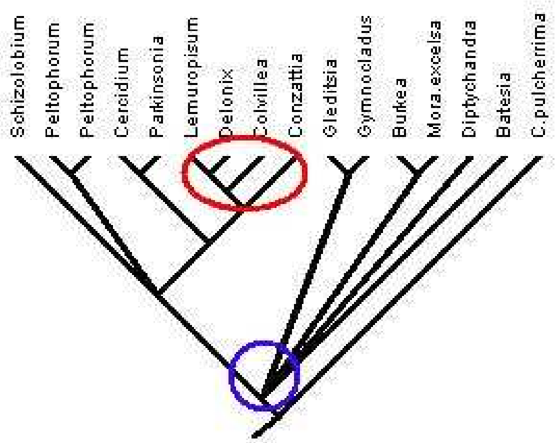


Figure 2 An example consensus tree. A consensus tree gives a simple summary of a set of trees. The region circled in red (upper circle) is fully resolved (all branchings are binary); all trees in the set have this same branching pattern. The blue circle (near the base) shows a polytomy; different trees in the set contain different branchings here. We computed this consensus tree with MacClade (Maddison & Maddison 1999) over trees found by Dr. Beryl Simpson (unpublished).

A consensus tree gives a useful but very rough summary of the content of a set of trees. My project explores new techniques for understanding large sets of phylogenetic trees, techniques based on the ideas of data visualization.

1.5 Data visualization background

Computers excel at the accurate computation and storage of numbers and have long since surpassed humans in this regard. When the goal of this computation is human understanding or insight, however, much of a computer's power is wasted. People are just not any good at integrating information from huge tables of information.

The ability of people to notice patterns or draw conclusions from data is dramatically increased if the data are presented visually (Ware 2000). A large part of our brains is devoted to visual processing. When we understand information visually, a huge amount of the processing is done automatically, below the conscious level, by neural machinery that is already hardwired for perceiving spatial relationships, contrasts in scale, clustering patterns, object definition, and so on.

However, because it was built to perceive the real world, there is an inherent limitation to the dimensionality of data that can be understood visually. The retina is after all only a two-dimensional input device, and the world our brains evolved to understand is three-dimensional. There is also a sort of primacy of spatial dimensions in visual perception. Perception is limited when a dimension is presented as variation along a scale of color, brightness, or shape

This limitation on the dimensionality of information that can be effectively understood by a human viewer presents a challenge to the visualization of a set of phylogenetic trees. The dimensionality of such data is extremely high. If trees are encoded using bipartitions, then tree-space has a binary-valued dimension for every possible bipartition in a tree, i.e. $2^{n-1} - 2$ dimensions. An approach that presents only a few dimensions at a time or that presents multiple views is simply insufficient. Instead, we had to look for a way to drastically reduce the dimensionality of our data while preserving the important information about it.

2 Methods

2.1 Project overview

I gathered data sets from professors in the school of biology who are working on systematics research: Dr. David Hillis, Dr. Robert Jansen, and Dr. Beryl Simpson. In order to construct visualizations, I computed the Robinson-Foulds distances between every pair of trees in each set and used an off-the-shelf MDS implementation (XGvis) to experiment and learn how useful the visualizations could be. After encouraging preliminary results and a positive response from the biologists, I implemented a more specialized MDS as part of a module for the Mesquite phylogenetic software system in order to make the visualizations generally available.

2.2 Multidimensional scaling (MDS)

We used multidimensional scaling to form pictures in the two or three dimensions to which people are accustomed. Multidimensional scaling is a technique for creating plots in R^k of data characterized by dissimilarities. For each pair of trees in a set, we can compute any of a number of distance metrics that tell us how dissimilar (or, equivalently, how similar) that pair of trees is. The most obvious such metric is called the Robinson-Foulds distance (Robinson and Foulds 1981), which is based on an encoding of trees with bipartitions. The Robinson-Foulds distance between two trees is the number of bipartitions that is present in only one of the trees. If C(T) is the set of bipartitions in tree T, then the R-F distance between trees t_1 and t_2 is given by

$$\left|C(t_1) - C(t_2)\right| + \left|C(t_2) - C(t_1)\right| .$$

In using multidimensional scaling, we discard information about each tree's absolute location in tree-space and instead only look at how far apart the trees are from one another. The goal is to assign each tree a location in R^2 or R^3 such that the distances between points in the plot resemble as closely as possible the corresponding distances between trees in tree-space. Similar trees (trees that agree closely on the evolutionary history under investigation) are drawn close together and dissimilar trees are drawn far apart.

We feel that this approach is justified for a few different reasons. Among the dimensions of variation for phylogenetic trees, there are not one or two that are more important than the others. MDS places all input dimensions on equal footing. Also, the judgments that need to be made regarding tree sets depend not on the sets absolute position in tree-space, but on the relative positions of trees within the set. Thirdly, we feel that multidimensional scaling is valuable because the dimensions along which phylogenetic trees vary are binary-valued, and we gain a lot of room by fitting them into a real-valued space.

2.2.1 Distortion in MDS

There are some sets of dissimilarity data that can be plotted in R^2 without any distortion; that is, the distances between points in the plane exactly match the given dissimilarities. In the typical case, however, a perfect embedding is impossible and some distortion in the distances must be introduced in order to draw all of the points in two dimensions. The quality of an embedding based on dissimilarity data can be judged by a stress function, which is simply a residual sum of squares (Buja et. al. 1998):

$$S_D(x_1,\ldots,x_n)=\left(\sum_{i\neq j=1\ldots n}\left(D_{ij}-\left\|x_i-x_j\right\|\right)^2\right)^{1/2},$$

where

D is the given dissimilarity matrix,
$(x_1,\ldots,x_n)$ are the point locations in the picture, and
S_D is the stress given the matrix D.

The bigger the stress function for a given embedding, the more the relationships between points in that embedding are distorted from their true separation. A Shepard diagram provides a nice visual representation of the errors in an embedding. It is a plot of the distances in the embedding against the distances given by the dissimilarity matrix (see Figure 3).

2.3 Implementation of MDS

My implementation of multidimensional scaling performs a heuristic iterative minimization of the stress function. The algorithm is:

1. Compute the distance between each pair of trees in the set. This is the target dissimilarity matrix.
2. Begin with a random assignment of point locations to trees.
3. For each pair of points, compute the vector difference between their current separation and their target separation.
4. For each point, find the vector sum of all the error vectors with which it is involved; this is the total residual “error” for that point—adding this sum to that point reduces the total error of the point’s position with respect to all the other points.
5. Scale the residual error vectors down. (By about 0.01—this is a user-defined parameter.)
6. Add the residual error vectors to the point locations.
7. Repeat Steps 3-6 until the user says stop.

The number of iterations required to reach a stable solution can vary quite a bit depending on the initial point locations. There is no guarantee that this algorithm will avoid local minima in the stress function, but we have run each embedding dozens of times and never ended up with more than one stable solution.

3 Results

3.1: Tree set visualizations

We used the MDS plotting algorithm to generate visualizations of several data sets obtained from researchers in the school of biology at U.T.

3.1.1 Islands and clusters in sets of trees

Maddison (1991) described the idea of islands of phylogenetic trees. An island is a set of trees connected under a simple branch rearrangement operation. Because the next tree to check in a tree search is often chosen by rearranging a good tree that has already been checked, islands can be found as a side effect of the tree search. Maddison observed that the best trees found in a search are usually all grouped into one island, but occasionally more than one island of most-parsimonious trees is discovered.

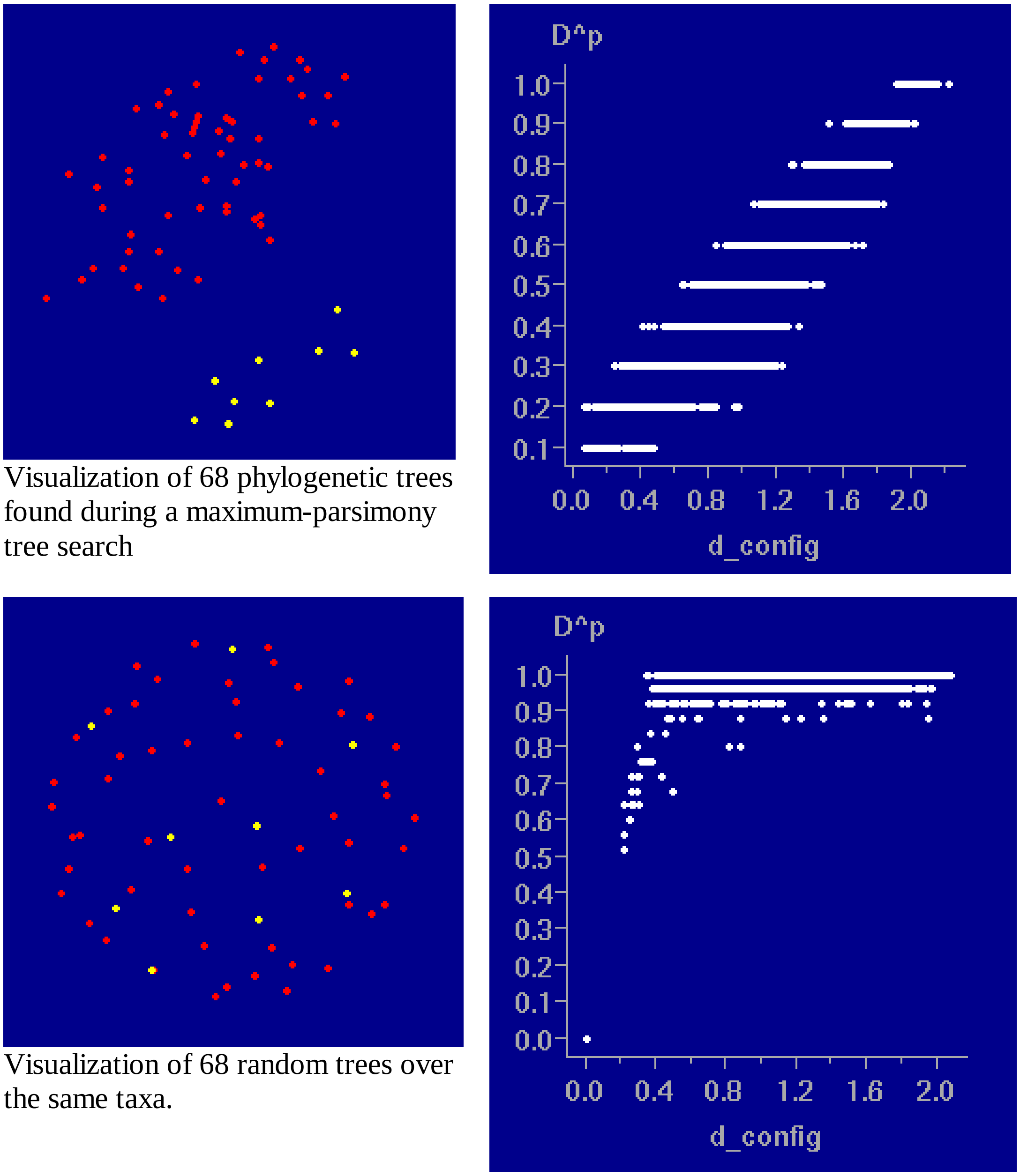
We were interested in knowing if islands within a set of most-parsimonious (or most-likely, etc.) trees were visually apparent in the pictures created by our visualization technique. In a study of several species of sunflowers, Jansen (1995) found more than 8000 equally parsimonious trees divided between two islands. We created visualizations for a sample of 68 of these trees and for 68 randomly generated trees over the same taxa (Figure 3).

We found that the islands discovered during the tree search were indeed visually separate from one other in the visualizations.

We also found that consensus trees computed for each island of trees in the picture were each more resolved than a consensus tree for the entire tree set (Figure 4). The clusters' consensus trees give a clearer idea of which taxa are contributing the most to a lack of resolution in the overall consensus tree and can help guide researchers in the collection of additional data to resolve the uncertain regions in the trees.

Our results suggest that islands are indeed a useful way of grouping phylogenetic trees, but that there is meaningful clustering of trees below the level of island. One island of trees in Jansen's data encompassed three distinct visual clusters. The consensus trees computed for each cluster within that island are more resolved than the consensus tree for the entire island and provide another level of detail in understanding the structure of the entire tree set.

Clusters of trees can be thought of as competing hypotheses regarding the evolutionary history of the taxa under investigation, with each cluster corresponding to a different evolutionary scenario. The trees within each cluster are often all minor rearrangements of one another and are all usually more similar to one another than they are to trees in other clusters. When a consensus tree is constructed for the entire tree set, the distinction between clusters is lost, and the small variation within each cluster is washed out by the larger variation between them.



Visualization of 68 phylogenetic trees found during a maximum-parsimony tree search

Visualization of 68 random trees over the same taxa.

Figure 3. Visualizations of a set of trees discovered during a maximum-parsimony search over several kinds of sunflower. Tree islands are indicated by color and are visually separated from one another. The Shepard diagrams on the right show the quality of the visualizations' embedding of the trees in 3D space. In a Shepard diagram, the horizontal axis indicates the distance between points in an embedding, and the vertical axis indicates the target distances. A straight line would correspond to a perfect embedding. The degree of divergence from a straight line shows how much the given distances are distorted in the picture. The horizontal banding in the Shepard diagrams reflects the fact that tree separation varies discretely. We used XGvis (Buja et. al. 1998) to create these visualizations.

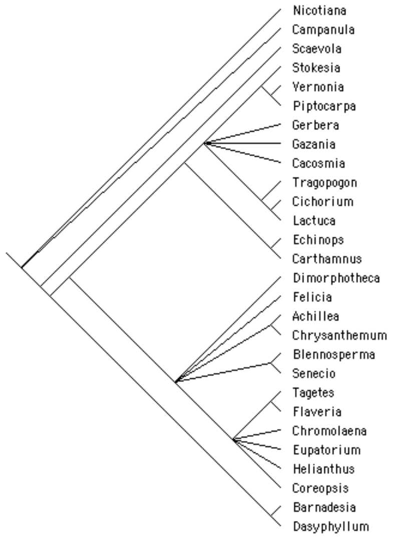
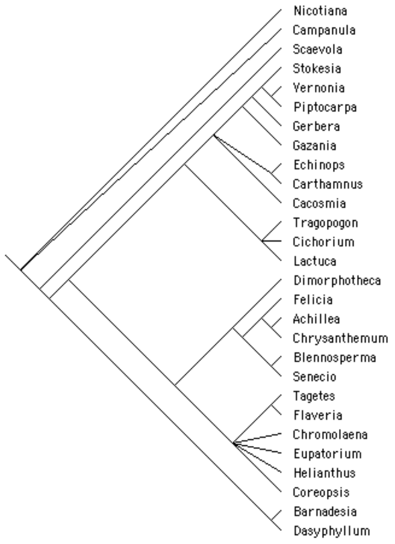
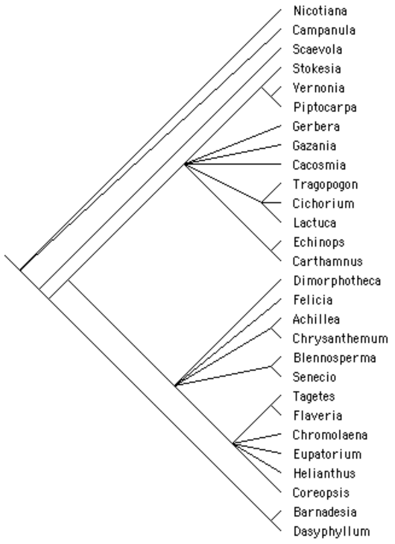
Strict consensus tree for the red island	Strict consensus tree for the yellow island	Strict consensus tree for the entire set
 Nicotiana Campanula Scaevola Stokesia Vernonia Piptocarpa Gerbera Gazania Cacosmia Tragopogon Cichorium Lactuca Echinops Carthamnus Dimorphotheca Felicia Achillea Chrysanthemum Blennosperma Senecio Tagetes Flaveria Chromolaena Eupatorium Helianthus Coreopsis Barnadesia Dasyphyllum	 Nicotiana Campanula Scaevola Stokesia Vernonia Piptocarpa Gerbera Gazania Echinops Carthamnus Cacosmia Tragopogon Cichorium Lactuca Dimorphotheca Felicia Achillea Chrysanthemum Blennosperma Senecio Tagetes Flaveria Chromolaena Eupatorium Helianthus Coreopsis Barnadesia Dasyphyllum	 Nicotiana Campanula Scaevola Stokesia Vernonia Piptocarpa Gerbera Gazania Cacosmia Tragopogon Cichorium Lactuca Echinops Carthamnus Dimorphotheca Felicia Achillea Chrysanthemum Blennosperma Senecio Tagetes Flaveria Chromolaena Eupatorium Helianthus Coreopsis Barnadesia Dasyphyllum

Figure 4. Consensus trees for the tree islands shown in Figure 3, computed using MacClade (Maddison 1999). Consensus trees computed for islands of trees are more resolved than the consensus tree for the entire set.

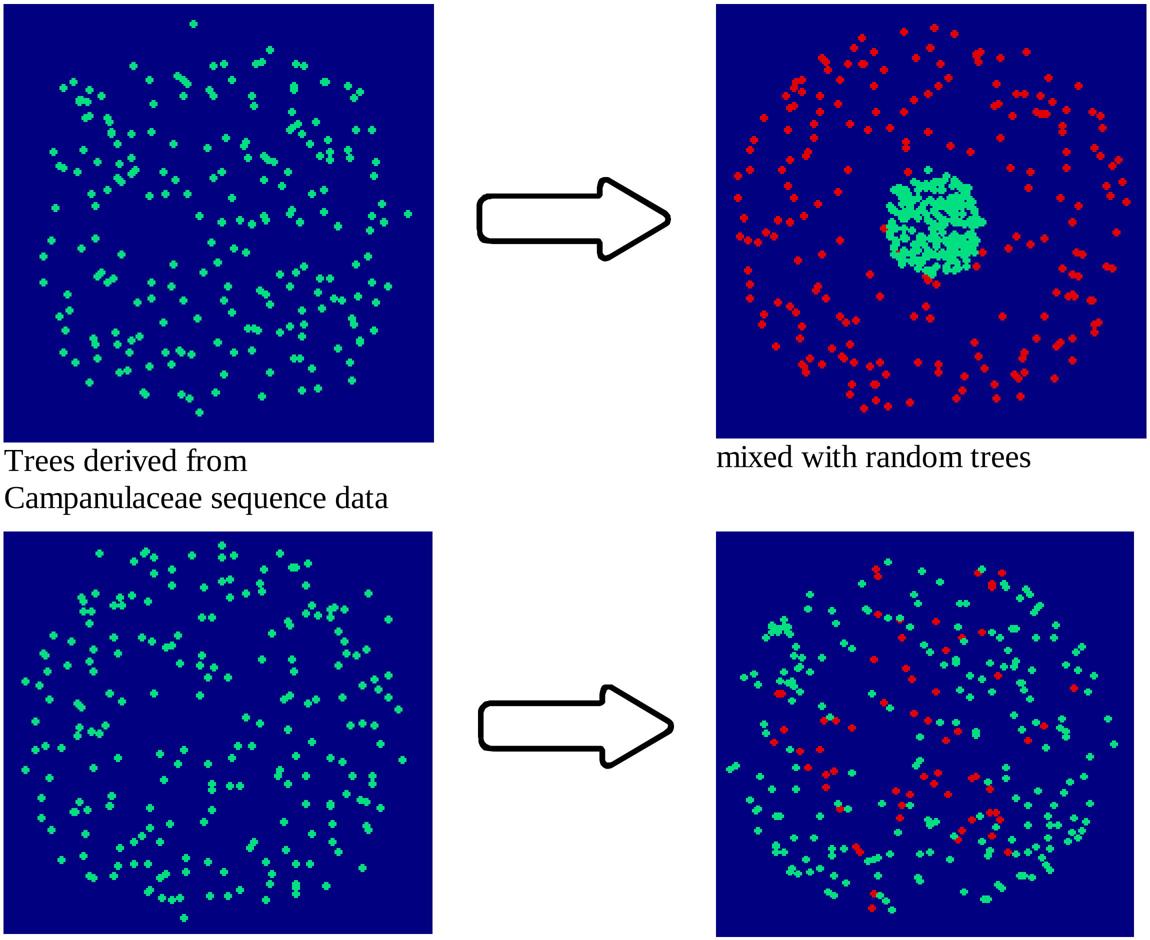
3.1.2 Visualizing tree density

A third case study shows the utility of our method to indicate the degree of tree agreement within a set of phylogenetic trees. An analysis of chloroplast DNA from several species of bellflowers resulted in 216 equally good trees. When we used our techniques to visualize these trees, we saw an undifferentiated ball similar to that seen for the random trees in Figure 3.

The undifferentiated ball indicates that there are no outstanding differences among the inter-tree distances in a group of trees. However, scale is ignored by our plotting method, and a very tight cluster of trees would appear as an undifferentiated ball if it were the only thing in the picture. The situation might be that you have a lot of variation among the trees of the set (All the inter-tree distances are large and more or less equal) or that you have very little (All the inter-tree distances are small and more or less equal). Both give the same picture. In the first case, the trees represent very different evolutionary histories for the taxa under investigation, and in the second, the trees agree closely.

In order to distinguish between these two cases, we added random trees to both kinds of set and repeated the visualization procedure (Figure 5). In the case of the trees that agreed closely with one another, the trees of interest contracted into a tight ball at the center of the picture. In the second case, the undifferentiated ball remained largely the same, with the trees of interest scattered uniformly throughout.

This result showed us that it is a good idea to include some indication of scale with the picture, either by adding random trees as we did in this experiment or accompanying the image with



Trees derived from Campanulaceae sequence data

mixed with random trees

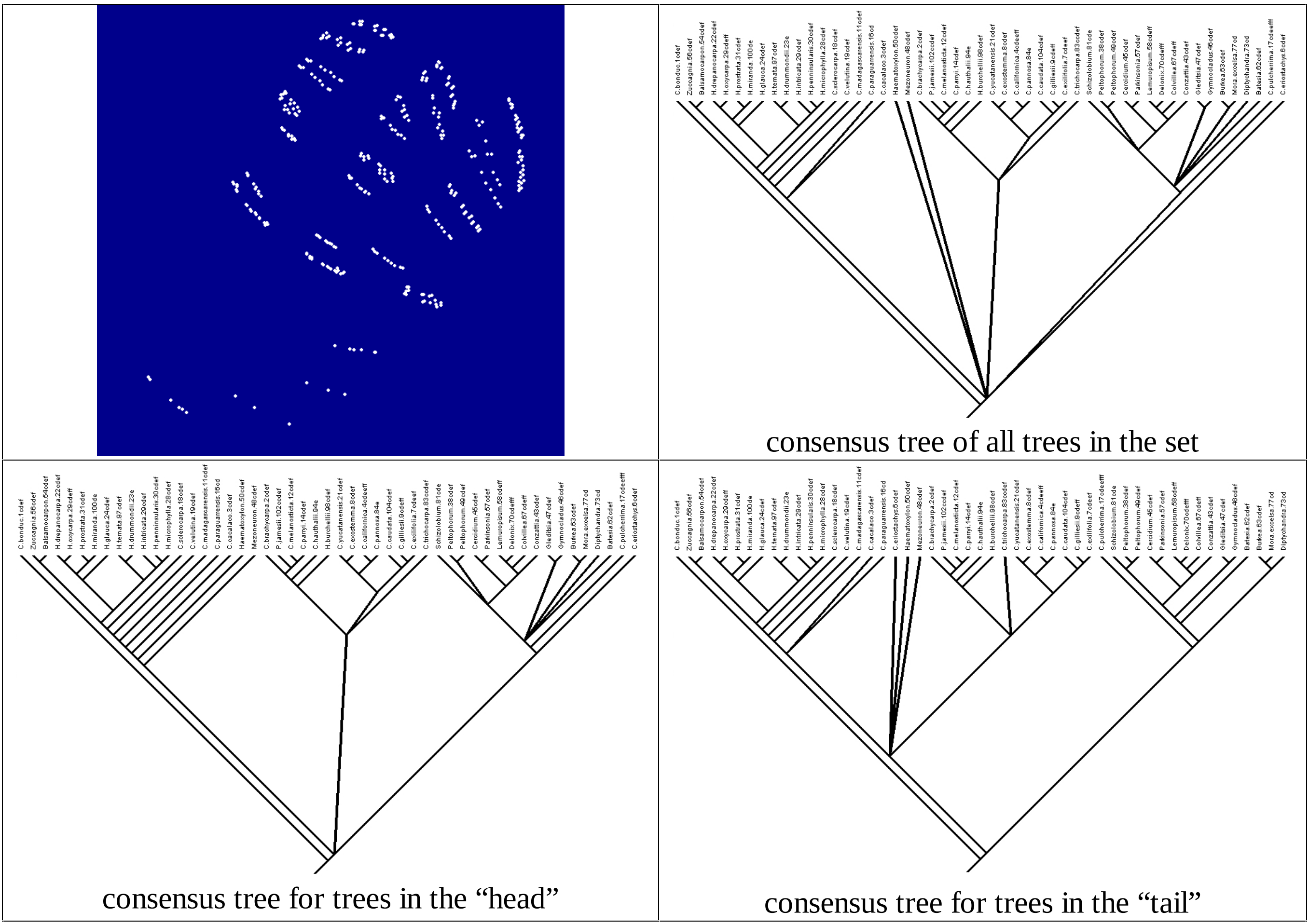
random trees

mixed with random trees

Figure 5. An demonstration of a method to distinguish between two types of tree set that appear similar under our visualization technique. By adding random trees to each set, we were able to distinguish between a group of trees that agrees closely and one that does not.

3.1.2 Another example embedding

In another example of the visualization of a tree set, we generated pictures of a group of trees that arose in research being performed on several bushes in the Caesalpinia genus by Dr. Beryl Simpson of U.T.'s botany department. She provided us with 342 equally parsimonious trees constructed over 51 taxa. Figure 6 shows an example picture along with some sample consensus trees.



consensus tree of all trees in the set

consensus tree for trees in the "head"

consensus tree for trees in the "tail"

Figure 6. A 2D embedding of 342 phylogenetic trees over the genus Caesalpinia with some sample consensus trees. As was the case for our first example with sunflowers, consensus trees of visual partitions of the picture are more resolved than the overall consensus tree.

3.2 Using MDS to visualize inter-taxon distances

Multidimensional scaling is applicable to any kind of difference data, not only difference data for phylogenetic tree sets. Differences between pairs of species can be computed directly from molecular sequence or morphological data. We can then apply our visualization technique to create pictures in which the points correspond to taxa rather than to trees.

3.2.1 Visualization of a broad set of taxa

Figure 7 shows a picture of a wide variety of species drawn in this way. The species differences were based directly on sequence differences in mitochondrial DNA common to all of the species. Such a picture can only show gross differences between species and is not useful in revealing small evolutionary separations, because homology (convergent evolution) can cause a pair of taxa to appear more closely related than they really are. This is why techniques like parsimony are used to measure the true degree of divergence between species.

This kind of picture *is* useful, however, in getting an overall idea of the degree of evolutionary divergence among a group of species. In a phylogenetic tree, information on the degree of divergence can only be shown clumsily, by using labeled branch lengths. A combination of a phylogenetic tree and an MDS-generated picture based on species differences can convey a more integrated understanding of the evolutionary past of a group of taxa.

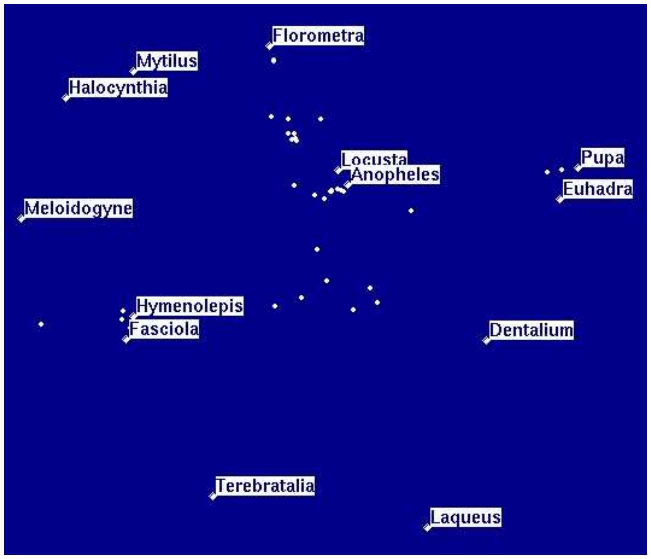


Figure 7. Visualization of taxa differences. Each point in the picture represents a single taxon. Similar taxa are drawn close together. A few taxa are labeled.

3.2.2 Direct visualization of taxa differences as a guide to tree reconstruction

As stated above, the biggest computational obstacle to phylogenetic reconstruction is the vast number of phylogenies to consider. The early identification of groups of taxa known to be related allows the problem to be broken up, drastically reducing the amount of computation necessary for a complete reconstruction.

In a demonstration to show how our visualization techniques can be useful in this regard, we set up some synthetic model evolutionary trees. We simulated the evolution of a gene sequence along the model trees and then visualized the resulting sequence differences (Figure 8). Monophyletic groups of species (a common ancestor species and all of its descendants) clustered clearly in the pictures.

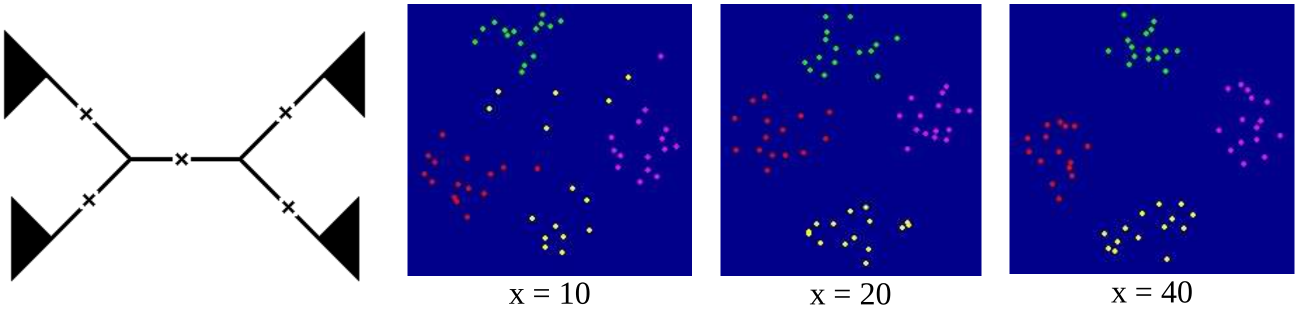


Figure 8. Visualization of species differences. The difference data were computed based on modeled evolution along the synthetic evolutionary tree shown on the right. This tree has long internal edges of length x that separate its taxa into four monophyletic groups. The branch lengths within each of the four sub-trees are fixed at four. A longer branch length corresponds to greater evolutionary separation.

Because the number of trees that can organize each sub-group of species is much smaller than the number of trees that can organize all of them together, a successful divide-and-conquer approach to phylogenetic reconstruction is possible in some situations. Tree reconstruction is undertaken on each group in isolation, followed by a search for the best way to connect the sub-trees (Figure 9). This is what is done on a very large scale when a scientist looks for a phylogenetic tree to organize salamanders, for example, but does not include any pine tree DNA in his data. It is not clear on how small of a scale it is possible to use divide-and-conquer as a phylogenetic reconstruction technique.

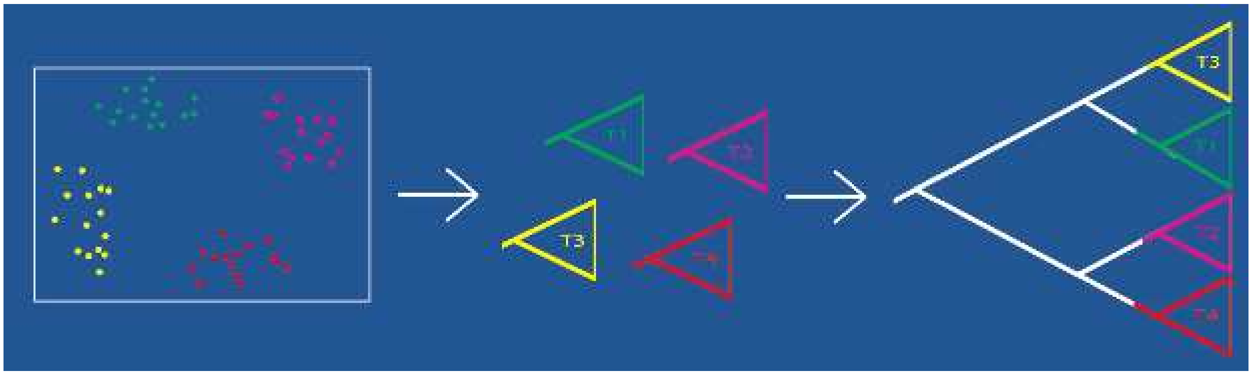


Figure 9. Divide-and-conquer method for phylogenetic tree reconstruction. First, monophyletic groups of taxa are found by clustering. Second, a tree search is conducted for the best tree to assemble each of the groups. Finally, another search is used to find the best way of connecting the sub-trees.

Such partitioning of taxa may be very useful to phylogenetic reconstruction, but it is not clear whether visualizations are very helpful in this regard. For this divide-and-conquer method to be correct, one must be assured that the sub-groups are each monophyletic, and the degree of clustering required for confidence in this fact is such that it probably could just as easily be recognized by automatic clustering procedures as by human viewers. On the other hand, visualization may reveal subtleties of the distribution that raw clustering would hide.

4 Software Tools

Visualization in phylogenetics is a new idea, so there isn't any software currently available for it. We did our experimentation and prototyping for the visualizations by using many different pieces of existing software and tying them together with simple scripts to handle data flow and formatting. Initial tree searches were done by various researchers in the school of biology using software designed for that purpose: PAUP* (Swofford 2000) or MacClade (Maddison 1999). We used a set of programs written by Daniel Huson (2000) to compute the inter-tree distances and to generate random trees. Finally, we used the generic data visualization package of XGvis (Swayne et. al. 1998) and XGobi (Buja et. al. 1998) to perform MDS on the difference data and plot and explore the resultant point sets. I wrote several simple programs to handle data flow between these programs and to translate data among the various required formats.

A goal of this project is a software tool that can be used by biology researches to explore their phylogenetic data sets visually. To this end, I have written an integrated program that ties the whole process together. I wrote this program as a module of the larger Mesquite system.

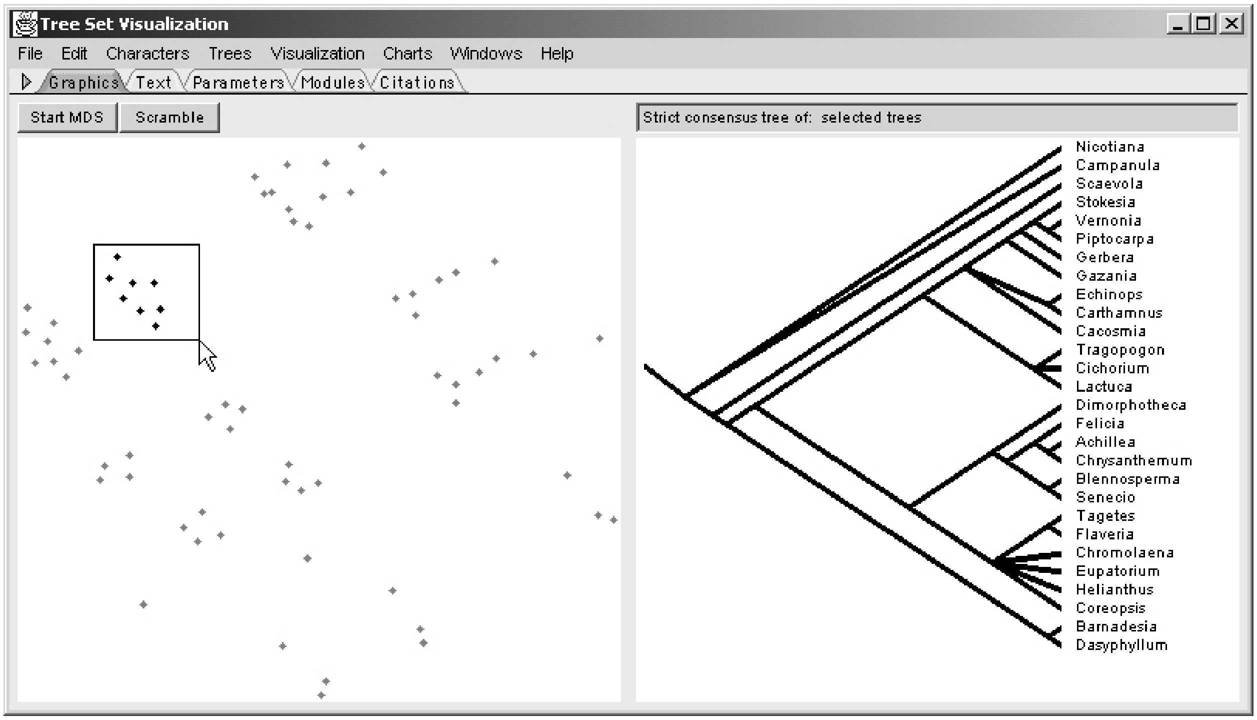
4.1 What is Mesquite?

Mesquite is a modular software system for phylogenetic analysis currently under development at the University of Arizona (Maddison and Maddison 2001). Mesquite modules already exist to perform all of the traditional phylogenetic computations: organization of sequence and data, tree to researchers using many different computer platforms. Mesquite is new and not as well established in the systematics world as some older packages (PAUP* and MacClade), but its modular nature made it easier for me to add my work to a larger piece of software and distributing the visualization module will give me the chance to have my ideas tried out in a real research environment. We do expect that Mesquite will become very popular for use in phylogenetic analysis.

4.2 My Mesquite module

My software adds tree set visualization as an integrated piece of a large phylogenetic analysis package (Figure 10). A user can take any group of trees with which he is working and create a 2D plotting of the trees based on Robinson-Foulds distances, or, because the Mesquite system is modular, any other distance metric for which a module has been written. My module will be distributed freely as a plug-in module for Mesquite. I will be available along with the main Mesquite software from Drs. Wayne and David Maddison at http://mesquite.biosci.arizona.edu/mesquite/download/download.html.

My program is interactive. The MDS algorithm is animated so that the user can watch and understand what is being done and can decide when a plotting has been reached that is good enough. Points in the embedding are selectable. When a single tree is selected, that tree is displayed, and when multiple trees are selected, their consensus tree is displayed. This connection between the point set display and the tree display allows effective exploration of the tree set and aids in the understanding of its structure.



5 Discussion

Success in data visualization is notoriously difficult to measure or evaluate. One cannot take an insight of a researcher and ask her, “Would you have realized this without the visualization?” Likewise, we can never be sure if a scientist gains more understanding from our pictures of tree sets that he would have with only the individual trees in the set and their consensus.

However, we are encouraged by the responses of the biologists with whom we have worked. There is excitement about the idea, and it is clear that a useful tool for understanding the results of phylogenetic analyses is needed. There is reason to have confidence in our technique. We have found a correspondence between the traditional logical structure of a tree set given by islands and the visual structure of our images. Generally, what is seen in the pictures has triggered associations with knowledge that a viewer already had about a set of trees. Only time will tell how useful this new method is, but our early results are encouraging, and further investigation is warranted.

Very large data sets and heavy computer analysis, though rare in the past, have recently become a common feature of biological research. Advances in automated experimental techniques, especially with regard to gene sequencing and biological imaging, have given rise to the fields of computational biology and bioinformatics. The data are out there. Visualizing phylogenetic tree sets is only a small example of the opportunities that abound for the application of existing and novel data visualization ideas in the field of biology.

6 Future Work

No software tool or research project is ever complete, of course. The methods of visualizing sets of phylogenetic trees need to be explored more thoroughly. There are many additional features that would improve my Mesquite module and, there is a lot of unexplored potential in the area of biological data visualization generally.

The Robinson-Foulds metric is a convenient and, we believe, meaningful measure of tree dissimilarity to use in constructing MDS-based visualizations. It is not without problems, however, as Penny and Hendy (1985) showed. Other distances metrics should be investigated in order to determine their utility in visualization construction. Alternatives to MDS in these visualization techniques also deserve attention; multidimensional scaling is only one way of mapping a point set into two or three dimensions;

There are many improvements and additional features that I think would improve my Mesquite module. Firstly, I think that 3D embeddings should be included with the 2D. Pictures in 3D can be rendered by MDS with less distortion than in 2D, and I think that a lot of the structure we’ve seen in our prototype 3D visualizations just isn’t captured in the corresponding 2D embeddings. Also, another link between the tree view and the embedding picture should be created. A user should be able to select taxa or edges in the tree view and restrict the MDS calculations to tree variation that involves those taxa or edges. This feature would allow better understanding of the connection between the variation in phylogenetic trees and the ambiguity in the data from which it arises. Finally, and more ambitiously, I would like to connect the tree set visualizations to the tree search routines, so that a tree search could be steered by the researcher and better trees could be found in the first place.

A big visualization challenge in biology is the "tree of life," a phylogenetic tree organizing *all* species. Even when extinct species are ignored, this is a truly enormous tree, with as many as ten million leaves, and creating a comprehensible visual representation of if will be a big challenge. There is also a lot of room for further investigation of tree set visualizations. Our work could be extended greatly through the exploration of other way of mapping trees into two or three dimensions. Multidimensional scaling is only one way of doing it, and it would be interesting to try other point assignment algorithms or additional tree comparison ideas.

7 Acknowledgements

I am very grateful to many people who worked together with me on the project. My supervising professor, Nina Amenta, has been generous with her time, advice, and understanding. She suggested this research direction to me in the first place, and her support over the past year has been invaluable. I also thank Tandy Warnow, for teaching me about the mathematics and algorithms and unrooted trees and for always brainstorming and throwing new ideas at us to keep us on our toes. Dr. Warnow's student, Li-San Wang, and Katherine St. John, a visitor from City College in New York, assisted with some side investigations along the way. Although he didn't work with me, I am indebted to Daniel Huson for his work on some of the tree software utilities I used in prototyping; they saved me a lot of time. I would like also to acknowledge Bob Jansen and Beryl Simpson for providing real data sets on which I could experiment and David Hillis for consultation on general systematic issues and for his enthusiasm in support of the project. Finally, I thank Drs. Wayne and David Maddison for the creation of Mesquite, without which my hopes of building a complete and usable implementation of this research would have been unfulfilled.

Bibliography

Buja, A., D. F. Swayne, M. Littman, and N. Dean. 1998. XGvis: interactive data visualization with multidimensional scaling. Tentatively accepted for publication in the Journal of Computational and Graphical Statistics.

Huson, Daniel H. 2000. The Tree Software Package.

Kim, Ki-Joong and Robert K. Jansen. 1995. ndhF sequence evolution and the major clades in the sunflower family. Proceedings of the National Academy of Sciences of the USA. Vol. 92, pp. 10379-10383.

Livingstone, M. and Hubel, D. 1988. Segregation of form, color, movement and depth: anatomy, physiology, and perception. Science, 240, 740-749.

Maddison, David R. 1991. The discovery and importance of multiple islands of most-parsimonious trees. Systemic Zoology, Vol. 40(3). pp. 315-328.

Maddison, W. P. and D. R. Maddison. 1999. MacClade 3.08. Interactive analysis of phylogeny and character evolution. Sinauer Associates, Sunderland, Massachusetts.

Maddison, W. P. and D. R. Maddison. 2001. Mesquite: a modular system for evolutionary analysis. Beta test Version 0.95d80. http://mesquite.biosci.arizona.edu/mesquite/mesquite.html

Penny, David and Hendy, M. D. 1985. The Use of Tree Compariuson Metrics. Systematic Zoology, 34(1):75-82.

Robinson, D.F. and L.R. Foulds. 1981. Comparison of phylogenetic trees. Math. Biosci., 53:131-147.

Swayne, D. F., D. Cook, and A. Buja. 1998. XGobi: interactive dynamic data visualization in the X window system. Journal of Computational and Graphical Statistics. 7(1).

Swofford, D. L. 2000. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Ware, Colin. 2000. Information visualization: perception for design. Academic Press.