# Phylogeny Visualization: iMap, an Interactive Heatmap

Kathleen Timmerman Tiffani Williams and Suzanne Matthews

***Abstract*—Phylogeny studies how organisms have evolved over time by combining the organisms into a phylogenetic tree. Branching in a tree represents an evolutionary event leading to two species. Visualization is an important part of phylogeny. Phylogenetic applications output so much data that it is hard for humans to interpret it manually. With visualization tools, people can draw conclusions about large amounts of data instantaneously. Heatmaps are a colorful representation of ranges of values that allow for the visual comparison of multiple attributes simultaneously. Current heatmap programs create a static representation of data. iMap allows users to interact with their heatmap, creating a customized viewing experience.**

**iMap takes as input a Robinson Foulds (RF) distance matrix of n x n, where n is the number of trees being compared. Each cell in a matrix is assigned a value based on how different two trees are from each other (a higher number indicates more differences, 0 indicates they are identical). Current heatmap visualizations of the RF distance matrix may list the trees arbitrarily. The interactive features in iMap allow the user to move the columns and rows to see if a pattern emerges. It will also allow users to zoom into specific areas of the map. iMap also has a clustering feature to locate similar trees. By using iMap, users can create customized heatmaps to better understand their results, and ultimately, communicate their findings to other researchers.**

***Index Terms*—Heatmap, Interactive, Phylogenetic, Visualization**

## I. Introduction

PHYLOGENY studies the evolutionary history of organisms. The ultimate goal of phylogeny is to create the tree of life. The tree of life would consist of a rooted phylogenetic tree that contains all living organisms, which is estimated to be between 5 and 100 million [2]. A rooted phylogenetic tree considers time as well as mutations as important. The root of the tree is the oldest organism and as the taxa get farther from the root the more they have evolved from that organism. Every time an evolutionary event happens the phylogenetic tree branches, representing the creation of a new species.

Phylogeny is constantly balancing the amount of time spent searching for most likely situation versus confidence that the answer found is the best answer. While all the answers can be looked at with a small number of taxa, the option of using the exhaustive method quickly disappears because the number of possible rooted trees, t, is equal to

$$t = (2n - 5)!!$$

where n is equal to the number of taxa.

Humans are very good at interpreting massive amounts of data when it is visually represented. By locating patterns in the way the best tree is located, several steps in the process of locating a really good tree might be able to be skipped, decrease the time spent finding it. Everytime the amout of time is needed to accomplish locating a really good tree, Phylogenists come closer to reaching their ultimate goal. For that reason, visualization programs have been created to view trees. One such item used to view trees is a heatmap.

## II. Background

### A. What is a Heatmap?

Heatmaps are a colorful representation of ranges of values that allow for the visual comparison of multiple attributes simultaneously. Items along one axis of a heatmap can be compared to items along the other axis. The comparison is numerical value that is assigned a color. Although three dimensional dimensional heatmaps do exist, in which case each spot on the map has a color and a height representing two different values, two dimensional heatmaps are much more common and what iMap uses. Since a value is represented by a color, a thirteen digit number can be condensed down to just a few pixels on the screen in a heatmap. This ability to condense down large amounts of data into a realitively small yet easy to interpret area without lossing any of the information is what makes heatmaps so important and powerful [5].

### B. What is a Phylogeny Heatmap?

Systematists often need to interpret a large set of phylogenetic trees. In order to do this, it is common to make a consensus tree. However, in doing so, much information about the sample being examined is lost [1]. Heatmaps can help solve this problem by providing a practicle way of comparing numerous trees at once.

In Phylogeny a heatmap is used to compare the distances between multiple trees at once by assigning a color to each value in a Robinson Foulds (RF) distance matrix, a t x t matrix where t is the number of trees being compared. The number in the matrix is based off of how different the two trees are.

#### B.1 Tree Bipartions

To determine the number that should be placed in the distance matrix, trees are bipartitioned and the bipartitions are compared. A tree is bipartitioned by removing each inner branch individually and listing the taxa on one side, a bipartition '|', and the taxa on the other side. For example, in Figure 1, if the red branch is removed from Tree X, the bipartition is 'AB |CDE'. Likewise if the blue branch is removed, the bipartition is 'ABE |CD'. The same technique can be used on Tree Y resulting in the following bipartitions: 'AE |BCD' and 'ABE |CD'. Then to determine how different two trees are, the set of bipartitions from each is subtracted from the set of bipartitions from

the other. The mean of the two values is found and used in the matrix. In this example, The set of bipartitions in tree X, not in tree Y is one, since they both have 'ABE |CD', but only tree X has 'AB |CDE'. Likewise the bipartitions in tree Y not in tree X is one, since only tree Y has the bipartition 'AE |BCD'. Therefore the mean is one and that is the number that would be placed in the distance matrix, because that is the number that represents the distance between them in tree space.

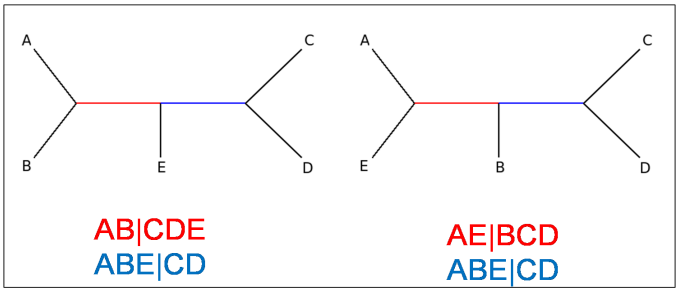


Fig. 1. Tree X and Tree Y Bipartions

### B.2 Robinson Foulds Distance Matrix

The RF distances matrix is comprised of a t x t matrix that compares each tree to every other tree in the set. So from the above example, where tree X and tree Y meet in the matrix, a one can be found. In Figure 2, where tree A is compared to tree B the value is two. That indicates that when A and B were bipartitioned, the difference of the sets was taken, then the mean was calculated, and the resulting number was two. A higher number indicates that two trees had fewer bipartions in common and are, therefore, farther apart in tree space.

To minimize steps for the user, iMap will create an RF distance matrix by running a tree file through HashRF. HashRF runs in $O(nt^2)$, making it one of the fastest programs for creating RF distance matricies [3].

| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0 | 2 | 1 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 1 | 5 |
| B | 2 | 0 | 1 | 5 | 4 | 5 | 4 | 5 | 4 | 5 | 1 | 5 |
| C | 1 | 1 | 0 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 0 | 5 |
| D | 5 | 5 | 5 | 0 | 1 | 1 | 2 | 1 | 2 | 0 | 5 | 1 |
| E | 5 | 4 | 5 | 1 | 0 | 2 | 1 | 2 | 1 | 1 | 5 | 2 |
| F | 5 | 5 | 5 | 1 | 2 | 0 | 1 | 1 | 2 | 1 | 5 | 1 |
| G | 5 | 4 | 5 | 2 | 1 | 1 | 0 | 2 | 1 | 2 | 5 | 2 |
| H | 5 | 5 | 5 | 1 | 2 | 1 | 2 | 0 | 1 | 1 | 5 | 0 |
| I | 5 | 4 | 5 | 2 | 1 | 2 | 1 | 1 | 0 | 2 | 5 | 1 |
| J | 5 | 5 | 5 | 0 | 1 | 1 | 2 | 1 | 2 | 0 | 5 | 1 |
| K | 1 | 1 | 0 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 0 | 5 |
| L | 5 | 5 | 5 | 1 | 2 | 1 | 2 | 0 | 1 | 1 | 5 | 0 |

Fig. 2. A Robinson Foulds Distance Matrix: The examples in this paper are based on this matrix.

### B.3 Phylogeny Heatmap

Once the entire distance matrix is filled out, each value can be assigned a color. This colored representation of the the distance matrix is the phylogeny heatmap. It can be read by locating the tree on the verticle axis and another tree on the horizontil axis, locating where the two trees intersect, and comparing that color to the key to see how far apart the two trees are in tree space. For example, in Figure 3 where tree A and tree C intersect is a bright blue color. According to the key, this is a distance of five indicating that the two trees are not close together. To display the greatest contrast, the colored key goes from the lowest distance to the highest distance, rather than using an absolute scale.

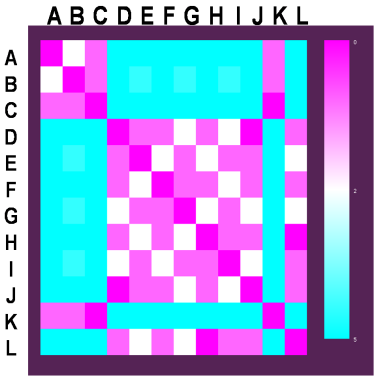


Fig. 3. A 12 x12 Phylogeny Heatmap

### C. Current Phylogeny Heatmap Programs

Current phylogeny heatmap programs will display the heatmap based off the distance matrix, but the diagram might be static. Other programs do not have intuitive user interfaces 'R'. Some programs, such as Spotfire can become expensive or are not specific to phylogeny [4]. iMap is a program specifically designed for viewing phylogeny trees. It provides a friendly user interface and also allows the user to interact with the heatmap. The user can zoom into an area to see it in greater detail, choose a tree to emphasized be reordering the rows and columns so that the trees are ordered according to how close they are to the cosen tree, and cluster the entire tree structure.

## III. Methods

iMap was designed to create an interactive heatmap specifically for phylogenetic research that had a friendly user interface.

### A. Programming Language

iMap was created in the processing language. Processing was chosen due to its ability to quickly and easily draw quality graphics. Processing is based on Java and is therefore an object oriented programming language.

### B. Zoom Feature

The zoom feature allows users to zoom into any one area of the heatmap to see that section in greater detail. This feature allows researchers to see areas of interest in greater detail and allows them to emphasize an area to other researchers. The user just specifies the upper-left hand cell and the lower right hand cell that they want to still be visable.

In Figure 4, the comparison of trees A, B, and C to themselves is zoomed in so that the area can be seen in more detail. From the detail, it becomes obvious that trees A and B are closer to tree C than they are to each other.

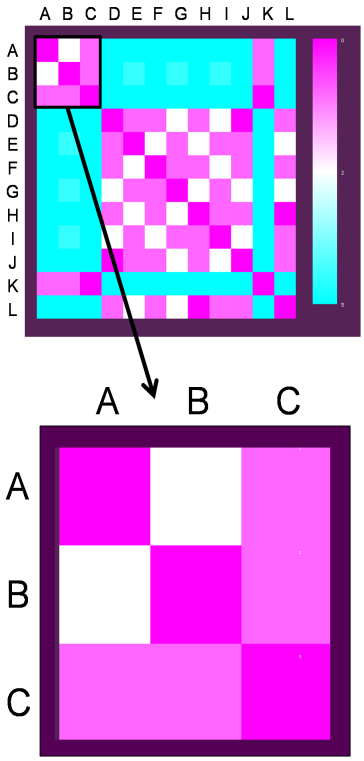


Fig. 4. A zoomed heatmap of trees A, B, and C

### C. Tree Emphasis Feature

The tree emphasis feature allows the user to choose a tree and have all the other trees ordered from most like it to least like it. The row of the chosen tree is moved to the top. Then all the columns are rearranged so the values go from smallest to largest. Finally, the remaining rows are ordered to match the order of the columns. In a way this is similar to clustering, but clustering considers how each tree compares to all the other trees to find out what trees are most alike. This feature only cares about how each tree compares to the tree of interest.

In Figure 5, the emphasis is place on tree G. Tree E, being most like G, is placed next,, and so on. The first column and row, will always show colors in the order of close together to far apart. This could be useful in looking at how a neighborhood of trees compares to the tree chosen for the next generation (See Results and Discussions).

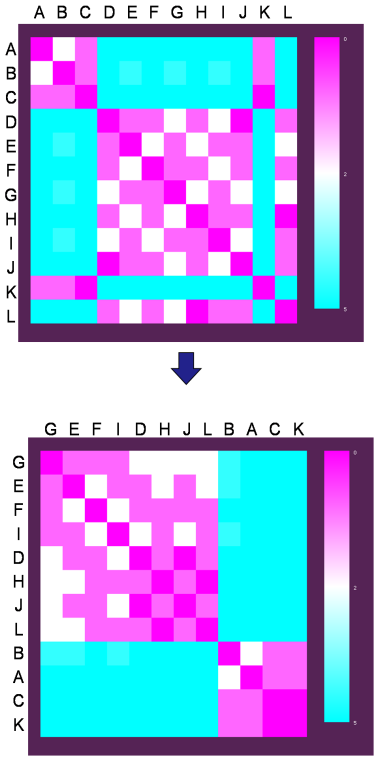


Fig. 5. A 12 X 12 Matrix with a emphasis on tree G

### D. Clustering Feature

The Clustering Feature compares each tree to every other tree. It reorganizes the trees so that trees that have the smaller distances between them appear next to each other. This is done by finding the smallest value in the distance matrix (ignoring the values of a tree compared to itself), and pairing the two trees associated with that value together. Then the distance between those two trees to every other tree is averaged, and the paired trees are treated as a single tree. Take the situation where trees E and G are paired. E is a distance of one away from J, and G is a distance of two away from J. So the tree pair EG would now be considered a distance of 1.5 away from J. This processes gets repeated until all the trees are connected to all the other trees.

In Figure 6, Trees Following Trees were most alike and therefore, placed next to each other: CK, JD, EG, and HL. The other trees were then added based on their closeness to these pairs, in the process described above. The lines connecting the labels create a phylogenetic tree describing the closeness of the pairs.

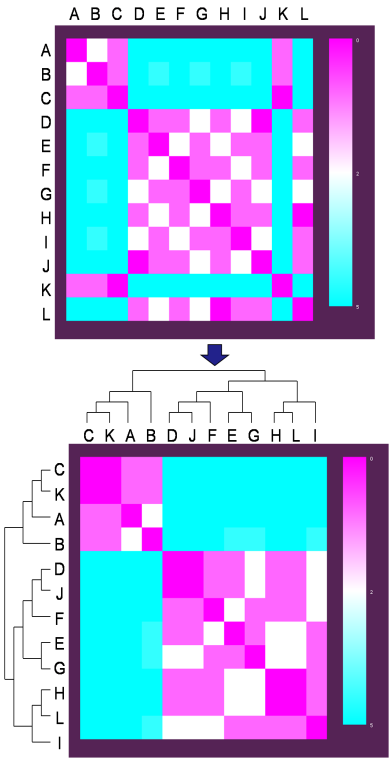


Fig. 6. A Clustering of a 12 X 12 Matrix

## IV. Results and Discussions

The interactive features in iMap will help research discover patterns that might have otherwise been unclear to them or a pattern that they had no way of observing before iMap. Below are some ideas on how the program might be used to discover some new pattern or build upon the existing structure.

### A. Neighborhoods

iMap allows the user to emphasize a specific tree. This feature could be used when looking at neighborhood to study how the tree chosen for the next generation compares to the other trees. Is it an outlier? Is it centrally located in the tree space explored? A better understanding of how the tree chosen compares to the other trees explored could lead to a pattern that could decrease the number of trees viewed to get the same results. This would quicken the search for the best tree, which is incredibly important in phylogeny research.

### B. 3D heatmap

iMap could be built upon to include a 3D feature that showed additional features. For example distances could be based on the height and then the color could tell if the tree included a specified bipartion.

## V. Conclusion

iMap is a program specifically for using heatmaps to study phylogeny. It interactive features will benefit researchers in discovering new patterns, seeing details more clearly, and sharing their discoveries with others.

## References

[1] Analysis and visualization of tree space. *Syst Biol 54*, 3 (2005), 471–82.

[2] of Life Web Project, T. What is phylogeny? *www.tolweb.org* (2004).

[3] Sul, S.-J., and Williams, T. L. An experimental analysis of robinson-foulds distance matrix algorithms. In *ESA '08: Proceedings of the 16th annual European symposium on Algorithms* (Berlin, Heidelberg, 2008), Springer-Verlag, pp. 793–804.

[4] Tibco. Spotfire webstore. *http://spotfire.tibco.com/Webstore/* (2009).

[5] Wilkinson, L., and Friendly, M. The history of the cluster heat map. *The American Statistician 63*, 2 (2009), 179–184.