Interactive visualization software for exploring phylogenetic trees and clades

Mark Derthick*

Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, 15213, USA

Received on November 8, 2007; revised on January 23, 2008; accepted on January 24, 2008

Advance Access publication February 8, 2008

Associate Editor: Martin Bishop

ABSTRACT

Summary: The Summary Tree Explorer (STE) is a Java application for interactively exploring sets of phylogenetic trees using two coupled representations: a node-and-link diagram and a textual list of common clades. Selection, pruning, filtering or re-rooting in one representation is immediately reflected in the other. While summary trees are more effective at showing the relationship among clades, they can only show a consistent subset of those that appear in the textual list. Working with both representations mitigates the disadvantages of having to choose just one.

Availability: STE, along with several sample datasets, is available at http://cityscape.inf.cs.cmu.edu/phylogeny/

Contact: mad@cs.cmu.edu

1 INTRODUCTION

Modern statistical methods can generate thousands or millions of distinct possible phylogenetic trees based on genomic data. It is difficult to provide an intuitive understanding of the relative frequency of so many trees, or even of subsets of them sharing some property of interest. Although innovative visualizations of ‘tree space’ are the focus of some research efforts (Amenta and Klingner 2002), the most commonly used summary representations continue to be summary trees and lists of common clades. The most common summary tree is the consensus tree, which includes the most commonly occurring clades in the tree set. A clade is the set of taxa descended from a single common ancestor (or for unrooted trees, a partition of the taxa separated from the rest by a single internal tree node). There is always a single unambiguous consensus tree that includes all clades occurring more often than some threshold, as long as the threshold is above 50% of the trees (‘Majority Rule consensus trees’). Thresholds below 50% may also be used, but there can be multiple incompatible clades meeting the threshold, so choices must be made. The usual rule is to greedily add clades in decreasing order of frequency, as long as they do not conflict with a previous clade. A consensus tree may be binary (‘fully resolved’), but in general will have internal nodes with multiple (unresolved) children. A sorted list of common clades includes all clades that occur more frequently than some threshold, including mutually inconsistent ones. We are not aware of another application that interactively couples a summary tree with a common clade list.

2 FEATURES AND ALGORITHMS

STE supports exploration through ‘what if’ questions based on the clades. For instance, Figure 1 shows STE applied to a set of animal species including 15 Cetaceada taxa (hippopotamuses, dolphins and whales). In a particular analysis, a user may be interested only in these species, and the others serve only as an outgroup to root the subtree.

First, the desired outgroup is selected (a). Pressing the ‘Set Outgroup’ button re-roots the tree (b). Pressing ‘Collapse Selected’ abbreviates the outgroup (c).

In (b), a whale clade of interest {6–13} has been colored in order to keep track of it. (It would be most natural to use one color here, and to do it after collapsing, but this order reduces the number of figures.) What about alternate phylogenies for this clade? In (c), the trees containing {6–13} have been filtered out (as shown by the red x), so that the consensus tree now shows ‘what if’ {6–13} is not a clade. The {6–7} clade now branches off above {4–5, 13}, as one might expect. The unresolved branch including the yellow taxa {10,12} is now fully resolved. Since there remain only three of the original 5000 trees (as shown at the top of the window on the right), it is not so surprising that it is resolved. However, the particular resolution may be of interest.

The full interface includes buttons and sliders on the left and the common clade list on the right. The ‘Clone Window’ button allows comparison of multiple ‘what if’ scenarios, and was used to generate (a) and (b). Other buttons allow coloring taxa using up to five colors, taxa selection, collapsing subtrees and setting the outgroup, all of which were illustrated above. In addition, the threshold for the consensus tree is continuously adjustable. There is also a slider to return to previous states of the analysis, with forward and backward buttons for single undo/redo. The summary tree can also be pruned. Finally, there are checkboxes to control the encoding of branch information. All three possible encodings are shown in the figures: the horizontal distance between tree nodes represents branch length; the amount of zigzag encodes the variance (zigzag length/horizontal distance = branch length plus one standard deviation/branch length); and the dash length relative to the distance between dashes is the relative frequency of the clade in the tree set. Solid lines indicate clades that occur in 100% of the
trees, which is true of all but three branches in (c). If relative frequency is not encoded with dashes, it is shown with a text label.

In (c), the mouse is over the clade {8–12}, so information about that clade is shown at the bottom of the window: it occurs in 100% of the 3 trees, and its branch length is 0.032 with a standard deviation of 0.0153. This clade is also highlighted in the common clade list on the right. Clades are named with taxon index ranges, so it is difficult to couple color exactly between the tree and the list. If all taxa in a common clade are colored the same way, the entire name is colored. Otherwise ranges or individual numerals are colored.

To the right of the common clade names are pairs of checkboxes to filter the tree set to those trees that include or don’t include the clade, respectively. In the example, the exclude box of {6–13} is checked. To the right of the checkboxes are numerical and graphical representations of the relative frequency of the clade in the entire set and in the filtered set. The difference is shown graphically with an arrow pointing from the unfiltered value to the filtered one. {6–13} of course goes to zero. The relative frequency of most of the other common clades changes significantly, as well, either increasing or decreasing. If the filtered relative frequency is 0% or 100%, adding that clade to the current filters results in either the same set or the empty set, neither of which is useful. Therefore, the checkboxes for such clades are removed.

3 IMPLEMENTATION
STE was built using Piccolo (Bederson et al., 2004) and requires Java 1.4 or later. It reads tree sets in Nexus, Newick, or Badger (Simon and Larget 2004) format, and can write out a pruned and/or filtered tree set in the same formats. It has been tested on tree sets as big as 50,000 100-taxa trees.

ACKNOWLEDGEMENTS
Thanks to Don Simon, Bret Larget, Jay Kadane, and David Baum for defining the problem, suggesting design improvements, and contributing code. This work was supported by National Institutes of Health (NIH) grant R01 GM068950-01.

Conflict of Interest: none declared.

REFERENCES