8.4 Tree balance statistics

If some lineages radiate more rapidly than other lineages, the tree will be more "unbalanced." Thus, measures of tree shape can provide insights to differential diversification of lineages.

A variety of programs are available to calculate tree balance statistics. Here, I provide a brief overview of some of the options. To illustrate the methods, I'll use a nexus file of Indo-European languages, which can be found here:

IEchrono2.nex

The tree was generated using Bayesian methods and dated. For more information, see Gray and Atkinson 2003.

**R: apTreeshape**

First, let's investigate some of the options in the R package *apTreeshape* (Bortolussi et al. 2006). Install the package and load it in the usual way:

```r
install.packages("apTreeshape")
library(apTreeshape)
```

This package uses a somewhat different tree format called "treeshape." This is simply a way to represent the tree in computer memory, and there are several different formats that are used. One option is to read the nexus file into a phylo object using the ape package, and then convert from phylo to treeshape, as done here:

```r
IEphy_phylo=read.nexus("/IEchrono2.nex")  # note that I changed the working directory # to the directory with this file.
IEphy_treeshape=as.treeshape(phy)         # this converts from phylo to treeshape
```

Plot the tree:

```r
plot.treeshape(IEphy_treeshape)           # you can also simply use "plot"
```

Now, we can calculate a variety of tree shape statistics, such as Colless' I. This statistic is the sum of the absolute values of the difference between the number of taxa in every pair of clades on the tree; thus, larger values indicate less balanced trees, which is consistent with differential diversification in the clade. Colless' I can be standardized to account for the fact that a larger tree has more nodes (see Agapow and Purvis 2002 and Bortolussi et al. 2006).
An easy way to obtain Colless’ I is with the “summary()” function, as follows:

```
summary(IEphy_treeshape)
```

You will see that there are 87 tips (languages, in this case), and Colless’ I is 461. We can also see that the expected value of this index is 291.4 under a Yule model (and 1438 under a “PDA” model, see below). This indicates that the Indo-European tree is less balanced than expected, but is this statistically significant? We have estimates of the standard deviation, but it is helpful to examine the distribution of the statistic empirically. For this, let’s use this code:

```
CollessDbn = sapply(rtreeshape(500, tip.number = 87, model = "yule"), FUN = colless)
hist(CollessDbn, freq = FALSE, main = "Colless' indices for randomly generated Yule trees", xlab = "Colless' indices", n = 50)
pValue = length(subset(CollessDbn, CollessDbn > colless(IEphy_treeshape)))/length(CollessDbn)
```

The first of these lines of codes creates 500 random trees with 87 tips under a Yule model, and calculates Colless’ I on each of those random trees. The values of Colless’ I are stored in the vector CollessDbn. The second line gives a histogram of the values. The final line calculates how many of the Colless statistics from the random trees are larger than the observed Colless statistic on IEphy_treeshape. Typing pValue gives (approximately) 0.018, suggesting that we can reject the null hypothesis (the p-value is determined by stochastic simulation, and so your p-value may differ slightly). We can see from the plotted output that the observed value (=461) lies in the outermost tail of the distribution of values.

We can also more simply use the function colless.test(). Thus, our alternative hypothesis is that the Indo-European tree is more balanced, and we use the following code to generate a one-tailed p-value based on 1000 replicates of the null (yule) model:

```
colless.test(IEphy_treeshape, model = "yule", alternative = "greater", n.mc = 1000)
```

You should see the following output:

```
Test of the yule hypothesis using the Colless index
Statistic = 461
Standardized Statistic = 1.948874
p-value = 0.014
alternative hypothesis: the tree is less balanced than predicted by the yule model
```

Note: the p-value was computed using a Monte-Carlo method

Note that this output also provides a standardized version of Colless’ statistic (=1.945). This can also be obtained with the colless() function, i.e. colless (IEphy_treeshape, norm="yule"), which standardizes (normalizes) based on a Yule model (see Bortolussi et al. 2006).

The apTreeshape package provides a variety of other functions and options, including comparison to “proportional to distinguishable arrangements” models (PDA), in which all trees are equally likely (Mooers and Heard, 1997), likelihood based tests to compare the fit of PDA and Yule models, and additional measures of balance (e.g., the functions sackin() and cladesize()). Although it is not comprehensive in its implementation of different tree balance statistics, an advantage is that it unleashes the flexibility of using R to study tree balance.

Other Options
SymmeTREE

The book discusses some of Brian Moore and colleagues' work on tree balance statistics, including their application to study primate phylogeny. The program SymmeTREE (Chan and Moore 2005) implements many of these methods. As noted by the authors in their documentation, the program allows users to address two questions: Has a tree experienced variation in diversification rates among its branches? If so, where in the tree have significant shifts in diversification rate occurred? The program is command line, but has a nice user’s guide, and even video tutorials. On page 6 of the manual is a useful flow chart showing the steps needed to import trees of different types.

Mesa

The program Mesa calculates a wide diversity of tree balance and related statistics, including Slowinski-Guyer imbalance metric. It runs on Mac and Windows, and also provides many other useful statistics and tree-editing features, including the potential to automate many of the functions. Worth checking out if you are working on tree balance related questions. See also Agapow and Purvis (2002) and Purvis et al. (2002).

Sources Cited


Contributed by: Charlie Nunn