strap: an R package for plotting phylogenies against stratigraphy and assessing their stratigraphic congruence: A tutorial

Mark A. Bell[1] and Graeme T. Lloyd[2,3]

[1]Department of Earth Sciences, University College London,
Gower Street, London, WC1E 6BT, UK; email: mark.bell521@gmail.com.
[2]Department of Earth Sciences, University of Oxford
South Parks Road, Oxford, OX1 3AN, UK; email: graemetlloyd@gmail.com.
[3]Current address: E8A 320, Department of Biological Sciences,
Faculty of Science, Macquarie University, NSW 2109, Australia

November 8, 2014

1 Installation and loading

Once the user has installed R (freely-available on all platforms from: `http://cran.r-project.org/`) and booted it up they can type the following to install the strap and dependent geoscale packages:

```
install.packages(c("geoscale", "strap"), dependencies=TRUE)
```

The package can then be loaded into the R environment by typing:

```
library(strap)
```

Some basic information about the package can be obtained by typing:

```
?strap
```

Clicking on the index link at the bottom of the help file will bring up a list of the major functions and data files included.

2 Importing example data

Before we can showcase the functions we need to import some data to apply them to. The package already contains two example data sets, one for lungfish (from Lloyd et al. 2012) and one for asaphid trilobites (from Bell and Braddy 2012). However, it is probably worthwhile to show how the user might import their own data and so here we include the raw data for both these examples in the Supplementary Information. These files can be viewed by opening them in a text editor to give the user an idea about how they should be formatted. Note that here we assume the tree file is in Newick format, and hence import into R using the function read.tree in the R package ape (Paradis et al. 2006), but data in #NEXUS format can also be imported using the read.nexus function in the same package. The ages files are comma-delimited and represent two columns of numbers (for first and last appearances, respectively, in millions of years), with row names that correspond exactly to the taxon names in the tree file

and column names that are precisely "FAD" (First Appearance Datum) and "LAD" (Last Appearance Datum). This formatting matches that used in the R package paleotree (Bapst 2012) and allows for easy swapping of time-scaled trees between packages. (An example of how the user might do this is included in the tutorial below.) Once downloaded these files should be moved to a folder on the user's hard drive and then within R the working directory should be set to that folder. This can be done either from the menu system or using the function setwd. Once this is done we can import the lungfish data by typing:

```
Dipnoi.ages <- read.table("Dipnoi.txt", header=T)
Dipnoi.tree <- read.tree("Dipnoi.tre")
```

The first line imports the age data and stores it in the variable Dipnoi.ages, and the second line imports a single phylogenetic tree and stores it in the variable Dipnoi.tree. We can repeat this with the asaphid data by typing:

```
Asaphidae.ages <- read.table("Asaphidae.txt", header=T)
Asaphidae.trees <- read.tree("Asaphidae.tre")
```

The first line imports the age data and stores it in the variable Asaphidae.ages, and the second line imports all of the equally parsimonious phylogenetic trees and stores them in the variable Asaphidae.trees. At any point the user can view the contents of a variable simply by typing its name. So, for example, the ages data for asaphid trilobites can be viewed by typing:

```
Asaphidae.ages
```

3 Time-scaling a phylogeny

Now we have some data we can try out the three main functions. First of all we will use the lungfish data and the time-scaling function DatePhylo with the default options:

```
Dipnoi.ts.tree <- DatePhylo(Dipnoi.tree, Dipnoi.ages)
```

Our newly time-scaled tree is now stored in the variable Dipnoi.ts.tree. We can compare our non-time-scaled tree with our newly time-scaled tree by typing their variable names:

```
Dipnoi.tree
Dipnoi.ts.tree
```

The main difference you should note is that the latter includes branch lengths and the former does not. We can see the branch lengths in millions of years by typing:

```
Dipnoi.ts.tree$edge.length
```

Note that many branches are zero million years in length. This is because the default time-scaling option is to treat each node as being only as old as its oldest descendant (Norell 1992; Smith 1994). Thus for each node the branch leading to its oldest descendant will necessarily have zero-length (except for terminal branches where the difference between the FAD and LAD is considered a range). Another useful feature, borrowed from the paleotree package (Bapst 2012), is the inclusion of the root age, which is essential for plotting our phylogeny against stratigraphy. This can be viewed by typing:

```
Dipnoi.ts.tree$root.time
```

4 Plotting a phylogeny against geological time

4.1 Temporal units

The function geoscalePhylo allows for a time-scaled phylogeny to be plotted against geologic time using either the current geologic time scale of Gradstein et al., 2012 or previously published time scales by the International Commisioin on Stratigraphy. The time scale that is plotted is comprised of a number of temporal components representing the different units that the geological time scale is divided into. There are five primary temporal units that can be included, each of which have two alternative names and are as follows: Eon (Eonothem), Era (Erathem), System (Period), Series (Epoch), and Stage (Age). These alternative names can be used interchangably i.e. both Eon and Erathem are accepted, however should both these alternative names be included then that temporal unit will only be included once. In addition, the order in which they are included into units does not affect the order in which they appear in the chart so units=c("Period","Epoch","Age") will produce the same results as units=c("Age","Epoch","Period") with the default order as they were listed previously with Eons plotted at the base and Stages at the top.

4.2 Basic options

Using our newly time scaled lungfish tree we can plot it against the current geologic time scale (Gradstein et al. 2012) using the default options of geoscalePhylo by typing:
The best way to visualise our time-scaled tree is by plotting it against geologic time using either the current geologic time scale (Gradstein et al. 2012), or previous versions. This can be done with the function geoscalePhylo. Using our newly time-scaled tree and the default options we can type:

```
geoscalePhylo(Dipnoi.ts.tree)
```

This will show a plot in a new window. Alternatively it can be exported as a PDF to the user's working directory by typing:

```
pdf("Dipnoi_tree_1.pdf", width=10, height=7)
 geoscalePhylo(Dipnoi.ts.tree)
dev.off()
```

The result should appear exactly as in Figure 1. Note that the tree has a large polytomy, but this isn't actually the case, rather it only appears that way due to the large number of zero-length branches. We can get around this problem by re-scaling the tree and employing the "equal" option (using a root length of 2 million years). This will enforce all branches to have a positive length:

```
Dipnoi.ts.tree <- DatePhylo(Dipnoi.tree, Dipnoi.ages, 2, "equal")
 pdf("Dipnoi_tree_2.pdf", width=10, height=7)
  geoscalePhylo(Dipnoi.ts.tree)
 dev.off()
```

The result should appear exactly as in Figure 2. Note that other options for DatePhylo can be accessed from the help file by typing:

```
?DatePhylo
```

Although we've removed our apparent polytomy the branch lengths in the Devonian are still very short. If we want to make them longer (for aesthetic or analytical purposes) then we can employ the minimum-branch length option (here we will make the minimum length 2 million years) in the timePaleoPhy function from the paleotree package by typing:

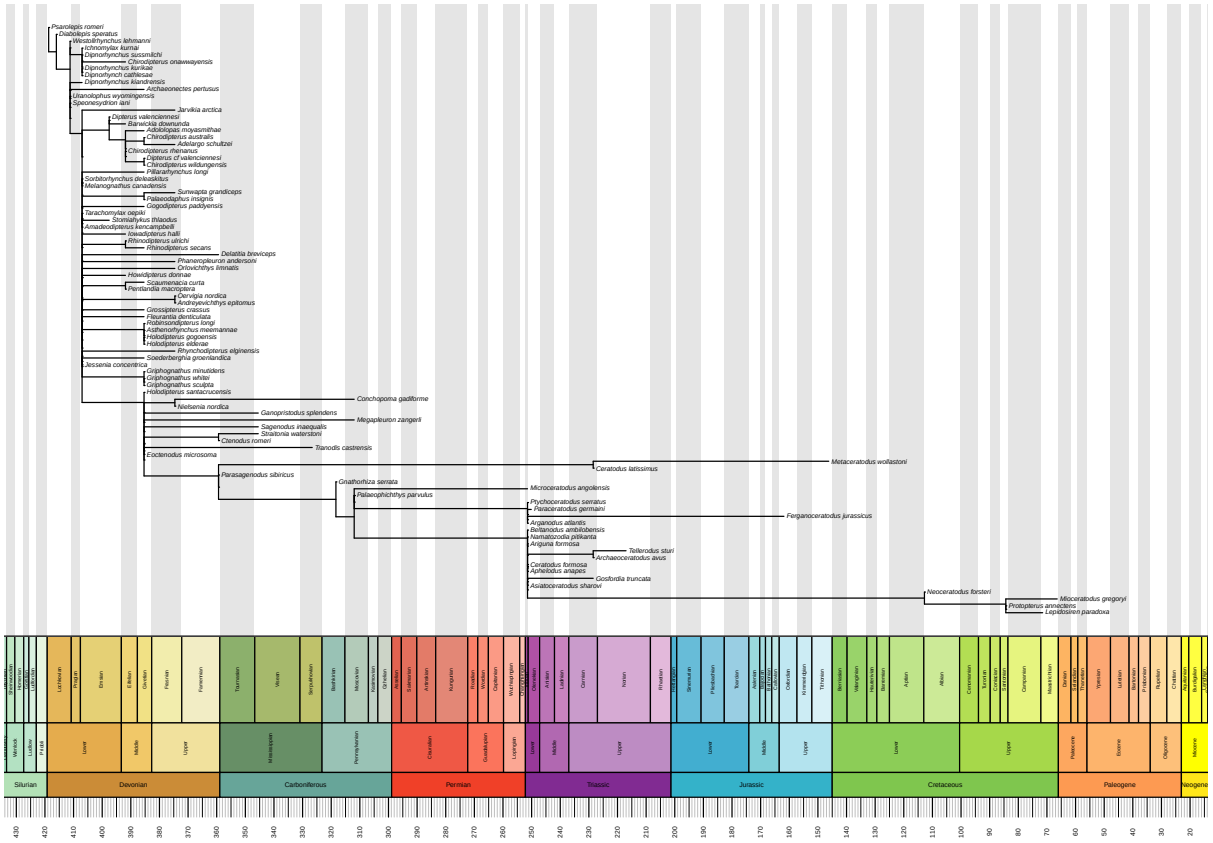

Figure 1: Plot of the lungfish (Lloyd et al. 2012) phylogeny (time-scaled using the DatePhylo function and the default options) plotted against stratigraphy using the function geoscalePhylo with the default options.

```
install.packages("paleotree" , dependencies=TRUE)
 library(paleotree)
  Dipnoi.ts.tree <- timePaleoPhy(Dipnoi.ts.tree, Dipnoi.ages, "mbl", 2)
   pdf("Dipnoi_tree_3.pdf", width=10, height=7)
    geoscalePhylo(Dipnoi.ts.tree)
   dev.off()
```

The result should appear exactly as in Figure 3. The exact branching sequence should now be much clearer, but we can still add some extra options to the figure. The following will ladderise the tree to the left and add the ranges of the taxa as thicker black bars:

```
pdf("Dipnoi_tree_4.pdf", width=10, height=7)
 geoscalePhylo(ladderize(Dipnoi.ts.tree, right=FALSE), Dipnoi$ages,
   cex.ts=0.5)
dev.off()
```

The result should appear exactly as in Figure 4.

4.3 Including a user-defined time scale.

In practice the user might also want to use their own time-scale. This might reflect local stratigraphy, a particular biozonation, or some other time-scale not included here. This can be added into the time scale by including "User" into the units argument. This requires a table with three columns, "Start", "End", and "Name", corresponding to the date of the base of each unit, top of each unit, and name to

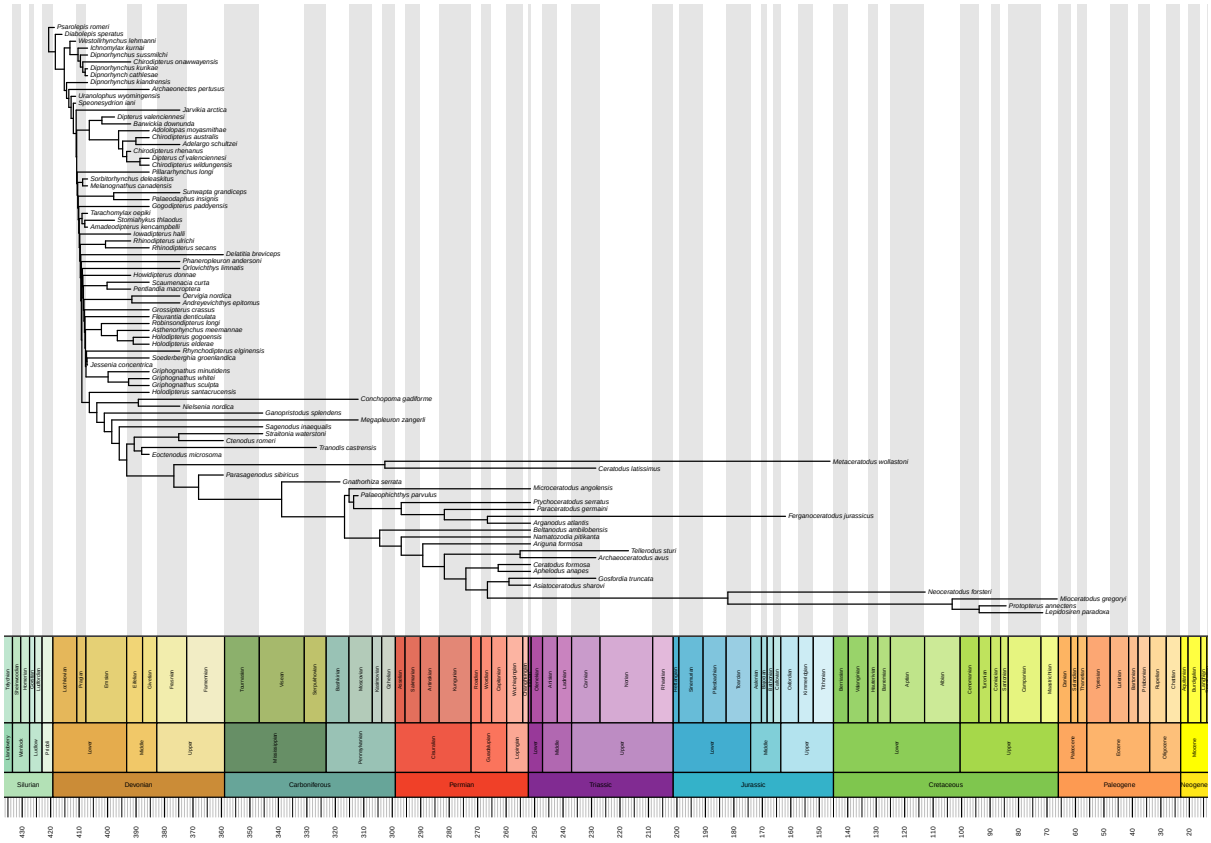

Figure 2: Plot of the lungfish (Lloyd et al. 2012) phylogeny (time-scaled using the DatePhylo function and the "equal" method with a root length of 2 million years) plotted against stratigraphy using the function geoscalePhylo with the default options.

be printed of each unit respectively. This table should then be assigned to the user.scale argument. As an example we include a table of ages for UK Ordovician stages. This can be viewed by typing:

```
UKzones
```

Using the asaphid trilobite data an example of how a user time-scale can be implemented is given by typing:

```
Asaphidae.tree <- DatePhylo(Asaphidae$trees[[1]], Asaphidae$ages,
 method="equal", rlen=1)
 pdf("Asaphidae_tree.pdf", width=10, height=7)
  geoscalePhylo(ladderize(Asaphidae.tree, right=FALSE), Asaphidae$ages,
   cex.ts=0.4, units=c("Period","Epoch", "User"), user.scale=UKzones,
    vers="ICS2009")
 dev.off()
```

The result should appear exactly as in Figure 5. Again, additional options for the function can be accessed by typing a question mark followed by its name:

```
?geoscalePhylo
```

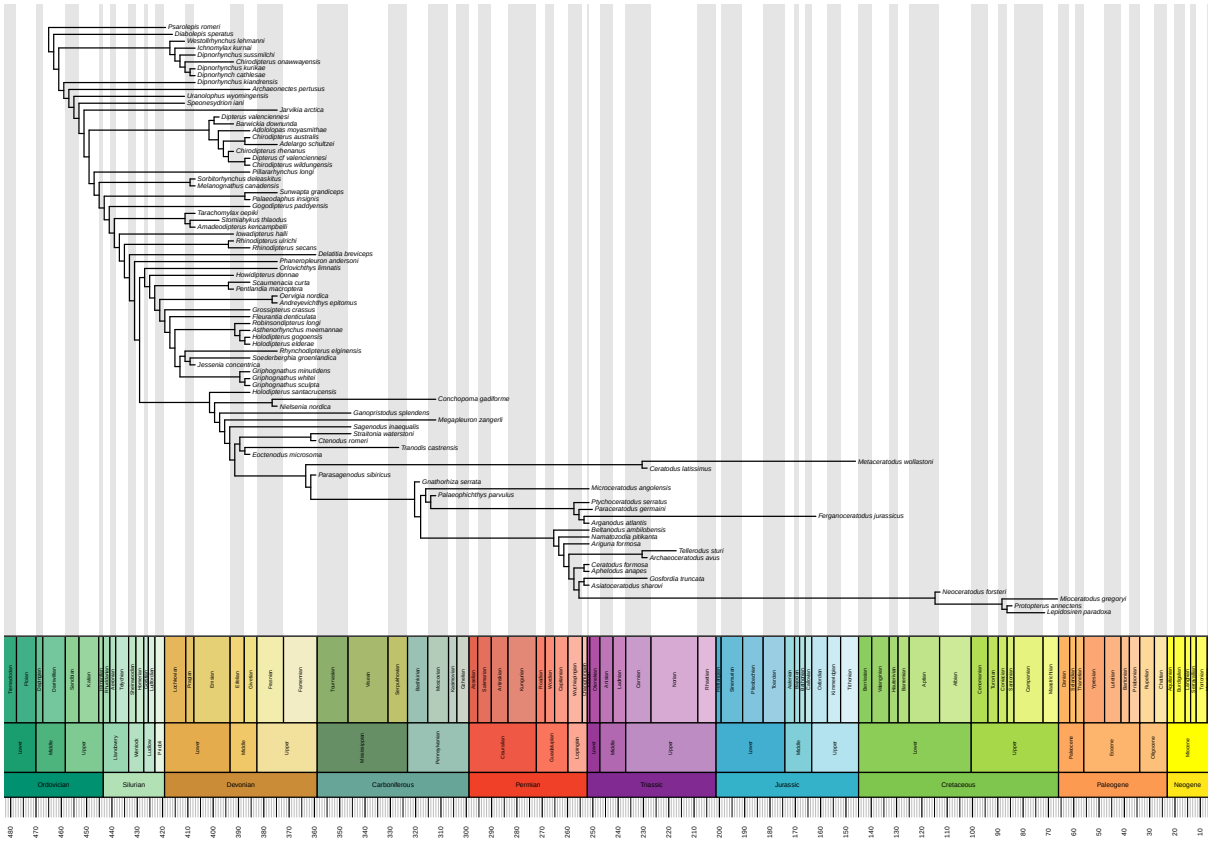

Figure 3: Plot of the lungfish (Lloyd et al. 2012) phylogeny (time-scaled using the timePaleoPhy function in the paleotree package (Bapst 2012) and the "mbl" method with a minimum branch length of 2 million years) plotted against stratigraphy using the function geoscalePhylo with the default options.

4.4 Additional options for geoscalePhylo

There are a couple of additional features available in geoscalePhylo for formatting the look of your tree. Firstly, the function plot.phylo allows a tree to be plotted in four different directions using the argument: "rightwards" (the default), "upwards", "downwards" and "leftwards". In all the previous examples the trees have been plotted using the default "rightwards" option but trees can also be plotted vertically using the direction argument as in the following:

```
pdf("Dipnoi_tree.pdf", width=10, height=7)
 geoscalePhylo(Dipnoi.ts.tree, Dipnoi.ages, cex.ts=0.4,
  direction="upwards")
dev.off()
```

The result should appear as in Figure 6 with the tree starting from the bottom of the plot and the time-scale on the left. Users may also wish to have the text in the time-scale orientated in a different direction rather than the default. Three arguments are available to alter the direction of the Epoch/Series, Age/Stage and the User temporal units called erotate, arotate and urotate respecitively. The default values are zero when direction is set to "upwards" and 90 when direction is set to "rightwards". Using the Asaphidae tree file as an example if we wanted to plot the tree horizontally but keep the names plotting horizontally we could use the following:

```
pdf("Asaphidae_tree.pdf", width=10, height=7)
 geoscalePhylo(ladderize(Asaphidae.tree), Asaphidae.ages,
```

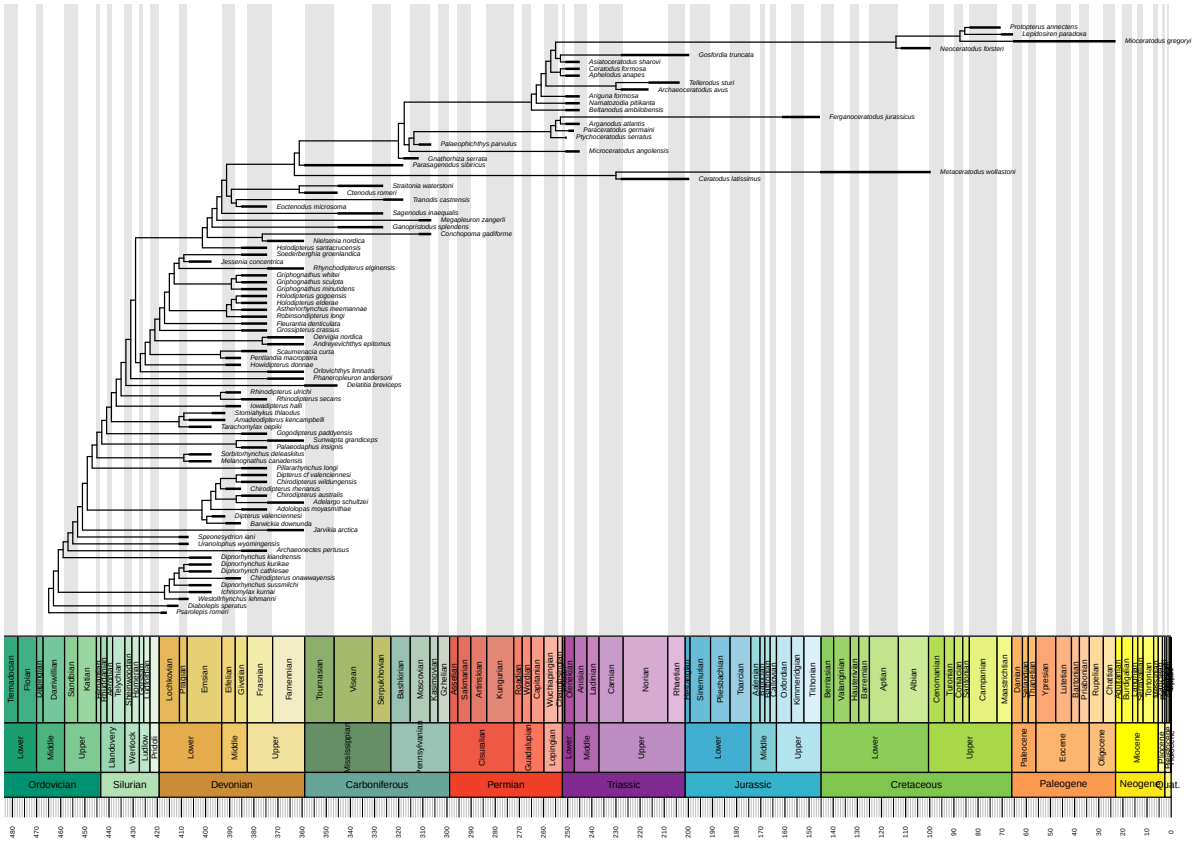

Figure 4: Plot of the lungfish (Lloyd et al. 2012) phylogeny (time-scaled using the timePaleoPhy function in the R package paleotree (Bapst 2012) and the "mbl" method with a minimum branch length of 2 million years) plotted against stratigraphy using the function geoscalePhylo with observed ranges of taxa plotted as black boxes and alternating grey and white boxes in the background denoting geologic stages.

```
    cex.ts=0.4, units=c("Period","Epoch","User"), user.scale=UKzones,
    vers="ICS2009", erotate=0,urotate=0)
  dev.off()
```

Figure 7 shows what the output of these arguments will look like compared to the previous example in Figure 5.

5 Assessing stratigraphic congruence of phylogenies

Now we can move on to the StratPhyloCongruence function and assess the stratigraphic congruence of our trees. We will employ the Asaphidae data set and the recommended options. These include: 1) resampling the input trees and both randomly bifurcating them and drawing random dates for the tips between a taxon's first and last appearances (options: hard=FALSE and randomly.sample.ages=TRUE), 2) using the default numbers of permutations (1,000) for both these resampled trees and the randomly generated trees used in the significance tests (options: samp.perm=1000 and rand.perm=1000), 3) fixing of the overall tree shape (in the random topologies; option: fix.topology=TRUE), and 4) fixing the outgroup taxon (option: fix.outgroup=TRUE). These last three options do the best job of ensuring our significance test is both accurate and as fair as possible (see above). The user can implement these options together by typing:

```
X <- StratPhyloCongruence(Asaphidae.trees, Asaphidae.ages, hard=FALSE,
```

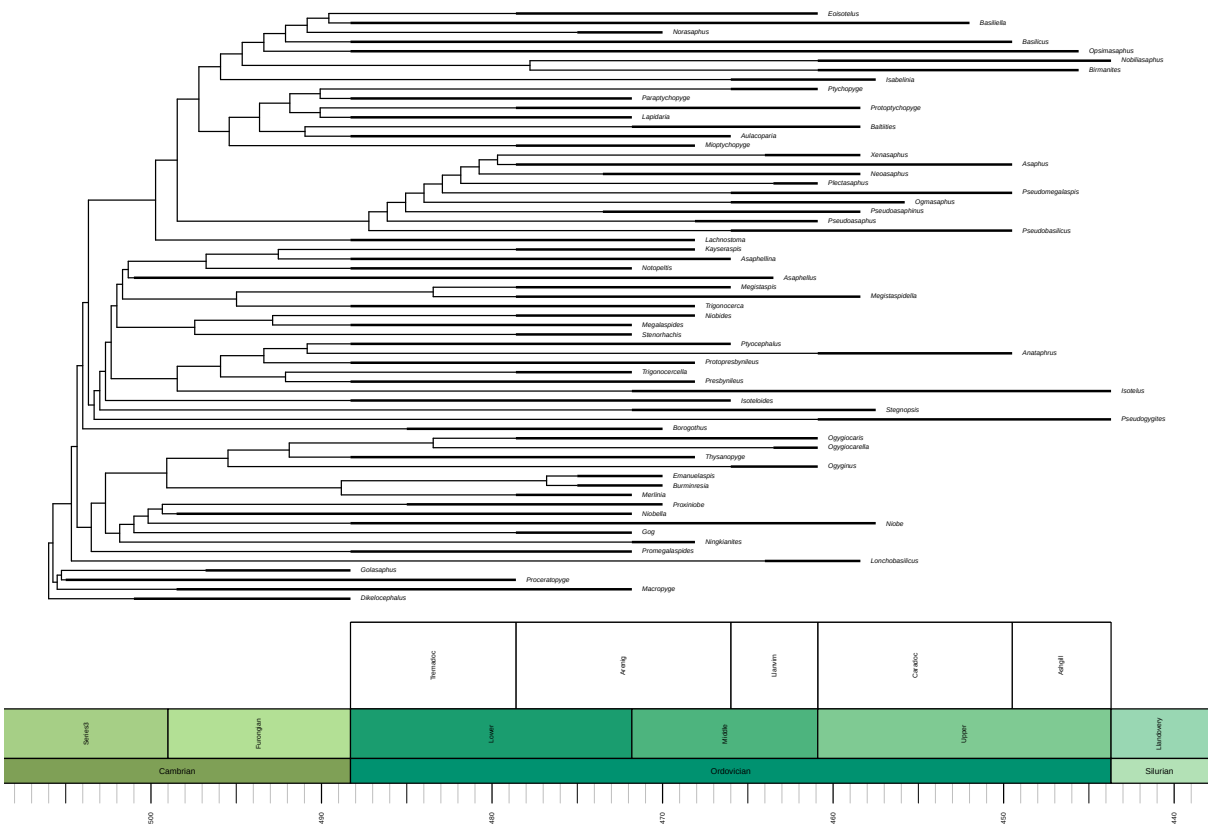

Figure 5: Plot of the asaphid trilobite (Bell and Braddy. 2012) phylogeny (time-scaled using the DatePhylo function and the "equal" method with a root length of 1 million years) and plotted against stratigraphy using the function geoscalePhylo. The Ordovician regional stages are included in the time-scale as white boxes.

```
    randomly.sample.ages=TRUE, fix.topology=TRUE, fix.outgroup=TRUE)
```

It is important to note that this function will probably take several minutes to run, depending on the power of the user's computer. After a short wait (when the randomly generated trees are created) the user should see a progress bar that will give some indication of the time it will take to complete the remainder of its tasks. Once complete the results are stored in a new variable called X. Six different outputs are generated by the function and these will be viewed one at a time below. First of all, and perhaps most important, are the results for the stratigraphic fit measures and significance tests for the input trees themselves:

```
X$input.tree.results
```

The user could also export this table as a comma-delimited text file (easily opened in software such as Microsoft Excel) by typing:

```
write.table(X$input.tree.results, "Asaphid_strat_congruence.csv", sep=",")
```

This is a table where each row corresponds to the result for an input tree (given in order) and each column is either a stratigraphic fit measure (SCI, RCI, GER, and MSM*), the significance test for that measure (p.SCI, p.RCI, p.GER, and p.MSM*), or a combined fit and significance test measure (GER* and GERt). Note that the p-values indicate the probability of the null, that the input tree has no better a fit to stratigraphy than a random topology. Thus significantly good fits to stratigraphy will be indicated by very small p-values. Importantly, the level of significance may vary between stratigraphic fit

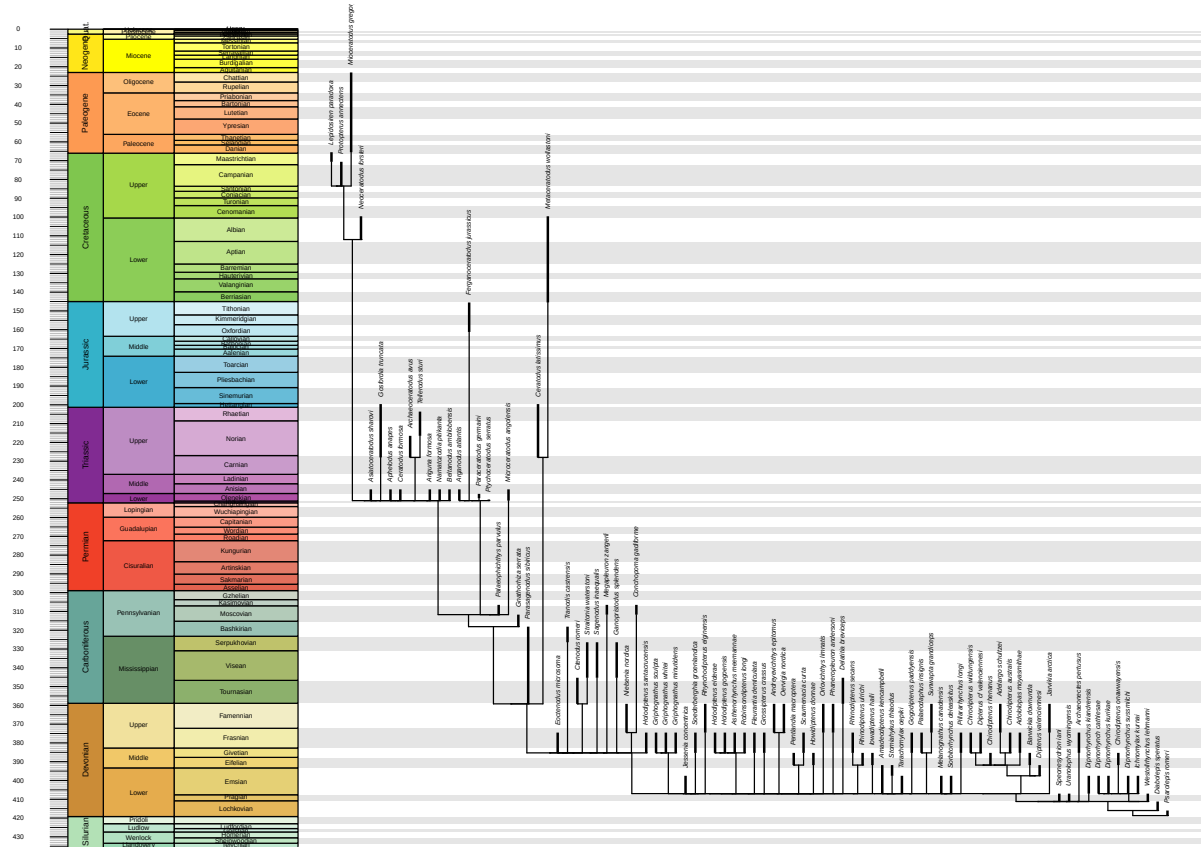

Figure 6: Plot of the lungfish (LLoyd 2012) phylogeny plotted plotted against stratigraphy using the geoscalePhylo function and the direction="upwards" option.

measures. This is apparent in our Asaphidae example where we find small p-values for the SCI and RCI measures, but comparatively larger ones for the GER and MSM*. A similar output for the resampled trees can be found by typing:

```
X$samp.permutation.results
```

The number of rows here corresponds to the number of sampling permutations used (the default being 1,000) and the columns are identical to the input tree results. Note that depending on the options used there may be no resampled trees and instead this variable will return, "NULL". Next the results for the random topologies can be found by typing:

```
X$rand.permutation
```

Again each row corresponds to a tree and as with the sampling permutations the number of rows is decided by the number of random permutations (again the default being 1,000). Unlike for the input and resampled trees, this time there are only four columns, one for each of the main stratigraphic fit measures. These are used to generate the distributions that are then used to test the significance of the input and resampled trees. Finally the function also returns the trees used for every calculation:

```
X$input.trees
X$samp.trees
X$rand.trees
```

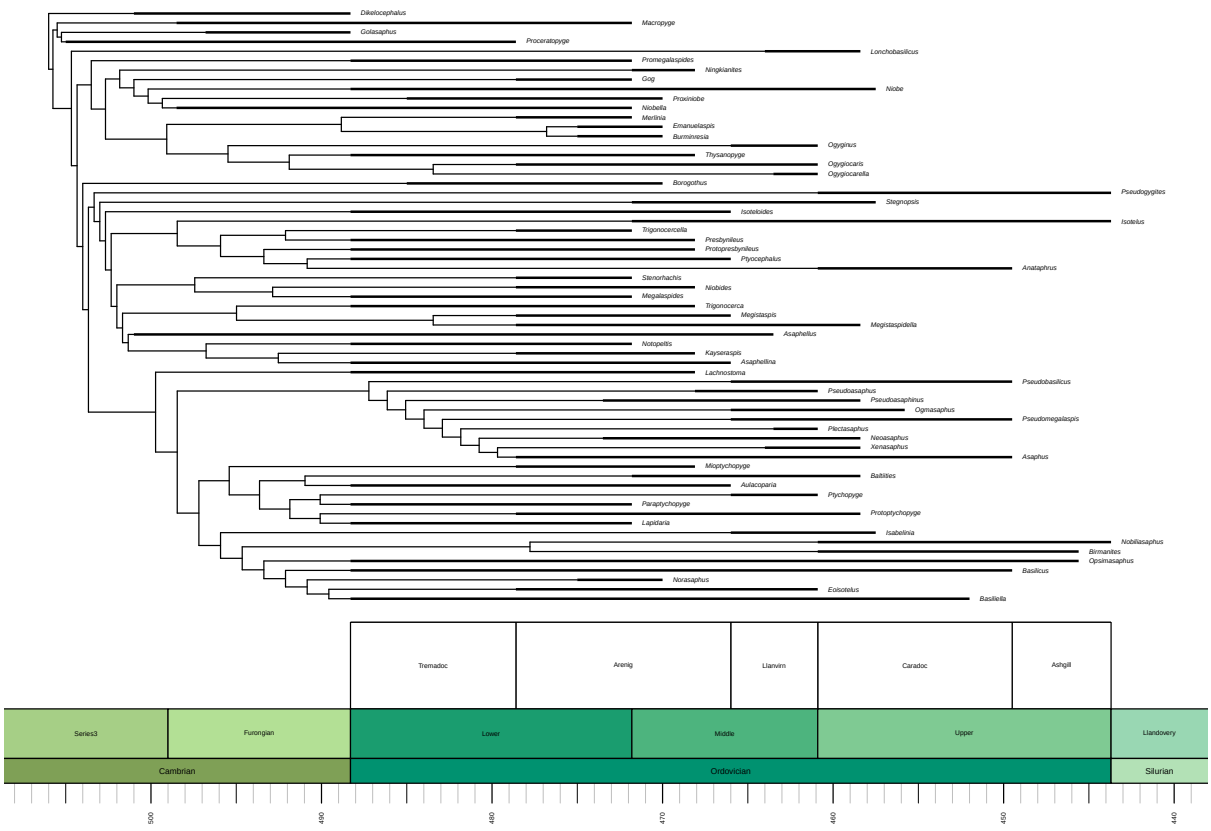

Figure 7: Plot of the asaphid trilobite (Bell and Braddy. 2012) phylogeny plotted against stratigraphy using the geoscalePhylo function with the text for the Epoch and User temporal units plotted horizontally using the erotate and urotate arguments respectively.

The first of these are the input trees (corresponding to the input tree results), the second are the resampled trees (corresponding to the sampled tree results), and the last are the randomly generated trees (corresponding to the randomisation results). Note that again if no resampling is done the sample trees will return "NULL". Crucially, these trees are all time-scaled according to the options chosen by the user. This allows the selection of, say, the input tree with the best RCI:

```
best.input.RCI <- X$input.trees[[which.max(X$input.tree.results[, "RCI"])]]
```

We can then plot this tree against geologic time as we did with our lungfish:

```
pdf("Asaphidae_tree_1.pdf", width=10, height=7)
 geoscalePhylo(ladderize(best.input.RCI), Asaphidae$ages, ranges=T,
  boxes="Age", cex.ts=0.5,vers="ICS2009")
dev.off()
```

The result should appear exactly as in Figure 8. Note that because we are again using the basic time-scaling method there are many zero-length branches giving the false impression of a polytomy. We can also visualise the results of our tests (this time using the SCI) with histograms by typing the following:

```
pdf("Asaphidae_strat_fit_SCI.pdf", width=10, height=7)
 par(mfrow=c(2, 1))
  cutoff <- qnorm(0.95, mean(X$rand.permutations[, "SCI"]),
```

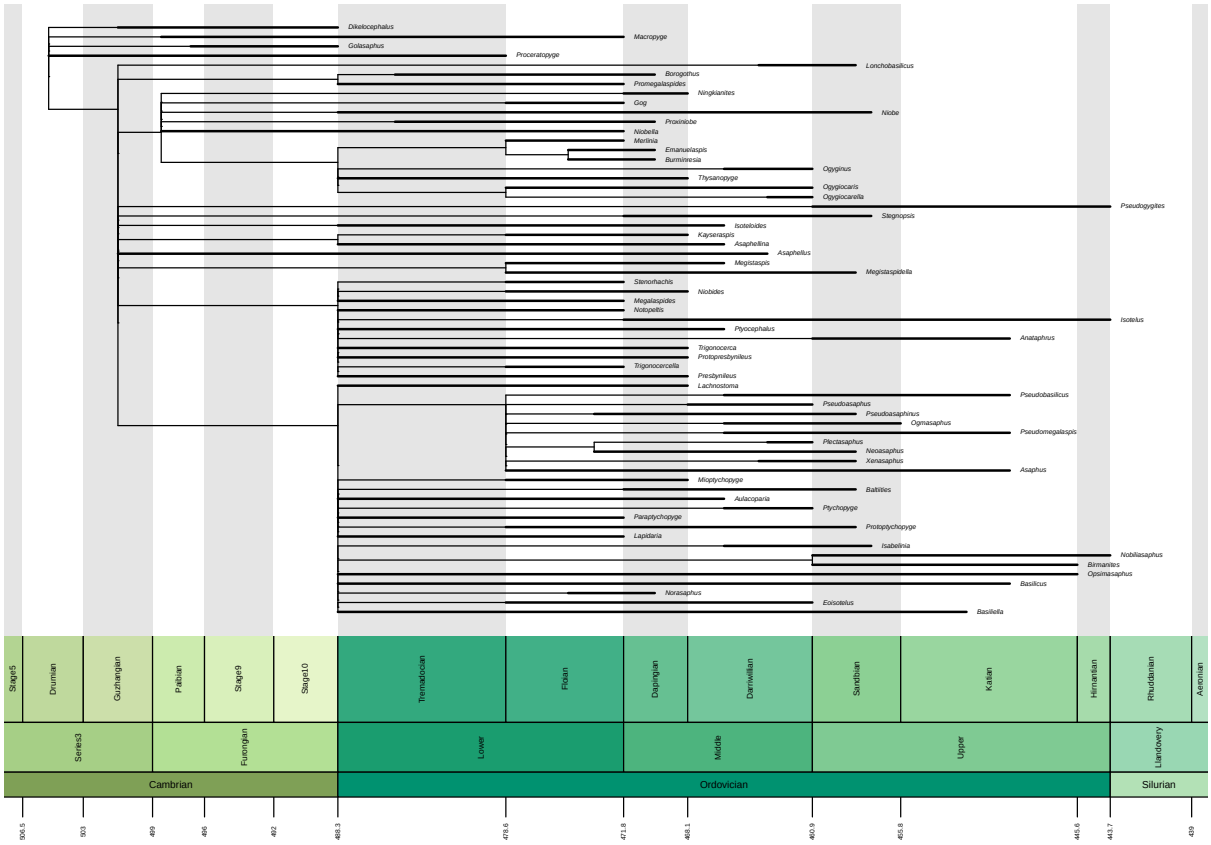

Figure 8: Plot of the asaphid trilobite (Bell and Braddy. 2012) phylogeny with the best RCI value (calculated using the StratPhyloCongruence function), time-scaled using the DatePhylo function and the default options, and plotted against stratigraphy using the function geoscalePhylo with the default options.

```
    sd(X$rand.permutations[, "SCI"]))
    hist(X$input.tree.results[, "SCI"], xlim=c(0, 1), xlab="SCI",
     main="Input trees", col="black")
      lines(x=c(cutoff, cutoff), y=c(0, 1000) , lty=2)
    hist(X$rand.permutations[, "SCI"], xlim=c(0, 1), xlab="SCI",
     main="Randomly generated trees", col="black")
      lines(x=c(cutoff, cutoff), y=c(0, 1000) , lty=2)
  dev.off()
```

The result should appear as in Figure 9, although the exact distribution of the random trees may vary. In the lower panel we can see the distribution of the SCI in the random topologies. The vertical dashed line indicates the critical value for a one-tailed test at an alpha of 0.05 (i.e., 95% of the random values will fall to the left and 5% to the right). Trees with an SCI that exceed this critical value may be considered to show significantly better fits to stratigraphy than random. The upper panel indicates the distribution of SCI amongst the input trees, with the same dashed line shown. From this we might infer that most (but not all) of our input trees have a significantly better fit to stratigraphy than expected at random. We can repeat this for a different measure (this time the GER) by typing:

```
pdf("Asaphidae_strat_fit_GER.pdf", width=10, height=7)
 par(mfrow=c(2, 1))
  cutoff <- qnorm(0.95, mean(X$rand.permutations[, "GER"]),
   sd(X$rand.permutations[, "GER"]))
```

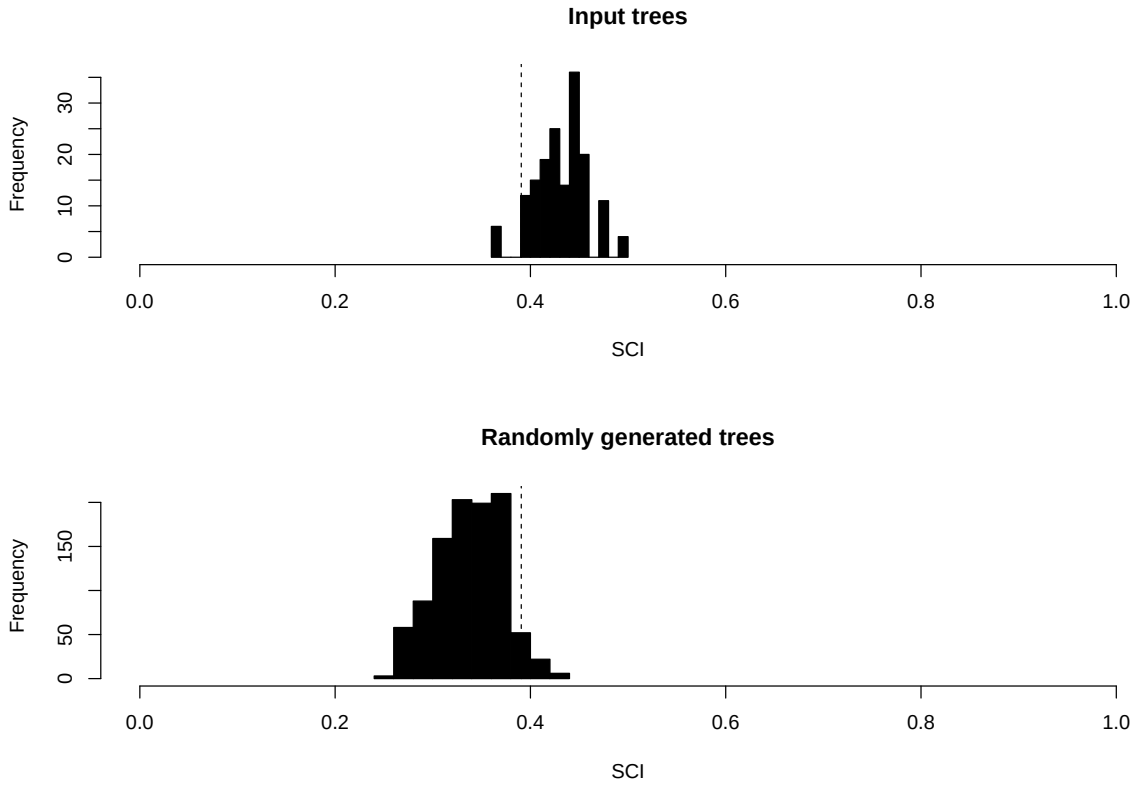

Figure 9: Histograms of the SCI values for the most parsimonious trees (top) for the asaphid trilobites (Bell and Braddy 2012) and randomly generated topologies (bottom) with the critical value (at an alpha of 0.05) for the one-tailed test drawn as a vertical dashed line. Note that most, but not all, input trees would reject the null hypothesis of having a stratigraphic fit that is no better than random.

```
  hist(X$input.tree.results[, "GER"], xlim=c(0, 1), xlab="GER",
   main="Input trees", col="black")
   lines(x=c(cutoff, cutoff), y=c(0, 1000) , lty=2)
 hist(X$rand.permutations[, "GER"], xlim=c(0, 1), xlab="GER",
  main="Randomly generated trees", col="black")
  lines(x=c(cutoff, cutoff), y=c(0, 1000) , lty=2)
dev.off()
```

The result should appear as in Figure 10, although again the distribution of the random trees may vary. Note that this time all of the input tree values are lower than the critical value and a significantly better fit to stratigraphy than random cannot be supported for any topology. As with the other functions the available options can be explored further by typing:

```
?StratPhyloCongruence
```

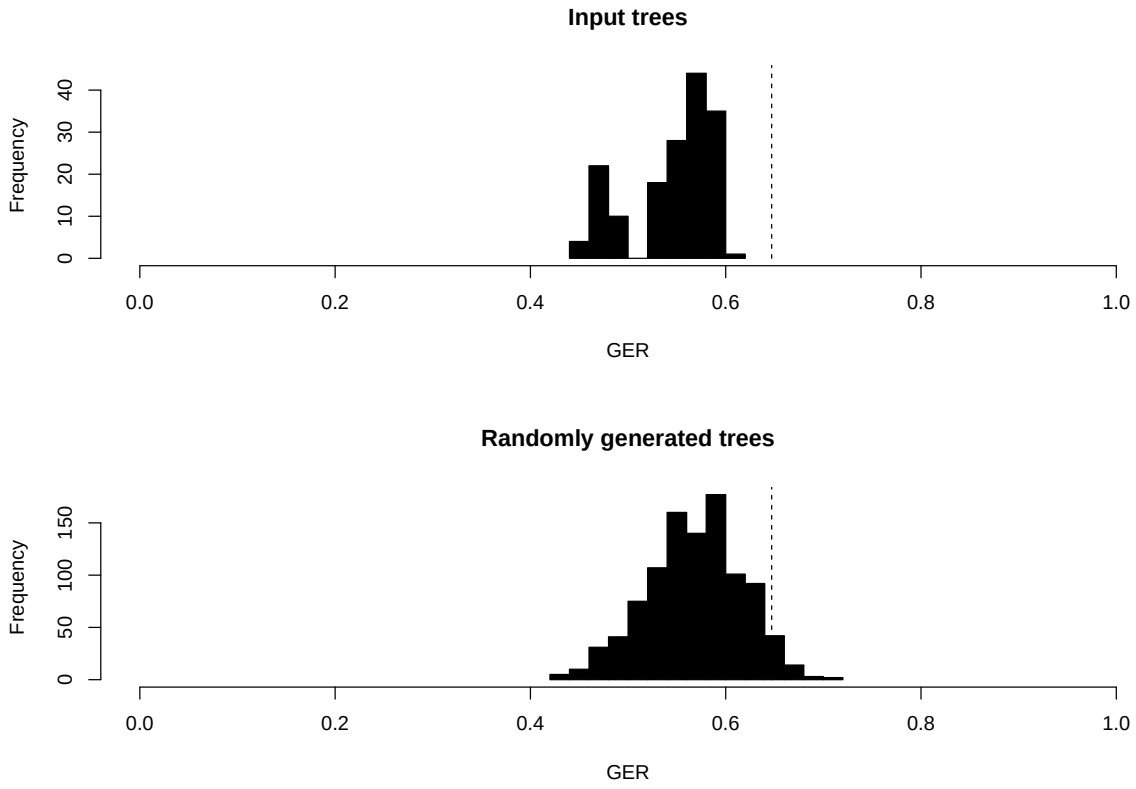

Figure 10: Histograms of the GER values for the most parsimonious trees (top) for the asaphid trilobites (Bell and Braddy 2012) and randomly generated topologies (bottom) with the critical value (at an alpha of 0.05) for the one-tailed test drawn as a vertical dashed line. Note that although the input data and random topologies are identical to those in Figure 9 for the GER metric no trees would reject the null hypothesis of having a stratigraphic fit that is no better than random..

6 References

BAPST, D. W. 2012. paleotree: an R package for paleontological and phylogenetic analyses of evolution. Methods in Ecology and Evolution, 3, 803-807.

BELL, M. A., and BRADDY, S. J. 2012. Cope's rule in the Ordovician trilobite Family Asaphidae (Order Asaphida): patterns across multiple most parsimonious trees. Historical Biology, 24, 223-230.

GRADSTEIN, F. M., OGG, J. M., and SCHMITZ, M. 2012. A Geologic Time Scale. Elsevier, Boston, USA.

LLOYD, G. T., DAVIS, K. E., PISANI, D., TARVER, J. E., RUTA, M., SAKAMOTO, M., HONE, D. W. E., JENNINGS, R., and BENTON, M. J. 2008. Dinosaurs and the Cretaceous Terrestrial Revolution. Proceedings of the Royal Society of London B, 275, 2483-2490.

NORELL, M. A. 1992. Taxic origin and temporal diversity: the effect of phylogeny. 89-118. In NOVACEK, M. J. and WHEELER, Q. D. (eds.). Extinction and Phylogeny. Columbia University Press, New York.

PARADIS, E., CLAUDE, J., and STRIMMER, K. 2006. APE: Analysis of Phylogenetics and Evolution in R language. Bioinformatics, 20, 289-290.

SMITH, A. B. 1994. Systematics and the Fossil Record: Documenting Evolutionary Patterns. Blackwell Science, Oxford, 223pp.