phylo-stability
Split-specific Bootstrap Measures for Quantifying Phylogenetic Stability and the Influence of Taxon Selection

Edward Susko
Department of Mathematics and Statistics, Dalhousie University

Introduction

The main programs, `bptaxon_split`, `rbic_taxon_split` and `taxa_split_support` implement the methods described in Wang et al. (2016); please cite this reference when using the software.

The program `bptaxon_split` obtains the `bptaxon_split` support measure for each split in an input tree and all taxa. Similarly, the program `rbic_taxon_split` obtains the `rbic_taxon_split` for each split in an input tree and all taxa. An additional program, `tree2treein` can be used to associate the labels of edges with the split labels in the output files.

The program `taxa_split_support` can be applied to the output of `bptaxon_split` or `rbic_taxon_split` to get average split support or average taxa support values.

Installation

The main programs need to be compiled from C source code. To install the programs

1. Download and unpack the software:

   ```
   $ tar zxf phylo-stability-v1.0.tar.gz
   ```

   This will create a directory `phylo-stability-v1.0` that contains the source code.

2. Change directories to `phylo-stability-v1.0` and create the main program files with the `make` command.

   ```
   $ cd phylo-stability-v1.0
   $ make
   ```

The default installation assumes that the gcc compiler is available. To use a different compiler change the variable `CC` in `Makefile`.

bptaxon_split

The program `bptaxon_split` obtains the `bptaxon_split` support measure for each split in an input tree and all taxa. It is run at the command line with

```
bptaxon_split -t treefile -b bootstrap_treefile -i sequence_file
```

Here `treefile` should give the name of a file containing the tree of interest in Newick format. `bootstrap_treefile` should give the name of a file containing the bootstrap trees. Each row of the file should contain a bootstrap tree in Newick format. `sequence_file` should give the name of the file with the sequence data, which should conform to PHYLIP standards. The output of the program is to the screen (stdout) and is a sequence of rows, with three entries in each row

```
taxon split bptaxon_split
```

Here `bptaxon_split` is the `bptaxon_split` support measure defined in Wang et al. (2016) and these are listed for each taxon x split combination. For each row the taxon name is indicated as taxon and the split number is indicated as split. The ordering of taxa matches the ordering in the sequence file. Which split corresponds to which edge can be seen after using the provided `tree2treein` program.

The ratites data considered in Wang et al. (2016) will be used as a running example. The sequence data for this example was stored in `24.dna`, the estimated tree in `RAxML_bestTree.24.out` and the bootstrap trees in `24.boot1000`. The program is run at the command line with

```
$ bptaxon_split -t RAxML_bestTree.24.out -i 24.dna -b 24.boot1000
Cassowary    24   100.00000
Emu          24   100.00000
...
BrushTurke   29    99.28750
Chicken      29    99.28750
...
Alligator    29    94.30000
...
```

For split 24, the measures were 100% regardless of the taxon considered. For split 29, however, there was some variation across taxa. The labels of the splits correspond to those of Figure 1 and were obtained with the `tree2treein` program.

Additional Notes about Input Format

The tree should conform to the Newick standard. The programs in PHYLIP (Felsenstein, 1989, 2004), TREE-PUZZLE (Schmidt et al. 2002) and PAML (Yang 1997, 2007), which can be used to obtain ML estimates of edge-lengths for the models described here, will output trees in this format. A discussion of this standard as implemented in PHYLIP is given at

`http://evolution.genetics.washington.edu/phylip/newicktree.html`

and a more formal description is available at

`http://evolution.genetics.washington.edu/phylip/newick_doc.html`

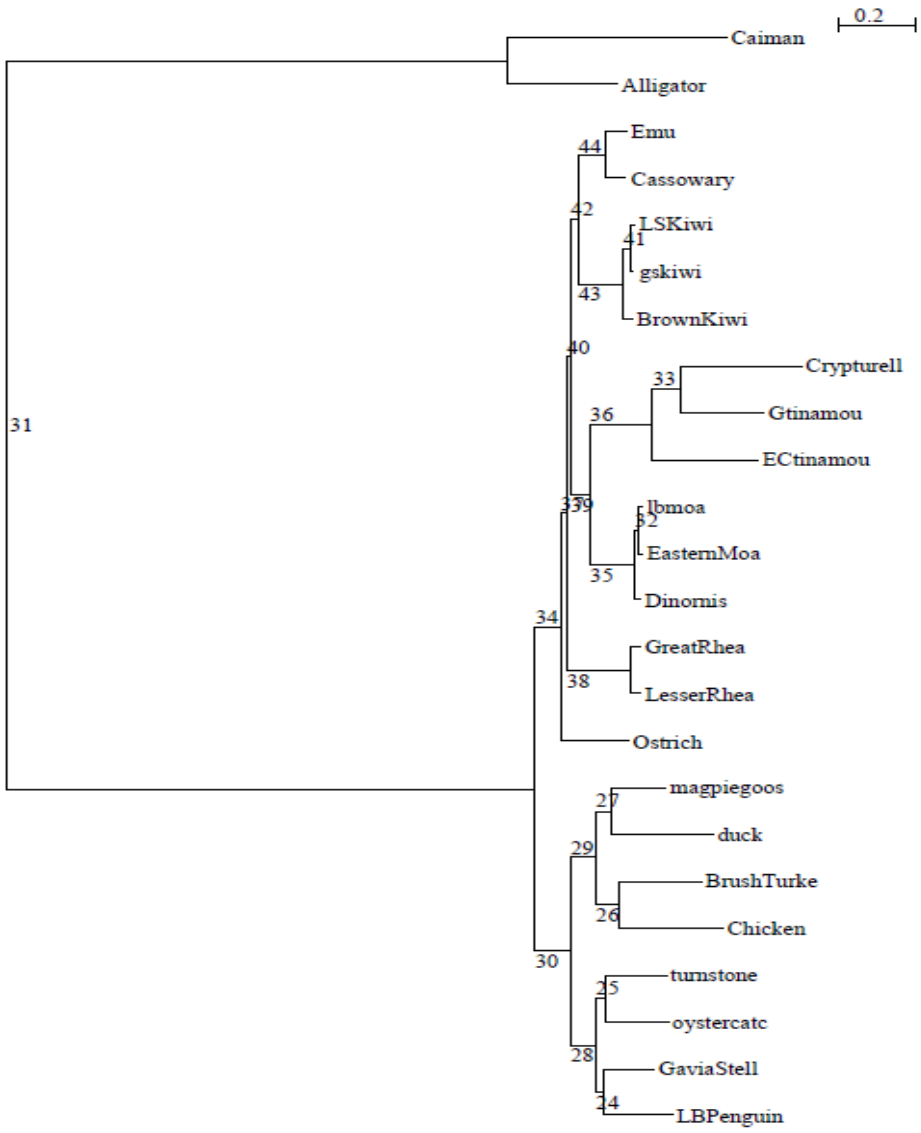

Figure 1: The ratites tree with labels.

The file should conform to PHYLIP standards for input with 10 character long names padded by blanks. The names should match the names used in the input treefile. Input can be either interleaved or sequential with one caveat: The lines 2 through $m + 2$, where m is the number of taxa, must contain the name of taxa followed by sequence data. For instance the start of a sequence file might be

```
6 3414
Homsa      ANLLLLIVPI LI...
Phovi      INIISLIIPI LL...
...
```

but not

```
6 3414
Homsa
ANLLLLIVPI LI...
Phovi
INIISLIIPI LL...
...
```

which would be allowed under the sequential format by PHYLIP.

tree2treein

The main programs used to obtain split-specific support measures need to provide labels for these splits. The program `tree2treein` can be used to see how these labels correspond to the edges in a tree. It is run at the command line with the command

```
$ tree2treein treefile ntaxa < sequence_file > treefile-labeled
```

Here `treefile` is the name of the newick treefile of interest, `ntaxa` is the number of taxa and `sequence_file` is the name of the same sequence file used by the main programs. For instance, for the ratites data,

```
$ tree2treein RAxML_bestTree.24.out 24 < 24.dna > outtreefile
```

created the labelled tree file in Figure 1. Newick format is used so that common tree drawing programs will bring up the tree with branch labels. For example with the program `njplot` (Perrière and Gouy 1996) available at

```
http://pbil.univ-lyon1.fr/software/njplot.html
```

the tree can be brought up with

```
$ njplot outtreefile
```

Clicking on the button labelled "Bootstrap" will indicate the labels of branches along them. Alternatively, the split labels can be obtained from the output to the screen which, for each split, lists the taxa on either side of the split.

rbic_taxon_split

The program `rbic_taxon_split` obtains the `rbic_taxon_split` support measure for each split in an input tree and all taxa. It is run at the command line with

```
rbic_taxon_split -t treefile -b bootstrap_treefile -i sequence_file
```

and uses the same input format as `bptaxon_split`.

The output of `rbic_taxon_split` differs slightly from `bptaxon_split`. As with `bptaxon_split`, the output is a sequence of rows, with three entries

```
taxon split rbic_taxon_split
```

giving the taxon, split and support measure. However, when a split corresponds to a cherry in a tree and the taxon is one of the two taxa in that cherry, the entry is -100.10000. This is because upon removing the taxon, the split becomes a terminal split and is certain to be present in all bootstrap trees. To avoid inflating taxon-specific measures for this taxon, such values of `rbic_taxon_split` are ignored in averaging.

The other negative entries arise when a taxon is not part of a cherry. In that case, removing the taxa gives rise to one edge that is the combination of the two edges the taxon's terminal edge was adjacent to. Entries corresponding to the taxon and one of these adjacent edges are indicated with a minus sign. The program `taxa_split_support` utilizes this information in constructing taxon-specific support measures; see Wang et al. (2016) for additional information on the issue.

Using the ratites data as an example:

```
 $ rbic_taxon_split -t RAxML_bestTree.24.out -i 24.dna -b 24.boot1000
Cassowary    24   100.00000
Emu          24   100.00000
LBPenguin    24  -100.10000
GaviaStell   24  -100.10000
...
BrushTurke   29    94.30000
Chicken      29    94.30000
...
Alligator    29    94.30000
Caiman       29    94.30000
...
Ostrich      34   -50.40000
```

```
...
Ostrich      37   -50.40000
```

By contrast with `bptaxon_split`, `rbic_taxon_split` shows no variation for split 29. Split 24 is cherry with LBPenguin and GaviaStell forming that cherry which is the reason for those negative entries. The terminal edge leading to Ostrich is adjacent to the edges 34 and 37 which is the reason those two entries are negative.

taxa_split_support

The program `taxa_split_support` uses the output of either `bptaxon_split` or `rbic_taxon_split` and averages measures over either taxa or splits to get split-specific or taxon-specific measures of support. It is run at the command line with line as

```
taxa_split_support -m ntaxa -i taxa_split_file [-s]
```

where `ntaxa` is the number of taxa and `taxa_split_file` is the name of file gving the the output of either `bptaxon_split` or `rbic_taxon_split` and averages measures over either taxa or splits to get split-specific or taxon-specific measures of support. The -s option need not be present. If it is present, the output will be the split-specific averages. If it is not present, the output will be the taxon-specific measures. The output is of the form

```
split support
```

when the `-s` option is used. Here split is the label of the split and support is the average support measure for that split. In the case that the `-s` option is not present the output is

```
taxon support
```

where taxon is the name of the taxon in the sequence file and support the average support measure, averaged over splits.

For instance assuming that `bptaxon_split` had been run with

```
$ bptaxon_split -t RAxML_bestTree.24.out -i 24.dna -b > bptaxon_split.out
```

which would put the output in the file `bptaxon_split.out`, split-specific measures could be obtained with

```
$ taxa_split_support -i bptaxon_split.out -m 24 -s
 24 100.00
 25 100.00
 26  99.70
...
 44 100.00
```

Taxon-specific measures could be obtained with

```
$ taxa_split_support -i bptaxon_split.out -m 24
Cassowary   92.84
Emu         92.84
EasternMoa  91.77
```

References

Perrière, G. and Gouy, M. (1996) WWW-Query: An on-line retrieval system for biological sequence banks. Biochimie, 78, 364-369.

Wang, H., Susko, E. and Roger, A.J. (2016). Split-specific Bootstrap Measures for Quantifying Phylogenetic Stability and the Influence of Taxon Selection. *Molecular Phylogenetics and Evolution*. **105**:114–125.