Software for using bootstrap support for splits to contruct confidence regions for trees

Edward Susko

Department of Mathematics and Statistics, Dalhousie University

Introduction

The main programs are

1. `sboot_data`
2. `dboot_data`
3. `min_BP`
4. `ranked_spl`
5. `gp_splits`

The programs `sboot_data` and `dboot_data` generate single and double bootstrap data sets. The program `min_BP` obtains all trees with bootstrap support at least as large as a threshold. The program `ranked_spl` obtains a pre-specified number of trees that have splits in bootstrap trees. The program `gp_splits` converts a file with trees into trees of groups.

Installation

To create the directory that contains the source code

```
$ gzip -d minBP_soft.tar.gz
$ tar xvf minBP_soft.tar
```

This will create the directory `minBP_soft`. To create the executables in this directory type

```
$ make
```

This will create the program files `sboot_data`, `dboot_data`, `min_BP`, `ranked_spl` and `gp_splits`, which should be copied to a directory in your PATH. The file `minBP_soft.tar` and the directory `minBP_soft` can then be removed.

Generating bootstrap files

Single bootstrap

The program `sboot_data` outputs the original data set and then `B` additional bootstrapped data sets to the screen. It can be called at the command line with

```
$ sboot_data seed B < infile
```

where `infile` is a standard PHYLIP format infile. An integer `seed` is used to set the seed for the uniform random number generator.

Double bootstrap

The program `dboot_data` outputs the original data set, `B` bootstrapped data sets and an additional `B*B` bootstrapped data sets coming from each of the `B` bootstrapped data sets. It can be called at the command line with

```
$ dboot_data seed B < datafile
```

where `infile` is a standard PHYLIP format infile and the integer `seed` is used to set the seed for the uniform random number generator. The data sets output will be ordered as follows

Table 1: The ordering of the bootstrapped data sets output by `dboot_data` and the ordering that estimated Newick format trees should follow when input to `min_BP` with `zeta` set to -1.

	original data
b1:	bootstrap data set 1 of original data
	...
bB:	bootstrap data set B of original data
	bootstrap data set 1 of data b1
	...
	bootstrap data set B of data b1
	bootstrap data set 1 of data b2
	...
	bootstrap data set B of data bB

Obtaining all trees with bootstrap support greater than a threshold

The program for obtaining trees with large minBP is min_BP.

```
$  min_BP zeta alpha B treefile < infile
```

The file infile should be a PHYLIP format infile and is only used to obtain the names of the taxa. The estimated trees are assumed stored as Newick format trees in treefile. These estimated trees can and must be obtained from other software. However, it is assumed that the first tree is the estimated tree for the original sequence data, and that the next B trees are the estimated trees from B bootstrapped replicates of the original data. In the case that a threshold is to be determined through double bootstrapping, the ordering of trees in the file should be as in Table 1.

The integer zeta should be -1 or the cutoff that is to be used. If zeta is non-negative, the program will find all trees with minBP at least as large as zeta. Here, bootstrap support for a split is the number of times the split arose in bootstrap samples. For instance, if B is 1000 and you desire all trees with minBP greater than or equal to 52% you should input 520. In the case that zeta is non-negative the parameter alpha is ignored although it should be present (set it to 0.05).

If zeta is -1, the program will determine the threshold required to construct a (1-alpha)×100% confidence set of trees. The output to the screen in this case will indicate what this threshold is.

The program min_BP creates two files minbp.minbp and minbp.newick. The file minbp.newick gives the list of trees with minBP at least as large as the threshold. The ordering of trees is from highest to lowest min_BP. Edge lengths for terminal branches is set to 0 and edge lengths for internal edges to the bootstrap support for those edges. The corresponding minbp values are given as a single column in minbp.minbp.

As an example consider the following. A treefile mtprot_10101_treefile was constructed as described in Table 1 with 100 bootstrap trees and, for each of these 100 additional bootstrap trees. We obtain the cutoff for a 95% confidence set of trees with

```
$ min_BP -1 0.05 100 mtprot_10101_treefile < mtprot.dat
With B=100, the minimum bootstrap support cutoff for a 95.0% confidence region is 6
2 trees have minBP greater than or equal to 6
```

The cutoff is 6% minBP. We could look at the trees in minbp.newick, however, one of the reasons for using only 100 bootstrap samples is that the double bootstrap then requires an additional 100×100 bootstrap samples. Having obtained a minBP threshold of 6% we can now apply this to a lager set of bootstrap samples. The treefile `mtprot_treefile` contains 1000 bootstrapped trees.

```
$ min_BP 60 0.05 1000 mtprot_treefile < ../data/mtprot.dat
3 trees have minBP greater than or equal to 60
$ cat minbp.newick
(Orycu:0.00000,(Homsa:0.00000,(Phovi:0.00000,Bosta:0.00000):0.99900):...);
(Orycu:0.00000,(Phovi:0.00000...);
(Orycu:0.00000,Musmu:0.00000,...);
$ cat minbp.minbp
623
299
65
```

The terminal edge lengths are all set to 0. The other edge lengths give the bootstrap support for that split as a percentage. For instance, considering the first tree, 99.9% of the bootstrapped trees had `Phovi` and `Bosta` split from the rest of the taxa.

Obtaining a list of trees ranked according to minimum bootstrap support

The program `ranked_spl` obtains `ntree` trees that have splits in the trees obtained from `B` bootstrap samples, ranked according to their minbp

```
$ ranked_spl ntree B treefile < infile
```

The file `infile` should be a PHYLIP format infile. The estimated trees are assumed stored as Newick format trees in `treefile`. These estimated trees can and must be obtained from other software. However, it is assumed that the first tree is the estimated tree for the original sequence data, and that the next `B` trees are the estimated trees from `B` bootstrapped replicates of the original data.

```
$ ranked_spl 10 1000 mtprot_treefile < mtprot.dat
$ cat minbp.newick
(Orycu:0.000...);
(Orycu:0.000...);
(Orycu:0.000...);
(Musmu:0.000...);
```

```
((Phovi:0.00...);
((Phovi:0.00...);
((Phovi:0.00...);
(Musmu:0.000...);
((Orycu:0.00...);
(Musmu:0.000...);
$ cat minbp.minbp
623
299
65
56
56
14
13
13
1
1
```

Minimum bootstrap support for trees of groups

The program `gp_splits` can be useful for summary when the number of trees in the treefile is fairly large. It is called at the command line with

```
$ gp_splits B ntrees gpfile < infile
```

The program assumes that `min_BP` has been run with parameter `B` and PHYLIP format `infile`. It assumes that the files `minbp.newick` and `minbp.minbp` contain the output from the call of `min_BP`. The file `gpfile` should contain the definitions of the groups and have the form

number of groups
first group name   number of members   member 1 name   member 2 name   . . .
. . .
last group name   number of members   member 1 name   member 2 name   . . .

The program creates two files `minbp.minbp_new` and `minbp.newick_new` that contain the Newick format trees of groups and the corresponding MinBP values for each tree.

```
$ ./min_BP -1 0.05 100 a13_cov1_10101_outtree < a13_cov1.phy.txt
With B=100, the minimum bootstrap support cutoff for a 95.0% confidence region is 1
264 trees have minBP greater than or equal to 1
$ cat a13_cov1.phy.txt
13 269
A_Sulsol   STL ...
A_Desmob   STM ...
A_Aerper   STL ...
A_Pyraer   STL ...
A_Thecel   STT ...
A_Pyrhor   STT ...
A_Pyrwoe   STT ...
A_Arcful   STL ...
A_Metjan   STT ...
A_Metcan   STT ...
A_Halhal   STM ...
A_Halmar   STL ...
A_Theaci   STL ...
$ cat a13_gpfile
6
DSAP 4 A_Desmob A_Sulsol A_Aerper A_Pyraer
P 3 A_Pyrhor A_Pyrwoe A_Thecel
M 2 A_Metjan A_Metcan
H 2 A_Halhal A_Halmar
Af 1 A_Arcful
Ta 1  A_Theaci
$ ./gp_splits 100 264 a13_gpfile < a13_cov1.phy.txt
There were 72 trees that had splits incompatible with the groups
There were 16 trees of groups
$ cat minbp.newick_new
(P:0.00000,(Af:0 ... );
(P:0.00000,Af:0. ... );
(M:0.00000,Af:0. ... );
(P:0.00000,(M:0. ... );
(M:0.00000,(P:0. ... );
((Af:0.00000,(M: ... );
((M:0.00000,H:0. ... );
((M:0.00000,H:0. ... );
((H:0.00000,(M:0 ... );
(H:0.00000,(DSAP ... );
(H:0.00000,(M:0. ... );
(H:0.00000,(DSAP ... );
(H:0.00000,((M:0 ... );
(H:0.00000,(M:0. ... );
(P:0.00000,H:0.0 ... );
(Af:0.00000,((M: ... );
```

Limitations

The maximum number of taxa that can currently be considered is 80.

Susko, E. (2006). Using minimum bootstrap support for splits to construct confidence regions for trees. *Evolutionary Bioinformatics Online*. **2**:137–151.