aBPn: Adjusted First-order Correct Bootstrap Probabilities for Splits
Version 1.1.2

Edward Susko

Department of Mathematics and Statistics, Dalhousie University

Introduction

The main program, `aBPn`, implements the methods described in Susko (2009a) which utilize the results of Susko (2009b); please cite these references when using the software. It is shown in these references that 1 minus the adjusted bootstrap probability (BP) values returned by `aBPn` can be interpreted as p-values for the null hypothesis that the split is not present. This is a desirable interpretation that is consistent with the ways in which uncertainty is assessed throughout science and medicine. An adjusted BP larger than 95% gives significant evidence that a split is present in a similar way that a p-value less than 0.05 gives significant evidence against a null hypothesis in traditional statistical tests.

While BP for splits can generally be obtained for any phylogenetic method applied to aligned characters, the methods implemented by `aBPn` are for maximum likelihood (ML) analyses only. Moreover, while the methods described in Susko (2009a) can be applied to almost any ML analysis, the implementation here is limited to some of the more common models. Finally, `aBPn` will not fit (estimate) an ML model and will not obtain BP. The input to the routine is the result of ML analysis as well as BP from that analysis. Packages such as PHYLIP (Felsenstein, 1989, 2004), TREE-PUZZLE (Schmidt et al. 2002) and PAML (Yang 1997, 2007) can be used to obtain the input for `aBPn`

Installation

The main program `aBPn` is compiled from C and Fortran 77 source code. It utilizes the Fortran 77 code developed by Alan Genz to implement the numerical integration algorithms described in Genz (2004).

In cases where there are a large number of taxa or amino acid data is being considered, it will usually be necessary to either use observed information matrices, which require that the original alignment be available, or use Monte Carlo approximations to the expected information matrices utilized in calculations. The latter option requires that the program `seq-gen` (Rambaut and Grassly 1997) is available. This program can be downloaded from the aesthetically pleasing web site

`http://tree.bio.ed.ac.uk/software/seqgen/`

Precompiled binaries are available for Windows. To install `aBPn` on other systems

1. Download and unpack the software:

```
$ tar zxf aBPn.tar.gz
```

This will create a directory **aBPn** that contains the source code and test input files.

2. Change directories to **aBPn** and create the main program file **aBPn** with the **make** command.

```
$ cd aBPn
$ make
```

This should produce the main program file **aBPn** which can be copied to a location in your PATH. To test the program, still in the **aBPn** directory, issue the commands

```
$ ./aBPn mtprot.ctl > mtprot.outtree
$ ./aBPn hiv1.ctl > hiv1.outtree
```

This will create Newick treefiles, **mtprot.outtree** and **hiv1.outtree** with adjusted BP as labels that should be comparable to those reported in Susko (2009b); see the section Output below for additional information.

The source code has been compiled and tested using gcc and gfortran (versions 4.4.0) on an CentOS Linux distribution (release 5.3). While the program has not been tested on another platform, it should compile under any Linux distribution as well as Mac OS X, assuming a fortran compiler is installed.

Input

The program can be run at the command line with the command

```
$ aBPn controlfile
```

All input to the routine is through a main control file, **controlfile**. The control file is similar in format to the control files used by the programs **baseml** and **codeml** in the PAML package (Yang 1997, 2007). For instance, the **model** variable specifies the substitution model and gives a subset of the models available in PAML, with the same numbering scheme. As a running example, consider one of the test files, **hiv1.ctl**:

```
* Example control file
ntaxa =  6                * number of taxa
treefile =  hiv1-BP1.tree * Tree with ML edge-lengths

nchar =  4                * nucleotide data
model =  7                * GTR model
frfile =  hiv1.fr         * file with frequencies
Qfile =     hiv1.Q        * file with rate matrix
```

```
alpha =  0.28173            * gamma model alpha parameter
ncatG =  8                  * number of categories for gamma model

nsim =  10000               * number of simulated pseudo data sets
iseed =  3655               * starting seed for random number generators
```

As with PAML control files, blank lines are allowed and all text following a '*' till the end of a line is treated as a comment. The word on the left of an equal sign gives a control variable and the word on the right gives the value of that variable. Spaces are required on both side of an equal sign. The order of variables is unimportant. The control variables are as follows. All variables not indicated as optional are required.

ntaxa: An integer giving the number of taxa.

treefile: The name of a file containing the tree of interest with ML estimates of edge lengths and BP as labels.

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

As an illustration consider the running example. The treefile is hiv1BP1.tree-:

((BRU:0.032918,NDK:0.03422)100:0.088856,Q23:0.038153,
(U455:0.037855,(90CF11697:0.026739,93TH057:0.021763)100:0.022687)76:0.014752);

The substring

(90CF11697:0.026739,93TH057:0.021763)100:0.022687

indicates that the taxa 90CF11697 and 93TH057 form a subtree, that the ML estimate of the length of the edge leading to 90CF11697 was 0.026739, the ML estimate of the length of the edge leading to 93TH057 was 0.021763 and the ML estimate of the length of the edge separating these two taxa from the rest was 0.022687. The label '100' for the edge separating 90CF11697 and 93TH057 from the rest indicates that BP for this edge was 100%. BP can be indicated in labels either in percentages, as in the above example, or as proportions between 0 and 1.

Allowable features of the Newick standard that will likely create difficulties are:

1. Quoted labels.

2. Nested use of the characters ']' and/or ']' in comments. The characters '[' and ']' can only be used to delimit comments and cannot be used within comments.

3. Long leaf labels. A limit of 10 non-null characters is allowed for leaf names.

4. Underscores are not converted to blanks.

nchar: An optional integer indicating that the model was for nucleotide data (nchar=4) or amino acid data (nchar=20). The default value is 4.

model: An integer code for the substitution model. For nucleotide data (nchar=4), the models currently implemented are

model	Model
0	JC
2	F81
3	F84
4	HKY
7	GTR

and for amino acid data (nchar=20) the models currently implemented are

model	Model
0	Poisson
1	Proportional
2	Empirical
3	Empirical+F
8	REVaa

The documentation for the PAML package gives a good description of the models listed and can fit all of them.

The GTR and REVaa models refer to the most general time-reversible models in the nucleotide and amino acid case, respectively. The Poisson and Proportional models are the analogues of the JC and F81 models for amino acid data. The Poisson and Proportional models have substitution probabilities

$$P_{ij}(t) = \begin{cases} \pi_j + (1 - \pi_j) \exp[-\mu t] & \text{if } i = j \\ \pi_j - \pi_j \exp[-\mu t] & \text{otherwise} \end{cases}$$

where $\mu = [\sum \pi_i (1 - \pi_i)]^{-1}$ and π_j gives the stationary of the jth amino acid. In the Poisson model, the frequencies are all $1/20$

When `model=2` or 3 an empirical model is fit. The model is specified by the variable `aaRatefile`. When `model=2`, the stationary frequencies are the stationary frequencies of the specified empirical model.

When `model=3`, the stationary frequencies must be specified in a file `frfile`. In this case, the empirical model is used to specify the exchangeabilities. The exchangeability of amino acid i and j is defined as

$$S_{ij} = Q_{ij}/\pi_j$$

where, for the specified empirical model, Q_{ij} is the rate of substitution from i to j and π_j is the stationary frequency j. When `model=3`, the rate of substitution from i to j is

$$\tilde{Q}_{ij} = S_{ij}\tilde{\pi}_j$$

where $\tilde{\pi}_j$ is the frequency of j specified in `frfile`.

`aaRatefile`: Only required for empirical amino acid models (`model=2` or 3 and `nchar=20`). The name of the empirical model to fit. The models currently implemented are

`dayhoff.dat`	Dayhoff or PAM	Dayhoff et al. (1978)
`jones.dat`	JTT	Jones et al. (1992)
`wag.dat`	WAG	Whelan and Goldman (2001)
`mtREV24.dat`	mtREV	Adachi and Hasegawa (1996)
`lg.dat`	LG	Le and Gascuel (2008)

The naming scheme was chosen to be consistent with PAML. However, `aaRatefile` is not actually the name of file, rather it identifies a model.

`frfile`: The name of a file containing the model's stationary frequencies nucleotides or amino acids, each separated by white space. It is not required for the JC, proportional and empirical models (`model=0` or `model=2`, `nchar=20`). In the example, the entry in the control file is

```
frfile =  hiv1.fr          * file with frequencies
```

and the file `hiv1.fr` contains

```
0.22093 0.16814 0.39233 0.2186
```

which give the frequencies with which the nucleotides A, C, G and T occurred in the alignment. Note that this ordering differs from the T, C, A and G ordering of PAML. Amino acids are ordered alphabetically:

```
    alanine, arginine, asparagine, aspartic, cysteine,
    glutamine, glutamic, glycine, histidine, isoleucine,
    leucine, lysine, methionine, phenylalanine, proline,
    serine, threonine, tryptophan, tyrosine, valine
```

which is the same ordering used by most phylogenetic packages including PAML, PHYLIP and TREE-PUZZLE.

Qfile: Only required for the general time reversible model, GTR or REVaa (model=7, nchar=4 or model=8, nchar=20). The name of a file containing the entries of the rate matrix separated by blanks.

The ordering of nucleotides or amino acids should match the ordering in frfile. In the example, the entry in the control file is

```
Qfile =     hiv1.Q        * file with rate matrix
```

and the file hiv1.Q contains the entries

```
-0.809338 0.1209 0.636418 0.05202
0.2821 -1.323457 0.089596 0.951762
1.142202 0.068915 -1.244881 0.033764
0.092375 0.72433 0.033407 -0.850113
```

Considering the (3,2) entry, we see that the rate of substitution from G to C is 0.068915.

kappa or ttratio: One of these is required for the F84 and HKY models (model=3 or 4 and nchar=4). A real number giving the κ parameter for the model. The F84 model has a rate matrix proportional to

$$
\begin{bmatrix}
\cdot & \pi_C & (1 + \kappa/\pi_R)\pi_G & \pi_T \\
\pi_A & \cdot & \pi_G & (1 + \kappa/\pi_Y)\pi_T \\
(1 + \kappa/\pi_R)\pi_A & \pi_C & \cdot & \pi_T \\
\pi_A & (1 + \kappa/\pi_Y)\pi_C & \pi_G & \cdot
\end{bmatrix}
$$

where $\pi_R = \pi_A + \pi_G$ and $\pi_Y = \pi_C + \pi_T$. The HKY model has a rate matrix proportional to

$$
\begin{bmatrix}
\cdot & \pi_C & \kappa\pi_G & \pi_T \\
\pi_A & \cdot & \pi_G & \kappa\pi_T \\
\kappa\pi_A & \pi_C & \cdot & \pi_T \\
\pi_A & \kappa\pi_C & \pi_G & \cdot
\end{bmatrix}
$$

The transition-transversion ratio (ttratio) is related to the κ parameter in the F84 model through

$$
R = \kappa \times \frac{\pi_A\pi_G/\pi_R + \pi_C\pi_T/\pi_Y}{\pi_R\pi_Y} + \frac{\pi_A\pi_G + \pi_C\pi_T}{\pi_R\pi_Y}
$$

where $\pi_R = \pi_A + \pi_G$ and $\pi_Y = \pi_C + \pi_T$. For the HKY model the relationship is

$$R = \kappa \times \frac{\pi_A\pi_G + \pi_C\pi_T}{\pi_R\pi_Y}$$

alpha: Only required if a gamma rates-across-sites model (Yang 1994) is desired. A real value giving the shape parameter of the gamma distribution.

ncatG: Only used if a discrete gamma rates-across-sites model was fit. Optionally, an integer giving the number of categories to use in the discrete approximation. The default is 4. The discrete gamma approximation used is the same as the default of PAML 4.2; the representative rate is the conditional mean for the class.

nrate: Optionally used to specify the number of rate classes in an arbitrary discrete rates-across-sites model. The variable is not used if **alpha** is set. If **alpha** is not set the default is to fit an equal rates model.

ratefile: Only used if **alpha** is not set and **nrate**> 1. The name of a file with rates and weights for the arbitrary discrete rates-across-sites model. Each line should give a rate followed by a corresponding weight. For example, the file

```
0.00026 0.25
0.02369 0.25
0.33291 0.25
3.64313 0.25
```

indicates a rates-across-sites model with four rates where, for instance, there is a 25% chance that a site will have a rate multiplier 3.64313. This consequently allows for implementations of the discrete gamma rates-across-sites models that give different discrete rates or weights than PAML to the same shape parameter.

tinfo: Optionally one of 0, 1 or 2. The default is 0.

Adjusted bootstrap support uses expected information matrices that can require substantial calculation when either the number of taxa is large or amino acid data is considered. Expected information matrices can be approximated through simulation. The value **tinfo**=0 indicates that calculation should be exact and the value **tinfo**=1 indicates that simulation should be used for calculation. The value **tinfo**=1 assumes that the program **seq-gen** is available.

An alternative is to use the observed information matrix. Where the expected information matrix is the expected value of the second derivative matrix of the log likelihood multiplied by -1, the observed information is the actual second derivative matrix of the log likelihood multiplied by -1. Theory indicates that, assuming the generating model is correct, the observed and expected information matrix should be approximately the same with large numbers of

sites and the use of either will yield first-order correct adjusted BP values. To use the observed information indicate option `tinfo=2`. This will allow the original alignment to be read in from the file `seqfile`. Note that this option could also be used to approximate the expected information matrix using a sequence file simulated from some other routine than `seq-gen`.

`nsite`: Only used if simulation is used to approximate the information matrix (`tinfo=1`). Optionally, the number of sites to use in information matrix simulation. The default is 10,000.

`seqfile`: This is the name of a file containing the sequence data that will be used for information matrix calculation when `tinfo=2`. If this is the sequence data that was used to obtain ML estimates, the information matrix obtained is the observed information.

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.

`nsim`: Optionally, the number of pseudo data sets to simulate. The default is 10,000. The `aBPn` routine repeatedly simulates trivariate normal variates. Each of these simulations is akin to the simulation of a data set; see Susko (2009a).

`iseed`: An integer that will set the seed. Changing this will change the sequence of random generations and can be useful in evaluating the sensitivity of the Monte Carlo portions of the calculations.

`ismle`: Optionally, one of 0 or 1. The default is 0. If set to 1, corrections are made for the input tree being the ML tree.

Output

The output is a tree (to the screen or stdout) and is of the same format as the input tree; see `treefilename`. BP will be replaced by adjusted BP as labels.

In the example, the `type` was 'l' and the treefile contained the tree

```
((BRU:0.032918,NDK:0.03422)100:0.088856,Q23:0.038153,
(U455:0.037855,(90CF11697:0.026739,93TH057:0.021763)100:0.022687)76:0.014752);
```

The output was

```
(Q23:0.03815,(U455:0.03785,(90CF11697:0.02674,93TH057:0.02176)100:0.02269)90:0.01475,
(BRU:0.03292,NDK:0.03422)100:0.08886);
```

From which we see that the BP of 76% for the split of U455, 90CF11697 and 93TH057 from the rest is adjusted upwards to 90%. Note that, although the trees are the same for both input and output, the Newick representations differ.

Limitations

Very few reasons for error are output. Analyses should not include identical sequences and very similar sequences can create difficulties. The number of sites should be larger than the number of taxa.

References

Adachi, J. and Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. J. of Mol. Evol. 42:459–468.

Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978). A model of evolutionary change in proteins. pp 345–352, Atlas of protein sequence and structure. Vol. 5, Suppl. 3. National Biomedical Research Foundation. Washington D.C.

Felsenstein, J. (2004). PHYLIP Phylogeny Inference Package (version 3.6). Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (version 3.2). Cladistics 5: 164-166.

Genz A. (2004). Numerical computation of rectangular bivariate and trivariate normal and t probabilities. Statistics and Computing. 14:251–264.

Jones, D.T., Taylor, W.R., and Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. CABIOS 8:275–282.

Kosiol, C. and Goldman, N. (2005). Different versions of the Dayhoff rate matrix. Mol. Biol. Evol. 22:193–199.

Le, S.Q. and Gascuel, O. (2008). An improved general amino acid replacement matrix. Mol. Biol. Evol. 25:1307–1320.

Rambaut, A. and Grassly, N. C. (1997). Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13: 235-238.

Schmidt, H. A., Strimmer, K., Vingron, M. and von Haesler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504.

Susko, E. (2009a). Bootstrap support is not first order correct. Syst. Biol. 58:211-233.

Susko, E. (2009b). First-order correct bootstrap support adjustments for splits that allow hypothesis testing when using maximum likelihood estimation. Mol. Biol. Evol. 27:1621–1629.

Whelan, S. and Goldman, N. (2001). A general empirical mdoel of protein evolution derived from multiple protein families using a maximum likelihood approach. Mol. Biol. Evol. 18:691–691.

Yang, Z. (2007). PAML 4: a program for phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Yang, Z. (1997). PAML: a program for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13:555–556.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates across sites: approximate methods. J. Mol. Evol. 39:306–314.