SDBP: An easy-to-use R program package for assessing reliability of estimated phylogenetic trees based on the speedy double bootstrap method

Aizhen Ren[1], **Takashi Ishida**[2], **and Yutaka Akiyama**[2]
[1]Department of Mathematical and Computing Science, Tokyo Institute of Technology,
W8-76, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8550, JAPAN
[2]Department of Computer Science, Graduate School of Information Science and Engineering,
Tokyo Institute of Technology, W8-76, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8550, JAPAN

Abstract— *Evaluating the reliability of estimated phylogenetic trees is of critical importance in the field of molecular phylogenetics, and for other endeavors that depend on accurate phylogenetic reconstruction. The bootstrap method is a well-known computational approach to assessing phylogenetic trees, and more generally for assessing the reliability of statistical models. However, it is known to be biased under certain circumstances, calling into question the accuracy of the method. Therefore, several advanced bootstrap methods have been developed to achieve higher accuracy, one of which is the speedy double bootstrap approach (sDBP-method). In the phylogenetic tree selection problem, it has been shown that the sDBP-method has comparable accuracy to the double bootstrap approach and is much more computationally efficient. In this study, we thus develop an R package named SDBP, which is an implementation of our sDBP-method on a statistical software R to assesse the reliability of phylogenetic trees. We are confident that biologists will benefit from our sDBP-method and SDBP package.*

Keywords: SDBP, Speedy double bootstrap method, Phylogenetic trees, Reliability, Rapid computation, R package

1. Introduction

The analytical methods used in the field of molecular phylogenetics are important basic tools for reconstructing the evolutionary history (phylogenetic relationships) of molecules and organisms. Molecular phylogenetic methods are primarily used in the context of biological systematics, but they also find applications in a wide variety of other fields as diverse as community ecology [1], biogeography [2] and proteomics, including inference of the similarity of protein-protein interactions [3]. Many methods for phylogenetic reconstruction have been developed and are in regular use [4]. However, those based on maximum likelihood estimation have proved most effective for reconstructing phylogenies using molecular sequence data (DNA, protein, etc.). Early work on this application of maximum likelihood was conducted by [5], whose approach involved computing the maximum likelihood value for many topologies and selecting the topology with the highest likelihood (the maximum likelihood (ML) tree) as the most probable candidate for the true topology.

It must be noted that maximum likelihood values are dependent on the particular characteristics of a random variable; that is, the molecular sequences that constitute the underlying data for phylogenetic reconstruction. Thus, some analysis of the statistical reliability of the estimated ML tree or multiple alternative trees should be undertaken. Statistical hypothesis testing is commonly used for this purpose, and the 'bootstrapping' technique is a well-known computational method for calculating reliability when a simple mathematical formula is difficult to derive. Bootstrapping is a resampling method that approximates a random sample by creating a bootstrap sample, generated by random sampling with replacement from the original single data set. In the context of phylogenetic tree selection, Felsenstein [6] proposed the use of bootstrapping to place confidence intervals on phylogenies. He defined the p-value of a tree according to a frequency called the bootstrap probability (BP); the proportion of bootstrap pseudoreplicates of the original data set in which the tree is found to be optimal. However, it is known that under some circumstances the naive bootstrap probability can be biased [7], [8]. Thus, some advanced bootstrap methods have been proposed, to achieve higher accuracy [9], [10], [11].

Among these, the double bootstrap method (DBP-method) [9], [10] has been shown to be third-order accurate and is potentially a useful measure of phylogenetic tree support. However, the method has a huge computational cost. To overcome the computational burden in the phylogenetic tree selection problem, we have previously proposed a 'speedy' double bootstrap (sDBP-method) method to compute the reliability of phylogenetic trees [12]. In the phylogenetic tree selection problem, our previous work [12] has been shown that the sDBP-method has comparable accuracy to the DBP-method and is much more computationally efficient. Because, it is well known that a good statistical method is not in itself sufficient, we also need to develop an easy-to-

use computer tool. We thus develop the R package named SDBP, which is an implementation of our sDBP-method on a statistical software R to assess the reliability of phylogenetic trees. We are confident that biologists, who may not have advanced computer skills, will benefit from our sDBP-method and SDBP package.

R is a language and environment for statistical computing and graphics. It is an open-source GNU project based on the S language and environment developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. We can summarize why we implemented our method in R as follows. At first, R provides a wide variety of statistical (linear and non-linear modeling, classical statistical tests, time-series analysis, classification, clustering, $\cdots$) and graphical techniques, and is highly extensible. In addition, it is important that R is not only applicable to statistical fields of research, but also to the biological field. Genome analysis, including GneABEL [13], and areas related to biotechnology also have a great many applicable R packages. Finally, R is available under the terms of the Free Software Foundation's GNU General Public License in source code form. It can be compiled and run on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows, and MacOS.

This paper is organized as follows. We first give some background, and then briefly introduce the mathematical theory of the sDBP-method and its algorithm for assessing the reliability of phylogenetic trees. Next, we describe the basic usage of our package SDBP using the mammalian mitochondrial data from [14]. Finally, we describe the results.

2. Theory and Algorithm

2.1 The reliability of a phylogenetic tree

In this study, homologous sites of aligned molecular sequence data are regarded as the units for sampling, and we use DNA data as our example for the following methodological descriptions. Suppose we have m homologous sequences, each with n nucleotide sites. These data can be represented as an $m \times n$ matrix $\mathbf{X} = \{x_{jh}\} = \{\mathbf{x_1}, \cdots, \mathbf{x_n}\}$, where $\mathbf{x}_h$ is the value of the h-th site and x_{jh} is one of the four deoxyribonucleotides (T, C, A, or G).

$$\begin{array}{rcccc} Species\ 1 & : & x_{11} & x_{12} & \cdots & x_{1n} \\ Species\ 2 & : & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & & & & \vdots & \\ Species\ m & : & x_{m1} & x_{m2} & \cdots & x_{mn} \end{array} \quad (1)$$

The log-likelihood can be expressed as

$$l(\theta; \mathbf{X}) = \sum_{h=1}^{n} log f(\mathbf{x}_h; \theta), \quad (2)$$

where $f(\mathbf{x}_h; \theta) = f(x_{1h}, x_{2h}, \cdots, x_{mh}; \theta)$ is the probability that at a particular homologous site, species 1 has base x_{1h}, species 2 has x_{2h} and species m has x_{mh}. The vector θ denotes unknown parameters such as the edge lengths (branch lengths) of a tree, and the base substitution rates along these branches. Here we assume that the base substitution rates have already been estimated, so θ denotes only the unknown edge lengths. For a given tree topology, θ is estimated by maximizing the log-likelihood, and the maximum log-likelihood of any tree topology i is given by

$$l_i(\hat{\theta}_i; \mathbf{X}) = \sum_{h=1}^{n} log f_i(\mathbf{x}_h; \hat{\theta}_i). \quad (3)$$

The topology with the highest value of $l(\hat{\theta}; \mathbf{X})$ is the maximum likelihood phylogeny (T_{ML}) for the data set $\mathbf{X}$, and is thus the most likely candidate for the true topology. To define null hypotheses for performing model comparisons, we must consider the true distribution for a random variable $\mathbf{x}$ can be expressed as

$$q(\mathbf{x}) \quad (4)$$

And the expectation of $l_i(\hat{\theta}_i; \mathbf{X}), i = 1, \cdots, K$ with respect to

$$(\mathbf{x_1}, \cdots, \mathbf{x_n}) \overset{i.i.d.}{\sim} q(\cdot) \quad (5)$$

can be expressed as

$$\mu_i \quad = \quad E_q[l_i(\hat{\theta}_i; \mathbf{X})] \quad (6)$$

If we assume that tree T_1 is the best topology, the null and alternative hypotheses will then be

$$H_1 : \mu_1 = max_{i=1,\cdots,K}\ \mu_i \quad vs. \quad H_1^A : \ others, \quad (7)$$

and we must continue performing these comparisons as many times as is necessary, assuming in turn that tree $T_i, i = 2, \cdots, K$ is the best topology. Note that the null hypothesis H_1 involves multiple comparisons with the "best" topology [15]. As can be seen from equation (7), the null contains $K - 1$ hypotheses such that

$$H_{1j} : \mu_1 \geq \mu_j, j = 2, \cdots, K. \quad (8)$$

The null hypothesis H_1 is a polyhedral convex cone and $B(h_1)$, which is the boundary of H_1, is nonsmooth at the vertex as well as on the faces of dimension less than $K - 1$. Shimodaira and Hasegawa [14] proposed a multiple comparisons procedure (the SH-test) to test H_1, but this was shown to be overly conservative because they assumed that the parameter configuration is $\mu_1 = \mu_2 = \cdots = \mu_K$, that is, the least favorable configuration or the vertex of $B(h_1)$ [16]. A different method (the AU-test), which uses a multiscale bootstrap technique to obtain third-order accurate p-values for testing the null hypothesis, has also been proposed [11]. In our previous work [12], we developed an algorithm using an advanced bootstrap method [10] that was also able to

provide third-order accurate p-values to assess statistical reliability of phylogenetic trees. We call it the speedy double bootstrap method, which will be considered in the following subsection.

2.2 The theory of speedy double bootstrap method

In this subsection, it is necessary to review the theory of the speedy double bootstrap method. For this, we start by explaining the third-order accurate p-value. It was first proposed by [17] for the multivariate normal model, which can be represented as

$$\mathbf{Y} \overset{i.i.d.}{\sim} N_t(\eta, I_t). \tag{9}$$

This normal model is a simplification of reality. Let $\mathcal{H} \subset R^t$ be an arbitrarily-shaped region with smooth boundaries denoted by $B(h)$. We want to calculate a p-value $p(y)$ for testing the null hypothesis $\eta \in \mathcal{H}$. According to [17], when the true parameter η is on the boundary surface $B(h)$, the third-order accurate p-value can be expressed as

$$p(y) = 1 - \Phi(d - c), \tag{10}$$

where d is the signed distance from y to $\hat{\eta}(y)$, with a positive or negative sign when y is outside or inside $\mathcal{H}$, respectively. The point $\hat{\eta}(y)$ is the closest point to y (in Euclidean distance) on the surface $B(h)$, and c in equation (10) is a quantity related to the curvature of $B(h)$ at the point $\hat{\eta}(y)$. The speedy double bootstrap method of [10] (named later by [12]) begins with a bootstrap resampling from the multivariate normal model with distribution

$$\mathbf{Y}^* \overset{i.i.d.}{\sim} N_t(\hat{\eta}(y), I_t). \tag{11}$$

It then uses Y^* to calculate d^*, which is the signed distance from Y^* to $B(h)$. According to [10], the third-order accurate p-value obtained by the sDBP-method can be expressed as

$$1 - \Phi(d - c) = P(d^* > d; \hat{\eta}(y)) + O(n^{-3/2}). \tag{12}$$

2.3 The algorithm for the speedy double bootstrap method for phylogenetic trees

We now return to the problem of phylogenetic trees, as seen in H_1 and the vector $(l_1, \cdots, l_K)$. We describe the algorithm using the sDBP-method to calculate the p-value of H_1. First, we find a vector corresponding to $\hat{\eta}(y)$ in equation (11). According to [18], the maximum log-likelihood vector

$$\mathbf{l} = (l_1(\hat{\theta}_1), \cdots, l_K(\hat{\theta}_K)) \tag{13}$$

asymptotically follows a multivariate normal distribution, the mean vector of which is

$$\mu = (\mu_1, \cdots, \mu_K). \tag{14}$$

Note that the vector $\mathbf{l}$ in equation (13) is an unrestricted maximum likelihood estimate for μ. Because we assumed that $\mu_1 = max_{i=1,\cdots,K}\ \mu_i$ in H_1, under this restriction the restricted estimator for μ can be estimated using the PAVA (pool adjacent violators algorithm) [19] method, and is expressed as

$$\hat{\mu} = (\hat{\mu}_1, \cdots, \hat{\mu}_K). \tag{15}$$

We then excise a subset $W \in \{1, \cdots, K\}$, including the element 1, so that

$$\begin{aligned} \hat{\mu}_1 &= \frac{\sum_{j \in W} l_j(\hat{\theta}_j)}{\#W}, \\ \hat{\mu}_j &= min(\hat{\mu}_1, l_j(\hat{\theta}_j)), \quad j \in \{2, \cdots, K\}. \end{aligned} \tag{16}$$

The vector $\hat{\mu} = (\hat{\mu}_1, \cdots, \hat{\mu}_K)$ corresponds to $\hat{\eta}(y)$. Also, the covariance matrix of the vector $(l_1, l_2, \cdots, l_K)$ can be estimated by $\Sigma = (\sigma_{ij})$, with σ_{ij} given as

$$\begin{aligned} \frac{n}{n-1} \sum_{h=1}^{n} \left[logf_i(\mathbf{x_h}; \hat{\theta}_\mathbf{i}) - \frac{1}{n} \sum_{h=1}^{n} logf_i(\mathbf{x_h}; \hat{\theta}_\mathbf{i}) \right] \\ \times \left[logf_j(\mathbf{x_h}; \hat{\theta}_\mathbf{j}) - \frac{1}{n} \sum_{h=1}^{n} logf_j(\mathbf{x_h}; \hat{\theta}_\mathbf{j}) \right]. \end{aligned} \tag{17}$$

We then need to calculate two other quantities corresponding to d^* and d in equation (12). To do this, we generate $B1$, for example 10000 bootstrap pseudoreplicates of the vector $(\hat{\mu}_1, \hat{\mu}_2, \cdots, \hat{\mu}_K)$ in equation (15). The pseudoreplicates $(\hat{\mu}_1^{*(b1)}, \cdots, \hat{\mu}_K^{*(b1)}), b1 = 1, \cdots, B1$ are sampled from

$$(\hat{\mu}_1^{*(b1)}, \cdots, \hat{\mu}_K^{*(b1)})^T \overset{i.i.d.}{\sim} N_K((\hat{\mu}_1, \hat{\mu}_2, \cdots, \hat{\mu}_K)^T, \Sigma), \tag{18}$$

where T represents the transpose, and Σ is used as above. The vectors $(\hat{\mu}_1^{*(b1)}, \cdots, \hat{\mu}_K^{*(b1)})$ constitute the first-order (first-tier) bootstrap pseudoreplicates. Now, d^* and d in equation (12) can be written as

$$\begin{aligned} d^{*(b1)} &= max_{j=2,\cdots,K} \hat{\mu}_j^{*(b1)} - \hat{\mu}_1^{*(b1)}, \\ d &= max_{j=2,\cdots,K} l_j - l_1. \end{aligned} \tag{19}$$

Next, we calculate the p-value for H_1, defined below and also denoted by $sDBP$:

$$sDBP = \frac{\#(d^{*(b1)} > d)}{B1}. \tag{20}$$

In exactly the same way as shown for H_1, we can apply the sDBP method to all other hypotheses $H_k, k = 2, \cdots, K$.

3. Implementation

3.1 Implementation in R

We have implemented the sDBP algorithm for phylogenetic inference as a R package. Our package is named `SDBP`, and calculates p-values for phylogenetic trees. It can be used in combination with several other functions or packages in R.

The package was written in the S language using the S3 object system, and consists of a number of user-level objects:

`sdbp`, `sdbpk`, `bpk`, `bp`, `dbpk`, and `mam20`. The following subsections describe how to use these user-level objects. The SDBP provides three types of p-value: the sDBP (speedy double bootstrap probability), the DBP (double bootstrap probability), and the BP (bootstrap probability).

3.2 Usage – Using the mammalian mitochondrial protein sequences

In this subsection, we explain how to use SDBP with the mammalian mitochondrial protein sequences data from [14]. This data set included in file mam15-files, which can be download from scaleboot Home Page.
http://www.is.titech.ac.jp/˜shimo/prog/scaleboot/index.html
Scaleboot also is an R package. The mammalian protein data set includes sequences of n = 3414 amino acids from six mammalian species (human, seal, cow, rabbit, mouse, and opossum). The proteins coded for in the mammalian mitochondrial genome are ND1, ND2, COX1, COX2, ATP8, ATP6, COX3, ND3, ND4L, ND4, ND5, and CYTB. The clade $\{seal, cow\}$ was significantly supported in preliminary analyses, so only the 15 unrooted trees (see Table 1) that included this clade were considered in our comparisons (the opossum is the outgroup). Now, the number of trees K is 15, and the sample size n is 3414. Hypothesis H_1 denotes that $\mu_1 = max_{i=1,\cdots,15}\ \mu_i$. Our aim is to calculate the p-values for hypothesis H_1 as well as $H_i, i = 2, \cdots, 15$.

In advance, we used the software package PAML [20], to calculate the site-wise log-likelihood for each tree. The output file is file mam15.lnf. The format of mam15.lnf is not available for our package, so the format was changed using CONSEL [21] by executing the command "seqmt $--$paml mam15.lnf". Thus we obtain the site-wise log-likelihood matrix saved in the file mam15.mt for each tree. The file mam15.mt obtained by CONSEL should be placed in the R work directory. The 15 tree topologies is in the file mam15.tpl, that can be found in mam15-files.

Our SDBP package is built under R version 3.0.0. Therefore, this R version (or later) is needed to install our package. For Windows OS, after booting R, choose the tab **Packages** in the upper tool-bar and select the tab **Install Package(s) from zip files** option, then choose the **SDBP_1.0.zip** file downloaded from CRAN, the official R package archive.

For using the command line on UNIX platforms to install the source version package **SDBP_1.0.tar.gz** downloaded from CRAN, just write the following command:

```
R CMD INSTALL SDBP_1.0.tar.gz
```

and boot R via the command line using the command.

```
R
```

Then, the following on the **R console** command line to load our package (the following command can be typed on both Unix and regular Windows machines):

```
> library("SDBP")# load our package
```

And then, read the data named mam15.mt.

```
# read scaleboot for reading .mt files
> library(scaleboot)
> dat<-read.mt(mam15.mt)
> dim(dat)# dat matrix demation
[1] 3414   15
```

Calculating the sDBP-value for each tree requires the following line. Thus our package is as easy-to-use as R package.

```
> result <- sdbp.default(dat)
> result
```

We performed this on a personal computer with the following specifications: 2.50 GHz CPU (Core (TM) i5-2520M CPU) and 8.00 GB RAM. The results are output in decreasing order of log-likelihood.

```
Call:
sdbp.default(dat = dat)

Speedy double bootstrap probabilities:
t1     t3     t2     t5     t6     t7
0.5828 0.3905 0.2237 0.1191 0.1109 0.0681 ...
```

Calculating the stand error for each value, we can use the command summary.

```
> summary(result)
```

The output is

```
Call:
speedy.default(dat = dat)

   stdErr p.value
t1 0.0049  0.5717
t3 0.0049  0.3928
t2 0.0041  0.2173
t5 0.0032  0.1136
...
attr(,"class")
[1] "summary.sdbp"
```

This command is for testing hypothesis H_1 in equation (7) for tree 1 in the topology file mam15.tpl, using the algorithm in subsection 2.3. Also, for testing hypotheses $H_i : \mu_i = max_{k=1,\cdots,15}\mu_k, i = 2, \cdots, 15$ for tree 2, $\cdots$, tree 15 in the topology file mam15.tpl, we repeatedly use the algorithm in subsection 2.3. However, the algorithm for testing one of the hypothesis H_i of $H_i, i = 2, \cdots, 15$ is a little different from the algorithm for testing hypothesis H_1. The difference is that we calculate the projection $\hat{\mu}$ of the maximum log-likelihood vector $\mathbf{l} = (l_1, \cdots, l_{15})$ for each hypothesis and the signed distances. For example, the projection vector $\hat{\mu}$ of the maximum log-likelihood vector $\mathbf{l} = (l_1, \cdots, l_{15})$ under hypothesis H_2 is obtained using the following equations.

$$\hat{\mu}_2 = \frac{\sum_{j \in W_2} l_j(\hat{\theta}_j)}{\#W_2}, \hat{\mu}_j = min(\hat{\mu}_2, l_j(\hat{\theta}_j)), \quad j \in \{1, 3, \cdots, 15\},$$

where W_2 is subset of the numbers $\{1, \cdots, 15\}$ including the element 2. For details of the implementation of the PAVA method, see our R source code in sdbp.R in SDBP. The signed distances are

$$d^{*(b1)} = max_{j=1,3,\cdots,15}\hat{\mu}_j^{*(b1)} - \hat{\mu}_2^{*(b1)}, d = max_{j=1,3,\cdots,15}l_j - l_2.$$

The similarity between testing $H_i, i = 2, \cdots, 15$ and H_1 is the covariance matrix Σ in equation (17).

When we want to calculate the reliability for one tree, for example tree 2, we can use the command `sdbpk`, with the output shown below. This command corresponds to testing hypothesis $H_2 : \mu_2 = max_{k=1,2,\cdots,15}\mu_k$ for tree 2 in the topology file mam15.tpl.

```
> result1 <- sdbpk(dat,2)
> result1

Call:
sdbpk(dat = dat, k = 2)

t2
0.2237
```

Then, calculating the bootstrap probability can use the command `bp`, again shown with the output.

```
> result2 <- bp(dat)
> result2

Call:
bp(dat = dat)

Bootstrap probabilities:
t1     t3     t2     t5     t6     t7
0.5794 0.3213 0.0342 0.0124 0.0279 0.0057 ...
```

4. Result

4.1 Analysis of mammalian mitochondrial protein sequences

Table 1 presents the results of our sDBP value calculations for the 15 phylogenetic trees analyzed in this study, along with values reported by [11] for traditional BP analyses and the AU-test. We also developed an algorithm for the regular double bootstrap approach for phylogenetic trees [12], although in this paper we have omitted the description of how the DBP were calculated. In Table 1, the original tree number in the file mam15.tpl is renamed in decreasing order of log-likelihood. The confidence sets of trees obtained by the sDBP algorithm and the DBP algorithm at $\alpha = 0.05$ were $\{1, 2, 3, 4, 5, 6, 7\}$ and $\{1, 2, 3, 5, 7\}$, respectively (Table 1). The sDBP tree set was thus slightly larger than the set selected by DBP. Tree 7 is the most strongly supported as T_{ML} by recent analyses incorporating additional sequence data [22], [23], [24], and our results for this tree indicate that sDBP=0.084>0.05 and DBP=0.056 $>$ 0.05. Our conclusions are thus not in contradiction with the latest data. For a confidence set of models, our sDBP algorithm gives a confidence set of candidate trees, and includes the "best" topology $\max_{i=1,\cdots,15}\mu_i$, with an error rate below the 0.05 level. Thus, our sDBP tree set does not immediately give the work for straightly gives the "best" topology.

4.2 Comparison of computational speed

For the sDBP algorithm, the DBP algorithm, the AU-test and the BP-test, we measured the time taken to calculate a p-value for tree 7 (see Table 1), based on the site-wise log-likelihood data. We used the RELL approximation method [18] with the BP-test, and conducted two separate sets of analyses. In the first set, we applied the sDBP algorithm with $B1 = 10^3$ pseudoreplicates, the DBP algorithm with $B1 = 10^3$ and $B2 = 10^3$ pseudoreplicates, and the BP-test with 10^3 pseudoreplicates. In the second set, we applied the sDBP algorithm with $B1 = 5 \times 10^3$ pseudoreplicates, the DBP algorithm with $B1 = 5 \times 10^3$ and $B2 = 5 \times 10^3$ pseudoreplicates, and the BP-test with 5×10^3 pseudoreplicates. The results of the two sets are shown in Table 2. This time, we used the command `sdbpk`, `bpk` and `dbpk` to measure the time. For measuring time of AU-test, we used the command `relltest` from R package scaleboot. For both sets, the BP-test was the fastest, followed by the sDBP algorithm, the AU-test then the DBP algorithm. For the first set of calculations (lower numbers of pseudoreplicates) the sDBP algorithm was 1021-fold faster than the DBP algorithm, and this advantage improved substantially for the second set (higher pseudoreplication), with the sDBP algorithm being 5076-fold faster than the DBP algorithm.

Table 1 Comparison of four different p-values from analyses of fifteen mammalian trees, based on protein sequence data from [14]. The p-values that are NOT significant at $\alpha = 0.05$ are emphasized in bold type.

Tree[a]	$\triangle l_i$	BP$_i^b$	DBP$_i^c$	sDBP$_i^d$	AU$_i^e$	Tree form[f]
1	-2.7	**0.579**	**0.607**	**0.576**	**0.789**	(((1(23))4)56)
2	2.7	**0.312**	**0.458**	**0.401**	**0.516**	((1((23)4))56)
3	7.4	0.036	**0.167**	**0.235**	**0.114**	(((14)(23))56)
4	17.6	0.013	0.041	**0.116**	**0.075**	((1(23))(45)6)
5	18.9	0.035	**0.082**	**0.110**	**0.128**	(1((23)(45))6)
6	20.1	0.005	0.031	**0.069**	0.029	(1(((23)4)5)6)
7	20.6	0.017	**0.056**	**0.084**	**0.101**	((1(45))(23)6)
8	22.2	0.001	0.007	0.042	0.009	((15)((23)4)6)
9	25.4	0.000	0.002	0.022	0.000	(((1(23))5)46)
10	26.3	0.003	0.011	0.023	0.028	(((15)4)(23)6)
11	28.9	0.000	0.003	0.013	0.003	(((14)5)(23)6)
12	31.6	0.000	0.001	0.004	0.001	(((15)(23))46)
13	31.7	0.000	0.002	0.005	0.001	(1(((23)5)4)6)
14	34.7	0.000	0.003	0.001	0.005	((14)((23)5)6)
15	36.2	0.000	0.001	0.000	0.002	((1((23)5))46)

[a] Trees are numbered by increasing order of $\triangle l_i = max_{j \neq i} l_j - l_i$, the difference between the log-likelihood value for a given tree and the largest value among all other trees.
[b] Bootstrap probability, calculated from 10000 pseudoreplicates (from Shimodaira (2002)).
[c] Double bootstrap probability, calculated from 25 million pseudoreplicates (B1 = 5 $\times$1000, $B2 = 5 \times 1000$).
[d] Speedy double bootstrap probability, calculated from 10000 pseudoreplicates (B1 = 10000).
[e] Multiscale bootstrap probability, calculated from 100000 pseudoreplicates (AU-test; from Shimodaira (2002)).
[f] Taxon labels: 1 = human, 2 = seal, 3 = cow, 4 = rabbit, 5 = mouse, 6 = opossum.

Table 2 Comparison of the BP, DBP, sDBP and AU methods, regarding their speed for computing a p-value for tree-7.

	BP	DBP	sDBP	AU	Speed increase (sDBP/DBP)
Time (secs)[a]	0.69	715	0.73	3.72	1021-fold
Time (secs)[b]	3.52	17921	3.53	14.39	5076-fold

[a] Case of $B1 = 10^3$, $B2 = 10^3$ pseudoreplicates
[b] Case of $B1 = 5 \times 10^3$, $B2 = 5 \times 10^3$ pseudoreplicates

5. Conclusion

As shown in the result section, the sDBP algorithm has comparable accuracy to the DBP algorithm and is much more computationally efficient for phylogenetic tree selection problem. For allowing researchers to apply the sDBP algorithm easily, we have developed an easy to use R package. We think this implementation of sDBP algorithm will be of further utilities to assessing the reliability of phylogenetic trees.

Availablity

The program is freely distributed under GNU General Public License (GPL) and can directly installed from CRAN,
http://cran.r.-project.org/
the official R package archive. The instruction and program source code are avaliable at
http://www.bi.cs.titech.ac.jp/sdbp/

References

[1] C. Webb, D. Ackerly, M. McPeek, and M. Donoghue, "Phylogenies and community ecology," *Annual Review of Ecology and Systematics*, pp. 475–505, 2002.

[2] E. Wiley, *Phylogenetics: The Theory and Practice of Phylogenetic Systematics*. Wiley-Interscience, New York, 1981.

[3] F. Pazos and A. Valencia, "Similarity of phylogenetic trees as indicator of protein–protein interaction," *Protein Engineering*, vol. 14, no. 9, pp. 609–614, 2001.

[4] J. Felsenstein, *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts, 2004.

[5] ——, "Evolutionary trees from dna sequences: A maximum likelihood approach," *Journal of Molecular Evolution*, vol. 17, no. 6, pp. 368–376, 1981.

[6] ——, "Confidence limits on phylogenies: An approach using the bootstrap," *Evolution*, pp. 783–791, 1985.

[7] D. Hillis and J. Bull, "An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis," *Systematic Biology*, vol. 42, no. 2, pp. 182–192, 1993.

[8] M. Sanderson and M. Wojciechowski, "Improved bootstrap confidence limits in large-scale phylogenies, with an example from neo-astragalus (leguminosae)," *Systematic Biology*, vol. 49, no. 4, pp. 671–685, 2000.

[9] P. Hall, *The bootstrap and Edgeworth expansion*. Springer Verlag, New York, 1992.

[10] B. Efron and R. Tibshirani, "The problem of regions," *Stanford Technical Report*, vol. 192, 1996. [Online]. Available: ftp://utstat.toronto.edu/pub/tibs/regions.ps.

[11] H. Shimodaira, "An approximately unbiased test of phylogenetic tree selection," *Systematic Biology*, vol. 51, no. 3, pp. 492–508, 2002.

[12] A. Ren, T. Ishida, and Y. Akiyama, "Assessing statistical reliability of phylogenetic trees via a speedy double bootstrap method," *Molecular Phylogenetics and Evolution*, vol. 67, pp. 429–435, 2013.

[13] Y. S. Aulchenko, S. Ripke, A. Isaacs, and C. M. van Duijn, "Genabel: an r library for genome-wide association analysis," *Bioinformatics*, vol. 23, no. 10, pp. 1294–1296, 2007.

[14] H. Shimodaira and M. Hasegawa, "Multiple comparisons of log-likelihoods with applications to phylogenetic inference," *Molecular Biology and Evolution*, vol. 16, pp. 1114–1116, 1999.

[15] J. Hsu, "Simultaneous confidence intervals for all distances from the " best "," *The Annals of Statistics*, pp. 1026–1034, 1981.

[16] H. Shimodaira, "Testing regions with nonsmooth boundaries via multiscale bootstrap," *Journal of Statistical Planning and Inference*, vol. 138, no. 5, pp. 1227–1241, 2008.

[17] B. Efron, "Bootstrap confidence intervals for a class of parametric problems," *Biometrika*, vol. 72, no. 1, pp. 45–58, 1985.

[18] H. Kishino, T. Miyata, and M. Hasegawa, "Maximum likelihood inference of protein phylogeny and the origin of chloroplasts," *Journal of Molecular Evolution*, vol. 31, no. 2, pp. 151–160, 1990.

[19] M. Ayer, H. Brunk, G. Ewing, W. Reid, and E. Silverman, "An empirical distribution function for sampling with incomplete information," *The Annals of Mathematical Statistics*, pp. 641–647, 1955.

[20] Z. Yang, "Paml: a program package for phylogenetic analysis by maximum likelihood," *Computer Applications in the Biosciences*, vol. 13, no. 5, pp. 555–556, 1997.

[21] H. Shimodaira and M. Hasegawa, "Consel: for assessing the confidence of phylogenetic tree selection," *Bioinformatics*, vol. 17, no. 12, pp. 1246–1247, 2001.

[22] Y. Cao, M. Fujiwara, M. Nikaido, N. Okada, and M. Hasegawa, "Interordinal relationships and timescale of eutherian evolution as inferred from mitochondrial genome data," *Gene*, vol. 259, no. 1, pp. 149–158, 2000.

[23] O. Madsen, M. Scally, C. Douady, D. Kao, R. DeBry, R. Adkins, H. Amrine, M. Stanhope, W. de Jong, and M. Springer, "Parallel adaptive radiations in two major clades of placental mammals," *Nature*, vol. 409, no. 6820, pp. 610–614, 2001.

[24] W. Murphy, E. Eizirik, W. Johnson, Y. Zhang, O. Ryder, and S. O'Brien, "Molecular phylogenetics and the origins of placental mammals," *Nature*, vol. 409, no. 6820, pp. 614–618, 2001.