Vol.2013-MPS-94 No.4
2013/7/22

SDBP: スピーディー・ダブルブートストラップ法に基づいた系統樹の信頼性評価のためのRパッケージ

任愛珍[1,a)]　石田 貴士[2,b)]　秋山 泰[1,2,c)]

概要：分子系統樹の信頼性評価は，分子系統学だけではなく，生物の系統樹に依存する他の分野にも重要である．ブートストラップ法はモデルの信頼性評価では一般的に使われている非常に有名な手法であるが，その精度が低いので，様々な試みがなされている．その中の１つに，我々が提案したスピーディー・ダブルブートストラップ法と呼ぶ手法がある．我々の評価によれば，系統樹の信頼性評価では，スピーディー・ダブルブートストラップ法は精度の面でも，速度の面でも優れた性質を持つ方法である．そこで，我々はスピーディー・ダブルブートストラップ法を R のパッケージとして実装した．本研究報告では，スピーディー・ダブルブートストラップ法の理論背景と系統樹の信頼性評価でのアルゴリズムを解説した後，実装された R パッケージ SDBP の使い方について説明する．

キーワード：SDBP，スピーディー・ダブルブートストラップ法，信頼性，高速計算，R パッケージ

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

Ren Aizhen[1,a)]　Takashi Ishida[2,b)]　Yutaka Akiyama[1,2,c)]

Abstract: Evaluating the reliability of estimated phylogenetic trees is of critical importance in the field of molecular phylogenetics. 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. And in this paper we briefly introduce the mathematical theory of the sDBP-method and its algorithm for assessing the reliability of phylogenetic trees. Then, we describe the basic usage of our package.

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

1. 導入

分子系統学で用いられる情報分析手法は，生物の進化の歴史（系統発生の関係）を再建するための基本的で重要な道具である．分子系統法は主に生物分類学の分野で使われ

[1] 東京工業大学 大学院情報理工学研究科　数理・計算科学専攻，Department of Mathematical and Computing Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology
[2] 東京工業大学 大学院情報理工学研究科 計算工学専攻，Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology
[a)] ren@bi.cs.titech.ac.jp
[b)] t.ishida@bi.cs.titech.ac.jp
[c)] akiyama@cs.titech.ac.jp

ているが，さらには，群集生態学や生物地理学などのさまざまな分野で広く応用されている．系統樹の推定には多くの方法が開発され，一般的に使われている．それらのうちで，最尤法による系統樹の推定は最も優れた性質をもつ手法であることが示されている．最尤法による系統樹の推定は 1980 年代に Felsenstein[1] によって初めて提案されて，次の段落でその計算法について簡単に説明する．

本研究では，例として DNA 配列データを使用して説明する．データセットとして m 本の相同配列からなる長さ n のアラインメントであとき，これを $m \times n$ の行列 $\mathbf{X} = \{x_{jh}\} = \{\mathbf{x_1}, \cdots, \mathbf{x_n}\}$ として表現する．ここで，x_{jh} は j 番目の配列の h 番明の塩基を意味する．記号 $\mathbf{x}_h$ でこのデータ行列の h 番目の列を表す．系統樹の対数尤度は $l(\boldsymbol{\theta}; \mathbf{X}) = \sum_{h=1}^{n} \log f(\mathbf{x}_h; \boldsymbol{\theta})$ である．ただし，$f(\mathbf{x}_h; \boldsymbol{\theta}) = f(x_{1h}, x_{2h}, \cdots, x_{mh}; \boldsymbol{\theta})$ は一つのサイトでの確率値である．ベクトル $\boldsymbol{\theta}$ は未知のパラメータで，枝の長さを表す．与えられたトポロジーに対して，パラメータベクトル $\boldsymbol{\theta}$ は対数尤度を最大化することによって推定され，$\hat{\boldsymbol{\theta}}$ と書く．そして，与えられた任意のトポロジー i の最大対数尤度は $l_i(\hat{\boldsymbol{\theta}}_i; \mathbf{X}) = \sum_{h=1}^{n} \log f_i(\mathbf{x}_h; \hat{\boldsymbol{\theta}}_i)$ になる．さらに $l_i(\hat{\boldsymbol{\theta}}_i; \mathbf{X}), i = 1, \cdots, K$ を最大化したトポロジーはデータ $\mathbf{X}$ に対する最大尤度系統樹 (T_{ML}) であり，このデータセットの時，トポロジー T_{ML} はデータへの当てはまりが一番良いトポロジーであると解釈することができる．いくつかの系統樹のトポロジーの信頼性評価をする時，データ $\mathbf{X} = \{\mathbf{x_1}, \cdots, \mathbf{x_n}\}$ の要素である $\{\mathbf{x_1}, \cdots, \mathbf{x_n}\}$ について以下のように仮定する．

$$(\mathbf{x_1}, \cdots, \mathbf{x_n}) \overset{i.i.d.}{\sim} q(\mathbf{x}) \qquad (1)$$

ここで $q(\mathbf{x})$ は未知の真の進化の確率過程の分布であり，実際には未知であるのだが，トポロジー (以下は系統樹とよぶ) の選択あるいは系統樹の信頼性評価では，その存在性を仮定して理論を展開する．系統樹の信頼性評価ではこの真の分布 $q(\mathbf{x})$ を利用して一番良い系統樹を定義する．理論的に保証される一番良い系統樹は　$\bar{k} = \underset{i=1,\cdots,K}{\arg\max} E_q[l_i(\hat{\boldsymbol{\theta}}_i; \mathbf{X})]$ で定義される．ただし，$E_q[l_i(\hat{\boldsymbol{\theta}}_i; \mathbf{X})] (= \mu_i)$ の期待値は式 (1) に関して取ったものである．分布 $q(\mathbf{x})$ は未知なので，この一番良いモデルは理論的に求めることができない．以下で説明するモデルの信頼性はこの最大尤度モデルに基づき，候補の K 個のモデルがそれぞれ一番良いモデルであると仮定した時のその仮説の確率値である．例えば，もし系統樹 T_1 が一番良い系統樹であるという仮説は $T_1 = T_{\bar{k}}$ と仮定されていると考えられる．つまり，帰無仮説と対立仮説は以下のように表すことができる．

$$H_1 : \mu_1 = \max_{i=1,\cdots,K} \mu_i \quad vs. \quad H_1^A : \ others, \qquad (2)$$

仮説 H_1 に対して多くの検定が提案されている．大別すると，多重比較とブートストラップングを使用した 2 種類の手法がある。ブートストラップングを使用した手法の中で，2002 年に，Shimodaira[1] が提案したマルチスケール・ブートストラップ法は精度が 3 次の確率を計算する手法である．我々の提案したスピーディー・ダブルブートストラップ法 [2] はマルチスケール・ブートストラップ法と同じく 3 次の精度の確率値を計算する手法である．しかも，スピーディー・ダブルブートストラップ法はマルチスケール・ブートストラップ法よりも系統樹の実問題においては，数倍ほど高速であった．そこで本研究では sDBP 法を SDBP と言う R パッケージに実装して，生物学者が sDBP 法を簡便に利用できるようにした．

2. 理論とアルゴリズム

2.1 スピーディー・ダブルブートストラップ法の理論背景

この節では，スピーディー・ダブルブートストラップ法の理論背景について述べる．Shimodaira [1] が Efron の多変量正規分布に従う確率変量の実関数における検定の理論 [3] を系統樹の信頼性評価の帰無仮説例えば H_1 のような帰無仮説に応用した．DNA 配列データ $\mathbf{X}$ に対して，変数変換 g が存在して，m 次元の確率ベクトル $\boldsymbol{Y} = g(\mathbf{X})$ に変換し，$\boldsymbol{Y}$ は平均ベクトルが $\boldsymbol{\eta}$，分散共分散行列が単位行列である m 次元の多変量正規分布に従うと仮定する．つまり，$\boldsymbol{Y} \sim N_m(\boldsymbol{\eta}, I_m)$ と仮定する．また任意の領域 $\mathcal{H}$ が存在して，滑らかな境界面 $B(h)$ をもつと仮定する．そして，帰無仮説 $\boldsymbol{\eta} \in \mathcal{H}$ の確率値 $p(\boldsymbol{y})$ を求める方法をあるメカニズムによって逆に H_1 に応用する発想である．論文 [3] によれば，真のパラメーター $\boldsymbol{\eta}$ が境界面 $B(h)$ にある時，3 次の精度の確率値は $p(y) = 1 - \Phi(d - c)$ と書ける．ただし，d は $\boldsymbol{y}$ から $\hat{\boldsymbol{\eta}}(\boldsymbol{y})$ までの符号付き距離で，$\boldsymbol{y}$ が領域 $\mathcal{H}$ の外にある時は正で，中にある時は負である．点 $\hat{\boldsymbol{\eta}}(\boldsymbol{y})$ は点 $\boldsymbol{y}$ の境界面 $B(h)$ への射影で，c は幾何量で境界面 $B(h)$ の点 $\hat{\boldsymbol{\eta}}(\boldsymbol{y})$ での曲率と関係がある量である．1997 年に Efron と Tibushirani の技術報告 [4] の中で，スピーディ・ブートストラップ法の原型となるアイデアが示されていた．彼らは以下の多変量正規分布からリサンプリングしている．

$$\boldsymbol{Y}^* \overset{i.i.d.}{\sim} N_t(\hat{\boldsymbol{\eta}}(\boldsymbol{y}), I_t). \qquad (3)$$

d^* はブートストラップ標本 $\boldsymbol{y}^*$ から境界面 $B(h)$ への射影 $\hat{\boldsymbol{\eta}}(\boldsymbol{y}^*)$ までの符号付き距離である．技術報告 [4] によると，3 次の精度の確率値を，以下のように得ることができる．

$$1 - \Phi(d - c) = P(d^* > d; \hat{\boldsymbol{\eta}}(\boldsymbol{y})) + O(n^{-3/2}). \qquad (4)$$

2.2 系統樹の信頼性評価のスピーディー・ダブルブートストラップ法のアルゴリズム

この節では系統樹の信頼性評価に対する議論に戻る．論

文 [2] では Efron と Tibshirani の技術報告 [4] の中で書かれていたのと同様のアイデアに基づき，仮説 H_1 の確率値の計算を実現した．実用問題に応用する時に解決しなければならない射影と符号付き距離の問題を解決する手法を提案しそれをスピーディー・ダブルブートストラップ法と名付けて提案をした [2]．まずは，式 (3) の射影 $\hat{\boldsymbol{\eta}}(\boldsymbol{y})$ と対応するベクトルを以下のような方法で解決した．論文 [5] によると，最大対数尤度ベクトル $\boldsymbol{l} = (l_1(\hat{\boldsymbol{\theta}}_1), \cdots, l_K(\hat{\boldsymbol{\theta}}_K))$ は漸近的に正規分布に従い，その平均ベクトルを $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_K)$ とするとき，ベクトル $\boldsymbol{l}$ は無制約の時の平均 $\boldsymbol{\mu}$ のデータ $\mathbf{X}$ からの推定値である．しかし，今は仮説 H_1 では，$\mu_1 = \max_{i=1,\cdots,K} \mu_i$ と仮定しているので，この制約の下での平均 $\boldsymbol{\mu}$ の制約付き最大尤度推定量は PAVA (pool adjacent violators algorithm) [6] を用いて推定でき，

$$\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \cdots, \hat{\mu}_K). \tag{5}$$

とおくと，以下のように求めることができる．

$$\begin{aligned} \hat{\mu}_1 &= \max_{W \subseteq \{1,2,\cdots,K\}} \frac{\sum_{j \in W} \hat{l}_j}{|W|} \\ \hat{\mu}_j &= \min\{\hat{\mu}_1, \hat{l}_j\}, j \in \{2, \cdots, K\} \end{aligned} \tag{6}$$

ここで，ベクトル $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \cdots, \hat{\mu}_K)$ は $\hat{\boldsymbol{\eta}}(\boldsymbol{y})$ と対応している．また，ベクトル $\boldsymbol{l}$ の従う多変量正規分布の分散共分散行列の推定を $\Sigma = (\sigma_{ij})$ と書き，その要素である σ_{ij} は以下の式によって推定できる [5].

$$\begin{aligned} \frac{n}{n-1} \sum_{h=1}^{n} \left[log f_i(\mathbf{x_h}; \hat{\theta}_\mathbf{i}) - \frac{1}{n} \sum_{h=1}^{n} log f_i(\mathbf{x_h}; \hat{\theta}_\mathbf{i}) \right] \\ \times \left[log f_j(\mathbf{x_h}; \hat{\theta}_\mathbf{j}) - \frac{1}{n} \sum_{h=1}^{n} log f_j(\mathbf{x_h}; \hat{\theta}_\mathbf{j}) \right]. \end{aligned} \tag{7}$$

次に，式 (4) の d と対応する量を計算しなければならない．これに対して，我々は以下の量を用いて計算する．

$$d = \max_{j=2,\cdots,K} l_j - l_1. \tag{8}$$

また，式 (4) の d^* と対応する量の計算法は d と対応する量と同じような計算法である．仮説 H_1 の確率値を以下のアルゴリズムで計算する．まず，仮説 H_1 に対して，ベクトル $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \hat{\mu}_2, \cdots, \hat{\mu}_K)$ の $B1$ 個のブートストラップ複製 $(\hat{\mu}_1^{*(b1)}, \cdots, \hat{\mu}_K^{*(b1)}), b1 = 1, \cdots, B1$ を以下の正規分布から発生する．

$$(\hat{\mu}_1^{*(b1)}, \cdots, \hat{\mu}_K^{*(b1)})^T \overset{i.i.d.}{\sim} N_K((\hat{\mu}_1, \hat{\mu}_2, \cdots, \hat{\mu}_K)^T, \Sigma), \tag{9}$$

ただし，記号 T は転置を表し，行列 Σ は式 (7) で定義されたものである．d^* と対応する量は以下のように計算する．

$$d^{*(b1)} = \max_{j=2,\cdots,K} \hat{\mu}_j^{*(b1)} - \hat{\mu}_1^{*(b1)}. \tag{10}$$

次に，仮説 H_1 の p-値を以下のように計算して，記号 $sDBP$ で表す．

$$sDBP = \frac{\#(d^{*(b1)} > d)}{B1}. \tag{11}$$

仮説 H_1 と全く同じ様に，我々はこのアルゴリズムを他の全ての仮説 $H_k, k = 2, \cdots, K$ に適用することができる．

3. 実装と利用例

3.1 R のパッケージへの実装

我々は sDBP 法の系統樹の信頼性評価でのアルゴリズムを R のパッケージに実装した．パッケージ名は SDBP である．パッケージ SDBP は六つのユーザーレベルオブジェクトから成っていて，それぞれ sdbp，sdbpk，bpk，bp，dbpk，mam20 である．この節でこれらのオブジェクトの使い方について説明する．また，パッケージ SDBP は 3 種類の確率値を計算できる：sDBP (speedy double bootstrap probability), DBP (double bootstrap probability), BP (bootstrap probability).

3.2 使い方 — 哺乳類のミトコンドリアの蛋白質アミノ酸配列データを用いて

この節で，我々はパッケージ SDBP の使い方について哺乳類のミトコンドリアの蛋白質アミノ酸配列データを用いて説明する [1]．このデータセットは“mam15-files”というフォルダの中に含まれている．このデータセットは 6 種類の哺乳類 (ヒト，アザラシ，ウシ，ウサギ，マウス，オポッサム) の長さ $n = 3414$ のアミノ酸配列である．分岐群 { アザラシ，ウシ } は事前の分析で強く支持されているため，この分岐群を含む 15 個の無根系統樹の確率値を計算した (表 1 を参照).

アミノ酸配列データをあらかじめソフトウェア PAML[7] と CONSEL を使って mam15.mt というファイルを準備する．この中に各系統樹のサイトあたりの最大対数尤度行列が保存されている．15 個の系統樹のトポロジーのファイルは mam15.tpl で，先述の mam15-files フォルダ内にある．

まず R を起動し，**R console** コマンドラインに次のコマンドを入力してパッケージ SDBP をロードする：

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

つぎに，データ mam15.mt を読み込む.

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

各系統樹の sDBP を計算するには以下のコマンドを入力する．この 1 行のコマンドだけで，sDBP の計算が完了する．

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

以下の出力では，最大対数尤度の大きい順でモデルが並べられている．

```
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 ...
```

コマンド sdbp.default は仮説 H_1 および仮説 $H_k, k = 2, \cdots, 15$ の確率値を計算するためのコマンドである．各確率の標準エラーを計算するには

```
> summary(result)
```

を用い，出力は

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

   stdErr p.value
t1 0.0049  0.5717
t3 0.0049  0.3928
...
attr(,"class")
[1] "summary.sdbp"
```

のように得られる．もし，特定の 1 つの系統樹，例えば系統樹 2 の確率値を計算するには，コマンド sdbpk(dat,2) を使う．このコマンドは仮説 H_2 の確率値 sDBP を計算する．

4. 結果と考察

4.1 哺乳類のミトコンドリアの蛋白質アミノ酸配列を用いて分析した結果

表 1 に，15 個の系統樹に対する DBP と sDBP の計算結果を示す．また，比較のために，論文 [1] で計算された BP と AU(approximately unbiased の略) も表 1 内に示す．表 1 の系統樹の番号はファイル mam15.tpl での系統樹の番号ではなく，最大対数尤度の大きい順で番号を改めてつけている．sDBP アルゴリズムと DBP アルゴリズムで計算した有意水準 $\alpha = 0.05$ での信頼集合はそれぞれ $\{1, 2, 3, 4, 5, 6, 7\}$ と $\{1, 2, 3, 5, 7\}$ であった．今回のデータでは sDBP の信頼集合は DBP の信頼集合より少し大きい．また，系統樹 7 は論文 [8], [9], [10] の最新のデータでは最大尤度系統樹と見られているが，この系統樹に対する我々の計算結果は sDBP=0.084>0.05 と DBP=0.056>0.05 であり，この生物学的な結果と矛盾していない．

5. 結論

統計学者および生物学者が sDBP アルゴリズムを簡単に使えるようにするため，我々は使いやすい R のパッケージを開発した．パッケージ SDBP は系統樹の信頼性評価の汎用のユーティリティになり得ると我々は考えている．

利用

このプログラムは，GNU 一般公衆利用許諾のもとで配布されて，直接 CRAN のホームページ http://cran.r.-project.org/ からダウンロードすることができる．

表 1 15 個の系統樹の 4 種の手法での p-値の比較

Table 1 Comparison of four different p-values from analyses of fifteen mammalian trees

系統樹 [a]	$\triangle l_i$	BP_i^b	DBP_i^c	$sDBP_i^d$	AU_i^e	系統樹の形 [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] 系統樹は $\triangle l_i = max_{j \neq i} l_j - l_i$ の小さい順で並べられた．
[b] ブートストラップ確率,10000 個の複製から計算された (Shimodaira (2002) から引用)．
[c] ダブル・ブートストラップ確率,2500 万個の複製から計算された ($B1 = 5 \times 1000$, $B2 = 5 \times 1000$)．
[d] スピーディー・ダブルブートストラップ確率,10000 個の複製から計算された ($B1 = 10000$)．
[e] マルチスケールブートストラップ確率,100000 個の複製から計算された (AU 検定; Shimodaira (2002) から引用)．
[f] 種のラベル: 1 = ヒト, 2 = アザラシ, 3 = ウシ, 4 = ウサギ, 5 = マウス, 6 = オポッサム．

参考文献

[1] Shimodaira, H.: An approximately unbiased test of phylogenetic tree selection, *Systematic Biology*, Vol. 51, No. 3, pp. 492–508 (2002).
[2] Ren, A., Ishida, T. and Akiyama, Y.: Assessing statistical reliability of phylogenetic trees via a speedy double bootstrap method, *Molecular Phylogenetics and Evolution*, Vol. 67, pp. 429–435 (2013).
[3] Efron, B.: Bootstrap confidence intervals for a class of parametric problems, *Biometrika*, Vol. 72, No. 1, pp. 45–58 (1985).
[4] Efron, B. and Tibshirani, R.: The problem of regions, *Stanford Technical Report*, Vol. 192 (online), available from ⟨ftp://utstat.toronto.edu/pub/tibs/regions.ps.⟩ (1997).
[5] Kishino, H., Miyata, T. and Hasegawa, M.: Maximum likelihood inference of protein phylogeny and the origin of chloroplasts, *Journal of Molecular Evolution*, Vol. 31, No. 2, pp. 151–160 (1990).
[6] Zhao, H.: Comparing several treatments with a control, *Journal of Statistical Planning and Inference*, Vol. 137, No. 9, pp. 2996–3006 (2007).
[7] Yang, Z.: PAML: a program package for phylogenetic analysis by maximum likelihood, *Computer Applications in the Biosciences*, Vol. 13, No. 5, pp. 555–556 (1997).
[8] Cao, Y., Fujiwara, M., Nikaido, M., Okada, N. and Hasegawa, M.: Interordinal relationships and timescale of eutherian evolution as inferred from mitochondrial genome data, *Gene*, Vol. 259, No. 1, pp. 149–158 (2000).
[9] Madsen, O., Scally, M., Douady, C., Kao, D., DeBry, R., Adkins, R., Amrine, H., Stanhope, M., de Jong, W. and Springer, M.: Parallel adaptive radiations in two major clades of placental mammals, *Nature*, Vol. 409, No. 6820, pp. 610–614 (2001).
[10] Murphy, W., Eizirik, E., Johnson, W., Zhang, Y., Ryder, O. and O'Brien, S.: Molecular phylogenetics and the origins of placental mammals, *Nature*, Vol. 409, No. 6820, pp. 614–618 (2001).