계통수 재구성에서 붓스트랩

정유진[1]*

요약**:** 붓스트랩은 진화유전학과 통계학에서 강력한 통계적 도구로 사용되고 있다. 특히 계통수의 재구성에서 신뢰도를 측정하기 위해 붓스트랩은 가장 많이 사용되는 도구 중 하나이다. 본 논문에서는 통계학에서의 붓스트랩의 원리 및 이를 활용한 계통수의 신뢰도 측정 방법에 대한 리뷰를 제공한다. 붓스트랩 비율과 붓스트랩 consensus 나무와 같은 계통수의 신뢰도 측정 방법을 소개하고, 12종의 영장류 DNA 데이터를 붓스트랩 분석한 예시를 제시한다. 또한 붓스트랩의 해석 , 이점 그리고 한계에 대해 논의한다.

키워드**:** bootstrap, phylogenetic tree, confidence

[1] *경기도 수원시 영통구 광교산로 154-42 경기대학교 응용통계학과*

*__Corresponding author__: yujinchung@kyonggi.ac.kr

서론

DNA 염기서열 데이터의 급속한 축적은 생물 간의 계통 관계를 재구성하는데 많은 활동을 자극하여 계통수(phylogenetic tree)의 재구성하는 방법의 개발에도 많은 관심을 불러 일으켰다. DNA 또는 RNA 서열 데이터로부터 추정된 계통수에 대한 신뢰도를 측정하기 위해 가장 일반적으로 사용되는 방법은 붓스트래핑으로 Efron(1979)이 처음 발명하고 Felsenstein(1985)이 계통 발생 문제에 처음 적용한 통계 기법이다.

”Bootstrapping”의 용어는 ”자신의 부츠끈을 당겨 자신을 들어올리다”라는 뜻의 오래된 표현에서 유래된 것으로 불가능한 일을 자신의 힘만으로 극복하려는 노력을 의미한다(bootstrapping 2023). 통계학에서 붓스트랩 방법은 Bradley Efron이 1979년에 처음 소개하였다(Efron 1979). 이 방법은 원래 표본에서 반복적으로 재표본(resampling)을 수행하여 분산과 편향성 추정에 집중되어 있다. Efron이 처음 붓스트랩을 소개한 이후 엄청난 양의 통계 연구가 붓스트랩 접근법의 유효성을 입증하는데 투자되었다. 대부분의 확률 모형과 대부분의 추정량에 대한 붓스트랩 표준오차는 실제 표준오차에 대한 좋은 추정치이며, 다른 정확도 측정값에 대해서도 마찬가지라는 것은 알려져있다(Efron 2003). 그로부터 44년이 지난 지금 붓스트랩은 가장 강력한 통계적 추론 방법 중에 하나로 널리 사용되고 있다.

붓스트래핑은 주어진 데이터 세트를 원래의 모집단을 대표하는 독립 표본으로 가정하고, 그 자료로부터 중복을 허용한 무작위 재추출로 복수의 자료를 작성하며, 자신의 데이터에서 포인트를 리샘플링하고 교체하여 원본 데이터와 동일한 크기의 일련의 부트스트랩 샘플을 생성하는 것을 포함한다. 각각의 표본을 분석하고, 그 결과 도출된 추정치 간의 차이를 통해 원본 데이터에서 추정한 오차의 크기를 나타낸다. 1985년 Felsenstein은 붓스트랩을 계통수의 신뢰도를 측정하는데 사용하는 것을 제안하였다 (Felsenstein 1985). 이 방법은 최대가능도 추정법(Maximum likelihood estimation), neighbor-joining, UPGMA, maximum parsimony 등과 같은 방법으로 계통수를 구할 때 신뢰도를 측정하기 위한 표준 수단으로 phylogenetics 분야에서 널리 사용되고 있다(Felsenstein 2004; Yang 2006).

본 논문에서는 통계학에서 처음 소개된 붓스트랩의 원리를 설명하고 이어서 Felsenstein의 붓스트랩

방법으로 계통수의 신뢰도를 측정하는 방법을 소개하겠다. 각 방법의 장단점에 대해 설명하고 12종의 영장류의 DNA 서열 데이터 분석 예제를 제시한다. 통계적 용어는 한국통계학회 용어집에 따라 사용하였다.

붓스트래핑

통계학에서는 일반적으로 관심 모수(parameter) θ를 모집단(population)으로부터 추출한 랜덤 표본 $X_1, \ldots, X_n$으로부터 추정하고, 그 추정량 $\hat{\theta}$의 정확도(accuracy) 혹은 반대로 불확실성(uncertainty)에 관심이 있다. 이러한 불확실성은 일반적으로 표준 오차(standard error)[1]나 신뢰 구간(confidence interval)으로 요약할 수 있다.

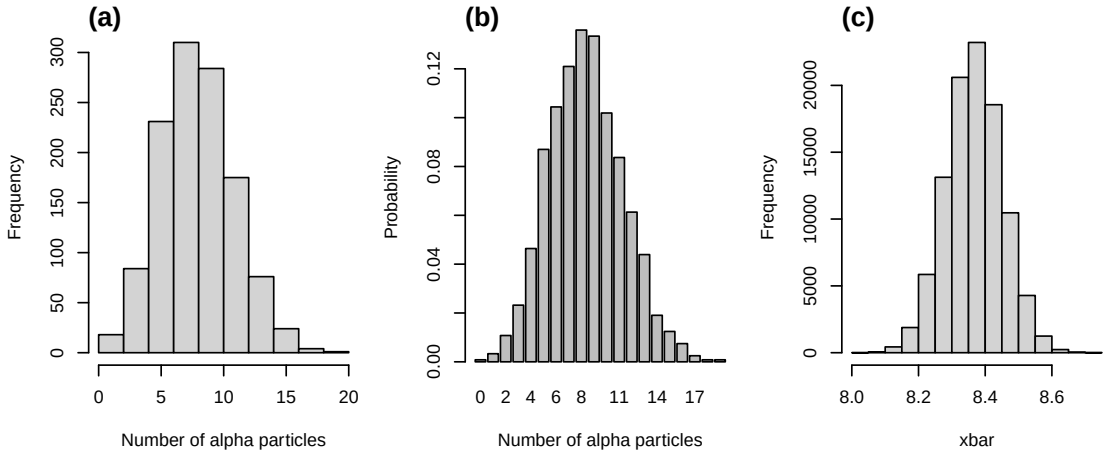

그림 1. (a) 알파 입자수 데이터의 히스토그램. (b) 각 자료가 $1/n$의 확률값을 가지는 경험적 분포. (c) 붓스트래핑으로 100,000번 반복 추출하여 구한 $\bar{X}$의 분포

예를 들어, $n = 1207$개의 알파 입자수[2](그림 1a)를 분석하기 위해, 모집단의 분포로 포아송(Poisson) 분포를 가정하였고, 포아송 분포의 모수 λ를 추정하고자 한다(Rice 2007). 모수 λ를 최대가능도추정량(maximum likelihood estimator; 이하 MLE)인 $\hat{\lambda} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = 8.37$으로 추정하였을 때, $\hat{\lambda}$의 표준 오차는 얼마일가? 포아송 분포의 모평균(population mean)과 모분산(population variance)은 모두 λ로 동일하므로, 중심 극한 정리(central limit theorem)에 의하여 $\hat{\lambda}$는 근사적으로 다음과 같이 정규분포를 따른다.

$$\sqrt{n}(\hat{\lambda} - \lambda) \sim N(0, \lambda).$$

위 정리에 의해 $\sqrt{n}(\hat{\lambda} - \lambda)$의 분산이 근사적으로 λ이므로 $\hat{\lambda}$의 표준 오차는 근사적으로 $\sqrt{\hat{\lambda}/n} = \sqrt{8.37/1207} = 0.083$로 추정한다. 마찬가지로 95% 신뢰구간은 $(8.37 - 1.96 \times 0.083, 8.37 + 1.96 \times 0.083) = (8.204, 8.530)$으로 구할 수 있다.

위 예제에서 $\hat{\lambda}$의 표준 오차를 붓스트래핑 방법으로 구할 수 있다. 붓스트래핑 방법의 원리는 모집단의 실제 분포를 추정하여 이 *추정 분포*로부터 같은 크기의 표본을 반복 추출하는 것이다. 실제 분포를 추정하

[1] 추정량 $\hat{\theta}$의 표준 편차(standard deviation)를 표준 오차라 부른다.

[2] 통계학자 J. Berkson이 1996년에 수행한 실험 데이터로 방사성 물질인 아메리슘-241(Americium-241)에서 방출되는 알파 입자수를 10초 간격으로 측정하였다.

는 방법으로는 경험적 분포(empirical distribution)가 있다. 경험적 분포는 관측한 자료 $X_1,\ldots,X_n$마다 확률 $1/n$을 가지는 분포로 알파 입자수 자료의 경험적 분포의 막대그래프는 그림 1b와 같다. 이 분포로부터 표본을 추출하는 것은 표본 $X_1,\ldots,X_n$으로부터 복원추출을 하는 것과 동일하다. 따라서 붓스트래핑으로 표준오차를 구하는 방법은 다음과 같다(Efron and Tibshirani 1993).

(1) 원자료로부터 복원 추출로 같은 표본 크기 만큼 추출한다. 이를 $X_1^*,\ldots,X_n^*$라고 하자.

(2) $X_1^*,\ldots,X_n^*$로부터 $\hat{\lambda}^* = \bar{X}^* = \dfrac{1}{n}\sum_{i=1}^{n} X_i^*$을 구한다.

(3) (1)∼(2)를 B번 반복하여 얻은 $\hat{\lambda}_1^*,\ldots,\hat{\lambda}_B^*$값들의 표본표준편차(sample standard deviation)를 다음과 같이 구하여 표준오차를 추정한다.

$$\sqrt{\frac{\sum_{i=1}^{B}(\hat{\lambda}_i^* - \overline{\hat{\lambda}^*})^2}{B-1}}$$

여기서 $\overline{\hat{\lambda}^*} = \dfrac{1}{B}\sum_{i=1}^{B}\hat{\lambda}_i^*$이다.

그림1c는 위와 같은 방법으로 (1)∼(2)를 100,000번 반복하여 얻은 $\bar{X}^*$의 히스토그램이다. 이 값의 표준편차, 즉 붓스트래핑으로 추정한 $\bar{X}$의 표준 오차는 0.084였다.

위와 같이 실제 모집단의 분포를 경험적 분포로 추정하는 방법을 비모수적 붓스트랩(nonparametric bootstrap)이라고 부른다. 만약 경험적 분포 대신 $\lambda = 8.37$인 포아송 분포, 즉 표본 평균값을 모수에 대입한 포아송 분포를 추정 분포로 사용하는 방법은 모수적 붓스트랩(parametric bootstrap)이라고 부른다.

비모수적 붓스트래핑 방법으로 $100(1-\alpha)\%$ 신뢰구간을 구하는 여러가지 방법 중 두 가지 간단한 방법은 다음과 같다.

(1) $\bar{X} \pm z(\alpha/2) \times \hat{se}$,

여기서 $z(\alpha/2)$는 표준정규분포의 상위 $\alpha/2$-분위수이고, $\hat{se}$는 앞에서 설명한 붓스트래핑 방법으로 구한 표준오차이다. 만약 $\bar{X}$의 분포가 정규분포와 유사할 때 사용하기 적합하다.

(2) $(\bar{X}^*(\alpha/2), \bar{X}^*(1-\alpha/2))$

여기서 $\bar{X}^*(\alpha/2)$와 $\bar{X}^*(1-\alpha/2)$는 각각 붓스트랩으로 구한 $\bar{X}^*$의 분포의 $\alpha/2$-분위수와 $(1-\alpha/2)$-분위수이다.

만약 $\bar{X}^*$의 분포가 $\bar{X}$를 중심으로 하는 정규분포와 유사한 형태이면 위의 두 신뢰구간은 비슷하다. 하지만 만약 정규분포의 형태가 아니면 첫번째 신뢰구간은 사용하면 안된다. 그림1c의 $\bar{X}^*$값으로부터 두 신뢰구간을 구하면 각각 (8.2014, 8.5326)와 (8.2013, 8.5327)로 유사하다. 이 외에도 붓스트랩 신뢰구간의 단점을 보완한 수많은 방법이 개발되었다 (DiCiccio and Efron 1992, 1996).

붓스트래핑은 여러가지 장점이 있다(Campbell and Torgerson 1999) 첫번째로 모집단의 분포의 모수적 형태를 가정하지 않아도 통계적 추정을 수행할 수 있다. 두번째로 복잡한 형태의 추정량에 대한 표준 오차나 신뢰구간을 추정하는데 유용하다. 세번째로 복잡한 이론적 배경 없이도 경험적 분포로 통계적 추정을

수행할 수 있다. 이처럼 통계적 추정에서 붓스트래핑은 강력한 도구이지만, 원자료에 주어진 정보 이상의 정보를 제공하지 않는 한계점이 있다. 따라서 원자료가 편향(biased)되어 있거나 모집단을 잘 대표하지 못하는 경우, 표본의 크기가 너무 작은 경우, 그리고 극단값을 추정하는 경우에는 붓스트래핑에 의한 추정치는 신뢰할 수 없다(Campbell and Torgerson 1999; Efron and Tibshirani 1993). 또한 추정해야 할 관심 모수가 너무 많은 경우에도 붓스트랩 사용에 주의를 기울여야한다(Efron 2003).

Felsenstein의 붓스트랩

1985년 Joe Felsenstein은 DNA 또는 RNA 서열 데이터로부터 추정된 계통수의 신뢰도를 측정하기 위해 붓스트랩 방법을 도입하였다(Felsenstein 1985). 이 방법은 최대가능도 추정법(Maximum likelihood estimation), neighbor-joining, UPGMA, maximum parsimony 등과 같은 방법으로 계통수를 구할 때 신뢰도를 측정하기 위한 수단으로 phylogenetics 분야에서 널리 사용되고 있다.

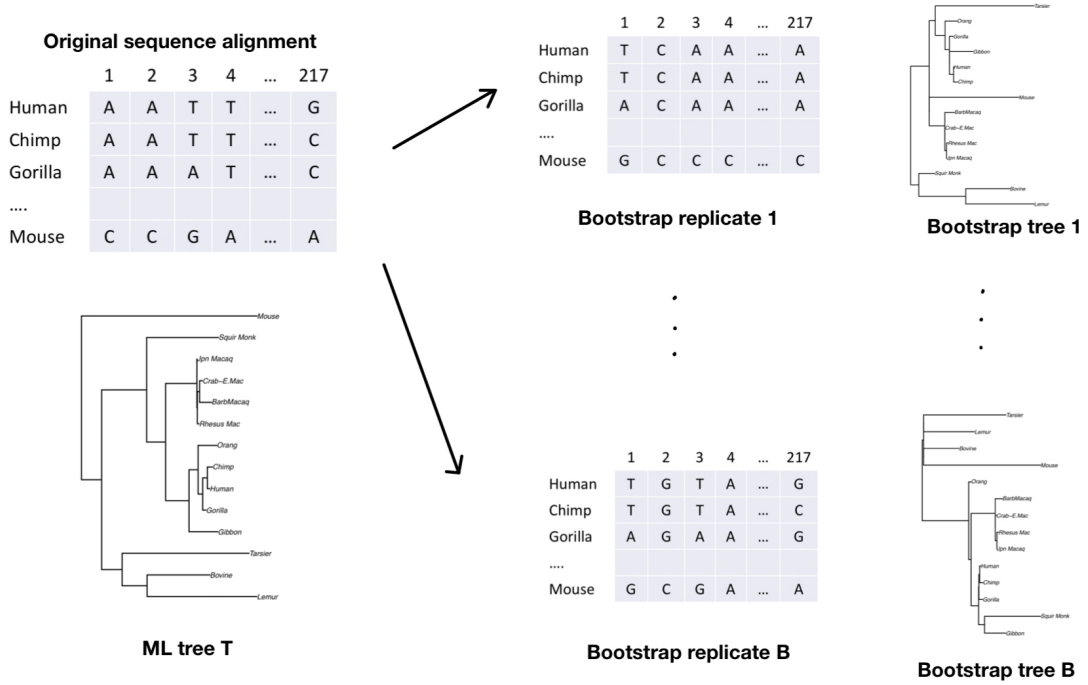

그림 2. 다중서열정렬로부터 추정한 최대가능도 나무(ML 나무)의 붓스트랩 절차. 원 다중서열정렬의 site들을 동일한 수의 열에 도달할 까지 복원추출하여 붓스트랩 복제본(bootstrap replicate)을 구한다. 각 정렬 복제본을 분석하여 ML 나무를 구한다. 이 과정을 B번 반복하여 구한 B개의 붓스트랩 나무를 요약하여 ML 나무의 각 clade에 대한 붓스트랩 비율을 구한다. 단, 본 그림의 원 염기서열 alignment은 붓스트랩 절차를 설명하기 위해 만든 유사 데이터(pseudo data)이다. 그리고 이 방법은 ML 나무 외 다른 방법으로 추정한 계통수에도 동일하게 적용할 수 있다.

여러 종의 DNA 염기 서열의 다중서열정렬(multiple sequence alignment)이 n개의 사이트로 구성되어 있다고 가정하자. 다중서열정렬의 사이트들을 $X_1,\ldots,X_n$으로 표기하고, 다중서열정렬 $\mathbf{X}=(X_1,\ldots,X_n)$로부터 구한 계통수를 $T=T(X_1,\ldots,X_n)$이라 하자. 계통수 T의 clade마다 신뢰도를 평가하기 위해 bootstrap 방법을 다음과 같은 절차로 수행한다(그림2).

Step 1. 각 사이트를 독립인 자료로 가정하여 원자료 $X_1,\ldots,X_n$으로부터 복원추출로 동일한 크기의 자료인

$X_1^*, \ldots, X_n^*$를 생성한다.

Step 2. $X_1^*, \ldots, X_n^*$로부터 계통수 $T^* = T(X_1^*, \ldots, X_n^*)$을 구한다.

Step 3. Step 1 $\sim$ Step 2를 B번 반복하여 $T_1^*, \ldots, T_B^*$ 계통수(이하 *붓스트랩 나무*)를 구한다.

Step 4. 원자료로부터 구한 계통수 T의 clade마다 똑같은 clade를 포함한 붓스트랩 나무의 비율을 구한다. 이를 각 clade의 붓스트랩 비율이라 부른다.

계통수를 unrooted tree로 추정하는 경우에는 분석 대상인 taxa 이중 분할하는 split의 붓스트랩 비율을 구한다. 본 논문에서는 편의상 unrooted tree인 경우에도 clade의 붓스트랩 비율이라고 지칭하겠다.

붓스트랩 비율은 해당 clade가 "참"인 계통수(true phylogenetic tree)[3]에 포함될 *신뢰도*(confidence)로 해석할 수 있다. 예를 들어, 한 clade의 붓스트랩 비율이 100%이면 B번의 반복시행 중 B번 모두 해당 clade의 분기점이 나타났다는 의미이고 붓스트랩 비율이 50%이면 B번의 반복시행 중 절반만 같은 분기점이 나타났다는 의미이다. 이 붓스트랩 비율은 해당 clade에 대한 "신뢰"평가로 해석할 수 있고, 붓스트랩 비율이 높을 수록 해당 clade가 더 신뢰할 수 있다는 것을 의미한다.

신뢰도 외에 붓스트랩 비율로부터 다양한 의미을 찾으려는 많은 연구가 있었다. 첫번째는 clade의 붓스트랩 비율을 실제 계통수가 해당 clade를 포함하는 확률로 해석하는 것이다. 계통수를 베이지안 방법으로 추정하는 경우 붓스트랩 방법과 유사하게 clade마다 비율을 구할 수 있는데, 이 경우 비율은 "참"인 계통수가 해당 clade를 포함할 사후확률(posterior probability)이다. 하지만 붓스트랩 비율과 사후확률 간 일반적인 연관성을 찾기는 어려웠다(Yang and Rannala 2005). 두번째는 붓스트랩 비율로부터 통계적 가설검정의 유의성을 구하는 것이다. 일반적으로 모평균에 대한 95% 신뢰구간은 통계적 가설검정에서 유의수준 5%에서 모평균에 대한 귀무가설(null hypothesis)을 기각하지 않는 값들의 집합이다. 이런 성질로부터 유추하여 계통수에서 clade의 붓스트랩 비율을 P라고 했을 때 $1-P$를 p-value로 사용하기 위한 수정된 붓스트랩 방법이 제안되었다(Susko 2010). 하지만 이러한 해석은 여전히 몇 가지 문제점이 있다. 그 중 하나로 통계적 가설검정에서 p-value는 귀무가설이 참이라고 가정했을 때 관측된 결과와 같거나 더 극단적인 결과가 관측될 확률을 의미하는데, 계통수 추정의 문제에서는 clade에 대한 올바른 귀무가설이 무엇인지가 명확하지 않는 것이다.

붓스트랩 절차에서 Step 4 대신 consensus 방법으로 붓스트랩 나무를 요약할 수도 있다. Consensus 나무를 구하는 방법은 여러가지가 있다. 첫번째 strict consensus tree는 모든 붓스트랩 나무에 포함된 clades만으로 계통수를 재구성한 나무이다. 두번재로 majority rule consensus tree는 과반수가 넘는 붓스트랩 나무가 포함하는 clades만으로 재구성한 계통수이다. 마지막으로 priority consensus tree는 majority rule consensus tree와 충돌하지 않는 clade들 중 붓스트랩 빈도가 높은 순으로 추가하며 계통수를 재구성한 것이다.

[3]참인 계통수(true phylogenetic tree)는 통계적 모형하에서 계통수로 통계적 추정 대상을 의미한다. 이는 실제로 발생한 진화과정과 다를 수 있다.

데이터 분석의 예

R 프로그램의 `phangorn`패키지에 있는 `primates` 데이터를 분석하였다[4]. 이 데이터는 12종의 영장류의 DNA 서열 데이터를 포함하여 총 217개의 사이트로 구성된 작은 데이터로, 이 예제는 영장류에 대한 최종적인 분석은 아니고 최대가능도 나무(maximum likelihood tree; 이하 ML 나무) 추정과 붓스트랩 분석의 여러 기본 단계를 수행하는 것을 보여주기 위한 것이다. 이 데이터를 분석하여 ML 나무를 구하기 위해 JC, F81, K80, HKY, GTR DNA 염기치환 모형과 사이트 간 진화속도의 이질성에 대한 범주가 네 개인 이산형 감마 모형의 가능한 조합을 후보로 하였다(서태건 2022). 베이지안 정보 기준(Bayesian information criterion, Schwarz (1978))을 기준으로 최종 모형을 선택하여 HKY DNA 염기치환 모형과 범주가 4개인 이산형 감마 모형으로 선택되었다. 그림 3는 이 모형하에서 구한 ML 나무를 보여준다.

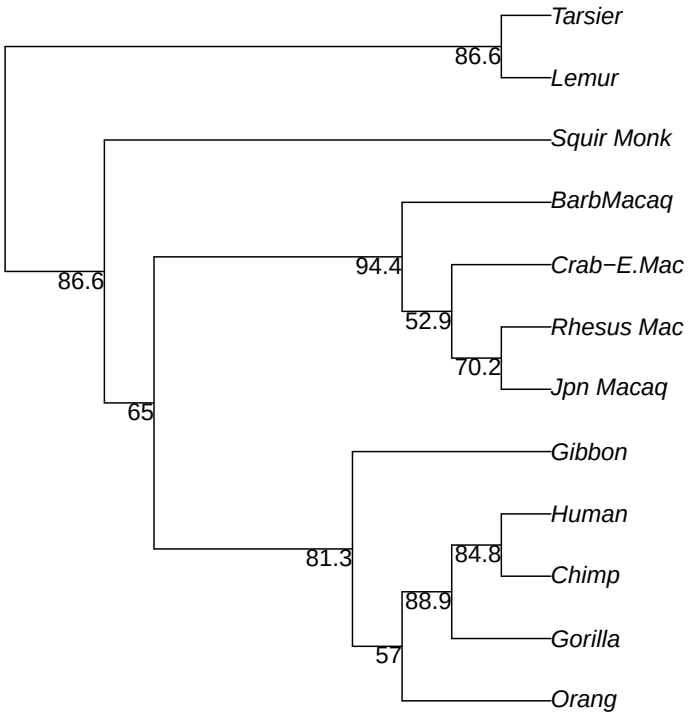

그림 3. 12종 영장류의 최대가능도 나무 위상(topology)과 붓스트랩 비율. ML 나무는 unrooted tree로 taxa 간 길이가 가장 긴 가지의 중간점(midpoint)을 root로 표현하여 그렸다.

붓스트랩 절차에 따라 Step 1 $\sim$ Step 2를 $B = 1{,}000$번 반복하여 붓스트랩 정렬을 생성하였다. 각 다중서열정렬로부터 ML 나무를 추정하여 1,000개를 붓스트랩 나무를 구하였다. 원 자료로 부터 구한 ML 나무(그림 3)의 clades의 붓스트랩 비율을 구하여, 붓스트랩 비율이 50%이상인 clades의 분기점에 붓스트

[4]본 예제에서 사용한 R 코드는 저자에게 요청시 제공할 수 있다

랩 비율을 표기하였다. 그림 3와 같이 일반적으로 50% 이상의 붓스트랩 값만 보고되고, 붓스트랩 비율은 붓스트랩 빈도(frequency)가 아닌 백분율로 표시하여 더 쉽게 읽고 다른 나무의 붓스트랩 결과와 비교가 가능하도록 한다.

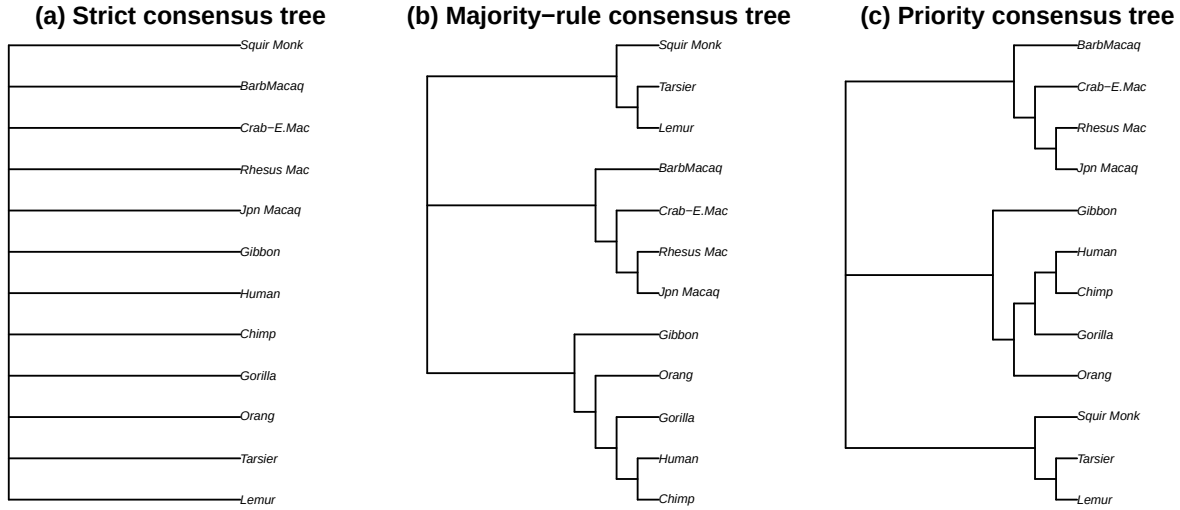

그림 4. `primates` 데이터의 세 가지 consensus trees: (a) strict consensus tree. (b) majority-rule consensus tree. (c) priority consensus tree.

그림 4은 앞에서 설명한 세 가지 consensus 나무를 보여준다. 그림 4a의 consensus 나무는 star tree 형태로 모든 붓스트랩 나무가 공통으로 포함하는 clade는 존재하지 않았다는 것을 알 수 있다. 그림 4b의 majority-rule consensus tree는 unrooted ML 나무 위상(topology)과 동일하다. 그 이유는 ML 나무(그림 3)가 포함하는 모든 clades는 모두 bootstrap 비율이 50%를 초과하기 때문이다. 같은 이유로 붓스트랩 비율이 큰 순으로 clade를 추가하여 재구성한 그림 4c의 priority consensus 나무도 ML 나무의 위상과 동일하다.

결론

본 논문에서는 통계학과 진화유전학에서 강력한 도구로 사용되는 붓스트랩의 원리를 예제와 함께 설명하였다. 계통수에 붓스트랩을 적용하는 방법을 설명한 Felsenstein의 논문은 역대 가장 많이 인용된 논문 중 하나로 붓스트랩은 계통수 재구성의 일반적인 도구가 되었다. 대부분의 계통수를 재구성하는 소프트웨어 패키지가 붓스트랩 알고리즘을 통합하고 있고 현재까지 많은 계통학 관련 학술지에서는 실제로 붓스트랩 값을 요구한다. 한 예로 이미 2000년 *Systematic Botany*와 2001년 *Systematic Biology*에 발표된 논문 중 최소 50%의 논문이 식물 계통학적 분석 결과를 제시했는데, 계통을 재구성한 두 저널의 모든 연구는 붓스트랩을 사용하여 clades의 붓스트랩 지지도를 측정하였다(Soltis and Soltis 2003). 또한 다양한 종에 대한 염기서열이 점점 더 많이 확보됨에 따라 빅데이터를 분석하여 수백 또는 수천 개의 분류군이 포함된 계통수의 재구성을 위한 프레임워크에 적용할 수 있도록 붓스트랩을 개선하는 노력도 계속 이어지고 있다(Lemoine *et al.* 2018).

Felsenstein은 붓스트랩 값을 정확도 측정이 아닌 반복성 측정으로 간주하였다(Felsenstein 1985). 다시 말해 붓스트랩 나무는 다중서열정리의 사이트를 반복적으로 리샘플링하여 구한 계통수들로 실제 계통

수가 아니다. 과거 이런 붓스트랩 값을 사용하여 정확도를 측정하려고 시도하거나 붓스트랩 신뢰도 값이 편향되었다는 비판도 있었다(Hillis and Bull 1993). 하지만 이후 붓스트랩 방법 자체가 편향된 것이 아니고 Felsenstein의 방법이 합리적인 수준의 신뢰도를 제공한다는 것이 입증되었다(Efron *et al*. 1996). 그럼에도 불구하고 실제 많은 연구자들이 붓스트랩을 통해 "진실"을 측정하기를 원하고 신뢰도를 제공하는 붓스트랩 값의 해석은 여전히 다소 직관적이지 않고 잘못 해석하기 쉽다. 이러한 논란 속에서도 현대에 붓스트랩을 표준 도구로 많이 사용하는 한 가지 이유를 Sanderson(1989)이 잘 요약하였다: "붓스트랩은 가정이 타당하다는 것을 인정하지 않더라도 미세변동(섭동; perturbation)에 대한 데이터의 견고성(robustness)을 평가하는 체계적인 방법을 제공하기 때문에 가치가 있다."

감사의 글

본 리뷰 논문의 주제를 제안하고 독려해주신 이화여대 김유섭 교수님과 논문의 질을 높여준 소중한 의견과 제안을 해주신 심사위원들에게 감사드린다. 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구이다(No. NRF-2021R1C1C1011250).

참고문헌

Bootstrapping. 2023 Dec. Wikipedia. https://en.wikipedia.org/wiki/Bootstrapping.

Campbell MK, Torgerson DJ. 1999. Bootstrapping: estimating confidence intervals for cost-effectiveness ratios. QJM: An International Journal of Medicine. 92(3):177-182.

DiCiccio T, Efron B. 1992. More Accurate Confidence Intervals in Exponential Families. Biometrika. 79(2):231-245.

DiCiccio TJ, Efron B. 1979. Bootstrap Confidence Interval. Statistical Science. 11(3):189-212.

Efron B. 1979. Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics. 7(1):1-26.

Efron B. 2003. Second Thoughts on the Bootstrap. Statistical Science. 18(2):135-140.

Efron B, Tibshirani RJ. 1993. An Introduction to the Bootstrap. Chapman & Hall/CRC.

Efron B, Halloran E, Holmes S. 1996. Bootstrap confidence levels for phylogenetictrees. Proceedings of the National Academy of Sciences. 93(23):13429-13429.

Felsenstein J. 1985. Confidence Limits on Phylogenies: An Approach Using the Bootstrap.Evolution. 39(4):783-791.

Felsenstein J. 2004. Inferring Phylogenies. Sinauer Associates, Inc., Sunderland, MA.

Hillis DM. Bull JJ. An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis. Systematic Biology. 42(2):182-192.

Lemoine F, Entfellner JD, Wilkinson E, Correian D, Felipe MD, de Oliveria T, Gascuel O. Renewing Felsenstein's Phylogenetic Bootstrap in the Era of Big Data. Nature. 556:452-456.

Rice JA. 2007. Mathematical Statistics and Data Analysis. Cengage Learning.

Sanderson MJ. 1989. Confidence limits on phylogenies: The bootstrap revisited. Cladistics. 5(2):113-129.

Schwarz G. 1978. Estimating the Dimension of a Model. The Annals of Statistics. 6(2):461-464.

Soltis PS. Soltis DE. 2003. Applying the Bootstrap in Phylogeny Reconstruction. Statistical Science. 18(2):256-267.

Susko E. 2010. First-Order Correct Bootstrap Support Adjustments for Splits that Allow Hypothesis Testing When Using Maximum Likelihood Estimation. Molecular Biology and Evolution. 27(7):1621-1629.

Yang Z. 2006. Computational Molecular Evolution. Oxford University Press.

Yang Z, Rannala B. 2005. Branch-Length Prior Influences Bayesian Posterior Probability of Phylogeny. Systematic Biology. 54(3):455-470.

서태건. 2022. DNA 염기치환 모형의 비교. 한국진화학회지. 1(1):88-104.

영문초록

Title: The Bootstrap in a Phylogenetic tree reconstruction

Abstract: Bootstrap is a powerful statistical tool in evolutionary genetics and statistics. In particular, it is one of the most popular tools for measuring confidence in phylogenetic tree reconstructions. This paper provides a review of the principles of the boostrap in statistics and how they can be used to measure the confidence of phylogenetic trees. We explain the bootstrap proportion of a clade and bootstrap consensus trees for measuring the confidence of phylogenetic trees, and provide an example of a bootstrap analysis of DNA data from 12 primate species. It also discusses the interpretation, benefits, and limitations of the bootstrap.

Authors: Yujin Chung [1,$]

Affiliation: [1] Department of Applied Statistics, Kyonggi University, Suwon, 16227, Republic of Korea

Corresponding author: [$] yujinchung@kyonggi.ac.kr