**Bootstrapping Phylogenies: Large Deviations and Dispersion Effects**

Michael A. Newton

*Biometrika*, Vol. 83, No. 2. (Jun., 1996), pp. 315-328.

Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28199606%2983%3A2%3C315%3ABPLDAD%3E2.0.CO%3B2-7

*Biometrika* is currently published by Biometrika Trust.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/bio.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org/
Thu Nov 2 12:33:06 2006

*Biometrika* (1996), **83**, 2, *pp.* 315–328
*Printed in Great Britain*

# Bootstrapping phylogenies: Large deviations and dispersion effects

BY MICHAEL A. NEWTON

*Department of Statistics, University of Wisconsin-Madison, 1210 West Dayton Street, Madison, Wisconsin 53706-1685, U.S.A.*

SUMMARY

A large deviation result is established for the bootstrap empirical distribution in a finite sample space, thereby validating both nonparametric and parametric bootstrapping in certain phylogenetic inference problems. The bias previously observed in the bootstrap distribution of the estimated tree topology is shown to stem from dispersion effects in the joint distribution of sample and bootstrap empirical distributions. Both results are examined for maximum likelihood estimation in a three-taxon model having particularly simple geometry. They are also applicable to discrete parameter problems outside of phylogenetic inference.

*Some key words*: Bias; Bootstrap efficiency; Cladistics; Discrete parameter space; DNA; Entropy; Molecular evolution; Relative entropy; Systematics; Tree topology.

## 1. INTRODUCTION

Ever-increasing volumes of molecular data are being used to infer evolutionary relationships among living populations and within families of genes. For example, differences in the nucleic acid sequences of extant species provide partial information about the phylogeny relating these species; that is to say, the nature of divergence from the single ancestral population to the present. A phylogeny consists of a discrete tree topology describing the pattern of relationships, and a set of continuous branch lengths indicating time into the past or the amount of evolution. Figure 1 shows a phylogenetic tree relating five primate species which was reconstructed using the method of maximum likelihood applied to mitochondrial restriction site data. In fact a variety of molecular data types are analysed and different methods are used to infer phylogenies. The reader is referred to Felsenstein (1983), Nei (1987) and Miyamoto & Cracraft (1990) for comprehensive reviews.

Felsenstein (1985) has advocated the use of Efron's (1979) bootstrap to assess the uncertainty in phylogeny reconstruction, and this method has become standard. While theory supports the use of bootstrapping in many models, e.g. Hall (1992), existing theory does not adequately support its use in phylogenetic inference. This is due in part to the unusual structure of the parameter space created by the discrete tree topologies, and in part to the type of questions involved. Furthermore, recent studies have suggested a bias in bootstrap estimates of the probability distribution of the estimated tree topology (Zharkikh & Li, 1992a, b; Hillis & Bull, 1993). Results presented in this paper give a theoretical underpinning, beyond first order, to the phylogenetic bootstrap and describe a source for the observed bias.

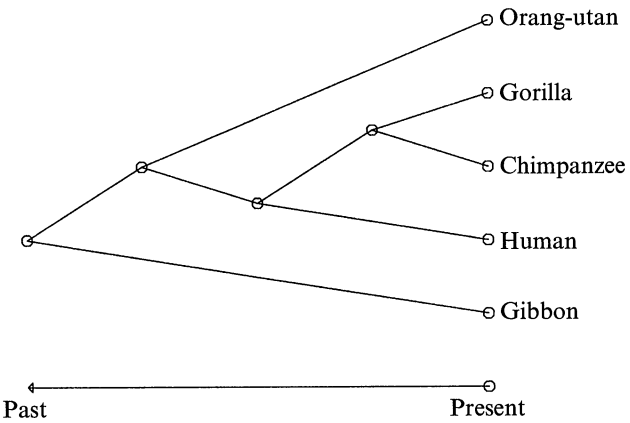


Fig. 1. A tree topology estimated from mitochondrial restriction site data (Felsenstein, 1992) for five primate species. Branch lengths are not estimated.

The approach is to study the probability of estimating any particular but incorrect tree topology. This is seen to be the probability that the empirical distribution of the sample lands in a certain set. Large deviation theory provides an asymptotic approximation to this sampling probability, and guarantees that the bootstrap analogue is at least as accurate. Insight on bias is gained by analysing the bootstrap empirical distribution and its relation to the empirical distribution of the sample.

In unpublished work, N. R. Chaganty and R. Karandikar have proved a general large deviation result for the bootstrap empirical measure from which some of the present results can be derived; see Barbe & Bertail (1995, p. 73). Their result is for function space valued random variables. Rather than proceeding from the general theory, we appeal to a theorem of Ellis (1984) concerning random vectors in $\mathscr{R}^d$.

After reporting the main theoretical results in the next section, we study their general implications for phylogenetic inference in § 3. Section 4 considers the particular case of maximum likelihood estimation in a three-taxon model where the large deviation approximation can be computed explicitly. Section 5 gives a brief discussion.

## 2. BOOTSTRAP EMPIRICAL DISTRIBUTION

Consider data $X_1, X_2, \ldots, X_n$ that are independent and identically distributed random variables on a finite sample space having $d > 1$ possible values. The set of possible probability measures for $X_i$ is taken as the compact subset of $\mathscr{R}^d$,

$$\mathscr{S}^d = \left\{ v = (v_1, v_2, \ldots, v_d) \in \mathscr{R}^d : v_j \geqslant 0, \sum_{j=1}^{d} v_j = 1 \right\},$$

with the relative topology. A nonempty subset $R \subset \mathscr{S}^d$ will be called a continuity set if it is contained in the closure of its interior. We suppose that a particular $P$ with $P_j > 0$ for all $j$ governs the $X_i$.

The relative entropy function,

$$I_P(v) = \begin{cases} \sum_{j=1}^{d} v_j \log(v_j/P_j) & \text{if } v \in \mathscr{S}^d, \\ \infty & \text{otherwise,} \end{cases}$$

is a convex nonnegative function, continuous on $\mathcal{S}^d$, with minimum 0 at $v = P$. By convention, we take $0 \log 0 = 0$.

Let $P_n \in \mathcal{S}^d$ denote the nonparametric maximum likelihood estimator of $P$:

$$P_n = (P_n(1), P_n(2), \ldots, P_n(d)),$$

where $nP_n(j)$ is the number of $X_i$ equal to the $j$th element of the sample space. Then

$$nP_n \sim \text{Multinomial}_d(n, P). \tag{1}$$

Of course, $P_n$ is the empirical distribution of the data.

A nonparametric bootstrap sample $Y_{n,1}, Y_{n,2}, \ldots, Y_{n,n}$ is a set of conditionally independent and identically $P_n$-distributed random variables, given the original data (Efron, 1979). The empirical distribution of the bootstrap sample, $Q_n$, then satisfies

$$nQ_n \sim \text{Multinomial}_d(n, P_n) \tag{2}$$

given $P_n$. The following result is proved by Sanov (1957).

THEOREM 1. *If $R \subset \mathcal{S}^d$ is a continuity set, then*

$$\lim_{n\to\infty} \frac{1}{n} \log \text{pr}(P_n \in R) = -I_P(R),$$

*where $I_P(R)$ is the minimum of $I_P$ over the closure of $R$.*

If the closure of $R$ contains $P$, then $I_P(R) = 0$, and thus Theorem 1 recapitulates the weak law of large numbers. Otherwise, we have an exponential decay to 0 of the probabilities, and the approximation $\text{pr}(P_n \in R) \doteq \exp\{-nI_P(R)\}$. The first new result, proved in the Appendix, is that bootstrapping approximates the sampling distribution of the empirical measure to the degree given in Theorem 1.

THEOREM 2. *For the nonparametric bootstrap,*

$$\lim_{n\to\infty} \frac{1}{n} \log \text{pr}(Q_n \in R | P_n) = -I_P(R) \tag{3}$$

*along almost every data sequence $X_1, X_2, \ldots$. If $P$ is not contained in the closure of $R$, then, marginally,*

$$\lim_{n\to\infty} \frac{1}{n} \log \text{pr}(Q_n \in R) = -J_P(R) > -I_P(R), \tag{4}$$

*where*

$$J_P(s) = \inf_{v \in \mathcal{S}^d} \{I_v(s) + I_P(v)\}, \tag{5}$$

*and $J_P(R)$ denotes the minimum of $J_P$ over the closure of $R$. Result (3) continues to hold for parametric bootstrap samples generated by a strongly consistent estimator of $P$.*

Generally, the conditional probability $\text{pr}(Q_n \in R | P_n)$ can be approximated by simulation. It is the limiting proportion, as the simulation size grows, of bootstrap distributions that land in $R$. The first result in Theorem 2 ensures a level of accuracy for the bootstrap which is not implied by earlier laws of large numbers or central limit theory. Both the parametric and nonparametric bootstrap accurately approximate exponentially small

probabilities in the sampling distribution of the empirical measure. Further, the approximation is automatic in the sense that the relative entropy $I_P(R)$ does not need to be computed. Result (4) confirms our intuition that the bootstrap distribution is more dispersed, marginally, than the true sampling distribution. In (5), the entropy $I_v(s)$ is defined as $\infty$ unless $s$ has zeros where $v$ has zeros.

The essence of bootstrapping is that $\text{pr}(Q_n \in R | P_n)$ is a computable surrogate for the unknown sampling probability $\text{pr}(P_n \in R)$. Indeed the approximation is rather accurate as Theorem 2 suggests. However, for the nonparametric bootstrap, we prove the following result in the Appendix.

THEOREM 3. *Suppose that $R \subset \mathcal{S}^d$ is a continuity set containing $P$ in its interior, and that the closure of $R$ is not all of $\mathcal{S}^d$. There exists $N$ such that, for all $n > N$,*

$$\text{pr}(Q_n \in R) = E\{\text{pr}(Q_n \in R | P_n)\} < \text{pr}(P_n \in R), \tag{6}$$

$$E\{\text{pr}(Q_n \in R | P_n) | P_n \in R\} < \text{pr}(P_n \in R). \tag{7}$$

Since $R$ contains $P$, all of these quantities converge to 1 as $n \to \infty$. Inequality (6) says that on average over data sets, the bootstrap probability underestimates the true sampling probability of a given set. Inequality (7) refines this by restricting the empirical distribution $P_n$ to be close to $P$.

## 3. PHYLOGENETIC INFERENCE

The motivation for Theorems 2 and 3 comes from the following problem. Suppose that data from each of $m$ populations or taxa can be represented as a sequence of length $n$, $(X_{1,j}, X_{2,j}, \ldots, X_{n,j})$, where $1 \leqslant j \leqslant m$ indicates the taxon and $X_{i,j}$ indicates a discrete datum observed at site $i$ in the $j$th taxon. For example, if the data are aligned DNA sequences, then each $X_{i,j}$ takes values in the four-letter alphabet $\mathcal{A} = \{A, C, T, G\}$. Alternatively, the $X_{i,j}$ may be binary indicators of the presence or absence of a particular base sequence cut by the $i$th restriction enzyme. We can express the complete data set as a sequence of column vectors

$$(X_1, X_2, \ldots, X_n) = \begin{pmatrix} X_{1,1} & X_{2,1} & \ldots & X_{n,1} \\ X_{1,2} & X_{2,2} & \ldots & X_{n,2} \\ \vdots & \vdots & & \vdots \\ X_{1,m} & X_{2,m} & \ldots & X_{n,m} \end{pmatrix},$$

one vector for each site. Each site vector $X_i$ can take on one of $d = \{\text{card}(\mathcal{A})\}^m$ possible values, and is the basic random element in what follows.

Most techniques of phylogenetic inference assume independence of $X_i$ from site to site in the sense that reconstruction is based on their empirical distribution $P_n$ and does not use information on the relative ordering of the sites. With this in mind, a natural regularity condition is to suppose that the method of estimating the tree topology induces a partition $R_1, R_2, \ldots, R_K$ of $\mathcal{S}^d$ formed by the rule

$$P_n \in R_k \Leftrightarrow \hat{\tau}(P_n) = \tau_k,$$

where $\tau_k$ is the $k$th tree topology, out of $K$, and $\hat{\tau}(P_n)$ is the estimated tree topology based on the data. See Felsenstein (1978b) on how $K$ grows with $m$. A further mild condition is that each set in the partition is a continuity set. Disjointness of the $R_j$ implies that a given

data set provides an unambiguous estimate of the topology. This may not be an important restriction, since some convention can be incorporated which breaks ties. Maximum likelihood estimation in an identifiable model satisfies this condition, as we show below.

Our second assumption is that $X_1, X_2, \ldots, X_n$ is a random sample from some $P \in \mathscr{P}^d$. Then $P \in R_k$ for some $k$, and so the topology estimator is inconsistent if the true topology $\tau = \tau_j$ for $j \neq k$. Compare with the inconsistency results of Felsenstein (1978a). Rates of evolution may vary between sites, as long as these rates are viewed as random so that the data at different sites are marginally identically distributed. Independence is a crucial assumption. The entire DNA sequence is often subsampled to make this assumption more reasonable.

Felsenstein (1985) describes the implementation of Efron's (1979) bootstrap in phylogenetic inference. The sampling distribution of the estimated topology is approximated by the empirical distribution of bootstrap estimates. Bootstrap samples are created by sampling $n$ sites with replacement from the observed sites. A topology estimator is applied to each bootstrap sample, inducing a distribution on the finite set of topologies which estimates the true sampling distribution of that topology estimator.

Although central limit theory is applicable to the empirical distributions $P_n$ and $Q_n$, it does not make sense to discuss $n^{-\frac{1}{2}}$-neighbourhoods of a discrete topology. Thus large deviation theory is appropriate. With the two regularity assumptions stated above, we have from Theorem 2 that

$$\lim_{n\to\infty} \frac{1}{n} \log \operatorname{pr}\{\hat{\tau}(P_n) = \tau_j\} = -I_P(R_j),$$

$$\lim_{n\to\infty} \frac{1}{n} \log \operatorname{pr}\{\hat{\tau}(Q_n) = \tau_j | P_n\} = -I_P(R_j),$$

the second limit being with $P$-probability one. When the true $P$ is contained in the closure of $R_j$, $I_P(R_j) = 0$ and we have the convergence in probability of the estimated topology to $\tau_j$ mimicked by the bootstrap. The bootstrap distribution concentrates on whatever topology is ultimately selected by $\hat{\tau}$. Bootstrapping will not diagnose inconsistency of the estimator. If $P$ is not contained in the closure of $R_j$, as happens if $\tau_j$ is an incorrect topology and $\hat{\tau}$ is consistent, then we observe the size of small probabilities and we see that the bootstrap matches.

Suppose that $\hat{\tau}$ is consistent and that $\tau$ is the correct tree topology. Theorem 3 implies that for large enough $n$,

$$E[\operatorname{pr}\{\hat{\tau}(Q_n) = \tau | P_n\} | \hat{\tau}(P_n) = \tau] < \operatorname{pr}\{\hat{\tau}(P_n) = \tau\}.$$

That is, the bootstrap probability assigned to topology $\tau$ underestimates the true sampling probability, on average, when the topology is correctly estimated. This bias does not persist, as both probabilities converge to 1 as $n \to \infty$, but may be significant for any fixed $n$.

## 4. Maximum likelihood and an example

One approach to phylogenetic inference is to model the evolutionary process so as to restrict $P$ to a parametric subset of $\mathscr{P}^d$. Felsenstein (1981, 1983, 1992), Barry & Hartigan (1987), Golding & Felsenstein (1990) and Navidi, Churchill & von Haeseler (1993), among others, have constructed parametric models and have advocated maximum likelihood estimation. Such models are based on branching Markov processes describing the evolution of each site from its ancestral state to its present distribution.

In parametric models, each topology $\tau_j$ corresponds to a subset of $\mathscr{S}^d$, that is the set of multinomial probability vectors obtained by varying the model parameters within the $j$th topology. Maximum likelihood creates a partition $\{R_1, R_2, \ldots, R_K\}$ of $\mathscr{S}^d$, where

$$R_j \supseteq \left\{ \rho \in \mathscr{S}^d : \sup_{v \in \tau_j} \prod_{l=1}^{d} v_l^{\rho_l} > \sup_{v \in \tau_k} \prod_{l=1}^{d} v_l^{\rho_l} \text{ for all } k \neq j \right\}.$$

Some convention must be established for boundary conditions, as is done in the example below.

For illustration, we present a simple model for three taxa and a binary alphabet $\mathscr{A} = \{0, 1\}$. There are three possible tree topologies relating the three taxa $A, B, C$, which we denote by

$$\tau_1 = ((A, B), C), \quad \tau_2 = ((A, C), B), \quad \tau_3 = ((B, C), A).$$

In topology $\tau_2$ for example, taxa $A$ and $C$ are more closely related than any other pair. The phylogeny's longest branch represents one unit of time, and the more recent divergence occurred at $t_0 \in (0, 1)$ time units before the present. Sites evolve independently and according to the same Markov process. The ancestral type at site $i$ is equivalent to a fair coin flip. Conditionally upon this value, two independent Markov processes record evolution along the branches. The transition matrix after $t$ time units is

$$\begin{pmatrix} 1-(1-e^{-\lambda t})/2 & (1-e^{-\lambda t})/2 \\ (1-e^{-\lambda t})/2 & 1-(1-e^{-\lambda t})/2 \end{pmatrix},$$

depending on a rate parameter $\lambda$. Along one branch, depending on the topology, the process splits at $t_0$ units before the present, and proceeds, conditionally independently, to the present.

By straightforward algebra, we obtain the marginal distribution of the present-day triples $X_i = (X_{i,1}, X_{i,2}, X_{i,3})$. Fourier analysis can be used to solve more complex models (Evans & Speed, 1993). Although there are $2^3$ possible values of $X_i$, symmetry reduces the sample space to four states in this case. Any state has the same probability as that obtained by interchanging labels 0 and 1. We may write the four states as

$$\sigma_1 = (0, 0, 0) \text{ or } (1, 1, 1), \quad \sigma_2 = (0, 0, 1) \text{ or } (1, 1, 0),$$

$$\sigma_3 = (0, 1, 0) \text{ or } (1, 0, 1), \quad \sigma_4 = (1, 0, 0) \text{ or } (0, 1, 1).$$

In general, the different topologies assess different probabilities on these states, allowing us to identify the most likely one given data. Table 1 shows probabilities of each state under each topology. State $\sigma_1$ is not informative since it has the same probability for all

Table 1. *Probabilities conferred on each state by each topology in the model*

| | State | | | |
|---|---|---|---|---|
| Topology | $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ |
| $((A, B), C)$ | $u$ | $v$ | $w$ | $w$ |
| $((A, C), B)$ | $u$ | $w$ | $v$ | $w$ |
| $((B, C), A)$ | $u$ | $w$ | $w$ | $v$ |

$u$, $(1+2e^{-2\lambda}+e^{-2\lambda t_0})/4$; $v$, $(1-2e^{-2\lambda}+e^{-2\lambda t_0})/4$; $w$, $(1-e^{-2\lambda t_0})/4$.

three topologies. Figure 2 shows this model as a subset of $\mathcal{S}^4$ when $t_0$, $\lambda$ and the topology are allowed to vary.

A topology $\tau_j$ is the maximum likelihood estimate if $P_n$ is closer to it than to either of the other two in terms of the distance $-\sum_{l=1}^{d} P_n(l) \log v_l$, where $v = (v_1, v_2, \ldots, v_d)$ varies within topologies. Introducing a convention for tie-breaking, we obtain the partition

$$R_1 = \{v \in \mathcal{S}^4 : v_2 \geqslant v_3,\ v_2 > v_4\},$$
$$R_2 = \{v \in \mathcal{S}^4 : v_3 \geqslant v_4,\ v_3 > v_2\},$$
$$R_3 = \{v \in \mathcal{S}^4 : v_4 \geqslant v_2,\ v_4 > v_3\}.$$

Figure 2(b) shows the boundary between these regions.

In this particularly simple model, the relative entropy function is readily minimised to obtain an explicit expression for the large deviation approximation. Consider the case where $P \in R_1$ and we are interested in the probabilities of incorrectly estimating topologies $\tau_2$ or $\tau_3$. Observe that $I_P(v)$ is minimised on the boundary connecting $R_1$ to either $R_2$ or $R_3$. In the first case, boundary points $v$ can be represented as $v = (a, b, b, 1 - a - 2b)$ with $(a, b)$ constrained to the triangle $\{(a, b) : a \geqslant 0, b \geqslant 0, a + 3b \geqslant 1, a + 2b \leqslant 1\}$. One obtains by routine calculus:

$$I_P(R_2) = \inf_{v \in R_2} I_P(v) = -\log\{\psi_1 + 2(\psi_2\psi_3)^{\frac{1}{2}} + \psi_4\},$$

where $P = (\psi_1, \psi_2, \psi_3, \psi_4)$. The chance of incorrectly estimating topology $((A, C), B)$ when the truth is $((A, B), C)$ is approximately $\exp\{-nI_P(R_2)\}$, and the nonparametric bootstrap provides at least as good an approximation. Similarly, $I_P(R_3) = -\log\{\psi_1 + 2(\psi_2\psi_4)^{\frac{1}{2}} + \psi_3\}$.

Table 2 presents a simple comparison of bootstrap and large deviation approximations to the probability of correctly estimating topology $\tau_1 = ((A, B), C)$ for different values of

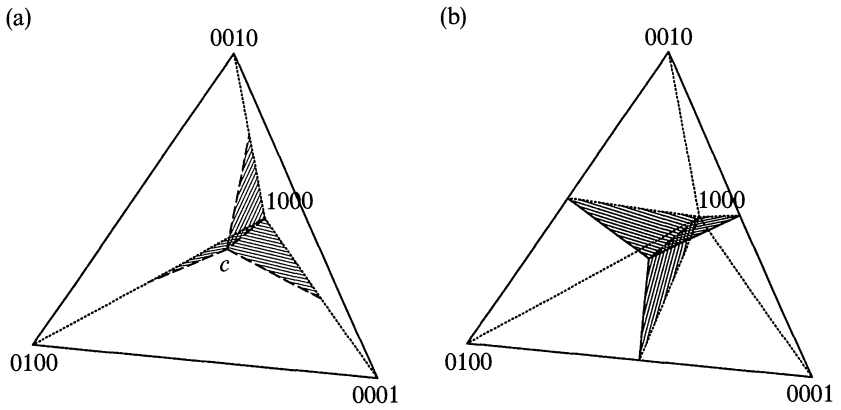


Fig. 2. Geometry of the 3-taxon model. (a) shows the model of Table 1 as a subset of $\mathcal{S}^4$. The vertex (1, 0, 0, 0) is farthest from view. Each point in the prism is a probability vector in $\mathcal{S}^4$. Each tree topology defines a planar region. For example, the topology $((A, B), C))$ corresponds to the region bounded by the line connecting (1, 0, 0, 0) to (0, 1, 0, 0). The centroid is $c = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. (b) shows how a maximum likelihood 'paper airplane' divides $\mathcal{S}^4$ into three regions. For example, if the empirical measure $P_n$ lands in the lower left region, then $((A, B), C)$ becomes the estimated topology. Points in a region are closer to the contained model subset than to either of the other two model subsets.

Table 2. *Comparison of approximations to the probability of estimating the true topology* $\tau_1 = ((A, B), C)$. *For each column, Monte Carlo estimates are based on* $5 \times 10^6$ *draws from the joint distribution of sample and bootstrap empiricals. Rows correspond to different probabilities in this joint distribution*

| | | $n = 100$ sites | | | | $n = 200$ sites | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $t_0 = \frac{1}{3}$ | | $t_0 = \frac{2}{3}$ | | $t_0 = \frac{1}{3}$ | | $t_0 = \frac{2}{3}$ | |
| | | $\lambda = 1$ | $\lambda = 2$ | $\lambda = 1$ | $\lambda = 2$ | $\lambda = 1$ | $\lambda = 2$ | $\lambda = 1$ | $\lambda = 2$ |
| Accuracy | | 0·997 | 0·932 | 0·739 | 0·488 | 1·000 | 0·988 | 0·859 | 0·554 |
| Average bootstrap | NP | 0·968 | 0·826 | 0·630 | 0·442 | 0·997 | 0·932 | 0·739 | 0·489 |
| | P | 0·980 | 0·831 | 0·641 | 0·420 | 0·999 | 0·945 | 0·766 | 0·467 |
| Average bootstrap | NP | 0·970 | 0·863 | 0·760 | 0·684 | 0·997 | 0·939 | 0·813 | 0·701 |
| (given correct) | P | 0·983 | 0·879 | 0·811 | 0·655 | 0·999 | 0·954 | 0·867 | 0·688 |
| Matching | NP | 0·969 | 0·845 | 0·720 | 0·655 | 0·997 | 0·935 | 0·783 | 0·665 |
| | P | 0·982 | 0·863 | 0·776 | 0·629 | 0·999 | 0·951 | 0·844 | 0·656 |
| Large deviation | | 0·977 | 0·584 | $<0$ | $<0$ | 1·000 | 0·913 | 0·233 | $<0$ |
| Empirical LD | | 0·852 | 0·425 | 0·056 | $<0$ | 0·983 | 0·721 | 0·240 | $<0$ |

'Accuracy' denotes the probability that the sample topology estimator is correct.
'Average boostrap' indicates the marginal probability that the bootstrap estimator is correct, and 'given correct' the conditional probability that the bootstrap estimator is correct given that the sample estimator is also correct.
'Matching' shows the probability that the boostrap and sample estimators coincide.
NP, nonparametric bootstrap; P parametric bootstrap.
'Large deviation' approximation is $1 - \exp\{-nI_P(R^2)\} - \exp\{-nI_P(R^3)\}$.
'Empirical LD' gives the average value of the sample large deviation approximation; that is the conditional expectation of $1 - \exp\{-nI_{P_n}(R^2)\} - \exp\{-nI_{P_n}(R^3)\}$ given correctness of the sample estimator.

the parameters. Parameters are chosen so that this sampling probability, sometimes called 'accuracy', is high, between about 0·5 and 1. Each tabulated probability is computed by averaging the indicators of an event in a large simulation ($5 \times 10^6$ draws) from the joint distribution of $(P_n, Q_n)$. Both nonparametric and parametric bootstrapping are considered, the latter being based on maximum likelihood estimates of the topology, $\lambda$, and $t_0$. The arbitrary tie-breaking convention used to define $\hat{\tau}(P_n)$ had an insignificant effect on the reported probabilities.

For all cases in Table 2, on average over data sets, the bootstrap estimate of accuracy, $\mathrm{pr}\{\hat{\tau}(Q_n) = \tau_1 | P_n\}$, underestimates the accuracy. This observation is consistent with (6). In the five cases having highest accuracy, this underestimation property of the nonparametric bootstrap continues conditionally on correct estimation of $\tau_1$. The conditional underestimation occurs to a lesser extent for the parametric bootstrap. Further, the average parametric bootstrap estimate is closer than the nonparametric bootstrap estimate to the accuracy, at least when the accuracy is high. An alternative bootstrap estimate of accuracy is $\mathrm{pr}\{\hat{\tau}(Q_n) = \hat{\tau}(P_n) | P_n\}$. On average, this also underestimates high accuracies, and tends to overestimate low accuracies.

As Theorems 1 and 2 suggest, small probabilities are the realm of large deviation approximations. Such approximations to the probability of incorrectly estimating the topology, should therefore be good when the accuracy is high. The results in Table 2 support this. In fact, the approximations are very poor for all but the highest accuracies.

That the bootstrap approximations are superior highlights the limits of large deviation theory to uncover the detailed structure of the bootstrap distribution.

There is a geometric explanation for bootstrap underestimation. When the probability of correctly estimating the topology is relatively high, $P_n$ will tend to be closer to the other topologies, i.e. to $R_j^c$, than $P$ is. The bootstrap proportion tends to be lower than the target probability because it is easier for the bootstrap sample to escape $R_j$. Figure 3 illustrates this argument by projecting the three-taxon model of Fig. 2 onto the facing triangle.

Figure 4 shows an estimate of the sampling distribution of the nonparametric bootstrap proportion for one case, the second column of numbers in Table 2. To compute this, 2000 samples $P_n$ are drawn, and, for each one leading to $\hat{\tau}(P_n) = \tau_1$, about 93%, 10 000 samples $Q_n$ are drawn from the bootstrap distribution, and the bootstrap proportion $\mathrm{pr}\{\hat{\tau}(Q_n) = \tau_1 | P_n\}$ is recorded. Note the significant skewness in this distribution. Although theoretically and from Table 2 there is a bias, in the mean sense, the median bootstrap proportion is very close to the true accuracy in this example.

## 5. Is there something wrong with the bootstrap?

The discreteness of the parameter space in phylogenetic analysis, and the need to associate measures of uncertainty with inferred tree structures, raise statistical questions that existing theory does not address. Efron's bootstrap naturally produces an estimate for the sampling distribution of any topology estimator, but several studies have suggested a bias. Zharkikh & Li (1992a, b) establish bias analytically for a particular estimator in a four-taxon model. Hillis & Bull (1993) make similar observations in a simulation study. Put simply, if the chance of correctly estimating the phylogeny is high, then, on average over data sets, the bootstrap proportion estimating this probability is smaller. Felsenstein &

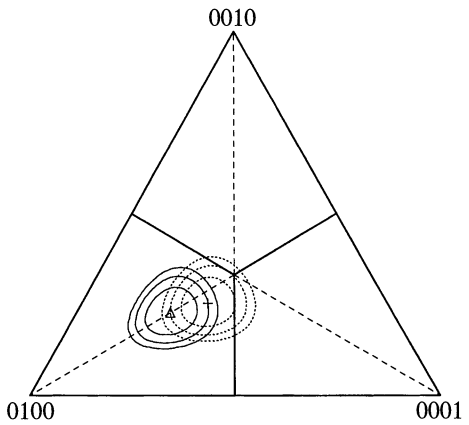


Fig. 3. Bootstrap underestimation: the triangle represents the projection of $\mathscr{S}^4$ of Fig. 2 onto the near face. Solid contours indicate the true sampling distribution of $P_n$, and hence of the topology estimator. Dashed contours indicate the bootstrap distribution of $Q_n$ conditional on a particular $P_n$, and are drawn to illustrate that probability in this distribution leaks out of region $R_1$ into regions leading to false topology estimates. Of course the contours mask the true discreteness of both distributions.

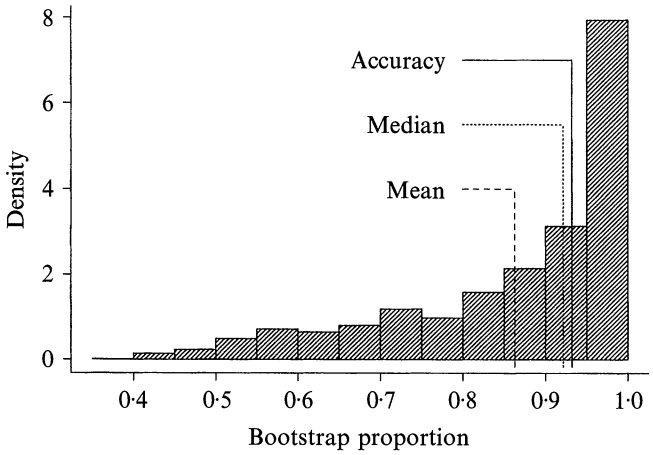


Fig. 4. Bootstrap bias: the histogram summarises 1855 draws for the conditional distribution of a bootstrap proportion, given correct estimation of the topology, in a case of high accuracy; see Table 2, second column of numbers.

Kishino (1993) reproduce the same behaviour in a simple model and suggest that the problem is not inherent to bootstrapping. Results presented here show that this bias is a general phenomenon, occurring for any number of taxa, for a wide range of topology estimators, and under mild assumptions on the sampling process. In fact, the results extend beyond phylogenetic analysis to general discrete parameter inference. This bias is a property of the joint distribution of sample and bootstrap empirical distributions, and diminishes with increasing sample size. Notwithstanding the bias, the conditional distribution of the bootstrap topology estimator accurately approximates the sampling distribution of the topology estimator.

Large deviation theory establishes these general statements. The large deviation approximation itself is often hard to calculate, and appears to be worse than the bootstrap approximation. We do not advocate the use of large deviations to estimate probabilities. The method is invoked simply to enable a general comparison of bootstrap and sampling probabilities.

That the bootstrap underestimates high accuracies, on average, has been interpreted as a conservative feature. One is tempted to conclude, upon observing a high bootstrap proportion, that the true accuracy is likely to be higher. This conclusion is unfounded, however, as the example of Fig. 4 demonstrates. The median bias, outside of our theoretical analysis, is much smaller than the mean bias in this example. When the true accuracy is high, it may be that about half the bootstrap proportions are smaller and about half are larger, even though on average in magnitude they are smaller.

When the number of tree topologies is large compared to the number of sites, it may be that no single tree yields a high bootstrap proportion. Thus the relevance of computations involving extreme probabilities is questionable. However, the systematist may be interested instead in the proportion of bootstrap trees in which a particular subgroup of taxa is monophyletic, that is to say, forms a branch by itself. Being an agglomeration of different tree topologies, this set may yield a high bootstrap proportion. Furthermore, the probability that the topology estimator possesses such a branch is simply the probability

that the empirical distribution lands in a certain set, and thus all the theory developed here applies.

Theorem 2 guarantees the quality of the parametric bootstrap in cases where the model is correctly specified, or more weakly, when the model produces a consistent estimator of $P$. Simulation results indicate a smaller bias than for the nonparametric bootstrap. Most users prefer the nonparametric bootstrap because it relies on fewer assumptions. It is worth noting, however, that parametric bootstrapping provides a more natural framework for testing hypotheses about models of evolution.

More work is needed to convert bootstrap sampling distributions into inference summaries such as hypothesis tests and confidence tests. Simply removing the bias may not be wise. For example, having the bootstrap sample size larger than $n$ would counteract the underestimation of high probabilities, but may not fairly reflect the uncertainty involved. See Zharkikh & Li (1995) for a different proposal. On another note, since the bootstrap is trying to mimic a sampling process, it is unwise to resample only informative sites from the raw data, as is sometimes advocated.

## Ackowledgement

The author thanks Joseph Felsenstein, Charles Geyer, David Mason, Robert Mau and Jens Præstgaard for extremely useful conversations on this topic, N. R. Chaganty for providing a copy his technical report with R. Karandikar, and the Editor and a referee for suggesting improvements. This work is a partial response to questions raised by Professor Felsenstein during several discussions on bootstrapping and evolutionary genetics.

## Appendix

### *Proofs*

The proof of Theorem 2 hinges on the following fundamental result. Following Ellis (1984), let $Z_1, Z_2, \ldots$ be $\mathscr{R}^d$-valued random variables, $Z_n$ being defined on a probability $(\Omega_n, \mathscr{F}_n, \tilde{P}_n)$. Let $\{a_n\}$ be a divergent sequence of positive integers, and for $t \in \mathscr{R}^d$, define extended-real-valued functions

$$c_n(t) = \frac{1}{a_n} \log \int \exp\langle t, Z_n \rangle \, d\tilde{P}_n,$$

where $\langle ., . \rangle$ is Euclidean inner product. Up to the constants $a_n$, these functions are the cumulant generating functions of the $Z_n$. Assume that, for all $t \in \mathscr{R}^d$,

$$c(t) = \lim_{n \to \infty} c_n(t)$$

exists, taking $c(t) = \infty$ if $c_n(t) = \infty$ for all $n > N_t$. As a further regularity condition, suppose that $c(t)$ is a closed convex function (Rockafellar, 1970, p. 52) and that the effective domain $\mathscr{D}(c) = \{t \in \mathscr{R}^d : c(t) < \infty\}$ has nonempty interior containing $t = 0$. The Legendre–Fenchel transform of $c(t)$, also called the conjugate of $c(t)$ or the entropy function, is, for $s \in \mathscr{R}^d$,

$$I(s) = \sup_{t \in \mathscr{R}^d} \{\langle t, s \rangle - c(t)\}.$$

Ellis (1984), extending Gärtner (1977), established the following large deviation theorem.

Theorem 4. *For any closed Borel set* $K \subset \mathscr{R}^d$,

$$\limsup_{n \to \infty} \frac{1}{a_n} \log \tilde{P}_n(Z_n / a_n \in K) \leqslant - \inf_{s \in K} I(s).$$

*If $c$ is differentiable on the interior of $\mathscr{D}(c)$ and is steep, see below, then, for any open Borel set $G \subset \mathscr{R}^d$,*

$$\liminf_{n\to\infty} \frac{1}{a_n} \log \tilde{P}_n(Z_n/a_n \in G) \geqslant -\inf_{s\in G} I(s).$$

Steepness is a property of the gradient of a convex function near its boundary. If $\mathscr{D}(c) = \mathscr{R}^d$ and $c$ is differentiable on all of this domain, then it is steep (Rockafellar, 1970, p. 251).

It is well known that Theorem 1 follows as a special case of Theorem 4. See for example Ellis (1985, pp. 250–3) or Bucklew (1990, pp. 25–8).

LEMMA 1. *Let $s$ and $t$ denote distinct points in the interior of $\mathscr{S}^d$. Construct by element-wise multiplication a third point $x = (s \times t)^{1/2}/c$ where $c = \langle s^{1/2}, t^{1/2} \rangle$ is a normalising constant. Then*

$$D := I_t(s) - I_t(x) - I_x(s) > 0.$$

*Proof.* Straightforward algebra gives

$$\begin{aligned} D &= \langle s, \log(s/t) \rangle - \langle s, \log(s/x) \rangle - I_t(x) \\ &= \langle s, \log(x/t) \rangle - I_t(x) \\ &= \langle s, \log\{s^{\frac{1}{2}}/(ct^{\frac{1}{2}})\} \rangle - I_t(x) \\ &= \tfrac{1}{2} I_t(s) - \log c - I_t(x). \end{aligned}$$

By Jensen's inequality, $\log c > -\frac{1}{2} I_t(s)$. A second application of Jensen's inequality gives $I_t(x) < \log\langle x^2, 1/t \rangle = -2 \log c$. Combining these three facts yields the result. □

*Proof of Theorem* 2. Identify $Z_n$ of Theorem 4 with the bootstrap empirical measure: $Z_n = nQ_n$. The sequence of probability spaces $(\Omega_n, \mathscr{F}_n, \tilde{P}_n)$ corresponds to bootstrap probability and is determined by (2). We may have a different sequence $\tilde{P}_n$ for each infinite realisation of data. Conditionally upon the data,

$$c_n(t) = \frac{1}{n} \log \int \exp\langle t, nQ_n \rangle \, dP_n = \log\langle e^t, P_n \rangle.$$

For each $t \in \mathscr{R}^d$, by the strong law of large numbers,

$$\lim_{n\to\infty} c_n(t) = c_I(t) = \log\langle e^t, P \rangle \tag{A1}$$

along all data sequences $X_1, X_2, \ldots$ except those in a null set $N$ depending, perhaps, on $t$. Then (A1) holds simultaneously for all $t$ in a countably dense subset of $\mathscr{R}^d$, except for data sequences in a null set. Since $c_I(t)$ is convex, it follows from Theorem 10.8 of Rockafellar (1970) that the limit holds simultaneously for all $t \in \mathscr{R}^d$ except for a null set of data sequences. Off this null set, the limit of $c_n(t)$ is as in Theorem 1, establishing the limit for the nonparametric bootstrap. The argument for the parametric case is identical.

To establish (4), $\tilde{P}_n$ now corresponds to the marginal distribution of $nQ_n$. The normalised cumulant generating function is, for $t \in \mathscr{R}^d$,

$$c_n(t) = \frac{1}{n} \log E\{\exp\langle t, nQ_n \rangle\} = \frac{1}{n} \log E\{E(\exp\langle t, nQ_n \rangle \mid P_n)\} = \frac{1}{n} \log E\{\langle e^t, P_n \rangle^n\},$$

noting (2). Hölder's inequality ensures that all functions in this sequence are convex. Applying the integration theorem of Varadhan (1966), see also Ellis (1985, p. 51) or Dembo & Zeitouni (1993, p. 120), we have

$$\lim_{n\to\infty} c_n(t) = \sup_{v \in \mathscr{S}^d} \{\log\langle e^t, v \rangle - I_P(v)\} =: c_J(t). \tag{A2}$$

Being the pointwise limit of convex functions, $c_J$ is also convex. Noting that $\log\langle e^t, v \rangle$ is itself the

conjugate of the relative entropy $I_v$, we have

$$c_J(t) = \sup_{v \in \mathscr{P}^d} \sup_{s \in \mathscr{P}^d} \{\langle s, t\rangle - I_v(s) - I_P(v)\} = \sup_{s \in \mathscr{P}^d} \{\langle s, t\rangle - J_P(s)\},$$

with $J_P$ as defined in (5). Thus $c_J(t)$ is the conjugate of the entropy function $J_P(s)$. Strict convexity of $I_v(s)$ translates into strict convexity of $J_P(s)$ for $s \in \mathscr{P}^d$, and therefore, by Theorem 26.3 of Rockafellar (1970), $c_J(t)$ is differentiable on $\mathscr{R}^d$. The conditions of Theorem 4 are thus satisfied, implying that, in the limit, $n^{-1} \log \text{pr}(Q_n \in R)$ exists and equals the infimum over $R$ of the marginal entropy function (5).

It remains to verify that for any continuity set $R$, not containing $P$ in its closure,

$$J_P(R) := \inf_{v \in R} J_P(v) < \inf_{v \in R} I_P(v) =: I_P(R). \tag{A3}$$

To establish this, note that

$$J_P(s) = \inf_{v \in \mathscr{P}^d} \{I_v(s) + I_P(v)\} \leqslant I_x(s) + I_P(x),$$

where $x = (s \times P)^{\frac{1}{2}}/c$. If $s$ is in the interior of $\mathscr{P}^d$ and $s \neq P$, then by Lemma 1 with $t = P$, $J_P(s) < I_P(s)$. If $s = P$, then both entropy functions equal zero. The last case has $s$ on the boundary of $\mathscr{P}^d$, where at least one entry equals zero. That $J_P(s) < I_P(s)$ can then be established by an argument similar to the one above, but using a different choice for $x$. Details are omitted. □

*Proof of Theorem* 3. Applying Theorems 1 and 2 to the complement of $R$, given $\varepsilon = I_P(R^c) - J_P(R^c) > 0$, there exists an $N$ such that, for all $n > N$,

$$|n^{-1} \log \text{pr}(P_n \in R^c) + I_P(R^c)| < \varepsilon/2, \quad |n^{-1} \log \text{pr}(Q_n \in R^c) + J_P(R^c)| < \varepsilon/2.$$

For such $n$, therefore,

$$\frac{1}{n} \log \text{pr}(Q_n \in R^c) > \frac{1}{n} \log \text{pr}(P_n \in R^c)$$

proving inequality (6).

For (7), it is equivalent to prove that, for sufficiently large $n$,

$$\frac{1}{n} \log E\{\text{pr}(Q_n \in R^c | P_n) | P_n \in \mathbb{R}\} > \frac{1}{n} \log \text{pr}(P_n \in R^c). \tag{A4}$$

By Theorem 1, the right-hand side tends to $-I_P(R^c)$. The left-hand side is

$$-\frac{1}{n} \log \text{pr}(P_n \in R) + \frac{1}{n} \log \int_R \text{pr}(Q_n \in R^c | P_n = v)\, d\mu_n(v),$$

where $\mu_n$ is given by (1). The first term above tends to 0 since $P$ is in the interior of $R$. The second term is amenable to Varadhan's (1966) integration theorem. The mode of the integrand dominates, and we obtain

$$\lim_{n \to \infty} \frac{1}{n} \log \int_R \text{pr}(Q_n \in R^c | P_n = v)\, d\mu_n(v) = \sup_{v \in R} \{-I_v(R^c) - I_P(v)\} = -\inf_{v \in R} \{I_v(R^c) + I_P(v)\}.$$

Thus (A4) and thus (7) follow if we can establish

$$\inf_{v \in R} \{I_v(R^c) + I_P(v)\} < I_P(R^c), \tag{A5}$$

or equivalently, if we can identify a point $x \in R$ for which

$$I_x(R^c) + I_P(x) < I_P(R^c). \tag{A6}$$

Let $s$ in the closure of $R^c$ denote a point where $P$ projects onto $R^c$, that is, such that

$$I_P(s) = \inf_{v \in R^c} I_P(v).$$

If $s$ has no zero elements, define $x = (s \times P)^{\frac{1}{2}}/c$, where $c$ normalises $x$ to sum to one. It follows, on representing $I_x(v) = \{I_P(v) + I_t(v)\}/2 + \log c$, that $x$ also projects onto $R^c$ at $s$. Therefore, $I_x(R^c) = I_x(s)$ and $I_P(R^c) = I_P(s)$. Inequality (A6), and thus the main result (7), follow by applying Lemma 1 with $t = P$. If $s$ has zero elements, then a more careful choice of $x$ is required, but the same argument works. □

## REFERENCES


BARBE, P. & BERTAIL, P. (1995). *The Weighted Bootstrap*, Lecture Notes in Statistics, **98**, Ed. J. Berger, et al. New York: Springer-Verlag.

BARRY, D. & HARTIGAN, J. A. (1987). Statistical analysis of hominoid molecular evolution. *Statist. Sci.* **2**, 191–210.

BUCKLEW, J. A. (1990). *Large Deviation Techniques in Decisions, Simulation, and Estimation.* New York: John Wiley.

DEMBO, A. & ZEITOUNI, O. (1993). *Large Deviations Techniques and Applications.* London: Jones and Bartlett.

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7**, 1–26.

ELLIS, R. S. (1984). Large deviations for a general class of random vectors. *Ann. Prob.* **12**, 1–12.

ELLIS, R. S. (1985). *Entropy, Large Deviations, and Statistical Mechanics.* New York: Springer-Verlag.

EVANS, S. N. & SPEED, T. P. (1993). Invariants of some probability models used in phylogenetic inference. *Ann. Statist.* **21**, 355–77.

FELSENSTEIN, J. (1978a). Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**, 401–10.

FELSENSTEIN, J. (1978b). The number of evolutionary trees. *Syst. Zool.* **27**, 27–33.

FELSENSTEIN, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Molec. Evol.* **17**, 368–76.

FELSENSTEIN, J. (1983). Statistical inference of phylogenies (with Discussion). *J. R. Statist. Soc.* A **146**, 246–72.

FELSENSTEIN, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**, 783–91.

FELSENSTEIN, J. (1992). Phylogenies from restriction sites: A maximum likelihood approach. *Evolution* **46**, 159–73.

FELSENSTEIN, J. & KISHINO, H. (1993). Is there something wrong with the bootstrap? A reply to Hillis and Bull. *Syst. Biol.* **42**, 193–200.

GÄRTNER, J. (1977). On large deviations from the invariant measure. *Theory Prob. Applic.* **22**, 24–39.

GOLDING, B. & FELSENSTEIN, J. (1990). A maximum likelihood approach to the detection of selection from a phylogeny. *J. Molec. Evol.* **31**, 511–23.

HALL, P. (1992). *On the Bootstrap and Edgeworth Expansion.* New York: Springer-Verlag.

HILLIS, D. M. & BULL, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.* **42**, 182–92.

MIYAMOTO, M. M. & CRACRAFT, J. (Ed.) (1991). *Phylogenetic Analysis of DNA Sequences.* Oxford University Press.

NAVIDI, W. C., CHURCHILL, G. A. & VON HAESELER, A. (1993). Phylogenetic inference: Linear invariants and maximum likelihood. *Biometrics* **49**, 543–55.

NEI, M. (1987). *Molecular Evolutionary Genetics.* New York: Columbia University Press.

ROCKAFELLAR, R. T. (1970). *Convex Analysis.* Princeton: Princeton University Press.

SANOV, I. (1957). On the probability of large deviations of random variables (in Russian). *Mat. Sb.* **32**, 11–44. English translation in *Selected Translations in Mathematical Statistics*, I (1961), Trans. D.E.A. Quade, pp. 213–44. Providence, RI: Am. Math. Soc.

VARADHAN, S. R. S. (1966). Asymptotic probabilities and differential equations. *Commun. Pure Appl. Math.* **19**, 261–86.

ZHARKIKH, A. & LI, W.-H. (1992a). Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. *Molec. Biol. Evol.* **9**, 1119–47.

ZHARKIKH, A. & LI, W.-H. (1992b). Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. II. Four taxa without a molecular clock. *J. Molec. Evol.* **35**, 356–66.

ZHARKIKH, A. & LI, W.-H. (1995). Estimation of confidence in phylogeny: The complete-and-partial bootstrap technique. *Mol. Phylogenet. Evol.* **4**, 44–63.


[*Received October* 1994. *Revised July* 1995]