On the interpretation of bootstrap trees: appropriate threshold of clade selection and induced gain

Berry Vincent and Gascuel Olivier

LIRMM

UMR 9928 UNIVERSITE MONTPELLIER II/CNRS

161, Rue Ada

34392 Montpellier cedex 5 France

email: {berry,gascuel}@lirmm.fr

tel: (33) 67 41 85 85

fax: (33) 67 41 85 00

Running Head: Interpretation of bootstrap trees

Key words: bootstrap method, threshold of clade selection, topological distance, Type I and Type II error, bias/variance compromise, maximum parsimony, neighbor-joining, computer simulations.

Abstract

In this study we address the problem of interpreting a bootstrap tree. The main issue is to choose the threshold of clade selection in order to separate reliable clades from unreliable ones, depending on their bootstrap proportion. This threshold depends on the chosen error measure. We investigate error measures that stem from a generalization of Robinson and Foulds's (1981) distance, used to quantify the divergence between the true phylogeny and the estimated trees. We propose two analytical approximations of the optimum threshold of clade selection to interpret (*i.e.*, reduce) the bootstrap tree. We performed extensive simulations along the lines of Kuhner and Felsenstein (1994), using the neighbor joining and the maximum parsimony methods. These simulations show that our approximations cause only small losses in quality when compared to the optimum threshold resulting from empirical observation. Next we measured the error reduction achieved when estimating the true phylogeny by the properly reduced bootstrap tree rather than by the complete original tree, obtained with a classical tree-building method. Our simulations on short sequences show that an error reduction of 39% is achieved with the parsimony method and an error reduction of 33% is achieved with the distance method, when the error is measured with the standard Robinson and Foulds distance. The observed error reduction is shown to originate from an important decrease in Type I error (wrong inferences), while Type II error (omitted correct clades) is only slightly increased. Greater error reduction is achieved when shorter sequences are used, and when more importance is given to Type I error than to Type II error. To investigate the causes of error from another point of view, we propose a general decomposition of the error expectation in two terms of bias, and one of variance. Results for these terms show that no fundamental bias is intro-

duced by the bootstrap process, the only source of bias being structural (lack of resolution). Moreover, the variance in the estimations is greatly reduced, providing another explanation for the better results of the reduced bootstrap tree compared with the original tree estimate.

Introduction

The bootstrap (Efron 1979; Efron and Tibshirani 1993) is a computer-based technique which makes it possible to characterize the behavior of almost any statistical estimate. Felsenstein (1985) introduced the use of the bootstrap method in the phylogenetic field to assess the reliability of estimated trees. Let t be an (unknown) actual phylogeny and $\hat{t}$ an estimate of this phylogeny obtained by using any tree-building method. Felsenstein's method consists in generating bootstrap samples by randomly selecting sites from the original data sample with replacement. These pseudosamples are then analyzed using the original tree-building method, each time giving a bootstrap tree $\hat{t}^*$. The series of bootstrap trees thus obtained is used to calculate the *bootstrap proportions* for the clades of the original estimate $\hat{t}$. The bootstrap proportion of a given clade is simply the proportion of times this clade is represented in the series of bootstrap trees, and may be thought of as a "confidence" assessment for the given clade. Felsenstein's method thus provides an *average bootstrap tree*, denoted $\bar{t}^*$ in the following, having the same topology as the original estimate $\hat{t}$, but whose internal branches are labeled by the bootstrap proportions.

Felsenstein's bootstrap method is simple and widely employed in phylogenetic studies. However, since Hillis and Bull's (1993) paper, an intense and rather technical discussion (for review, see Sanderson 1995) has been engaged in as to its deep statistical meaning. In fact, there are three points of view about the bootstrap method and the meaning of the bootstrap proportion:

(1) Efron (1979) and Felsenstein (1985) viewed the bootstrap proportion as a *repeatability* measure. The basic idea is that the observed distribution of the bootstrap tree $\hat{t}^*$ (for the various pseudo-samples) around the original

estimate $\hat{t}$ enables one to infer the unobservable distribution of the estimate $\hat{t}$ around the true tree t. In other words, when the bootstrap proportion of a given clade of $\hat{t}$ is high, we are quite confident that this clade would be inferred again if another original data sample was available and analyzed by the same tree-building method. When the bootstrap proportion is low, the repeatability is judged to be low and the corresponding clade may be suspected to be erroneous. Thinking of the bootstrap proportion as a measure of repeatability therefore leads to discarding clades which would not be inferred sufficiently frequently, when using other data sets. Note that this interpretation already amply justifies the use of the bootstrap method. If the data contain little "phylogenetic signal" and if the reconstruction method selects trees in a near-random way, the bootstrap method will detect the flaw and lead to an irresolution, which is better than a false resolution.

(2) The second interpretation (Sanderson 1989; Zharkikh and Li 1992a,b; Hillis and Bull 1993) is much more ambitious than the previous one. It consists in looking at the bootstrap proportion as a measure of *accuracy* or, in other words, as an estimate of the probability for a clade inferred by $\hat{t}$ to be in the true tree. Clearly, such a view requires the reconstruction method to be consistent, *i.e.*, to converge toward the true tree as more and more data are available. In fact, an inconsistent (but convergent) method associated to a strong phylogenetic signal (*e.g.*, Felsenstein 1978) iteratively finds the same erroneous clades for the various pseudosamples, and these clades have both a null probability of being correct and a high bootstrap proportion. The simulations of Hillis and Bull (1993) for parsimony showed that under realistic conditions where the method is mostly consistent, the high bootstrap proportions tend to underestimate the probability of clades being correct. This behavior was predicted by Zharkikh and Li (1992a,b) in the four taxa

case. Felsenstein and Kishino (1993) explained this phenomenon on the basis of a simplified model, while Efron *et al.* (1995) showed that it should not be seen as a general property of the bootstrap method, but as only true on average and in the rather favorable evolutionary conditions tested.

(3) Finally, the bootstrap proportion can be assimilated to the confidence level of a usual statistical hypothesis testing (Felsenstein and Kishino 1993; Zharkikh and Li 1995; Efron *et al.* 1995). It should be pointed out that such a view also requires the reconstruction method to be consistent. However, it may be distinguished from the previous view in that it is now a question of the conditional probability of a decision, *e.g.*, the observed clade is correct, under some null hypothesis, *e.g.*, the clade is incorrect, instead of simply considering the probability for this clade to be correct. The difficulty with this third approach is that the original bootstrap method, as employed by Felsenstein (1985), is not directly related to hypothesis testing in the usual statistical meaning (Efron *et al.* 1995). However, the latter showed that the bootstrap proportion is a first order approximation of the true confidence level. Moreover, they proposed a more elaborate bootstrap method for obtaining a second order approximation of the confidence level. On the other hand, Zharkikh and Li (1995) proposed an approach called the "complete-and-partial" bootstrap technique, based on a simplified analysis and on the estimate of the number of alternative resolutions for the clades. They evaluated their approach on a simple (5 taxa and molecular clock) case, showing satisfactory behavior of the proposed approximations.

Therefore, despite the variety of these points of view, it appears that discarding the clades of $\hat{t}$ with low bootstrap proportions is a well founded practice. It may be expected that such clades have a rather low chance of being correct, and also that most incorrect clades will be detected in the process. It follows that a gain (*i.e.*, a reduction in error) may be expected when estimating the true phylogeny, if instead of considering the complete original tree $\hat{t}$, we only consider its clades supported by sufficiently high bootstrap proportions. However some caution is required in order not to discard too many correct clades because of low bootstrap proportions. Our aim in this study is to quantify the extent of the mentioned error reduction under realistic evolutionary conditions. For this purpose, given a threshold S such that the clades with bootstrap proportions smaller than S are no longer considered, we evaluate:

- Type I error decrease, *i.e.*, the number of wrong clades of the original tree $\hat{t}$ which are discarded;
- Type II error increase, *i.e.*, the number of clades of the true tree t which are no longer considered but are present in the original tree $\hat{t}$;
- The total error reduction (or augmentation) according to some error measure combining Type I and Type II errors, the most classical being Robinson and Foulds's (1981) where both errors are equally weighted.
- The best value for the threshold S of clade selection given the chosen error function.

Methodological and technical aspects of this study are detailed in the next section, then results are discussed and followed by conclusion.

Material and Methods

Overview

Let $\bar{t}^*_S$ be the S *reduced bootstrap tree*, *i.e.*, the tree obtained by only retaining the clades of $\bar{t}^*$ (or equally of $\hat{t}$) supported by bootstrap proportions greater than or equal to S. To measure the error reduction obtained by using $\bar{t}^*_S$ rather than $\hat{t}$ to estimate the true phylogeny, we performed extensive computer simulations, along the lines of Kuhner and Felsenstein (1994). We first detail this simulation scheme. The extent of the error reduction depends on the error measure. Thus, in the next part, we present the error measure we adopted, namely the Robinson and Foulds (1981) distance to the true tree, or a generalization which enables different weights to be given to Type I and Type II errors. The extent of the error reduction also depends on the threshold used to select clades in the bootstrap tree, and the optimum threshold of clade selection depends on the weight given to both types of error. In the third part, we provide simple analytical approximations of this optimum threshold. Finally, we present a bias/variance decomposition of the error expectation, which makes it possible to analyze the sources of error from another point of view.

Scheme of simulations

Kuhner and Felsenstein (1994) reported quite a complete simulation study on the accuracy of various usual reconstruction methods, and we tried to follow their framework as much as possible. The basis of the simulations consists in setting down conditions of evolution (evolutionary rates and sequence lengths) and then randomly generating phylogenies and data sets under these conditions.

Figure 1 describes the way $\hat{t}$ and $\bar{t}^*_S$ were compared. For each condition of evolution, 50 phylogenies t were randomly generated and served each as a support for producing 100 original data sets or *samples*. Each data sample was analyzed by a tree-building method to obtain an *original tree estimate* $\hat{t}$, and bootstrapped to generate 100 *pseudosamples*. These pseudosamples were separately analyzed by the same reconstruction method, producing as many *bootstrap trees* $\hat{t}^*$. With 100 replicates the standard deviation of the bootstrap proportion, bp, reaches a maximum of 5% (for $bp \approx 50\%$). This was sufficient for our purposes and realistic from a computational point of view. Branches of the average bootstrap tree $\bar{t}^*$ having bp lower than a fixed threshold S were reduced, providing a *reduced bootstrap tree* $\bar{t}^*_S$. For a properly chosen threshold S, we compared $\hat{t}$ and $\bar{t}^*_S$ on the basis of their respective topological distance to the true tree, this distance being a combination of Type I and Type II errors. Bias and variance were also studied.

To generate phylogenies we used the RANTREE and DNATREE programs not included in the PHYLIP package but kindly provided by M.K. Kuhner and J.F. Felsenstein. These programs generate random topologies of ten-taxa by an iterative-branching, poissonian process that mimics to some extent biological evolution, giving a wide range of different trees. Each generated topology serves as a support for simulating the evolution of an ancestral DNA sequence using the two-parameter model of Kimura (1980) with a transition/transversion ratio of 2.0. This initial sequence contains molecular characters with equal proportions between the four states. The probability for a substitution event on a given branch is the product of the branch length expressed in time units and of a given substitution rate. Two different substitution rates (0.01 and 0.1) may be used, providing *slow* and *fast* branches.

The expectation of the number of substitutions per site along a lineage (from root to leaf) is between 0.0193 and 0.193, depending on the chosen substitution rate. Data sets are composed by taking sequences obtained at the leaves, for four different conditions:

1. low substitution rate for all branches;
2. high substitution rate for all branches;
3. low substitution rate for half the branches and high substitution rate for the other half;
4. low substitution rate for half the sites and high substitution rate for the other half.

We tested each condition with four sequence lengths (100, 300, 1000 and 3000 nucleotides). To ensure independence we generated new phylogenies for the different conditions of evolution.

Both a maximum parsimony (MP) and a distance (D) heuristic algorithm were used to reconstruct phylogenies. The maximum-likelihood was not included in the simulations due to the heavy computational expense it requires. For the distance method, we chose the NEIGHBOR program of PHYLIP related to the NJ method of Saitou and Nei (1987). It appears to perform well and its simplicity makes it very fast to execute. As Kuhner and Felsenstein (1994) did previously, we supplied it with corrected distances, using the correct transition/transversion ratio (2.0). The DNAPARS heuristic program from PHYILP was chosen for MP, since we observed almost no difference with an optimum algorithm during preliminary simulations (as similarly observed by Kuhner and Felsenstein, 1994). Moreover, this choice helped reduce the computational burden. Most phylogenetic programs were taken from the PHYLIP package version 3.5 (1993). All others are available on request. Finally, more than 16 million trees of ten species were inferred (1.01 million for each condition of evolution), which represents about a year of computation on a SUN SPARC 5 station.

Error measure

We now detail the Robinson and Foulds distance and its generalization, used to quantify the error of the estimated trees.

Each branch of a phylogeny partitions the species into two sets, according to the two subtrees resulting from the deletion of this branch. Thus, any phylogeny can be considered as a set of bipartitions, as induced by its branches. All trees we compare are phylogenies on the same set of species. They always have the same external branches, so that only *non-trivial* bipartitions, *i.e.*, induced by internal branches, are relevant. We will only consider these non-trivial bipartitions. For each estimated tree, $\hat{t}$ or $\bar{t}^*_S$, we measured the Type I and Type II errors. Type I error, denoted e_1, is the number of incorrect bipartitions of $\hat{t}$ (or $\bar{t}^*_S$), *i.e.*, we have

$$e_1(\hat{t}) = \Big|\{\, b \in \hat{t} / \, b \notin t \,\}\Big|$$

where b denotes any bipartition. Type II error, denoted e_2, is the number of bipartitions of the true tree not belonging to the estimated tree, *i.e.*, we have

$$e_2(\hat{t}) = \Big|\{\, b \in t / \, b \notin \hat{t} \,\}\Big|.$$

With ten species, a fully resolved tree contains 7 non-trivial bipartitions, so that e_1 and e_2 belong to $\{0, 1, \ldots, 7\}$. Note that if $\hat{t}$ is fully resolved, it includes the same number of branches as t, so that $e_1(\hat{t}) = e_2(\hat{t})$. Note also that $\bar{t}^*_S \subseteq \hat{t}$ (in terms of bipartition set), so that we always have $e_1(\hat{t}) \geq e_1(\bar{t}^*_S)$ and $e_2(\hat{t}) \leq e_2(\bar{t}^*_S)$.

The two error terms can be combined in several ways to express a distance between trees. The most classical is Robinson and Foulds's (1981), denoted e_{RF}, which gives equal weight to both types, *i.e.*, we have

$$e_{RF}(\hat{t}) = e_1(\hat{t}) + e_2(\hat{t}) = |t \oplus \hat{t}| \, .$$

$e_{RF}(\hat{t})$ ranges here in $\{0, \ldots, 14\}$.

A generalization consists in choosing a different weight for the two terms. In phylogenetic studies, e_1 is usually given more importance since it is preferable to forget correct bipartitions than to infer false ones. A solution to take this into account is to express the overall error as

$$\lambda \, e_1(\hat{t}) \; + \; e_2(\hat{t}) \qquad \text{with} \qquad \lambda \geq 1 \, ,$$

so that we can express the fact that Type I error is given λ times more importance than Type II error. However, to normalize the range of values given to this generalized error, denoted e^λ_{RF}, we use the expression

$$e^\lambda_{RF}(\hat{t}) = \frac{2}{\lambda + 1} \left(\lambda \, e_1(\hat{t}) + e_2(\hat{t}) \right) \, . \tag{1}$$

The standard Robinson and Foulds distance is then obtained for $\lambda = 1$, and the error is still comprised between 0 and 14. Note that if $\hat{t}$ is fully resolved, $e^\lambda_{RF}(\hat{t}) = e_{RF}(\hat{t})$. We consider λ values taken in the range $[1, 10]$. Higher values, with $\lambda = \infty$ as an extreme, would lead the Type II component to be negligible. The best estimate of t, i.e., which minimizes the overall error (1) would then be the "star" topology which induces only trivial bipartitions.

We were also interested in measuring the extent of the error reduction induced using the reduced bootstrap tree $\bar{t}^*_S$ rather than the original tree

$\hat{t}$. More precisely, for each condition of evolution, on the basis of $\bar{e}^{\lambda}_{RF}$, the average error observed over the 50×100 data samples, we examined the absolute error reduction $\bar{e}^{\lambda}_{RF}(\hat{t}) - \bar{e}^{\lambda}_{RF}(\bar{t}^{*}_{S})$, and the relative error reduction $(\bar{e}^{\lambda}_{RF}(\hat{t}) - \bar{e}^{\lambda}_{RF}(\bar{t}^{*}_{S}))/\bar{e}^{\lambda}_{RF}(\hat{t})$.

Choosing the threshold of clade selection

Given the generalized expression (1) of the Robinson and Foulds error and a chosen value for λ, the natural question that arises concerns the choice of the optimum threshold, denoted $S(\lambda)$, to reduce the average bootstrap tree $\bar{t}^{*}$, in order to minimize the expected error. For this purpose, consider the analogy with a simple two-sided game. A player bets repeatedly on the outcome of an event, having true and false as possible results with known respective probabilities p and $1-p$. The player commits a Type I error, e_1, whenever he chooses "true" and the event is false, and a Type II error, e_2, whenever he chooses "false" whereas the event is true. Suppose that the error function, e^{λ}, is given as $e^{\lambda} = \lambda e_1 + e_2$, *i.e.*, is the form of expression (1) without the normalization term, which is useless here. The question now is how to choose the best course open to the player as a function of p, *i.e.*, how to determine the optimum threshold $S(\lambda)$ so that the expected error, $E(e^{\lambda})$, will be minimized. For a fixed S, there are two possibilities:

case 1: if $p \geq S$ then "true" is decided so that the only possible error is of Type I, and committed with a probability $1-p$. Thus $E(e^{\lambda}) = \lambda(1-p)$;

case 2: if $p < S$ then "false" is always decided so that a Type II error is committed each time the event is true, that is with probability p. The expected error is then $E(e^{\lambda}) = p$.

Figure 2 details the expected error as a function of p, the two straight lines representing the case 1 and case 2 respectively. Let S be the threshold corresponding to the intersection of both lines, and S' be an arbitrary threshold. The bold part of each line represents the error to be expected if S' is used. We observe that the use of S instead of S' would lessen the expected error on the part $[S, S']$, and induce the same expected error elsewhere. So it clearly appears that the optimum threshold $S(\lambda)$ is situated at the intersection of both lines. It is analytically obtained by solving the equation $S(\lambda) = \lambda(1 - S(\lambda))$, which leads to $S(\lambda) = \frac{\lambda}{\lambda+1}$.

In phylogenetic studies, we are concerned with deciding which clades of the estimated tree $\hat{t}$ should be retained (case 1) and which discarded (case 2). Let us consider a particular clade, and p the probability for this clade to be true. We do not know p as in the previous simpler case but we can estimate it, in a certain way, by the bootstrap proportion, bp, of the clade. If we assume $p = bp$, the previous analysis leads to discarding the clades of $\hat{t}$ with $bp < \frac{\lambda}{\lambda+1}$. We then dispose of a first approximation of the optimum threshold of clade selection, applicable to bp

$$S_1(\lambda) = \frac{\lambda}{\lambda + 1}.$$

In fact, several authors (Zharkikh and Li 1992a,b; Felsenstein and Kishino 1993; Hillis and Bull 1993) have shown that when the reconstruction method is globally consistent, there is a correlation between p and bp but no equality. Thus $S_1(\lambda)$ is suspected to be only a crude approximation. Suppose that now we have an (increasing) function f, so that $f(bp)$ can be considered as an estimate of p. In this case the clades would be retained when satisfying

$$f(bp) \geq \frac{\lambda}{\lambda + 1} \quad \text{or equivalently} \quad bp \geq f^{-1}\left(\frac{\lambda}{\lambda + 1}\right),$$

and discarded otherwise. The previously mentioned authors provided simulations, and some analyses of simplified cases, which show that the general aspect of the function f is sinusoidal, as represented in Figure 3, so that we can choose f as being the "centered and reduced" cosine function:

$$f(x) = \frac{-\cos(\pi x) + 1}{2} .$$

With this in mind, we obtain a more sophisticated approximation for $S(\lambda)$, namely

$$S_2(\lambda) = \arccos\left(1 - \frac{2\lambda}{\lambda + 1}\right) / \pi .$$

Note that for the standard ($\lambda = 1$) Robinson and Foulds distance, a 50% threshold is indicated by both $S_1(\lambda)$ and $S_2(\lambda)$. This corresponds to the well known majority-rule. We then expect that $\bar{t}^*_{50\%}$ will be roughly the most accurate tree for this measure.

The alternative to these analytical approximations lies in simulations. For each method and any given $\lambda \in \{1, 1.5, 2, 2.5, 3, 3.5, 4, 5, 6, 7, 8, 9, 10\}$, we collected the 16 optimum thresholds $S^1(\lambda)$, ..., $S^{16}(\lambda)$ observed respectively for the 16 conditions of evolution studied. We then obtained the *empirical* threshold

$$S_e(\lambda) = \frac{1}{16} \sum_{i=1..16} S^i(\lambda) .$$

To assess the efficiency of these approximations ($S_1(\lambda)$, $S_2(\lambda)$ and $S_e(\lambda)$), we measured the *loss in quality* using the following formulae:

$$L_A(S_a) = max_{\substack{\lambda \in \{1,\dots,10\} \\ i \in \{1,2,\dots,16\}}} \{\bar{e}^{\lambda}_{RF}(\bar{t}^*_{S_a(\lambda)}) - \bar{e}^{\lambda}_{RF}(\bar{t}^*_{S^i(\lambda)})\} \tag{2}$$

and

$$L_R(S_a) = max_{\substack{\lambda \in \{1,\dots,10\} \\ i \in \{1,2,\dots,16\}}} \left\{\frac{\bar{e}^{\lambda}_{RF}(\bar{t}^*_{S_a(\lambda)}) - \bar{e}^{\lambda}_{RF}(\bar{t}^*_{S^i(\lambda)})}{\bar{e}^{\lambda}_{RF}(\bar{t}^*_{S^i(\lambda)})}\right\} \tag{3}$$

where $a \in \{1, 2, e\}$. In other words, we measured the maximum absolute (2) and relative (3) increases in error induced by using the approximation rather than the optimum threshold, over the various studied conditions of evolution and values for λ. The lower the error increase, the better the approximation.

Bias and variance

We also inquired about a possible bias for the various phylogenetic reconstruction methods used during the simulations. Method bias was first detected with the technique introduced by Kuhner and Felsenstein (1994). According to this measure, a tree-building method is said to be biased for a given true phylogeny if the cloud of inferred trees is not detected as centered on this phylogeny. This is determined by comparing $\bar{e}_{RF}(\hat{t})$, the average Robinson and Foulds distance between an estimated tree and the true tree, to $\bar{d}_{RF}(\hat{t}_i, \hat{t}_j)$, the average distance between two estimated trees, $\hat{t}_i$ and $\hat{t}_j$. If $\bar{e}_{RF}(\hat{t})$ is greater than $\bar{d}_{RF}(\hat{t}_i, \hat{t}_j)$, the cloud of inferred trees is judged not to be centered on the true tree, and bias is pronounced. We quantified the bias for a given method under any given evolutionary condition by counting the number of phylogenies for which it appeared biased (note that $\hat{t}$, $\hat{t}_i$ and $\hat{t}_j$ in previous expressions indicate original trees as well as reduced bootstrap trees). First results showed that reduced bootstrap trees are almost always biased according to this measure. The bias is simply explained (as will be detailed later) by the fact that these trees are not fully resolved. The Kuhner and Felsenstein (1994) measure does not make it possible to take this effect into account directly. We then investigated a bias measure which is more adapted to tree estimates.

In fact, in the statistical theory of estimation, the notion of bias is naturally tied to that of variance. For example, when considering the mean square error as loss function, a well known result is that the expected error of any estimate is the sum of a bias term and a variance term. The former measures the distance between the average value of the estimate and the true value to be estimated. The latter measures the dispersion of the estimate around this average value. An empirical rule is that the importance of both terms must be sufficiently balanced for the total expected error to be low. An important literature has been devoted to unbiased estimates (with minimum variance). But it is also well known (Kendall and Stuart 1973, p21-22) that better estimates (*i.e.*, with smaller expected error) may be obtained by admitting some bias. For example, consider a normal variate x having null expectation and unknown variance ($x \approx \mathcal{N}(0, \sigma)$). The best unbiased estimate of the variance σ^2 is given by the formula $\sum x_i^2/s$, where s is the sample size. But it may easily be shown that the best, slightly biased estimate is given by $\sum x_i^2/(s+2)$. This estimate outperforms the former because of a smaller variance (Kendall and Stuart 1973, p34).

When speaking of trees, the same phenomenon may be expected. Given a loss (or error) function, here the Robinson and Foulds distance or its generalization (1), we may prove a similar decomposition in bias and variance terms, as we shall see.

Let t be an unknown phylogeny inducing n non-trivial bipartitions on a given set of species. Consider as well a given reconstruction method and the disposition of data samples, with fixed length, obtained from t through any sampling process. Each time a new data set is analyzed, the method produces a new estimate $\hat{t}$, our aim being to characterize the distribution of $\hat{t}$. Let B represent the set of all possible non-trivial bipartitions. Viewing each tree as a set of bipartitions leads us to introduce the set of events attached to the bipartitions of B. The event $\{b \in \hat{t}\}$ is true when the corresponding bipartition is present in $\hat{t}$, and otherwise false. We may now define the weight, $W(b)$, of a bipartition b, as the probability of the event $\{b \in \hat{t}\}$, *i.e.*,

$$W(b) = Pr\{b \in \hat{t}\}\,.$$

Note that the weight distribution W is not a probability distribution over the set of events $\{b \in \hat{t}\}$ (this will be made clear below). Let $I_{\{b\in\hat{t}\}}$ be the indicator function attached to the event $\{b \in \hat{t}\}$, *i.e.*, $I_{\{b\in\hat{t}\}} = 1$ when the event is true and $I_{\{b\in\hat{t}\}} = 0$ otherwise. Using this notation, the number of bipartitions of a given tree $\hat{t}$ is simply the sum over B of the values of all these indicator functions. It follows that the expected size (number of bipartitions) of $\hat{t}$, $E(|\hat{t}|)$, is given by

$$E\left(|\hat{t}|\right) = E\left(\sum_{b\in B} I_{\{b\in\hat{t}\}}\right) = \sum_{b\in B} E(I_{\{b\in\hat{t}\}}) = \sum_{b\in B} W(b)\,, \tag{4}$$

i.e., is simply the sum of all weights $W(b)$. Moreover, this illustrates that W is not a probability distribution ($\sum W(b) \neq 1$). Using these weights, we may express the expectations of Type I and Type II errors:

- Type I error is the sum of the indicator functions of the events $\{b \in \hat{t} \wedge b \notin t\}$, so that the expectation of Type I error is

$$E\left(e_1(\hat{t})\right) = E\left(\sum_{b\in B-t} I_{\{b\in\hat{t}\}}\right) = \sum_{b\in B-t} W(b) \tag{5}$$

- Type II error is the sum of the indicator functions of the events $\{b \notin \hat{t} \wedge b \in t\}$. Expectation of Type II error, is then

$$E\left(e_2(\hat{t})\right) = E\left(\sum_{b\in t}\left(1 - I_{\{b\in\hat{t}\}}\right)\right) = n - \sum_{b\in t} W(b)\,. \tag{6}$$

Using the weight distribution W we can also define bias and variance terms as shown in the following.

- To define a variance term, we first need a notion of centrality, $\mathcal{M}$, which is defined as the sum of the n most important weights. In other words, $\mathcal{M}$ equals the total weight of the n most probable bipartitions and expresses, in this way, the center of the weight distribution. Now the variability, V, can be thought of as the tendency of the method to infer something different from this central point:

  $$V = \left(\sum_{b \in B} W(b) \right) - \mathcal{M}\,.$$

  A method that always proposes the same bipartitions, whatever the data set, has a null variance term V, since these bipartitions are accounted for in $\mathcal{M}$. The other extreme is a method that infers all bipartitions with an equal probability, inducing a uniform weight distribution. The method then has a maximum variance term, since the n weights accounted for in $\mathcal{M}$ are insignificant compared with the exponential number of weights of the other bipartitions.

- A bias term, A, expressing the *acentrality* of the true phylogeny t in the weight distribution of inferred bipartitions, can be defined in the following way

  $$A = \mathcal{M} - \left(\sum_{b \in t} W(b) \right)\,.$$

  If a tree inferred by the method is regularly close to t, the bipartitions of t are of significant weights and A is low. On the contrary, if the inferred tree rarely contains bipartitions of t, A is close to $\mathcal{M}$ and almost maximum. Clearly A corresponds to the general idea of bias , *i.e.* as expressed before, to the distance between the central value of the estimate and the value being estimated.

- The second bias term that we define is called *structural* (SB). This term expresses lack of adequation between the structure inferred and the topology being estimated, measuring the loss in the number of bipartitions. It is obtained by using the expected size of $\hat{t}$ (4), through the following expression

$$SB = n - E\left(|\hat{t}|\right) = n - \sum_{b \in B} W(b)\,.$$

  SB is null if all inferred trees have the same number n of bipartitions as t. It increases when less resolved trees are proposed, as is the case for reduced bootstrap trees.

Using these three bias and variance terms we derive a decomposition of the expected Robinson and Foulds error, in rewriting expressions (5) and (6) for Type I and Type II errors.

$$\begin{aligned}
E(e_1(\hat{t})) &= \sum_{b \in B-t} W(b) \\
&= \sum_{b \in B-t} W(b) - \left(\sum_{b \in B} W(b) - \mathcal{M}\right) + \left(\sum_{b \in B} W(b) - \mathcal{M}\right) \\
&= \left(\mathcal{M} - \sum_{b \in B} W(b) + \sum_{b \in B-t} W(b)\right) + \left(\sum_{b \in B} W(b) - \mathcal{M}\right) \\
&= \qquad\qquad A \qquad\qquad + \qquad\qquad V\,. \qquad (7)
\end{aligned}$$

and:

$$
\begin{aligned}
E(e_2(\hat{t})) &= n - \sum_{b \in t} W(b) \\
&= n - \sum_{b \in t} W(b) - \left(\sum_{b \in B} W(b) - \mathcal{M} \right) + \left(\sum_{b \in B} W(b) - \mathcal{M} \right) \\
&= \left(n - \sum_{b \in B} W(b) \right) + \left(\mathcal{M} - \sum_{b \in t} W(b) \right) + \left(\sum_{b \in B} W(b) - \mathcal{M} \right) \\
&= \quad SB \quad + \quad A \quad + \quad V\,. \qquad (8)
\end{aligned}
$$

From equations (1), (7) and (8), we obtain the expression of the expected generalized Robinson and Foulds error

$$
E(e^{\lambda}_{RF}(\hat{t})) = 2 \left(V + A + \frac{SB}{\lambda + 1} \right) . \qquad (9)
$$

For the standard ($\lambda = 1$) Robinson and Foulds distance, the expected error is thus $2V + 2A + SB$. Moreover, giving more importance to Type I error (*i.e.*, increasing λ) lowers the influence of the structural bias. This decomposition will enable us, in the following section, to further explain results observed with our simpler first measure of bias, taken from Kuhner and Felsenstein (1994). Note also that the usual tree-building methods such as MP and NJ are usually *convergent*, *i.e.*, tend to infer repeatedly the same (fully-resolved) tree as longer sequences are available. The variance gets close to zero as the number of sites grows, and the structural bias is null, so that the only (potentially) remaining term of expected error is acentrality. A consistent method converges toward the true tree and, therefore, it has asymptotically a null acentrality. Conversely, an inconsistent method has a significant acentrality, even for long sequences.

Results

We first report the simulation results concerning approximations of the optimum threshold of clade selection. We then focus on the error reduction obtained by using the properly reduced bootstrap tree rather than the original one. Next, we provide bias/variance results for the standard Robinson and Foulds error. Further remarks on the 95% threshold reduced tree and on the influence of other parameters conclude this section.

Optimum threshold of clade selection

We have previously provided two analytical approximations and the way to calculate an empirical approximation of the optimum threshold of clade selection. We evaluated the loss in quality for these various approximations through the measures L_A (2) and L_R (3), calculated on the basis of 1600 trees and 160.000 data samples. Table 1 details the results. For each method (D, resp. MP), one of the simple *a priori* approximations ($S_1(\lambda)$, resp. $S_2(\lambda)$) clearly outperforms the other and, more important, is almost as efficient as the empirical approximation which is obtained *a posteriori*. $L_A(x)$'s are error variations comprised in the range $[0, 14]$, so that observed values can be considered as low. For example, consider the loss in quality induced by $S_2(\lambda)$ with MP. This loss is of 0.20 clades (Table 1) and is obtained when $\lambda = 3.5$, the substitution rate is low and sequences have 300 sites. This means that under these evolutionary conditions, the bootstrap tree reduced by $S_2(\lambda)$ is 0.20 clades further from the true tree than if reduced by the optimum threshold $S^i(\lambda)$. More precisely, having $\lambda = 3.5$, $\bar{e}^{\lambda}_{RF}(\bar{t}^*_{S^i(\lambda)}) = 1.61$ clades and $\bar{e}^{\lambda}_{RF}(\bar{t}^*_{S_2(\lambda)}) = 1.81$ clades, so that even in this worst case, the trees $\bar{t}^*_{S^i(\lambda)}$ and $\bar{t}^*_{S_2(\lambda)}$ can be judged to be approximately the same distance from the true tree. Relatively low values were also observed in terms of the more severe relative measure: $L_R(S_1) = 8\%$ with D, and $L_R(S_2) = 13\%$ with MP.

To explain these good results we traced the empirical function linking the bootstrap proportion and the probability of the corresponding clades being correct (Figure 3). Similar work was previously done by Hillis and Bull (1993) for the parsimony case, on the basis of simulations for nine-taxa and four-taxa phylogenies. Figure 3 also displays the two functions f_1 and f_2, from which the two approximations $S_1(\lambda)$ and $S_2(\lambda)$ are obtained. The sinusoidal function is seen to be a rather good approximation of the empirical function for parsimony, while identity plays the same role for the distance approach (for $bp > 60\%$). The similar performance of $S_1(\lambda)$ and $S_e(\lambda)$ for the distance method observed previously, and of $S_2(\lambda)$ and $S_e(\lambda)$ for the parsimony one are thus partly explained.

The high quality of the approximations is also linked to the form of the curve representing the average error of the reduced bootstrap tree in relation to the threshold of clade selection. Figure 4 details this curve for the previously mentioned worst case (MP method, $\lambda = 3.5$, low rates, 300 sites), for which we have $S^i(\lambda) \approx 59\%$, $S_e(\lambda) \approx 65\%$, $S_2(\lambda) = 69\%$ and $S_1(\lambda) = 78\%$. $S_2(\lambda)$ appears to be more accurate than $S_1(\lambda)$, which is consistent with the previous observations (Table 1 and Figure 3). Moreover, the flatness of the curve around the optimum threshold $S^i(\lambda)$ explains why the resulting error is moderately sensitive to the exact choice of the threshold. A relatively large difference between thresholds only induces a small difference at the error level. From a mathematical standpoint, the loss in quality induced by the approximations is only of the second order relative to the gap between these approximations and $S^i(\lambda)$.

Extent of the error reduction

We now examine the extent of the error reduction induced by using the properly reduced bootstrap tree. We first detail results for generalized error, focusing on peculiar values of λ. Afterwards, results will only concern the standard Robinson and Foulds distance, for which a $\approx 50\%$ optimum threshold was shown.

Generalized error

Table 2 provides absolute and relative error reduction obtained by using $\bar{t}^*_{S(\lambda)}$ instead of $\hat{t}$. According to the previous results, $S_1(\lambda)$, resp. $S_2(\lambda)$, was used with D, resp. MP, to approximate the optimum threshold of clade selection. Both error reduction measures are detailed for the 16 conditions of evolution studied, and for several values of λ (1, 2 and 5). The extent of the error reduction depends greatly on the conditions of evolution. The bootstrap process appears to be more useful when short sequences are used. Besides, more significant error reductions are generally observed with MP than with D. Increasing λ also leads to more significant error reductions with both tree-building methods, in all cases. Since $e^{\lambda}_{RF}(\hat{t})$ is independent of λ, this last remark implies that $e^{\lambda}_{RF}(\bar{t}^*_{S(\lambda)})$ decreases as λ increases. In fact, for extremely high values of λ (and for a high number of bootstrap replicates), $\bar{t}^*_{S(\lambda)} \approx \bar{t}^*_{100}$ is expected to be the star topology. In this case, Type I error is null and the influence of Type II error is negligible, so that we expect $e^{\lambda}_{RF}(\bar{t}^*_{S(\lambda)}) \approx 0$. However, for more reasonable values of λ, two tendencies are opposed in each error component: Type I error decreases but its weight, $\frac{2\lambda}{\lambda+1}$, increases, while Type II error increases, but its weight, $\frac{2}{\lambda+1}$, decreases. We therefore could not predict *a priori* the results of Table 2, which seems to indicate that the error reduction obtained by using the properly reduced bootstrap tree increases with λ in a monotonous way.

Standard Robinson and Foulds distance

From now on we focus on the standard ($\lambda = 1$) Robinson and Foulds error. Tables $3 - 6$ quantify the error reduction induced by using the near-optimally reduced bootstrap tree, $\bar{t}^*_{50\%}$, rather than the original tree $\hat{t}$. Note that we are already assured of an error reduction since $\hat{t}$ can be viewed as equal to $\bar{t}^*_0$, a non-optimally reduced tree from previous results. These tables also detail the average error $\bar{e}_{RF}$ (decomposed into $\bar{e}_1$ and $\bar{e}_2$) of the estimated trees under the several conditions of evolution studied. Recall that values for $\bar{e}_{RF}$ range in $[0, 14]$, and values for $\bar{e}_1$ and $\bar{e}_2$ in $[0, 7]$. Note also that the star topology (*i.e.* the fully unresolved phylogeny) induces an error of 7 ($e_1 = 0$ but $e_2 = 7$). Average errors observed for the original trees are comparable to those of Felsenstein and Kuhner (1994), except perhaps in the case of unequal rates within sites where the convergence seems to be faster. For sequences of 100 nucleotides, $\bar{t}^*_{50\%}$ reduced bootstrap trees inferred with D and MP are closer to the true phylogeny than are the original trees of the various reconstruction algorithms which these authors studied. For sequences of 300 or more nucleotides, they are on average (over the different tested rates of evolution) as efficient as the original trees inferred by the maximum likelihood method.

In discarding clades from the original tree estimate $\hat{t}$, we expected that $\bar{t}^*_{50\%}$ would induce smaller Type I error and greater Type II error. The tables however show that these two tendencies do not have the same range, *i.e.*, Type I error is greatly reduced in $\bar{t}^*_{50\%}$, whereas Type II error is only slightly increased. As an extreme example, consider the case of low substitution rate where $\bar{e}_1(\hat{t}) = \bar{e}_1(\hat{t}) \approx \bar{e}_2(\bar{t}^*_{50\%})$, while $\bar{e}_1(\bar{t}^*_{50\%}) \approx 0$. On the whole, the relative error reduction obtained in discarding clades of $\hat{t}$ with $bp < 50\%$ ranges from a minimum of 3% (for D, high substitution rate and with long sequences) to a maximum of 39% (for MP, low substitution rate and with short sequences).

Bias / variance

Still considering the standard Robinson and Foulds distance, we now focus on bias/variance measures. Our results obtained for $\hat{t}$ using the measure of Kuhner and Felsenstein are similar to those of their study. In Table 7, we only provide the results for the case of unequal rates within branches, which is more realistic than the cases linked to the molecular clock hypothesis. These results illustrate general tendencies observed for $\hat{t}$ and $\bar{t}^*_{50\%}$ under other conditions. $\bar{t}^*_{50\%}$ always appears clearly more biased than $\hat{t}$, the gap being larger for shorter sequences.

When analyzing the results through the more adequate measures resulting from our bias/variance error decomposition, we can determine the real source of the previously observed bias for $\bar{t}^*_{50\%}$. As an example, in Table 8, we detail the average values observed for the bias and variance terms in the case of unequal substitution rates per branches, with 100 sites. Under this condition of evolution, reduced trees were most often detected as having bias in the sense of Kuhner and Felsenstein (1994). Weights $W(b)$ were approximated by their observed frequencies and the terms $\overline{A}$, $\overline{SB}$, $\overline{V}$ were averaged over the 50 phylogenies generated for this condition of evolution. This explains the coincidence of the $\bar{e}_{RF}$ values with those of Table 5. As expected, $\bar{t}^*_{50\%}$ presents much less variance than $\hat{t}$ with both MP and D tree-building methods. This results from the fact that $\bar{t}^*_{50\%}$ only contains clades with relatively high bp, which are likely to be inferred again from other data samples. On the other hand, $\bar{t}^*_{50\%}$ shows more structural bias since it usually has some unresolved parts (≈ 2 clades on average), whereas $\hat{t}$ is fully resolved. Low values are observed for acentrality, expressing absence of fundamental bias for the various estimates ($\bar{t}^*_{50\%}$ and $\hat{t}$). This is the general tendency observed for this term during the simulations, a maximum of 1.75 being observed for a phylogeny detected as clearly biased according to Kuhner and Felsenstein's measure. These good results for acentrality are explained by the fact that the reconstruction methods are globally consistent (and convergent) for the evolutionary conditions we examined. As previously stated, this fact implies that the acentrality term must be low. More critical conditions would likely lead to acentrality being a greater error component.

Further remarks

95% *threshold and statistical hypothesis testing*

Table 9 displays the Type I and Type II errors observed for the 95% reduced bootstrap tree. For the standard Robinson and Foulds error, the performance of $\bar{t}^*_{95\%}$ is expected to be poor in terms of distance to the true tree, since 95% is far from the optimum threshold ($\approx$ 50% in this case). Tables $(3-6, 9)$ show that it is effectively the case, and we have: $\bar{e}_{RF}(\bar{t}^*_{95\%}) >> \bar{e}_{RF}(\bar{t}^*_{50\%})$ and almost always $\bar{e}_{RF}(\bar{t}^*_{95\%}) > \bar{e}_{RF}(\hat{t})$. However, 95% is the usual threshold considered in statistical hypothesis testing. Following Felsenstein and Kishino (1993) and others, we may consider that the null hypothesis of the test is that the given clade of $\hat{t}$ is incorrect. In this case, the bootstrap proportion is thought to be approximatly equal to the confidence level. In other words, we hope that the probability of $\bar{t}^*_{95\%}$ committing a Type I error will be smaller but close to 5%. Since $\hat{t}$ contains 7 (non-trivial) bipartitions, this probability is estimated by $\bar{e}_1(\bar{t}^*_{95\%})/7$. In fact, it may be seen (Table 9) that this quantity is usually very close to 0%, and reaches a maximum value of 2% (in the case of unequal substitution rates per branches). This indicates that building a test directly based on bootstrap proportions is well founded (for the conditions of evolution studied here) because $\bar{e}_1/7 < 5\%$, but is surely very conservative, since $\bar{e}_1/7 << 5\%$. This leads to questioning the *power* of the test, *i.e.*, the probability of retaining a true clade, an estimate of which is $1 - \bar{e}_2(\bar{t}^*_{95\%})/7$. From Table 9, we observe that $\bar{e}_2$, and thus the power of the test, varies greatly depending on substitution rates and on sequence lengths, but not depending on the tree-building method. The power of the test based on the bootstrap proportion is observed to be very low for short sequences (3.3%-38% with 100 sites) but much more acceptable for long sequences (75%-93% with 3000 sites). These remarks underline the interest of recent studies (Efron *et al* 1995; Zharkikh and Li 1995) which attempt to provide better estimates of the confidence level, than the one obtained by the simple bootstrap proportion.

Influence of the reconstruction method

Tables $3-6$ show that the original estimate $\hat{t}$ has more or less the same performance when inferred with the parsimony or with the distance method. However, the error reduction obtained by using the reduced bootstrap tree $\bar{t}^*_{50\%}$ is always (except once) more significant with the former. The bootstrap process seems thus to have a different influence on the parsimony method than on the distance method. We observe that more clades are discarded with MP than with D (see differences in e_1 and e_2 from Tables $3-6$, and in SB from Table 8). Among other explanations, the choice of the 50% threshold seems to be the crux of this phenomenon. Indeed, from Figure 3 we observe that it is much more suitable for MP than for D, for which an approximately 55% threshold would have been more appropriate. This higher threshold discards more clades and would probably give as favorable results as those observed for MP with a 50% threshold.

Influence of the conditions of evolution

Increasing sequence length reduces the average error of all inferred trees under all tested conditions of evolution, so that the inference methods appear globally consistent for these conditions. It follows that the original trees are quite accurate estimates of the true tree for long sequences, leaving little room for improvements. Moreover, for long sequences, the bootstrap resampling leads to less variability between samples, so that the relative efficiency of $\bar{t}^*_{50\%}$ is lowered. As a consequence of these two effects, the gap between $\bar{t}^*_{50\%}$ and $\hat{t}$ for sequences of 3000 sites is reduced to the point where the error of both estimated trees is roughly similar. However, the two error components are equal for $\hat{t}$, while Type I error is still inferior to Type II error for $\bar{t}^*_{50\%}$.

The gap observed between $\hat{t}$ and $\bar{t}^*_{50\%}$ also varies depending on the substitution rates, but is usually lower when evolutionary conditions are more favorable for phylogenetic reconstruction. The most favorable condition studied is high uniform substitution rate, and $\hat{t}$ is in this case almost as efficient as $\bar{t}^*_{50\%}$. In contrast, the highest values of $\bar{e}_{RF}$ are obtained under low substitution rate, and it is precisely the condition under which the difference between $\hat{t}$ and $\bar{t}^*_{50\%}$ is the greatest (39% relative error reduction with MP and 33% with D for 100 sites). This condition requires long sequences for accurate estimations because of the small proportion of informative sites present in the sequences.

Conclusion

In the present study we have considered the problem of correctly interpreting the tree obtained by the bootstrap method (Felsenstein 1985). The main issue lies in the choice of a threshold of clade selection, applicable to bootstrap proportions, so that unreliable clades are discarded, and yet maximum information about the true tree is preserved. Our results can be summarized as follows: (1) we propose two analytical approximations of the optimum threshold to reduce the bootstrap tree in order to minimize its distance to the true tree. These thresholds depend on the exact distance chosen as error measure. Using extensive computer simulations, they were shown to be near-optimum, one for the parsimony method and one for the neighbor-joining method, leading only to slightly more error than that achieved by using the empirically determined optimum threshold. (2) We studied the reduction in distance to the true tree obtained by using the properly reduced bootstrap tree rather than the original tree. For the standard Robinson and Foulds distance, for which we found an optimum threshold of $\approx 50\%$, our simulations revealed that, for short sequences and low substitution rate, the relative error reduction is of 39% with the parsimony method, and of 33% with the distance one. In most cases, Type I error is greatly decreased while Type II error is only slightly increased. The error reduction is more significant when short sequences are used and when more importance is given to Type I error than to Type II error. (3) Decomposing the error expectation of the estimated trees into one term of variance and two of bias, we showed that the bootstrap process does not introduce any fundamental bias. The only source of bias is structural (lack of resolution). Moreover, variability is greatly decreased, providing another explanation for the better results of the reduced bootstrap tree compared with the original tree estimate.

This indicates that significantly positive results are achieved by properly reducing the original tree obtained by traditional tree-building methods. These latter have in common the defect of providing fully resolved trees with some extremely variable (unreliable) parts, although it is well known that information about each branch of the tree to be estimated is usually not equally present in the data. For this reason, it seems natural to investigate methods for reducing trees inferred by the traditional reconstruction methods. The bootstrap method (Felsenstein 1985) is demonstrated here to be very efficient for this purpose. However other methods, either statistical or combinatorial, could be designed to reduce inferred trees or to directly infer partially resolved trees. We believe this particular area deserves further research.

Acknowledgments

We are grateful to Mary K. Kuhner and Joe Felsenstein for providing their phylogeny generation programs and for discussing the software. We thank A. Zharkikh and Gilles Caraux for helpful comments on preliminary versions of this manuscript. We also thank all members of our laboratory for letting us use the machines during the very long computation time that was necessary. This research was supported in part by the $GREG$ and the IA^2 network.

LITERATURE CITED

EFRON, B. 1979. Bootstrap Methods: another look at the jackknife. Ann. Statist. **7**:1–26.

EFRON, B. and R. TIBSHIRANI. 1993. *An Introduction to the Bootstrap.* Chapman & Hall, 450p.

EFRON, B., E. HALLORAN and S. HOLMES. 1995. Bootstrap confidence levels for phylogenetic trees. ech. Rep. **179**, Standford University.

FELSENSTEIN, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Syst. Zool. **27**:401–410.

——————. 1985. Confidence Limites on Phylogenies: an approach using the bootstrap. Evolution **39**:783–791.

——————. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. distributed by the author. Department of genetics, University of Washington, Seattle.

FELSENSTEIN, J. and H. KISHINO. 1993. Is there something wrong with the bootstrap on phylogenies? a reply to Hillis and Bull. Syst. Biol. **42**:193–200.

HILLIS, D. M. and J. J. BULL. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. **42**:182–192.

KENDALL, M.G. and A. STUART, 1973. The advanced theory of statistics. Vol 2, chap. 17, Charles Griffin (ed.), London.

Kimura, M. 1980. A simple method for estimating evolutionary rates base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

Kuhner, M. K. and J. Felsenstein. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. **11**:459–468.

Robinson, D. F. and L. R. Foulds. 1981. Comparison of phylogenetic trees. Math. Biosci. **53**:131–147.

Saitou, N. and M. Nei. 1987. The neighbor-joining method: a new method for reconstruction phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

Sanderson, M.J. 1989. Confidence limits on phylogenies: the bootstrap revisited. Cladistics **5**:113–129.

———— . 1995. Objections to bootstrapping phylogenies: a critique. Syst. Biol. **44**:299–320.

Zharkikh, A. and W. Li. 1992a. Statistical properties of bootstrap estimation of phylogenies variability from nucleotide sequences. I. four taxa with a molecular clock. Mol. Biol. Evol. **9**:1119–1147.

———— . 1992b. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences: II. four taxa without a molecular clock. J. Mol. Evol. **35**:356–366.

———— . 1995. Estimation of confidence in phylogeny: the complete-and-partial bootstrap technique. Mol. Phylogenet. Evol. **4**:44–63.

Table 1

Losses in quality associated with the various approximations of the optimum threshold of clade selection

	L_A		L_R	
threshold	MP	D	MP	D
$S_1(\lambda)$	0.46	**0.16**	24%	**8%**
$S_2(\lambda)$	**0.20**	0.20	**13%**	18%
$S_e(\lambda)$	0.18	0.10	13%	6%

NOTE. - Maximum absolute (L_A) and relative (L_R) losses in quality observed for the two analytical ($S_1(\lambda)$ and $S_2(\lambda)$) and the empirical ($S_e(\lambda)$) approximations of the optimum threshold of clade selection. This maximum is obtained over the 16 conditions of evolution and for $\lambda \in \{1, 1.5, 2, 2.5, 3, 3.5, 4, 5, 6, 7, 8, 9, 10\}$. L_A is measured in number of clades and ranges in $[0, 14]$. Results are detailed separately for the parsimony (MP) and the distance (D) tree-building methods.

Table 2

Absolute and relative error reduction of the reduced bootstrap tree $\bar{\mathbf{t}}^*_{\mathbf{S}(\lambda)}$ over the original tree estimate $\hat{\mathbf{t}}$

A. Parsimony method

		$\lambda = 1$		$\lambda = 2$		$\lambda = 5$	
rates	sites	rel	abs	rel	abs	rel	abs
low	100	39%	3.26	56%	4.74	75%	6.29
	300	33%	1.49	51%	2.37	69%	3.19
	1000	24%	0.51	45%	0.99	65%	1.42
	3000	20%	0.14	39%	0.31	59%	0.47
high	100	18%	0.51	36%	1.02	61%	1.71
	300	9%	0.11	25%	0.33	51%	0.68
	1000	4%	0.03	21%	0.16	48%	0.36
	3000	4%	0.01	21%	0.06	46%	0.13
high-low per branches	100	27%	1.45	46%	2.49	69%	3.75
	300	21%	0.76	40%	1.43	64%	2.31
	1000	14%	0.27	31%	0.59	55%	1.04
	3000	4%	0.03	14%	0.15	31%	0.34
high-low per sites	100	24%	1.07	42%	1.90	66%	2.97
	300	16%	0.32	33%	0.64	58%	1.13
	1000	8%	0.06	23%	0.18	49%	0.39
	3000	6%	0.01	17%	0.05	42%	0.13

Table 2

B. Distance method

rates	sites	$\lambda = 1$ rel	$\lambda = 1$ abs	$\lambda = 2$ rel	$\lambda = 2$ abs	$\lambda = 5$ rel	$\lambda = 5$ abs
low	100	33%	2.63	49%	4.00	72%	5.92
	300	24%	1.17	38%	1.93	64%	3.21
	1000	11%	0.15	22%	0.36	51%	0.87
	3000	5%	0.03	19%	0.15	49%	0.40
high	100	9%	0.24	25%	0.67	53%	1.41
	300	6%	0.09	22%	0.36	49%	0.79
	1000	3%	0.02	18%	0.10	46%	0.27
	3000	3%	0.01	23%	0.09	54%	0.20
high-low per branches	100	17%	0.86	37%	1.93	65%	3.37
	300	9%	0.28	29%	0.87	59%	1.77
	1000	9%	0.20	28%	0.60	58%	1.22
	3000	7%	0.08	25%	0.28	53%	0.61
high-low per sites	100	14%	0.59	32%	1.37	61%	2.57
	300	6%	0.13	23%	0.47	51%	1.05
	1000	3%	0.03	21%	0.20	49%	0.48
	3000	3%	0.02	23%	0.14	55%	0.34

NOTE. - Relative (rel) and absolute (abs) reduction in distance to the true tree obtained by using the properly reduced bootstrap tree rather than the original tree estimate. λ is the importance attributed to Type I error in relation to that given to Type II error. The optimum threshold of clade selection was approximated by $S_1(\lambda)$ for D and $S_2(\lambda)$ for MP.

Table 3

Accuracy of estimated trees for the standard Robinson and Foulds distance under low (0.01) substitution rate

A. Parsimony method

	Error reduction			$\hat{t}$		$\bar{t}^*_{50\%}$		
#Sites	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1 = \bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$
100	**39%**	47%	-8%	**8.42**	4.21	**5.16**	0.29	4.87
300	**33%**	45%	-12%	**4.66**	2.33	**3.13**	0.23	2.90
1000	**24%**	39%	-15%	**2.11**	1.06	**1.61**	0.23	1.37
3000	**20%**	39%	-19%	**0.79**	0.40	**0.64**	0.09	0.55

B. Distance method

	Error reduction			$\hat{t}$		$\bar{t}^*_{50\%}$		
#Sites	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1 = \bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$
100	**33%**	37%	-4%	**8.22**	4.11	**5.48**	1.03	4.44
300	**24%**	31%	-7%	**5.02**	2.51	**3.84**	0.96	2.88
1000	**11%**	18%	-7%	**1.69**	0.85	**1.52**	0.55	0.97
3000	**5%**	16%	-11%	**0.79**	0.40	**0.76**	0.27	0.49

NOTE. - Table 3 – 6 give the average standard ($\lambda = 1$) Robinson and Foulds error, $\bar{e}_{RF}$, induced by the original tree, $\hat{t}$, and by the reduced bootstrap tree, $\bar{t}^*_{50\%}$. This error is decomposed into Type I error (incorrect clades) and Type II error (omitted correct clades), denoted $\bar{e}_1$ and $\bar{e}_2$. The relative reduction in error obtained by using $\bar{t}^*_{50\%}$ rather than $\hat{t}$, is also detailed, together with its distribution between Type I and Type II error components.

Table 4

Accuracy of estimated trees for the standard Robinson and Foulds distance under high (0.1) substitution rate

A. Parsimony method

	Error reduction			$\hat{t}$		$\bar{t}^*_{50\%}$		
#Sites	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1 = \bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$
100	**18%**	34%	-16%	**2.82**	1.41	**2.32**	0.46	1.85
300	**9%**	26%	-17%	**1.32**	0.66	**1.20**	0.32	0.88
1000	**4%**	16%	-12%	**0.75**	0.38	**0.72**	0.26	0.47
3000	**4%**	14%	-10%	**0.28**	0.14	**0.27**	0.10	0.17

B. Distance method

	Error reduction			$\hat{t}$		$\bar{t}^*_{50\%}$		
#Sites	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1 = \bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$
100	**9%**	22%	-13%	**2.69**	1.35	**2.44**	0.75	1.70
300	**6%**	18%	-12%	**1.53**	0.76	**1.44**	0.48	0.95
1000	**3%**	15%	-12%	**0.58**	0.29	**0.56**	0.20	0.36
3000	**3%**	12%	-9%	**0.32**	0.16	**0.31**	0.12	0.19

NOTE. - See Note to Table 3.

Table 5

Accuracy of estimated trees for the standard Robinson and Foulds distance when substitution rates vary among branches

A. Parsimony method

	Error reduction			$\hat{t}$		$\bar{t}^*_{50\%}$		
#Sites	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1 = \bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$
100	**27%**	38%	-11%	**5.38**	2.69	**3.92**	0.63	3.29
300	**21%**	36%	-15%	**3.59**	1.80	**2.82**	0.49	2.33
1000	**14%**	27%	-13%	**1.91**	0.95	**1.64**	0.44	1.20
3000	**4%**	10%	-6%	**1.07**	0.54	**1.04**	0.43	0.60

B. Distance method

	Error reduction			$\hat{t}$		$\bar{t}^*_{50\%}$		
#Sites	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1 = \bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$
100	**17%**	26%	-9%	**5.21**	2.60	**4.29**	1.24	3.05
300	**10%**	20%	-10%	**3.04**	1.52	**2.74**	0.91	1.83
1000	**9%**	20%	-11%	**2.11**	1.06	**1.91**	0.62	1.29
3000	**7%**	18%	-11%	**1.13**	0.56	**1.04**	0.36	0.68

NOTE. - See Note to Table 3.

Table 6

Accuracy of estimated trees for the standard Robinson and Foulds distance when substitution rates vary among sites

A. Parsimony method

	Error reduction			$\hat{t}$		$\bar{t}^*_{50\%}$		
#Sites	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1 = \bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$
100	**24%**	38%	-14%	**4.51**	2.25	**3.43**	0.53	2.89
300	**16%**	31%	-15%	**1.94**	0.97	**1.64**	0.37	1.27
1000	**8%**	20%	-12%	**0.75**	0.38	**0.69**	0.22	0.47
3000	**6%**	19%	-13%	**0.32**	0.16	**0.29**	0.10	0.20

D. Distance method

	Error reduction			$\hat{t}$		$\bar{t}^*_{50\%}$		
#Sites	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1 = \bar{e}_2$	$\bar{e}_{RF}$	$\bar{e}_1$	$\bar{e}_2$
100	**14%**	26%	-12%	**4.19**	2.10	**3.62**	1.03	2.59
300	**7%**	20%	-13%	**2.04**	1.02	**1.91**	0.61	1.29
1000	**3%**	13%	-10%	**0.97**	0.49	**0.94**	0.36	0.58
3000	**3%**	11%	-8%	**0.62**	0.31	**0.60**	0.24	0.36

NOTE. - See Note to Table 3.

Table 7

Kuhner and Felsenstein (1994) bias measure

	Parsimony		Distance	
#Sites	$\hat{t}$	$\bar{t}^*_{50\%}$	$\hat{t}$	$\bar{t}^*_{50\%}$
100	0	41	1	20
300	3	22	0	6
1000	5	15	2	6
3000	9	15	1	3

NOTE. - Number of true phylogenies out of 50 for which the estimated trees ($\hat{t}$ and $\bar{t}^*_{50\%}$) appear to be biased. Distances between trees are measured with the standard Robinson and Foulds distance. Displayed values are obtained for the case of unequal substitution rates per branches.

Table 8

Bias/variance decomposition of the standard Robinson and Foulds average error

	Parsimony		Distance	
	$\hat{t}$	$\bar{t}^*_{50\%}$	$\hat{t}$	$\bar{t}^*_{50\%}$
$2\overline{V}$	5.20	1.06	5.04	2.28
$\overline{SB}$	0	2.66	0	1.81
$2\overline{A}$	0.18	0.20	0.17	0.20
$\overline{e}_{RF}$	5.38	3.92	5.21	4.29

NOTE. - Standard Robinson and Foulds average error for sequences of 100 sites and unequal substitution rates per branches. Error is decomposed for the original tree ($\hat{t}$), and for the properly reduced bootstrap tree ($\bar{t}^*_{50\%}$), into a variance term ($2\overline{V}$), a structural bias term ($\overline{SB}$) and an acentrality term ($2\overline{A}$).

Table 9

Type I and Type II errors for $\bar{\mathbf{t}}^*_{\mathbf{95\%}}$

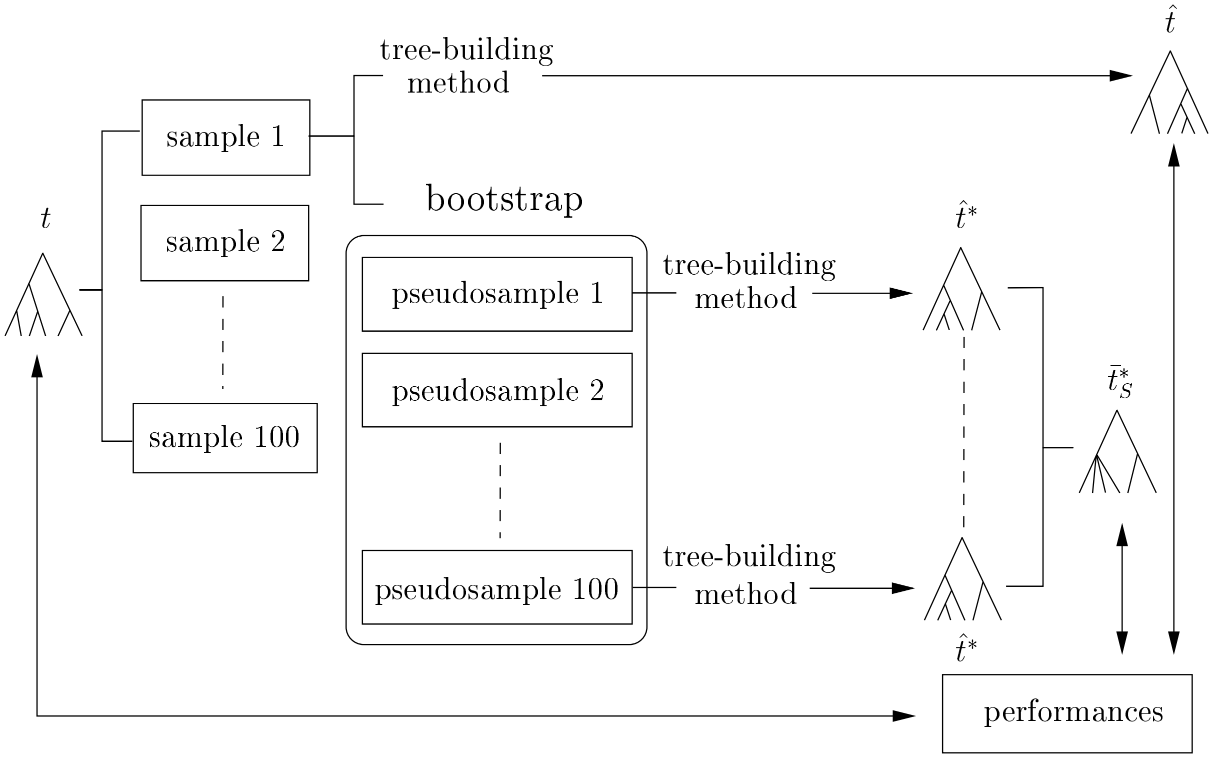
		Parsimony		Distance	
rates	sites	$\bar{e}_1$	$\bar{e}_2$	$\bar{e}_1$	$\bar{e}_2$
low	100	0.00	6.77	0.06	6.71
	300	0.00	5.76	0.05	5.66
	1000	0.00	3.21	0.02	2.90
	3000	0.00	1.53	0.00	1.52
high	100	0.00	4.48	0.05	4.34
	300	0.00	2.50	0.01	2.83
	1000	0.01	1.29	0.00	1.17
	3000	0.00	0.58	0.00	0.51
high-low per branches	100	0.00	5.26	0.04	5.35
	300	0.00	4.40	0.03	3.88
	1000	0.03	2.73	0.02	2.86
	3000	0.14	1.43	0.01	1.77
high-low per sites	100	0.00	5.66	0.02	5.48
	300	0.00	3.40	0.01	3.64
	1000	0.00	1.64	0.01	1.72
	3000	0.00	0.74	0.00	0.92

Fig. 1.- Scheme of simulations. We compared the original tree estimate $\hat{t}$ to the reduced bootstrap tree $\bar{t}^*_S$ (for a chosen threshold S of clade selection) with respect to their ability to recover the true phylogeny t.

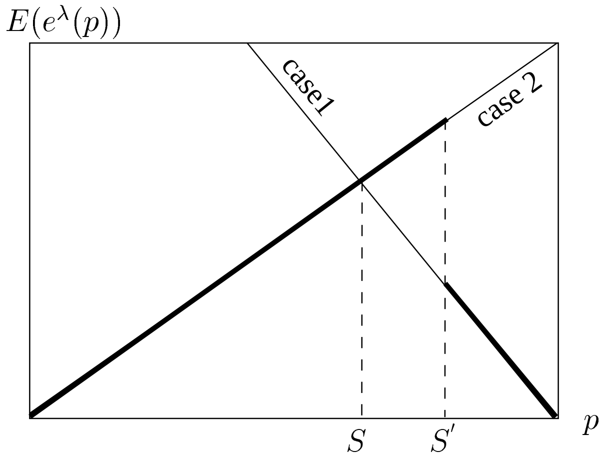
Fig. 2.- Expected error $E(e^\lambda(p))$ in a game where a player bets repeatedly on the outcome of an event (true, false) having known probability p. The player uses a threshold S: if $p \geq S$ (case 1) he always decides "true", and otherwise (case 2) "false". The expected error in case 1 is given by the line $E(e^\lambda(p)) = \lambda(1-p)$, and in case 2 by the line $E(e^\lambda(p)) = p$. The optimum threshold is the abscissa, S, of the intersection of both lines.

Fig. 3.- Probability (p) for a clade to be true depending on its bootstrap proportion (bp). f_e^D and f_e^{MP} are the empirical functions for D and MP respectively, while $f_1(x) = x$ and $f_2(x) = \frac{-cos(\pi x)+1}{2}$ are estimates of these functions. The approximations $S_1(\lambda)$, $S_2(\lambda)$ and $S_e(\lambda)$ of the optimum threshold of clade selection are obtained by applying the inverse of these functions to $\frac{\lambda}{\lambda+1}$.

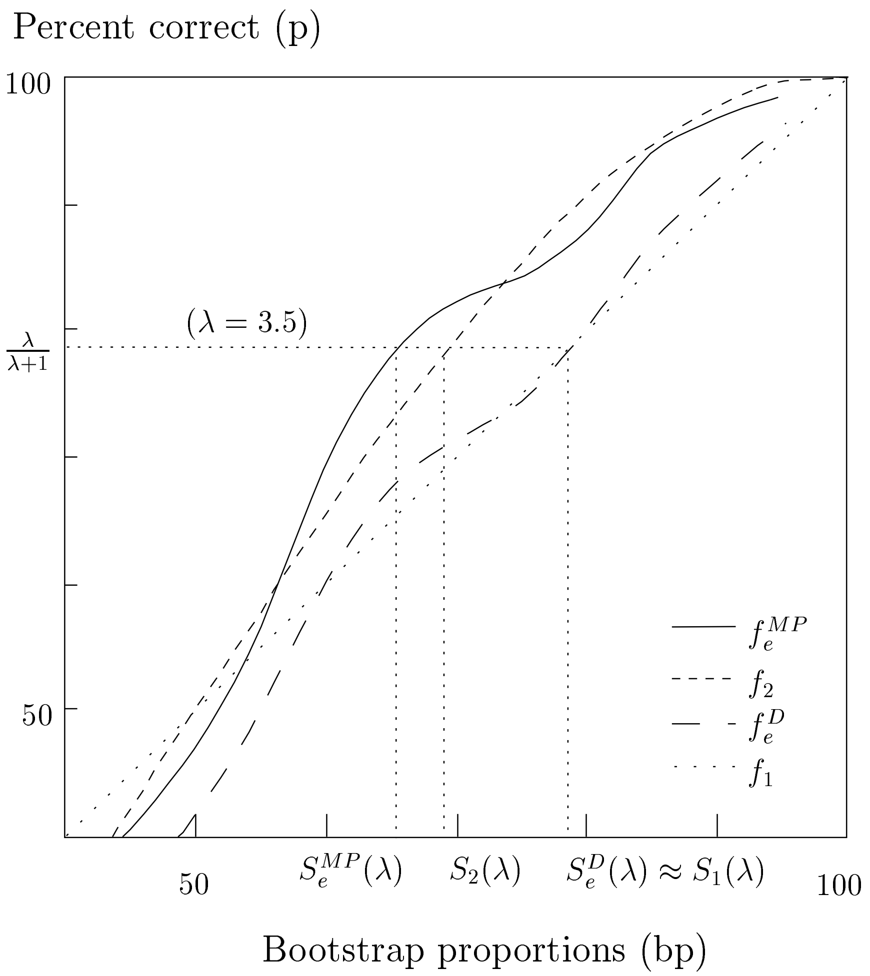
Fig. 4.- Average error ($\bar{e}^\lambda_{RF}$) of the reduced bootstrap tree $\bar{t}^*_S$, represented as a function of the threshold S of clade selection ($\lambda = 3.5$, low substitution rates, 300 nucleotides, MP). $S^i(\lambda)$ is the optimum threshold. $S_2(\lambda)$ is the recommended approximation for MP. $S_e(\lambda)$ is the empirical threshold. $S_1(\lambda)$ is the simplest possible approximation. δ is the observed absolute loss in quality obtained with $S_2(\lambda)$.



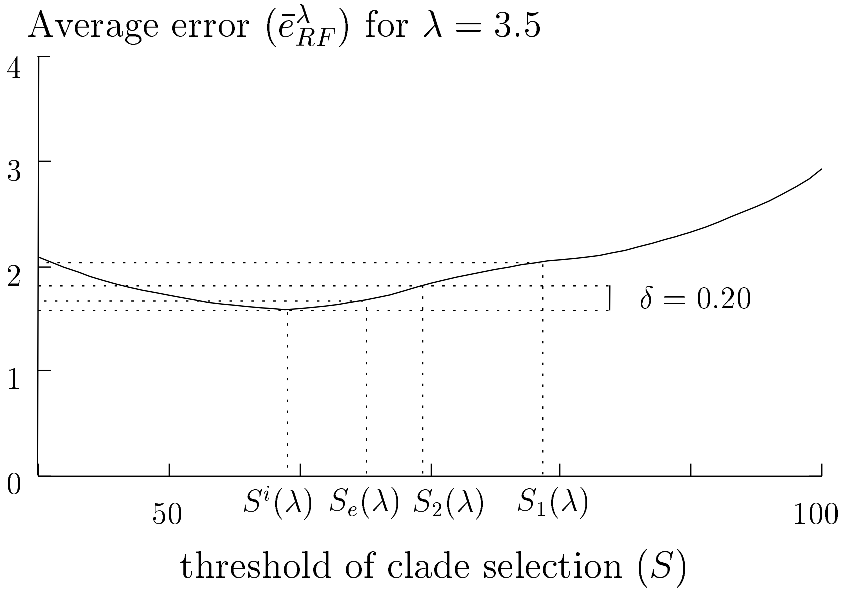
Berry and Gascuel - Figure 1



Berry and Gascuel - Figure 2



Berry and Gascuel - Figure 3



Berry and Gascuel - Figure 4