Statistical Science
2003, Vol. 18, No. 2, 241–255

Bootstrapping Phylogenetic Trees: Theory and Methods

Susan Holmes

Abstract. This is a survey of the use of the bootstrap in the area of systematic and evolutionary biology. I present the current usage by biologists of the bootstrap as a tool both for making inferences and for evaluating robustness, and propose a framework for thinking about these problems in terms of mathematical statistics.

Key words and phrases: Bootstrap, phylogenetic trees, confidence regions, nonpositive curvature.

1. AN INTRODUCTION TO SYSTEMATICS

The objects of study in systematics are binary rooted semilabeled trees that link species or families by their coancestral relationships. For example, Figure 1 shows a tree with seven strains of HIV.

Two leaf vertices (taxa) that share a parent in the tree are supposed to be descended from the same ancestor. The ancestors, and indeed the whole tree, have to be inferred in the absence of relevant fossil data. Today, the data used to build the trees are aligned DNA or protein sequences. These are usually represented as matrices of letters, where the rows are labeled by species and the columns represent positions in the genomic sequence; many of the letters in a column are the same (see Table 1).

Trees have been used in court cases and environmental surveillance. Some examples of current applications include:

- The whale watch team builds phylogenetic trees from their own databases to classify whale meat they sample from Japanese fish markets (Baker and Palumbi, 1994; Baker, Lento, Cipriano and Palumbi, 2000). Over the last eight years they have found, among others species, blue whale, humpback, minke whales, beaked whales and dolphins. They report regularly to the International Whaling Commission (Lento et al., 1998).
- Immunologists, microbiologists and epidemiologists are interested in the origins of strains of HIV, tuberculosis, influenza and other fast evolving bacteria and viruses. The data in Table 1 came from a public database of HIV sequences available at LANL (2002).

Phylogenetic trees are the main object of publication in many biology journals such as *Systematic Biology*, *Molecular Biology and Evolution*, *Molecular Phylogenetics and Evolution*, *Trends in Research in Ecology and Evolution*, *Journal of Molecular Evolution*, *Evolution*, *Molecular Ecology* and *American Journal of Botany*.

Although several methods for tree estimation (or inferring trees) are currently available, very little inferential theory is available for quantifying uncertainty for these trees. The most widely used tool for inference is a version of the bootstrap introduced by Felsenstein (1983). Figure 1 shows bootstrap values along the edges of the tree. These are generally required for publication of a tree estimate in the same way clinical trials require publication of p values.

The confidence statements made about such trees are my main focus. Biologists have also begun to adopt Bayesian methods based on Markov chain Monte Carlo computations that use parametric evolutionary models (Li, Pearl and Doss, 2000; Mau, Newton and Larget, 1999; Yang and Rannala, 1997), but I will not discuss these methods in depth here. A recent issue of *Systematic Biology* (Volume 51, Number 5, October 2002) had many interesting articles on parametric Bayesian approaches.

The tree in Figure 1 shows bootstrap values at the inner nodes; for example, 93 means that the species `CONS-CPZ` and `CONS O4` were siblings in 93% of

Susan Holmes is Associate Professor, Department of Statistics, Stanford University, Stanford, California 94305, detached from the Biometry Department of INRA, Montpellier, France (e-mail: susan@stat.stanford.edu).

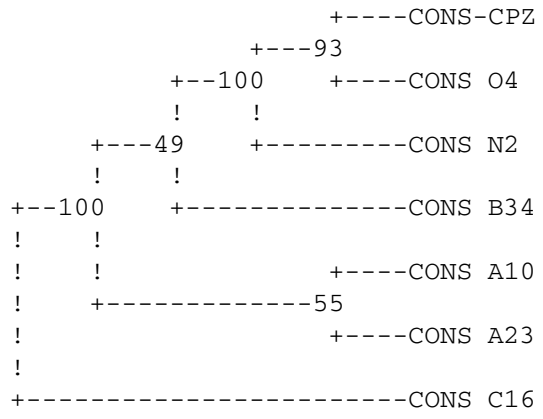

FIG. 1. *Tree with bootstrap values.*

the bootstrap replications; 49 means that the sequences CONS B34, CONS N2, CONS-CPZ and CONS O4 were grouped together in what is called a monophyletic (a group containing the most common ancestor of a given set of taxa and all the descendents of that most recent common ancestor) clade in 49% of the bootstrap replications. The method of bootstrapping is the multinomial nonparametric bootstrap as applied in the binomial setting. For each bootstrap simulation step, a new data matrix is accumulated by choosing columns from the original data matrix at random with replacement, and this is repeated until there are as many columns in the new matrix as were in the original data.

In Section 2, I explain the data and the estimation problem from a statistical viewpoint. I analyze some of the ways in which biologists make use of bootstrap values in Section 3. Section 4 shows how some statistical concepts, such as sufficiency and correction terms, can be used in this problem. Section 5 rephrases the problem in geometrical terms and Section 6 gives some constructive suggestions on how bootstrapping can be used for trees. It also summarizes some of the caveats presented in this article.

Some of the challenges involved in giving a statistical justification for the basic bootstrapping procedure are reviewed first.

TABLE 1
Partial HIV amino acid sequences from LANL (2002)

1	CONS A10	KKEEEEALLT GADTVVVLEE INLGGKKKPK
2	CONS A23	RREEEEALLT GADTVVVLEE INLGGKKKPK
3	CONS B34	KKEEEEALLT GADTVVVLEE MNLGGRKKPK
4	CONS C16	KKEEEEALLT GADTVVVLEE INLGGKKKPK
5	CONS N2	RREEEEALLT GADTVVVIEE ?QLGGKKKPK
6	CONS O4	CCEEEEVLLT GADTVVVLNN IQLGGKTTPK
7	CONS-CPZ	??EEEEALLT GADTVVVIDD IQLGG?RRPK

2. FROM MOLECULAR DATA TO PHYLOGENETIC TREE

Suppose an observed data set made on s species and n characters is given: the characters are DNA or amino acids. The sequences, one for each species, are aligned before this stage of analysis. For simplicity's sake, the aligned sequences are considered to be of equal length, thus providing a matrix block, for which each column is often called a character. Table 1 is a little example taken from LANL (2002). The Markovian evolutionary model (Li, 1997) is a low-dimensional (one to six parameters) model for such data. The metric evolutionary tree (binary rooted tree with edge lengths) can also be considered a parameter in this model. If the metric tree is supposed known and the mutation rates are known, data can be generated by choosing a character from the stationary distribution of the transition matrix (suppose A were drawn). The ancestral column at the root would then be all A's, but each row would have a different path to follow through the tree and changes would occur with probabilities proportional to the edge lengths. This simulation procedure is available through the SEQGEN program (Rambaut and Grassly, 1997).

From the observed data, the maximum likelihood method attempts to estimate the tree by choosing the tree with the highest probability of occurring, given the data, either using prior estimates of the relevant mutation rates or using the data to estimate these parameters. Maximum parsimony is a nonparametric method that uses a different criterion: it ignores all evolutionary models and searches for the tree with the least number of mutations along its branches needed to explain the data. Distance-based methods are semiparametric methods that use the evolutionary model to estimate distances between sequences and then use methodology akin to hierarchical clustering to build the tree.

I will not go into the details of how a tree is estimated from these data: book-long treatments exist (Felsenstein, 2003; Page and Holmes, 2000) and a survey for statisticians was presented by Holmes (1999). The parameter estimated from the data is a binary rooted tree with the labels (taxa: species or populations) at the leaves, with edge lengths (this is called the metric tree). If the method does not provide edge lengths, by default we set all the edge lengths to 1. Even ignoring edge lengths and considering only the branching order of the tree, the size of the space is huge. There are an exponentially large [there are $(2n-3)\times(2n-5)\times\cdots\times 3\times 1=(2n-3)!!$ (Schröder,

1870) different ones] number of combinatorial tree forms. I denote the metric tree estimate by $\hat{\tau}$, the true tree by τ and the space of metric trees $\mathcal{T}$, sometimes with an index n to denote the number of leaves.

After deciding which estimator to use, a natural followup question is, "How variable is the estimate?" Making a confidence statement about the parameter itself poses numerous quandaries where the classical paradigms, both Bayesian and frequentist, suffer from a lack of theory about these general non-Euclidean parameters. Possible questions are the following:

- What is a sampling distribution in $\mathcal{T}$?
- What is the natural notion of variability in $\mathcal{T}$?

These questions would be solved naturally if it were known how to represent a probability distribution $\mathbb{P}$ and distance d in $\mathcal{T}$. If the sampling distribution $\mathbb{P}_n$ of $\hat{\tau}$ were known, the bias and variance could be written $E_{\mathbb{P}_n} d(\hat{\tau}, \tau)$, $E_{\mathbb{P}_n} d^2(\hat{\tau}, \tau)$. At this stage, a Bayesian could look at samples from a posterior distribution on trees. In a frequentist approach, the bootstrap usually comes in and provides an estimate for bias as $E_{\mathbb{P}_n} d(\hat{\tau}^*, \hat{\tau})$ and an estimate of variance as $E_{\mathbb{P}_n} d(\hat{\tau}^*, \hat{\tau})^2$.

Because this theory is lacking, biologists have simplified their questions: the simplest one concerns the presence/absence of a certain monophyletic group c, called a clade. For instance, in Figure 1 the hypothesis to test is whether the clade $c =$ `(CONS A10, CONS A23)` exists in the true tree τ.

Felsenstein (1983) first introduced the use of the nonparametric bootstrap to assess what biologists call *repeatability*: the probability that another such sample shares the clade with the original sample. In statistical terms, this is denoted by $\mathbb{P}_n(c \in \hat{\tau} | c \in \hat{\tau}_0)$. Other biologists hoped that the bootstrap would provide estimates of what they called *accuracy* (Hillis and Bull, 1993), that is, $\mathbb{P}(c \in \tau | c \in \hat{\tau}_0)$. The two quantities are linked since the tighter the sampling distribution around τ, the more probable it is that the same clade appears again in a second sample, and the more probable it is that the clade from the estimated tree is a true clade. However, as in all statistical hypothesis testing, only $P(\text{data} | H_0)$ is available; an estimate of $P(H_0 | \text{data})$ requires further information.

2.1 Nonidentically Distributed, Nonindependent Columns

A first statistical reservation to the vanilla multinomial nonparametric bootstrap of the columns of an aligned set of DNA or protein sequences is the assumption of independent, identically distributed columns. It is well known that these assumptions are violated; Fitch (1971a, b), for example, showed the existence of regions that covaried. Some regions are highly conserved (most of the columns are identical all the way down) and some columns are highly variable, so the columns are not identically distributed either. These departures from the i.i.d. situation have been modeled in many ways. Various suggestions for more believable models have been included in different procedures:

- Hidden Markov model for rate variation along the sequences as discussed by Yang (1994) and Felsenstein and Churchill (1996) (see Durbin, Eddy, Krogh and Mitchison, 1998, for a review).
- A Gamma distribution model of rate variation along sequences was suggested by Yang (1994).
- Fitch and Markowitz (1970), Lockhart et al. (1996) and Tuffley and Steel (1998) have built *covarion* models to deal with sites that vary together.
- Changes in rate variation can be detected and modeled using hot spot or change point detection as, for instance, in Tang and Lewontin (1999).

The block bootstrap (Künsch, 1989) as explained clearly by Efron and Tibshirani (1993, pages 98–102) can provide a nonparametric equivalent for the multinomial under dependence of the data. Parametric bootstrapping using models such as those from Felsenstein and Churchill (1996) is also available in SEQGEN (Rambaut and Grassly, 1997). In practice there seems to be a strong preference for nonparametric resampling. This may be due to the misguided intuition that using a model that injects more variability gives safer, more conservative answers.

2.2 Consequences of Dependency

Unfortunately, dependency assumptions destroy the classical validity of the bootstrap. There are no clear theoretical bounds on the difference in bootstrap results between the i.i.d. and dependent cases in this context.

Generally speaking, when the columns are dependent, there are fewer effectively independent components (using the block bootstrap makes this clearer). A smaller sample means that the estimates are less accurate, and the bootstrap is no exception. In classical statistical problems such as estimating the mean, the nonparametric bootstrap itself has errors on the order of $1/n$. If the *effective* n is much smaller than the number of columns, the actual error may be very large, and the estimation error also may be exceed-

ingly large as pointed out, for example, in Nei, Kumar and Takahashi (1998). Thus the asymptotic consistency results are even less relevant here than in a simple univariate problem. The parameter is of very high dimension and in the realm of Freedman and Peters (1984a, b), where the size of the parameter space is large compared to the amount of data.

3. WHAT IS THE BOOTSTRAP SUPPOSED TO TELL US?

If a statistician is presented with a problem where the columns are observations to be resampled, the question that arises immediately is which distribution are the columns being sampled from? Certainly no simple random sampling is in effect, as was clearly pointed out by Sanderson (1995), so the classical paradigm by which the bootstrap is justified as proposing an approximation to the sampling distribution of the estimate is not in order here. However, since this method is so widely used by biologists, it is worth asking which questions it is *actually* answering, to better understand why they consider it an essential feature of phylogenetic methodology.

3.1 Stability or Reliability

Here are some quotes from the systematics literature:

> Boostrapping measures how consistently the data support given taxon bipartitions (Hedges, 1992).
>
> This is not a test of how accurate your tree is; it only gives information about the stability of the tree topology (the branching order), and it helps assess whether the sequence data is adequate to validate the topology (Berry and Gascuel, 1996).
>
> Bootstrap values are a measure of support of a given edge like the measure introduced by Bremer (1988) that asks how many more steps a parsimony tree must be for a given edge to disappear.
>
> High bootstrap values (close to 100%) mean uniform support, i.e., if the bootstrap value for a certain clade is close to 100%, nearly all of the characters informative for this group agree that it is a group (Berry and Gascuel, 1996).

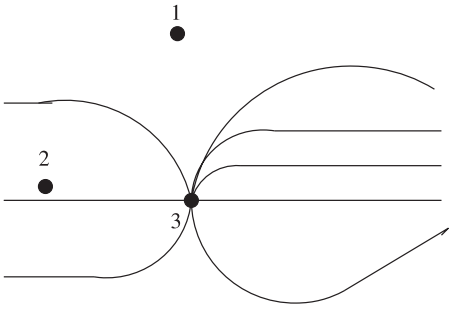

FIG. 2. *A partition of tree space.*

> The bootstrap test which is a crude way of testing interior branches, is applicable to all tree-building methods and is easy to use. Although this test has been shown to be conservative under certain theoretical frameworks, a conservative test is preferable in real data analysis, because the evolution of actual DNA (or protein) sequences never follows any mathematical model available (Nei, Kumar and Takahashi, 1998).

In fact, if a more attentive look at the actual way the bootstrap is used is taken, it is seen much more as a measure of robustness of the estimator with regard to small changes in the data. Could a small plausible perturbation of the data give a different result? The best way to think about this question is graphically: partition the parameter space of all possible binary trees into regions, each region corresponding to a different tree topology (I abbreviate "different tree topologies" to "different trees").

In Section 5 it is demonstrated that this can be justified mathematically. For the time being, Figures 2 and 3 are a sufficient schematization. The partitioning

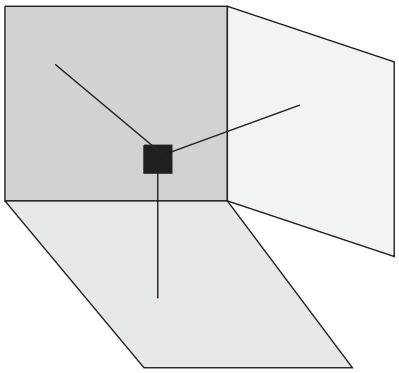

FIG. 3. *Data (the black rectangle) being projected on three possible trees.*

```
  4.60
Taxon 1   AATAATCACACAAGTATATTGTTCTTTAAACCTTGCAAAGAACCCAATATCTACTTCTGA
Taxon 2   GGCAATTATGTAAGTATATTGTTATTTAAGCACTGCAGTGAACCCCGTCTCTACAGCTGA
Taxon 3   GGTGGCCTCGTAAGTACCTTGTTCTGTAGACATTGCAGATAACCCCGTATGTACATCTCA
Taxon 4   AACGATCACGTAAGTGTACTGTTCTTCGAACATAGAGGAGAAGACCGTATCTCCATCGGG
```

FIG. 4.

represented by Figure 2 differs from Efron, Halloran and Holmes (1996, Figure 3) in that we are not considering this to be a partition of the data space, but of the parameter space onto which the data are considered to be projected.

If the data are quite far from being treelike, the data may be hesitating between several different trees. Figure 3 shows such a situation schematically. Making a small perturbation of the data could give very different trees if the estimation function near that data is not continuous, as pictured in the example shown in Figure 3. Suppose that the data are at the black rectangle, and the estimator $\hat{\tau}$ projects the data onto the closest point on one of the flaps. Then the estimate could oscillate between the tree flaps if the data are only slightly perturbed, thus giving a discontinuous estimator. (This can only be rigorously corrected for by making the estimate a mixture of trees; for instance, in this case $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ for the weights associated to the three components.)

Here is an example of such a data set. The original DNA data are shown in Figure 4. The output from PHYLIP (see the Appendix) with the parsimony criterion gives the tree described in Figure 5. Bootstrap of the Example 2 data gives, with the DNAPARS parsimony program, the results presented in Figure 6.

```
One most parsimonious tree found:

  +--------Taxon 4
  !
--3     +--Taxon 3
  !  +--2
  +--1  +--Taxon 2
     !
     +-----Taxon 1

requires a total of 43.000
```

FIG. 5.

If we make one change of nucleotide on the last letter of the second column from an A to a G, then the PHYLIP output becomes as shown in Figure 7.

This second data set is an obvious case of the mixture of three trees, as described above, and the bootstrap detects it. Bootstrap of the second data set (with the changed nucleotide) yields the results presented in Figure 8.

The bootstrap does detect the mixture, not quite with the right proportions, but indicates the presence of alternative choices which get lost when just one tree is chosen.

This discontinuous behavior is also present when maximum likelihood estimation is used, since the log likelihoods can be very close, even though the trees are

```
Sets included in the consensus tree
(* means on one side, · on the other)

Set                    How many times
(species in order)      (out of 1000)
4321
·***                        1000.0
·**·                         572.66

Sets NOT included in consensus tree:
Sets                   How many times
(species in order)      (out of 1000)
4321
··**                        215.16
·*·*                        212.16
```

```
CONSENSUS TREE:
The numbers at the forks indicate
the number of times the group con-
sisting of the species which are
to the right of that fork occurred
among the trees, out of 1000 trees

 +--------------Taxon 4
 !
 !    +---------Taxon 1
 +1000.0
      !    +----Taxon 2
      +572.7
           +----Taxon 3
```

FIG. 6.

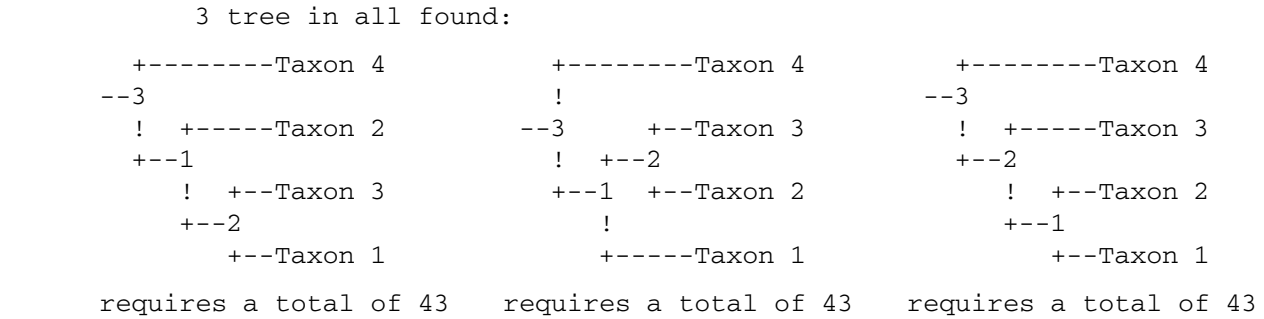

FIG. 7.

not close in tree space. Consequently, the likelihood contours are also discontinuous. This is responsible for the problem of *islands* such as those cited in Maddison (1991).

The quality that biologists call stability or reliability is what statisticians call *robustness*. The question addressed here is, "Do the perturbed data all project into the same region?" In more precise terms, it would also be comforting to know which columns produce these discontinuities. PHYLIP (see the Appendix) has such a tool in DNACOMP that uses a compatibility estimation to give an indication as to whether each column agrees with the final tree or not. This is at least a first possible level of analysis of residuals or a search for points with high leverage as in regression. The output in Table 2 shows that only the third and fourth columns disagree with the final tree.

As can be seen in Figure 2, some of the boundaries between tree regions are curved. In addition, points such as point 3 border more than two regions. For both of these reasons the simple bootstrap can be applied as a perturbation tool to assess the stability (in the sense of continuity, a small perturbation in the data that produces only a small perturbation in the estimate) of the estimator. We can consider that we are exploring the neighborhood of the estimator by a simulation experiment as suggested by Efron and Tibshirani (1998). We can also try to describe the regions and their neighborhoods mathematically; this requires some extra work, as we will see in Section 5. To quantify robustness, a notion of distance in parameter space is necessary; this will also be provided in Section 5. The more precise study of influence functions (Huber, 1996; Hampel, Ronchetti, Rousseeuw and Stahel, 1986) is not available here because a notion of derivative in $\mathcal{T}$ is unavailable.

TABLE 2
Output from DNACOMP *for the above data*

```
   01234567890123456789012345678901234567890
   *----------
 0 ! YYNNYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY
40 !YYYYYYYYYYYYYYYYYYYY
```

```
Sets included in the consensus tree
(* means on one side, · on the other)

Set                   How many times
(species in order)     (out of 1000)
4321
·***                       1000.0
·**·                        336.83

Sets NOT included in consensus tree:
Set                   How many times
(species in order)     (out of 1000)
4321
··**                       336.33
·*·*                       326.83
```

```
CONSENSUS TREE:
The numbers at the forks indicate
the number of times the group con-
sisting of the species which are
to the right of that fork occurred
among the trees, out of 1000 trees

 +--------------Taxon 4
 !
 !          +----Taxon 1
 !    +336.8
 +1000.0    +----Taxon 2
      !
      +---------Taxon 3
```

FIG. 8.

3.2 Statistical Insight for the Simplest Case: One Clade

As pointed out by Efron, Halloran and Holmes (1996), the bootstrap compares the bootstrapped values to the original estimate, not to the truth. The essential bootstrap identity is that the bootstrapped values are to the original estimate what the original estimate is to the true parameter.

Some parameter has to be chosen to evaluate the sampling distribution; often in the case of real-valued parameters this may be the variance of the estimate. If this variance is small, then we can conclude that the parameter has a high probability of being close to the original estimator, so a notion of variability would be very useful.

From a statistical point of view the simplest possible context in which to study the bootrapping for trees is in the statistical test of the presence of a certain monophyletic group c. Suppose that we have decided ahead of time that we want to approximate the probability that an observed clade c is in fact in the true tree τ. A binary parameter θ_c can be defined that takes on the value 1 if the clade c is present in the true tree and 0 if not. Then the question posed by Hillis and Bull (1993) about the accuracy of the method $[P(\theta_c = 1|\hat{\theta}_c = 1)]$ can be answered through the parametric bootstrap by generating data with many different known trees τ and using the SEQGEN program of Rambaut and Grassly (1997) to generate data sets with τ as the generating process. Then compute for each the bootstrap samples $P(\hat{\theta}_c = 1|\hat{\theta}^*_c = 1)$. We can thus compare the accuracies and the estimated bootstrap estimates: what happens depends on τ, its inner and pendant edge lengths (pendant edges are the edges of degree 1), its topology and the estimation method used.

Even though we are only interested in whether or not a clade is present, we have to consider the number of possible alternative trees, with and without the clade, because this influences the bootstrap's validity (Zharkikh and Li, 1995). In fact, it is as if we were trying to estimate only one component p_j of a multinomial parameter $(p_1, p_2, \ldots, p_{2(n-3)!!})$; we still have to know how many p_i's are big enough to compete with p_j.

Rodrigo (1993), Efron, Halloran and Holmes (1996) and Zharkikh and Li (1995) suggested postprocessing the data to recalibrate by taking into account the number of neighbors and the curvature of the boundaries. Li and Zharkikh (1995) coined the phrase "number of effective neighbors" and called this number K. This accounts both for the number of neighboring regions and for the probability assigned to the regions. There are some trees (in particular balanced trees when estimating with parsimony) that are combinatorially neighbors, but that have such light densities they can be neglected. Zharkikh and Li (1995) proposed to compute this number by simulation in a procedure called the *complete and partial bootstrap*. Efron, Halloran and Holmes (1996) "looked out" on the regions neighboring the boundary of the estimated tree by first searching for the boundaries using the bootstrap samples that have trees different from the estimated one and then using a binary search to find data that are as close as possible to the boundary. The borderline data are then used to generate more bootstrap resamples that will be as different as possible and provide an empirical profile of the neighborhood. This is then used to recalibrate the original bootstrap values.

Hillis and Bull (1993) and Zharkikh and Li (1995) reported that the bootstrap estimates of repeatability were biased. Efron, Halloran and Holmes (1996), Zharkikh and Li (1995) and Newton (1996) clarified some of the reasons for this apparent bias, due to several facts that we will revisit:

- Directly comparing the truth to the bootstrap samples.
- Not accounting for the other edge lengths in the tree that change the number of possible neighbors and thus the baseline.
- Not accounting for the curvature of the boundary between the regions that define different trees.

4. HOW CAN WE EFFECTIVELY SUMMARIZE BOOTSTRAP DATA?

Most biologists agree that the simplest possible probability distribution on tree space—the uniform distribution—is not relevant; a slightly more realistic one is the Yule process (see Aldous, 2001). Building a probability distribution on trees is a complex procedure. Further, choosing optimal trees in a model cannot, in general, be decomposed into simpler problems. This is the essence of what constitutes computationally intractable problems. The estimation of both the maximum likelihood tree and the parsimony tree have been proven to be intractable.

Now let us come back to actually trying to summarize a bootstrap resampling distribution on trees—a Bayesian posterior distribution or a distribution that could be used to build a frequentist confidence region. Classically, sufficient statistics arise from a model. We can also go the other way, following the statistical

mechanics paradigm of deciding which features are relevant for a tree, call these $S_1(\tau), S_2(\tau), \ldots, S_k(\tau)$, and then forming the exponential family $Z^{-1} \cdot \exp(\sum \theta_i S_i(\tau))$ based on these, where Z is the partition constant that makes this a probability measure. See Lauritzen (1988) for background.

For this to give an effective model, the distribution must be well characterized by a few summaries, the sufficient statistics. One example is the exponential family model $P(\tau) = L\exp(-\lambda d(\tau, \tau_0))$ defined by Billera, Holmes and Vogtmann (2001) in analogy to Mallows' model. The sufficient statistic is

$$\sum_{i=1}^{k} d(\tau_i, \tau_0) = S_k$$

for a collection of k trees and a central tree τ_0. So the estimation depends only on the distances between trees d. This reduces the data to one number S_k. Of course, without the assumption of the symmetrical distribution, sufficient statistics are much more complex.

Note that the bootstrap distribution is often summarized by just the frequencies of edges or clades as in Figure 1. These numbers do not constitute sufficient statistics for the complete bootstrap distribution. This has been implicitly understood by several authors who work on trees. For instance, Penny and Hendy (personal communication) have proposed a nearest neighbor (NN) bootstrap which counts how many times an edge *or* a *neighboring* split occurs (for an example of its use, see Cooper and Penny, 1997). We see later that a suitable geometrical enhancement of the mathematical picture of tree space makes such a NN bootstrap natural.

4.1 Multiple Testing

Showing all the bootstrap values on the tree simultaneously as in Figure 1 makes for an easy misinterpretation. Users may believe that these values can be used together. If they are considered as ersatz p values, this carries the flavor of multiple testing without correction and so should be avoided. When more than one edge is of interest, a multidimensional approach involving a certain amount of nonstandard geometry is preferable, as we see in Section 5.

4.2 Clade Frequencies as a First Level Approximation

Considering just the clade frequencies as a first order approximation can be justified by considering a decomposition of a set of trees $\mathcal{X}$ by a Fourier type analysis in tree space. (This Fourier analysis is not the same as that proposed by Hendy and Penny, 1993 and Hendy, Penny and Steel, 1994.) Bootstrap clade frequencies just count the binomial counts of presence/absence of a given clade in a set B of trees obtained, for instance, by bootstrap simulation. The set of these trees can be considered as a function from the set of all trees into the integers. To each tree we associate its frequency of appearance in the set. Diaconis and Holmes (1998, 2002) showed that the space of all combinatorial trees on n leaves $\mathcal{T}_n$ can be represented as the quotient of the symmetric group on $2(n-1)$ by the subgroup $B_{2(n-1)}$ that leaves the pairs

$$\{(1,2),(3,4),(5,6),\ldots,(2n-3,2n-2)\} \quad \text{invariant.}$$

This is called the matching representation of trees, where the tree is replaced by all its sibling pairs, including the inner nodes.

Suppose we are trying to describe a set of 1000 bootstrap trees. This is a function from tree space to $\mathbb{R}$, where each tree is associated to the number of times it occurs among the 1000 trees (we allow a fractional number of trees to be counted, because sometimes the output from the bootstrap functions can be fractional). The decomposition of functions on tree space was given by Diaconis and Holmes (1998). It is a direct sum decomposition of all functions on tree space,

$$\mathcal{L}(\mathcal{T}_n) = \bigoplus_{\lambda \vdash (n-1)} S^{2\lambda},$$

where the sum is over all partitions λ of $n-1$ (i.e., all vectors of integers that sum to $n-1$), $2\lambda = (2\lambda_1, 2\lambda_2, \ldots, 2\lambda_k)$ and $\mathcal{S}^{2\lambda}$ is the associated irreducible representation of the symmetric group $\mathfrak{S}_{2(n-1)}$. The first few terms in the decomposition can be interpreted as follows:

- For $\lambda = (n-1)$, $S^{2\lambda}$ counts the number of trees in the data set.
- For $\lambda = (n-2, 1)$, $S^{2\lambda}$ counts the number of times each particular sibling occurs.
- For $\lambda = (n-3, 1, 1)$, $S^{2\lambda}$ counts the number of times the sibling pair (i, j) occurs at the same time as the pairs (k, l).
- For $\lambda = (n-3, 2)$, $S^{2\lambda}$ counts the number of times the sibling pair (i, j, k, l) occurs as a clade.

It is in this sense that the sibling pair frequencies are the first order approximation to the complete distribution on trees. It would be useful to be able to say what proportion of the information contained in the data set can be reconstructed just by the sibling pair counts. This Fourier type decomposition closely follows the

analysis of ranking data provided by Diaconis (1989). To the other extreme is the idea that we could keep all the data and make useful multidimensional summaries of it.

5. GEOMETRICAL REPRESENTATION

Trees are high-dimensional parameters, even if they do not lie naturally in a Euclidean space, so the best frequentist confidence statements that can be made about them rely on the notion of a confidence region $\mathcal{R}_\alpha$ defined by statements of the form

$$\mathbb{P}(\tau \in \mathcal{R}_\alpha) = 1 - \alpha$$

(Tukey, 1975). Green (1981) suggested the use of successively peeled convex hulls as ersatz confidence regions. Take, for instance, the convex hull of all the parameter estimates in a high-dimensional space. Then omit the points on the boundary and construct the next hull, which may contain 90% of the data, for instance, so that it is a nonparametric 90% convex envelope. This envelope can then be used to test whether a certain tree is in the envelope or not. To give a more rigorous discussion of these continuity–sensitivity issues, it would be necessary to define both relevant probability measures and satisfactory metrics on $\mathcal{T}_n$.

Distances can also be useful for summarizing the bootstrap sampling distribution. In this section, we fill in the set of combinatorial trees, which is a discrete set, making the space of trees a continuous space with a meaningful distance that is always defined. This work was presented in its mathematical technicality by Billera, Holmes and Vogtmann (2001).

We start by explaining intuitively what we would like our geometrical representation to provide. Our goal is to give a rigorous definition of the boundaries depicted in Figure 2. Every tree exists as a point in its own region. The boundaries between regions represent an area of uncertainty about the exact branching order. In biological terminology, this is called an unresolved tree. Two neighboring regions represent neighboring trees. The notion of neighboring is nearest neighbor interchange, the rotation distance in $\mathcal{T}_n$. This defines two trees as neighbors if you can get from one to the other by contracting an edge to have length zero and then reexpanding this vertex of degree 4 so that it has degree 3 again. In Figure 9, we see one such change. This seems to be the most widely accepted neighborhood relationship in biology, although other distances could also be used to define a metric tree space in the same way. The natural way to vary

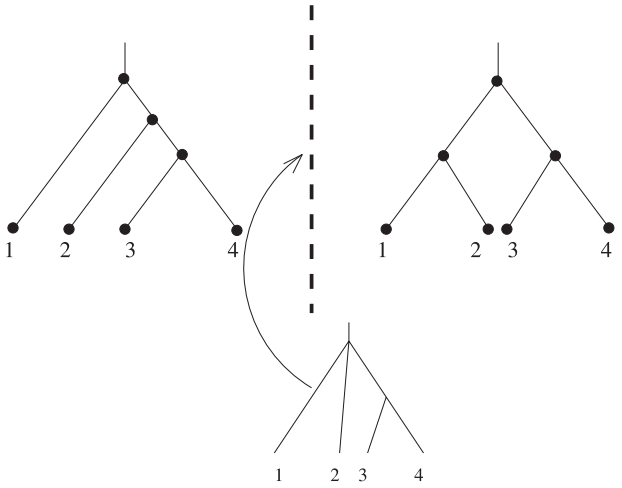

FIG. 9. *Two tree regions separated by a boundary—a degenerate tree.*

closeness to the boundary or unresolved tree is to make the edge lengths e decrease linearly in the direction of the boundary.

There are three rooted binary semilabeled trees on three leaves. We arrange them along three half lines that meet at the origin which represents the star tree. For this article, to make our geometrical space slightly simpler, we restrict ourselves to trees with finite branch lengths. By standardizing the combinational trees to all have edge lengths of 1, we can build the space of trees with n leaves as a cube complex $\mathcal{T}_n$, where the cubes are all of dimension $n-2$. For rooted binary trees with four leaves, we have a set of squares pasted together by two edges each. Each square in Figure 10 corresponds

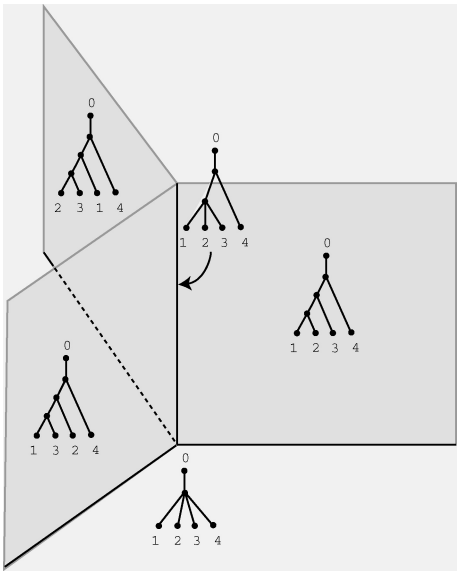

FIG. 10. *Three neighboring quadrants.*

to a different branching order and the position within the square is determined by the coordinates, each representing one of the two inner edge lengths.

Note that the boundary is shared by two other trees. The pendant branch lengths do not appear in this geometrical representation. (To obtain a complete coordinate system of binary semilabeled trees, one would have to take the product of $\mathcal{T}_n$ with $\mathbb{R}^n$.) All quadrants have to have the star tree as one of their corners, so that particular point will have 15 neighboring quadrants. This generalizes and explains why, at the star tree, the origin of our space, there are exponentially many cubes attached. On the other hand, a degenerate tree with only one nonzero edge is represented as a point on the segment boundary to three quadrants; thus its neighborhood contains three "flaps."

In fact, if we have a four leafed tree, but are sure what the outgroup is, the relevant space is the space of rooted trees on three leaves. This embedding, shown geometrically in Figure 11, is important if we consider the problem of finding all the trees with a given edge as in Section 2. It is the cube complex $[0, 1] \times \mathcal{T}_{n-1}$ embedded in $\mathcal{T}_n$. This is important to consider when we talk about the boundary region between the trees that have the edge c and those that do not.

As we saw in Section 3, Zharkikh and Li (1995) did a simulation study to find how many trees neighbor a given tree. This has consequences for the quality of the bootstrap estimate as also was pointed out in Efron, Halloran and Holmes (1996). The geometrical picture allows us to just count the number of neighbors. We can see that for a tree on four leaves, there can be either no neighbors except trees with the same

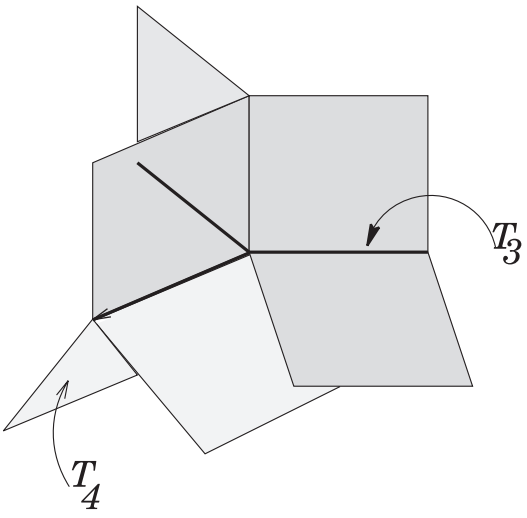

FIG. 11. *Embedding of $\mathcal{T}_3$ in $\mathcal{T}_4$.*

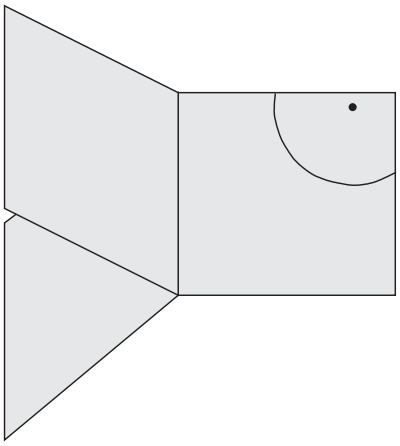

FIG. 12. *One tree in the neighborhood.*

branching pattern as in Figure 12 or three neighboring combinatorially different trees as in Figure 13 or 15 neighboring trees (if all the edges are small and we are close to the star tree).

Of course, for a tree with two inner edges, this is the only possible way to have these two edges small. This same notion of neighborhood containing 15 different branching orders applies to all trees on as many leaves as necessary, but which have two contiguous "small edges" and all the other inner edges significantly larger than 0. This picture of tree space frees us from having to use simulations to find out how many different trees are in a neighborhood of a given radius r around a given tree. All we have to do is check how many contiguous edges in the tree are smaller than r. Say there is just one set of small contiguous edges of size n_r. Then the neighborhood contains

$$(2n_r - 3)!! = (2n_r - 3) \times (2n_r - 5) \times \cdots \times 3 \times 1$$

different types of trees. Thus a point very close to the

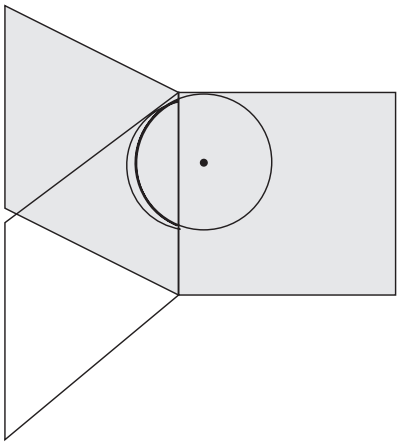

FIG. 13. *A tree with two neighbors.*

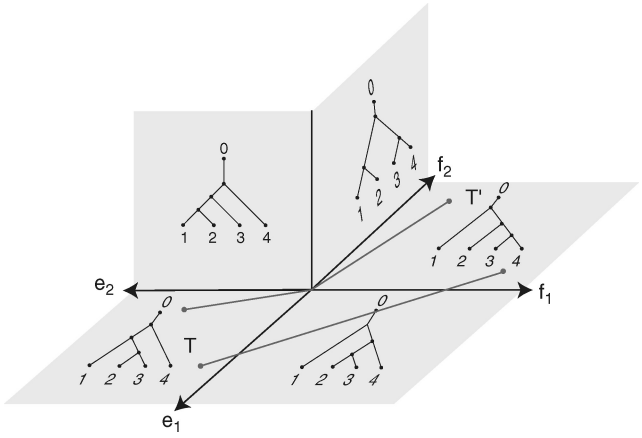

FIG. 14. *Cone path and geodesic path in* $\mathcal{T}_4$.

star tree at the origin will have an exponential number of neighbors.

Billera, Holmes and Vogtmann (2001) have shown that the geodesic distance induced by this filling of tree space always exists. There is always a path going through the star tree to go from one tree to another. Sometimes this is the shortest path, thus the distance; sometimes there is a shorter path as can be seen in Figure 14.

The last element necessary to make a rigorous picture of tree space is the probability measure. We can define such a measure in the parametric mutation model of maximum likelihood estimation of trees. Figure 15 presents a picture of the likelihood contours in the four leaf case for the Markovian evolution model with two parameters. Many other uses of tree space were described by Billera, Holmes and Vogtmann (2001).

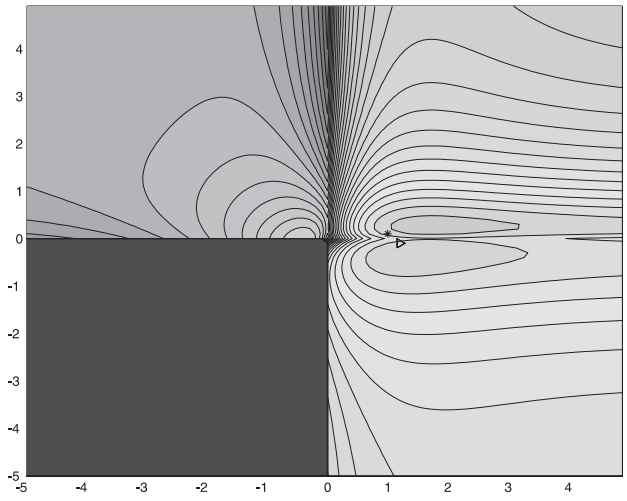

FIG. 15. *Likelihood contours.*

6. SUMMARY: HOW SHOULD ONE BOOTSTRAP?

First the simulation methods involved in resampling should follow the biological knowledge as closely as possible. Block bootstrapping (Künsch, 1989) with blocks that have size on the same order as the size of dependent blocks as estimated from real data can be used. Alternatively, simulations can follow covarion models (Fitch and Markowitz, 1970) as in Lockhart et al. (1996) and Tuffley and Steel (1998). These lean toward models where the observations—columns of the sequence matrix—are not identically distributed, but depend on a covariable. In their studies, the covariate was binary, that is, could be either on or off. This is also modeled by "hot spots" along the sequences (see Tang and Lewontin, 1999) and Gamma rate variations for different sites as in Yang (1994). Any of these models can be bootstrapped using a parametric bootstrap. It is more coherent to use the same model at the resampling stage as at the estimation stage. Thus the parametric bootstrap as implemented in SEQGEN by Rambaut and Grassly (1997) is coherent when doing a maximum likelihood estimation, and even allows inclusion of rate heterogeneity models such as Yang (1994) and Felsenstein and Churchill (1996).

A justification for using the multinomial bootstrap may be that the mutation model itself is being tested. This leads to confusing conclusions because the alternative is not explicitly defined.

Can we switch paradigms as we deem fit? Starting with a parametric model for evolution, such as the one parameter Jukes–Cantor model, does it make sense, after a tree has been estimated, to switch to a different paradigm at the validation stage and use a nonparametric bootstrap to compute confidence levels on the tree? This is an open problem. Some statisticians often switch paradigms in the middle of their studies from data analytic to parametric to nonparametric. Usually what actually can be proved is that if the parametric model is correct, there is a loss in power (*sensu statisticae strictu*) when switching to a nonparametric procedure. As for the mixture between Bayesian and frequentist, empirical Bayes (Robbins, 1980, 1983, 1985) is a typical example of the loose boundaries that exist when choosing different paradigms at different stages of an analysis.

Having settled on the process for generating the new data, one can think about the statistic to bootstrap. The simplest case where only one clade is of interest can be handled by recalibrating Felsenstein's (1983) repeatability index as was done by Efron, Halloran

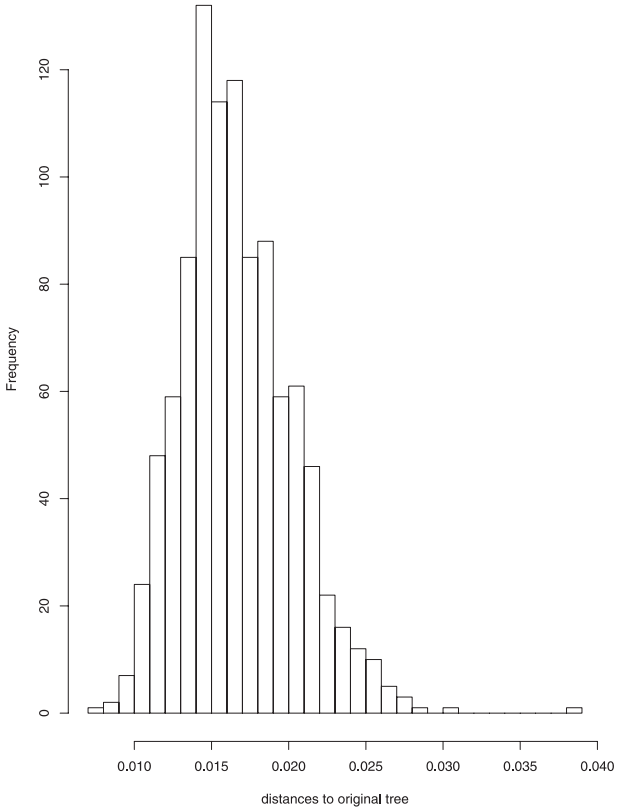

FIG. 16. *Bootstrap histogram of* $d(\hat{\tau}^*, \hat{\tau})$.

and Holmes (1996) (for which a PYTHON/R/PHYLIP program is available from this author) or Zharkikh and Li (1995). However, this recalibration does not apply to more extensive use of the bootstrap. In particular, if several edges are of interest, multiple testing problems loom and no correction methods are readily available.

The geometric perspective outlined in Section 5 does provide two alternatives. We can use the distance d and look at the distribution of distances, as for instance in the histogram of Figure 16. This gives the average distance $\bar{d}(\hat{\tau}^*, \hat{\tau}) = 0.167$ and a 95th percentile of 0.023. Such numbers allow the use of the bootstrap to test various hypotheses.

In the maximum likelihood context, it is possible to construct likelihood contours (Ramsay, 1978) such as those in Figure 15 for the bootstrap resampling distribution following Hall (1987). Another method now available through the geometric method is to use convex hulls. These exist in $\mathcal{T}_n$ because the space has nonpositive curvature as proved in Billera, Holmes and Vogtmann (2001). Thus methodology based on hull peeling as in Liu, Parelius and Singh (1999) could provide a 95% convex hull and one could answer questions such as, "Does the star tree belong to the 95% convex hull?" If it does, the data are probably very far from being treelike. Another question that could be answered is, "Is the convex hull elongated along a certain direction?"

7. FUTURE RESEARCH

It would be meaningful to consider bootstrapping the whole procedure, alignment and treebuilding included as, for instance, in Gong (1986).

In the statistical literature, the theorems that justify the use of the bootstrap usually state that the distribution of the distance between the true parameter and the estimate can be well approximated by the distribution between the estimate and the bootstrap resampled estimate—something that can be summarized as

$$\text{Distribution}(d(\tau, \hat{\tau})) \approx \text{Distribution}(d(\hat{\tau}, \hat{\tau}^*)).$$

However, most of the theoretical work involves an assumption of independent, identically distributed variables and some strong assumptions on the properties of the distance. No actual theory in the phylogenetic context exists at present, although referring to theoretical arguments in other cases does provide useful insight into the most sensible simulations to undertake.

Considering the resampling procedure as a means for making a small plausible perturbation of the original data and looking at how much the tree changes, or more precisely whether or not a clade disappears, is in fact a way to understand the estimator's continuity (in the mathematical sense) by simulation and provides quite a bit of information.

There has also been phylogenetic work on ideas very similar to Breiman (1996) in statistical learning theory. Berry and Gascuel (1996) used this idea to find more robust trees by taking consensii of bootstrapped data. The method for making consensii is quite crude (majority rule consensus chooses all the edges that have more than 50% support). A more refined method could use a geometric consensus using the distance defined in Section 5 and minimizing either $\sum d(\tau_i, \tau)$ or $\sum d^2(\tau_i, \tau)$.

The biggest change in statistics as practiced by statisticians over the last 30 years has been a decrease in the use of p values. Tukey and his co-workers in exploratory data analysis have shown the importance of keeping as much of the data in mind as possible. A picture is worth a thousand words, and geometry has much to offer in the complex multidimensional

analysis of DNA sequences through trees, graphs and their assorted confidence regions. Some open problems for mathematically inclined colleagues include the following:

- Can we give quantitative bounds on how much the new assumptions involved in more realistic models of sequence distribution change the bootstrap distributions of the distances?
- How can we represent confidence regions and neighborhoods of trees graphically?
- Can we extend some of the work on bootstrapping trees to networks? This would be useful for analyzing regulatory networks, but also would enable us to test whether data are actually treelike.

APPENDIX

Some noncommercial programs for constructing, evaluating and visualizing phylogenetic trees. A very complete list, including the commercial packages not mentioned here, can be found at *http://evolution.genetics.washington.edu/phylip/software.html.*

PHYLIP (Author: Joe Felsenstein.)
Available on almost all machines as source and executables, its different programs allow computations of parsimony, distance matrix methods, maximum likelihood and other methods on a variety of types of data, including DNA, RNA and protein sequences. It implements the nonparametric bootstrap and consensus programs used in the examples here. Website: *http://evolution.genetics.washington.edu.*

TREE-PUZZLE (Authors: Heiko A. Schmidt, Korbinian Strimmer, Martin Vingron and Arndt von Haeseler.)
Maximum likelihood phylogenetic analysis using quartets and parallel computing. Website: *http://www.tree-puzzle.de/.*

MR. BAYES (Authors: John Huelsenbeck and Fredrik Ronquist.)
This program implements a parametric Bayesian method for finding trees by Monte Carlo Markov chains that provide both a Bayesian estimate and the associated posterior distribution. Website: *http://morphbank.ebc.uu.se/mrbayes/.*

LVB (Author: Daniel Barker.)
Contruction of phylogenies using simulated annealing. Website: *http://sapc34.rdg.ac.uk/lvb/.*

SEQGEN (Authors: Andrew Rambaut and Nick Grassly.)
Parametric bootstrap for generating nucleotide sequences from a given phylogeny or mixture of phylogenies. As well as the more classical parametric models, they allow rate heterogeneity among sites. Website: *http://evolve.zoo.ox.ac.uk/software/Seq-Gen/Seq-Gen.html.*

TREEVIEW (Author: Rod Page.)
For visualizing trees on a PC or Mac, or with Linux/Unix with TREEVIEW X. Website: *http://taxonomy.zoology.gla.ac.uk/rod/treeview.html.*

APE (Analyses of phylogenetics and evolution) (Authors: Emmanuel Paradis, Korbinian Strimmer, Julien Claude, Yvonnick Noel and Ben Bolker.)
APE is an R package that provides functions for reading, writing, plotting and manipulating phylogenetic trees, but not estimating them. It allows analyses of comparative data in a phylogenetic framework. Website: *http://cran.r-project.org/src/contrib/PACKAGES.html#ape.*

PAL (Authors: Alexei Drummond, Ed Buckler and Korbinian Strimmer.)
A collection of Java classes for use in molecular phylogenetics that enables maximum likelihood, neighbor joining and least squares analysis. Website: *http://www.cebl.auckland.ac.nz/pal-project/.*

NONA (Author: Pablo Goloboff.)
Computes maximum parsimony and consensus trees on Windows machines only. Website: *http://www.cladistics.com/about_nona.htm.*

ACKNOWLEDGMENTS

I thank George Casella for giving me the opportunity to write this survey. Marc Coram and Persi Diaconis patiently read several versions of this text. Brad Efron and Warren Ewens had several helpful discussions on the bootstrap. My 2002 summer research students, Martina Koeva, Aaron Staple and Henry Towsner, all helped with computational aspects of this project.

This research was partly funded by Grant NSF DMS-00-72569.

REFERENCES

ALDOUS, D. (2001). Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statist. Sci.* **16** 23–34.

BAKER, C., LENTO, G., CIPRIANO, F. and PALUMBI, S. (2000). Predicted decline of protected whales based on molecular genetic monitoring of Japanese and Korean markets. *Proc. Roy. Soc. London Ser. B* **267** 1191–1199.

BAKER, C. and PALUMBI, S. (1994). Which whales are hunted? A molecular genetic approach to monitoring whaling. *Science* **265** 1538–1539.

BERRY, V. and GASCUEL, O. (1996). On the interpretation of bootstrap trees: Appropriate threshold of clade selection and induced gain. *Molecular Biology and Evolution* **13** 999–1011.

BILLERA, L., HOLMES, S. and VOGTMANN, K. (2001). Geometry of the space of phylogenetic trees. *Adv. in Appl. Math.* **27** 733–767.

BREIMAN, L. (1996). Bagging predictors. *Machine Learning* **24** 123–140.

BREMER, K. (1988). The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution* **42** 795–803.

COOPER, A. and PENNY, D. (1997). Mass survival of birds across the cretaceous–tertiary boundary: Molecular evidence. *Science* **275** 1109–1113.

DIACONIS, P. (1989). A generalization of spectral analysis with application to ranked data. *Ann. Statist.* **17** 949–979.

DIACONIS, P. and HOLMES, S. (1998). Matchings and phylogenetic trees. *Proc. Nat. Acad. Sci. U.S.A.* **95** 14,600–14,602.

DIACONIS, P. and HOLMES, S. (2002). Random walks on trees and matchings. *Electronic Journal of Probability* **7** 17 pages.

DURBIN, R., EDDY, S., KROGH, A. and MITCHISON, G. (1998). *Biological Sequence Analysis.* Cambridge Univ. Press.

EFRON, B. and TIBSHIRANI, R. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, London.

EFRON, B. and TIBSHIRANI, R. (1998). The problem of regions. *Ann. Statist.* **26** 1687–1718.

EFRON, B., HALLORAN, E. and HOLMES, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proc. Nat. Acad. Sci. U.S.A.* **93** 13,429–13,434.

FELSENSTEIN, J. (1983). Statistical inference of phylogenies (with discussion). *J. Roy. Statist. Soc. Ser. A* **146** 246–272.

FELSENSTEIN, J. (2003). *Inferring Phylogenies.* Sinauer, Boston.

FELSENSTEIN, J. and CHURCHILL, G. A. (1996). A hidden Markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution* **13** 93–104.

FITCH, W. (1971a). The nonidentity of invariable positions in the cytochromes *c* of different species. *Biochemical Genetics* **5** 231–241.

FITCH, W. (1971b). Rate of change of concomitantly variable codons. *Journal of Molecular Evolution* **1** 84–96.

FITCH, W. M. and MARKOWITZ, E. (1970). An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics* **4** 579–593.

FREEDMAN, D. A. and PETERS, S. C. (1984a). Bootstrapping a regression equation: Some empirical results. *J. Amer. Statist. Assoc.* **79** 97–106.

FREEDMAN, D. A. and PETERS, S. C. (1984b). Bootstrapping an econometric model: Some empirical results. *J. Bus. Econom. Statist.* **2** 150–158.

GONG, G. (1986). Cross-validation, the jackknife, and the bootstrap: Excess error estimation in forward logistic regression. *J. Amer. Statist. Assoc.* **81** 108–113.

GREEN, P. J. (1981). Peeling bivariate data. In *Interpreting Multivariate Data* (V. Barnett, ed.) 3–19. Wiley, New York.

HALL, P. (1987). On the bootstrap and likelihood-based confidence regions. *Biometrika* **74** 481–493.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (1986). *Robust Statistics*: *The Approach Based on Influence Functions*. Wiley, New York.

HEDGES, S. B. (1992). The number of replications needed for accurate estimation of the bootstrap p-value in phylogenetic studies. *Molecular Biology and Evolution* **9** 366–369.

HENDY, M. D. and PENNY, D. (1993). Spectral analysis of phylogenetic data. *J. Classification* **10** 5–23.

HENDY, M. D., PENNY, D. and STEEL, M. A. (1994). A discrete Fourier analysis for evolutionary trees. *Proc. Nat. Acad. Sci. U.S.A.* **91** 3339–3343.

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

HOLMES, S. (1999). Phylogenies: An overview. In *Statistics in Genetics* (M. E. Halloran and S. Geisser, eds.) 81–119. Springer, New York.

HUBER, P. J. (1996). *Robust Statistical Procedures*, 2nd ed. SIAM, Philadelphia.

KÜNSCH, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* **17** 1217–1241.

LANL (2002). HIV database. Available at URL: http://hiv-web.lanl.gov/content/hiv-db/mainpage.html.

LAURITZEN, S. L. (1988). *Extremal Families and Systems of Sufficient Statistics. Lecture Notes in Statist.* **49**. Springer, Berlin.

LENTO, G. M., CIPRIANO, F., PATENAUDE, N. J., PALUMBI, S. R. and BAKER, C. S. (1998). Taking stock of minke whale in the North Pacific: The origins of products for sale in Japan and Korea. Technical Report SC/50/RMP15, Scientific Committee, International Whaling Commission.

LI, S., PEARL, D. K. and DOSS, H. (2000). Phylogenetic tree construction using Markov chain Monte Carlo. *J. Amer. Statist. Assoc.* **95** 493–508.

LI, W. H. (1997). *Molecular Evolution.* Sinauer, Boston.

LI, W. H. and ZHARKIKH, A. (1995). Statistical tests of DNA phylogenies. *Systematic Biology* **44** 49–63.

LIU, R. Y., PARELIUS, J. M. and SINGH, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference (with discussion). *Ann. Statist.* **77** 783–858.

LOCKHART, P. J., LARKUM, A. W. D., STEEL, M. A., WADDELL, P. J. and PENNY, D. (1996). Evolution of chlorophyll and bacteriochlorophyll: The problem of invariant sites in sequence analysis. *Proc. Nat. Acad. Sci. U.S.A.* **93** 1930–1934.

MADDISON, D. (1991). The discovery and importance of multiple islands of most parsimonious trees. *Systematic Zoology* **40** 315–328.

MAU, B., NEWTON, M. A. and LARGET, B. (1999). Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics* **55** 1–12.

NEI, M., KUMAR, S. and TAKAHASHI, K. (1998). The optimization principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small. *Proc. Nat. Acad. Sci. U.S.A.* **95** 12,390–12,397.

NEWTON, M. A. (1996). Bootstrapping phylogenies: Large deviations and dispersion effects. *Biometrika* **83** 315–328.

PAGE, R. and HOLMES, E. (2000). *Molecular Evolution*: *A Phylogenetic Approach*. Blackwell Science, Oxford, UK.

RAMBAUT, A. and GRASSLY, N. C. (1997). Seq-gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computational Applied Bioscience* **13** 235–238.

RAMSAY, J. O. (1978). Confidence regions for multidimensional scaling analysis. *Psychometrika* **43** 145–160.

ROBBINS, H. (1980). An empirical Bayes estimation problem. *Proc. Nat. Acad. Sci. U.S.A.* **77** 6,988–6,989.

ROBBINS, H. (1983). Some thoughts on empirical Bayes estimation. *Ann. Statist.* **11** 713–723.

ROBBINS, H. (1985). Linear empirical Bayes estimation of means and variances. *Proc. Nat. Acad. Sci. U.S.A.* **82** 1571–1574.

RODRIGO, A. G. (1993). Calibrating the bootstrap test of monophyly. *International Journal for Parasitology* **23** 507–514.

SANDERSON, M. J. (1995). Objections to bootstrapping phylogenies: A critique. *Systematic Biology* **44** 299–320.

SCHRÖDER, E. (1870). Vier combinatorische Probleme. *Zeitschrift für Mathematik und Physik* **15** 361–376.

TANG, H. and LEWONTIN, R. (1999). Locating regions of differential variability in DNA and protein sequences. *Genetics* **153** 485–495.

TUFFLEY, C. and STEEL, M. (1998). Modeling the covarion hypothesis of nucleotide substitution. *Math. Biosci.* **147** 63–91.

TUKEY, J. (1975). Mathematics and the picturing of data. In *Proc. International Congress of Mathematicians* (R. D. James, ed.) 523–531. Vancouver.

YANG, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Molecular Evolution* **39** 306–314.

YANG, Z. and RANNALA, B. (1997). Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. *Molecular Biology and Evolution* **14** 717–724.

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