Statistical Science
2003, Vol. 18, No. 2, 256–267

Applying the Bootstrap in Phylogeny Reconstruction

Pamela S. Soltis and Douglas E. Soltis

Abstract. With the increasing emphasis in biology on reconstruction of phylogenetic trees, questions have arisen as to how confident one should be in a given phylogenetic tree and how support for phylogenetic trees should be measured. Felsenstein suggested that bootstrapping be applied across characters of a taxon-by-character data matrix to produce replicate "bootstrap data sets," each of which is then analyzed phylogenetically, with a consensus tree constructed to summarize the results of all replicates. The proportion of trees/replicates in which a grouping is recovered is presented as a measure of support for that group. Bootstrapping has become a common feature of phylogenetic analysis. However, the interpretation of bootstrap values remains open to discussion, and phylogeneticists have used these values in multiple ways. The usefulness of phylogenetic bootstrapping is potentially limited by a number of features, such as the size of the data matrix and the underlying assumptions of the phylogeny reconstruction program. Recent studies have explored the application of bootstrapping to large data sets and the relative performance of bootstrapping and jackknifing.

Key words and phrases: Bootstrap, phylogeny, support, jackknife.

INTRODUCTION

Systematic biologists for centuries have striven to expose the "natural order" of living things, and for the past 150 years (since Darwin, 1859) this endeavor has focused largely on inferring phylogeny—that is, the evolutionary history of living organisms, or the "tree of life" (see Box 1). Many methods, in addition to intuition, have been developed for use in phylogeny reconstruction. Explicit cladistic reconstruction of phylogenetic trees can be traced largely to the pioneering work of Hennig (1966). Early efforts to reconstruct phylogeny were based on morphological data, but as molecular characters became accessible, they were quickly integrated into phylogenetic analyses. With the increasing emphasis on tree reconstruction, questions arose as to how confident one should be in a given phylogenetic tree and how support for phylogenetic trees should be measured. Cavender (1978, 1981) developed an approach to assess how many steps longer a tree needed to be than the most parsimonious (i.e., shortest) tree(s) to be significantly longer than the shortest tree(s), and Templeton's (1983) test measured whether one tree is significantly better supported than another tree.

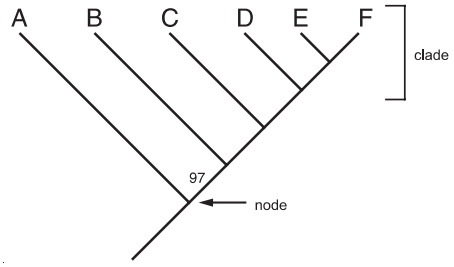
The bootstrap was introduced by Efron (1979; see also Efron and Gong, 1983, Diaconis and Efron, 1983) to obtain estimates of error in nonstandard situations by resampling the data set many times to provide a distribution against which hypotheses could be tested. Very soon after the introduction of the bootstrap, Penny, Foulds and Hendy (1982) and Penny and Hendy (1985) applied it to questions in phylogeny reconstruction, and Felsenstein (1985) formally proposed bootstrapping as a method for obtaining confidence limits on phylogenies. Phylogeny reconstruction uses a matrix of taxa (usually species

Pamela S. Soltis is Curator of Molecular Systematics and Evolutionary Genetics, Florida Museum of Natural History and the Genetics Institute, University of Florida, Gainesville, Florida 32611 (e-mail: psoltis@flmnh.ufl.edu). Douglas E. Soltis is Professor of Botany, Department of Botany and the Genetics Institute, University of Florida, Gainesville, Florida 32611.

BOX 1. *Basic terms and concepts in phylogenetics*

Clade—a monophyletic group; i.e., an ancestor and all of its descendants
Character—an attribute of an organism (or group of organisms); it may be a morphological, anatomical or chemical feature or a nucleotide position, etc.
Node—a branching point on a phylogenetic tree
Phylogeny—evolutionary history of a group of organisms
Phylogenetic tree—a diagram depicting an interpretation of phylogeny
Taxon—a group of related organisms; e.g., *Homo sapiens* is a taxon at the rank of species, *Homo* is a taxon at the rank of genus, etc.

In the figure on the right, A–F represent taxa related as indicated by the branches. E and F are sister species and are thus a clade; D, E and F form a clade; C, D, E and F form a larger clade, etc. Each node, or internal branching point, represents an ancestor of the clade that lies "above" it. Thus, the node indicated by the arrow is the common ancestor of taxa A–F. Clade B–C–D–E–F is supported by a bootstrap value of 97%.



or populations) and characters (today, generally DNA sequence data; but also morphological, chemical or other non-DNA characters). Felsenstein suggested that bootstrapping be applied across characters—that is, the characters are sampled from the data matrix with replacement to produce replicate "bootstrap data sets," each of which is then analyzed phylogenetically, with a consensus tree constructed to summarize the results of all replicates (Boxes 2 and 3). The proportion of trees/replicates in which a clade is recovered is presented as a percentage and referred to variably as the bootstrap value, bootstrap percentage (BP) or, less commonly, bootstrap *p*-value.

Following Felsenstein's (1985) explicit description of the procedure, bootstrapping became a common tool in phylogeny reconstruction. The fact that readily available packages of phylogenetic software such as PHYLIP (Felsenstein) and PAUP 3.1/PAUP* (Swofford, 1998) incorporated a bootstrapping algorithm greatly facilitated the widespread application of the bootstrap. One reason for the popularity and importance of the bootstrap in systematic applications is nicely summarized by Sanderson (1989): "even if one is not willing to accept the validity of its assumptions, the bootstrap is valuable because it provides a systematic method of assessing the robustness of a data set to perturbation."

Phylogenetic trees are typically presented with bootstrap values associated with the nodes (Figure 1), and

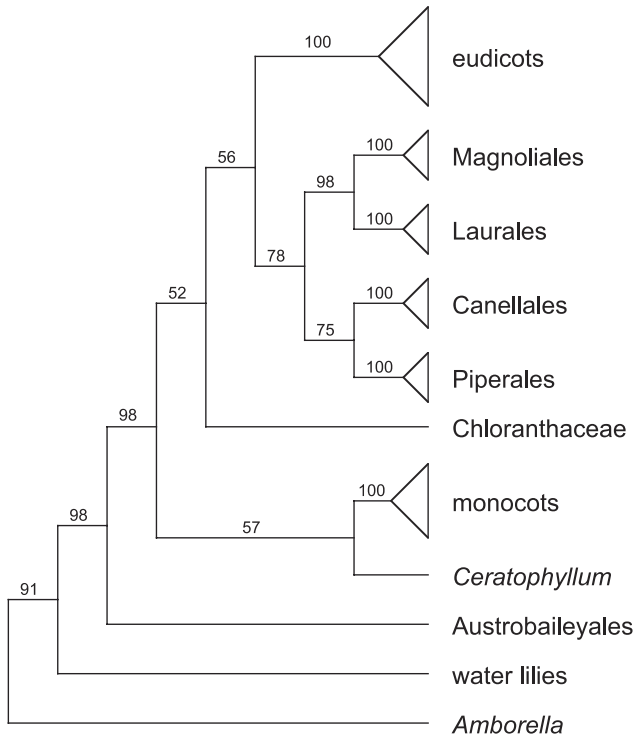


FIG. 1. *Phylogenetic tree, based on DNA sequences from* 11 *genes* (*greater than* 15,000 *bp*), *showing relationships among the basal branches of flowering plants* (*simplified and redrawn from Zanis et al.*, 2002). *Numbers above branches are bootstrap values. Note that some nodes receive bootstrap values of* 100%, *indicating strong support for these nodes, whereas other nodes receive much weaker support* (*e.g.*, 52%, 56%).

BOX 2. *Constructing bootstrap data sets in phylogeny reconstruction*

Constructing bootstrap data sets. The original data set of 4 taxa (A–D) each with 10 nucleotide characters is bootstrapped across characters (with replacement) to produce bootstrap pseudoreplicates. Each pseudoreplicate contains each of the 4 original taxa, but some original characters are represented more than once and some not at all.

Original Data Set

Taxa	Characters
	1 2 3 4 5 6 7 8 9 10
A	C G A A C C A C T T
B	C G A A C C G G T T
C	G G T A C C G G A T
D	G C T A G C G C A T

Bootstrap Data Sets

Bootstrap Pseudoreplicate 1:

Taxa	Characters
	8 10 7 4 1 10 2 8 5 3
A	C T A A C T G C C A
B	G T G A C T G G C A
C	G T G A G T G G C T
D	C T G A G T C C G T

Bootstrap Pseudoreplicate 2:

Taxa	Characters
	1 8 10 4 2 9 2 8 5 6
A	C C T A G T G C C C
B	C G T A G T G G C C
C	G G T A G A G G C C
D	G C T A C A C C G C

Bootstrap Pseudoreplicate 3:

Taxa	Characters
	3 2 5 7 1 6 9 4 4 10
A	A G C A C C T A A T
B	A G C G C C T A A T
C	T G C G G C A A A T
D	T C G G G C A A A T

Bootstrap Pseudoreplicate 4:

Taxa	Characters
	7 8 5 8 9 6 4 10 1 5
A	A C C C T C A T C C
B	G G C G T C A T C C
C	G C G A C C A T G C
D	G C G C A C A T G C

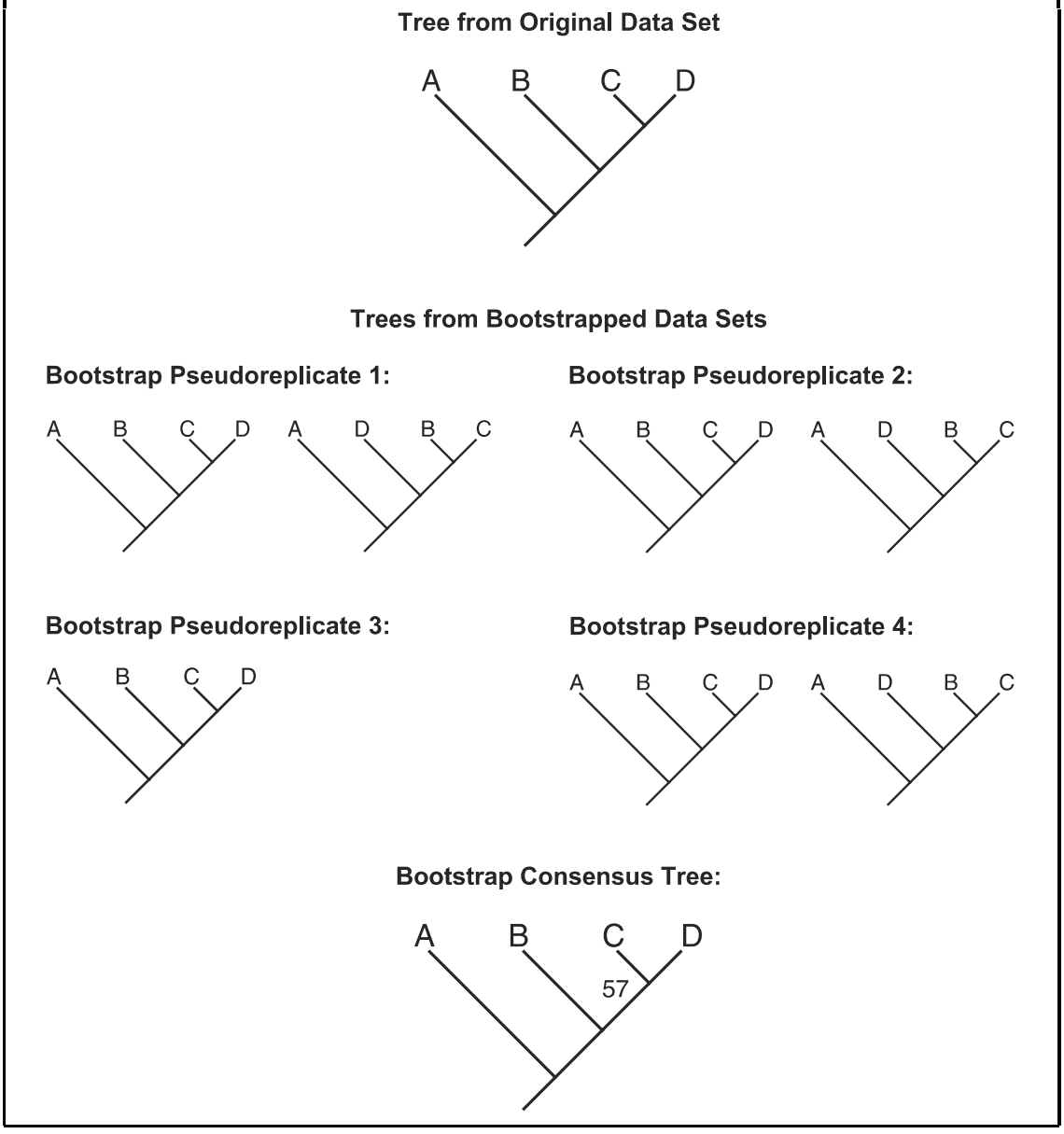
many systematics journals actually require bootstrap values or an alternative measure of support on trees. The monumental impact of the bootstrap in systematics is provided by a perusal of articles in mainline systematics journals, such as *Systematic Botany* and *Systematic Biology*. Of the papers published in 2000 in *Systematic Botany* and in 2001 in *Systematic Biology*, at least 50% of the papers presented the results of phylogenetic analyses: all studies in both journals that reconstructed phylogeny used bootstrapping to measure nodal support.

INTERPRETATION OF BOOTSTRAP VALUES

The interpretation of bootstrap values has been both murky and controversial. Felsenstein (1985) proposed that bootstrap values of 95% or greater be considered statistically significant and indicate "support" for a clade; alternative nodes can be rejected if they occur in less than 5% of the bootstrap estimates. However, bootstrap confidence levels apply to single nodes—they are not joint confidence statements. Thus, although two clades may each be supported at 95% and are thus not contradictory, the confidence interval that includes

BOX 3. *Analyzing bootstrap data sets and the resulting phylogenetic trees*

Analyzing bootstrap data sets. Each bootstrap data set is analyzed phylogenetically. The phylogenetic trees resulting from parsimony analysis of each bootstrap data set are shown. The trees from all four of these data sets are summarized in a consensus tree, typically a 50% majority rule consensus tree that shows all clades recovered in at least 50% of all trees. Below, taxon A was designated the outgroup, forcing B–C–D together; thus, no bootstrap value is given for B–C–D.



both clades may be only 90%, and the joint confidence drops as additional nodes are considered. Joint confidence will thus be necessarily low for a large tree, even if all nodes are strongly supported. A majority-rule consensus tree summarizing all of the bootstrap replicates provides a set of noncontradictory nodes, each with a rejection probability below 50%, and can be interpreted as an "overall bootstrap estimate of the phylogeny" (Felsenstein, 1985, page 786).

Many practitioners would like bootstrap values to measure the "truth." However, Felsenstein (1985) explicitly stated that bootstrapping provides a confidence

interval that contains *the phylogeny that would be estimated from repeated sampling of many characters from the underlying set of all characters*, NOT the true phylogeny. Thus, Felsenstein viewed bootstrap values as measures of *repeatability* rather than measures of *accuracy*.

Hillis and Bull (1993) used simulation studies and a laboratory-generated phylogeny to assess bootstrap values as measures of repeatability and accuracy and considered the precision of these estimates. They used Felsenstein's definitions of repeatability and accuracy and defined precision as "the degree to which bootstrap proportions based on a finite set of pseudosamples are expected to match the values that would be obtained from an infinite set of pseudosamples" (Hillis and Bull, 1993, page 183). Hillis and Bull contend that bootstrap proportions are highly imprecise, except when the parametric values are near 0 and 1. The high variance and corresponding imprecision of bootstrap proportions were also described by Hedges (1992; although he referred to "precision" as "accuracy" in this paper). Hillis and Bull concluded that the imprecision of bootstrap proportions impairs their use as measures of repeatability and suggested that analysis of a new data set of independent characters (i.e., a true replicate rather than a pseudoreplicate) could not be expected to yield bootstrap proportions similar to those achieved with the initial data set. However, their studies showed that, for conditions they considered "typical" of most phylogenetic analyses, bootstrap proportions are biased and conservative measures of accuracy; that is, under certain conditions, bootstrap proportions of 70% or more usually correspond (with greater than or equal to 95% probability) to a "real" clade. Thus, Hillis and Bull (1993) challenged Felsenstein's original interpretations of bootstrap values: contrary to Felsenstein, bootstrap values are poor measures of repeatability and poor, but conservative, measures of accuracy.

Many systematists have adopted Hillis and Bull's "70%" value as an indication of support, but without regard for the conditions under which this value was obtained, that is, equal rates of change, symmetric phylogeny and internodal change of 20% or less of the characters. At least the first two conditions are quite unrealistic for real phylogenies, and when all of these conditions are not met, bootstrap values of 50% or more may be overestimates of accuracy (Hillis and Bull, 1993). Thus, bootstrap values are perhaps poor measures of accuracy as well. For a well-supported clade, bootstrap values will almost always underestimate both accuracy and repeatability (Sanderson and Wojciechowski, 2000, and references therein).

Further analytical, theoretical and statistical work has helped to refine the meaning of bootstrap values as applied to phylogenies. Zharkikh and Li, using analytical and simulation approaches for the four-taxon case (1992a, b) and a simple multinomial model (1995), also found bootstrap values to be biased and conservative measures of accuracy, at least under conditions when the phylogenetic method is consistent. This bias was also noted by Felsenstein and Kishino (1993), using a simple case that did not even involve the bootstrap, and was interpreted as an effect of placing a probability value on a prespecified hypothesis. Further statistical analysis (Efron, Halloran and Holmes, 1996; Newton, 1996) identified the source of bias in bootstrapping in phylogenetics. Efron, Halloran and Holmes (1996) demonstrated that the bootstrap method itself is not biased; rather, the bias is due to the nature of phylogenetic problems and the implementation of bootstrapping in phylogeny reconstruction. They concluded that bootstrapping following Felsenstein (1985) provides a "reasonable first approximation" (Efron, Halloran and Holmes, 1996, page 13,429) to confidence levels of observed clades, with more ambitious bootstrapping providing improved confidence levels and an interpretation more in line with standard concepts of confidence levels and hypothesis testing. The apparent bias observed earlier was shown to result from dispersion effects in the joint distribution of sample and bootstrap empirical distributions (Newton, 1996). Regarding the interpretation of "typical" bootstrap values, Felsenstein and Kishino (1993) suggested that a conservative but straightforward approach, given a bootstrap proportion P for a given clade, is to consider $1 - P$ as the probability of a Type I error, that is, falsely accepting a clade that is not there.

Is there consensus on interpretation of bootstrap values? Not exactly. Despite continued, in fact, expanded, use of bootstrapping on phylogenies, the interpretation of bootstrap values remains open to discussion. Although many biologists may interpret $1 - P$ as a highly conservative measure of the probability of Type I error, we think that most systematists simply view bootstrap values as relative assessments of clade support (cf. Sanderson, 1989; Hillis and Bull, 1993) rather than strict statistical values, despite Felsenstein and Kishino's (1993) modification and suggestion. In this sense, consensus has been reached among practitioners, if not among statisticians and theoreticians.

LIMITATIONS OF THE BOOTSTRAP

The usefulness of bootstrap values for assessing even relative confidence in clades is limited by the application of the bootstrapping procedure to topologies rather than single variables (i.e., nodes), the effects of varying numbers of characters on bootstrap values, statistical bias with increased taxon sampling and the underlying assumptions and properties of the data and phylogeny reconstruction algorithm. As noted above, the bootstrap as applied by Felsenstein (1985) to phylogenies provides assessments of support for specific clades, one at a time, rather than a joint confidence statement for the entire tree. Thus, although support for each of several clades may be high, joint confidence in the interrelationships among these clades cannot be adequately assessed. Furthermore, if bootstrap values are interpreted in a strict statistical manner, bootstrapping presents a "multiple tests" problem; that is, by chance one in 20 clades will show significance at the 95% level. Because the problem is too complex for a correction factor, Felsenstein (1985) assumes an a priori test, basically clade by clade, rather than multiple comparisons.

The magnitude of bootstrap values, and thus their usefulness for assessing support for nodes, is affected by the number of characters—both supporting the clade of interest and the data set as a whole. Felsenstein (1985) showed that for "perfectly Hennigian data" (i.e., no conflict among characters) at least three characters are needed to provide 95% support. With real data, conflict among characters may require greater numbers of characters to achieve 95% support. This "rule of three" is practical but may be far too conservative (cf. Sanderson, 1989). For example, nonconflicting groups supported by only two characters each will necessarily receive less than 95% support. This problem frequently arises in studies of closely related species that have not diverged extensively (Figure 2). One solution is to generate more characters for phylogenetic analysis. Although most molecular data sets, with their large numbers of nucleotides, generally are not plagued by this problem, it still arises occasionally in recently diverged groups and is often a serious limitation for morphological analyses.

Bootstrap values may also be affected artifactually by the total number of characters in the data set (e.g., reviewed in Harshman, 1994; Farris et al., 1996). Bootstrap support for a clade may decrease with the addition to the matrix of (i) characters that are compatible with but not informative for that node (Faith and Cranston, 1991), (ii) autapomorphies (Carpenter, 1992) or (iii) invariant characters (Kluge and Wolf, 1993). The expected bootstrap value (Harshman, 1994) for a node given a matrix of n characters, r of which support the node of interest, is $1 - (1 - r/n)^n$. When r is fixed, the bootstrap value decreases as n increases, even if the additional characters do not contradict any of the r characters supporting the node. The addition of such irrelevant (to the node in question) characters increases the pool of characters that may be selected for a bootstrap pseudoreplicate, thus decreasing the chance that a given relevant character will be selected. However, because the number of chances to be selected is greater for a larger data set, Harshman (1994) suggested that the two possible effects of irrelevant characters might counteract each other. To address this problem, he analyzed two exact simple cases and two empirical data sets and concluded that the effect of irrelevant characters on bootstrap values is small over the range of numbers of characters

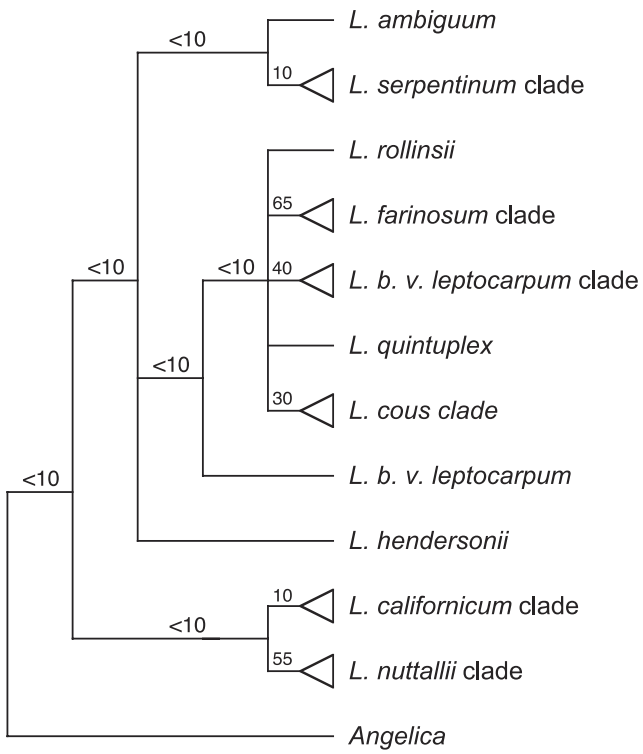


FIG. 2. *Strict consensus of two most-parsimonious trees showing phylogenetic relationships among species of the plant genus Lomatium* (*desert parsley*) *from western North America inferred from chloroplast DNA data. Although the data set had few conflicting characters and only two most-parsimonious trees were generated, bootstrap values are low because of the small number of characters supporting each node. This problem is typical of recently diverged groups* (*redrawn from Soltis and Novak*, 1997). *Numbers above branches are bootstrap values.*

likely to be included in real data sets. But, because the number of irrelevant characters may differ among nodes, the effect of increasing *n* may also differ among nodes. To equalize the effect, Harshman recommended adding 1000 invariant dummy characters to a matrix. Of course, this addition may alter all bootstrap values across the tree (cf. Farris et al., 1996). In contrast to Harshman's conclusions, Carpenter's (1996) analysis of several real data sets found large effects of irrelevant characters. This discrepancy between Harshman's (1994) and Carpenter's (1996) conclusions may reflect underlying differences in the proportion of relevant characters in the data sets analyzed.

Bootstrap support also decreases with increasing taxon sample size (Sanderson and Wojciechowski, 2000), despite evidence from both simulation and empirical studies for increased phylogenetic accuracy with increased taxon sampling (e.g., Hillis, 1996; Graybeal, 1998; Soltis et al., 1998). Sanderson and Wojciechowski's (2000) thorough analysis of the effects of sampling on bootstrap proportions clearly demonstrates that bootstrap support is a function of both taxon number and search algorithm, with all methods (various parsimony approaches and neighbor joining) showing decreased bootstrap support as the number of taxa increased. However, corrected bootstrap proportions, following Efron, Halloran and Holmes (1996; see above), remained near 95% as taxon sample size increased, even though the conventional bootstrap values ranged from 67% (for all 140 taxa) to 92% (a random sample of 16 taxa). This statistical bias may likewise be overcome through iterated bootstrapping (sensu Rodrigo, 1993) or by the complete and partial bootstrap method (Zharkikh and Li, 1995). Unfortunately, all of these correctional methods are computationally intensive, and it is precisely those large data sets that are themselves computationally challenging that would benefit most from such corrections. Fast search options (e.g., bootstrapping under parsimony without branch swapping, in PAUP* 4.0, Swofford, 1998; parsimony jackknifing, Farris et al., 1996; see below), coupled with correctional methods, may provide support levels that are less affected by increasing numbers of taxa.

The results of phylogenetic bootstrapping are only as good as the data and tree reconstruction algorithm. As Felsenstein (1985) noted, the character data are assumed to be independent and identically distributed, an assumption that is generally accepted but rarely tested, especially for nucleotide characters. Violation of this assumption will necessarily affect, to varying extents, the outcome of tree reconstruction and inferences of support from bootstrapping or any other method. Felsenstein (1985, 1988) considered correlations among characters to be the "most serious challenge" to bootstrapping in phylogeny reconstruction. Furthermore, bootstrapping can be applied to any of the commonly used methods of phylogeny reconstruction, from clustering to neighbor joining, maximum parsimony and maximum likelihood. Because these methods differ dramatically in their performance, for example, when evolutionary rates are unequal (e.g., Felsenstein, 1978; Huelsenbeck and Hillis, 1993; Huelsenbeck, 1995) or when different programs are used (e.g., Farris et al., 1996), bootstrap support may thus vary with the approach (or program!). The selection of an appropriate phylogenetic method is thus also necessary for bootstrap values to be at all informative. Finally, the specifications of the bootstrap analysis (e.g., type of branch swapping, if any) may affect the magnitude of bootstrap values (e.g., Sanderson and Wojciechowski, 2000; Mort, Soltis, Soltis and Mabry, 2000; DeBry and Olmstead, 2000; see below).

These limitations of the bootstrap are recognized even by proponents of bootstrapping. Many other systematists (see, e.g., Carpenter, 1992, 1996; Kluge and Wolf, 1993) object philosophically to the use of bootstrapping—or other statistical approaches—to phylogeny reconstruction. They argue that the support for cladistic hypotheses is evidenced by the degree of corroboration (cf. Hennig, 1966; Wiley, 1975; Platnick and Gaffney, 1977, 1978; Farris, 1983; Kluge, 1997, 1999) and that the application of a statistical method requires the specification of an underlying probability distribution—which is impossible for a phylogeny (e.g., Carpenter, 1992). Carpenter (1992, 1996) and Kluge and Wolf (1993) further point to violations of the assumptions of bootstrapping and other resampling techniques, emphasizing especially that the characters are not independent and identically distributed. Furthermore, bootstrapping assumes a close correspondence between the empirical distribution of characters in the data set and the distribution of a larger data set (see Felsenstein, 1985); this assumption may be violated in real data sets. Although this violation is widely recognized, its effects are unknown, causing some (e.g., Wendel and Albert, 1992) to reject the use of bootstrapping as a method for assessing support.

FAST METHODS

During the past few years, systematists have begun to analyze data sets containing hundreds of taxa.

For example, recent analyses of flowering plant phylogeny have involved 500 or more species (e.g., Chase et al., 1993; Soltis, Soltis and Chase, 1999; Soltis, et al., 2000; Savolainen et al., 2000), and an analysis of green plants and cyanobacteria contained 2,538 *rbcL* sequences (Källersjö et al., 1998). From a computational standpoint, standard bootstrap, as well as jackknife, analyses pose a potential problem, especially for large data sets. If done thoroughly and with a sufficiently large number of replicates (see Hedges, 1992), these analyses can be extremely time consuming and, for very large data sets, are not even practical.

One method for getting around the computational intensity of estimating nodal support using conventional bootstrap analysis is to analyze a larger number of replicate data sets and employ a simpler search strategy, with minimal or no branch swapping. "Branch swapping" refers to rearrangements of branches in a tree during the search for an optimal tree(s). For example, a "twig" is moved from "branch A" to "branch B," and a new optimality score is computed. If the score of the new tree with the swapped branch is worse than the score of the previous tree, then swapping of additional branches will continue on the original tree. If the score improves when the twig is swapped, the new tree is retained, and swapping begins on the new tree. This process continues until a single optimal tree (or a group of equally optimal trees) is obtained, or until the search is stopped. Several "fast" (i.e., "no-swapping") approaches have been proposed. These fast methods sacrifice thorough searches per bootstrap replicate for an increased number of bootstrap replicates, perhaps 1000 or more. They are computationally faster and easier than standard methods and can be used to analyze data sets of hundreds of taxa.

However, because fast methods, in the extreme, do not employ branch swapping, the possible effects—and their magnitude—of reduced search intensity on bootstrap values are not obvious. This issue has been explored using both empirical (Mort et al., 2000; Sanderson and Wojciechowski, 2000) and simulated (DeBry and Olmstead, 2000) data.

Mort et al. (2000) compared support values from bootstrap and jackknife analyses with fast searches (i.e., without branch swapping) with those from more thorough searches (i.e., with branch swapping) and explored the effect of increasing the number of replicates on bootstrap and jackknife support values. The fast methods provide estimates similar to, although generally less than, those obtained with branch swapping bootstrap analyses (Mort et al., 2000). Although statistically different, support values obtained from fast methods typically differed from those obtained with branch swapping by only a few percentage points at those nodes with bootstrap values greater than 85%. However, the difference between values from fast and more thorough searches was greater at those nodes with weaker support. Furthermore, a relatively small number of replicates, perhaps as few as 500, may be sufficient to obtain repeatable values (i.e., values did not change with 1000 or more replicates), in contrast to Hedges' (1992) recommendation that 2000 replicates are necessary to ensure a 95% confidence interval for a node receiving 95% support. Mort et al. (2000) concluded that standard bootstrapping should be employed whenever possible, but that the fast methods certainly are suitable for large data sets.

Sanderson and Wojciechowski (2000) also compared the effects of search strategies on bootstrap values, using nucleotide data across a range of taxon sampling intensities. At all sampling intensities, more thorough search algorithms yielded higher bootstrap values, with the differences ranging from 3% for small taxon samples (16 taxa) to 9% for larger samples (73 taxa) and the full data set (140 taxa). Sanderson and Wojciechowski (2000) attributed the differences to the ability of the more thorough searches to find a more similar set of trees than the less thorough algorithms, thus producing higher support values.

Simulation studies have also been conducted to evaluate the performance of the fast bootstrap (DeBry and Olmstead, 2000). "Reduced-effort bootstrapping" (i.e., with little or no branch swapping) did not generate inflated support values and produced bootstrap values similar to those computed with branch swapping at nodes with values greater than 90%. Furthermore, although the no-swapping approach generally produced lower values than the more thorough searches, the magnitude of the differences varied among nodes and among general categories of support. Therefore, no correction term could be devised to compensate for the differences caused by reduced search effort, a conclusion likewise reached by Mort et al. (2000).

Although additional investigation into the effects of search intensity on bootstrap values is needed, those studies to date concur that no-swap methods produce more conservative bootstrap values than more thorough searches. These differences are quite small for nodes with values greater than 90% but increase, somewhat unpredictably, at nodes with weak support. The effects of taxon number, tree shape and patterns

and rates of molecular evolution on no-swap relative to branch-swap bootstrap values deserve further study.

JACKKNIFING VERSUS BOOTSTRAPPING IN PHYLOGENY RECONSTRUCTION

Jackknife resampling, in which either characters or taxa are resampled without replacement (Efron, 1979, 1982; Efron and Gong, 1983; for a review of the jackknife, see Miller, 1974), has also been used to assess stability of and support for phylogenetic trees (Mueller and Ayala, 1982; Lanyon, 1985; Felsenstein, 1988).

The suitability of jackknifing versus bootstrapping in phylogeny reconstruction has received little discussion. Felsenstein (1985) posed that "one might wonder whether the jackknife would be a viable alternative to the bootstrap," but he favored the bootstrap to the "classical" jackknife, in which a single character is deleted per replicate. Single-character deletion from matrices with large numbers of characters would produce very similar trees from the respective replicates and would therefore not provide especially useful measures of support. Felsenstein (1985) suggested that one way to make the jackknife vary as much as the bootstrap would be to delete half of the characters, at random, in each replicate.

Farris et al. (1996) explored jackknifing further and concluded that 50% deletion is too severe. They incorporated their ideas for implementing jackknife resampling with parsimony analysis in the program JAC, which reads a matrix of nucleotide sequences, performs jackknife resampling, reconstructs the phylogeny with or without branch swapping and presents a tree showing group frequencies of 50% or more. PAUP* (Swofford, 1998) also has a jackknife option that can be used with parsimony, maximum likelihood or distance-based phylogeny reconstruction.

OTHER APPLICATIONS: PARAMETRIC BOOTSTRAPPING

An alternative to the standard, nonparametric bootstrap for testing specific hypotheses of relationship is the *parametric bootstrap* (Efron, 1985), in which a single data set can be used to parameterize a model of sequence evolution. This model is then used to simulate new, independent data sets, each of which is analyzed in turn to generate a distribution against which a specific hypothesis can be tested (see, e.g., Bull et al., 1993).

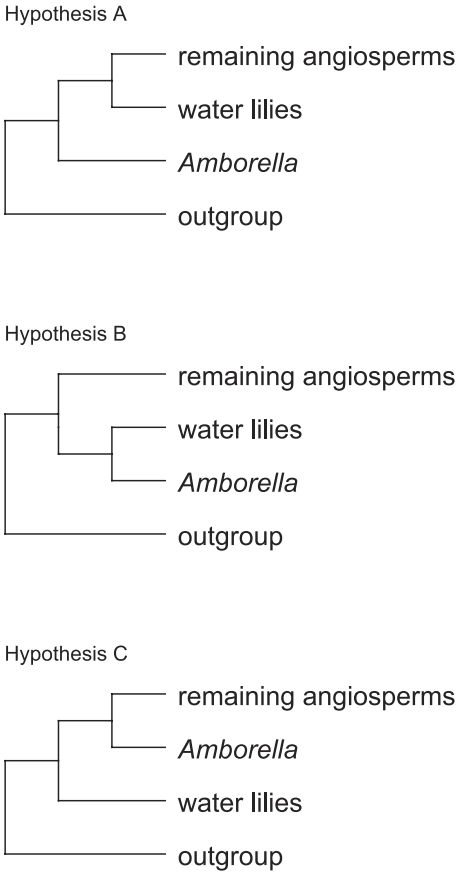


FIG. 3. *Alternative hypotheses for the root of flowering plant phylogeny* (*redrawn from Zanis et al*., 2002).

Zanis et al. (2002) applied the parametric bootstrap to the problem of the root of the flowering plants. Most phylogenetic analyses place *Amborella* alone as the sister to all other flowering plants (hypothesis A), whereas a few analyses place a clade of *Amborella* + water lilies (hypothesis B) or water lilies alone (hypothesis C) as sister to all other flowering plants (Figure 3; reviewed in Zanis et al., 2002). To test whether the latter two hypotheses were significantly different from the hypothesis of *Amborella* alone as sister to all other flowering plants, Zanis et al. (2002) parameterized a model of molecular evolution across the flowering plants, used this model to simulate independent data sets, analyzed these data sets using parsimony and maximum likelihood and concluded that hypothesis C could be rejected under both parsimony and likelihood and that hypothesis B could be rejected under parsimony but not under likelihood (Figure 4). Likelihood ratio tests with parametric bootstrapping allow for tests of specific hypotheses that cannot be addressed via nonparametric bootstrapping (see Huelsenbeck and Crandall, 1997).

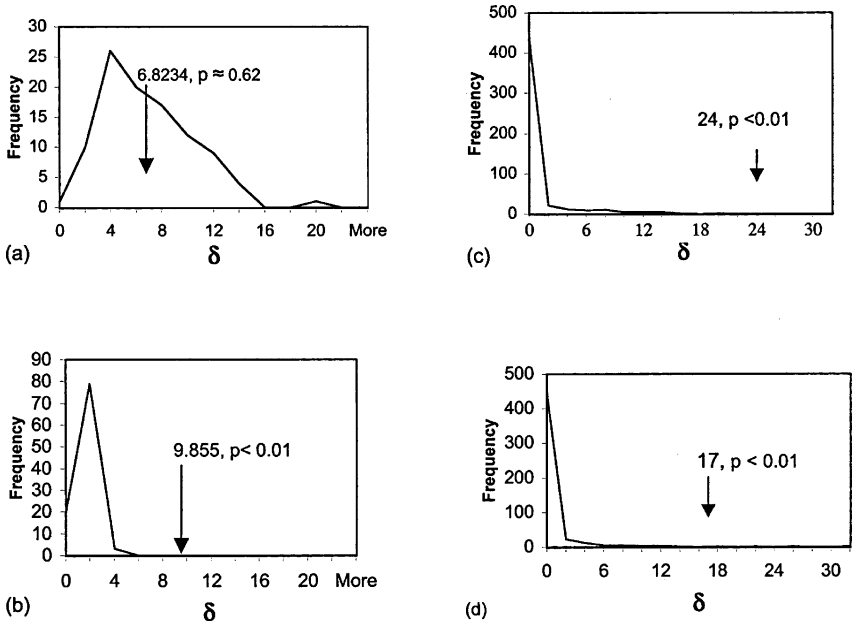


FIG. 4. *Results of likelihood ratio tests of the root of the flowering plants using the parametric bootstrap. Each histogram shows the distribution of* δ*, the likelihood ratio test statistic, which is the difference between the optimal tree(s) supporting hypothesis* B *or* C *and hypothesis* A *for data simulated under the assumption that hypothesis* B *or* C*, respectively, is correct. The observed value for the real data is indicated by the arrow.* (a) *Hypothesis* B *vs. hypothesis* A*, using maximum likelihood.* (b) *Hypothesis* C *vs. hypothesis* A*, using maximum likelihood.* (c) *Hypothesis* B *vs. hypothesis* A*, using parsimony.* (d) *Hypothesis* C *vs. hypothesis* A*, using parsimony* (*from Zanis et al.*, 2002).

SUMMARY

The application and usefulness of the bootstrap in phylogeny reconstruction have been extensively discussed and debated (e.g., Sanderson, 1989, 1995; Carpenter, 1992, 1996; Zharkikh and Li, 1992a, b; Felsenstein and Kishino, 1993; Hillis and Bull, 1993; Farris et al., 1996). Despite concerns, controversy and confusion over interpretation of bootstrap values, bootstrap analyses have played a prominent role in many phylogenetic studies and likely will remain a key method for assessing nodal support in phylogenetic trees. Given that most systematists are most interested in identifying well-supported groups, the ease of bootstrap analysis using computer packages such as PAUP* (Swofford, 1998) ensures that the bootstrap will continue to be widely used in systematic circles for the foreseeable future.

ACKNOWLEDGMENTS

We thank George Casella, Matt Gitzendanner, Mike Sanderson and an anonymous reviewer for helpful comments on the manuscript. This work was supported in part by NSF Grants DEB-0090283 and PGR-0115684.

REFERENCES

BULL, J. J., CUNNINGHAM, C. W., MOLINEUX, I. J., BADGETT, M. R. and HILLIS, D. M. (1993). Experimental molecular evolution of bacteriophage T7. *Evolution* **47** 993–1007.

CARPENTER, J. M. (1992). Random cladistics. *Cladistics* **8** 147–153.

CARPENTER, J. M. (1996). Uninformative bootstrapping. *Cladistics* **12** 177–181.

CAVENDER, J. A. (1978). Taxonomy with confidence. *Math. Biosci.* **40** 271–280.

CAVENDER, J. A. (1981). Tests of phylogenetic hypotheses under generalized models. *Math. Biosci.* **54** 217–229.

CHASE, M. W., SOLTIS, D. E., OLMSTEAD, R. G., MORGAN, D., LES, D. H., MISHLER, B. D., DUVALL, M. R., PRICE, R. A., HILLS, H. G., QIU, Y.-L., KRON, K. A., RETTIG, J. H., CONTI, E., PALMER, J. D., MANHART, J. R., SYTSMA, K. J., MICHAELS, H. J., KRESS, W. J., KAROL, K. G., CLARK, W. D., HEDRÉN, M., GAUT, B. S., JANSEN, R. K., KIM, K. J., WIMPEE, C. F., SMITH, J. F., FURNIER, G. R., STRAUSS, S. H., XIANG, Q.-Y., PLUNKETT, G. M., SOLTIS, P. S., SWENSEN, S. M., WILLIAMS, S. E., GADEK, P. A., QUINN, C. J., EGUIARTE, L. E., GOLENBERG, E., LEARN, G. H., JR., GRAHAM, S. W., BARRETT, S. C. H., DAYANANDAN, S. and ALBERT, V. A. (1993). Phylogenetics of seed plants: An analysis of nucleotide sequences from the plastid gene *rbcL*. *Annals of the Missouri Botanical Garden* **80** 528–580.

DARWIN, C. (1859). *On the Origin of Species by Means of Natural Selection*. J. Murray, London.

DEBRY, R. W. and OLMSTEAD, R. G. (2000). A simulation study of reduced tree-search effort in bootstrap resampling analysis. *Systematic Biology* **49** 171–179.

DIACONIS, P. and EFRON, B. (1983). Computer-intensive methods in statistics. *Scientific American* **249** 116–130.

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

EFRON, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*. SIAM, Philadelphia.

EFRON, B. (1985). Bootstrap confidence intervals for a class of parametric problems. *Biometrika* **72** 45–58.

EFRON, B. and GONG, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *Amer. Statist.* **37** 36–48.

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.

FAITH, D. P. and CRANSTON, P. S. (1991). Could a cladogram this short have arisen by chance alone? On permutation tests for cladistic structure. *Cladistics* **7** 1–28.

FARRIS, J. S. (1983). The logical basis of phylogenetic analysis. In *Advances in Cladistics* **2** (N. I. Platnick and V. A. Funk, eds.) 7–36. Columbia Univ. Press.

FARRIS, J. S., ALBERT, V. A., KÄLLERSJÖ, M., LIPSCOMB, D. and KLUGE, A. G. (1996). Parsimony jackknifing outperforms neighbor-joining. *Cladistics* **12** 99–124.

FELSENSTEIN, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* **27** 401–410.

FELSENSTEIN, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39** 783–791.

FELSENSTEIN, J. (1988). Phylogenies from molecular sequences: Inference and reliability. *Annual Review of Genetics* **22** 521–565.

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

GRAYBEAL, A. (1998). Is it better to add taxa or characters to a difficult phylogenetic problem? *Systematic Biology* **47** 9–17.

HARSHMAN, J. (1994). The effect of irrelevant characters on bootstrap values. *Systematic Biology* **43** 419–424.

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.

HENNIG, W. (1966). *Phylogenetic Systematics*. Univ. Illinois Press, Urbana.

HILLIS, D. M. (1996). Inferring complex phylogenies. *Nature* **383** 130–131.

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.

HILLIS, D. M. and DIXON, M. T. (1989). Vertebrate phylogeny: Evidence from 28S ribosomal DNA sequences. In *The Hierarchy of Life* (B. Fernholm, K. Bremer and H. Jörnvall, eds.) 355–367. Elsevier, Amsterdam.

HUELSENBECK, J. P. (1995). Performance of phylogenetic methods in simulation. *Systematic Biology* **44** 17–48.

HUELSENBECK, J. P. and CRANDALL, K. A. (1997). Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review of Ecology and Systematics* **28** 437–466.

HUELSENBECK, J. P. and HILLIS, D. M. (1993). Success of phylogenetic methods in the four-taxon case. *Systematic Biology* **42** 247–264.

KÄLLERSJÖ, M., FARRIS, J. S., CHASE, M. W., BREMER, B., FAY, M. F., HUMPHRIES, C. J., PETERSEN, G., SEBERG, O. and BREMER, K. (1998). Simultaneous parsimony jackknife analysis of 2538 *rbcL* DNA sequences reveals support for major clades of green plants, land plants, seed plants, and flowering plants. *Plant Systematics and Evolution* **213** 259–287.

KLUGE, A. G. (1997). Testability and the refutation and corroboration of cladistic hypotheses. *Cladistics* **13** 81–96.

KLUGE, A. G. (1999). The science of phylogenetic systematics: Explanation, prediction, and test. *Cladistics* **15** 429–436.

KLUGE, A. G. and WOLF, A. J. (1993). Cladistics: What's in a word? *Cladistics* **9** 183–199.

LANYON, S. (1985). Detecting internal inconsistencies in distance data. *Systematic Zoology* **34** 397–403.

MILLER, R. G. (1974). The jackknife—a review. *Biometrika* **61** 1–15.

MORT, M. E., SOLTIS, P. S., SOLTIS, D. E. and MABRY, M. (2000). Comparison of three methods for estimating internal support on phylogenetic trees. *Systematic Biology* **49** 160–171.

MUELLER, L. D. and AYALA, F. J. (1982). Estimation and interpretation of genetic distance in empirical studies. *Genetical Research* **40** 127–137.

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

PENNY, D., FOULDS, L. R. and HENDY, M. D. (1982). Testing the theory of evolution by comparing phylogenetic trees constructed from 5 different protein sequences. *Nature* **297** 197–200.

PENNY, D. and HENDY, M. D. (1985). Testing methods of evolutionary tree construction. *Cladistics* **1** 266–278.

PLATNICK, N. I. and GAFFNEY, E. S. (1977). Review of *The Logic of Scientific Discovery* and *Conjectures and Refutations*, by K. R. Popper. *Systematic Zoology* **26** 361–365.

PLATNICK, N. I. and GAFFNEY, E. S. (1978). Evolutionary biology: A Popperian perspective. *Systematic Zoology* **27** 138–141.

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

SANDERSON, M. J. (1989). Confidence limits on phylogenies: The bootstrap revisited. *Cladistics* **5** 113–129.

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

SANDERSON, M. J. and WOJCIECHOWSKI, M. F. (2000). Improved bootstrap confidence limits in large-scale phylogenies, with an example from Neo-Astragalus (Leguminosae). *Systematic Biology* **49** 671–685.

SAVOLAINEN, V., CHASE, M. W., MORTON, C. M., HOOT, S. B., SOLTIS, D. E., BAYER, C., FAY, M. F., DE BRUIJN, A., SULLIVAN, S. and QIU, Y.-L. (2000). Phylogenetics of flowering plants based upon a combined analysis of plastid *atpB* and *rbcL* gene sequences. *Systematic Biology* **49** 306–362.

SOLTIS, D. E., SOLTIS, P. S., MORT, M. E., CHASE, M. W., SAVOLAINEN, V., HOOT, S. B. and MORTON, C. M. (1998). Inferring complex phylogenies using parsimony: An empirical

approach using three large DNA data sets for angiosperms. *Systematic Biology* **47** 32–42.

SOLTIS, D. E., SOLTIS, P. S., CHASE, M. W., MORT, M. E., ALBACH, D. C., ZANIS, M., SAVOLAINEN, V., HAHN, W. H., HOOT, S. B., FAY, M. F., AXTELL, M., SWENSEN, S. M., PRINCE, L. M., KRESS, W. J., NIXON, K. C. and FARRIS, J. S. (2000). Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. *Botanical Journal of the Linnean Society* **133** 381–461.

SOLTIS, P. S. and NOVAK, S. J. (1997). Polyphyly of the tuberous Lomatiums (Apiaceae): cpDNA evidence for morphological convergence. *Systematic Botany* **22** 99–112.

SOLTIS, P. S., SOLTIS, D. E. and CHASE, M. W. (1999). Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature* **402** 402–404.

SWOFFORD, D. L. (1998). PAUP* 4.0: Phylogenetic analysis using parsimony (and other methods), Beta version 4.0. Sinauer, Sunderland, MA.

TEMPLETON, A. R. (1983). Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. *Evolution* **37** 221–244.

WENDEL, J. F. and ALBERT, V. A. (1992). Phylogenetics of the cotton genus (Gossypium): Character-state weighted parsimony analysis of chloroplast-DNA restriction site data and its systematic and biogeographic implications. *Systematic Botany* **17** 115–143.

WILEY, E. O. (1975). Karl R. Popper, systematics, and classification: A reply to Walter Bock and other evolutionary taxonomists. *Systematic Zoology* **24** 233–243.

ZANIS, M. J., SOLTIS, D. E., SOLTIS, P. S., MATHEWS, S. and DONOGHUE, M. J. (2002). The root of the angiosperms revisited. *Proc. Nat. Acad. Sci. U.S.A.* **99** 6848–6853.

ZHARKIKH, A. and LI, W.-H. (1992a). Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. *Molecular Biology and Evolution* **9** 1119–1147.

ZHARKIKH, A. and LI, W.-H. (1992b). Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. II. Four taxa without a molecular clock. *J. Molecular Evolution* **35** 356–366.

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.