Syst. Biol. 49(4):671–685, 2000

Improved Bootstrap Confidence Limits in Large-Scale Phylogenies, with an Example from Neo-Astragalus (Leguminosae)

MICHAEL J. SANDERSON[1,3] AND MARTIN F. WOJCIECHOWSKI[1,2]

[1]*Section of Evolution and Ecology, University of California, Davis, California 95616, USA*
[2]*Museum of Paleontology and University/Jepson Herbaria, University of California, Berkeley, California 94720, USA*

Abstract.—Phylogenetic analyses of large data sets pose special challenges, including the apparent tendency for the bootstrap support for a clade to decline with increased taxon sampling of that clade. We document this decline in data sets with increasing numbers of taxa in *Astragalus*, the most species-rich angiosperm genus. Support for one subclade, Neo-Astragalus, declined monotonically with increased sampling of taxa inside Neo-Astragalus, irrespective of whether parsimony or neighbor-joining methods were used or of which particular heuristic search algorithm was used (although more stringent algorithms tended to yield higher support). Three possible explanations for this decline were examined, including (1) mistaken assignment of the most recent common ancestor of the taxon sample (and its bootstrap support) with the most recent common ancestor of the clade from which it was sampled; (2) computational limitations of heuristic search strategies; and (3) statistical bias in bootstrap proportions, especially that from random homoplasy distributed among taxa. The best explanation appears to be (3), although computational shortcomings (2) may explain some of the problem. The bootstrap proportion, as currently used in phylogenetic analysis, does not accurately capture the classical notion of confidence assessments on the null hypothesis of nonmonophyly, especially in large data sets. More accurate assessments of confidence as type I error levels (relying on iterated bootstrap methods) remove most of the monotonic decline in confidence with increasing numbers of taxa. [Bootstrap; phylogeny reconstruction; species richness; taxon sampling.]

Astragalus L. is a vast assemblage of >2,500 species and 250+ sections (Lock and Simpson, 1991; Mabberly, 1997) in the angiosperm family Leguminosae (Fabaceae). Distributed mainly in cool arid regions of the Northern Hemisphere and South America, *Astragalus* is especially diverse in southwest Asia (~1,000–1,500 species), the Sino-Himalayan region (500 species), western North America (~400 to 450 species), and along the Andes in South America (100–150 species). It is also diverse in Mediterranean climates along the west coasts of North and South America and in Europe. Many *Astragalus* species are narrow endemics, often preferentially distributed in marginal habitats or associated with specialized substrates. However, many temperate herbaceous angiosperm genera have similar ecological and biogeographic characteristics without displaying the species-richness of *Astragalus*. *Astragalus*, therefore, provides an opportunity for studying evolutionary processes on a nearly unique scale. At the same time, it represents a challenge to the prevailing taxon sampling strategy in phylogenetics, which, almost of necessity, relies on sampling a modest number of taxa (Kim, 1998).

[3]Author for correspondence. E-mail: mjsanderson@ucdavis.edu

An important clue to phylogenetic relationships in *Astragalus* is a close correlation between chromosome number and geographic distribution. Of the ~2,000 Old World species, all but 22 have euploid numbers based on $n = 8$ (among those assayed). Of the 500 New World species, all but 13 have numbers in an aneuploid series, with $n = 11$–15. Previous molecular phylogenetic analyses based on fairly small taxon samples supported the monophyly of the almost exclusively New World aneuploid species. This group, referred to as "Neo-Astragalus", is nested well within Old World euploid taxa. Wojciechowski et al. (1993) sequenced nuclear rDNA internal transcribed spacer (ITS) regions for 14 aneuploid and 12 euploid *Astragalus* and found bootstrap support for Neo-Astragalus at the 88% level. Corroboration was found in independent chloroplast restriction fragment length polymorphism data sets (Sanderson and Doyle, 1993; Liston and Wheeler, 1994).

Recently we completed a much more intensive molecular phylogenetic study in *Astragalus*, increasing the taxon sampling by fivefold (Wojciechowski et al., 1999). Once again, Neo-Astragalus is a clade, but the support—as measured by bootstrap proportions (BP; Felsenstein, 1985)—declined to 64–73%, depending on the search strategy.

This decline is more than can be explained by the binomial sampling variance of BP that stems from the finite number of bootstrap replicates (Hedges, 1992) and raises several issues about the assessment of confidence in phylogenetic analyses of large sets of taxa. The purpose of this paper is to reevaluate the monophyly of this keystone group within *Astragalus* in light of a much more intense taxon sample, and weigh its support in the context of the possible methodological problems that arise in the analysis of large data sets. Basically, we ask whether the apparent decline in support is real or an artifact of measures of support or strategies of taxon sampling. We regard our preliminary finding as an important challenge to our long-term strategy for tackling sampling in *Astragalus*—that is, identifying strongly supported clades in limited taxon samples and then designing future sampling strategies around those results. Lecointre et al. (1993; see also Poe, 1998) suggested that taxon sampling can influence the amounts of support for a clade; if their warnings hold even for strongly supported but sparsely sampled clades, then the phylogenetic analysis of species-rich clades may be even more difficult than previously imagined.

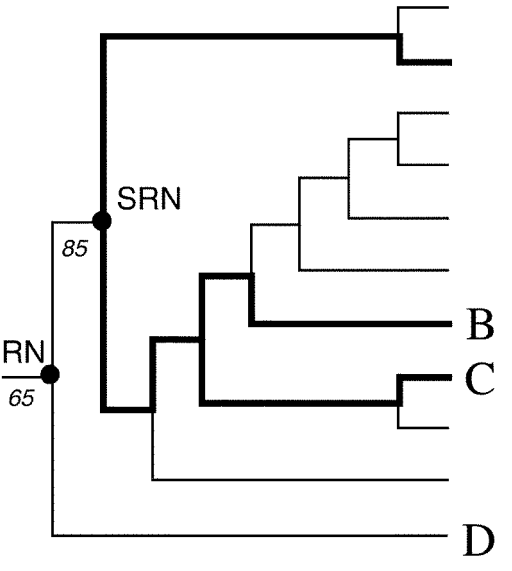


FIGURE 1. Explanation 1: Taxon sampling and node recognition. The most recent common ancestor of a small sample of taxa, {A, B, C}, here labeled SRN (sample root node), may be different from the most recent common ancestor of the larger clade it is thought to represent, labeled RN (root node). If the smaller clade has greater bootstrap support than the larger clade, support will be seen to decline once additional sampling (say, by adding taxon D) begins to span the true common ancestor of the larger clade.

We consider three distinct explanations for the observed difference in bootstrap support between our sparse and more complete taxon samples.

Explanation 1: The result is real but relevant to the wrong node in the tree.—The support for the smaller, sparsely sampled clade is high, but that clade does not correspond to Neo-Astragalus in the sense entailed by a larger sample of taxa. Instead, what was previously considered "Neo-Astragalus" is nested within Neo-Astragalus, and Neo-Astragalus as a whole is poorly supported (Fig. 1). The most recent common ancestor of a small sample of taxa may not be the same as that for the larger sample from which it is drawn, and obviously their supports may differ. Our earlier finding (Wojciechowski et al., 1993) was contingent on the particular sample of species chosen, and presumably a different sample, spanning the root node of Neo-Astragalus, would have had less support.

Explanation 2: The result is an artifact of a progressive deterioration in the ability of heuristic tree search algorithms to find the optimal tree as the number of taxa increases.—For combinatorial optimization algorithms such as parsimony, likelihood, and certain distance methods, no efficient methods are known that can guarantee an optimal solution when the number of taxa exceeds ~8–20 taxa (Swofford et al., 1996). A group that is monophyletic in the most-parsimonious tree might not be monophyletic in suboptimal trees, including those that are derived from bootstrapped data sets from the original data matrix. If this happens, many more instances of nonmonophyly will be discovered among the bootstrap replicates than would be found had the optimal trees been uncovered. After all, there are many more ways for a group *not* to be monophyletic than to be monophyletic. If this pattern gets worse with increasing numbers of taxa, a decline in support would result.

Explanation 3: The result is an artifact of the so-called bias in bootstrap proportions that has been discussed in the theoretical literature.—Bootstrap support for a clade is often less than the probability that the clade is found on the true tree (Zharkikh and Li, 1992a, 1995; Hillis and Bull, 1993; Felsenstein and Kishino, 1993; Efron et al., 1996; Newton, 1996). This effect

may become more pronounced as the number of taxa sampled increases. Although initial discussions concluded that the bias was in the direction of underestimating true accuracy (as long as support was >50%), Efron et al. (1996) cited examples in which this bias can go in either direction. As yet, investigations of this bias have not considered the effect of sampling intensity.

This bias may well depend on the number of taxa in a clade. As sampling intensity within a clade increases, eventually at least one taxon may be discovered that has undergone a simultaneous reversal in almost all of the characters that were synapomorphies of the clade (Fig. 2). The probability of this happening depends on the number of synapomorphies supporting the clade, the average rates of evolution (probabilities of homoplasy) in those characters, and possibly the pattern of support within the clade itself. Sampling of such a rogue taxon would immediately disrupt the bootstrap support for the clade, because many bootstrap replicates would sample the reversed characters but omit the one lingering synapomorphy that places the rogue within the group. Unfortunately, although this argument is couched in terms of a single problematic taxon, the effect might be diffused over many taxa. Thus, even a systematic search for, and elimination of, one or a few rogue taxa might not alleviate the problem.

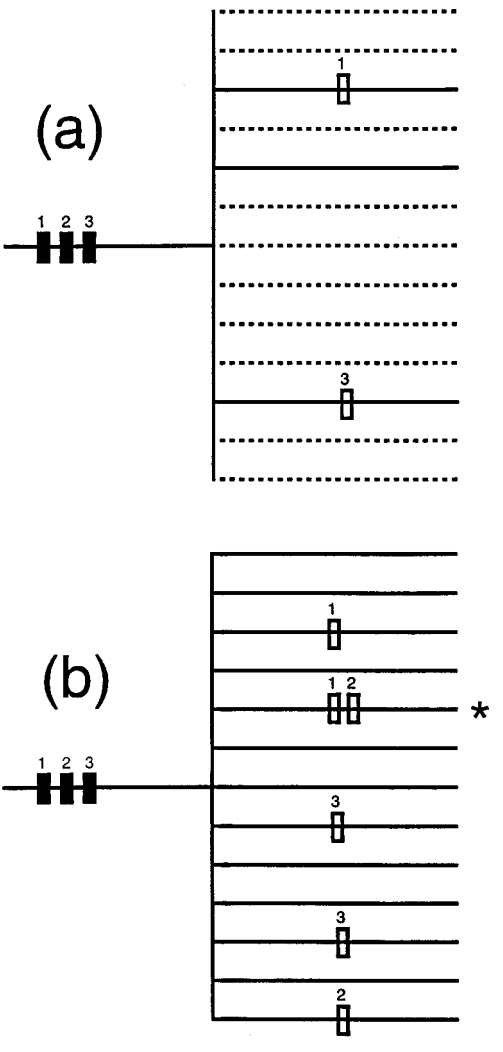


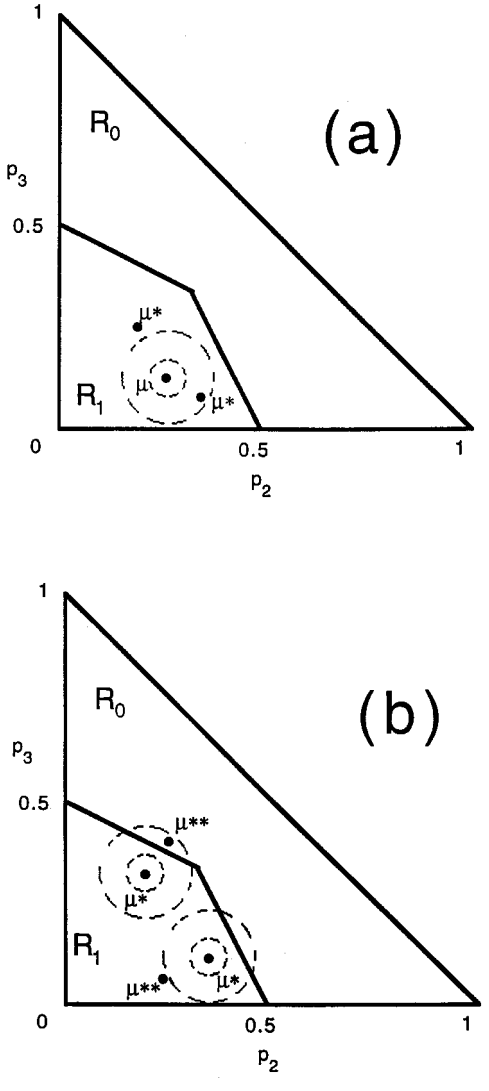
FIGURE 2. Explanation 3: Taxon sampling, homoplasy, and random reversals. Taxon sampling from a large, rapidly diversifying, clade may eventually uncover a taxon that has undergone reversals (open tick-marks) in nearly all of the synapomorphies (sold tick-marks) that ordinarily would have suggested its membership in that clade. Because bootstrap proportions are tied to the precise taxon membership of a clade, this means that even if just one taxon periodically "jumps out" of the clade in some bootstrap replicates, the support for the entire clade will diminish. Solid lines represent lineages leading to sampled taxa; dashed lines are lineages leading to unsampled taxa. (a) Only one of the three indicated synapomorphies of the clade is reversed in each of two sampled lineages. (b) In a larger sample of taxa, one taxon has lost two of the synapomorphies (lineage marked by asterisk). This taxon is highly likely to jump out of the clade and hence lower the clade's support in bootstrap replicates. Other apomorphies of these lineages are not shown. Note that a complementary pattern in which outgroups with random homoplasies "jump into" the clade can also occur as taxon sampling intensifies.

Framework for Investigating the Properties of Bootstrapping

Efron et al. (1996) proposed a useful geometric framework for examining the statistical properties of bootstrapping in phylogenetics, relating it to the so-called "problem of regions" (Efron and Tibshirani, 1996). We have found it useful to combine their framework with a specific ("*K*-alternative") model of phylogenetic data proposed by Zharkikh and Li (1995). Together, these two elements provide much-needed intuition about the problem at hand.

Consider just two phylogenetic hypotheses regarding a prespecified set of taxa (Fig. 3). Hypothesis H^0 is that the taxa are not a clade; H^1 is that the taxa are a clade. A character "supports" a clade if it has the derived state for all and only the taxa in that clade. A character "class" is a set of one or more

characters that have exactly the same distribution of states among taxa and hence support the same clade. The *K*-alternative model specifies that every data matrix consists of exactly K classes of characters, in which one class supports monophyly of the clade of interest, and $K - 1$ classes conflict with it. Under the simplification of parsimony considered by Zharkikh and Li (1995), a data set will support monophyly if and only if there are more characters supporting H^1 than conflicting with it. To be precise, if the proportion of characters in the matrix supporting monophyly is p_1 and the proportions of conflicting characters are $p_2, \ldots, p_K$, then H^1 is supported if and only if $p_1 > \max(p_2, \ldots, p_K)$.



Note that true parsimony does not adhere to this, because a clade can be recovered on the most-parsimonious tree even if no character by itself "supports" it.

To visualize the geometric aspect of the problem, we restrict attention to the case of $K = 3$. Figure 3 shows a frequency diagram in which all possible data matrices can now be arrayed in a triangular region of a plane. No axis for the frequency of p_1 need be shown because it is determined by $1 - p_2 - p_3$. The space is divided into two subregions, R_0 and R_1, corresponding to the data matrices that parsimony suggests support H^0 and H^1, respectively. A point in this space, labeled μ, corresponds to a data matrix with the character frequencies indicated at that point. Associated with the point is also the tree or trees estimated from such data sets. When it is necessary to distinguish trees from data matrices, the data matrix will be denoted by **M.**

To understand the statistical properties of a bootstrap test, we must imagine the outcome of such tests performed repeatedly in

FIGURE 3. Bootstrapping and the "problem of regions" (Efron and Tibshirani, 1996). The space of all data sets can be divided into regions, each of which corresponds to whether it supports some prespecified group as a clade. Any data set in R_1 supports this group as a clade; any data set in R_0 does not. Diagram shows a concrete example in which the axes represent the frequencies of characters in a hypothetical data set space. Characters in character class 1 support the clade and characters in classes 2 and 3 conflict with it. Frequencies of character classes 2 and 3 are p_2 and p_3, respectively (note that $p_1 = 1 - p_2 - p_3$). (a) The point μ, which is in R_1, is the "actual" true frequency of characters in the hypothetical universe of characters from which data sets are drawn. Points μ^* are two observed data sets obtained by some process of sampling from μ. Sampling is illustrated schematically by circular contour lines around μ. "Accuracy" is generally taken to be the proportion of μ^* that fall in R_1. (b) Conventional bootstrap resampling produces another distribution of points around each μ^*. The pseudo-data sets generated are labeled μ^{**}. The fraction of pseudo-data sets falling in R_1 is taken as the bootstrap proportion (BP), which is sometimes viewed as an estimator of accuracy. However, BP themselves have a probability distribution across the samples of μ^*. The discrepancy between BP and accuracy as measured in simulations (Hillis and Bull, 1993) arises because the expected BP (the proportion of μ^{**} values that fall in R_1) will often be less than the accuracy (the proportion of μ^* that fall within R_1). This is because the μ^* that are sampled from μ will often be closer to the boundary than μ is, especially when the average distance between μ^* and μ is large because of sampling error in drawing the data sets from the underlying character universe.

the long run on some sample of data sets. The classical notions of type I and type II error refer to the probability of mistakes being made in repeated statistical experiments. Here the notion is that some underlying universe of character data exists from which a sample–the observed data matrix–has been drawn. In Figure 3 the universe of character data is a point labeled μ in the space, and any sample from it (i.e., any real data set) is labeled μ^*. The statistical properties of a bootstrap test that uses some data set μ^* can be determined by examining its properties over the set of all possible μ^* sampled from μ.

In conventional hypothesis testing, a "test" is a procedure that leads to a choice between hypotheses. In this respect, the bootstrap procedure most widely used in phylogenetics (Felsenstein, 1985), which reports a BP for a clade, is not a test unless it is coupled with a decision rule, such as "reject monophyly for BP <95%". Some of the controversy regarding the properties of the BP has resulted from inattention to the difference between the BP itself and the tests based on it (Felsenstein and Kishino, 1993). For example, the average BP (determined across many μ^*) will often be less than the probability that data sets sampled from μ (the μ^* themselves) support monophyly (see Fig. 3; Zharkikh and Li, 1992a, 1995; Felsenstein and Kishino, 1993; Hillis and Bull, 1993; Efron et al., 1996). This is the genesis of the now often-stated claim that bootstrapping underestimates the true support for a group. However, Felsenstein and Kishino (1993) point out that this phenomenon is not especially relevant if one interprets BP in terms of the type I error of a *test* based on the BP.

Let us construct a hypothesis test based on the bootstrap proportion, called BP(x), where x is a cutoff value. For example, the test BP(80) specifies that a group will be considered monophyletic if its BP exceeds 80%. The type I error, α, of this test is the probability that if H^0 is true, a mistaken inference of monophyly will be made (Fig. 4). Felsenstein and Kishino (1993) suggest that a BP of $1 - x$ should be interpreted as a type I error of x. Even if the true type I error is less than this (as was the case in examples discussed in their paper—but that need not always be true—see Efron et al., 1996), the test is conservative. However, even a conservative test can be bad if it causes one to accept the null

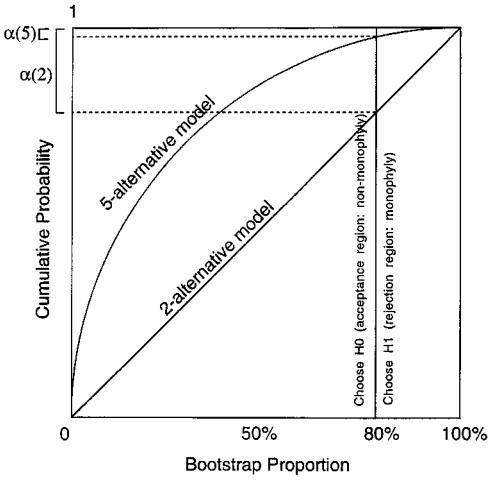


FIGURE 4. Statistical distribution of BP under a theoretical model described by Zharkikh and Li (1995). Graphs show the cumulative distribution function of BP, which is a distribution over many repeatedly drawn random data sets. As the number of character classes that support alternative and mutually exclusive topologies increases (e.g., from 2 to 5), the distribution of BP shifts to the left. The mean BP decreases, but it also causes the α-level for a given test, such as BP(80), to decline as well. Thus a consideration of these two notions of confidence can be seemingly contradictory (Felsenstein and Kishino, 1993). The α-level for a k-alternative model is indicated as $\alpha(k)$. The theoretical model assumes that the point lies at the point where all character classes have equal sampling probability (i.e., at the corner in the middle of the triangle in Fig. 3). Increased taxon sampling may effectively lead to an increase in the number of alternatives, causing a downward shift in BP.

hypothesis incorrectly too often. This motivates investigations of the actual type I error of the BP(x) test, for which one needs to know the probability distribution of BP(x) over repeated samples of real data sets, μ^*, from μ. Zharkikh and Li (1995) showed that as K increases in the K-alternative model, the cumulative probability distribution of BP shifts to the left, such that the expected BP declines. Consequently, α also declines for the same cutoff value of the BP test, meaning the type I error for the test BP(x) is $< 1 - x$, rather than being equal to it (Fig. 4). Other behavior is possible in more realistic models (Efron et al., 1996).

To understand what is happening to conventional BP as taxon sampling is increased, it will therefore be necessary to have a better estimate of type I error than that provided by the quantity $1 - x$. This improved estimate is described below under Corrected Bootstrap Confidence Limits.

Materials and Methods

Overview of Methodology

Throughout this paper, we use the terms "data set size", or "sample size", or "sampling intensity" to refer to the number of taxa sampled—not the number of characters, which remains fixed throughout. We used several computational methodologies to evaluate the relative merits of the three explanations outlined above. First, we examined BP in replicate random taxon samples (from the new data set) of the size of our earlier published work (14 Neo-Astragalus and 12 other *Astragalus*; Wojciechowski et al., 1993). This tested whether or not the high bootstrap support for Neo-Astragalus in that data set was peculiar to that set of taxa or was characteristic of other samples of that size. Next, we examined bootstrap support in taxon samples of progressively larger size to determine whether support declined more or less monotonically with increasing sample size or dropped abruptly when reaching some threshold. Third, we considered several different heuristic search strategies in parsimony and neighbor-joining analyses. Finding the most-parsimonious tree is an NP-hard problem (Graham and Foulds, 1982), meaning no known "efficient" or polynomial-time algorithm is known to give an exact solution (Garey and Johnson, 1979). Heuristic parsimony searches and neighbor-joining can be considered polynomial–time approximations for finding the most-parsimonious trees. Comparison of algorithms of different stringency may therefore shed some light on possible artifacts stemming from failure of algorithms to find optimal trees. Finally, we implemented the iterated bootstrap procedure of Efron et al. (1996), which should give more accurate confidence limits on trees, to determine whether the iterated procedure is less sensitive to sampling intensity than is conventional bootstrapping.

Details of the new, expanded ITS data set and how it was obtained are discussed elsewhere (Wojciechowski et al., 1999), along with phylogenetic and systematic implications. Complete aligned sequences in the form of a NEXUS file (Maddison et al., 1997) can be obtained at our *Astragalus* website (http://loco.ucdavis.edu/astragalus/astragalus_home.htm, and http://www.utexas.edu/ftp/depts/systbiol/). The file includes 140 sequences, of which 116 are species of *Astragalus*, with the remainder belonging to other Astragalean genera and one more distant outgroup, *Caragana*. Of the *Astragalus* sampled, 79 are presumptive members of Neo-Astragalus based on their aneuploid chromosome numbers and New World distribution, and the remaining 37 are Eurasian species.

Phylogenetic Methods

Phylogenies were reconstructed by using maximum parsimony and neighbor-joining methods (Swofford et al., 1996). A set of heuristic search strategies varying in stringency was used with the following PAUP* options: (1) FAST-1 (ADDSEQ = RANDOM, NREPS = 1, SWAP = NONE, MAXTREES = 1; (2) FAST-10 (ADDSEQ = RANDOM, NREPS = 10, SWAP = NONE, MAXTREES = 10; corresponding to the search options in so-called "fast" bootstrapping); (3) SIMPLE-0 (ADDSEQ = SIMPLE, SWAP = NONE, MAXTREES = 1); (4) SIMPLE-1 (ADDSEQ = SIMPLE, SWAP = TBR, MAXTREES = 1); (5) NJ (neighbor-joining with Kimura two-parameter distances). These heuristic options are less stringent than many reported in the literature in analyses of smaller data sets, which routinely include TBR swapping with MAXTREES set to ≥100. However, because the sampling experiments described below entailed >300,000 separate heuristic searches, considerable care had to be exercised in selecting strategies that would finish in reasonable run times (in seconds to minutes rather than hours). Sampling strategy 4, which has limited TBR swapping on only one tree, still seems to capture much of the ability of swapping to find trees (R. Olmstead, pers. comm.).

In NJ analyses, pairwise distances were calculated by using the Kimura two-parameter model (Li, 1997). Relatively simple distance corrections have an advantage of lower variance than those with more parameters (Kumar et al., 1993; Zharkikh, 1994).

Analysis of BPs

Bootstrap support for Neo-Astragalus was examined in the full 140-taxon ITS data set with an analysis of 1,000 replicates for each of the five algorithms described above. The influence of taxon sampling intensity was investigated by examining BP in data sets

randomly drawn from the 140-taxon data set. Taxon samples of sizes 13, 19, 25, 33, 53, and 73 were examined (chosen for programming convenience). At each size, 100 random taxon samples were drawn (without replacement), and the BP (based on 100 replicates) was calculated for each of the algorithms described above. The set of randomly chosen taxa at a given sampling intensity was kept constant across all algorithms.

Sampling was "stratified" to focus on the putative clade Neo-Astragalus and its relationships to other *Astragalus*. The same three outgroups were included in all taxon samples (*Caragana arborescens*, *Colutea arborescens*, and *Oxytropis campestris*). For all sample sizes except that with 73 taxa, equal numbers of taxa were drawn from Eurasian *Astragalus* and from New World aneuploid (Neo-Astragalus) taxa; for example, for 53 taxa, the sampling was 3 + 25 + 25 from outgroups, euploids, and aneuploids, respectively. For samples of size 73, the number of aneuploids was increased to 45, but the number of euploids was held at 25. Because the original 140-taxon data set has more than twice as many New World aneuploids as Eurasian taxa (79 vs. 37), for the largest subsample it seemed reasonable to sample more intensively from Neo-Astragalus than to include all euploids in any given sample (and hence remove taxon sampling as a factor among the euploids).

Random taxon samples were generated by using the program r8s, available from M.J.S. at http://loco.ucdavis.edu/r8s/r8s.html. This program generates stratified or unstratified sets of taxon identification numbers corresponding to those in a NEXUS-formatted file (Maddison et al., 1997). It also generates the appropriate taxon deletion and restoration syntax as PAUP block commands to accomplish the sampling. All of these commands are appended to the original NEXUS file, and the bootstrap tree resulting from each taxon sample is appended to a file (using the SAVEBOOTP = BRLENS options to store the confidence estimate for every node). Next, these trees are once more imported into r8s, which summarizes the bootstrap support for the node that consists of the common ancestor of all (sampled) taxa of interest—in this case, the putative members of Neo-Astragalus. This is done by supplying r8s with a list of all the taxa from the data matrix that are in the aneuploid group, assigning a node name to the most recent common ancestor of whatever aneuploids are actually present in that taxon sample on that tree, and extracting the bootstrap support for that named clade from the stored treefile. Note that the named node is either the most recent common ancestor of all aneuploids (assuming monophyly) or a node descended from that node; in either case, the node would be considered the most recent common ancestor of aneuploids in that sample of taxa (see Fig. 1).

In all, five algorithms were examined at six sampling intensities for 100 random taxon samples (at 100 bootstrap replicates each), for a total of $5 \times 6 \times 100 \times 100 = 300{,}000$ heuristic searches. In addition, another 5 (algorithms) × 1,000 (bootstrap replicates) = 5,000 searches were undertaken with the full 140-taxon data matrix.

Corrected Bootstrap Confidence Limits

Several procedures have been proposed for improving the accuracy of the bootstrap proportion as an estimator of $1 - \alpha$, where α is the type I error. Rodrigo (1993), Zharkikh and Li (1995), and Efron et al. (1996) proposed variations that rely on multiple rounds of ("iterated") bootstrapping. The theoretical results from Zharkikh and Li emphasize their simplified *K*-alternative model (see above). Real data sets pose challenges to their simplified analytical results, however, which are outlined in some detail in their paper (1995:54). More complex bootstrap procedures can be invoked, including the "complete-and-partial bootstrap" they themselves proposed. The simplest, however, was proposed by Rodrigo (1993). Imagine taking a data set in the region of the null hypothesis, bootstrapping it, and calculating the BP on monophyly. Repeating this many times would provide a distribution of BP under the null hypothesis (Fig. 4). This could be simulated by an iterated bootstrap procedure (Rodrigo, 1993) in which *N* replicated bootstraps are taken from the original data set and the group is checked for monophyly in each. Some pseudo-data sets will generate trees without the group monophyletic and these are set aside. Then the data sets that were set aside are bootstrapped again *N* times each, which allows the distribution of BP to be estimated.

One problem with this method is that some of the first-round pseudo-data sets in which the group is not a clade may be unusually "bad", in the sense that they may be deep in the parameter space of the null hypothesis and far from the boundary between R_0 and R_1 (see Fig. 5); this would result in a very low estimate of type I error. Efron et al. (1996) outlined a procedure that constructs better null pseudo-data sets and also requires fewer total bootstrap replicates than iterated bootstrapping. This is the method we adopt here. The intuition is illustrated in Figure 6, based on Efron et al. (1996). Under the null hypothesis, the true tree (corresponding to the underlying universe of characters), μ_0, lies in R_0 (point not shown). The observed data set, μ^* however, might lie in either region. Regardless, conventional bootstrapping resamples from μ^*, generating a cloud of points around μ^*. The proportion of these points that fall in R_1, μ^{**}, is the conventional (Felsenstein, 1985) bootstrap estimate of $1 - \alpha$. Efron et al. (1996) suggest that a better bootstrap can be obtained by bootstrapping from pseudo-data sets along the boundary between R_0 and R_1. This "simulates" a process of randomly sampling real data sets from a point truly located along the boundary. Efron et al. (1996) assess the curvature of the boundary, which is used in an analytical correction formula to estimate the magnitude of the shifted bootstrap distribution.

The details of this are as follows. In the first round of bootstrapping, an observed data matrix, **M*** (which corresponds to the point, μ^*, in Fig. 6), is bootstrapped N times, which generates a set of pseudo-data sets {**M****}. Some fraction, BP, of these will have the group monophyletic—this is the conventional BP. The remaining fractions, which do not have the group monophyletic, are pulled aside for further analysis. Each is used to construct a pseudo-data set that is on or very close to the boundary between R_0 and R_1. Each pseudo-data set is constructed as a

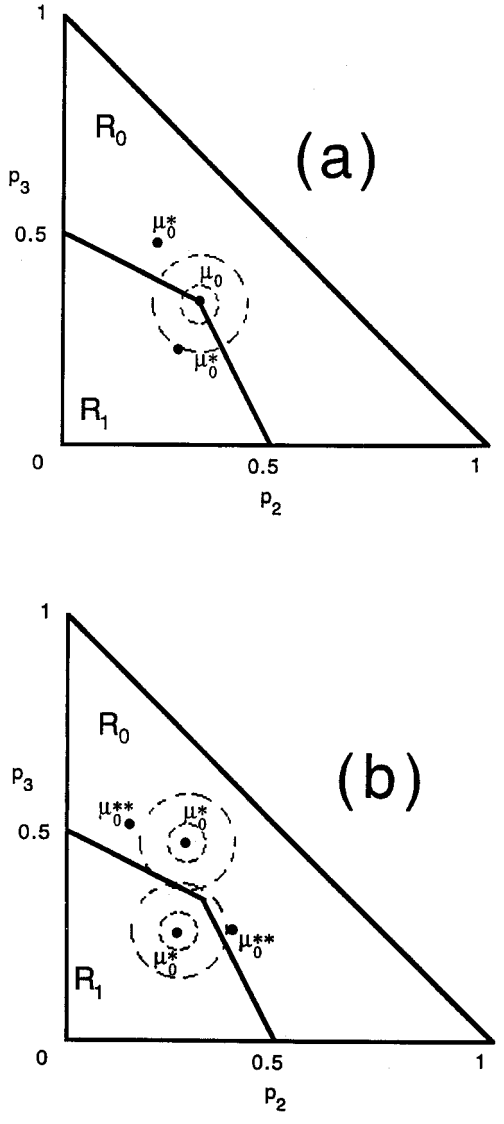


FIGURE 5. Estimation of type I error in the problem of regions. Here type I error is the probability of mistakenly inferring that a group is a clade—belongs in R_1—given that it is not a clade. The calculation of type I error is always in the context of a test or decision rule, such as BP(95): "accept monophyly if the BP exceeds 95%". We wish to know how often that rule will lead us astray when the group actually is not a clade. (a) The underlying character universe is represented by the point μ_0 (analogous to μ in Fig. 3). Unlike the case in Figure 3, in which this point was in R_1, here we choose a point in the region of the null hypothesis that is as close to the boundary as possible, to obtain the maximum value (worst case) of type I error. Data sets sampled from this point are indicated by points labeled μ_0^* (two are shown). (b) Bootstrap proportions are calculated for an observed data set, μ_0^*, by generating pseudo-data sets labeled μ_0^{**}. A brute force way to obtain the type I error would be to look at many real data sets, μ_0^*, and see how often the BP test described above leads to an acceptance of monophyly. The cumulative distribution of BP over many μ_0^* would indicate whether type I error was close to 5% (i.e., 1 - 0.95 for the BP(95) test) or not. Note that the curvature of the boundary inward toward R_1 means that >50% of sampled μ_0^* characters lie in R_0 and these sample data sets will tend to have lower a BP, because a majority of their respective distributions of pseudo-data sets will tend to lie within R_0.

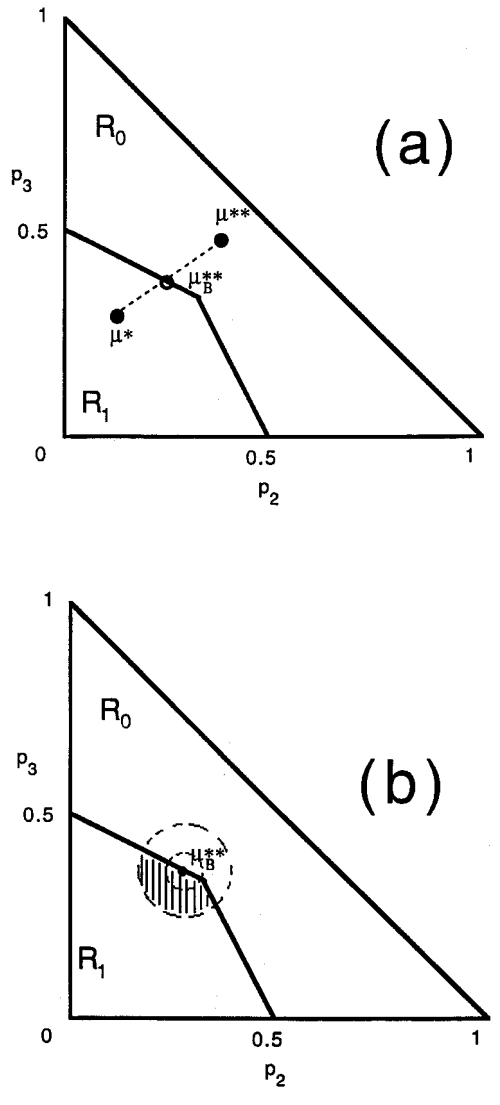


FIGURE 6. Estimation of type I error by an iterated bootstrap procedure. Instead of obtaining repeated samples of real data sets as in Fig. 5, Efron et al. (1996) suggested using a modified bootstrap method to estimate the true type I error. (a) The first step is to obtain one or more pseudo-data sets (μ_B**) along the boundary between the two regions. This is done be interpolating between the original data set, μ*, and one of the resampled pseudo-data sets that happened to land in R_0, μ**. (b) Bootstrapping around μ_B** can then give an estimate of the curvature of the boundary. It is the curvature that explains the shift in type I error rates away from first-order predictions. The deviation from 50% of the proportion of pseudo-data sets, μ_B**, that land in R_1 is an indication of curvature. In the figure, the proportion is <50% because the boundary curves in toward R_1. The curvature is used in correction formulas to obtain a better estimate of type I error (see text). Efron et al. (1996) recommend sampling the curvature at many points along the boundary.

weighted combination of the original data set (in which the clade was supported) and the bootstrapped data set (in which the clade was not supported). Thus, a new matrix $\mathbf{M_B}^{**} = \omega\mathbf{M}^{**} + (1 - \omega)\mathbf{M}^{*}$ is constructed by finding ω (with $0 \leq \omega \leq 1$) such that $\mathbf{M_B}^{**}$ sits right on the border between R_0 and R_1 (Fig. 6). The actual linear combination is implemented by concatenating the two data matrices, $\mathbf{M}^{**}$ and $\mathbf{M}^{*}$, and assigning a weight, either ω or $1 - \omega$, to every character in $\mathbf{M}^{**}$ or $\mathbf{M}^{*}$, respectively. Weights were treated as weights—not counts—in PAUP* with an equal probability of being included or not in a bootstrap replicate. The constant, ω, is found by a simple linear search. For a set of evenly spaced values of ω on [0,1], a tree search is implemented by using $\mathbf{M_B}^{**}$, and the value of ω selected is that at which the focal group switches from being monophyletic to not. To avoid slight biases related to the discreteness of the values of ω, the two values of ω that bracket the value where the switch occurs are randomly selected with equal probability. Efron et al. (1996) describe a more efficient binary search strategy, (but we did not use it because the weighting function was not always monotonic (see below).

Each of the boundary matrices, $\{\mathbf{M_B}^{**}\}$, is then subjected to a second round of bootstrap analysis, determining what proportion of replicates has the group monophyletic, and the results are averaged across the matrices. This average is referred to as BP_0. If the boundary between regions is flat, BP_0 is expected to be 50%, and standard BP values provide a reasonable estimate of $1 - \alpha$ for the null hypothesis. As BP_0 deviates from 50%, indicating a curved boundary, the value of BP, BP, must be corrected for this curvature. The shift in the distribution in the BP away from the straight line shown in Figure 4 depends on the extent of this curvature. For multinomial sampling, Efron et al. (1996) suggest a correction factor that allows $BP_C = 1 - \alpha$, the corrected BP, to be determined from BP and BP_0. The formula is written in terms of standard normal deviates; therefore, for example, $z_0 = z(BP_0) = \Phi^{-1}(BP_0)$, and $z_{BP} = z(BP) = \Phi^{-1}(BP)$, where $\Phi$is the standard normal distribution function. Then, the corrected z-score is approximately as follows:

$$z_C = \frac{z_{BP} - z_0}{1 + a(z_{BP} - z_0)} - z_0$$

where a is a constant that depends on the direction of the boundary from the original data set (Efron, 1987), which is calculated from each boundary matrix as described in Efron et al. (1996:7089) and then averaged across all matrices. In our data, a is negligible, and the correction formula reduces to $z_C = z_{BP} - 2z_0$. When the corrected z_C is obtained, the corrected BP_C (or $\alpha = 1 - BP_C$) is determined simply by converting the z-score back to a percentile.

Software Implementation and Distribution

Efron et al.'s (1996) procedure was implemented in software available from M.J.S.; it requires integration of three separate programs. At the core is the UNIX version of PAUP* vers. 4.0 (Swofford, pers. comm.). Because PAUP* does not implement certain features needed, such as multinomial sampling from a given frequency distribution, or random taxon sampling, these functions were coded as ANSI C modules in the software package, r8s (M.J.S.). Finally, a UNIX C-shell script was written to generate necessary NEXUS files on the fly for the two rounds of bootstrapping needed, to oversee runs of the r8s program, and to pass results to and from different routines.

Because this procedure is extremely computer-intensive, limited analyses were performed at different sampling intensities (Table 1). From 10 to 20 random taxon replicates were done for 16, 22, 30, 40, 50, and 60 taxa. In each, 200 first-round replicates were followed by boundary searches involving 25 evenly spaced points on the interval [0,1]. Second round analyses consisted of 100 replicates for every boundary matrix. For the full data set, 500 first-round replicates were used.

TABLE 1. Bootstrap corrections for Neo-Astragalus at different sampling intensities and in the full data set. All values (except for the full data set) are means of 10 random taxon samples in which the complete correction procedure was implemented. BP, observed bootstrap proportion; BP_0, second round bootstrap proportion; z_{BP}, z-score for BP; z_0, z-score for BP_0; z_C, z-score given by the correction formula (see text); BP_C, the corrected percentile calculated from z_C.

Number of taxa	BP	BP_0	z_{BP}	z_0	z_C	BP_C
16	0.92	0.39	1.44	−0.27	1.98	0.98
22	0.89	0.40	1.23	−0.25	1.73	0.96
30	0.86	0.40	1.09	−0.24	1.57	0.94
40	0.83	0.36	0.97	−0.36	1.69	0.95
50	0.80	0.32	0.84	−0.46	1.76	0.96
60	0.80	0.33	0.84	−0.43	1.70	0.95
140 (full)	0.67	0.29	0.44	−0.55	1.55	0.93

Efron et al. (1996) did not provide details on the implementation of their method, and the molecular data set they examined had only 11 sequences. Our implementation revealed several practical problems that have to be handled in larger or more complex data sets. Several of these were discovered by testing theoretical predictions based on Zharkikh and Li's (1995) formulation of the K-alternative model. First, multiple optimal trees can cause biases in the estimation of BP_0. If the heuristic search strategy selected allows multiple equally parsimonious trees, and if the implementation of Efron et al.'s method counts clades as present only if they are found in all equally parsimonious trees (or, equivalently, in their strict consensus), then the estimate of BP_0 will be biased downward, because many data sets on the boundary will generate pseudo-data sets that will produce sets of conflicting equally parsimonious trees. If we count the number of replicates in which these trees have the focal clade, BP_0 will be $<50\%$ because many trees will have unresolved polytomies in their consensuses. One way around this problem is either to give individual trees fractional weights (which is done in PAUP's bootstrap procedure but is difficult to implement manually) or to keep only one tree from each search.

Second, certain search strategies are inherently biased in pernicious ways. This was also discovered by noting a departure from theoretical expectations under the equiprobable K-alternative model. In particular, nonrandom addition sequences without branch swapping and without multiple equally parsimonious trees tended to be biased either for or against certain clades in model data sets as a result of the order of taxon input. This was sufficient to increase or decrease estimates of BP_0 by 10–20%. The solution was either to include branch swapping or to use random addition sequences.

Finally, some random addition search strategies (without branch swapping) did not result in an accurate estimate of the boundary matrix, because their stochastic nature caused the focal clade to appear and disappear sporadically as the weighting parameter, ω, was tuned from 0 to 1. This would not have been noted in Efron et al.'s binary search; it was discovered by canvasing

the whole interval in an exhaustive search. The solution to this problem was to include branch swapping with random addition sequences, although this did not completely eliminate the noise. However, branch swapping in the largest data sets simply took too much time, so a compromise was to increase the number of first-round replicates, screen the boundary searches for nonsmooth behavior, omit those replicates, and then proceed to the second round. Overall, the optimal combination of search parameters (given the constraints of computer time!) was ADDSEQ = RANDOM, NREPS = 10, SWAP = TBR/NONE (depending on data set size), MAXTREES = 10.

RESULTS

Complete 140-Taxon Data Set

Neo-Astragalus is a clade in all most-parsimonious trees for all algorithms and is also a clade in the neighbor-joining analysis. Figure 7 shows results from a SIMPLE-0 search. Bootstrap support ranged from 64% to 73%, depending on algorithm (Table 2). SIMPLE-1 gave the most support; FAST-1 the least. Other clades on the tree are supported at higher levels, including the next node out, which is supported at 92%, and the bulk of the entire genus *Astragalus*, supported at 86%. Little resolution within Neo-Astragalus is evident, except for two clades, and those each contain both North and South American species. For more discussion of the systematic implications of these results, see Wojciechowski et al. (1999).

Taxon Subsampling

We focus on bootstrap support for Neo-Astragalus. Figure 8 shows support for this clade as a function of taxon sample size and search algorithm. Bootstrap support for Neo-Astragalus declines monotonically as sample size increases, irrespective of algorithm. Algorithm SIMPLE-1 (the algorithm with branch swapping) has the most support and generally either NJ or FAST-1 has the least. Support drops from between 89–95% in the smallest samples to between 70–80% in the largest subsample (73 taxa), which is about half the size of the full data set. The data points at 25 taxa reflect the approximate sampling intensity of our earlier study, which had 26 species of *Astragalus* and a support level of 88%, comparable with that found here.

TABLE 2. Conventional bootstrap proportions for Neo-Astragalus as a function of search algorithm in full 140-taxon data matrix (1,000 replicates). For description of algorithms, see text.

Algorithm	Bootstrap proportion (*BP*)
FAST-1	0.64
FAST-10	0.70
SIMPLE-0	0.65
SIMPLE-1	0.73
NJ (+K2P)	0.66

Corrected Bootstrap Proportions

The magnitude of the Efron et al. (1996) correction factor depends on the intensity of taxon sampling (Table 1). The value of BP_0 moves farther away from 50% as the number of taxa increases, meaning that z_0 becomes more and more negative and the quantity $z_{BP} - 2z_0$ gets larger; this latter increases the estimate of the true confidence value, thereby overcoming most of its apparent decline. Corrected bootstrap support for Neo-Astragalus in the full data set was 93%, which is somewhat less than corrected values for small subsets of taxa but is much greater than the uncorrected value of 67%.

DISCUSSION

These results argue for the rejection of explanation 1—that the support inferred for Neo-Astragalus in our earlier study (Wojciechowski et al., 1993) was contingent on the particular sample of taxa chosen. In 100 random samples of 25 taxa (about the same size as our 1993 study), the average bootstrap support for Neo-Astragalus was 87% for a "quick" search strategy of SIMPLE-1 (compared with 88% in the original study). Support is probably slightly greater for more exhaustive searches, such as those performed in our earlier paper. Variability in the support for this clade did exist among the taxon samples. Many samples had support as low as 80%; others were as high as 100%. Although some portion of this difference is a result of the simple binomial sampling variance, reflecting the use of only 100 bootstrap replicates (e.g., a BP of 90% has an expected 95% confidence interval of $\pm 6\%$ for 100 replicates), the rest undoubtedly indicates a real variability in the lineages sampled. Nonetheless, the *expected* support for a small sample of taxa is about what we observed for the

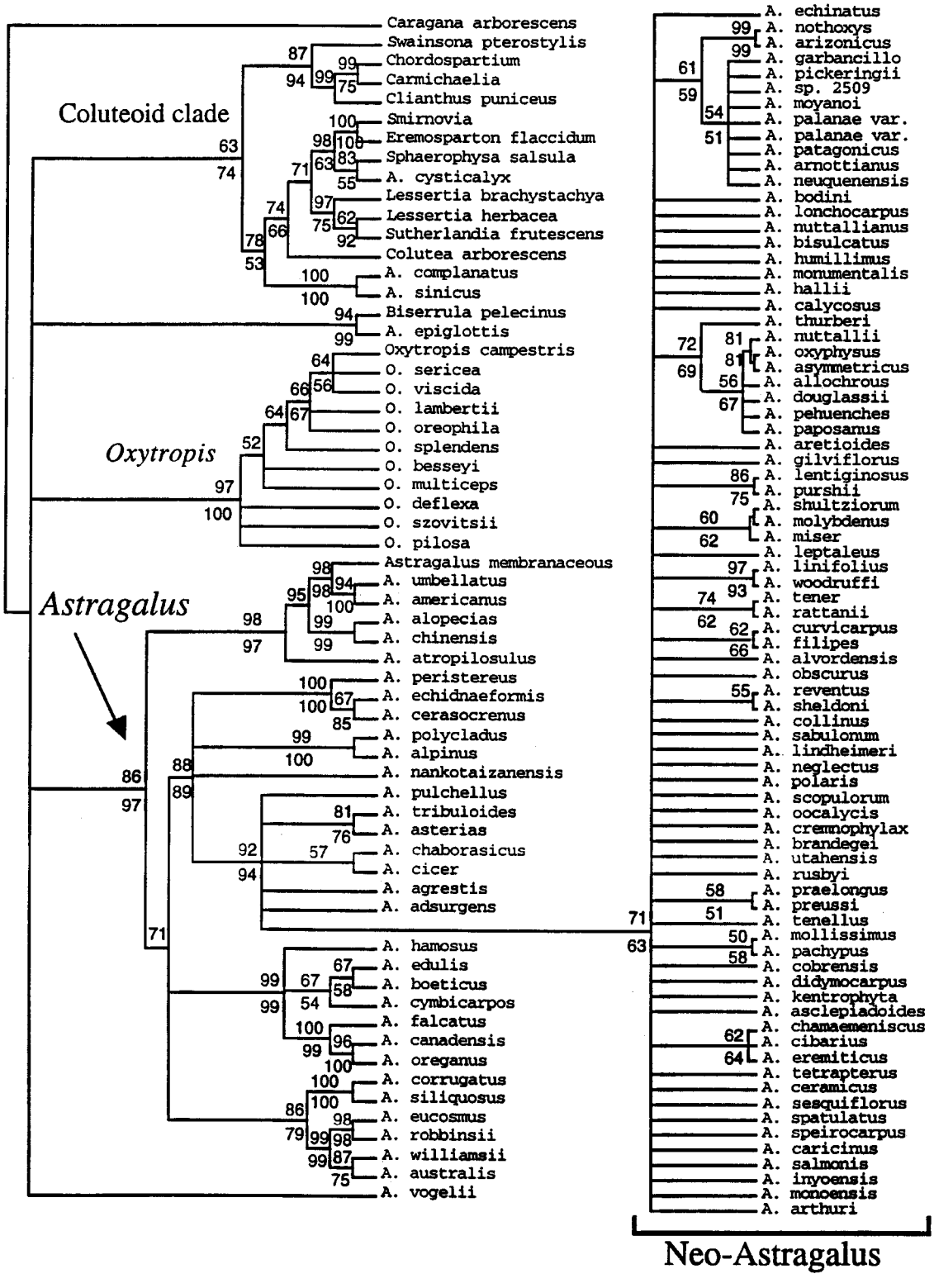


FIGURE 7. Majority rule bootstrap consensus tree based on maximum parsimony analysis of 140-taxon ITS data set (500 replicates, uncorrected BP derived by using SIMPLE-0 search strategy are shown above nodes; uncorrected BP values found by NJ search strategy are shown below nodes (Wojciechowski et al., 1999).

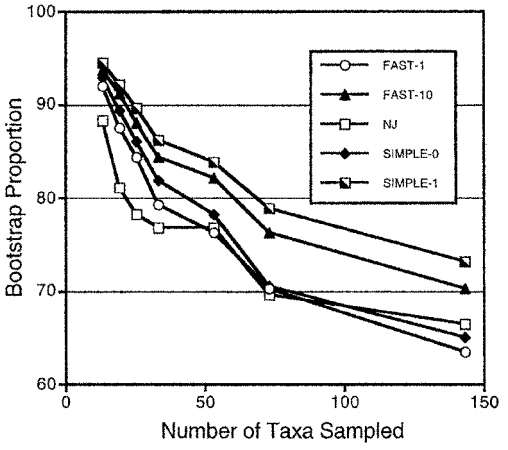


FIGURE 8. Uncorrected bootstrap proportions for the clade Neo-Astragalus in taxon subsamples of various sizes for five search algorithms. Each data point represents the mean over 100 random taxon samples of 100-replicate bootstrap runs except that, for the full data set (the points farthest to the right), one bootstrap run on the same full data set was done (1,000 bootstrap replicates).

particular sample used in our earlier study. Apparently, therefore, our original sample was not especially biased.

It is more difficult to reject unequivocally explanation 2—that heuristic searches are so ineffective with larger numbers of taxa that bootstrap support must decline. Results shown in Figure 8 make it clear that even among different heuristic search algorithms, bootstrap support varies to some extent. The differences are not huge, but they are consistent: More exhaustive strategies tend to narrow in on a set of trees that are more similar to one another and hence a higher proportion supports a particular clade. This is true at all sampling intensities. Interestingly, however, there is little indication that the difference in support between the best and worst algorithms changes with sampling intensity, as might be expected if the decline were driven by failure of the algorithms to find optimal trees, a process that presumably would worsen more quickly for the weaker algorithms. Moreover, heuristic strategies perform quite well in comparison with exact strategies, at least in the range of <20 taxa, where it is possible to check their performance given enough patience and judicious use of branch and bound algorithms. Over this range, the results of increased sampling still show a marked decline in the average bootstrap support. Assuming that at least SIMPLE-1 is performing fairly well, it is difficult to explain its decline as a mere failure to find most-parsimonious trees. A limited number of branch-and-bound searches bears this out, but this is a difficult issue to test because we cannot guarantee exact solutions over most of the interesting range covered by this study. The corrected bootstrap support for Neo-Astragalus declines very slightly with increasing numbers of taxa (Table 1), which may reflect some residual effect of algorithmic shortcomings.

On the other hand, explanation 3—that the decline in support stems partly from the statistical behavior ("bias") of the bootstrap proportion—must be given credence. The corrected $1 - \alpha$ level support for Neo-Astragalus remains close to 95% (Table 1), even though the conventional bootstrap support declines markedly (Fig. 8).

The BP has often been interpreted in the context of accuracy or repeatibility (e.g., Hillis and Bull, 1993; but see also Felsenstein and Kishino, 1993). The accuracy of a phylogenetic inference method is the probability (across randomly sampled data sets, μ^*, sampled from μ—and *not* bootstrapped pseudo-data sets, μ^{**}) that monophyly of some group will be inferred when true. Repeatibility is the mean BP for that group calculated across randomly sampled data sets. For a well-supported clade, the BP taken from a single data set will almost always underestimate both the accuracy and repeatibility (Zharkikh and Li, 1992a, 1995; Felsenstein and Kishino, 1993; Hillis and Bull, 1993; Efron et al., 1996; Newton, 1996). This stems directly from basic considerations about resampling in the set of discrete regions corresponding to the space of data sets (see Fig. 3).

A different, but more conventional, interpretation of the BP would be that it provides an estimate of $1 - \alpha$, where α is the probability that if the group were *not* a clade, we would have mistakenly inferred that it was a clade (i.e., made a type I error) by using a test that said to accept monophyly if $BP > 1 - \alpha$. We are happy to find a data set in which α is low, say, $<5\%$, because this means it is highly unlikely that by chance alone we would have obtained such a data set if the group were really not a clade. Referring to Figure 6, we see that the way to estimate α is to examine hypotheses along the boundary between regions R_0 and R_1 and than look at the probability of obtaining the observed BP value under

this null hypothesis. When the boundary is straight, this turns out to be exactly the same as $1 - BP$. However, in Figure 6, the BP is actually less than $1 - \alpha$, when α is corrected, because the boundary curves in toward R_1.

The biology enters in by altering the geometry of the sample space as a function of taxon sampling intensity. This is tied to the number of alternative hypotheses about monophyly of the relevant taxa. Figure 3 contains three alternative kinds of characters, one that supports monophyly and two that refute it. With only two alternatives the boundary is straight, but with three there is a corner, and the proportion of the frequency space occupied by the monophyly hypothesis, R_1, decreases from 50%. As more alternatives are added, the corner gets "sharper" in higher-dimensional space, and the hypervolume of R_1 relative to R_0 continues to get smaller. The BP declines, because more of the resampled data sets fall outside of R_1 in R_0. From the perspective of data sets (hypotheses) on the boundary, however, the picture is different. As the volume of R_1 declines, bootstrapped data sets taken from the boundary data sets have a harder and harder time reproducing the BP value associated with the original data set. Most of their resampled pseudo-data sets land in R_0, and thus their average BP values go down, forcing their cumulative distribution leftward (as in Fig. 4) and hence keeping true type I errors much less than would be expected from the conventional BP.

The formal mathematical relationship between Felsenstein's (1985) BP and Efron et al.'s (1996) correction is that the latter provides "second-order" accurate confidence limits on the null hypothesis, whereas the former is only first-order accurate (Efron et al., 1996). As the number of characters increases, the two should converge to the same value. However, for a fixed amount of character data, as here, when the number of taxa increases, our results suggest that the two estimates may diverge from one another. At some point, third-order and higher terms may be necessary to obtain sufficiently accurate confidence estimates (Efron and Tibshirani, 1996).

These conclusions have implications for the analysis of large data sets in general. Many data sets now exist with >100 taxa, and several molecular studies of 500+ taxa have been the subject of intense scrutiny in recent years (e.g., Rice et al., 1997). One characteristic of these studies is very low bootstrap scores (Soltis et al., 1998), which are often taken as conservative estimates of true bootstrap support, on the basis of previous theoretical considerations (Felsenstein and Kishino, 1993; Hillis and Bull, 1993; Zharkikh and Li, 1992a,b, 1995). However, better estimates of statistical confidence probably can be obtained by procedures such as Efron et al.'s (1996) method, the complete and partial bootstrap method (Zharkikh and Li, 1995), or iterated bootstrapping (Rodrigo, 1993). None of the latter methods has been examined relative to the question of taxon sampling intensity, however. The stumbling block in application of all these methods (ironically) is that although they are much more computationally intense than ordinary bootstrapping, they appear to be that much more necessary in just those circumstances in which the problem is already computationally difficult enough—large data sets. Perhaps this limitation can be circumvented by relying on very quick search algorithms (i.e., with little or no branch swapping), coupled with multiple rounds of bootstrapping. These may well yield results that are still too conservative, but as the present case study suggests, these alternative approaches are still much more indicative of true support than are BP. We know much more when a clade is conservatively supported at the 90% level (i.e., that it is between 90% and 100%) than when it is conservatively supported at the 60% level (and thus constrained only to lie in the broad interval of 60–100%).

Of course, corrections to bootstrap estimates cannot compensate for lack of information. If the number of characters is held constant, and the number of taxa increases, accuracy of phylogeny reconstruction can decline (Kim, 1998; Poe and Swofford, 1999; Bininda-Emonds et al., in press). This probably means an eventual decline even in *corrected* bootstrap values, if averaged over all clades in the tree. For any particular well-supported clade, in a data set with few taxa, however, most of the decline in uncorrected bootstrap support when taxa are added may simply be due to statistical bias.

ACKNOWLEDGMENTS

We are grateful to B. Efron and S. Holmes for discussion of these issues, and to P. Soltis and J. Trueman for thoughtful reviews. This work was supported by National Science Foundation grant DEB-9407824 to the

authors and a University of Arizona Foundation grant to M.F.W.

REFERENCES

BININDA-EMONDS, O., S. G. BRADY, J. KIM, AND M. J. SANDERSON. Scaling of accuracy in extremely large phylogenies. Pacific Symposium on Biocomputing. In press.

EFRON, B. 1987. Better bootstrap confidence intervals. J. Am. Stat. Assoc. 82:171–185.

EFRON, B., AND R. TIBSHIRANI. 1996. The problem of regions. Stanford Technical Report no. 192. Available from ftp://utstat.toronto. edu/pub/tibs/regions.ps.

EFRON, B., E. HALLORAN, AND S. HOLMES. 1996. Bootstrap confidence levels for phylogenetic trees. Proc. Natl. Acad. Sci. USA 93:7085–7090.

FELSENSTEIN, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791.

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

GAREY, M. R., AND D. S. JOHNSON. 1979. Computers and intractability. W. H. Freeman, San Francisco.

GRAHAM, R. L., AND L. R. FOULDS. 1982. Unlikelihood that minimal phylogenies for a realistic biological study can be reconstructed in reasonable computational time. Math. Biosci. 60:133–142.

HEDGES, S. B. 1992. The number of replications needed for accurate estimation of the bootstrap *P* value in phylogenetic studies. Mol. Biol. Evol. 9:366–369.

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

KIM, J. 1998. Large-scale phylogenies and measuring the performance of phylogenetic estimators. Syst. Biol. 47:43–60.

KUMAR, S., K. TAMURA, AND M. NEI. 1993. MEGA: molecular evolutionary genetic analysis, vers. 1.0. Pennsylvania State Univ., University Park, Pennsylvania.

LECOINTRE, G., H. PHILIPPE, H. L. V. LE, AND H. LE GUYADER. 1993. Species sampling has a major impact on phylogenetic inference. Mol. Phylog. Evol. 2:205–224.

LI, W.-H. 1997. Molecular evolution. Sinauer Press, Sunderland, Massachusetts.

LISTON, A., AND J. A. WHEELER. 1994. The phylogenetic position of the genus *Astragalus* (Fabaceae): evidence from the chloroplast genes *rpo*C1 and *rpo*C2. Biochem. Syst. Ecol. 22:377–388.

LOCK, J. M., AND K. SIMPSON. 1991. Legumes of West Asia, a check list. Royal Botanic Gardens, Kew, England.

MABBERLEY, D. J. 1997. The plant-book, a portable dictionary of the vascular plants, 2nd edition. Cambridge University Press, Cambridge.

MADDISON, D. R., D. L. SWOFFORD, AND W. P. MADDISON. 1997. NEXUS: an extensible file format for systematic information. Syst. Biol. 46:590–621.

NEWTON, M. A. 1996. Bootstrapping phylogenies: large deviations and dispersion effects. Biometrika 83:315–328.

POE, S. 1998. Sensitivity of phylogeny estimation to taxonomic sampling. Syst. Biol. 47:18–31.

POE, S., AND D. L. SWOFFORD. 1999. Taxon sampling revisited. Nature 398:299–300.

RICE, K. A., M. J. DONOGHUE, AND R. G. OLMSTEAD. 1997. Analyzing large data sets: *rbc*L 500 revisited. Syst. Biol. 46:554–563.

RODRIGO, A. 1993. Calibrating the bootstrap test of monophyly. Int. J. Parasitol. 23:507–514.

SANDERSON, M. J., AND J. J. DOYLE. 1993. Phylogenetic relationships in North American Astragalus (Fabaceae) based on chloroplast DNA restriction site variation. Syst. Bot. 18:395–408.

SOLTIS, D. E., P. S. SOLTIS, M. E. MORT, M. W. CHASE, V. SAVOLAINEN, S. B. HOOT, AND C. M. MORTON. 1998. Inferring complex phylogenies using parsimony: an empirical approach using three large DNA data sets for angiosperms. Syst. Biol. 47:32–42.

SWOFFORD, D. L., G. K. OLSEN, P. J. WADDELL, AND D. M. HILLIS. 1996. Phylogeny reconstruction. Pp. 407–514 *in* Molecular systematics, 2nd edition (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer, Sunderland, Massachusetts.

WOJCIECHOWSKI, M. F., M. J. SANDERSON, B. G. BALDWIN, AND M. J. DONOGHUE. 1993. Monophyly of aneuploid *Astragalus* (Fabaceae): evidence from nuclear ribosomal DNA internal transcribed spacer sequences. Am. J. Bot. 80:711–722.

WOJCIECHOWSKI, M. F., M. J. SANDERSON, AND J.-M. HU. 1999. Evidence on the monophyly of *Astragalus* (Fabaceae) and its major subgroups based on nuclear ribosomal DNA ITS and chloroplast DNA *trn*L intron data. Syst. Bot. 24:409–437.

ZHARKIKH, A. 1994. Estimation of evolutionary distances between nucleotide sequences. J. Mol. Evol. 39:315–329.

ZHARKIKH, A., AND W.-H. LI. 1992a. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. Mol. Biol. Evol. 9:1119–1147.

ZHARKIKH, A., AND W.-H. LI. 1992b. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. II. Four taxa without a molecular clock. J. Mol. Evol. 35:356–366.

ZHARKIKH, A., AND W.-H. LI. 1995. Estimation of confidence in phylogeny: the complete-and-partial bootstrap technique. Mol. Phylogenet. Evol. 4:44–63.

Received 4 January 1999; accepted 17 May 1999
Associate Editor: R. Olmstead