

Minimal Values for Reliability of Bootstrap and Jackknife Proportions, Decay Index, and Bayesian Posterior Probability

Richard H. Zander

*Missouri Botanical Garden
P.O. Box 299, St. Louis, MO 63166-0299 U.S.A.
Email: richard.zander@mobot.org*

Received: 15 September 2003 - Accepted: 15 November 2003

Abstract

Although optimal cladograms based on real data sets are readily demonstrated to be well loaded with phylogenetic data, statistical means of evaluating dependability of details of branch arrangements have been problematic. Exact values of four measures of branch arrangement reliability - nonparametric bootstrap and jackknife proportions, the Decay Index, and Bayesian posterior probabilities - were obtained from artificial 4-taxon data sets predetermined to have .95 confidence limits through a separate standard: an exact binomial calculation. Minimum values required for a .95 binomial confidence interval for each of these four metrics for internode lengths of 3 through 60 steps varied between 1.00 and .88 for bootstrap and jackknife; for Decay Index between 3 and 15; and for Bayesian posterior probabilities between 1.00 and .91. Binomial analysis involved the relative support for the optimal branch arrangement and a pool of support for the two alternative arrangements obtainable through nearest neighbor interchange with a null at probability 1/3. Any imbalance between the numbers of steps for the two non-optimal branch arrangements (of the three possible arrangements) lowers these four reliability measurements without affecting the binomial confidence interval, but such low values alone do not necessarily mean high reliability. In the literature, if any of these four common branch reliability measures and the branch lengths are given, unambiguous maximum binomial confidence intervals can now be estimated for those cladogram internodes. More exact confidence levels can be ascertained by recalculation with constraint trees using nearest neighbor interchange.

Introduction

Confidence intervals have been a sticking point in phylogenetic estimation. Although many methods using parametric and nonparametric statistics have been devised to estimate signal or compare signal in whole trees, measurement of the reliability of individual branch arrangements remains controversial (as discussed, e.g., by Zander, 2001). In this paper, a direct approach is taken to provide a rule of thumb for translating published common measures of branch arrangement reliability--bootstrap (BP) and jackknife (JP) proportions, Decay Index (DI), and Bayesian posterior

probabilities (BPP)--into binomial confidence intervals (CI), or for calculating this proposed new standard from data sets.

A general review of these measures follows: Non-parametric bootstrap proportions (BPs) using maximum parsimony analysis of data sets are obtained by resampling with replacement of the columns in the data matrix. Since their description by Felsenstein (1985), they have been a commonplace means of gauging the reliability of the internodes of cladograms generated by maximum parsimony or likelihood analysis. They are usually summarized in a consensus tree, for

instance, a majority-rule consensus tree. Felsenstein was careful to define bootstrap analysis in phylogenetic estimation as not the chance of being correct but as "a confidence interval within which is contained not the true phylogeny, but the phylogeny that would be estimated on repeated sampling of many characters from the underlying pool of characters." Larget et al. (2002) (and others) pointed out that, as opposed to Bayesian posterior probabilities, BPs are distinguished in not having a strong existing theory to justify their interpretation of uncertainty. The fact that nonparametric bootstrapping is used successfully in other fields as an alternative to other nonparametric analytical techniques (Efron and Tibshirani, 1993), e.g., Chi-squared or signed rank tests, has led researchers (Sanderson, 1989, 1995; Zharkikh and Li, 1995, and others) to examine its role as a confidence interval (CI).

Hillis and Bull (1993) demonstrated with simulations that BPs are generally lower than equivalent CIs, with reliable (.95 and above) CIs reached by BP values of .70 and above, given that rates of evolution are not high and disparity of rates between lineages is not great. Marvaldi et al. (2002) considered nodes with BP of $> .50$ to be "well supported." La Farge et al. (2002) used as robust support BPs of $\geq .70$, moderate as $< .70$ and $\geq .50$, and weak as $< .50$. Fishbein et al. (2001) used BP values of $< .70$ for weak support, 70--84 for moderate support, and $\geq .85$ as strong support. Generally, in the literature, the empirical ranges of BP support accepted as adequately reliable approximately match the ranges of variation in observed BP support in authors' cladograms, with authors having mostly well supported cladograms more inclined to rigorous requirements for acceptable BPs. Correction formulae have been proposed (Efron et al., 1996; J. Farris in Salamin et al., 2002; Rodrigo, 1993; Sanderson and Wojciechowski, 2000; Zharkikh and Li, 1995) that purport to provide the equivalent in CIs for BP values. Most are methodologically complex, often requiring special software or complex data treatment. Jackknife proportions are similar to BPs as reliability measures though obtained in a

resampling method that does not involve replacement.

The Decay Index (also known as Bremer Support) (Bremer 1988, 1994; DeBry 2001; Giribet 2003; Morgan 1997) is another measure of clade reliability commonly defined as the number of steps needed to "relax" parsimony until a given branch arrangement collapses in a consensus tree. For the simple artificial data sets in the present study it merely means that the optimal branch is n steps (the DI) longer than the next alternative branch arrangement. In any case, the DI does not take into account that there are three possible branch arrangements for any one internode. Thus, a decay index of 3 for an optimal branch 5 steps in length may mean in the worst case that the two alternative branches each have support of 2 steps. The chance of 5 steps occurring by chance alone out of 9 steps with a null hypothesis of random generation of steps (as discrete evolutionary events) at 1/3 probability is .15, which is too great to be considered statistically significant. With larger data sets, a branch arrangement may break down with a less relaxed (or less longer) tree due to homoplasy in distant branches, but this is not then a function of individual internode reliability. The local definition of the DI as used in this study is more meaningful. According to Goloboff and Farris (2001) "Bremer support generates absolute values to which a tree is suboptimal compared with another tree. A defect of that method is that it does not always take into account the relative amounts of evidence contradictory and favorable to the group. For example, according to the Bremer metric, a group supported by 2 uncontradicted characters is less well supported than a group supported by 100 and contradicted by 97. The first group, however, is relatively well supported, whereas in the second case there is about as much evidence in favor of the group as against it." Although Goloboff and Farris indeed pointed out this problem, they ignored critical statistical aspects in that they evaluated support for only two of the three possible alternative branch arrangements, suggesting a corrected Bremer support (RFD) as the ratio between the favorable and the contradictory evidence. The

importance of knowing the internal branch length when evaluating the meaning of the Decay Index has been emphasized by Deby (2001) and Wilkinson (2003).

The Bayesian Posterior probability as assigned to internal branches in consensus trees is now common in the literature with the more general use of Bayesian Markov chain Monte Carlo analytical software such as MrBayes (Huelsenbeck and Ronquist, 2001) and BAMBE (Simon and Larget, 1998). These programs are attractive because the BPP is in practice taken as the equivalent of the more familiar CI. Aside from problems with model and prior selection (Huelsenbeck et al., 2002), over-credibility (unexpectedly high scores) has been reported (Yoshiyuki et al., 2002) as has under-credibility (Wilcox, et al., 2002). Douady et al. (2003) compared nonparametric BPs and BPPs, and found that apparent conflicts in topology obtained with Bayesian analysis were reduced with bootstrapping, and that these two measures may prove to be valid upper and lower bounds to node reliability, but cannot be directly compared. Jordan et al. (2003) found maximum likelihood BPs to be generally lower than BPP scores of the same internodes, pointing out that BPP never involves subsampling and so always uses all the data.

Small, artificial 4-taxon data sets were devised so as to produce .95 CIs by exact binomial calculations (EBCs) based on the relative support for the three possible arrangements of the 4 taxa (through nearest neighbor interchange). The binomial CI is the probability that the data was not generated by chance alone. The CIs obtained were compared to BP, JP, DI and BPP values estimated for these data sets. The rationale for this choice of dataset is that any resolved, complex cladogram (based here on artificial data but also applicable to real data) may be decomposed into a series of four taxa or lineages connected by one internode. Through nearest neighbor interchange (NNI), parsimony analysis with constraint trees (constraining the full tree) can measure branch support for each of the three possible arrangements. The importance of

evaluating the distribution of contrary support among similar but competing hypotheses has been advanced recently by Zander (2001) and Wilkinson et al. (2003)

I propose that analysis of the relative support for the three different arrangements at any one internode of a cladogram (e.g., three terminal nearest neighbor lineages in rooted cladograms) is sufficient for reliability estimation of the one optimal branch arrangement. For this, the rationale is that we are gauging a posteriori the reliability of a particular branch arrangement in a given "best" tree, not continuing a search for a best tree. The reliability of such fixed branch arrangements is then a local problem, and considering the influence on the reliability of one internode by unrelated lineages (in that cladogram) requires consideration of other, non-optimal cladograms and is not productive. The question of interest is not "is the optimal cladogram wrong," but "is each arrangement of three terminal nearest neighbor branches reliably supported against contrary evidence from the two nearest neighbor lineages"? We may find that some internodes are not well supported but that does not mean they are incorrect (for instance, due to a short time between speciation events), and those that are well supported may not be correct (for instance, because of multiple-test problems). Reliability is defined here as support at a .95 CI, a minimal reliability value generally acknowledged as maximal acceptable risk in science. Reliability measures other than the binomial CI may not have the same statistical basis or be to some extent ambiguous, and "reliable" then may merely signal an author's degree of confidence in the hypothesis. Ideally, with any definition, reliable hypotheses are those that one chooses to act on (e.g., base further theorization, such as biogeographic analysis, on the branch arrangement) given the risk involved.

Another concept introduced here is the use of a binomial CI. Although a multinomial distribution is commonly used to determine if a die is loaded or for the description of phylogenetic data (Huelsenbeck, et al. 2000), a binomial calculation is sufficient for hypotheses of two categories. A multinomial

analysis (e.g. a Chi-squared test) is used when the spread of values from several categories matter, but the support in numbers of steps for the three alternative arrangements of an optimal tree involves theory that expects phylogenetic loading on only one arrangement, with any support for either of the two alternative arrangements being due to chance alone. We can use two categories because the data for the two alternative arrangements can be pooled a posteriori.

The analysis involves an alternative hypothesis of support for the optimal arrangement being due to shared ancestry at some CI, versus a null hypothesis of a star, where all support in terms of shared synapomorphic pairs of character states is equal and any variation is due to chance alone. The artificial data sets in Table 4, for instance, were all composed to generate a binomial CI of .95 (or a bit over that figure) but provide multinomial CIs from .89 in the shortest optimal branch length to .76 for the longest (by exact multinomial calculation, SISA, 2003). The EBC examines the chance by chance alone that AC successes (steps) would occur out of AB+AC+BC trials.

This paper compares BP, JP, DI and BPP to a separate standard, the binomial CI, using specially defined data sets of different sizes to test possible variation in results correlated with optimal branch length, and of imbalance in support for the three possible alternative arrangements by NNI.

Methods

Artificial datasets with four taxa were invented, using one two-state character for each column. These were not simulations with expectations, but tests of how reliability measures would differ with different numbers of columns (characters) and various character state proportions, all characters unweighted. The outgroup, D, was of all zeros, and 1's and 0's represented various proportions of uniquely shared support for the three other taxa, here named A, B and C. The ratios of shared support are cited in this paper by the formula AB:AC:BC, where AB represents the number of steps of

support for sister taxa in the optimal branch arrangement ((AB)C,D), and always is assigned the most support. Taxa A and C always share the second largest number of steps

```
D 000000000000000000000000000000000
A 1111111111111111111111100000000
B 1111111111111110000000011111111
C 0000000000000000111111111111111
```

Example data set, where AB:AC:BC = 15:08:07, resulting in the optimal cladogram of Fig. 1.

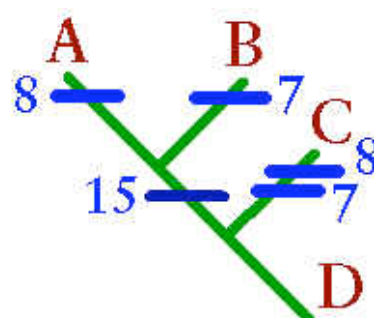


FIGURE 1. Example optimal cladogram (with one internal node) in which AB shares 15 advanced traits, but A and C share 8, and B and C share 7 other unique advanced traits. If the chance is < .05 that the support for the putative shared ancestry of AB can occur by chance alone (as 15 successes out of 30 trials at probability 1/3), then the cladogram must be collapsed to a star (the null hypothesis). In this case, that probability is .0425 and the cladogram ((AB)C,D) is a reliable reconstruction conditional on the data.

To investigate the effect on BP and BPP that an imbalance between the support in numbers of steps for the two alternative arrangements ((AC)B,D) and ((BC)A,D) that values AC and BC might have, 14 data sets with 50 states supporting ((AB)C,D), with varying numbers from 25 to 50 supporting ((AC)B,D), and none supporting ((BC)A,D) were created (see Table 1). Another 14 data sets were devised with 50 states supporting ((AB)C,D), and values for AC and BC always adding to 100 but with imbalanced ratios of AC:BC from 25:25 to 50:00 (see Table 2). Data sets for AB = 45 at predetermined binomial CI of .95 or .94

demonstrated the sensitivity and importance of AC and BC imbalance (see Table 3).

To compare the binomial CI with BP, JP, DI and BPP, twelve data sets were created to span the usual range of branch lengths in phylogenetic estimation, 3 through 60 steps supporting the optimal ((AB)C,D) arrangement, with the value of AB of 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55 and 60 steps in the 12 sets (see Table 4). Character states supporting A and C, and B and C, were selected such that with increasing AC + BC an EBC just reached a .95 CI for the optimal arrangement for that data set (the example data set above has a binomial CI of .957; one with a ratio of 15:8:8 would have a binomial CI of .942). That is, in terms of Bernoulli trials, such that AB or greater successes would occur by chance alone at most only 1 in 20 times in AB + AC + BC trials. The values AC and BC were made as close to equal as possible to avoid the ambiguity in BP, JP, DI and BPP as discussed below.

The probability (for that number or greater of shared characters of the optimal branch being generated by chance alone out of the total number of synapomorphies shared by all pairs of taxa) appropriately used in the binomial calculation is 1/3. This is under the null hypothesis of a star and a random generation of the same number of synapomorphies equally supporting the three possible arrangements and all variation being random. The alternative hypothesis is that two of the taxa have shared ancestry, demonstrated by significantly higher shared support. Support for each of the two suboptimal branch arrangements would be expected to be fairly balanced as a binomial distribution (with random generation through parallelism in real data).

Given that the artificial data sets were extremely simple, reanalysis though NNI and constraint was unnecessary for determining relative support of the alternative arrangements because support in numbers of steps could be read by eye from the data sets. If real data sets were used, constraint trees that fixed the entire tree with only one branch pair switched could be used

to calculate the length of a new internode supporting the new NNI arrangement (see Zander, 2001, 2003.) Nonparametric BPs were produced for the single internode of most-parsimonious cladograms for each artificial 4-taxon data set with PAUP* (Swofford, 1998) with 2000 bootstrap replications, fast heuristic search, tree-bisection-reconnection (TBR) algorithm, unordered, steepest descent not in effect, under ACCTRAN and MulTrees, and trees unrooted. Jackknife analyses with PAUP* also were done in the same manner. No islands of maximum parsimony were possible with these tiny data sets. Decay Index values were checked with PAST (Hammer and Harper, 2003) using branch-and-bound and increasing the longest tree kept until the single tree became two trees. The difference between the shortest tree and the tree length where the consensus was a star was the DI; with very simple trees such as these, the DI is basically the difference between AB, which is the number of steps supporting ((AB)C,D) and the next highest support value, which supports ((AC)B,D). Bayesian posterior probabilities were obtained with MrBayes (Huelsenbeck and Ronquist, 2001), set to "filetype = standard" (the morphology option assuming uniform priors), but otherwise default options, running 300,000 generations, and deleting 50 initial trees after burn-in. This number of generations was easily enough to reach convergence and 50 were sufficient to exclude the burn-in. Occasionally these simple data sets were run with many more generations (to 4 million) but there was no or little difference in BPPs associated with length of run.

Results

Effect of Imbalance in support of the two alternative arrangements

Table 1 gives binomial CIs, BPs and BPPs for 14 data sets with only two combinations of the three terminal taxa sharing characters. Thus, AB:AC:BC where AB = 50 uniquely shared character states; AC is varied from 25 to 50; and BC always = 0). The confidence limit by EBC for all these

data sets at the expected null hypothesis of probability 1/3 for AB was .99+, reflecting that 50 or more successes would be significantly rare out of a total number of trials ranging from 75 to 100. The BP, however, varied almost the same as EBC did at a null hypothesis of probability 1/2 (as though there were only 2 sides to the loaded phylogenetic coin instead of 3), probably because bootstrapping responded to the fact that BC was zero throughout, and there appeared to be only 2 categories. Although it is demonstrable that BP is an acceptable substitute for a nonparametric test with two combinations of shared traits, an EBC at probability 1/3 would have more correctly reflected the reliability of this data set given a theoretical evolutionary model involving 3, not 2 alternatives to the optimal branch arrangement. The BPP mostly agreed with the binomial CI at probability 1/3 (correctly), but drifted to .50 when data was strongly skewed. This is probably infrequent in real data sets.

TABLE 1. Nonparametric bootstrap proportions and Bayesian posterior probabilities for a 4-taxon data set with support for only 2 of the 3 possible arrangements ((AB)C,D) and ((AC)B,D). Binomial confidence intervals at 1/3 probability are always high, but the binomial CI calculated at 1/2 probability varies almost exactly with BP, and to some extent with BPP.

AB:AC:BC	Binomial CI at 1/3 Prob.	Binomial CI at 1/2 Prob.	BP	BPP
50:25:00	.99+	.99	1.00	1.00
50:26:00	.99+	.99	1.00	1.00
50:28:00	.99+	.99	.99	1.00
50:30:00	.99+	.98	.99	1.00
50:32:00	.99+	.97	.98	1.00
50:34:00	.99+	.95	.95	1.00
50:36:00	.99+	.92	.94	1.00
50:38:00	.99+	.88	.89	1.00
50:40:00	.99+	.83	.86	1.00
50:42:00	.99+	.77	.81	1.00
50:44:00	.99+	.70	.74	.99
50:46:00	.99+	.62	.66	.96
50:48:00	.99+	.54	.58	.82
50:50:00	.99+	.46	.51	.50

Table 2 gives binomial CIs, BPs and BPPs for 14 data sets each with 100 columns of traits uniquely shared among each of the three pairs. The data sets were created assigning AB 50 shared traits and by variously splitting the remaining 50 character state pairs among the two suboptimal branch arrangements. Thus, AB:AC:BC where AB = 50, and AC + BC = 50. Binomial CIs at 1/3 probability for all data sets are .99+ while those calculated at 1/2 probability are all .46 (i.e., 50 or more out of 100 is rare at probability 1/3 but common at probability 1/2). The BP values were approximately those produced by EBC at 1/2 probability in Table 1, and likewise largely ignore BC values. This clearly demonstrates (compare 50:25:00 in Table 1 with 50:25:25 in Table 2) that the bootstrap as implemented in PAUP* calculates BPs as if EBCs were done on a sliding scale from 1/3 to 1/2 probability depending on imbalance between AC and BC, and contributes to the lowering of the BP towards an EBC at 1/2 probability. Again, the BPP agreed with the binomial CI at probability 1/3 but skewed towards a CI at probability 1/2 when there was great imbalance between AC and BC.

TABLE 2. Here AB + AC + BC always sum to 100 steps and AB is always 50, so CI at 1/3 probability and that at 1/2 are always the same for all data sets, but BP, and BPP to a lesser extent, varies with degree of imbalance between values of AC and BC.

AB:AC:BC	Binomial CI at 1/3 Prob.	Binomial CI at 1/2 Prob.	BP	BPP
50:25:25	.99+	.46	1.00	1.00
50:26:23	.99+	.46	1.00	1.00
50:28:22	.99+	.46	.99	1.00
50:30:20	.99+	.46	.99	1.00
50:32:18	.99+	.46	.99	1.00
50:34:16	.99+	.46	.96	1.00
50:36:14	.99+	.46	.93	1.00
50:38:12	.99+	.46	.92	1.00
50:40:10	.99+	.46	.86	1.00
50:42:08	.99+	.46	.83	.99
50:44:06	.99+	.46	.73	.98
50:46:04	.99+	.46	.65	.94
50:48:02	.99+	.46	.56	.81
50:50:00	.99+	.46	.51	.51

The above specially constructed data sets demonstrate that imbalance between support for the two suboptimal branch arrangements at any internode lowers the BP and some extent the BPP. This is true although the binomial CI may remain the same (Table 2). This source of ambiguity can only be dealt with by determining BP and BPP values equivalent to an acceptable binomial CI by using artificial data sets with AB:AC:BC ratios that have $AC \approx BC$. These provide a maximum binomial CI unambiguously implied by that BP or BPP.

Table 3 compares reliability measures determined from ratios of support for 4 data sets, each of 45 steps in the optimal branch arrangement, but with different values for AC and BC, producing either a .95 (one chance in 20 of being generated by chance alone) or a .94 (one chance in 17.5 of being by chance alone) binomial CI. In Table 3, for the same optimal branch length, different BPs can imply the same CI and the same BP can imply different CIs. The ratio 45:35:29 (with a binomial CI of .95), and 45:33:33 (with a binomial CI of .94), both have a BP of .85. Only the BP that matches a ratio of AC as close as possible to BC can unambiguously imply (be explained by) a particular maximum binomial CI. For the ratios in Table 3, for example, only a BP of .87 implies a binomial CI of .95, not a BP of .85, which only sometimes implies a support ratio that attains a binomial CI of .95. Jackknife values were not calculated as they are similar to BP at least for these 4-taxon artificial data sets, as indicated in Table 4.

TABLE 3. Comparison of binomial CI, BP, BPP and DI for different data sets with a 45-step optimal branch internode. The AC and BC imbalance critically affects the CI.

AB:AC:BC	Binomial CI	BP	DI	BPP
45:32:32	.95	.87	13	.948
45:33:31	.95	.87	12	.944
45:33:33	.94	.85	12	.928
45:35:29	.95	.85	12	.935

If the support ratios are known, as in the artificial 4 taxon data sets of Table 3, the DI

is simply AB minus the next greatest support value among the alternative arrangements; only a DI of 13 unambiguously implies a binomial CI of .95 for a branch length of 45, since the DI as a measure a reliability may also be ambiguous if AC and BC are imbalanced.

The BPP responds to some extent to imbalance in AC and BC, and for a 45-step branch length, nothing lower than a BPP of .948 unambiguously implies a binomial CI of .95. At this branch length, however, there is little problem (but see Table 4).

Minimal measures for reliable branch arrangements

Identifying imbalance in AC and BC in published papers is impossible (without recalculation from the actual data set). Table 4 provides a guide to the minimum values for BP, JP, DI and BPP that imply binomial CI equivalents of at least .95. Various optimal internode lengths are given for artificial data sets where AB is the support for the optimal branch arrangement, and AC is set as close to BC as possible. These data sets were selected as the ones for which increasing the sum of AC + BC just reaches a binomial CI of .95 when the data AB versus AB + AC + BC was analyzed with EBC. Internode lengths of AB of 3 and 4 steps are included because 3 is the shortest length of an internode that returns a binomial CI of .95 (maximum binomial CI for an internode of 1 step is .66, of 2 steps is .89). Then, data sets with 5 through 60 steps in the optimal internode are described; interpolation of branch lengths not multiples of 5 is necessary if the chart is used to check actual information given in the literature. The BP can be calculated with very small internode lengths and four taxa because the generalized attraction of branches in many-taxon data sets with a paucity of information (e.g. morphological data) is absent.

TABLE 4. Internal branch lengths with the maximal AB:AC:BC ratios that attain a .95 binomial CI with $AC \approx BC$ at that length, and minimum BP, JP, BPP and DI values needed to attain that level of confidence unambiguously.

Length of Optimal Branch AB	Max. AB:AC:BC Needed for .95 CI	Min. BP Needed	Min. JP Needed	Min. DI Needed	Min. BPP.
3	03:00:00	1.00	1.00	3	.99
4	04:01:00	.95	.79	3	.99
5	05:01:01	.95	.90	4	1.00
10	10:04:04	.91	.92	6	1.00
15	15:08:07	.91	.91	7	.99
20	20:12:11	.89	.89	8	.98
25	25:16:15	.89	.89	9	.98
30	30:20:19	.89	.89	10	.97
35	35:24:23	.88	.89	11	.95
40	40:28:27	.89	.89	12	.96
45	45:32:32	.87	.88	13	.95
50	50:37:36	.87	.87	13	.93
55	55:41:40	.87	.87	14	.92
60	60:45:45	.88	.88	15	.91

BPs for short optimal internode lengths (3 to 15 steps) are generally similar to the equivalent binomial CI. Those for longer optimal internode lengths are about 10 percentage points lower, with .87 to .89 being the range of minimal BPs that ensure reliability at a .95 binomial CI. JPs for these same very simple data sets are nearly identical with the BPs, except in the shortest trees where the paucity of data affects the methods differently.

The required DI unambiguously implying a .95 binomial CI increases evenly about every 5 steps in optimal branch length. BPPs are overly credible for optimal branch lengths of 30 and less, but underestimate the binomial CI in longer branches. According to Huelsenbeck (pers. comm.) the prior probably overwhelms the data when they are very few. This is, however, probably not due to overspecification of a model since the priors were specified as uniform. Although less affected by imbalance in AC and BC than BP, BPPs are nevertheless not immune, and ambiguity is a factor to consider.

DISCUSSION

Common measures of internal branch reliability

All common measures of branch reliability are affected by imbalance in AC and BC. If one is able with real data sets to determine directly by analysis of constraint trees the internode length ratios for AB:AC:BC, then far more exact CIs can be calculated. The binomial CI values in Table 4, however, may be good approximations for the majority of published BPs, JPs, DIs and BPPs, given that AC and BC are presumably generated randomly through parallelism, and should seldom be strongly imbalanced. One problem is that much modern literature does not give internode lengths, and these have to be guessed at, for instance, from the total tree length and numbers of internodes, or relative length of phylogram branches.

Hillis and Bull (1993) indicated that BP values of .70 and above imply CIs of .95 and above in their simulations. This may be explained in two ways: (1) Their results are quite understandable in view of the effect of imbalance in AC:BC ratios in lowering BP values. An artificial data set constructed in the manner above with the ratio 30:29:10 has a binomial CI of .95 and a BP of .55 but the ratio of AC:BC, or 29:10, has a binomial CI (EBC at probability 1/2) of .99 so that it itself has the equivalent of a phylogenetic signal (or sample error, etc.) under the null that support for AC and BC are equal, and generated randomly (in theory) by parallelism. The minimum ratio in which that secondary null hypothesis cannot be rejected at .95 is 35:29:18 (also CI .95), which has a BP of .77, itself undoubtedly a low value. Hillis and Bull, apparently, did not take into account the fact that imbalances in AC and BC cause ambiguity by lowering BPs, and ratios with AC \approx BC must be used to determine minimum BP for reliable CIs. Thus, BP values between their figure of .70 and at least that of .87 can match both .95 and lower binomial CIs for the internode lengths investigated here. (2) According to Hillis and Bull "... estimated internal branches with bootstrap proportions above 70% represent true clades over 95% of the time" Their CIs therefore address a different question than that of the present paper, since they measured the number of

times the "true" clade was returned, not the reliability of that clade. Apparently, few true clades with CIs of less than .95 but correct >.70 BP were included in their simulations.

Braun and Kimball (2002) found that, when analyzing their data under transversion parsimony, BPs were obtained of >.70 that matched Bayesian analyses using a complex model resulting in >.95 posterior probability. Clearly this equivalency is attainable in practice, but one cannot gauge a binomial CI by the BP or the BPP in the literature without knowing the branch lengths.

In sum, confounding ambiguities due to multiple possible values of BP for a given binomial CI is one of the problems of using BPs as reliability measures in parsimony studies. It is possible, however, to see if a minimal binomial CI for reliability is attained when BPs and their associated optimal internode lengths are given. If standard parsimony software were able to compare constraint trees more easily to obtain AB:AC:BC ratios, exact internodal binomial CIs could be quickly calculated through EBC.

The DI has been criticized for its deficiencies as a metric in accurately estimating reliability (Oxelman et al. 1999, Rice et al. 1997, and Yee 2000). It can be directly calculated (as a local DI, unaffected by homoplasy elsewhere in the tree) if both AB and AC (where AC here always is the larger value of AC and BC) are known (determined with NNI and constraint trees). If AB and the branch length are known, Table 4 can provide an estimate of the equivalent maximum unambiguous binomial CI.

Interpretation of large Decay Index values in the literature is especially problematic. For example, Ahonen et al. (2003) published a cladogram from molecular data for which only Decay Index branch support was given, and which ranged from 1 to 170 in value. No branch lengths were specified and the branches were not proportional to length. The Decay Index for internodes in this cladogram can be calculated, however, by dividing the tree length, given as 9458 steps, by the number of internal branches, 15, to

obtain the average branch length, or 631. The branch with a Decay Index of 170 would then have a length of 631 (probably longer in view of the large DI, but this is a minimum fail-safe figure) and the two alternative branch arrangements could then each be as many as 461 (631 less 170) steps in length (the worst case scenario). The ratio on which to base a binomial calculation is then 631:461:461, or 631 successes out of 1553 at probability 1/3, which provides a .99+ binomial CI. The ratio that would actually return a binomial CI of .95 is 631:581:580; thus, the minimum Decay Index value needed for a reliable branch arrangement in this cladogram is 50 (631 less 581). Of the 15 internodes, 6 were below 50 and are, on the basis of this procedure, unreliable.

The BPP is a new and promising measure of branch arrangement reliability, it is especially attractive in that it is easily interpreted as the standard CI. Although a Bayesian credible interval is philosophically and in part methodologically different (Lewis, 2001) from the frequentist CI, both are certainly viewed in the same manner when retrodicting single historical events, as in phylogenetic estimation. Confidence intervals in phylogenetic estimation are more like Bayesian credible intervals than they are measures of confidence in frequentist long-run predictions from long-run data. The present paper demonstrates that BPPs are apparently over-credible for branch lengths of 30 and less, and they underestimate the binomial CI at branch lengths of 40 and greater. They, like other common measures of reliability, are lowered by imbalance between AC and BC (support for the two alternative branch arrangements of nearest neighbors).

The binomial confidence interval

The conditional probability of reconstruction (CPR) as used for molecular data (Zander, 2001) is essentially the same as the binomial CI but uses a Chi-squared calculation instead of a binomial calculation; confidence intervals obtained with the former are somewhat less forgiving. Felsenstein (2003: 331) criticized the CPR, in part invoking his "S-statistic," which involves the difference between the number

of steps favoring the best tree and the next best tree. These data alone are insufficient for an adequate test for a 4-taxon tree given that three arrangements are possible. He was correct, however, in pointing out that expectation of equal numbers of changes in the interior branches of all three trees will not hold in non-clocklike cases, and indeed a clock must be assumed in the simple form of the binomial CI described here. For evaluating the reliability of internodes of cladograms obtained without assumption of a clock, however, the same limitations on non-clocklike behavior can also be assumed for the binomial CI. Felsenstein pointed out that long-branch attraction may make the binomial CI declare the wrong tree significantly supported, but this is contrary to the idea that finding the reliability of one branch arrangement in a particular tree is a local problem. For the purposes of the binomial analysis, namely assessing the phylogenetic import of one branch arrangement, the optimal tree is assumed to be the true tree except for the branch arrangement studied. The validity of settling on that one optimal tree is part of another, different investigative process. Finally, the CPR was not presented by me as a Bayesian posterior (beyond my pointing out that all phylogenetic retrodiction involves betting on reconstruction of a single historical event), so Felsenstein's objection that it was so offered but lacked a prior distribution is superfluous.

The maximal values obtained from Table 4 for the binomial CI are limited by the multinomial nature of evaluation for the Decay Index and generation of Bayesian MCMC posterior probabilities, methods that respond to the spread of support for the three possible arrangements of any internode in the optimal arrangement. In the case of nonparametric bootstrapping, the lowest value supporting one the three alternative branch arrangements (here, BC) is largely ignored. The binomial CI is also reasonable in that it returns higher values than nonparametric BP, which is known to be somewhat low, and generally lower values than BPP, known to be over-credible. If there are multiple shortest trees, analyzing one of them is doubtless sufficient because internodes that reduce to polytomies in

consensus trees will be poorly supported in any case.

In practice, there are two ways of measuring the reliability of local branch arrangements: (1) Table 4 can be used, with interpolation for branch lengths that are not a multiple of 5, to gauge the maximum unambiguous binomial CI of internal branch arrangements in published cladograms when a common measure of branch reliability (BP, JP, DI or BPP) and the branch lengths are both given. (2) A more exact estimation of reliability could be had by recalculating from the original data set. Support values for the three possible arrangements at any internode are obtained with recalculation with fully constrained trees (i.e. from a full tree description formula) after NNI. For complex data sets, a graphics program like PageView (Page 1996, 2001) would provide accurate constraint formulae after manual branch swapping. EBC from those values gives the binomial CI.

Generalized weighting is difficult to evaluate locally in a cladogram. The binomial CI is acceptable if the optimal and the two alternative branch lengths of an internode truly sum individual evolutionary events that are either stochastically produced or are ancestrally shared traits, and which are unweighted except as clearly reflecting likelihood (where a weight of 2 is actually equivalent to the chance of 2 traits randomly occurring together as synapomorphies, or the level of support of two state changes downweighted by 0.5 represents the support of only a single state change of other characters). If cladograms involving weighting of data sets are analyzed, then this warning is important; the CI of 4 successes out of 9 trials at 1/3 probability (.65) is quite different from that of 40 out of 90 trials (.99), where the proportions stay the same but the characters are up-weighted. Down-weighting is apparently less problematic than up-weighting. Likelihoods of molecular state changes may differ among sites, or by synonymous/nonsynonymous and transition/transversion bias, and assuming equal likelihoods/weighting may be misleading as is the case with morphological characters (Waters et al., 2002). Although

data may be available to infer transformation likelihoods for the cladogram overall, evaluating differential likelihoods of advanced characters associated with only the sites supporting the three arrangements at one internode is difficult since the data for inferring those particular likelihoods are then few. One might assume for the difficult problem of estimating reliability at one internode of a molecularly based tree that, pending development of adequate theory and possibly normalization procedures for weighted phylogenetic distance, rare and common state changes are represented in about equal proportions in support for each arrangement, and that their likelihoods are uniform or at least uninformative in the Bayesian sense.

Although binomial CIs, when approximated with the information in Table 4, are a better measure of branch arrangement reliability than the BP, JP, DI or BPP, it remains true that the total reliability for any complex lineage or cladogram is the product of all internode CIs multiplied by the probability of the branch arrangement being correct (when being incorrect causes a change in elements of interest in the cladogram) for each of every variable affecting the solution: the NP-complete problems of best fit and alignment (Marshall, 1997), sample error, weighting, model selection, convergence, long-branch attraction, and parallelism and introgression (Avice, 1994; Doyle, 1992; Templeton, 1986), while lineage sorting (Hudson, 1992; Lyons-Weiler and Milinkovitch, 1997; Maddison, 1996; Pamilo and Nei, 1988) can produce false positive results for different gene data sets for the same taxon. Conditional on the data set and with some attention paid to these other factors, high

binomial CIs as judged by BPs, JPs, DIs and BPPs in the literature, coupled with knowledge of the branch lengths, should give good indication of reliability of the tree or of parts of the tree.

Four key concepts involved in this paper are: (1) support values for each of the three possible branch arrangements of four taxa are sufficient to inform a reliability measure, (2) on such values depend a choice between, not the three arrangements which were previously dealt with in the optimization procedure, but the optimal arrangement (alternative hypothesis, e.g. Fig. 1) and a star, (3) gauging reliability of local branch arrangements on a tree is a local problem, and (4) a binomial confidence interval is justified by an expectation of phylogenetic loading on only one of the three possible arrangements. The reliability of individual branch arrangements in the vast store of published literature with parsimony results can now be evaluated through one or the other of the above techniques. The results are, of course, conditional on the data set, and on various experimental and regularity assumptions.

Acknowledgements

I appreciate several helpful and often illuminating communications on this topic from J. Felsenstein, J. B. Huelsenbeck, B. Larget, J. Lyons-Weiler, and M. Wilkinson over the past few years, and I value the comments of reviewers. Calculations of CIs were done with the EBC function of VassarStats (Lowry, 2000) or with the Z-ratio calculator when values were large.

Literature Cited

- Ahonen, I., J. Muona & S. Piippo. 2003. Inferring the phylogeny of the Lejeuneaceae (Jungermanniopsida): A first appraisal of molecular data. *Bryologist* 106: 297--308.
- Avice, J. C. 1994. *Molecular Markers, Natural History and Evolution*. Chapman and Hall, New York.
- Braun, E. L. & R. T. Kimball. 2002. Examining basal avian divergences with mitochondrial sequences: model complexity, taxon sampling, and sequence length. *Syst. Biol.* 614--625.
- Bremer, K. 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution* 42: 795--803.

- Bremer, K. 1994. Branch support and tree stability. *Cladistics* 10: 295-304.
- DeBry, R.W. 2001. Improving interpretation of the decay index for DNA sequence data. *Syst. Biol.* 50: 742--752.
- Douady, C. J., F. Delsuc, Y. Boucher, W. F. Doolittle & E. J. Douzery. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol. Biol. Evol.* 20: 248--254.
- Doyle, J. J. 1992. Gene trees and species trees: molecular systematics as one-character taxonomy. *Syst. Bot.* 17: 144--163.
- Efron, B., E. Halloran & S. Holmes. 1996. Bootstrap confidence intervals for phylogenetic trees. *Proc. Natl. Acad. Sci. USA* 93: 7085--7090.
- Efron, B. & R. J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman and Hall, New York.
- Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39: 783--791.
- Felsenstein, J. 2003. *Inferring Phylogenies*. Sinauer Associates, Inc., Sunderland, Massachusetts
- Fishbein, M., C. Hibsich-Jetter, D. E. Soltis & L. Hufford. 2001. Phylogeny of Saxifragales (Angiosperms, Eudicots): analysis of a rapid, ancient radiation. *Syst. Biol.* 50: 817--847.
- Giribet, G. 2003. Stability in phylogenetic formulations and its relationship to nodal support. *Syst. Biol.* 52: 554-564.
- Goloboff, P. A. & J. S. Farris. 2001. Methods for quick consensus estimation. *Cladistics* 17: 526-534.
- Hammer, Ø. & D. A. T. Harper. 2003. PAST v. 1.12. <http://folk.uio.no/hammer/past>.
- Hillis, D. M. & J. J. Bull. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.* 42: 182--192.
- Hudson, R. R. 1992. Gene trees, species trees and the segregation of ancestral alleles. *Genetica* 131 509--512.
- Huelsenbeck, J. P., B. Larget, R. E. Miller & F. Ronquist. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. *Syst. Biol.* 51: 673--688.
- Huelsenbeck, J. P., B. Rannala & B. Larget. 2000. A Bayesian framework for the analysis of cospeciation. *Evolution* 54: 352--364.
- Huelsenbeck, J. P. & F. Ronquist. 2001. MrBayes: v30B4. Bayesian Analysis of Phylogeny. University of California, San Diego, and Dept. of Systematic Zoology, Uppsala University.
- Jordan, S., C. Simon & D. Polhemus. 2003. Molecular systematics and adaptive radiation of Hawaii's endemic damselfly genus *Megalagrion* (Odonata: Coenagrionidae). *Syst. Biol.* 52: 89--109.
- La Farge, C., A. J. Shaw & D. H. Vitt. 2002. The circumscription of the Dicranaceae (Bryopsida) based on the chloroplast regions *trL-trnF* and *rps4*. *Syst. Bot.* 27: 435--452.
- Larget, B., D. L. Simon & J. B. Kadane. 2002. Discussion on the meeting on "Statistical modeling and analysis of genetic data." *J. Roy. Statist. Soc. B* 64 (Part 4): 764--767.
- Lewis, P. O. 2001. Phylogenetic systematics turns over a new leaf. *Trends Ecol. Evol.* 16: 30--37.
- Lowry, R. 2000. VassarStats: Web site for statistical computation. Department of Psychology, Vassar College, Poughkeepsie, New York. <http://faculty.vassar.edu/~lowry/VassarStats.html>, Jan. 25, 2000.
- Lyons-Weiler, J. & M. C. Milinkovitch. 1997. A phylogenetic approach to the problem of differential lineage sorting. *Mol. Biol. Evol.* 14: 968--975.
- Maddison, W. P. 1996. Molecular approaches and the growth of phylogenetic theory. In: *Molecular zoology: advances, strategies, and protocols* (eds. J. D. Ferraris & S. R. Palumbi.), pp. 47-63. Wiley-Liss, New York
- Marshall, C. R. 1997. Statistical and computational problems in reconstructing evolutionary histories from DNA data. *Computing Sci. Stat.* 29: 218--226.
- Marvaldi, A. E., A. S. Sequeira, C. W. O'Brien & B. D. Farrell. 2002. Molecular and morphological phylogenetics of weevils (Coleoptera, Curculionoidea): do niche shifts accompany diversification? *Syst. Biol.* 51: 761--785.

- Morgan, D. R. 1997. Decay analysis of large sets of phylogenetic data. *Taxon* 46: 509--517.
- Oxelman, B., M. Backlund & B. Bremer. 1999. Relationships of the Buddlejaceae s. l. investigated using parsimony, jackknife and branch support analysis of chloroplast *ndhF* and *rbcl* sequence data. *Syst. Bot.* 24: 164--182.
- Page, R. D. M. 1996. TREEVIEW: An application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences* 12: 357-358.
- Page, R. 2001. TreeView (Win32). Ver. 1.6.6 University of Glasgow, Glasgow.
- Pamilo, P. & M. Nei. 1988. Relationships between gene trees and species trees. *Mol. Biol. Evol.* 5: 568--583.
- Rice, K. A., M. J. Donoghue & R. G. Olmstead. 1997. Analyzing large data sets: *rbcl* 500 revisited. *Syst. Biol.* 46: 554--563.
- Rodrigo, A. G. 1993. Calibrating the bootstrap test of monophyly. *Int. J. Parasitol.* 23: 507--514.
- Salamin, N. T. R., Hodkinson & V. Savolainen. 2002. Building supertrees: an empirical assessment using the grass family (Poaceae). *Syst. Biol.* 51: 136--150.
- 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. *Syst. Biol.* 44: 299--320.
- Sanderson, M. J. & M. F. Wojciechowski. 2000. Improved bootstrap confidence limits in large-scale phylogenies, with an example from neo-*Astragalus* (Leguminosae). *Syst. Biol.* 49: 671--685.
- Simon, D. & B. Larget. 1998. BAMBE: Bayesian Analysis in Molecular Biology and Evolution. Ver. 1.01 beta, October 1998. Dept. Math. and Computer Sci., Duquesne Univ.
- SISA. 2003. Multinom, an exact multinomial calculator. SISA: Simple Interactive Statistical Analysis. <http://home.clara.net/sisa/pasprog.htm>. August 13, 2003.
- Swofford, D. L. 1998. PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Ver. 4. Sinauer Associates, Sunderland, Massachusetts.
- Templeton, A. 1986. Relation of humans to African apes: a statistical appraisal of diverse types of data. In: *Evolutionary processes and theory* (eds. S. Karlin & E. Nevo), pp. 365-388. Academic Press., New York.
- Waters, M. W., T. Saruwatari, T. Kobayashi, I. Oohara, R. M. McDowall & G. P. Wallis. 2002. Phylogenetic placement of retropennid fishes: Data set incongruence can be reduced by using asymmetric character state transformation costs. *Syst. Biol.* 51: 432-449.
- Wilcox, T. P., D. J. Zwickl, T. Heath, and D. M. Hillis. 2002. Phylogenetic relationship of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. *Molecular Phylogenetics and Evolution* 25: 361--371.
- Wilkinson, M., F.-J. Lapointe & D. J. Gower. 2003. Branch lengths and support. *Syst. Biol.* 52: 127--130.
- Yee, M. S. Y. 2000. Tree robustness and clade significance. *Syst. Biol.* 49: 829--836.
- Yoshiyuki, S., G. V. Glazko, and M. Nei. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proc. Nat. Acad. Sci.* 99: 16138--16143.
- Zander, R. H. 2001. A conditional probability of reconstruction measure for internal cladogram branches. *Syst. Biol.* 50: 425--437.
- Zander, R. H. 2003. Reliable phylogenetic resolution of morphological data can be better than that of molecular data. *Taxon* 52: 109--112.
- Zharkikh, A. & W.-H. Li. 1995. Estimation of confidence in phylogeny: Complete-and-partial bootstrap technique. *Mol. Phylog. Evol.* 4: 44--63.