Available online at www.sciencedirect.com
SCIENCE DIRECT®
Cladistics 19 (2003) 324–332
Cladistics
www.elsevier.com/locate/yclad

Improvements to resampling measures of group support

Pablo A. Goloboff,[a,*] James S. Farris,[b] Mari Källersjö,[b] Bengt Oxelman,[c] Martín J. Ramírez,[d] and Claudia A. Szumik[a]

[a] *Instituto Superior de Entomología "Dr. Abraham Willink," Facultad de Ciencias Naturales, Miguel Lillo 205, 4000 San Miguel de Tucumán, Argentina*
[b] *Molekylarsystematiska Laboratoriet, Naturhistoriska Riksmuseet, Stockholm, P.O. Box 50007, S-104 05 Stockholm, Sweden*
[c] *Department of Systematic Botany, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, SE-752 36 Uppsala, Sweden*
[d] *American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024, USA*

Accepted 7 May 2003

Abstract

Several aspects of current resampling methods to assess group support are reviewed. When the characters have different prior weights or some state transformation costs are different, the frequencies under either bootstrapping or jackknifing can be distorted, producing either under- or overestimations of the actual group support. This is avoided by symmetric resampling, where the probability p of increasing the weight of a character equals the probability of decreasing it. Problems with interpreting absolute group frequencies as a measure of the support are discussed; group support does not necessarily vary with the frequency itself, since in some cases groups with positive support may have much lower frequencies than groups with no support at all. Three possible solutions for this problem are suggested. The first is measuring the support as the difference in frequency between the group and its most frequent contradictory group. The second is calculating frequencies for values of p below the threshold under which the frequency ranks the groups in the right order of support (this threshold may vary from data set to data set). The third is estimating the support by using the slope of the frequency as a function of different (low) values of p; when p is low, groups with actual support have negative slopes (closer to 0 when the support is higher), and groups with no support have positive slopes (larger when evidence for and against the group is more abundant).

One of the fundamental aspects of parsimony analysis is the evaluation of group support. To the extent that there is more evidence in favor of it, and less evidence against, a group is said to be better supported.

Jackknifing (Farris et al., 1996; Lanyon, 1985) and bootstrapping (Felsenstein, 1985) are widely used to measure support, but they sometimes produce illogical results. In this paper, we discuss these problems and explore possible solutions. Some of the proposed solutions are less than ideal, but we discuss them because they may eventually lead to better ones. The methods described have been implemented in several computer programs, available from P.A.G. and J.S.F.

Terminology and conventions

Throughout the paper, Jackknifing is indicated as **J** and bootstrapping as **B**; the jackknife frequency of group AB under a deletion probability p is indicated as $j_{(\mathrm{AB},p)}$ and its bootstrap frequency for a Poisson distribution of mean m as $b_{(\mathrm{AB},m)}$. For **J**, the resampling strength is the probability p of deleting (or changing the weight) of a given character; the probability of the character retaining its original weight is $q = 1 - p$.

The exact calculation of frequencies was done by enumerating all possible weight combinations (2^n in the case of jackknifing, 3^n in the case of the symmetric resampling described below) for the n characters in the matrix. This is illustrated in Fig. 1, where each row in the second column shows a rearrangement, as a weight vector; the first row represents the resampled matrix where all characters have weight 0, the second row

* Corresponding author.
E-mail address: instlillo@infovia.com.ar (P.A. Goloboff).

0748-3007/$ - see front matter © 2003 The Willi Hennig Society. Published by Elsevier Science (USA). All rights reserved.
doi:10.1016/S0748-3007(03)00060-4

#	rearrangement	prob.	is group G present?	term
1	0000000 … 000	p^n	no	$X_{1,p}$
2	0000000 … 001	$p^{n-1}.q$	no	$X_{2,p}$
3	0000000 … 010	$p^{n-1}.q$	no	$X_{3,p}$
4	0000000 … 100	$p^{n-1}.q$	no	$X_{4,p}$
5	0000000 … 011	$p^{n-2}.q^2$	yes	$Y_{1,p}$
6	0000000 … 110	$p^{n-2}.q^2$	no	$X_{5,p}$
7	0000000 … 101	$p^{n-2}.q^2$	yes	$Y_{2,p}$
8	0000000 … 111	$p^{n-3}.q^3$	no	$X_{7,p}$
9	0000001 … 000	$p^{n-1}.q$	yes	$Y_{3,p}$
10	0000001 … 001	$p^{n-2}.q^2$	no	$X_{8,p}$
11	0000001 … 010	$p^{n-2}.q^2$	yes	$Y_{4,p}$
12	0000001 … 100	$p^{n-2}.q^2$	yes	$Y_{5,p}$
13	0000001 … 110	$p^{n-3}.q^3$	yes	$Y_{6,p}$
14	0000001 … 101	$p^{n-3}.q^3$	no	$X_{9,p}$
…	… …			
2^n	1111111 … 111	q^n	yes	$Y_{k,p}$

$$j_{(G,p)} = \sum_{i=1}^{k} Y_{i,p}$$

Fig. 1. Example of an exact calculation of frequencies; see text for details.

represents the matrix where all characters have weight 0 except the last one, etc. The probabilities (third column) are easily obtained for each possible rearrangement of the weight vector. For each rearrangement, it is necessary to calculate whether the group of interest, G, is supported by the resulting sample of characters. Define $Y_{i,p}$, as the probability of the ith rearrangement displaying the group, for resampling strength p; likewise for $X_{i,p}$, for the ith rearrangement *not* displaying the group. The rearrangements are orderly generated, because each must be examined only once. Then, $j_{(G,p)} = \sum Y_{i,p}$. For all the exact calculations of frequencies, the equal symbol (=) and four significant digits are used, while the symbol "approximately equal to" (≈) and only two significant digits are used to denote empirical estimations (in all cases, 10,000 replications were used). All the examples with multistate characters assume that the characters are nonadditive.

The consensus for each resampled data set can be estimated by means of approximate, more superficial searches (Farris et al., 1996). This is intended only as a practical approximation for large data sets, since more exhaustive analyses would be prohibitively time consuming. Although useful in practical terms, such a heuristic approximation may introduce a systematic bias, either under- or overestimating the actual group frequency. Problems can also arise when the group frequencies are calculated for each replication (as in PAUP* and Phylip); only using the strict consensus for each replication properly evaluates support. These problems were pointed out by Goloboff and Farris (2001) and by De Laet et al. (2002). In this paper, both the exact and the empirical estimations use searches exhaustive enough to guarantee that the correct strict consensus tree is found in each case.

Uninformative characters (and characters irrelevant to the monophyly of a group) can influence **J** and **B** as they were originally proposed (Carpenter, 1996; Farris et al., 1996; Harshman, 1994). Each of these two methods can be corrected for the influence of irrelevant characters by making the weight probability of each character independent. Farris et al. (1996) were the first to suggest this for **J**, and, to produce group frequencies more comparable to those obtained under **B**, they proposed a deletion (zero weight) probability of e^{-1}. For **B**, Harshman (1994) suggested generating the resampled matrices by (imaginarily) adding an infinite number of autapomorphies to the data set before resampling. Harshman's suggestion produces a weight distribution corresponding exactly to a Poisson distribution of mean 1 (see Farris et al. in Horovitz, 1999) and is thus easily implemented. Throughout this paper, when we refer to **J** or **B**, we always mean the procedures modified to make irrelevant characters uninfluential. The problems that we point out, however, will also exist for the original procedures.

Background

The amount of support for a group is the result of the interaction between the characters that favor the group and those that contradict it. However, because of character interaction, it is often impossible to evaluate relative amounts of favorable and contradictory evidence by simply counting characters. An example is Fig. 2, where group EFGH has positive overall support. The only character that provides a synapomorphy for the EFGH branch is character 1. Character 1 by itself does not provide support for EFGH, since eliminating

any of the other characters (which seem otherwise irrelevant to monophyly of EFGH) also eliminates EFGH. The only character that contradicts EFGH is character 1, the same character that appears as its synapomorphy. Increasing the weight of character 1 causes EFGH to become unsupported. The example of Fig. 2 shows that, for a given group, it may not be possible to divide the characters into those that are favorable, contradictory, or irrelevant: character 1 would fit in two categories at the same time. In other cases, a character that seems irrelevant to monophyly of a group must nonetheless be included (or excluded) for the group to have some positive support (e.g., the apparent synapomorphy of BC in Fig. 11 seems otherwise irrelevant to the monophyly of EF, but whether the character is included in the matrix determines whether the group is present in the most parsimonious trees).

Since a direct count of the characters that actually favor or contradict the group is not possible, essentially all methods to measure support do so indirectly. An example is the Bremer support and its variants, where the support is measured by comparing the fit of the data to optimal and suboptimal trees. The absolute (Bremer, 1988, 1994) and relative (Goloboff and Farris, 2001) Bremer supports measure two different aspects of group support. One aspect of the support is the absolute amount of favorable evidence, measured by the absolute Bremer support. The other aspect is the ratio between favorable and contradictory evidence, measured by the relative Bremer support. Ideally, these two quantities should be measured separately, because they represent two aspects of the support that can vary independently, but in practical terms it will often be preferable to combine them in a single value. **J** and **B** provide a single measure.

Several interpretations of **B** have been advanced by different authors (see Berry and Gascuel, 1996). A common interpretation (Efron, 1979; Felsenstein, 1985) is that **B** measures the probability of recovering a given group if a data set for the same organisms is to be sampled again from scratch (that is, a measure of stability under those specific circumstances). If a data set is to be sampled again from scratch, a supported group may indeed be less likely to be recovered again than some other unsupported group. Thus, neither lower frequencies for better supported groups nor the influence of autapomorphies are necessarily a problem when **B** is intended as a measure of stability (whether the assumptions of the method are met and whether there is any use for methods that predict what could happen if systematists were to throw away all their data in the future are different questions).

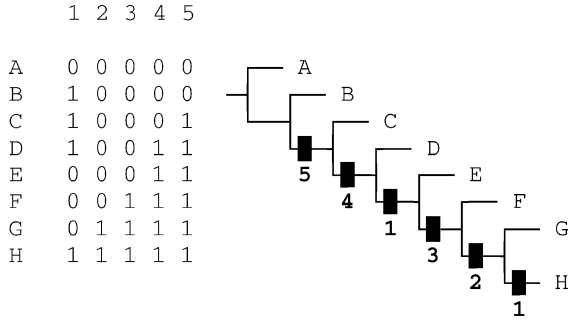

Fig. 2. A case where the characters in the matrix cannot be divided into favorable, indifferent, or contradictory of a given group. The only character that supports EFGH is the same character that contradicts it.

Support, however, is logically different from stability. Stability can be defined only by reference to some factors (e.g., a group stable under addition of characters may be very unstable under addition of taxa or under recoding of some characters). Unlike stability, support depends exclusively on presently available evidence (and, of course, assumptions or theories used to interpret that evidence). As first proposed explicitly by Farris et al. (1996), resampling methods such as **J** or **B** can be used to indirectly detect the relative amounts of favorable and contradictory evidence, rather than as statistical measures of confidence or stability. As discussed by Farris et al. (1996), resampling evaluates support because the frequency with which replicates display a given group will be determined by the relative amounts of favorable and contradictory evidence. While some strong statistical assumptions are necessary to interpret **J** or **B** as confidence levels, no statistical assumptions are necessary to interpret them as simply measuring the observed amount of support (Farris, 2002:352). However, some possible situations (unproblematic for a statistical interpretation) become problematic from this perspective; for example, the measure should never indicate a group contradicted by the data as "better supported" than a group with positive support. Since we intend to measure support (not stability or confidence), we propose in this paper possible ways to correct these problems.

Although well-supported groups will often survive sensitivity analyses (sensu Wheeler (1995); changing the parameters of the analysis), this need not be so. A group that shows up in all the parameter space may nonetheless be poorly supported (e.g., a group supported by 10 A → T transitions and contradicted by 9 A → T transitions). Similarly, a group with high "support" (in the sense used here) may nonetheless be unstable to changes in some parameters (e.g., a group supported by 0 → 1 changes in several additive characters with some members also having state 2 has a high "support" but is lost if the character is made nonadditive). This difference does not indicate conflict between sensitivity analysis and measures of support. Evaluating a hypothesis requires considering the evidence in the light of accepted ("background") knowledge or theories (see Farris, 1995; Popper, 1972). Whether an observation is seen as

corroborating or contradicting a hypothesis may itself depend on what knowledge is taken as accepted. Conclusions that depend critically on accepting some prior knowledge for which there is little basis are themselves weakly established, and this is what sensitivity analysis examines. In contrast, measures of support (such as **J** or Bremer support) evaluate quantities more directly related to the evidence itself.

Weights and symmetry

When some of the characters have higher weights or costs, both **J** and **B** can lead to wrong conclusions with regard to support. Consider the case in Fig. 3, where one character of weight N is in conflict with N characters of weight one. Under **J**, the group BC will not be supported by a given resampled matrix when the first character is eliminated (p) or when no character is eliminated (q^{N+1}). Thus, $j_{(\mathrm{BC},p)} = 1 - (p + q^{N+1})$; for a sufficiently large N, $j_{(\mathrm{BC},p)}$ tends to q (normally set to 0.6321). Thus, when N is large, group BC will appear as relatively well supported—but in fact the original data do not support this group. Under **B**, the group CD instead will appear as supported; in a given resampled matrix, group CD will be displayed when the weight has been increased for more than one of the (many) characters supporting CD and not for the (single) character supporting BC. The exact calculations are more difficult in this case, but for $N = 10$, $b_{(\mathrm{CD},1)} \approx 0.52$.

A possible correction for cases such as this, implemented in PAUP* (Swofford, 1998), is decomposing the characters in several variables of unit weight or cost (including N copies for a character with weight N or decomposing additive characters in binary variables). This correction is inapplicable to analyses under implied weights (since the character weights are determined during the analysis) or to step-matrix characters (where different transformations have different costs). A worse problem is that it alters the results in undesirable ways. Consider Fig. 4, which is a randomly generated data set (0 and 1 equiprobable for each cell). The consensus tree for that data set has two groups, both poorly supported; $b_{(\mathrm{ECF},1)} \approx 0.47$ and $b_{(\mathrm{CF},1)} \approx 0.45$. If the matrix is reanalyzed by including 10 copies of each character, the groups ECF and CF appear now as strongly supported; $b_{(\mathrm{ECF},1)} \approx 0.97$ and $b_{(\mathrm{CF},1)} \approx 0.93$. It is hardly surprising

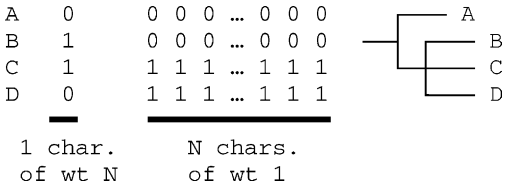

Fig. 3. A hypothetical example of character conflict.

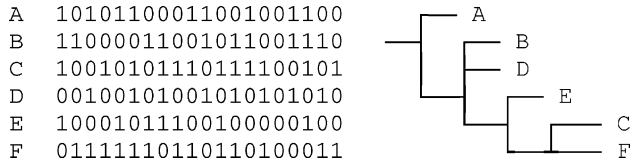

Fig. 4. A randomly generated data set (0 and 1 equiprobable for each cell), demonstrating the problem of replacing characters with weight N by binary characters of weight 1. The two groups in the consensus are poorly supported; both have a low frequency under resampling, but when all columns are multiplied by 10, the frequencies approach 1.

that the decomposition fails, since (to start with) the example of Fig. 3 showed that, for **J** or **B**, X characters of weight N are *not* equivalent to N characters of weight X.

The influence of the weights or costs can be eliminated if the resampling is done in such a way that the probability of increasing the weight of the character equals the probability of decreasing it. This also explains the difference in the error produced in **J** and **B**. The mean in **J** equals $1 - p$; a character has a chance p of being deleted (downweighted) and a chance q of being retained (upweighted relative to the mean); the asymmetry in **J** is q/p, or 1.71. In **B** (with Poisson of mean 1), the probability of deletion (downweighting) is 0.3679, and the probability of weight 2 or more (upweighting relative to the mean) is 0.2642; the asymmetry is thus 1.39. The Poisson distribution is closer to being symmetric than the deletion only, and this is why the deviation from equal probabilities for BC and CD is less pronounced for **B**. In the Poisson distribution, the probability of downweighting is more than the probability of upweighting, while in **J** the opposite is true; this is why **J** wrongly favors BC, and **B** wrongly favors CD.

Although **B** with a Poisson of mean 1 is close to symmetric, it is not actually so. The method could be corrected by using a Poisson distribution with a proper mean, such that this symmetry occurs. The symmetry also occurs under *J*, when $p = q = 0.5$, but this is a very strong resampling function; only extremely well supported groups will survive such a resampling, and the problems pointed out in the next section become more acute.

The required symmetry can be most easily obtained by modifying **J**, so that a character could be either upweighted (by a given factor) or deleted with equal probability, p. Consider first the case where a character can only be upweighted and never deleted, with probability p. For the example in Fig. 3, the group CD will be absent whenever the first character is upweighted (p) or when no character is upweighted (q^{N+1}), so that $j_{(\mathrm{CD},p)} = 1 - (p + q^{N+1})$. The probability of obtaining group CD in a given replicate with upweighting equals the probability of obtaining the group BC with deletion. If both actions are done at the same time, the effects will cancel. Thus, p will be the same for deletion or

upweighting, and $q = 1 - 2p$; when resampling a matrix, each character has a probability $2p$ to be changed, and if changed, it can be duplicated or deleted with equal probability. We will call this method the symmetric resampling, or SR; $sr_{(G,p)}$ denotes the frequency of group G when probability of up or downweighting equals p. As in **J**, the frequency for a group supported by a single uncontradicted character will be $1 - p$. SR can be applied to any kind of weighting scheme, such as successive weighting (Farris, 1969), implied weighting (Goloboff, 1993), or weighting of state transformations (including asymmetries in transformation costs).

Supported groups with low frequencies

Another problem with resampling methods in general arises when groups of low frequencies are considered more carefully. When the search for each resampled matrix is careful enough (and the resampling function is symmetrical), any group with frequency above 0.5 is certain to be supported, but the opposite is not true: many supported groups have frequencies well below 0.5. The frequency for actually supported groups can be calculated, regardless of whether it is above or below 0.5, but it is nonetheless difficult to interpret those values. Examples of groups with actual support but resampling frequency below 0.5 were provided by Harshman (1994).[1] An example is shown in Fig. 5; the group BC is supported, but $sr_{(BC,0.33)} = 0.3792$; the unsupported groups (BD, BE, etc.) have a SR frequency (for $p = 0.33$) of 0.0226. In general, when there are N mutually incompatible groups each supported by the same number of characters, each individual group can have at the most a frequency of $1/N$.

That condition does not represent the only problematic situation. Consider the case of Fig. 6, where group EF is supported, but $sr_{(EF,0.33)} = 0.2626$. Note that the three characters are required for the group EF to be supported; most possible combinations of weights between 0 and 2 (except for the original weights) will cause EF to be absent. Also note that group EF is not actually contradicted by any character; it is simply ambiguously supported. An even lower frequency can occur; for Fig. 7, group HI is supported and uncontradicted, but $sr_{(HI,0.33)} = 0.0798$. This is more than a problem of scale; for Fig. 7, contradicted group EFG (with $sr_{(EFG,0.33)} = 0.1611$) has twice the frequency of supported HI.

Examples such as these imply that the frequency itself does not provide an accurate measure of degree of support. Consider the case in Fig. 7, where group HI, with some actual support, has $sr_{(HI,0.33)} = 0.0798$.

[1] Harshman, however, did not draw any special conclusion from the existence of groups with actual support but frequency below 0.5.

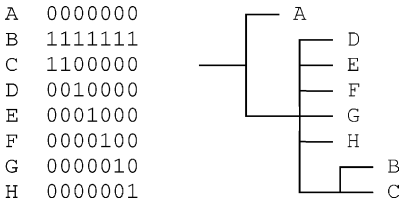

Fig. 5. Supported group BC has frequency below 0.5 under symmetric resampling.

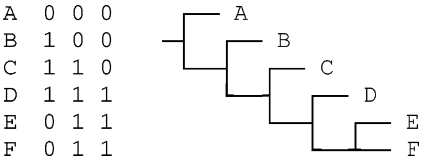

Fig. 6. Supported group EF has frequency below 0.5 under symmetric resampling.

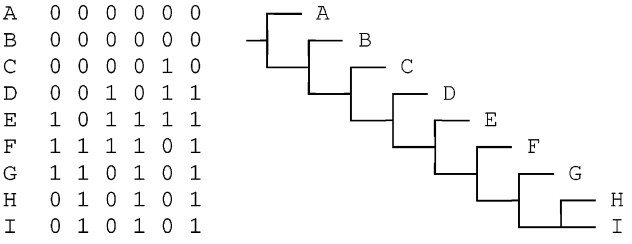

Fig. 7. Supported group HI has frequency below 0.5 under symmetric resampling. Contradicted group EFG has a higher frequency (0.1611) than supported group HI (0.0798).

Imagine that N characters joining HI and N characters joining GH are added to the matrix. As N becomes larger, $sr_{(HI,0.33)}$ approaches 0.5. For HI, adding those characters actually lowers support *and* increases frequency.

Three ways to solve this problem have been examined: (1) calculating the difference in frequencies between the group and the most frequent contradictory group, (2) calculating the frequency for low values of p, and (3) calculating the frequency slopes as a function of p.

Frequency differences

Groups with zero support can have a frequency up to 0.5. Consider the case where N characters support a BC partition and N characters support a CD partition. For a large N and any p, both $sr_{(BC,p)}$ and $sr_{(CD,p)}$ tend to 0.5. While for each group the frequency itself is well above zero, the difference between $sr_{(BC,p)}$ and $sr_{(CD,p)}$ is (sampling error aside) exactly zero—just like the actual support. This suggests that what actually measures the support is not the frequency itself, but instead the

difference in frequency between a group and the most frequent contradictory group. That interpretation casts doubts on the use of the absolute frequencies as measures of support, even in the case of frequencies above 0.5. A group with frequency 0.6 will be quite poorly supported if it is contradicted in 0.4 of the replicates, while it will be quite strongly supported if it is never contradicted; using just the absolute frequency will simply not distinguish between these two situations.

We will call the difference in frequencies GC (for "Group present/Contradicted"), and the difference in frequencies for a given group AB under resampling probability p will be called $gc_{(AB,p)}$. Such a measure varies between −1 and 1, and it can produce meaningful evaluations of the support for all groups, not only those with absolute frequency above 0.5. GC values of −1, 0, and 1 indicate (respectively) maximum contradiction, indifference, and maximum support. Consider the example in Fig. 5: $gc_{(BC,0.33)} = sr_{(BC,0.33)} - sr_{(BD,0.33)} = 0.3566$. For the examples in Figs. 6 and 7, since groups EF or HI never appear contradicted in any replication, $gc_{(EF,0.33)} = sr_{(Er,0.33)} = 0.2626$ and $gc_{(HI,0.33)} = sr_{(HI,0.33)} = 0.0798$. Contradicted group EFG of Fig. 7 (which had a higher frequency than HI) has $gc_{(EFG,0.33)} = -0.4487$.

Note that GC must count the number of occurrences of individual cases of contradiction, not simply the number of times that a group is contradicted. In the example of Fig. 8 (similar to Fig. 5, but with more characters), the group BC appears supported in ≈0.28 of replicates and contradicted in ≈0.37. If simply the number of times that BC is contradicted are counted, the impression that BC is unsupported would be given. However, each of the groups that contradicts BC (BD, BE, BF, BG, and BH) individually appears in only ≈0.07 of the cases. Even if more replicates display some contradictory group, the (supported) group BC is recovered in more replicates than any of the individual contradictory groups.[2]

The GC value for a group is more easily calculated considering that the group is contradicted when the consensus directly contradicts it (thus counting cases where the group is unresolved in the consensus as neither favorable nor contradictory). However, the consensus may display the group as unresolved even in cases where no actual underlying tree displayed the group. An example is in Fig. 9. When resampling, if the second character dominates over the first (either because it has been upweighted or because the first character has been deleted), the shortest trees display ED. If instead the first character dominates over the second, none of the possible multiple shortest trees (shown in Fig. 10) displays group DE, but their consensus is unresolved. Thus, if the GC difference is calculated using the strict consensus for each replication, $gc_{(DE,0.33)} = 0.3333$, even when group DE is unsupported. To prevent this, the number of occurrences of a group must be counted as cases where the group occurs in some of the most parsimonious trees for the resampled data set (and regardless of whether the group represents a zero-length branch). If that is done, it is seen that $gc_{(DE,0.33)} = 0$ (group DE occurs in 0.3333 of replicates, just like contradictory groups CD or EF; note that in this case the sum of frequencies of all possible groups may not add to 1). The GC value calculated using the strict consensus, which we will call GC′, may in practice be a good empirical approximation to the actual GC value, since it is much more easily obtained. The GC′ values, however, are only

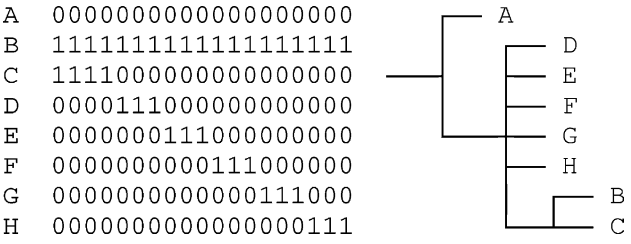

Fig. 8. Example to demonstrate that GC values must be calculated by counting individual instances of contradictory groups. BC is contradicted in more replications than it is supported, but all the individual groups that contradict it (BD, BE, BF, BG, and BH) have frequencies well below the frequency of BC. *Note*: resampled Bremer support (obtained by calculating the average Bremer support for each of the resampled matrices) for group BC is −0.47.

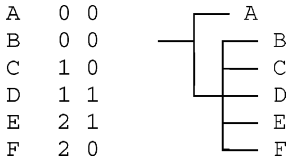

Fig. 9. A case where computing GC using the strict consensus for each replication misleadingly indicates positive support for group DE.

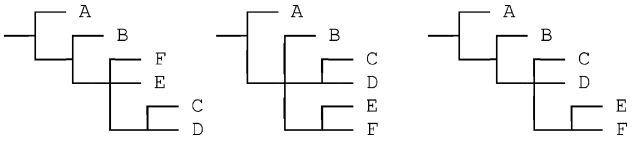

Fig. 10. The three equally parsimonious trees for the data set of Fig. 8, when the first character dominates over the second.

[2] In this example, the uncorrected GC values are similar to those obtained by calculating a resampled Bremer support (i.e., the average value of Bremer support for the group in the resampled matrices), which produces a "negative" support for group BC. However, the uncorrected GC values and resampled Bremer supports are not equivalent. Symmetric resampling of the Bremer supports has additional problems as measure of support: when there are N_{CD} apparent synapomorphies for CD and N_{BC} for group BC, the resampled Bremer support is simply $N_{CD} - N_{BC}$ (regardless of the ratio N_{CD}/N_{BC}).

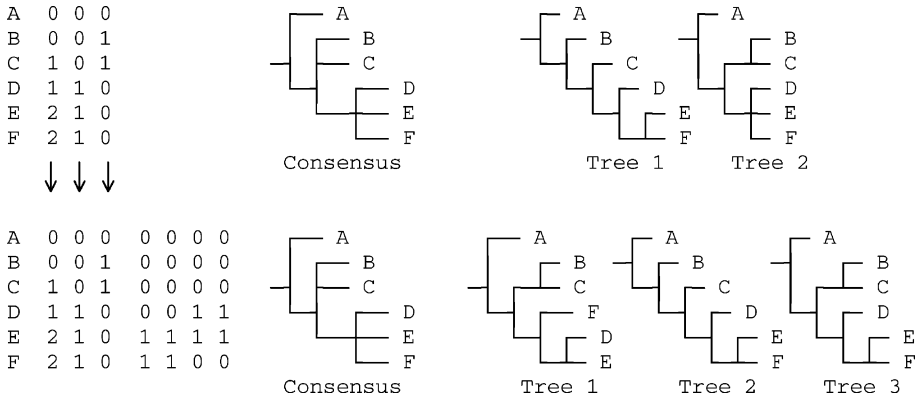

Fig. 11. Two cases where GC values are misleading. Group EF (appearing in only one of the two most parsimonious trees, shown at the right of the consensus) is unsupported, but it has a positive GC value, for both matrices.

estimations (under some situations, biased) of the actual GC.

There are some cases where the GC values are misleading. An example is shown in Fig. 11 (top). The consensus does not display group EF. Group EF appears in some of the most parsimonious trees for the data set (when C is closer to DEF than to B, optimization of the first character unambiguously indicates state 2 as a synapomorphy of group EF), but not in all (when C is closer to B than to DEF, state 2 of the first character can be considered plesiomorphic within DEF, even when E and F are placed together). In the resampled data sets, the group EF may be supported or ambiguous, but never contradicted; $gc_{(EF,0.33)} = 0.2233$, thus wrongly indicating that EF is supported. In this case, EF is unsupported because of ambiguity, not because it is contradicted by some characters. If these were the only circumstances where GC can attribute support to unsupported groups, this would mean that the values for actually supported groups will normally not be biased (since for actually supported groups, only actual contradiction—which precludes this situation—would lower the GC values). That, however, is not the case. Modifying the example by adding some characters (Fig. 11, bottom), the amount of evidence contradicting EF is almost the same as the evidence favoring it, and the consensus is still the same; unsupported EF is ambiguous because of character conflict, but $gc_{(EF,0.33)} = 0.1867$.

Low resampling strengths

An alternative solution to the problem of groups with positive support but low frequencies comes from considering the differences in frequency at low resampling strengths—that is, when the probability p of up- or downweighting is very low. Fig. 12 illustrates $sr_{(G,p)}$ for different values of p. By necessity, if group G is supported, $\lim_{p\to 0} sr_{(G,p)} = 1$, and if group G is unsupported, $\lim_{p\to 0} sr_{(G,p)} = 0$. This implies that, for any data set, there is a resampling strength where no supported group has a frequency below 0.5, and then there may be a given p (call it p_r) below which the resampling frequency will rank all the groups in the correct order of support. How close p_r is to 0 will depend on the data set. If the support for some of the groups is extremely low (but positive), p_r may be very close to 0.

Using low values of p, however, has an undesirable effect on the precision with which the estimation can be done. Using a low p has the effect that the resampling will not be able to discriminate differences in support among groups with relatively high support; in practice, if few replications are done, they will all have (estimated) frequencies of 1. The differences in (real) frequency will normally be very small; three groups with relatively high, medium, and low support may have frequencies of 0.9999, 0.9995, and 0.9990. The number of replications necessary to estimate the frequencies with such a degree of precision may be prohibitively large. Using very low values of p will effectively identify groups with extremely low (but positive) support, but the groups with support above a certain threshold will all be considered as equivalent. In some sense, this is the opposite of what happens under larger values of p, which cannot discriminate among groups with low support; the emphasis of the study may require evaluation of groups in one or the other category.

Frequency slopes

Further consideration of Fig. 12 suggests that the trajectories of the group frequencies, as a function of p, may themselves provide information on support. Groups with positive support always have negative slopes; groups with no overall support (i.e., ambiguous or contradicted) have negative or positive slopes, depending on the value of p. However, for low values of p, all groups with no support will necessarily have positive

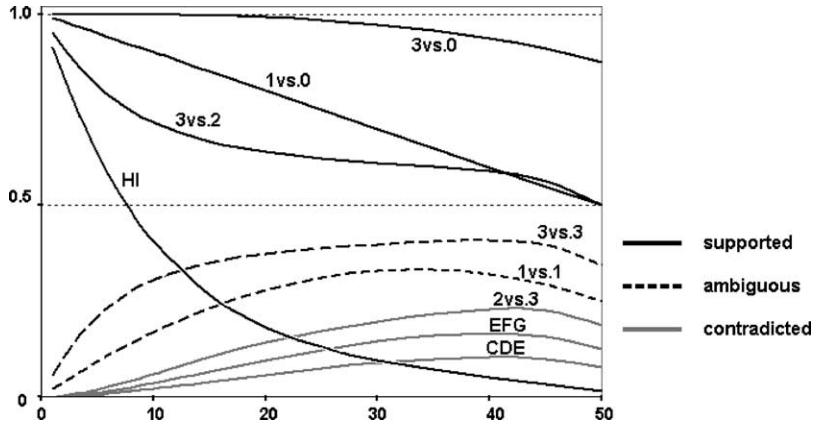

Fig. 12. Curves of frequency under symmetric resampling, as a function of different values of p. The groups HI, EFG, and CDE correspond to the example in Fig. 7.

slopes (since $\lim_{p\to 0} \mathrm{sr}_{(\mathrm{G},p)} = 0$), and all groups with positive support will necessarily have negative slopes (since $\lim_{p\to 0} \mathrm{sr}_{(\mathrm{G},p)} = 1$). Groups with high support have negative slopes that approach 0, and groups with low support have more negative slopes. Groups with slopes that are very close to 0 could be either strongly supported, very ambiguous, or strongly contradicted groups; the frequencies for such groups will be close to 1.0, 0.5, or 0 (respectively). The slope could be introduced in the support measure in several ways, for example, by multiplying the absolute frequencies by a factor that depends on the slope.

In principle, estimating the slope of $\mathrm{sr}_{(\mathrm{G},p)}$ accurately will require a significant amount of computational effort. This would require computing $\mathrm{sr}_{(\mathrm{G},p)}$ for different (but close) values of p and then using those values in a regression. Each of the values of $\mathrm{sr}_{(\mathrm{G},p)}$ will, however, have a significant error; estimating the slope implies calculating $\mathrm{sr}_{(\mathrm{G},p1)} - \mathrm{sr}_{(\mathrm{G},p2)}$ when p_1 approaches p_2. The error in $\mathrm{sr}_{(\mathrm{G},p1)} - \mathrm{sr}_{(\mathrm{G},p2)}$ can be up to twice the error in each estimation; what is worse, the magnitude of that error may be very large relative to $p_1 - p_2$.

For small matrices, an exhaustive enumeration can be used to compute the slope, using the approach illustrated in Fig. 1. If the number of changed/unchanged characters is recorded for each matrix examined, it is possible to calculate the actual values of $Y_{i,p}$ (and $X_{i,p}$) for any value of p, without the need to repeat the tree-searching calculations. For small matrices, this exhaustive enumeration can be used to compute the slope almost exactly.

For matrices with larger numbers of characters, exhaustive enumeration is not possible. An estimation obtained by sampling from among possible rearrangements of the weight vector (as before, each rearrangement must be examined only once and must have the same probability to be sampled) and calculating the estimated frequency as $\mathrm{sr}'_{(\mathrm{G},p)} = \sum X_{i,p} / (\sum X_{i,p} + \sum Y_{i,p})$ (note that the denominator does not add to 1) will work only for very large numbers of replications. When the number of replications (= rearrangements of the weight vector) is very large, $\mathrm{sr}'_{(\mathrm{G},p)}$ converges to $\mathrm{sr}_{(\mathrm{G},p)}$ (and $\sum X_{i,p} + \sum Y_{i,p}$) approaches 1), but the estimator is significantly biased for smaller sample sizes.[3] Aside from the bias, it has a significant dispersion, and, therefore, even for relatively modest numbers of characters (40 or 50), different estimations based on 1000 replications produce very different results; much larger numbers of replications are necessary to produce more stable results. Although estimations of the slope for $p = 0.3333$ should be more accurate (since in that case all rearrangements have the same probability), the slope at that point will not always produce proper evaluations of support, since the frequency of unsupported groups sometimes peaks below $p = 0.3333$ and thus has negative slopes at that point.

A more accurate and less biased estimation can be obtained by doing a normal estimation of the group frequency under a change probability p (i.e., simply generating matrices where each character is duplicated or removed with probability p, allowing duplicate matrices to be examined) and then extrapolating the probabilities associated with each replication to the vicinity of p. Thus, the frequency estimated at point p is (as usual) the proportion of replications that displayed the group, and at point p' the estimated frequency is $\mathrm{sr}'_{(\mathrm{G},p')} = (\sum v_i)/(\sum v_i + \sum w_i)$, where $v_i = Y_{i,p'}/Y_{i,p}$ if

[3] It is easy to show that the estimator is biased, for a sample size of 1. Imagine that there are N_y rearrangements of the weight vector that display group G (with an average associated probability $Y_{i,p}$ equal to α) and N_n rearrangements that do not (with an average associated probability $X_{i,p}$ equal to β). Then, sampling one rearrangement, $\mathrm{sr}_{(\mathrm{G},p)}$ will be estimated as 1 with frequency $N_y/(N_y + N_n)$ and as 0 with frequency $N_n/(N_y + N_n)$; $\mathrm{sr}_{(\mathrm{G},p)}$ is then estimated (on average) as $N_y/(N_y + N_n)$, but the true frequency of the group is $\alpha * N_y$. The estimator is unbiased only when $\alpha = 1/(N_y + N_n)$; it is easy to see that this also implies that $\beta = 1/(N_y + N_n)$, and then the estimator is unbiased only when $\alpha = \beta$ (which is guaranteed only under $p = 1/3$).

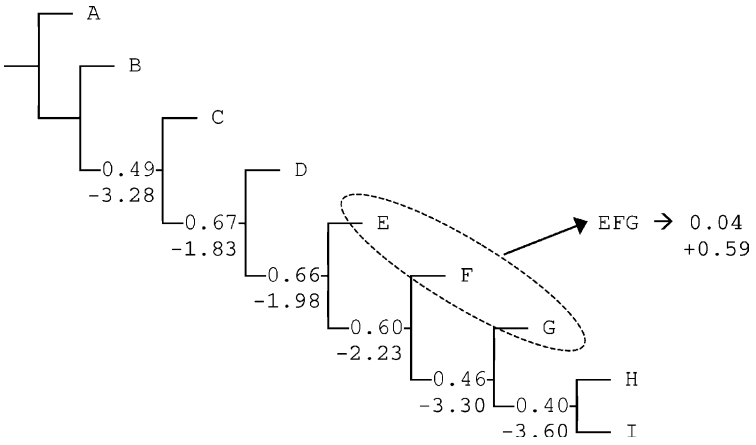

Fig. 13. Frequencies (on branches) and slopes (below branches), for $p = 0.10$, for the data matrix of Fig. 7, estimated with 10,000 replications. The least supported group is group HI (with the most negative slope); group EFG (contradicted by the data, with a higher frequency at $p = 0.33$) has a lower frequency at $p = 0.10$ and a positive slope.

replication i displays the group ($v_j = 0$ otherwise), and $w_i = X_{i,p'}/X_{i,p}$ if replication i does not display the group ($w_i = 0$ otherwise).[4] Using this procedure (for $p = 0.10$), the unsupported group EF of Fig. 11 (which appeared as supported with GC) is properly identified for both matrices, with a slope of 1.07 (frequency ≈0.15) for the top matrix, and 1.63 for the bottom one (frequency ≈0.35). As another example, Fig. 13 shows the value of slopes and frequencies (at $p = 0.10$), for the data set of Fig. 7; group HI is correctly shown as supported (with a negative slope, −3.60) and contradicted group EFG (which had a higher frequency than HI at $p = 0.33$) has a lower frequency and a positive slope (+0.59). Contrast this with a hypothetical case where 15 characters support group AB and 15 support BC (so that neither AB nor BC have any actual support); at $p = 0.10$, both AB and BC have a frequency ≈0.42 (slightly above the frequency of supported HI), but they have a positive slope (+0.38).

Acknowledgments

We acknowledge support from CONICET (PIP 4974, PEI 0324/97, and Beca Postdoctoral), Agencia Nacional de Promoción Científica y Tecnológica (PICT 98 01-04347), and NFR Grant 10204. P.A.G. benefited from discussions with M.I. Giannini, N. Giannini, and J. Miller. The extensive comments from one of the reviewers (Taran Grant) were very helpful. We also thank the editor, A. Kluge, for help in improving the manuscript.

[4] This method is also biased, as easily shown for sample sizes of 1. In this case, the ratios between $Y_{i,p'}$ and $Y_{i,p}$ or between $X_{i,p'}$ and $X_{i,p}$ are irrelevant; the value of $sr_{(G,p)}$ will be estimated always as 0 or 1, and $sr_{(G,p')}$ will be estimated (on average) as $sr_{(G,p)}$—regardless of the true value of $sr_{(G,p')}$.

References

Berry, V., Gascuel, O., 1996. On the interpretation of bootstrap trees: appropriate threshold of clade selection and induced gain. Mol. Biol. Evol. 13, 999–1011.

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.

Carpenter, J., 1996. Uninformative bootstrapping. Cladistics 12, 215–220.

De Laet, J., Farris, J., Goloboff, P., 2002. PAUP and Phylip can attribute support to unsupported groups. Cladistics.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. Ann. Stat. 7, 1–26.

Farris, J., 1969. A successive approximations approach to character weighting. Syst. Zool. 18, 374–385.

Farris, J., 1995. Conjectures and refutations. Cladistics 11, 105–118.

Farris, J., 2002. RASA attributes highly significant structure to randomized data. Cladistics 18, 334–353.

Farris, J., Albert, V., Källersjö, M., Lipscomb, D., Kluge, A., 1996. Parsimony jackknifing outperforms neighbor-joining. Cladistics 12, 99–124.

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

Goloboff, P., 1993. Estimating character weights during tree search. Cladistics 9, 83–91.

Goloboff, P., Farris, J., 2001. Methods for quick consensus estimation. Cladistics 17, S26–S34.

Harshman, J., 1994. The effect of irrelevant characters on bootstrap values. Syst. Biol. 43, 419–424.

Horovitz, I., 1999. A report on "One Day Symposium on Numerical Cladistics". Cladistics 15, 177–182.

Lanyon, S., 1985. Detecting internal inconsistencies in distance data. Syst. Zool. 34, 397–403.

Popper, K., 1972. Conjectures and Refutations: The Growth of Scientific Knowledge. Routledge and Kegan Paul, London.

Swofford, D., 1998. PAUP*: Phylogenetic Analysis Using Parsimony (* and other methods), Version 4. Sinauer associates, Sunderland, MA.

Wheeler, W., 1995. Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data. Syst. Biol. 44, 321–331.