

Resampling measures of group support: a reply to Grant and Kluge

Martín J. Ramírez

Museo Argentino de Ciencias Naturales—CONICET, Avenue Angel Gallardo 470, C1405DJR Buenos Aires, Argentina

Accepted 31 August 2004

Abstract

Grant and Kluge (2003) associated resampling measures of group support with the aim of evaluating statistical stability, confidence, or the probability of recovering a true phylogenetic group. This interpretation is not necessary to methods such as jackknifing or bootstrapping, which are better interpreted as measures of support from the current dataset. Grant and Kluge only accepted the absolute Bremer value as a measure of group support, and considered resampling methods as irrelevant to phylogenetic inference. It is shown that under simple circumstances resampling indices better reflect the degree of support than Bremer values. Grant and Kluge associated the resampling methods (and the use of measures of group support in general) with what they call a “verificationist agenda”, where strongly supported groups are first detected, and then protected against additional testing. They propose that identifying weakly supported groups, and then concentrating additional tests on them, will better serve science. Both programs are actually equivalent, and inert as to the selection of methods to estimate group support. The ranking of groups under a range of resampling strength is proposed as an additional criterion to evaluate resampling methods. A reexamination of the slope of symmetric resampling frequency as a function of resampling strength suggest that slopes can be problematic as well as a measure of group support.

© The Willi Hennig Society 2005.

In a recent contribution to this journal, Grant and Kluge (2003; GK henceforth) presented a critique of support measures obtained with resampling procedures (such as jackknifing or bootstrapping). In the immediately preceding issue of *Cladistics*, Goloboff et al. (2003; GEA henceforth) made a detailed discussion of these methods, analyzed biases and artifacts, and proposed several methodological developments that may improve resampling measures. Given that GK did not have the opportunity to discuss the ideas and arguments worked out in GEA¹, it is opportune to present a brief discussion here. GK (p. 411) came to the shocking conclusion that resampling estimations of group support belong to a group of methods that “are neither scientific

nor heuristic”, but “amount to mere sophistry and are irrelevant to phylogenetic inference”.

I want to distinguish three relatively independent issues in the construction of their arguments, and treat them separately in the following sections: (1) The *interpretation* of support measures (e.g., as a probability of recovering the true phylogenetic group). (2) Their *accuracy* in obtaining adequate measures of degree of support (i.e., if they measure correctly what they are intended for). (3) The *applications* of support measures (e.g., as information in the decision-making for the selection of groups of interest or for taxonomic changes).

Interpretation

GK discussed two alternative interpretations of resampling measures of group support. By one interpretation, a clade frequency obtained, e.g., from bootstrapping, is interpreted as a statistical measure of stability, confidence, or probability of recovering a true

Corresponding author.

E-mail address: ramirez@macn.gov.ar

¹They discussed contributions as of 2002, although some of the most relevant arguments were already available in Goloboff and Farris (2001), and in Farris (2002).

phylogenetic group. In a second interpretation, a resampling measure is construed as the degree of support from the presently available evidence, that is, support from the current dataset. GK (pp. 383, 411) concurred with Farris et al. (1996), Farris (2002, p. 351–352), and with GEA (p. 326), in this second interpretation of support. As noted by Farris (2002) and GEA, interpreting support measures as statistical confidence levels requires a series of strong assumptions that are not necessary when they are conceived as an observed amount of support. Resampling methods employ procedures similar to those used for statistical estimation, but they are not naturally attached to probabilistic interpretations. Recent developments are less amenable to statistical interpretation: a symmetric resampling (GEA, p. 327), by combining the properties of bootstrapping and jackknifing, neutralizes biases produced by differential weights or costs; and the difference in frequencies of a group and the most frequent contradictory group may be a better indication of support for groups with low, but positive support (GEA, p. 328). Much of the criticisms in GK are restricted to resampling measures as confidence values, and do not need to be addressed here.²

Kluge (2002, p. 591) has previously stated that bootstrap and jackknife, even when interpreted “as measuring hypothesis support, they assume a conditional (frequency) interpretation of probability, where no frequency exists”. (In GK such an assertion is only linked to applications of resampling with a confidence interpretation.) Kluge did not explain why such an assumption is necessary; there is nothing in the methods explored by Farris et al. (1996) and GEA that requires it. The dataset is not supposed to be a random sample, and more important, whether yet unknown data have the same or a different distribution is irrelevant to the estimation of the support from the *present* dataset. Of course we have the general expectation that today’s strongly supported groups will not be strongly contradicted in the future, but this falls far from interpreting support levels strictly as confidence values. It seems that Kluge’s objection is based on associating resampling procedures such as jackknifing with assigning probabilities or frequencies to singular historical events that have already occurred. He used a simple example from human history (Kluge, 2002, p. 586) to discuss the issue of probabilistic inference of historical events, which can also be used to illustrate the issue of support. The example involves the hypothesis that William defeated Harold in the battle at Hastings, in the year 1066. Kluge mentions some written records that are congruent or

supportive with that story. Let’s suppose that we want to evaluate the support for the hypothesis in the records, and collect more data to that purpose. Imagine three possible results, all involving independent written records that we will consider as equally informative and trustable:

(a) Eight records support the hypothesis.

(b) Three records support the hypothesis; one record reports Harold’s death in 1062.

(c) Eighteen records support the hypothesis; 15 records place Harold living in Spain between 1060 and 1085 as a shoemaker, and 15 more place Harold in Belgium in the same period, as a baker.

The three scenarios corroborate the hypothesis, but with different strengths. In scenario (a) the hypothesis is supported by many records and never contradicted, in (b) is moderately supported, and contradicted once, and in (c) is supported, but there are two alternative hypotheses that are only slightly suboptimal. If we calculate jackknifing frequencies (symmetric resampling, $P = 0.33$) for the hypothesis being supported more often than contradicted, we obtain 0.99, 0.77, and 0.56, respectively. This produces a ranking of scenarios according with the proportions of supporting and conflicting evidence. By doing this, we have not assumed that the known records are a representative sample of the records that may exist or have existed; we just describe the present situation with the available data. Strong assumptions have been made (the discretization of written records; that they are equally trustable and informative), but we avoid any statistical interpretation of the results, that is, we do not try to calculate the probability that the hypothesis has actually occurred, which would require even stronger assumptions. Further records may as well change the support values, or suggest a different hypothesis altogether. In this simple historical example, the use of a resampling technique to estimate proportions of supporting and conflicting evidence is superfluous. We could just count the number of supporting and conflicting pieces of evidence, and construct a simple formula to obtain some sensible index. However, in all but the simplest phylogenetic datasets, the characters interact in complex ways, such that it is not possible to classify them as favoring, opposing, or irrelevant for a given monophyly hypothesis (GEA, p. 326). For example, wings support monophyly of pterygote insects, but at the same time contradicts it, because some derived insects do not have wings (e.g., fleas).

A further set of objections pertains to resampling measures without a statistical interpretation. In this line, GK’s criticisms are difficult to discuss, because they only express very general ideas (frequencies are not obtained directly from the complete dataset; frequencies are not logically related to objective support) without explaining how or when resampling

²GK (p. 396) also criticized resampling measures because “characters are not necessarily independent”. This is a critique of most quantitative methods related to phylogeny reconstruction (including those embraced by GK) and should be discussed in a broader context.

measures may depart from a correct description of degrees of support.

During the construction of a resampling measure, the pseudoreplicates are intended to produce controlled alterations in the relative influence that the characters have on the dataset, thus permitting some of the character conflict and interactions to be made manifest, and to be measured in the process. These complex interactions are so far not detectable in the complete dataset. GK associated resampling methods with partitioned analyses:

“[Computer-intensive sampling] methods fail to measure objective support and therefore lack heuristic. For example, the parsimony jackknife relies on sampling frequencies derived from partitioned analyses and never evaluates the congruence of all critical evidence in a simultaneous test; therefore it cannot be said to measure objective support.” (GK, p. 397)

In a typical partitioned analysis, two or more subsets of the data are analyzed separately (different markers, sequences versus morphology, morphology versus behavior, etc.), and their results combined in some way. When many partitions converge in supporting a given group, it is interpreted as a high support level from independent sources. In a partitioned analysis, a very special emphasis is placed in the independence of the lines of evidence. Jackknifing does not have the same procedures or the aim of a partitioned analysis. The “partitions” (the pseudoreplicates), have a lot of overlapping among each other, and are not independent; and there are high numbers of pseudoreplicates, all drawn from the same dataset. Without more elaboration, this association of jackknifing with partitioned analyses simply relates to the general idea that some specific character interactions are only apparent when the dataset is complete, or almost complete.

It is curious that for other methodological procedures GK accepted that it is useful to leave part of the data to express its own hierarchic structure. They conferred a “high heuristic value” to the “long-branch extraction” procedure of Siddall and Whiting (1999) of pruning branches of the tree to evaluate if they suffer from branch attraction. By this procedure, two terminals or sets of terminals suspected of branch attraction are alternatively eliminated from the dataset; the trees are recalculated from the reduced dataset versions, and compared with the trees from the complete dataset. In other words, long-branch extraction involves a partitioned analysis at the level of taxa.

GEA (pp. 330–332) proposed that the slope of the resampling frequency as a function of the resampling strength may work as a promising measure of group support (see also Miller, 2003 for a similar approach to total tree support). The slope is measured for low resampling strengths, that is, where the probability P of up- or down-weighting tends to zero, and the dataset

tends to integrity. Estimating slopes requires a high computational effort, and GEA (p. 331) proposed some possible approximations. A simplification of the calculation of slopes can take advantage of the fact that the exact frequencies under $P = 0$ are known (1 for supported groups, 0 for unsupported ones); one additional point at a low resampling strength will suffice to estimate the slope. In the present context, this is relevant because the slope is calculated from a reference point representing the complete dataset ($P = 0$), thus satisfying GK’s idea that the complete dataset should be taken into account. Frequency slopes are still not well explored, and seem to have problems as well (see below, Fig. 2).

The last of GK’s criticisms of resampling measures without statistical interpretation is also very general and unspecific:

“[Jackknifing] “support” values cannot be interpreted as assessing the relative objective support provided by those data because [...] resampling frequencies are logically unrelated to degree of corroboration” (p. 396)

This looks like an unfair description of the state of the art in resampling measures, even as to 2002. As can be seen in GEA and the references therein, a lot is known about the logical connection between resampling and support. This understanding permitted the development of corrections of biases with uninformative characters in bootstrap (Harshman, 1994), the deduction of resampling strengths for making jackknife and bootstrap values comparable under certain circumstances (Farris et al., 1996), and correction for the effect of implied weights and ordered states (Farris et al. in Swofford, 1998; Horovitz, 1999). Perhaps GK do not see the connection between resampling measures and support because they concentrated in a “redefinition” of support (p. 383, “the degree to which critical evidence refutes competing hypotheses”) which seems very similar to a rewording of the Bremer support (Bremer, 1988, 1994), the only support measure admitted by GK as having some “heuristic value”. A limitation with Bremer values is that they only describe the absolute support, not taking into account the ratio between favorable and contradictory evidence (Goloboff and Farris, 2001; GEA). For this discussion, it seems worth examining the situations where Bremer and a resampling measure produce different rankings of groups.

Accuracy

The most basic test for a measure of support is the agreement with the strict consensus tree produced by the dataset; spurious groups with positive support, and supported groups with negative values are indicative of

A	0000	0	00	000	0000000000000000
B	1000	0	00	000	0000000000000000
C	1000	1	00	000	0000000000000000
D	1000	1	00	000	0000000000000000
E	0100	0	00	000	0000000000000000
F	0100	0	11	000	0000000000000000
G	0100	0	11	000	0000000000000000
H	0010	0	00	001	0000000000000000
I	0010	0	00	110	0000000000000000
J	0010	0	00	111	0000000000000000
K	0001	0	00	000	0000001111111111
L	0001	0	00	000	11111111110000
M	0001	0	00	000	11111100001111

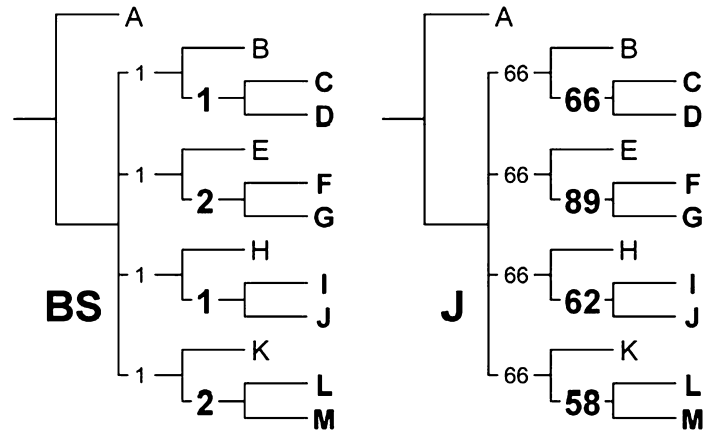


Fig. 1. Example where Bremer support values (BS) and jackknifing frequencies (J) produce different rankings of groups. The low J-value for group LM reflects more accurately the proportion of supporting and conflicting characters. J under symmetric resampling, $P = 0.33$, calculated with TNT (Goloboff et al., 2004).

problems with the measure. A good measure also has to produce values in direct relation to the degree of positive support and conflict. This relation is difficult to define in real datasets, but can be understood in simpler, artificial situations; when a method is known to fail in a simple situation, it is easier to trace the same type of artifact in a real dataset. The example of Fig. 1 is construed such that both the number and proportion of characters supporting and contradicting four groups of two species each gives an intuitive idea about the logically expected ranking of groups according to support values. CD and FG are supported by one and two characters, respectively, without conflict. IJ and LM are supported, but also contradicted by one or more characters; there are no character interactions among these four groups. Bremer support and jackknifing frequencies differences produce different rankings of groups: While Bremer support does not discriminate among groups FG and LM, jackknifing detects the high proportion of cases in which group LM is contradicted, thus assigning a lower support value, even lower than that of groups CD and IJ. The jackknifing value depends both on the absolute support (compare CD and FG), and the ratio between favorable and contradictory characters (compare FG and LM). In this example, jackknifing gives a more sensible ranking of groups according to the relation between supporting and conflicting characters, while the Bremer support only gives an indication about the absolute support. The extended practice of informing the Bremer support and a resampling measure (the bootstrapping is most commonly used) seems well advised, because they express different aspects of group support.

It is possible that a better comprehension of the issue of support will lead to more accurate measures related to the Bremer values and its variants (e.g., the relative Bremer support of Goloboff and Farris, 2001). The

“redefinition” of support given by GK (p. 383) as “the degree to which critical evidence refutes competing hypotheses” it is however, not very helpful without further elaboration.³ Since the number of possible trees for a given dataset is fixed, supporting some of the trees can only be made at the expense of opposing other trees. GK (p. 383) further clarified: “A hypothesis is unsupported if it is either (1) decisively refuted by the critical evidence or (2) contradicted by other, equally optimal hypotheses (i.e., evidence is ambiguous, such as when multiple most-parsimonious cladograms obtain); otherwise it is supported.” This is of no help for the discussion of group support indices, which are intended to measure the degree of support of the *supported* groups.

Ranking of groups under variable resampling strengths

Resampling methods may indeed produce spurious results; GEA have introduced several refinements to protect against methodological artifacts, but problems still remain, especially in the weakly supported groups. Those measures are utilized, not because they are perfect, but because other known measures (fundamentally, Bremer support and its variants) also have limitations, and the systematists prefer to report some approximate measures of support than no measure at all. The most evident way to detect artifacts with support measures is by comparison with the strict consensus tree, looking for spurious groups with positive values, and supported groups with negative or zero

³GK introduced this definition in a section discussing *group* support measures. However, their examples and some passages (p. 383, mentioning the consistency index) suggest that they may be referring to global *tree* support at the same time.

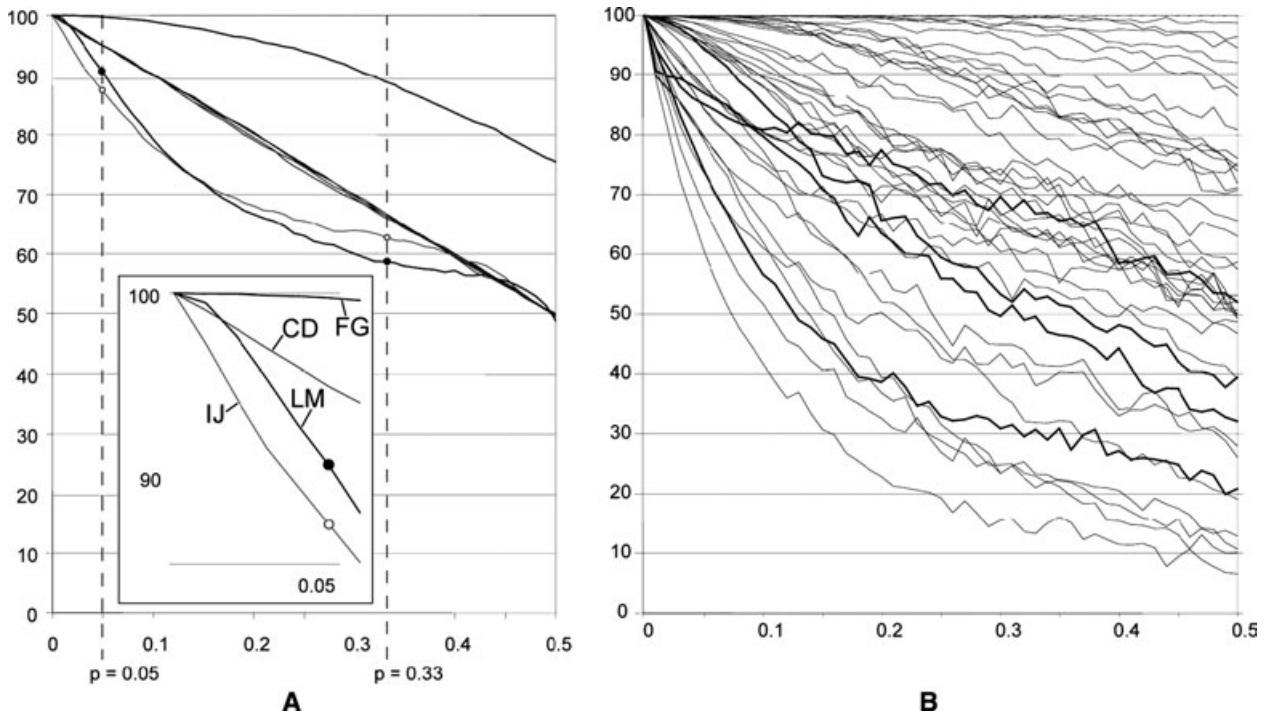


Fig. 2. Profiles of jackknifing frequencies under symmetric resampling, as a function of different resampling strengths ($0 \leq P \leq 0.5$). (A) Dataset of Fig. 1. Groups IJ and LM rank differently at different resampling strengths. The slope for group LM changes drastically from $P = 0.02$. (B) Dataset of araneomorph spiders (Ramírez, 2000). Several groups invert ranking order as P changes. The four thicker profiles indicate that the slopes can change drastically in the vicinity of $P = 0$. Frequencies calculated from 10 000 replicates, except in (B), for $0.13 \leq P \leq 0.50$, 1000 replicates, all calculated with TNT (Goloboff et al., 2004).

values. These cases are easily detected, but the same effect may subtly distort the support values without producing visible artifacts. It is then useful to explore additional criteria to evaluate the performance of support measures.

In addition to the comparison with the strict consensus, the trajectories of group frequencies against resampling strength are also informative for detecting problems with the resampling methods. For example, if group A is better supported than group B under $P = 0.33$, but the situation is reversed under $P = 0.20$, then we have a problem: either the method produces inconsistent results, or one of the two resampling strengths is inappropriate. Because there is not a “natural” resampling strength for the evaluation of group support, we should expect that the ranking of groups are consistent in some range of resampling strengths, at least between 0 and P . If the frequency is taken as an indication of support, then the trajectories should not cross each other in the range of reasonable probabilities of up- or down- weighting (e.g., $0 = P = 0.33$). When two trajectories intersect, it means that the ranking of the two groups are inverted before and after the point of intersection (Fig. 2), which means that the method produces different rankings at different resampling strengths. For example, the groups IJ and LM of

Fig. 1 rank differently under $P = 0.33$ and $P = 0.05$ (Fig. 2A). A close examination of the profiles for lower values of P indicates that the use of frequency slopes may have its caveats as well. Some group profiles are irregular, and the slope in the vicinity of $P = 0$ is quite different than the slope at slightly larger values of P (Fig. 2B).

Applications

GK emphasized against certain uses of support measures, in what they label as a “verificationist agenda” (see also Kluge, 1997). They associate resampling methods with this general intention (p. 411):

“According to our analysis, the most common interpretations of the results of data exploration in phylogenetic systematics are mistaken. Results of data exploration are primarily used to highlight strongly supported hypotheses as more accurate, reliable, or probably true and in effect protect those hypotheses from refutation by indicating that they are beyond additional testing.”

“Instead of drawing attention to strongly supported clades, we suggest that methods of data exploration be used to further the goals of science by highlighting weakly supported hypotheses by indicating cases in which choice among competing hypo-

theses is ambiguous or hypotheses have been less severely tested (tests have been less decisive), and therefore scientific inquiry aimed at them is likely to be more fruitful.”

The most important point here is that before *using* any support measures (regardless of the intended use or application), these measures need to be calculated first. GK manifested a special interest in the identification of weakly supported groups, but this of course presupposes the identification of the more strongly supported ones. The methods for obtaining support values are largely independent of their intended uses, and once these values are calculated, they can be used for a wide diversity of purposes, and assist in many, even opposite decisions. The alleged “verificationist” accusation of resampling methods is not related to their foundation, but to alleged uses or personal interests. It is important to consider the methodologies in the light of their intended uses (e.g., preferring a method more precise in the range of low support values, if the interest is detecting the most weakly supported groups; see GEA, pp. 328, 330), but their claim is so vague and general that it does not seem relevant for the argument. Their own recommendations for the application of support measures further illustrates some problems in their philosophical case:

“Pursuit of problems on the basis of degree of corroboration is defensible only in so far as the emphasis is placed on weakly corroborated hypotheses being more easily refuted and not on strongly corroborated hypotheses being more accurate or certain or less worthy of testing, as such a verificationist perspective would be contrary to the necessarily critical nature of science.” (p. 383)

“Our concept of support is heuristic in that it identifies cases in which refutation of competing hypotheses is weak, because weakly corroborated hypotheses may be more easily refuted than those that are more strongly corroborated. Rather than underscoring strongly supported clades by indicating them with asterisks and arrows, providing detailed discussions, and formalizing them with special taxonomic ranks, we believe that science would be better served by focusing on weakly supported clades and the potential means of more severely testing them (Kluge, 1997).” (pp. 383–384)

Similarly, Kluge (1997, p. 352) recommends:

“One approach is to focus the reanalysis on the characters diagnostic of those clades which are relatively weakly corroborated, as determined by unweighted branch support (Bremer, 1994 [...]).”

What they present as totally opposing conceptions and programs are actually the same: concentrate effort, time, and resources in testing the weakly supported groups, and leave the more strongly supported clades alone.

GK (p. 411) also criticized the use of support measures to assist taxonomic decisions:

“This misapplication of support is exemplified by common taxonomic practice, wherein strongly supported groups are

recognized formally, while weakly supported groups remain nameless and are thus hidden, often allowing paraphyletic groups to be retained. Such formal recognition effectively protects those so-called reliable groups from future refutation by fiat, i.e., by imposing legally the principle of stability, while the groups that are especially interesting scientifically are simply ignored. This practice is generally defended in the interest of “conservatism,” but we fail to see how this justifies overturning empirical evidence.”

According to GK’s recommendations, systematists should apply new names to weakly supported groups, and then select those groups as a priority for future research cycles. This contrast with the common practice of being cautious in giving new names to groups that one considers conflictive (and have good reasons to expect changes), especially if one or somebody else has started further research concentrated on that subject. At the time of deciding on taxonomic changes, the systematists use support measures as one element, among many others. For example, in the context of a revision of the spider subfamily Amaurobioidinae, I decided to maintain the genus *Sanogasta*, which is paraphyletic in terms of *Arachosia* (Ramírez, 2003). The relations are (*Sanogasta pehuenche* (*Arachosia* (other *Sanogasta*))); *Arachosia* is a very well supported group. The branches defining the paraphyly of *Sanogasta* are weakly supported, and changed resolutions very frequently during the construction of the dataset. There are many species of *Sanogasta* and *Arachosia* not included in that generic revision, and I saw little advantage in creating a new genus for *Sanogasta pehuenche*, based on what seems to be the ephemeral result of a poor taxon sampling. I rather decided to acknowledge that more work is needed to solve the issue satisfactorily. Contrasting with GK’s lecture, I perceive that leaving *Sanogasta* paraphyletic calls more strongly for attention as an unsolved problem, while creating a new generic name for *Sanogasta pehuenche* would give the impression that the issue is settled.

Publication

GK objected to the undue pressure of some scientific journals in demanding the inclusion of a wide range of analyses (p. 380):

“[M]ethods of data exploration have achieved greater prominence—so much so that empirical phylogenetic studies are judged less than “cutting edge” when data exploration is absent or insufficient and are even denied publication unless they meet the criterion of being unusually thorough explorations of the data.”

For the completeness of the argument, it should be mentioned that such editorial boards also press for the *exclusion* of their less favored methodologies. In my opinion, it would be sad if contributions employing

resampling measures of group support are questioned by journal boards with the use of obscure accusations such as “frequentist”, “verificationist”, “inductive”, or “unscientific”, based on the arguments of GK. Their critique points to idiosyncratic interpretations of resampling measures that are not necessarily nor universally associated with the methods, while not addressing their foundations or implementations. Their concern in the detection of weakly supported groups presupposes the detection of better supported ones, and their own recommendation for the use of support values is equivalent to the very same applications they criticize on philosophical grounds. They have not construed a convincing interpretation of resampling measures and their applications in the light of the particular brand of normative philosophy that they endorse.

Acknowledgments

To Roberto Keller, Gustavo Hormiga, Pablo Goloboff, Lone Aagesen, and two anonymous reviewers for comments and ideas on the manuscript. This research was supported by CONICET (PEI 6558).

References

- 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.
- Farris, J.S., 2002. RASA attributes highly significant structure to randomized data. *Cladistics*, 18, 334–353.
- Farris, J.S., Albert, V.A., Källersjö, M., Lipscomb, D., Kluge, A.G., 1996. Parsimony jackknifing outperforms neighbor-joining. *Cladistics*, 12, 99–124.
- Goloboff, P.A., Farris, J.S., 2001. Methods for quick consensus estimation. *Cladistics*, 17, S26–S34.
- Goloboff, P.A., Farris, J.S., Källersjö, M., Oxelman, B., Ramírez, M.J., Szumik, C.A., 2003. Improvements to resampling measures of group support. *Cladistics*, 19, 324–332.
- Goloboff, P.A., Farris, J.S., Nixon, K., 2004. TNT: Tree analysis using new technology. Program and documentation, available from the authors, and at <http://www.zmuc.dk/public/phylogeny/>.
- Grant, T., Kluge, A.G., 2003. Data exploration in phylogenetic inference: scientific, heuristic, or neither. *Cladistics*, 19, 379–418.
- Harshman, J., 1994. The effect of irrelevant characters on bootstrap values. *Syst. Biol.* 43, 419–424.
- Horowitz, I., 1999. A report on “One Day Symposium on Numerical Cladistics”. *Cladistics*, 15, 177–182.
- Kluge, A.G., 1997. Sophisticated falsification and research cycles: consequences for differential character weighting in phylogenetic analysis. *Zool. Scripta*, 26, 349–360.
- Kluge, A., 2002. Distinguishing “or” from “and” and the case for historical identification. *Cladistics*, 18, 585–593.
- Miller, J.A., 2003. Assessing progress in systematics with continuous jackknife function analysis. *Syst. Biol.* 52, 55–65.
- Ramírez, M.J., 2000. Respiratory system morphology and the phylogeny of haplogyne spiders (Araneae, Araneomorphae). *J. Arachnol.* 28, 149–157.
- Ramírez, M.J., 2003. A cladistic generic revision of the spider subfamily Amaurobioidinae (Araneae, Anyphaenidae). *Bull. Am. Mus. Nat. Hist.* 277, 1–262.
- Siddall, M.E., Whiting, M.F., 1999. Long-branch abstractions. *Cladistics*, 15, 9–24.
- Swofford, D., 1998. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4. Sinauer Associates, Sunderland, MA.