WILEY-BLACKWELL

Cladistics 24 (2008) 1051–1064

Cladistics

10.1111/j.1096-0031.2008.00231.x

Clade support measures and their adequacy

Taran Grant[a,*] and Arnold G. Kluge[b]

[a]*Faculdade de Biociências, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, RS, Av. Ipiranga 6681, 90619-900, Brazil;* [b]*3140 Dolph Drive, Ann Arbor, MI 48103, USA*

Accepted 15 May 2008

Abstract

In addition to hypothesis optimality, the evaluation of clade (group, edge, split, node) support is an important aspect of phylogenetic analysis. Here we clarify the logical relationship between support and optimality and formulate adequacy conditions for support measures. Support, S, and optimality, O, are both empirical knowledge claims about the strength of hypotheses, h_1, h_2, … h_n, in relation to evidence, e, given background knowledge, b. Whereas optimality refers to the *absolute* strength of hypotheses, support refers to the *relative* strength of hypotheses. Consequently, support and optimality are logically related such that they vary in direct proportion to each other, $S(h \mid e,b) \propto O(h \mid e,b)$. Furthermore, in order for a support measure to be objective it must quantify support as a function of explanatory power. For example, Goodman–Bremer support and ratio of explanatory power (REP) support satisfy the adequacy requirement $S(h \mid e,b) \propto O(h \mid e,b)$ and calculate support as a function of explanatory power. As such, these are adequate measures of objective support. The equivalent measures for statistical optimality criteria are the likelihood ratio (or log-likelihood difference) and likelihood difference support measures for maximum likelihood and the posterior probability ratio and posterior probability difference support measures for Bayesian inference. These statistical support measures satisfy the adequacy requirement $S(h \mid e,b) \propto O(h \mid e,b)$ and to that extent are internally consistent; however, they do not quantify support as a function of explanatory power and therefore are not measures of objective support. Neither the relative fit difference (RFD; relative GB support) nor any of the parsimony (bootstrap and jackknife character resampling) or statistical [bootstrap character resampling, Markov chain Monte Carlo (MCMC) clade frequencies] support measures based on clade frequencies satisfy the adequacy condition $S(h \mid e,b) \propto O(h \mid e,b)$ or calculate support as a function of explanatory power. As such, they are not adequate support measures.

The development of quantitative phylogenetics continues to be driven by the need for explicit concepts and assumptions in methods. Debate has appropriately focused on the optimality criteria employed in choosing among competing hypotheses. The evaluation of clade (group, edge, split, node) support is also an important component of phylogenetic analysis, yet comparatively little attention has been paid to the conceptual bases for the many methods used to measure support. Indeed, the concept of support has been a frequent topic in the philosophical and statistical literature, but there has been limited discussion about what is meant by "support" in phylogenetics (Grant and Kluge, 2003; Wilkinson et al., 2003) or the conditions that must be met for a given support measure to be considered adequate.

We recognize that the concept of support is a matter of definition. Consequently, we use widely accepted logical and epistemological arguments to clarify the relationship between support, optimality, and objectivity, which limits the scope that those definitions can reasonably take within the context of objective knowledge. Based on this relationship, we formulate adequacy conditions for support measures and evaluate the adequacy of several popular support measures. We also propose equivalent support measures for parsimony, maximum likelihood, and Bayesian phylogenetic inference. We classify measures as either parsimony support measures or statistical support measures in reference to the optimality criteria employed. As employed here, the hypothesis that explains

*Corresponding author:
E-mail address: taran.grant@pucrs.br

the evidence with the fewest transformation events (steps) is optimal under parsimony, whereas the hypothesis of greatest posterior probability (Bayesian inference) or maximum likelihood is optimal according to statistical criteria. Numerous efforts have been made to understand parsimony as a method of statistical estimation (i.e. by identifying the conditions under which the most parsimonious solution is also the best statistical estimate), specifically a maximum likelihood method (e.g. Farris, 1973; Goldman, 1990; Tuffley and Steel, 1997; Goloboff, 2003), but advocates of parsimony generally deny the applicability of statistical methods because of the detailed information about evolutionary processes that is required by the statistical models (Farris, 1983), the necessary uniqueness of the historical events phylogenetic methods aim to discover (Siddall and Kluge, 1997; Kluge, 2002), and the superfluity of the assumptions of statistical models (Kluge and Grant, 2006).

Support, optimality, and other concepts

In the most general terms, support, S, is an empirical knowledge claim about the strength of competing hypotheses, h_1, h_2, . . . h_n, in relation to a body of evidence, e, given specific background knowledge, b (which constitutes the minimal assumptions required of a scientific inference). Together, e, h, and b are the necessary and sufficient parameters for inferring phylogeny (Kluge, 1997). Symbolically, this may be expressed as

$$S(h \mid e, b).$$

Optimality, O, is also an empirical knowledge claim about the strength of competing hypotheses in relation to a body evidence, given specific background knowledge, expressed as

$$O(h \mid e, b).$$

Optimality is assessed by applying an optimality criterion, a rule with an epistemic justification, for identifying and selecting the strongest hypotheses. Popular optimality criteria in phylogenetic inference include parsimony, maximum likelihood, and posterior probability, and an extensive literature exists on the theoretical bases and relative merits of these criteria.

Given that both support and optimality are empirical knowledge claims about the strength of hypotheses in relation to evidence given background knowledge, support and optimality are logically related such that they vary in direct proportion to each other, or

$$S(h \mid e, b) \propto O(h \mid e, b).$$

The terms "support" and "optimality" are often used interchangeably because of the close logical relationship between the two concepts. For example, in maximum likelihood inference the maximum likelihood hypothesis is both the optimal hypothesis and the most supported hypothesis (see below), and the clades present in the optimal cladogram are also supported clades. Similarly, reports on optimality, such as tree length, extra-steps, branch lengths, and consistency and retention indices, are often discussed in the context of support (Grant and Kluge, 2003; Egan, 2006). However, support and optimality are not synonymous. Optimality is concerned with the *absolute* strength of hypotheses, whereas support is concerned with the *relative* strength of hypotheses. That is, optimality criteria identify the strongest hypothesis (e.g. the most parsimonious tree), whereas support measures evaluate the strength of some hypothesis (usually the optimal hypothesis) relative to one or more competing hypotheses (e.g. the frequency of cladograms that contain specified clades in a parsimony jackknife sample). Furthermore, following the conceptualizations and terminology developed by Grant and Kluge (2003, 2005), optimality is scientific, whereas support is heuristic.[1] As such, although optimality and support are logically related, optimality is epistemologically prior to support. Finally, in phylogenetic analysis support is usually (but not always, e.g., the total support of Källersjö et al., 1992) concerned with individual clades rather than whole topologies.

The directly proportional relationship between optimality and support may be illustrated with an example under parsimony. Figure 1 (modified from Ramírez, 2005) shows the optimal tree of 35 steps and the tree-lengths (steps) of the relevant competing hypotheses for the terminals H, I, J and K, L, M. These tree-lengths provide a basis to rank hypotheses according to their absolute strength, i.e. their optimality:

$$\begin{aligned} &O(h_{(\ldots(\mathrm{H(IJ)}))} \mid e, b) = O(h_{(\ldots(\mathrm{K(LM)}))} \mid e, b) > \\ &> O(h_{(\ldots(\mathrm{I(HJ)}))} \mid e, b) > O(h_{(\ldots(\mathrm{L(KM)}))} \mid e, b) = \\ &= O(h_{(\ldots(\mathrm{M(KL)}))} \mid e, b). \end{aligned}$$

[1] This use of "heuristic" is in the sense of "allowing or assisting to discover" (*Oxford English Dictionary*), as applied in standard English and philosophy of science (e.g. Lakatos, 1978). Related but specialized meanings of "heuristic" occur in many fields, including law, engineering, psychology, and computer science. It has been pointed out by W. Wheeler (pers. commun.) that our usage may lead to confusion given the common application in phylogenetics of "heuristic" in the specialized sense of computer science, i.e. in reference to any method that aims to solve a problem, often through trial and error, but ignores whether the solution can be proven correct, e.g. a heuristic algorithm. Ideally, we would replace one of these uses with a different word to avoid confusion. However, we have failed to discover a replacement that retains the intended meaning and connection to the broader literature, and we therefore rely on context for clarification.

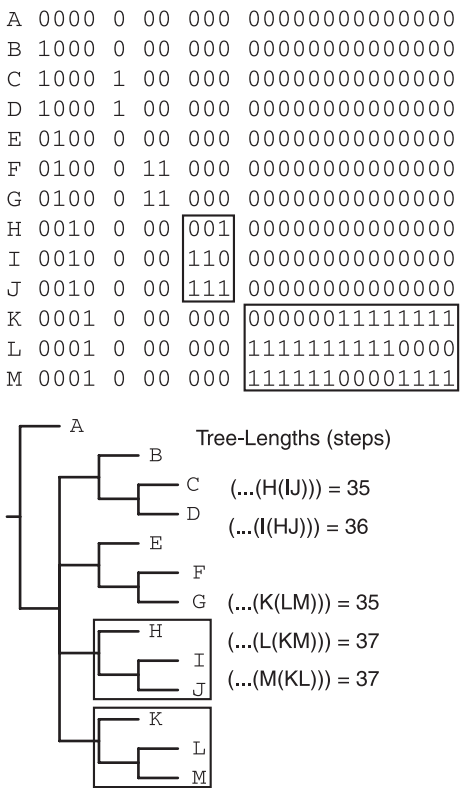

Fig. 1. Contrived example showing the optimal tree of 35 steps and tree-lengths for relevant alternative trees (in parenthetic notation). Groups discussed in the text are enclosed in boxes. Modified from Ramírez (2005).

Clades IJ and LM are both present in the optimal tree. Clade HJ is present in a tree one step longer and therefore less optimal than the tree with IJ and LM, but that tree is one step shorter than the optimal trees for KM and KL, which are equally suboptimal. As such, the rank order of the optimal trees that contradict these clades present in the most parsimonious tree is HJ > KM = KL. Although the trees that contain clades IJ and LM are equally optimal and therefore have the same *absolute* strength, the extent of the suboptimality of the contradictory hypotheses provides a basis for ranking IJ and LM according to their *relative* strength, i.e. their support. Because the optimal trees that contradict LM are more suboptimal than the optimal tree that contradicts IJ, it follows that the relative strength of LM is greater than the relative strength of IJ, LM > IJ. As such, support and optimality vary in direct proportion to each other; the greater the optimality of a given tree, the greater the support for its constituent clades.

The accuracy of a hypothesis relates only to its correspondence to truth regardless of the existence or amount of evidence for or against it. At least logically, support (and optimality) and accuracy are unrelated. Reliability, stability, and robustness ultimately relate to the concept of utility, which is also logically unrelated to support (and optimality), as noted by Hacking (1965).

The frequentist concepts of confidence interval and significance level also differ from support, as does the Bayesian concept of credibility interval. All of these concepts aim to prevent rejection of one hypothesis in favour of another unless the rival is "much better supported" (Hacking, 1965, p. 89), as assessed by estimating the probability of obtaining the observed magnitude of support by chance, according to some assumed distribution. For example, in maximum likelihood inference the maximum likelihood hypothesis has the greatest support, yet a likelihood ratio test may find that the degree of support is not significant at the 0.05 level because the test statistic (e.g. $L_1 - L_2$, where L is the log-likelihood estimate of each of two hypotheses) falls outside the 95% confidence interval.

Objectivity

Any concept or measure of support that aims to be objective must relate to a theory of objectivity. According to Popper's (1979, 1983) theory of knowledge, knowledge claims are either objective or subjective (see summary of opposing characteristics in Table 1). As such, this theory is subject to the logical law of the excluded middle, asserting either *p* or not-*p* and logically excluding middle cases, such as knowledge claims being partly objective and partly subjective. This does not deny the problem of cognitive filtering or the distinction between phenomena (things as they are perceived) and noumena (things in themselves), which have preoccupied philosophers for centuries; objective knowledge as knowledge that is demonstrably true, even by degrees, is unattainable (Watkins, 1997).

Instead, knowledge claims are objective if and only if they are open to and withstand rational criticism (Popper, 1979).[2] Objective empirical knowledge is sought with deductive logic and is controlled actively by test—rational criticism involving observation and experiment. It is concerned with the relationship between hypotheses and evidence given background

[2]Popper's theory of objective knowledge is not to be confused with objectivity of observation in science, where *repeatability* of observation across independent investigators is sought, or with the variability in perception among scientists, such as due to cognitive and cultural biases (see Kearney, 2007). Typically, the repeatability of measurements and counts used as evidence in systematics is sought in instrumentation, precisely defined rules, and other technicalities. Whereas objectivity of observation and theory-free evidence were central to numerical taxonomy, phylogenetics (e.g. *sensu* Farris, 1983, p. 1) emphasizes testability and causal explanation founded in evolutionary epistemology (contra Kearney, 2007).

Table 1
Some distinguishing characteristics of two kinds of empirical knowledge claims (Popper, 1979, 1983)

Objective	Subjective
Consists of causal explanations that may be true but whose truth is unprovable and therefore is scientifically irrelevant, i.e. they are eternally conjectural	Consists of either acausal descriptions (instrumentalism) or estimations whose truth is inductively provable
Methods of explanation are demonstrative; they consist of a logical deduction, one whose conclusion is the *explanandum*—the statement of the thing to be explained—and whose premises constitute the *explanans*—a statement of the explaining laws and initial conditions	Methods of estimation are non-demonstrative; they consist of probabilistic assessments of instances of the same kind
Hypotheses are preferred on the basis of their explanatory power	Hypotheses are preferred on bases other than explanatory power, including theory compatibility, probability, predictivity, stability, reliability, descriptive efficiency, and elegance of hypotheses and/or methods
Methods are constrained by the ontological and epistemological considerations required to achieve explanatory power	Methods are unconstrained by ontological and epistemological considerations and need only be believed to be relevant
Results can only be changed by evidence, i.e. by empirical refutation; personal belief is irrelevant	Results can be changed by intersubjectivity (i.e. consensus of personal beliefs)
Empirical evidence must be a severe test of a hypothesis in order to corroborate that hypothesis	Empirical evidence need only be an instantiation (i.e. a perceptually similar instance) of a hypothesis in order to verify that hypothesis
The results of previous tests are irrelevant	The results of previous tests are relevant
Does not involve expectations	Involves expectations, to which belief is closely connected
Exactness, or precision, is valued insofar as it facilitates testing, i.e. a precise claim may be more easily refuted than a vague claim	Exactness, or precision, is valued for its relationship to accuracy

knowledge, not with belief or accuracy. It is based on the power of causal theories to explain the critical evidence (those observations that have the potential, through causal entailment, to refute a particular theory), i.e. their explanatory power. In the absence of the logical restrictions required of objective knowledge claims, there is nothing to prevent the pervasion of idiosyncratic definitions of support.

Empirical knowledge claims, including claims of support, that are not derived from explanatory power (or logically equivalent functions; see Popper, 1983) are subjective. In practice, subjective knowledge often derives from a passive, progressive consolidation of confirming instances that leads to varying degrees of belief in a hypothesis. As such, subjective knowledge is often expressed probabilistically in terms of certainty and confidence, and nowhere does subjectivism have more influence than in the application of the probability calculus. In addition, exactness, or precision, is valued, if only as demanded by that calculus. Ultimately, subjective knowledge is dependent on one's belief in the accuracy of a hypothesis rather than the logic of deduction and the relationship between hypotheses and evidence given background knowledge.

Given the minimal evolutionary assumptions of descent with modification, explanatory power is operationalized in phylogenetics by summing the patristic distance (character-state transformations, steps, tree-length) of a given phylogenetic hypothesis in explaining the observed character variation (Kluge and Grant, 2006). The fewer the character-state transformations a phylogenetic hypothesis postulates in explaining the total (unpartitioned, combined) evidence, the greater its explanatory power and, in turn, its degree of optimality. Methods that incorporate assumptions about the probability of specific classes of transformations in the process of evolution (e.g. maximum likelihood and Bayesian methods) or the relative merits of classes of characters or transformations (e.g. weighted-parsimony methods) do so at the expense of explanatory power because they (1) include extra assumptions, in addition to the background knowledge of descent with modification (what is necessary and sufficient), and (2) postulate superfluous hypotheses of character transformations.

The above explication of explanatory power in phylogenetic inference differs from that of Farris et al. (1995, p. 218; see also Farris, 1983), who explicated explanatory power in terms of minimizing "the number of independent *ad hoc* hypotheses of homoplasy" instead of "minimizing the number of transformation events required to explain the character-states of the terminal taxa as hypotheses of homology, where the concept of homology is restricted to just those inherited 'things' shared by species" (Kluge and Grant, 2006,

p. 276; see also Grant and Kluge, 2004; Kluge, 2007). Whereas the former is dependent on the perceived similarity of objects (i.e. similar objects that evolved independently), the latter stresses the causality of evolutionary events.

Grant and Kluge (2003, p. 383) proposed an objective concept of support in phylogenetics, which they defined as "the degree to which critical evidence refutes competing hypotheses". Although this concept of support is objective and the interpretation was clarified by subsequent text, the choice of wording is unfortunate because a hypothesis that is refuted to a greater degree has less explanatory power and less support than a hypothesis that is refuted to a lesser degree. We therefore rephrase our definition of objective support in logically equivalent (Popper, 1983) but clearer terms as *the relative explanatory power of competing hypotheses* (see also Grant and Kluge, 2007).

The subjectivity of knowledge claims is rarely acknowledged. Bayesian inference aims to quantify personal belief in hypotheses and is therefore openly subjective, but the knowledge claims that result from operationalist or instrumentalist approaches that eschew ontology and assert that their operations stand on their own (e.g. Giribet and Wheeler, 2007) are also subjective because there is no basis for hypothesis preference beyond one's personal belief in the supremacy of the particular operations. Although predictive (or retrodictive) accuracy and robustness or stability have been cited as the basis for such belief (Wheeler and Blackwell, 1984; Goloboff, 1993; Siddall, 1995, 2002), they confuse repeatability with objectivity in the same way that description may be confused with explanation (Popper, 1957, p. 124; Hull, 1974, p. 97; Farris, 1979, pp. 512–514) and correlation may be confused with causation (e.g. Miller, 2003, p. 63). Application of confidence intervals and significance levels is subjective as well, due in part to the arbitrary selection of "acceptable" risk of type I or type II error.

Adequacy conditions

A variety of criteria have been operationalized as support measures. To evaluate competing criteria it is standard practice to formulate and apply a set of adequacy conditions. Such adequacy conditions provide a logical justification for the criteria deemed adequate, although they do not justify the resulting empirical knowledge claims as true (as was claimed by the discredited philosophy of justificationism; Notturno, 2003).

We propose two adequacy conditions, both of which must be met in order for a method to provide a measure of objective support. The first is a general requirement that the measure satisfy the relation $S(h \mid e,b) \propto O(h \mid e,b)$, i.e. regardless of the criterion employed in the evaluation of hypothesis optimality (e.g. parsimony, maximum likelihood, or Bayesian criteria) or support, support and optimality must vary in direct proportion to each other. This adequacy condition ensures internal logical consistency and prevents paradoxical situations in which suboptimal hypotheses are considered to have greater support than optimal hypotheses (although this latter requirement would be met by any strictly increasing monotonic function, not only direct proportionality; W. Wheeler, pers. commun.). The second adequacy condition required for support measures to be objective is that they quantify support in terms of explanatory power.

Parsimony support measures

For a given clade present in a most parsimonious tree, Goodman–Bremer (GB) support is defined as

$$S' - S$$

where S denotes the length of the most parsimonious tree(s) and S' is the length of the most parsimonious tree(s) that lacks that clade (Goodman et al., 1982; Bremer, 1988, 1994; Källersjö et al., 1992; see also Grant and Kluge, 2008). In other words, for each clade in the most parsimonious cladogram GB support measures the difference in the patristic distance of that cladogram and the optimal cladogram that lacks each of those clades. This allows all clades in the optimal tree(s) to be ranked according to their relative strength (Fig. 2). GB support is proportional to hypothesis optimality, such that the adequacy condition $S(h \mid e,b) \propto O(h \mid e,b)$ is satisfied.

As noted above, explanatory power is operationalized in phylogenetics as the summed patristic distance, or tree-length, of a given phylogenetic hypothesis in

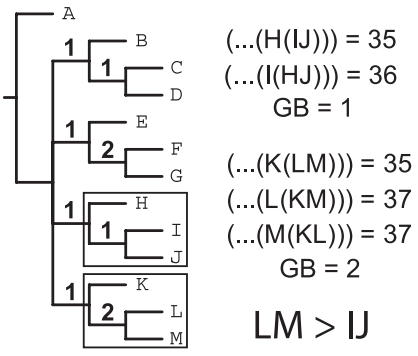

Fig. 2. Most parsimonious tree and Goodman–Bremer (GB) support values for the contrived example in Fig. 1. The relative strength and GB are greater for LM than for IJ. Groups discussed in the text are enclosed in boxes. Modified from Ramírez (2005).

explaining the observed character variation in light of the background knowledge of descent with modification (Kluge and Grant, 2006). The fewer the character-state transformations (steps) a phylogenetic hypothesis postulates in explaining the total (unpartitioned, combined) evidence, the greater its explanatory power. Given this direct relationship between steps and explanatory power, GB support satisfies the second adequacy condition of calculating support as a function of explanatory power, making it an adequate measure of objective support.

Rather than evaluating support as the difference in patristic distance, as done by GB support, the REP support index (ratio of explanatory power; Grant and Kluge, 2007) calculates support as the ratio of the patristic distances of the optimal tree and the optimal tree that lacks a given clade, which reduces to

$$\mathrm{REP} = (S' - S)/(G - S)$$

where G denotes the maximum number of steps required to explain aligned character-states [Farris, 1989; for the more general case in which alignment is not assumed prior to phylogenetic analysis replace G with X, the number of steps required to explain each unaligned character-state (e.g. each unaligned nucleotide) as uniquely evolved]. The numerator of this expression is GB support, so the REP support for a group is equal to its GB support divided by the difference in length between the least parsimonious tree (G or X) and the most parsimonious tree (S). When $S' = S$, the REP support value is 0 (as is the GB support); when $S' = G$ (or X), the REP support value is 1.

Like GB, REP quantifies support as a function of the relative explanatory power of the competing hypotheses, and the ranking of clades for a given dataset is identical for both measures, such that $S(h \mid e,b) \propto O(h \mid e,b)$ is also satisfied. However, by standardizing GB support relative to the best and worst possible explanation for each dataset, REP support also allows meaningful comparison of support across datasets (Grant and Kluge, 2007). As shown in Fig. 3, the effect of multiplication of identically distributed data is different for these two measures. A property of GB (and other methods, including those that calculate clade support from jackknife and bootstrap resampling or MCMC sampling; Wheeler and Pickett, 2008) is that multiplied datasets are assigned higher levels of support, whereas REP support remains constant. Neither of these effects is intrinsically good or bad, and the different behaviour of GB and REP provides different analytical insights.

Goloboff and Farris (2001, p. S30) stated that a "defect of [GB support] is that it does not always take into account the relative amounts of evidence contradictory and favorable to the group." They proposed that "[t]his problem is diminished if support is calculated as the ratio between the amounts of favorable and contradictory evidence", which led them to define the relative fit difference (RFD) index:

$$\mathrm{RFD} = (F - C)/F.$$

Given two trees, F is the sum of the differences in steps (fit differences) of the characters that favour the more parsimonious topology (i.e. the steps of the characters that fit the more parsimonious tree better than they fit the less parsimonious tree), and C is the sum of the differences in steps of the characters that contradict the most parsimonious topology (i.e. the steps of the characters that fit the less parsimonious tree better than they fit the most parsimonious tree). When a group has no favourable evidence or is contradicted by as much evidence as favours it, RFD = 0; when a group is favoured and uncontradicted, RFD = 1. Goloboff and Farris (2001) noted that $F - C$ is equal to GB (i.e. $S' - S$), and Goloboff et al. (2003, p. 326) referred to the RFD based on comparison to the optimal hypothesis as "relative Bremer support". Goloboff and Farris (2001; see also Goloboff et al., 2003; Ramírez, 2005) did not explicate why support should be calculated on the basis of the ratio of F and C instead of explanatory power or why GB support is defective.

GB support includes all evidence and, therefore, necessarily takes into account the evidence contradictory and favourable to a group. However, it does so in a way that satisfies both adequacy conditions identified above, whereas the RFD does not. As shown in Fig. 4, RFD group ranking can violate the condition $S(h \mid e,b) \propto O(h \mid e,b)$. Furthermore, as shown in Fig. 5, the RFD does not quantify support as a function of explanatory power. The RFD assigns an uncontradicted group corroborated by a single synapomorphy and an uncon-

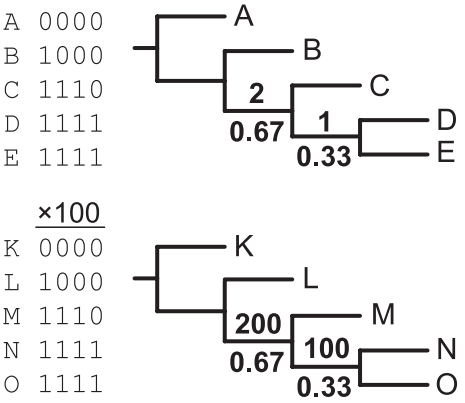

Fig. 3. Contrived datasets comparing Goodman–Bremer (GB) support and REP support. The upper dataset consists of four characters for five terminals. The lower dataset consists of the same characters repeated 100 times. GB values are 100 times greater for the lower dataset, but the REP support values are identical. GB values above branches. REP support values below branches.

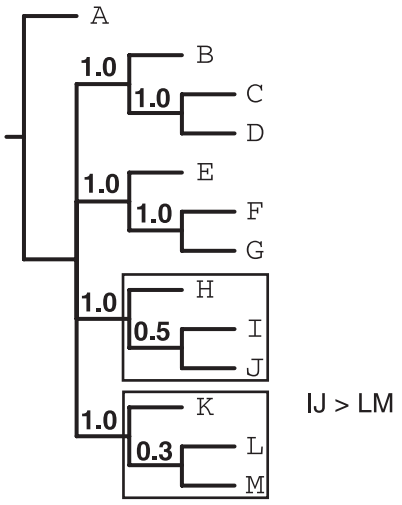

Fig. 4. Most parsimonious tree and relative fit difference (RFD) values for the contrived example in Fig. 1. Clades discussed in the text are enclosed in boxes. Modified from Ramírez (2005).

tradicted group corroborated by 100 synapomorphies the same support value, scoring both as maximally supported, even though the relative strength of the latter hypothesis is clearly greater. The problematic behaviour of the RFD is further illustrated by adding a single contradictory synapomorphy to each case. Whereas GB support values indicate the relative strength of the competing hypotheses and rank clades accordingly, the RFD jumps from scoring both clades equally to ranking the two clades at opposite extremes.

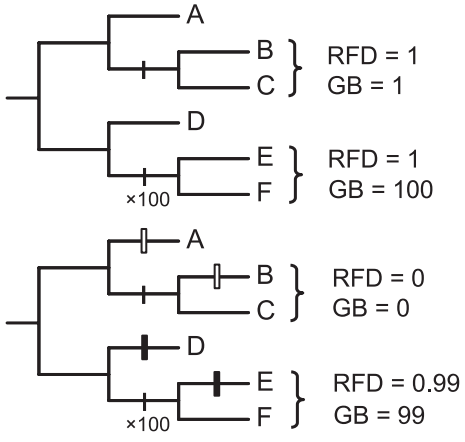

Fig. 5. Illustration of the difference between Goodman–Bremer (GB) support and the relative fit difference (RFD). In the upper example, clades AB and EF have equal (the maximum possible) RFD values, even though EF is delimited by 100 times as many synapomorphies as AB. GB correctly indicates the greater explanatory power of EF. Adding a single contradictory synapomorphy causes RFD to jump from ranking both groups as maximally supported to ranking them at opposite extremes. GB does not exhibit this extreme behaviour.

Bootstrap (Felsenstein, 1985) and jackknife (Farris et al., 1996) character resampling methods are commonly employed as parsimony support measures. Elsewhere we have criticized the statistical interpretation of resampling methods (Grant and Kluge, 2003). Goloboff et al. (2003) and Ramírez (2005) also rejected a statistical interpretation and instead argued that resampling frequencies quantify support as the relative amount of favourable and contradictory evidence for each group present in the optimal topology. Other methods have also been proposed to assess this relationship, such as spectral analysis (Hendy and Penny, 1993), and Lento et al. (1995, p. 41) claimed that the only difference between their PB values, defined as "the frequency of support minus the frequency of conflict (S − C)", and bootstrap values is that "PB values, derived from the HadTree spectrum, take not only support for an edge but also contradictory signal values for that edge into account." As such, the relationship between these measures and the parameter they aim to measure warrants inspection.

Ramírez (2005) demonstrated that clade rank order can be contradictory for jackknife resampling and GB support (Fig. 6). This example also demonstrates that jackknife resampling does not necessarily vary in direct proportion to optimality, so jackknife resampling does not satisfy the adequacy condition $S(h \mid e,b) \propto O(h \mid e,b)$. Furthermore, according to Goloboff et al. (2003) and Ramírez (2005), jackknife resampling quantifies support as a relation between partitions of evidence for and against a hypothesis, whereas explanatory power is assessed in reference to the evidence treated as a single partition, the total evidence. Jackknife resampling does not measure support as a function of explanatory power and therefore cannot be defended in terms of objective knowledge.

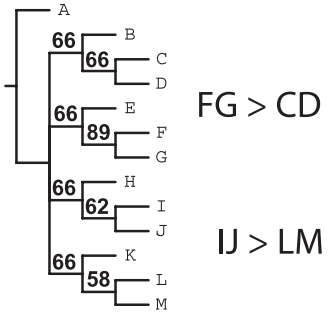

Fig. 6. Most parsimonious tree and jackknife support values for the contrived example in Fig. 1. Jackknife support is greater for IJ than for LM, even though the relative strength of LM is greater than that of IJ (see Fig. 2 and text). According to the RFD, CD and FG are both maximally supported, but according to jackknife resampling neither group receives maximum support and the support is less for CD than for FG (Fig. 4). Modified from Ramírez (2005).

Statistical support measures

Maximum likelihood support

Hacking (1965) related likelihood and support through the "law of likelihood" and proposed the likelihood ratio as a measure of support (p. 71, italics in original):

The law of likelihood: *If h and i are simple joint propositions included in the joint proposition e, then e supports h better than i if the likelihood ratio of h to i exceeds 1.*

Symbolically, the likelihood ratio support measure derived from the law of likelihood is

$$P(e \mid h_h, b)/P(e \mid h_i, b).$$

Accordingly, the maximum likelihood hypothesis has the greatest support, regardless of the significance of the degree of support (which is evaluated in relation to an assumed distribution, i.e. a likelihood ratio test).

The above is a general measure of support, but in practice one is usually interested in the support for the maximum likelihood hypothesis, h, versus the optimal contradictory hypothesis, h'. This may be expressed as

$$P(e \mid h, b)/P(e \mid h', b).$$

In the case of phylogenetic analysis, the degree of support for each clade in the maximum likelihood tree may be assessed by calculating the likelihood ratio of the maximum likelihood tree to the optimal tree(s) that lacks each clade in the maximum likelihood tree (the likelihood ratio support measure). In practice, calculations usually employ log-likelihoods, and the likelihood ratio support measure is therefore the anti-log of the difference in log-likelihoods. A simpler alternative is to report the log-likelihood of the likelihood ratio support measure (i.e. log-likelihood difference support measure), as done by Meireles et al. (1999; see also Lee and Hugall, 2003), who employed this measure to estimate the strength of grouping for clades in their maximum likelihood tree.

Instead of calculating support as the ratio of likelihoods (or the difference of log-likelihoods), support may be calculated for each clade in the maximum likelihood tree as the difference in likelihoods of the maximum likelihood tree and the optimal tree(s) that lacks each clade in the maximum likelihood tree (this support measure is equivalent to GB support in parsimony analysis):

$$P(e \mid h, b) - P(e \mid h', b).$$

In these measures hypothesis optimality and support are both calculated as a function of the likelihood score of competing hypotheses, so these measures satisfy the adequacy condition $S(h \mid e,b) \propto O(h \mid e,b)$.

Bootstrap character resampling (Felsenstein, 1985) was proposed to estimate confidence intervals but is also commonly interpreted as a maximum likelihood support measure. The use of bootstrap resampling in maximum likelihood to estimate confidence intervals has been questioned for a variety of reasons (e.g. Goldman, 1993; Holmes, 2003), but an additional criticism applies when bootstrap frequencies are interpreted as support values. As discussed above for parsimony support measures and shown in the example in Fig. 7 and Table 2, bootstrap frequencies may not vary in direct proportion to hypothesis optimality, thus violating the adequacy condition $S(h \mid e,b) \propto O(h \mid e,b)$.

Although the likelihood ratio and likelihood difference support measures satisfy the first adequacy condition $S(h \mid e,b) \propto O(h \mid e,b)$, they do not satisfy the second adequacy condition of quantifying support in terms of explanatory power. To begin with, maximum likelihood does not maximize explanatory power because it does not discern between critical evidence (severe tests) and mere data (Farris et al., 2001; Kluge, 2001; *contra* de Queiroz, 2004). Furthermore, phylogenetic applications of maximum likelihood require assumptions beyond those necessary to make a logically valid inference (Kluge and Grant, 2006); thus, even when maximum likelihood and parsimony agree on the optimal tree (for a review see Goloboff, 2003), the maximum likelihood solution has less explanatory power than the parsimony solution because it rests on additional assumptions about the evolutionary process. Additionally, many of those model assumptions concern details

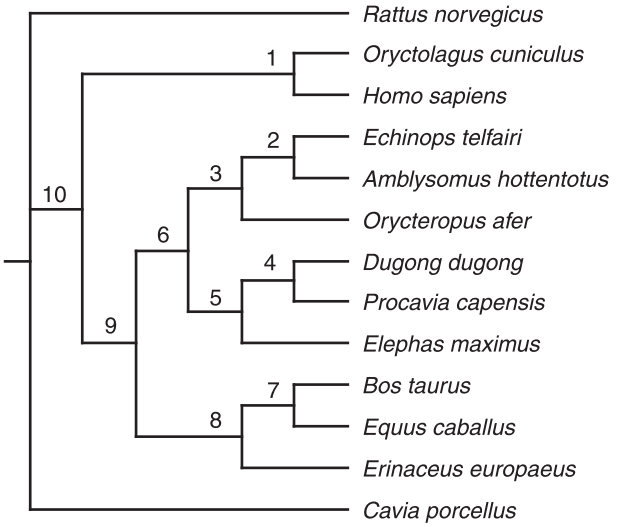

Fig. 7. Maximum likelihood tree (log-likelihood = −6068.56918) with numbered clades shown in Table 2. Data consist of 1134 aligned nucleotide characters of nuclear gene A2AB from 13 mammal species (Stanhope et al., 1998; Treebase matrix accession number M624, terminals with missing data removed). Analyses were performed using Garli 0.951 (Zwickl, 2006) under default model assumptions.

Table 2
Support measures for clades in the maximum likelihood tree in Fig. 7 (log-likelihood = −6068.56918) and log-likelihoods of optimal trees contradicting each of those clades. The rank order of clades according to their bootstrap frequencies is 6 = 10 > 5 > 1 > 4 > 7 > 3 = 8 > 2 > 9. The rank order of clades according to their log-likelihood difference support (LLD) values is 10 > 6 > 2 > 5 > 8 > 1 > 7 > 9 > 4 > 3. Analyses were performed using Garli 0.951 (Zwickl, 2006) under default model assumptions

Clade	Bootstrap frequency (×100)	Log-likelihood of contradictory tree	LLD support value
1	79	−6072.50919	3.94001
2	51	−6103.99962	35.43044
3	67	−6068.57111	0.00193
4	74	−6069.17414	0.60496
5	97	−6085.97068	17.4015
6	100	−6103.99969	35.43051
7	68	−6071.78813	3.21895
8	67	−6073.30530	4.73612
9	39	−6069.36604	0.79686
10	100	−6218.17912	149.60994

about the evolutionary process that are counterfactual, i.e. contradicted by empirical evidence (Siddall and Kluge, 1997). For these reasons, the support measures derived from maximum likelihood analysis in phylogenetic inference cannot be defended as quantifying objective support.

Bayesian support

A Bayesian measure for the degree of support (or confirmation) a piece of evidence provides for a given hypothesis remains an active area of research (Crupi et al., 2007). However, regardless of the measure used to calculate the evidential support for a hypothesis, a Bayesian measure to evaluate the relative degree of support of two competing hypotheses is simply the ratio of their posterior probabilities (although implementations may focus on the likelihood ratio; e.g. Hasegawa and Kishino, 1989), expressed symbolically as

$$P(h_1 \mid e,b)/P(h_2 \mid e,b).$$

The Bayesian support for the optimal hypothesis, h, relative to the optimal contradictory hypothesis, h', is then

$$P(h \mid e,b)/P(h' \mid e,b).$$

Alternatively, Bayesian support may be calculated as the difference in those posterior probabilities:

$$P(h \mid e,b) - P(h' \mid e,b).$$

Most current Bayesian implementations use an MCMC sampler to generate a sample of topologies (and other parameters) in relation to their prior probabilities and likelihoods (Huelsenbeck et al., 2002). Provided that several conditions are met (e.g. the correct acceptance probability was specified, sampling was sufficient to reach convergence; Tierney, 1994; Mossel and Vigoda, 2006), the frequency distribution of the sample parameter values approximates their posterior probability density. For example, the frequency with which a given topology is sampled is an estimate of its posterior probability, and the topology with the greatest frequency may be interpreted as the most probable point estimate of phylogeny, i.e. the maximum posterior probability topology (MAP; Rannala and Yang, 1996; Yang and Rannala, 1997).

A simple measure of the degree of Bayesian support for a given clade in the MAP may therefore be calculated as either the ratio or the difference of the posterior probabilities of the MAP and the most probable topology that lacks that clade.[3] Because the frequency of topologies in the MCMC sample is an estimate of their posterior probabilities, Bayesian support may be calculated either by dividing the frequency of the optimal topology by the frequency of the most frequent topology that lacks the clade of interest (posterior probability ratio support measure) or by subtracting the frequency of the best topology that lacks each clade of interest from the frequency of the optimal topology (posterior probability difference support measure). In this approach hypothesis optimality and support are both evaluated as a function of the posterior probability of the competing hypotheses, so this measure satisfies the adequacy condition $S(h \mid e,b) \propto O(h \mid e,b)$. The posterior probability ratio support measure is equivalent to REP support in parsimony and likelihood ratio support in maximum likelihood, and the posterior probability difference support measure is equivalent to GB support in parsimony.

This ability to calculate support directly from the MCMC sample, without additional searching, has been considered an important practical advantage of Bayesian phylogenetic analysis over maximum likelihood (Larget and Simon, 1999; Randle et al., 2005). However, to our knowledge none of the Bayesian support measures previously proposed in phylogenetics assesses support as suggested above. Instead, they employ the frequency of clades among the sampled trees (i.e. the MCMC clade frequencies) as a Bayesian support measure. Following Larget and Simon (1999), MCMC clade frequencies are often interpreted as clade posterior

[3]Another alternative, not explored here, is to compare the two topologies using Bayes factor, which measures the ratio of the posterior odds to the prior odds and therefore evaluates the change in support due to the evidence (Kass and Raftery, 1995; Lavine and Schervish, 1999), and to assign that value as the support for the clade in question (for similar applications in phylogenetics see Huelsenbeck, 2001; Suchard et al., 2005).

probabilities, whereby the posterior probability of each clade is taken as the sum of the posterior probabilities of all the topologies that contain that clade; however, the statistical validity of this interpretation is questionable given that sampling is done over individual topologies, not individual clades. Bayesian sampling theory justifies interpreting the frequency of topologies in the MCMC sample as posterior probabilities, but topologies are sets of multiple clades, and the relationship between the frequency of clades and the frequency of topologies is complex. For example, given that the most probable topology is composed of clades, it follows that the clades that constitute the most probable topology are themselves most probable. However, the frequency of contradictory clades may be greater because the sum of suboptimal topologies with a given contradictory clade may be greater than the frequency of the optimal topology (Fig. 8; for an empirical example see Wheeler and Pickett, 2008). The extent to which this complex relationship leads to under- or over-estimation of clade posteriors has not been investigated. It has been shown that uniform topological priors are not equal or equivalent to uniform clade priors (Pickett and Randle, 2005), which are in fact impossible to specify (Steel and Pickett, 2006). Regardless of the consequences for Bayesian inference (Alfaro and Holder, 2006; Velasco, 2007), this illustrates the difference between clade probabilities and topology probabilities.

As a measure of Bayesian support, MCMC clade frequencies do not satisfy the adequacy condition

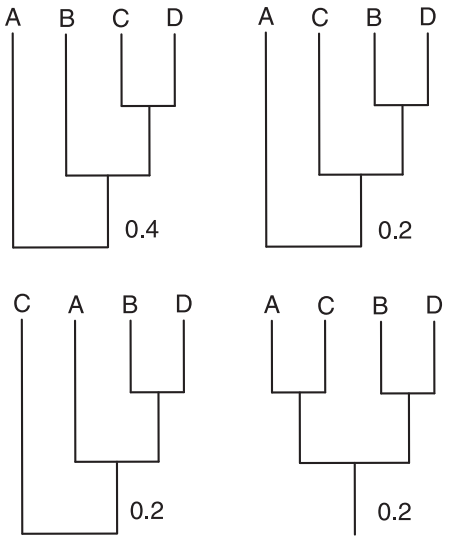

Fig. 8. Contrived example illustrating the potential conflict between MCMC clade frequencies and MCMC topology frequencies. The most frequent (0.4) and probable topology possesses clade CD and lacks clade BD. The less frequent alternative topologies (0.2 each) are less probable but all possess clade BD, resulting in an MCMC clade frequency of 0.6, greater than the frequency of the optimal clade CD. For an empirical example see Wheeler and Pickett (2008).

$S(h \mid e,b) \propto O(h \mid e,b)$ because, as noted above, MCMC clade frequencies do not necessarily vary directly with optimality (posterior probability). An important consequence of this is that clades present in the most probable estimate of phylogeny (and which are therefore most probable) may be absent from the majority rule consensus representation of trees in the MCMC sample (i.e. the Clade–Bayes topology; Wheeler and Pickett, 2008) due to low frequency among suboptimal trees, while clades that are absent from the most probable estimate (and which are therefore relatively improbable) may be present in the Clade–Bayes topology due to higher frequencies among suboptimal trees. Therefore, the common practice in Bayesian analysis (e.g. Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) of drawing phylogenetic inferences from Clade–Bayes topologies may be misleading (Wheeler and Pickett, 2008). More fundamentally, Clade–Bayes topologies have no associated optimality values, which are necessary to compare and test competing hypotheses (Wheeler and Pickett, 2008).

Bayesian inference aims to quantify belief in hypotheses and is therefore concerned with subjective knowledge. Furthermore, like other statistical approaches to phylogenetic inference, Bayesian methods require assumptions beyond those necessary to make a logically valid inference (Kluge and Grant, 2006), and many of those assumptions about the evolutionary process are counterfactual, i.e. contradicted by empirical evidence (Siddall and Kluge, 1997). For these reasons, no support measure derived from Bayesian analysis quantifies objective support. As such, Bayesian support measures do not satisfy the second adequacy condition of quantifying support in terms of explanatory power, although the posterior probability ratio and posterior probability difference support measures satisfy the adequacy condition $S(h \mid e,b) \propto O(h \mid e,b)$.

Discussion

The adequacy of support measures

The two adequacy conditions we formulated provide a common framework for the analysis of support measures, independent of the optimality criterion employed. The first adequacy condition, that support and optimality vary in direct proportion to each other, or $S(h \mid e,b) \propto O(h \mid e,b)$, ensures internal logical consistency. The second adequacy condition, that support must be quantified as a function of explanatory power, is required for the method to be a measure of objective support. Methods that eschew objectivity as a requirement of scientific knowledge claims, such as Bayesian inference, will obviously also eschew the requirement that support measures quantify objective

support, but they are still bound by internal logical consistency to satisfy the first adequacy condition, $S(h \mid e,b) \propto O(h \mid e,b)$.

Several of the parsimony and statistical support measures examined satisfy the first adequacy condition. The parsimony support measures of GB and REP support satisfy this adequacy condition and also quantify support as a function of explanatory power, making them measures of objective support. The equivalent statistical support measures are the likelihood ratio and likelihood difference support measures in maximum likelihood and the posterior probability ratio and posterior probability difference support measures in Bayesian inference. These statistical support measures satisfy the adequacy condition $S(h \mid e,b) \propto O(h \mid e,b)$, but they do not quantify support in terms of explanatory power and therefore do not measure objective support.

Neither the RFD nor any of the parsimony or statistical support measures based on clade frequencies (character resampling clade frequencies in parsimony and maximum likelihood, MCMC clade frequencies in Bayesian inference) satisfy the adequacy condition $S(h \mid e,b) \propto O(h \mid e,b)$. In parsimony analysis, Goloboff et al. (2003) noted the difficulty in interpreting low ($< 50\%$) resampling frequencies of optimal hypotheses and proposed a number of methods to mitigate this problem in the extremes (i.e. where the resampling frequency of optimal hypotheses is less than for contradictory suboptimal hypotheses), but those solutions do not address the more general problem of resampling frequencies not varying in proportion to hypothesis optimality. Indeed, Ramírez (2005) underscored the potentially contradictory ranking of clades as a strength of resampling methods. In Bayesian analysis, the common practice of employing the majority rule consensus of MCMC topologies (the Clade–Bayes topology) as the basis for phylogenetic inferences is inadvisable as it can contradict the optimal, most probable topology.

In addition to violating this general adequacy condition, the RFD and clade frequency methods do not calculate support as a function of explanatory power. Consequently, none of these measures produces objective knowledge claims. Minimally, there is a contradiction when hypothesis optimality is assessed according to one criterion (explanatory power, likelihood, posterior probability) and hypothesis support is assessed according to a contradictory criterion (the ratio of F and C, clade frequency). Regardless of the intended use of character resampling measures and the RFD in parsimony, their interpretation requires clarification, especially when one considers that the rank order of hypotheses for the two measures may disagree, even though they both aim to calculate support as the relative amount of evidence contradictory and favourable to a hypothesis (whether directly or indirectly; see Goloboff et al., 2003, p. 326). As illustrated in Figs 4 and 6, according to jackknife resampling, support is less for clade CD than for clade FG and neither clade receives maximum support, but according to the RFD CD and FG are both maximally supported.

In spite of the inadequacy of the RFD and clade frequency as support measures, they may be useful for other purposes. In the case of character resampling methods, at least one such purpose is immediately apparent. Although it has come to be recognized primarily as a method of measuring support for groups present in the optimal solution, Farris et al. (1996) introduced the parsimony jackknife as a more efficient method than neighbour-joining for analysing large datasets, and the various improvements to resampling methods proposed by Goloboff et al. (2003) and Ramírez (2005) only increase its importance in that regard. As such, the parsimony jackknife may be a heuristic method insofar as it helps to identify most parsimonious solutions (as do other heuristic procedures such as trajectory searches), but once the preferred hypothesis has been identified it is inadequate as an objective support measure. Similarly, MCMC clade frequencies are not adequate measures of Bayesian support and are questionable estimates of clade posterior probabilities, but they may be useful as credibility intervals for the clades of the most probable tree (cf. Alfaro and Holder, 2006), and MCMC sampling is useful in estimating topology posterior probabilities.

Regardless of the optimality criterion employed, the support measures that satisfy the adequacy condition $S(h \mid e,b) \propto O(h \mid e,b)$ are all derived from the comparison of optimal and suboptimal hypotheses. For a given clade present in the optimal hypothesis, this comparison may be expressed as the ratio of optimality scores of the optimal hypothesis (h) and the optimal hypothesis that lacks that clade (h'), $O(h \mid e,b)/O(h' \mid e,b)$, or the difference between the optimality scores of those hypotheses, $O(h \mid e,b) - O(h' \mid e,b)$. The REP, likelihood ratio (calculated as the difference in log-likelihoods), and posterior probability ratio support measures calculate support as a ratio using parsimony, maximum likelihood, and posterior probability as optimality criteria, respectively. Similarly, the GB, likelihood difference, and posterior probability difference support measures calculate support as a difference using parsimony, maximum likelihood, and posterior probability as optimality criteria, respectively.

Wilkinson (1994, p. 362) also noted that measures equivalent to GB support could be developed under any optimality criterion, and he proposed a "general support measure as the difference in optimized quantity between the most optimal tree that includes a particular item of cladistic information and the optimal tree that does not include this item". This measure is not strictly equivalent to GB support because "the most optimal tree that includes a particular item of cladistic information" may

not be the globally optimal tree, and the definition does not consider the possibility of calculating support as a ratio instead of a difference, but the approach is generally consistent with ours.

As the difference in steps, GB support provides clearly interpretable information that is not provided by REP support (Grant and Kluge, 2007). However, it is unclear what advantage may obtain from calculating statistical support as a difference instead of a ratio, and in maximum likelihood the size of numbers makes subtraction of raw likelihood scores impractical. Nevertheless, like the statistical support measures based on ratios, these measures satisfy the adequacy condition $S(h \mid e,b) \propto O(h \mid e,b)$ but do not quantify support in terms of explanatory power.

Concepts of support

Although we considered several parsimony and statistical support measures, many other criteria of support are often employed, either formally or informally, in phylogenetic inference (for references see Grant and Kluge, 2003). Clades delimited by "unique and unreversed" or relatively less homoplastic character-states are often considered more strongly supported, as are clades delimited by the origin of especially complex character-states. Clades delimited by synapomorphies from multiple partitions (e.g. morphological and DNA sequence characters, multiple loci, transitions and transversions, first, second, and third positions) are often considered better supported. Clades delimited by characters that are adaptive may be interpreted as especially strongly supported, as may those delimited by character-states that are believed to be non-adaptive or neutral. Hypothesis longevity is often employed implicitly as an informal criterion of support, with groups that have been recognized for a long time deemed more credible than novel hypotheses of relationship. Agreement with biogeographical scenarios is another criterion of support, as is taxic agreement, whereby strong support is attributed to groups that share the same biogeographical or coevolutionary patterns. Many of these criteria are used qualitatively but could be transformed into quantitative measures. This list is far from exhaustive, but it shows that many approaches to inferring support can be and have been employed in phylogenetic inference, all of which may be evaluated in terms of the adequacy conditions proposed above.

In one of the few studies to discuss the concept of support explicitly, Wilkinson et al. (2003, p. 129) considered evidence to support a hypothesis "to the extent that the hypothesis provides a better explanation of the evidence than some alternative hypothesis", which appears to be similar to the concept we endorse. However, Wilkinson et al.'s (2003) interpretation of both "a better explanation" and "the evidence" departs significantly from ours. As they clarified (p. 129), "[w]hen we say that a character (or a data matrix) supports a hypothesis, we imply that it fits that hypothesis better than it fits one or more incompatible hypotheses", and their concept of support was intended to apply "to the entire data set treated as a whole, to subsets of the data (by taxa or characters), and to individual characters". This conceptualization of support in terms of fit and partitions instead of explanatory power and the total (unpartitioned, combined) evidence can lead to paradoxical situations in which a refuted, suboptimal hypothesis may be "supported" and a corroborated, optimal hypothesis may be "unsupported" [thus violating the adequacy condition $S(h \mid e,b) \propto O(h \mid e,b)$], and also in which one or more characters may both refute and "support" a hypothesis, depending on which other evidence is included in the partition (Goloboff et al., 2003).

The relevance of support

Throughout this paper we have examined the adequacy of support measures, but we have paid little attention to their relevance to the science of phylogenetic systematics. Ramírez (2005, p. 88, italics in original) observed that "before *using* any support measures (regardless of the intended use or application), these measures need to be calculated first" (see also Sanderson, 1995, p. 311), but this does not make the case for the scientific relevance of these methods. Aesthetics aside, methods have no intrinsic value; they are valuable only to the extent that they further the goals of science, either directly by testing hypotheses (i.e. they are scientific) or indirectly by pointing to new or highly testable problems and hypotheses (i.e. they are heuristic; for details see Grant and Kluge, 2003). Support measures are not scientific because they do not test phylogenetic hypotheses. However, support measures are heuristic because by evaluating the relative degree of evidential support they identify those hypotheses that require less evidence to be overturned (Grant and Kluge, 2003).

Ramírez (2005) also noted that one cannot identify more weakly supported groups without also identifying those that are more strongly supported. However, the heuristic interpretation proposed by Grant and Kluge (2003) does not merely replace a *positive* interpretation of support (i.e. claiming greater reliability for more strongly supported groups) with a *negative* interpretation of support (i.e. claiming less reliability for more weakly supported groups). Rather, it is mute on the question of reliability; objectively, all groups present in the most parsimonious tree are least refuted and provide the best explanation of the critical evidence. Nevertheless, although there is no rational basis to claim that more weakly supported groups are less reliable, less contradictory evidence is required to overturn them,

which provides a rational basis for designing future studies (see also Grant and Kluge, 2007).

Consequently, even when support is conceptualized objectively as the relative explanatory power of competing hypotheses, degree of support does not provide a rational basis for greater confidence or disbelief in a group as more or less accurate, reliable, probable, or worthy of formal taxonomic recognition. All groups present in the strict consensus of most parsimonious trees are supported by the available evidence and are objectively optimal. Should this criterion of objectivity be abandoned as the basis for recognition or acceptance, then, minimally, workers must indicate the degree of support required for a group to merit recognition—1% better than competing hypotheses? 0.5% better? 10% better?—and why that particular value is preferred.

As noted by Sanderson (1995) and Grant and Kluge (2007), the interpretation of support in systematics as an indication of the relative strength of hypotheses and not as a predictor of future samples or the probability that a hypothesis is true is related, both conceptually and historically, to the interpretation of support in statistics according to the logic of maximum likelihood (e.g. Hacking, 1965). We have attempted to establish a general framework for the analysis of support by relating the concepts of optimality and support, formulating adequacy conditions that may be applied to all support measures in systematics—independent of concept, interpretation, or optimality criterion—and developing equivalent support measures in parsimony, maximum likelihood, and Bayesian analysis.

Acknowledgements

We are grateful to Pablo Goloboff, Kurt Pickett, Martín Ramírez, Mark Siddall, Leo Smith, Mark Wilkinson, anonymous reviewers, and especially Ward Wheeler for sharing their critical insights into the problem of support. However, what is published remains our responsibility entirely.

References

Alfaro, M.E., Holder, M.T., 2006. The posterior and the prior in Bayesian phylogenetics. Annu. Rev. Ecol. Syst. 37, 19–42.

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.

Crupi, V., Tentori, K., Gonzalez, M., 2007. On Bayesian measures of evidential support: theoretical and empirical issues. Philos. Sci. 74, 229–252.

Egan, M.G., 2006. Support versus corroboration. J. Biomed. Inform. 39, 72–85.

Farris, J.S., 1973. A probability model for inferring evolutionary trees. Syst. Zool. 22, 250–256.

Farris, J.S., 1979. The information content of the phylogenetic system. Syst. Zool. 28, 483–519.

Farris, J.S., 1983. The logical basis of phylogenetic analysis. In: Platnick, N.I., Funk, V.A. (Eds.), Advances in Cladistics. Columbia University Press, New York, pp. 7–36.

Farris, J.S., 1989. The retention index and the rescaled consistency index. Cladistics 5, 417–419.

Farris, J.S., Källersjö, M., Albert, V.A., Allard, M.W., Anderberg, A., Bowditch, B., Bult, C., Carpenter, J.M., Crowe, T.M., De Laet, J., Fitzhugh, K., Frost, D.R., Goloboff, P.A., Humphries, C.J., Jondelius, U., Judd, D., Karis, P.O., Lipscomb, D., Luckow, M., Mindell, D.P., Muona, J., Nixon, K.C., Presch, W., Seberg, O., Siddall, M., Struwe, L., Tehler, A., Wenzel, J.W., Wheeler, Q.D., Wheeler, W.C., 1995. Explanation. Cladistics 11, 211–218.

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

Farris, J.S., Kluge, A.G., Carpenter, J.M., 2001. Popper and likelihood versus "Popper*". Syst. Biol. 50, 438–444.

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

Giribet, G., Wheeler, W.C., 2007. The case for sensitivity: a response to Grant and Kluge. Cladistics 23, 294–296.

Goldman, N., 1990. Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analyses. Syst. Zool. 39, 345–361.

Goldman, N., 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36, 182–198.

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

Goloboff, P.A., 2003. Parsimony, likelihood, and simplicity. Cladistics 19, 91–103.

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

Goloboff, P.A., Farris, J.S., Källersjö, M., Oxelman, B., Ramírez, M.J., 2003. Improvements to resampling measures of group support. Cladistics 19, 324–332.

Goodman, M., Olson, C.B., Beeber, J.E., Czelusniak, J., 1982. New perspectives in the molecular biological analysis of mammalian phylogeny. Acta Zool. Fennica 169, 19–35.

Grant, T., Kluge, A.G., 2003. Data exploration in phylogenetic inference: scientific, heuristic, or neither. Cladistics 19, 379–418.

Grant, T., Kluge, A.G., 2004. Transformation series as an ideographic character concept. Cladistics 20, 23–31.

Grant, T., Kluge, A.G., 2005. Stability, sensitivity, science and heurism. Cladistics 21, 597–604.

Grant, T., Kluge, A.G., 2007. Ratio of explanatory power (REP): a new measure of group support. Mol. Phylogenet. Evol. 44, 483–487.

Grant, T., Kluge, A.G., 2008. Credit where credit is due: the Goodman–Bremer support metric. Mol. Phylogenet. Evol., doi: 10.1016/j.ympev.2008.04.023.

Hacking, I., 1965. The Logic of Statistical Inference. Cambridge University Press, Cambridge.

Hasegawa, M., Kishino, H., 1989. Confidence limits of the maximum-likelihood estimate of the hominoid tree from mitochondrial-DNA sequences. Evolution 43, 672–677.

Hendy, M.D., Penny, D., 1993. Spectral analysis of phylogenetic data. J. Classif. 10, 5–24.

Holmes, S., 2003. Bootstrapping phylogenetic trees: theory and methods. Stat. Sci. 18, 241–255.

Huelsenbeck, J.P., 2001. A Bayesian perspective on the Strepsiptera problem. Tijdschr. Entomol. 144, 165–178.

Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754–755.

Huelsenbeck, J.P., Larget, B., Miller, R.E., Ronquist, F., 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. 51, 673–688.

Hull, D.L., 1974. Philosophy of Biological Science. Prentice-Hall, Inc., Englewood Cliffs, NJ.
Källersjö, M., Farris, J.S., Kluge, A.G., Bult, C., 1992. Skewness and permutation. Cladistics 8, 275–287.
Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Am. Stat. Assoc. 90, 773–795.
Kearney, M.. 2007. Philosophy and phylogenetics: historical and current connections. In: Hull, D.L., Ruse, M.. (Eds.), Cambridge Companion to the Philosophy of Biology. Cambridge University Press, Cambridge, MA, pp. 211–232.
Kluge, A.G., 1997. Testability and the refutation and corroboration of cladistic hypotheses. Cladistics 13, 81–96.
Kluge, A.G., 2001. Philosophical conjectures and their refutation. Syst. Biol. 50, 322–330.
Kluge, A.G., 2002. Distinguishing "or" from "and" and the case for historical identification. Cladistics 18, 585–593.
Kluge, A.G., 2007. Completing the neo-Darwinian synthesis with an event criterion. Cladistics 23, 613–633.
Kluge, A.G., Grant, T., 2006. From conviction to anti-superfluity: old and new justifications for parsimony in phylogenetic inference. Cladistics 22, 276–288.
Lakatos, I., 1978. The Methodology of Scientific Research Programmes. Cambridge University Press, Cambridge.
Larget, B., Simon, D., 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. 16, 750–759.
Lavine, M., Schervish, M.J., 1999. Bayes factors: what they are and what they are not. Am. Stat. 53, 119–122.
Lee, M.S.Y., Hugall, A.F., 2003. Partitioned likelihood support and the evaluation of data set conflict. Syst. Biol. 52, 15–22.
Lento, G.M., Hickson, R.E., Chambers, G.K., Penny, D., 1995. Use of spectral analysis to test hypotheses on the origin of pinnipeds. Mol. Biol. Evol. 12, 28–52.
Meireles, C.M., Czelusniak, J., Schneider, M.P.C., Muniz, J.A.P.C., Brigido, M.C., Ferreira, H.S., Goodman, M., 1999. Molecular phylogeny of ateline New World monkeys (Platyrrhini, Atelinae) based on γ-Globin gene sequences: evidence that *Brachyteles* is the sister group of *Lagothrix*. Mol. Phylogenet. Evol. 12, 10–30.
Miller, J.A., 2003. Assessing progress in systematics with continuous jackknife function analysis. Syst. Biol. 52, 55–65.
Mossel, E., Vigoda, E., 2006. Limitations of Markov chain Monte Carlo algorithms for Bayesian inference of phylogeny. Ann. Appl. Probab. 16, 2215–2234.
Notturno, M.A., 2003. On Popper. Wadsworth, Thompson Learning, Toronto.
Pickett, K.M., Randle, C.P., 2005. Strange Bayes indeed: uniform topological priors imply non-uniform clade priors. Mol. Phylogenet. Evol. 34, 203–211.
Popper, K.R., 1957. The Poverty of Historicism. Routledge and Kegan Paul, London.
Popper, K.R., 1979. Objective Knowledge: An Evolutionary Approach. Oxford University Press, New York.
Popper, K.R., 1983. Realism and the Aim of Science. Routledge, London.
de Queiroz, K., 2004. The measurement of test severity, significance tests for resolution, and a unified philosophy of phylogenetic inference. Zool. Scr. 33, 463–473.
Ramírez, M.J., 2005. Resampling measures of group support: a reply to Grant and Kluge. Cladistics 21, 83–89.
Randle, C.P., Mort, M.E., Crawford, D.J., 2005. Bayesian inference of phylogenetics revisited: developments and concerns. Taxon 54, 9–15.
Rannala, B., Yang, Z., 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J. Mol. Evol. 43, 304–311.
Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.
Sanderson, M.J., 1995. Objections to bootstrapping: a critique. Syst. Biol. 44, 299–320.
Siddall, M.E., 1995. Another monophyly index: revisiting the jackknife. Cladistics 11, 33–56.
Siddall, M.E.. 2002. Measures of support. In: DeSalle, R., Giribet, G., Wheeler, W.C. (Eds.), Techniques in Molecular Systematics and Evolution. Birkhäuser Verlag, Basel, pp. 80–101.
Siddall, M.E., Kluge, A.G., 1997. Probabilism and phylogenetic inference. Cladistics 13, 313–336.
Stanhope, M.J., Waddell, V.G., Madsen, O., de Jong, W., Hedges, S.B., Cleven, G.C., Kao, D., Springer, M.S., 1998. Molecular evidence for multiple origins of Insectivora and for a new order of endemic African insectivore mammals. Proc. Natl. Acad. Sci. U.S.A. 95, 9967–9972.
Steel, M., Pickett, K.M., 2006. On the impossibility of uniform priors on clades. Mol. Phylogenet. Evol. 39, 585–586.
Suchard, M.A., Weiss, R.E., Sinsheimer, J.S., 2005. Models for estimating Bayes factors with applications to phylogeny and tests of monophyly. Biometrics 61, 665–673.
Tierney, L., 1994. Markov chains for exploring posterior distributions. Ann. Stat. 22, 1701–1762.
Tuffley, C., Steel, M., 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull. Math. Biol. 59, 581–607.
Velasco, J.D., 2007. Why non-uniform priors on clades are both unavoidable and unobjectionable. Mol. Phylogenet. Evol. 45, 748–749.
Watkins, J., 1997. Popperian ideas on progress and rationality in science. Crit. Rationalist 2, 2–11.
Wheeler, Q.D., Blackwell, M. 1984. Cladistics and the historical component of fungus-insect relationships. In: Wheeler, Q.D., Blackwell, M. (Eds.), Fungus–Insect Relationships: Perspectives in Ecology and Evolution. Columbia University Press, New York, pp. 5–41.
Wheeler, W.C., Pickett, K.M., 2008. Topology-Bayes versus Clade–Bayes in phylogenetic analysis. Mol. Biol. Evol. 25, 447–453.
Wilkinson, M., 1994. Common cladistic information and its consensus representation: reduced Adams and reduced cladistic consensus trees and profiles. Syst. Biol. 43, 343–368.
Wilkinson, M., Lapointe, F.-J., Gower, D.J., 2003. Branch lengths and support. Syst. Biol. 52, 127–130.
Yang, Z., Rannala, B., 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Mol. Biol. Evol. 14, 717–724.
Zwickl, D., 2006. Genetic algorithm approaches for the analysis of large biological sequence datasets under the maximum likelihood criterion. PhD thesis, University of Texas at Austin, Austin.