Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Biomedical Informatics 39 (2006) 72–85
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Methodological Review

Support versus corroboration

Mary G. Egan*

Division of Invertebrate Zoology, American Museum of Natural History, 79th Street, Central Park West, New York, NY 10024, USA

Received 22 September 2005
Available online 7 December 2005

Abstract

Numerous metrics have been developed that attempt to assess the reliability of phylogenetic trees. Several of these commonly used measures of tree and tree branch support are described and discussed in the context of their relationship to Popperian corroboration. Claims that measures of support indicate the accuracy of phylogenetic trees or provide information for tree choice are rebutted. Measures of support are viewed as being of heuristic value within a given phylogenetic framework for describing the precision of the data based on perturbations to the data. However, no direct link is observed between the calculation of measures of support and corroboration. Direct measures of support, but not re-sampling or randomization methods, may play a more specific role in phylogenetic inference by providing the tools to search for falsifiers that could be the subject of future rounds of hypothesis testing.

Keywords: Phylogeny; Parsimony; Corroboration; Support; Robustness; Branch support; Bremer support; Bootstrap; Jackknife; Permutation

1. Introduction

Phylogenetic methods have come to be used by researchers in numerous areas of study. In addition to providing systematists information for classification, phylogenetic trees have been used in a wide range of studies that make inferences about the evolution of morphological and molecular characters. Numerous comparative studies incorporate information from phylogeny for testing hypotheses regarding a range of topics including: the evolution of development, biodiversity assessment, the history of transposable elements, biogeographic patterns of species distributions, the track of infectious disease epidemics, patterns of adaptive radiation, evaluation of extinction risk as well as the evolution of body size and genome size (see [1–6] for examples). The reliability of the inferences made is contingent upon the underlying tree. Not surprisingly, numerous metrics have been developed that attempt to assess the reliability of phylogenetic trees.

Some of these measures result in indices of support for entire trees, other measures apply to particular groupings of taxa, still others attempt to provide a measure of the potential phylogenetic information content of a data matrix prior to tree building. Some measures take a direct approach by examining the relative amount of character support and conflict on the tree. Other measures use statistical tests and construct confidence intervals for examining departures from a null distribution of randomized data.

Among the numerous methods of phylogenetic inference, character-based methods such as parsimony [7–9] and maximum likelihood [10,11] are the most widely used. In addition, Bayesian analysis [12], which has a long history in statistics, has more recently come to be used for phylogenetic analysis [13–16]. Most of the measures of support described in this review can be applied to both parsimony and maximum likelihood methods of phylogenetic inference. However, these character-based methods that differ in their approach to tree building also differ with respect to the interpretation of the meaning of measures of support.

What is the goal of calculating measures of support? What do these measures of support tell us? Do they indicate the accuracy of the hypothesis (the true tree)? Do they provide information for choosing among alternate hypotheses (the best tree)? In addition to these

* Fax: +1 212 769 5277.
E-mail address: egan@amnh.org.

1532-0464/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2005.11.007

differences in terms of the interpretation of the meaning of support values, there are differences in terms of the definition of support. Support (or robustness), is commonly viewed in the literature as the stability of the hypothesis (tree) to perturbations in the data, parameters or analysis. Less commonly seen in the bioinformatics literature is the parsimony perspective which views support in terms of corroboration (sensu Popper [17,18], see also [19,20]). The varied answers to these questions affect decisions regarding choice of metric. This in turn may influence the direction for the development of new bioinformatic tools for phylogenies.

1.1. Goodness of fit metrics—tree statistics

Although not measures of support per se [21], in addition to tree length, tree fit statistics are among the most commonly reported. The consistency index (ci) [8] is a measure of the fit of a character to the tree. It is defined as m/s, in which m is the minimum amount of change that the character may exhibit on any tree and s is the observed amount of change for the character on the most parsimonious tree. The complement of the consistency index ($1 - \text{ci}$) is a measure of the amount of homoplasy. When there is no homoplasy, in other words, the character fit is perfect, $\text{ci} = 1$. This can also be calculated for sub-optimal trees.

The retention index (ri) [22] is a measure of the amount apparent synapomorphy that is retained as synapomorphy on the tree. It is defined as $(g - s)/(g - m)$; m and s are the same as for the ci and g is the greatest amount of change the character may exhibit on any tree. The ri can vary between 0 and 1. If the minimum amount of change observed for the character on the tree (s) is equal to the greatest amount of change possible on any tree (g), the retention index will be zero. If the minimum amount of change observed is the minimum possible ($s = m$) then $\text{ri} = 1$, indicating that there is no homoplasy.

Ensemble values for the consistency and retention indices, denoted CI and RI respectively, can be calculated for the entire suite of characters on the tree by first summing the individual values of m, s, and g for all characters and then calculating the indices in the same manner as for the individual character fits.

There are several known problems with the ci. Noting that the ci can never reach zero since it can never be less than m/g, Farris developed the rescaled consistency index (rc or RC) [22] which is equal to the product of the ci and ri. It is recognized that CI decreases when the number of taxa increases. In addition CI values are influenced by uninformative characters which result in an inflation of the CI. Therefore CI values are generally calculated after the exclusion of uninformative characters. Archie [23] proposed the Homoplasy Excess Ratio Maximum HERM (an estimate of his Homoplasy Excess Ratio) to address the problems with CI. However, HERM is identical to RI [22,24].

1.2. Re-sampling measures of support

1.2.1. Bootstrap

One of the measures of support most commonly seen in the literature is the nonparametric bootstrap of Felsenstein [25], see also [26,27]. Bootstrapping was first used in statistics [28–31] for parameter estimation, in cases where the distribution is unknown. By sampling with replacement from the rows in the matrix to generate many pseudoreplicate matrices, a null distribution is generated from the data at hand. Bootstrapping, as applied to phylogenetic trees involves the generation of numerous pseudoreplicate data matrices of the same size as the original matrix by re-sampling characters (columns), with replacement, from the original matrix. Re-sampling with replacement is, in effect, a random weighting strategy in which the weight of some characters is increased and others are effectively given a weight of zero [32]. Trees are built (by whichever method of phylogenetic inference) from each pseudoreplicate matrix as well as the original matrix and the topologies screened. This poses computational problems since a large number of pseudoreplicate matrices would need to be analyzed. One solution would be to perform simple searches without branch swapping [33]. The proportion (expressed as a percentage) of bootstrap replicates in which a given clade appears is used as an indicator of support and has been proposed as a confidence interval [25,34] for the accuracy of the clade. Results are expressed as a consensus of trees (often a majority-rule consensus) showing those clades that are present in greater than 50 percent of the bootstrap replicates.

1.2.2. Jackknife

Derived from statistical jackknife analysis [35], Lanyon [36] first proposed the use of the taxonomic jackknife for detecting inconsistency in distance data; evidenced by regions of the tree that are unstable to taxon removals. The premise behind it is that in the absence of inconsistency within the data set, there is sufficient redundancy in the signal such that the removal of a single taxon should not disrupt the topology of the tree (except for the node containing the removed taxon). Pseudoreplicate matrices are generated, each lacking a different ingroup taxon and trees are generated for each of these matrices. As originally proposed, results were assessed by constructing a jackknife strict consensus (JSC) tree of the trees from all analyses, and graphically distinguishing stable from unstable nodes. The JSC tree would contain the "set of nodes consistent with or shared by all pseudoreplicate trees" [36, p. 399]. Later, Siddall [37] objected to the view [36] that a strict consensus of jackknife replicates could be a better estimate of phylogeny than the most parsimonious trees, and proposed the jackknife monophyly index (JMI) as an alternative that uses the frequency with which a clade appears in the jackknife pseudoreplicate trees as an indication of the stability of that clade to this type of data perturbation. Siddall [37] was explicit in stating that the JMI was not to be

interpreted as a statistical test, or viewed in relation to a null distribution, nor used for comparing support values across data sets. Rather that it should be viewed as an "analytical tool designed to provide some insight into the affect of homoplasy on the relative stability of clades in most parsimonious trees" [37, p. 47].

Jackknife analyses also can be performed by re-sampling characters instead of taxa. Parsimony jackknifing [38–40] is used to assess the stability of clades to character removal. Informative characters are sequentially removed and phylogenetic analyses performed on the perturbed matrices in order to identify clades that are lost in resulting strict consensuses of trees. The clade stability index (CSI) [41] is defined as the minimum number of character removals required for clade loss divided by the number of informative characters. While similar to jackknifing (as statistical re-sampling), it differs in that there is a direct measure of support that identifies the number of (and specifically which) characters contribute support. This type of analysis would be computationally intensive. To test for all possible single character removals would require trees to be generated for as many matrices as there are characters, and this would need to be repeated to examine all possible two-character removals, etc. In addition, the removal of one character at a time may be of limited utility because it may not result in much variation among resulting trees [25]. An alternative is to direct the searches by first identifying characters that are synapomorphies of clades and target those characters for removal [41], or to remove subsets of characters [40]. Extensions of the CSI include the character removal index (CRI) [41] which is the minimum number of character removals required to collapse a node and the data set removal index (DRI) [42]; the minimum number of data set removals required for node collapse.

Parsimony jackknifing [40] generates replicate matrices by randomly and independently removing groups of characters, where each character would have the same probability of removal. Parsimony jackknifing has been used as a faster alternative to traditional tree searches (that use extensive branch swapping) for identifying well supported nodes [40,43–45], however, the advent of more efficient tree search algorithms such as the parsimony ratchet [46], and search strategies using tree drifting and fusing and sectorial searches [47] may make this use of the jackknife unnecessary. In addition, there are philosophical concerns about the use of measures of support for informing decisions about tree choice [37,48,49].

1.3. Randomization/permutation methods

Randomization methods differ in several respects, from re-sampling methods. Re-sampling methods attempt to assess support based on the stability of the hypothesis to perturbations of the data. Re-sampling with replacement (i.e., bootstrap) can be thought of as testing the stability of the hypothesis to differential weighting of the characters. Re-sampling without replacement (i.e., jackknife) is often thought of as a test of the stability of the hypothesis to the addition of data. Although, in reality, each jackknife replicate consists of the analysis of a matrix that has been reduced in size (in which either taxa or characters have been removed). Therefore, the proportion of jackknife replicates in which a given clade is present, is an indication of how well that clade would have been resolved had there been less data available; not how well that clade might be resolved if more data were available in the future. In contrast, randomization methods calculate support based on significant departures from random structure, and support is calculated for the entire tree rather than for nodes on the tree. In addition, support is evaluated in terms of the ability of the data matrix to provide information for tree choice.

1.3.1. Distribution of tree length skewness and the permutation tail probability

Several authors [50–54] have examined the distribution of tree lengths as an indicator of what has been variously referred to as: phylogenetic signal, informativeness, or decisiveness. Beginning with the premise that the preferred tree is the one with the fewest evolutionary changes, the shortest tree, methods such as the distribution of tree length skewness (DTL) [50–52], and the permutation tail probability (PTP) [53,54], also known as the Archie and Faith and Cranston method (AFC), look to the skewness of tree length distributions for information for tree choice. Matrices that result in strongly left-skewed distributions, with an attenuated tail, are considered to be more informative (or demonstrate greater phylogenetic signal) because they allow discrimination among near-optimal solutions. In DTL skewness, all possible trees are generated (from the actual matrix) and the skewness of the distribution of their tree lengths is calculated using the g_1 statistic [55]. DTL skewness g_1 is considered significant if the observed g_1 falls in the tail of the distribution of g_1 values obtained from the analysis of matrices of random characters [51]. The significance of DTL skewness has also been evaluated relative to simulated data [52].

The AFC method is viewed as a significance test of phylogenetic structure. Unlike DTL skewness, which compares the length of the shortest tree to the lengths of all possible trees generated from the original data matrix, AFC evaluates significance based on the departure in length of the actual tree from the distribution of tree lengths generated from data in which the character states are permuted among taxa.

1.3.2. Relative apparent synapomorphy analysis

Relative apparent synapomorphy analysis (RASA) [56,57] was presented as a statistical tool for the a priori testing of the amount of phylogenetic signal (considered by Lyons-Weiler et al. [56, p. 749], to be equivalent to character covariation) in a data matrix. "its use would then prevent the construction (and application) of a potentially spurious phylogeny" [56, p. 754]. It was presented as a faster alternative, which can be solved in polynomial time, to

measures of tree support (such as tree length distribution skewness [51,52] and PTP [53,54,58] methods) in which the finding of MP trees for matrices with large numbers of taxa is an NP-complete problem. Similar to tree taxon analysis [59,60], a character state shared between two taxa (i and j) to the exclusion of any other taxon is considered an apparent synapomorphy. RAS (apparent "cladistic" similarity) is the sum of these three taxon statements across all characters for which i and j share a character state to the exclusion any other taxon. Phenetic similarity (E) is defined as the number of characters that support the three taxon statements. The observed rate of increase in apparent similarity as a function of phenetic similarity (the slope of RAS_{obs} plotted against E_{obs}) is compared to the slope of a null distribution of RAS_{null} plotted against E_{null}. Significance is assessed by the test statistic (t_{RASA}) approximating the Student's t distribution. The original null model, of an equiprobable distribution of RAS and E [56] was later modified [57,61] to a permutation model [53] to better approximate the Student's t distribution.

1.4. Statistical and philosophical objections to probabilistic methods

There are drawbacks to the use of the bootstrap, and re-sampling methods in general, for confidence intervals [25,48,62–64] due to the violation of the assumptions of the statistic. Re-sampling methods assume that characters are independent and identically distributed (iid), furthermore that the samples are collected randomly, and finally that the data sample is sufficiently large such that the re-sampling with replacement from the data at hand is the same as sampling from the parametric distribution. In other words, re-sampling methods assume that a large number of characters are sampled randomly by the investigator from a pool of possible (uncorrelated) characters that have the same distribution. The requirement for a large sample set is likely to be violated for the generally small size (albeit increasing in size) of current phylogenetic matrices. In addition, it seems certain that neither morphological nor molecular characters are selected at random by the investigator. Perhaps "the most serious challenge to the use of bootstrap methods" [25, p. 785], and by extension re-sampling methods in general, is the violation of iid. While a multinomial distribution is assumed for sequence data [25], it "might be argued that this presupposes that the same probabilistic evolutionary process is operating in all of the characters, which is extremely unrealistic" [25, p. 785]; especially in the case of data sets with multiple loci or coding regions in which different codon positions may be subject to different constraints. In addition, there is the problem that occurs when characters are correlated which results in an overestimate of the actual number of independent characters and an upward bias in terms of level of support. In fact, evaluations of the statistical properties of the bootstrap have shown that it can be a biased estimator [65], but see also [66–69].

Philosophical objections to re-sampling methods stem from their implicit "truth" claims. It has been argued that phylogenetic inference should be viewed as a matter of statistical inference in which an unknown quantity is estimated, with some uncertainty, by means of probabilistic models of evolution [27, p. 523]. Support in this context can be viewed as an evaluation of the variance of the estimate. Central to statistical methods of inference is the concept of statistical consistency; as sample size increases, the estimate will more closely reflect the true value of the parameter being estimated—well-supported trees accurately reflect the true tree. There have been many studies comparing and evaluating the different methods of phylogenetic inference based on statistical consistency, in other words, their ability to accurately recover the true tree—generated from simulated data or a known phylogeny (see [70], for a recent review of these types of studies). There are also numerous studies evaluating measures of support in light of statistical consistency [71–74]. Practitioners of parsimony methods take the view that the true tree, the actual pattern of evolutionary events, is unknowable and reject the idea that measures of support indicate the accuracy of the tree. The use of probabilistic methods for evaluating singular historical events is questioned [48,63,75]; as is their use in tests of the significance of phylogenetic structure [19,48,76–80].

In addition, several methods have been proposed, such as parsimony jackknifing [40], some consensus methods, and the methods of implied weights [81] (see also [82]), support weighting [83] and the a posteriori successive approximations weighting [84], that are designed to provide the means for choosing among equally parsimonious trees. The justification for these methods of tree choice is that, unlike probabilistic methods, the criterion for choice is based on character fit. However the use of these methods is controversial since they represent a shift in optimality criterion from "shortest" trees minimizing ad hoc assumptions of homoplasy to "heaviest" trees maximizing fit [49,62].

The justification for this approach has its grounding in the philosophy of Popper. Its emphasis is on refutation as a means toward the growth of knowledge. Popper's refutationist, deductive method was proposed as a rational solution to the problem of induction. Inductive reasoning based on the premise that the future will be like the past, would derive universal laws from repeated singular observations. The problem of induction can be stated as: "Are we justified in reasoning from [repeated] instances of which we have experience to other instances [conclusions] of which we have no experience?" [85, p. 4]. Popper's reformulation of the problem is stated as: "Can the claim that an explanatory universal theory is true be justified by 'empirical reasons;' that is, by assuming the truth of certain test statements or observation statements (which, it may be said, are 'based on experience')?" [85, p. 7]. His answer to these questions is No, "no number of true test statements would justify the claim that an explanatory universal

theory is true'' [85, p. 7], or even probable. However, test statements (evidence) may allow ''*us to justify the claim that an explanatory universal theory is false*.'' [85, p. 7], italics in original.

Randomization methods, like re-sampling methods are verificationist in nature, and are objected to based on their assertion that the methods indicate something about the accuracy of the hypothesis. These methods ''involve statistical tests for departure of data sets from random structure, and some authors speak of *confidence* in monophyletic groups supported by data sets for which the null hypothesis of randomness is rejected. . .it is one thing to demonstrate that a data set is nonrandom in structure, and quite another to argue that the structure in the data is an accurate reflection of patterns of shared ancestry'' [39, p. 193] (italics in original).

In addition, randomization methods claim to provide information for choosing the best tree. But do measures of support provide information for tree choice? Parsimony, taking as a given that the true tree is unknowable, and that parsimony may or may not recover the true tree, operates by logical consistency (See Farris' classic paper describing this approach [86]; see also Kluge [9]) by restricting itself to a methodological approach that minimizes the number of assumptions that are made to explain the data. In other words, it restricts itself to attempts to find the tree that is the best explanation of the data by minimizing ad hoc assumptions (explaining away of contradictory evidence) of homoplasy, This Popperian preference for the least refuted hypothesis is linked to Hennig's [7] auxiliary principle. Rather than assume, a priori, an independent origin (convergence) for a given character, parsimony maximizes character congruence and minimizes homoplasy. The parsimony criterion provides information for tree choice; the best tree is the least refuted tree, the one with the fewest number of evolutionary changes (or steps, i.e., the shortest tree). This is not to say that homoplasy is assumed to be minimal nor that evolution is parsimonious. Farris [86] gives the analogy of regression analysis in which a regression line is drawn by minimizing variance around the line while remaining neutral as to the magnitude of the variance (see Phillips this volume, for a review of the literature on homoplasy). The parsimony criterion is merely used as a methodological, decision-making tool that maximizes conformity to evidence [48,79]. PTP methods appear to be grounded in parsimony however, in justifying methods such as the PTP in terms of Popperian corroboration, - corroboration has been re-interpreted.

2. Direct measures of support

2.1. Branch support

Branch support (BS) [87,88] is another commonly reported measure of support. Also referred to as Bremer support in recognition of its author [78] and as decay index [89]; branch support provides a direct measure of the number of character steps contributing to clade support by comparing the length of the most parsimonious (MP) tree(s) to the MP tree(s) lacking a given node, thus providing a measure of the stability of a given clade to relaxation of the parsimony criterion. It can be implemented as originally described [87] by performing tree searches for all trees one step longer than the original MP tree and constructing a strict consensus of resulting trees. Those nodes no longer present in the consensus of trees one step longer than the MP tree are given a BS value of one (as originally defined, those clades that are not present in all equally parsimonious trees [i.e., are collapsed in a strict consensus of MP trees] have a BS value of zero). The same procedure is followed searching for all trees up to two steps longer followed by three steps longer, and so forth until there are no nodes remaining to be collapsed (i.e., branch support has been calculated for all nodes). This procedure may be time consuming for nodes with a large amount of character support. As an alternative, BS can also be calculated by searching for all trees lacking a given node and subtracting the length of the MP tree from the lengths of the shortest trees lacking that node [42,90]. Total support (TS) is the sum of BS values for all nodes on the tree [78]. The total support index [88] scales the TS to tree length and the proportional support index [91] scales TS to the total potential support, defined as the product of CI and tree length. Branch support calculations can be implemented in PAUP*, [92] by using the constraints feature and sequentially performing searches for MP trees lacking each node of interest. Constraint files are generated manually or by using applications such as AutoDecay [93] and TreeRot [94] which generate a tree file that can be read into PAUP*. This file contains a tree for each possible node showing all nodes unresolved except for the node of interest. Anti-constraint searches (ENFORCE Converse) are performed for each of these trees and compared to the MP tree(s). BS can be calculated directly in both NONA [95] and TNT [96].

Lee [97] proposed a significance test (clade significance index) of BS in which Templeton tests [98] are sequentially performed for each clade on the MP tree compared to trees lacking each node of interest. If the MP tree is better supported than the shortest tree lacking the node of interest at $P < 0.05$, based on the Templeton test, then the clade is considered to show significant support. Where multiple equally parsimonious trees and/or multiple constraint trees exist, Templeton tests would be performed on all possible pairwise comparisons of trees and the least significant P-value used as an indication of clade significance.

2.2. Linked branch support and linked branch support index (LBSI)

Noting that BS scores for nodes within a tree may not be independent, insofar as the collapse of a single node may result in the simultaneous collapse of additional nodes (or alternatively, make the collapse of another node more

costly), Gatesy suggested [99] that the calculation of linked branch support (LBS) along with BS may provide a better indication of the stability of nodes (as well as the entire tree) to the relaxation of parsimony. Linked branch support [99] is calculated in a similar manner to BS except that the length of the MP tree(s) is compared to constraint trees lacking more than one node at a time. Calculations can be made for all possible combinations of compatible constraints to examine how the stabilities of nodes are affected by interaction.

Gatesy [99] traced the idea for examining compatible constraints to discussions of the T-PTP [58] and by extension the AG T-PTP [100], however neither of these papers discussed the implications of clumped versus dispersed homoplasy shown in the analyses of LBS. T-PTP and AG T-PTP can be ultimately traced to BS [87] and Total Support (TS) [78,88] in that the T-PTP and AG T-PTP evaluate BS and TS in the context of the randomization test proposed by Archie [53], later referred to as PTP [54] (see [101] for a history of the naming of these methods). T-PTP and AG T-PTP are in effect significance tests of BS and TS respectively and if compatible constraints were examined by the same methodology, it would constitute a significance test of LBS. These tests differ from the Wilcoxon rank-sum test [98] used by Lee [97] in that the lengths of the MP tree and constraint trees are compared to the respective lengths of trees recovered from multiple analyses of matrices in which the character states have been randomly permuted within each character.

2.3. Partitioned branch support

Partitioned Bremer support [102], also known as partitioned branch support (PBS) [90], examines the relative contribution of data partitions to branch support within a simultaneous analysis framework. Derived from BS, in which comparisons are made between the length of the MP tree(s) and the shortest trees lacking a given node, PBS is calculated by subtracting the length of the partition on the MP tree (based on the simultaneous analysis of the entire data matrix) from its length on the shortest trees (based on simultaneous analysis) lacking a given node. The difference in lengths is the contribution of a given partition to branch support at that node on the simultaneous analysis tree. PBS can be positive (indicating character support), negative (indicating conflict) or zero. The sum of PBS for all partitions at a node is equal to BS at that node. The calculation of PBS for less well supported nodes (based on low BS values, for example) can distinguish between low support values due to a lack of character information and poor support values due to character conflict among partitions. If more than one constraint tree is found, PBS is generally calculated for each and averaged [102], however the range of PBS values could also be reported [103]. PBS has also been scaled to the number of phylogenetically informative characters or to the minimum number of steps in an effort to allow comparisons across analyses [104,105].

2.4. Partitioned hidden branch support and hidden branch support

Hidden support (HBS) and partitioned hidden support (PHBS) [42,90,106] further extend the concept of BS and PBS by evaluating the amount of support for the simultaneous analysis tree that would have remained hidden if the data partitions were analyzed separately. This approach has also been implemented within a maximum likelihood framework [107,108]. Hidden support/conflict (HBS) is quantified as the difference between the support (BS) for a given node in the combined analysis tree and the sum of the BS values for that node in each of the separately analyzed partitions. Similarly, PHBS is equal to the PBS (for a given partition at a given node) minus the BS for that node in a separate analysis of that partition. The sum of the PHBS values across all partitions for a given node is equal to the HBS at that node.

2.5. Nodal data set influence

Nodal data set influence (NDI) [42] explores the effect of data set removal on branch support (BS) scores. It is similar to PBS in that it is a measure of the support provided by a particular data partition at a particular node. However it differs from PBS in method of calculation. NDI is equal to the BS value for a given node based on the combined analysis of all data partitions minus the BS value for a given node based on the combined analysis of a perturbed matrix that is lacking the partition of interest. In this sense it is similar to the jackknife in that perturbed matrices are sequentially generated, each lacking a partition (or combination of partitions) of interest. Each perturbed matrix is analyzed and BS values calculated and compared to the BS scores obtained from the combined analysis of all the data. In general PBS and NDI will give similar results however NDI summarizes the amount of BS, PHBS as well as the amount of HBS that the partition brings out in other data sets in combined analysis [42]. If BS increases with the addition of the data partition, NDI is positive. Conversely, if BS decreases with the addition of the data partition, NDI will be negative.

3. Measures of support and support (corroboration)

3.1. Corroboration

In answer to the question of what is support, it seems clear that none of the measures of support is directly linked with corroboration (sensu Popper). Paraphrasing [17–20,49,85,86,109], corroboration provides a logical framework for the growth of knowledge with refutation as the means. The degree to which a theory is falsifiable is directly related to its ability to be corroborated. In this context bold conjectures that go beyond currently accepted background knowledge are to be preferred since they would have greater falsifiability. In other words, the probability of the

evidence should be low given background knowledge alone; since evidence that has a high probability relative to background knowledge goes little beyond what is already known. In addition, the background knowledge (that is assumed but not tested) should not be determinate to the outcome of the analysis. Matrix permutation measures of support such as the PTP that would consider a null distribution as background knowledge fail in this respect due to this lack of independence. Re-sampling methods and randomization/permutation measures of support also fail since their approach is an inductive and verificationist search for truth or some probability of truth. Even if these methods did not violate the assumptions of the statistics and could be shown to be statistically consistent, they would not be logically consistent with corroboration in which all theories are conjectural and truth is unknowable. Measures of support do not indicate the accuracy of the tree.

3.2. Degree of corroboration

On the face of it, this seems an unsatisfactory state of affairs. At some level a research program is interested in having some level of confidence that the hypothesis (tree), for example, is a reliable enough approximation with which to move forward; to make classifications or other inferences. The logical framework of Popper does not imply that a preferred theory is in fact true. Neither does it speak to the falsity of a theory. However, the framework provides the methodology for preference among competing theories. Degree of corroboration is a report of the past performance of the theory; how well it has withstood critical tests. However it is not a report of the future performance of a theory (i.e., how it may stand up to future tests). Critical tests, refers to sincere attempts to refute the theory, these would be based on all available data. It is the evidence that provides potential falsifiers of the theory. However the presence of a falsifier (character conflict—character that supports an alternate topology) does not prove the theory false (absolute falsification). The weight of evidence provides the means for choosing among competing theories. To arbitrate weight of evidence requires the minimization of ad hoc assumptions. For example, given two trees, neither of which is perfectly congruent (i.e., each has some amount of characters that conflict with the topology), which is the better tree? Which is the preferred theory? At this stage, character conflict can be seen as homoplasy (independent origin of the character), or as error in the identification or coding of the character. Regardless, characters that conflict may support an alternate topology. In order to tentatively accept any theory in the face of conflicting evidence would require that each instance of conflicting evidence be explained away as error, thus protecting the theory from these falsifiers. This explaining away of conflicting evidence on a case by case basis is ad hoc since it is not based on any outside information that these data are in fact in error. The requirement for the minimization of ad hoc assumptions prevents the process of choosing among theories from becoming no more than the irrational belief in the superiority or truthfulness of one theory over another with no reference to evidence. Critical tests and a minimization of ad hoc assumptions allows for choice among topologies in the face of conflicting evidence. Preference is for the least refuted theory at a given time with a given set of evidence. Methods that attempt to maximize the probability of the evidence given the hypothesis by incorporating models of sequence evolution as background knowledge are logically inconsistent with degree of corroboration for choosing among hypotheses. The incorporation of a model of evolution or a null distribution as background knowledge excludes the parameters of the model from testing. In addition, since the model is deterministic to the outcome, the process is circular. Furthermore, the use of a model to "correct for" substitution rates, for example, immunizes the hypothesis from critical tests by excluding potential falsifiers thereby losing the logical basis for preference.

3.3. Stability

If measures of support neither provide information for tree choice nor information about the accuracy of the tree, what is the role, or place, of measures of support in phylogenetic inference? At some level, each of the measures of support describes the precision of the data. In addition, these measures describe different aspects of precision by assessing stability to perturbations. Most of the measures can be calculated within either a parsimony or maximum likelihood framework of phylogenetic inference. Even the direct measures of branch support and partitioned branch support that have traditionally been used within a parsimony framework, have counterparts; Likelihood Support and Partitioned Likelihood Support [107,108] that can be assessed within a Likelihood framework. It may be that within a given method of phylogenetic inference, the calculation of some suite of measures of support may give a general indication of the magnitude or lack of conflict of the character support. So each of these measures that explore stability to perturbation may have a general heuristic value but are limited in what they can specifically say about the data and hypothesis.

3.4. Research cycles and the search for falsifiers

There may be a place in phylogentic inference where measures of support can play a more specific rather than a general role. Phylogenetic analysis has been described as cyclic research cycles of sophisticated falsification [49] rather than an endpoint. Fig. 1 is a schematic of the strategy described by Kluge [49]. The cycles begin (in the upper left of the figure) with selection of taxa, character identification, coding and matrix assembly. Continuing with the inner layer of the spiral, character coding is followed by hypothesis testing (tree search). After each round of tree

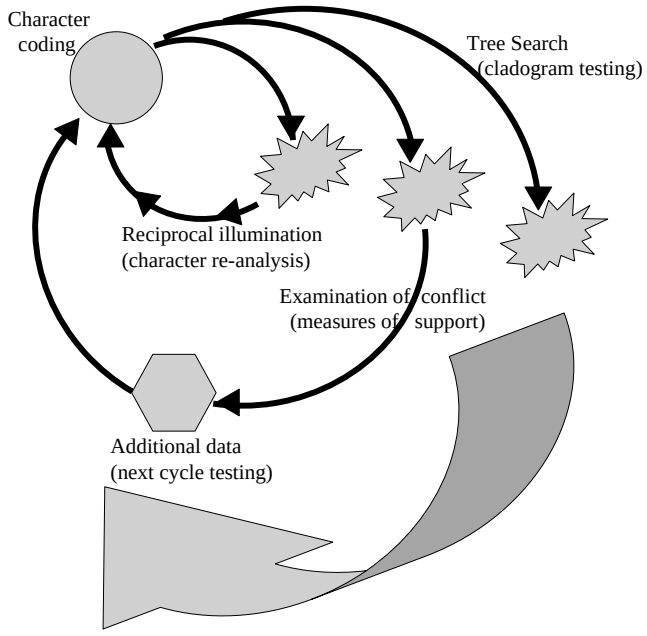

Fig. 1. Schematic of the cycle of phylogentic inference showing the role of measures of support. See text for description. Adapted from [49].

search, measures of support may be calculated to identify character conflict. Conflicting characters (ad hoc explanations of homoplasy) are re-examined for error (in identification or coding, for example). This stage of character reanalysis is referred to as reciprocal illumination in reference to Hennig [7]. Where there is evidence of error, characters are re-coded. Measures of support are used to identify potential homoplasy, pointing out areas where additional data may be incorporated, or where character coding may need to be re-examined. The cycle continues, including all previous data as well as new data, in rounds of reciprocal illumination and hypothesis testing.

Unlike re-sampling and randomization approaches, direct measures can provide specific information about which characters conflict. In other words, re-sampling and randomization methods may provide a general sense that the hypothesis is unstable to perturbation, leading to a conclusion that more data may be required. Direct measures can identify the specific characters associated with the instability. It is this aspect that may allow direct measures of support to play a role in one aspect of phylogenetic inference. Direct measures provide the means to screen for potential falsifiers, the identification of which may lead to the identification of more specific targets (critical tests) for successive rounds of hypothesis testing. In other words, by identifying specific conflict, direct measures may be useful for identifying data that may constitute critical tests in the next round.

4. Discussion

What is the goal of calculating measures of support? What do these measures of support tell us? Support is commonly viewed in the literature as the stability of the hypothesis to perturbations in the data, parameters or analysis (i.e., relaxation of the parsimony criterion) or through comparison to randomized data. While a stable hypothesis is desirable for making inferences about evolutionary history and formal classifications (see [110–112] for discussions of the use of phylogenies for classification), stability in and of itself is not the goal of systematics [62,112] insofar as hypotheses should not be stable to the addition of disconfirming evidence [113, p. 142]. Within the cladistic parsimony framework, the operations of phylogenetic inference (and tree choice) and the calculation of measures of support are generally separate operations [49]. Calculations of measures of support, as commonly defined, do not influence tree choice, nor do they indicate the accuracy of the hypothesis. These measures are viewed as a secondary operation used to identify the amount and distribution of homoplasy on the tree. They are used as a means of heuristic data exploration that may serve to identify regions of the tree that are less well corroborated and perhaps more likely to be overturned by subsequent data.

Support within the cladistic parsimony framework is not considered the same as statistical support [19–21,48,49,63,79,86,112,114]. Rather, support for a given hypothesis of relationships refers to Popperian [18] corroboration [18,21,63,86,110,115]. In essence, the best supported hypothesis is the one that has been corroborated by virtue of having been subjected to and withstood the severest of tests. Tests consist of evidence (characters), and the severest tests make use of all available evidence. Given the differences in the definition of what support is, it follows that there are differences in the interpretation of what these measures of support mean and what, if any, actions should be taken based upon them. In a footnote, in his new Appendix ix, on corroboration, Popper states that the corroboration of the hypothesis (denoted x) by the evidence does not speak to the acceptability of the hypothesis unless the evidence "represents the (total) results of sincere attempts to refute x" [18, p. 402], parentheses in original. The total evidence paradigm of cladistic parsimony [42,62,112,116–120] can be viewed as stemming from the writings of Popper [18]. This viewpoint has implications for the use of a priori weighting procedures insofar as the downweighting or removal of conflicting characters could constitute the removal of the very characters that have the ability to refute the hypothesis. This exclusion of evidence could render the hypothesis untestable (by protecting the hypothesis from conflicting evidence) and therefore unable to be corroborated since testability (falsifiability) is central to corroboration [18–20,86,114,115]. Furthermore, the discarding of evidence would require justification. For example, some types of data are considered unreliable; the downweighting of transitions or third positions is justified by the assumption that these types of characters become saturated with substitutions (resulting in fewer observed changes than expected). This justification adds to the background knowledge that is assumed (but not tested). Bold hypotheses, ones that are logically

improbable given background knowledge alone, are seen as the preferred means to scientific advance [17,18,20]. Although Popper wrote of logical probability and derived formulae for the corroboration of the hypothesis given the evidence and the background knowledge, the formulae were used as a device to explain the method and not an advocacy statistical probability [85,86].

In terms of stability or robustness to data perturbation (i.e., commonly referred to as support) it may not be the overall amount of homoplasy that is important [76,79,88,99]. Instead, it may be the distribution of homoplasy that plays a critical role in tree stability [42,76,79,88,90,99]. Therefore, the preferred measures of support (assuming stability to data perturbation is considered of value) will be those that explore the varying effects and distribution of homoplasy on the tree. Direct measures of support such as BS and its derivatives (LBS, PBS, PHBS and HS) seem to provide this information. They allow for an exploration of support as well as conflict, and provide a means to quantify the amount of and specifically which characters contribute support/conflict. This ability to search for potential falsifiers suggests that direct measures of support have a role in the region of reciprocal illumination.

The desire in calculating measures of support is to have some sense of the reliability of the tree and by extension, the reliability of inferences made based on the tree. In cases where phylogenetic analyses result in a single most parsimonious tree some nodes may be better supported than others (based on some measure of support) however, regardless of the amount of support indicated by these metrics, the tree represents the best explanation of all the data and is therefore the best available hypothesis of relationships. Where multiple most parsimonious trees obtain from the analysis, each is an equally valid (albeit somewhat contradictory) explanation of the data. While a single tree is desirable for making classifications or other inferences (and it can be argued that, in the case of multiple trees, they cannot all be right) a summary tree (whether based on consensus, weighting, measures of support, etc.) is a less valid explanation of the data.

Measures of support may best be thought of as a tool for data exploration to identify those areas of the cladogram less well supported by the data and therefore possibly more likely to be overturned by the addition of subsequent data (additional tests). These metrics may serve best as stopping rules identifying less well supported areas that could be the focus of additional data collection [21,32]. Although most of the measures of support refer to perturbations to the data, and might lead to the conclusion that more data (i.e., additional gene regions, for example) may be needed, the importance of increased taxonomic sampling should not be underestimated [121,122].

In the final paragraph of Logic of Scientific discovery, Popper concluded: "Science never pursues the illusory aim of making its answers final, or even probable. Its advance is, rather, towards an infinite yet attainable aim: that of ever discovering new, deeper, and more general problems, and of subjecting our ever tentative answers to ever renewed and ever more rigorous tests" [18, p. 281]. To answer the question posed earlier, from a cladistic parsimony perspective, a well supported tree is neither the "true" tree nor the best tree. A well supported tree is one that has been corroborated by virtue of having been subjected to and withstood the severest of tests consisting of all available evidence. Measures of support do not speak to the accuracy of the tree nor do they provide information for tree choice, however, they may provide information about the degree of corroboration of the best tree and point to areas of interest for future testing.

Acknowledgments

The author is grateful for valuable discussions with Rob DeSalle and Al Phillips, and wishes to thank John Gatesy and an anonymous reviewer for critical reading of the manuscript and helpful suggestions for its improvement.

Appendix A. Computational limitations

A.1. Tree search

Has the best tree been found? The problem of finding the best tree for a given set of data is common to most methods of phylogenetic analysis as well as for the calculation of measures of support. For data sets with few taxa it is possible to find all trees and then choose the best (in the case of parsimony, the tree with the fewest changes, the shortest tree). However, for data sets with large numbers of taxa the enumeration of all possible trees is an NP-complete problem since there will be more possible trees than can ever be examined (see [123–125] for discussions of the effect of increased numbers of taxa on the numbers of trees). Therefore, heuristic methods are often used to explore tree space. Branch swapping algorithms, which pursue the first better (i.e., shorter) topology may only find locally optimal trees yet not find globally optimal trees. One approach that attempts to address this problem of becoming stranded on an island of suboptimal trees [126] is to begin with multiple different starting points (random addition replicates). Computational time also can be decreased by combining data sets to both increase signal and perhaps minimize the effect of homoplasy [127]. Further improvements to tree search include the parsimony ratchet [46] and strategies that use faster alternatives to TBR swapping such as tree drifting and fusing and sectorial searches [47]. Once a starting tree is generated, the ratchet weights a small, random portion of the characters and performs tree-bisection-and-reconnection (TBR) swapping on the starting tree (it has been suggested that limiting swapping on the weighted data may further improve speed [47]), holding a few of the resulting trees. Weights are reset to zero and swapping is performed on the trees held in the

previous step. The few trees that are retained from this round of TBR swapping are used as the starting trees for the next iteration. The ratchet increases the speed of the analyses (allowing numerous iterations to be performed) by sampling many tree islands but only retaining a few trees from each island, allowing for a more rigorous exploration of tree space. Concern for the rigor of tree search is important for evaluating measures of support since those measures that use heuristic searches (rather than exact enumeration of all possible trees) will result in measures that are only estimates of support [21].

A.2. Cumbersome tools for calculating measures of support

Phylogenetic analyses are performed using several different applications for data matrix assembly and editing, tree building, tree viewing and for the calculation of measures of support (see Table 1 for a list of web addresses for software information and download sites). In general, matrices are edited and manipulated using menu interface programs like MacClade [128] for the Mac platform or WinClada [129] for the pc platform. Matrices are then executed in tree search programs. Upon completion of tree search, users return to MacClade or WinClada to examine trees (WinClada has the ability to directly spawn pc format tree search programs).

Hennig86 [130] one of the earliest tree search algorithms was written for DOS; while still available, the functionalities are incorporated into later programs such as NONA [95] which runs under windows making use of windows memory management but is command driven. NONA performs fast tree searches by implementing indirect calculation of tree lengths [131], (see [132] for a review of tree search algorithms) as well as the parsimony ratchet [46]. PAUP* [92] allows tree search under different methods of inference (distance, likelihood and parsimony—for Bayesian analyses see MrBayes [15]) and was the first to be available with a pull-down menu interface for Mac OS 9. It is also available with a command line driven interface for the pc and for Mac OS X. While batch files can be written and executed in both NONA and the command line versions of PAUP*, the ease of use of the menu interface may account for the widespread popularity of PAUP* (despite limitations in terms of speed of tree search; since the menu driven version of PAUP* does not implement the parsimony ratchet [46] or other faster tree search algorithms [47,81,132]). The parsimony ratchet [46] can be implemented for pc's using Winclada [129], NONA [95] and TNT [96]; PAUPRat [133] can be used on the mac for generating the text file of ratchet commands for implementation using PAUP*. TNT (Tree analysis using New Technology) [47,96] is the first tree search package available for the pc that has a full menu interface. It is available for Windows (95/98/NT4/NT2000 and XP) as well as for Linux and Mac OS X that are command line driven versions. Unlike other available applications, TNT combines

Table 1
List of websites for phylogeny software and tools

Software download sites	
Tree search	
Hennig86	http://www.cladistics.org/education/hennig86.html
Nona	http://www.zmuc.dk/public/phylogeny/Nona-PeeWee/
PAUP*	http://paup.csit.fsu.edu/
TNT	http://www.zmuc.dk/public/phylogeny/TNT/
Matrix editing	
MacClade	http://macclade.org/
WinClada	http://www.cladistics.com/aboutWinc.htm
TNT	http://www.zmuc.dk/public/phylogeny/TNT/
Ratchet commands	
PAUPRat	http://www.ucalgary.ca/~dsikes/software2.htm
PRAP	http://www.botanik.uni-bonn.de/system/downloads/PRAP
Support calculations	
AutoDecay	http://www.bergianska.se/index_kontaktaoss_torsten.html
PAUP*	http://paup.csit.fsu.edu/
PRAP	http://www.botanik.uni-bonn.de/system/downloads/PRAP
RandomCladistics	http://research.amnh.org/~siddall/rc.html
TNT	http://www.zmuc.dk/public/phylogeny/TNT/
TreeRot	http://people.bu.edu/msoren/TreeRot.html
Tree viewing/manipulation	
MacClade	http://macclade.org/
WinClada	http://www.cladistics.com/aboutWinc.htm
TNT	http://www.zmuc.dk/public/phylogeny/TNT/
TreeView	http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

Additional resources (Sites with descriptions and links to software and phylogeny information) Felsenstein's Phylogeny Programs website http://evolution.genetics.washington.edu/phylip/software.html http://evolution.genetics.washington.edu/phylip/software.html#new Willi Hennig Society's links http://www.cladistics.org/education.html Cladistics.com's listings http://www.cladistics.com/ http://www.cladistics.com/Software.html Siddall's Methods in Systematics http://research.amnh.org/~siddall/methods/ Tree of Life's links http://tolweb.org/tree/home.pages/linksev.html.

matrix editing, fast tree search and tree viewing and manipulation in a single package. Ease of use is an important factor since users from many diverse fields of study (with perhaps limited computer programming skill levels) use tree analyses as part of their research. Surveying all MP trees along with apomorphy lists and their character fit statistics is perhaps the most direct method for evaluating support and conflict among trees; yet one drawback common to all current software packages is the lack of tools for easily viewing all MP trees.

Most of the applications mentioned above provide for the calculation of tree length and fit statistics as well as some measures of support such as bootstrap and character jackknife analyses. However, the calculation of additional measures of support would require the use of helper applications or manually written and implemented batch files of commands, and manual parsing of log files into spreadsheet applications for calculations of support indexes. For example, AutoDecay [93] can be used together with PAUP* to generate the constraint files needed for the calculation of BS. AutoDecay will also parse the resulting log file and output the results in tree format, with nodes labeled with BS values, which can be viewed using TreeView [134,135]. BS can be calculated and viewed on the tree using the pull-down menus in TNT but the calculation of PBS would require the user to write and execute a command file and manually parse the log file. PBS can be calculated in PAUP* using TreeRot [94] as the helper application. TreeRot generates the command files to be implemented in PAUP* and parses the resulting log file into a text file that can be imported into a spreadsheet for final calculations.

As with the original phylogenetic analysis for finding the shortest tree(s), rigorous tree searches are necessary for finding the shortest trees lacking a given node. Failure to find the shortest trees lacking a node of interest will result in an overestimate of BS and PBS for that node. Both AutoDecay and TreeRot by default only perform a few (10-20) random addition replicate heuristic searches for each constraint analysis, however the command files can be edited prior to analysis to implement more rigorous tree searches (yet neither application implements the parsimony ratchet [46]). In addition, both applications, by default, generate constraints from a single tree (the first tree in the tree file). Where more than a single MP tree has resulted from phylogenetic analyses, calculations of BS and PBS using AutoDecay or TreeRot would need to be repeated for each MP tree. Alternatively, a strict consensus of all MP trees could be used as the input tree. This would allow calculations of BS and PBS for all supported nodes (*ie*: those nodes with a BS > 0). TreeRot does examine the multiple constraint trees and automatically calculates the mean PBS value but does not output the range of values obtained. Furthermore, TreeRot can give misleading results when calculating PBS for matrices with large numbers of partitions. One solution would be to perform the calculations on batches of partitions ($\sim$10 at a time) however the command file would need to be edited in order for the comparison MP tree to be calculated from all the data (not just the data pertaining to the subset of partitions). PRAP (Parsimony Ratchet Analyses using PAUP*) [136], written in Java, allows for the calculation of BS values with tree searches that implement the parsimony ratchet. PRAP also outputs tree support values within tree files that can be opened and edited in TreeView or TreeGraph [137], with support values displayed at the nodes, as well as Nexus format output. Analyses and calculations of PHBS and HS are generally accomplished by batch command file, manual parsing of log files and pasting the results into a spreadsheet for calculations. However, a soon to be available helper application, ASAP (Automated Simultaneous Analyses Phylogenies) [138], written in Perl, implements the parsimony ratchet for tree searches and performs BS operations as well as PBS, PHBS and HS, parses the log files and outputs a spreadsheet format results file. In addition, ASAP automates many of the matrix manipulations that are currently performed manually, such as the assembly of the simultaneous analysis matrix from numerous data partitions, and the enumeration of character partitions. Execution of ASAP will assemble the matrix, enumerate the character partitions, execute PAUP*, perform rigorous phylogenetic analyses implementing the parsimony ratchet, save trees and tree scores, calculate measures of support (including BS, PBS, PHBS and HS) and output final results; without further manual intervention.

References

[1] Eldredge N, editor. Systematics, Ecology and the Biodiversity Crisis. New York: Columbia University Press; 1992.
[2] Novacek MJ, Wheeler QD, editors. Extinction and Phylogeny. New York: Columbia University Press; 1992.
[3] Harvey PH, Leigh Brown AJ, Maynard Smith J, Nee S, editors. New Uses for New Phylogenies. New York: Oxford University Press; 1996.
[4] Harvey PH, Pagel MD. The comparative method in evolutionary biology. New York: Oxford University Press; 1991.
[5] Givnish TJ, Sytsma KJ, editors. Molecular evolution and adaptive radiation. Cambridge: Cambridge University Press; 1997.
[6] Humphries CJ, Parenti LR. Cladistic biogeography: interpreting patterns of plant and animal distributions. 2nd ed. New York: Oxford University Press; 1999.
[7] Hennig W. Phylogenetic systematics. Urbana: University of Illinois Press; 1966.
[8] Kluge AG, Farris JS. Quantitative phyletics and the evolution of anurans. Syst Zool 1969;18(1):1–32.
[9] Kluge AG. What is the rationale for 'Ockham's Razor' (a.k.a. parsimony) in phylogentic inference? In: Albert VA, editor. Parsimony, phylogeny, and genomics. New York: Oxford University Press; 2005. p. 15–42.
[10] Felsenstein J. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst Zool 1973;22(3):240–9.
[11] Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 1981;17:368–76.
[12] Bayes T. An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F.R.S. Communicated by Mr. Price, in a Letter to John Canton, A.M.F.R.S. Philosophical Transactions (1683–1775) 1763;53:370–418.

[13] Yang Z, Rannala B. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Mol Biol Evol 1997;14:717–24.
[14] Larget B, Simon D. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol Biol Evol 1999;16:750–9.
[15] Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogeny. Bioinformatics 2001;17:754–5.
[16] Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 2001;294:2310–4.
[17] Popper KR. Realism and the aim of science. London: Routledge; 1992.
[18] Popper KR. The logic of scientific discovery. New York: Routledge; 1992.
[19] Farris JS. Conjectures and refutations. Cladistics 1995;11:105–18.
[20] Kluge AG. Testability and the refutation and corroboration of cladistic hypotheses. Cladistics 1997;13(1-2):81–96.
[21] Grant T, Kluge AG. Data exploration in phylogenetic inference: scientific, heuristic, or neither. Cladistics 2003;19(5):379–418.
[22] Farris JS. The retention index and the rescaled consistency index. Cladistics 1989;5:417–9.
[23] Archie JW. Homoplasy excess ratios: new indices for measuring levels of homoplasy in phylogenetic systematics and a critique of the consistency index. Systematic Zoology 1989;38(3):253–69.
[24] Farris JS. The retention index and homoplasy excess. Syst Zool 1989;38(4):406–7.
[25] Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 1985;39:783–91.
[26] Sanderson MJ. Confidence limits of phylogenies: the bootstrap revisited. Cladistics 1989;5:113–29.
[27] Felsenstein J. Phylogenies from molecular sequences: inference and reliability. Annu Rev Genet 1988;22(1):521–65.
[28] Efron B. Bootstrap methods: another look at the jackknife. Ann Stat 1979;7:1–26.
[29] Efron B, Gong g. A Leisurely look at the bootstrap, the jackknife, and cross-validation. Am Statist 1983;37:36–48.
[30] Efron B, Tibshirani R. An introduction to the bootstrap. London: Chapman & Hall; 1993.
[31] Efron B. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. Biometrika 1981;68(3):589–99.
[32] Siddall ME. Measures of support. In: DeSalle R, Giribet G, Wheeler WC, editors. Techniques in molecular systematics and evolution. Basel: Birkhäuser Verlag; 2002. p. 80–101.
[33] Mort ME, Soltis PS, Soltis DE, Mabry ML. Comparison of three methods for estimating internal support on phylogenetic trees. Syst Biol 2000;49(1):160–71.
[34] Sanderson MJ. Objections to bootstrapping phylogenies: a critique. Syst Biol 1995;44(3):299–320.
[35] Tukey JW. Bias and confidence in not quite large samples. Ann Math Stat 1958;29:614.
[36] Lanyon SM. Detecting internal inconsistencies in distance data. Syst Zool 1985;34(4):397–403.
[37] Siddall ME. Another monophyly index: revisiting the jackknife. Cladistics 1995;11:33–56.
[38] Penny D, Hendy M. Estimating the reliability of evolutionary trees. Mol Biol Evol 1986;3(5):403–17.
[39] Davis JI, Frohlich MW, Soreng RJ. Cladistic characters and cladogram stability. Syst Bot 1993;18(2):188–96.
[40] Farris JS, Albert VA, Källersjö M, Lipscomb D, Kluge AG. Parsimony jackknifing outperforms neighbor-joining. Cladistics 1996;12(2):99–124.
[41] Davis JI. Character removal as a means for assessing stability of clades. Cladistics 1993;9(2):201–10.
[42] Gatesy J, O'Grady P, Baker RH. Corroboration among data sets in simultaneous analysis: hidden support for phylogenetic relationships among higher level artiodactyl taxa. Cladistics 1999;15(3):271–313.
[43] Källersjö M, Farris JS, Chase MW, Bremer B, Fay MF, Humphries CJ, et al. Simultaneous parsimony jackknife analysis of 2538 *rbcL* DNA sequences reveals support for major clades of green plants, land plants, seed plants and flowering plants. Plant Syst Evol 1998;213:259–87.
[44] Lipscomb DL, Farris JS, Kallersjo M, Tehler A. Support, ribosomal sequences and the phylogeny of the eukaryotes. Cladistics 1998;14(4):303–38.
[45] Källersjö M, Albert VA, Farris JS. Homoplasy increases phylogenetic structure. Cladistics 1999;15(1):91–3.
[46] Nixon KC. The parsimony ratchet, a new method for rapid parsimony analysis. Cladistics 1999;15(4):407–14.
[47] Goloboff PA. Analyzing large data sets in reasonable times: solutions for composite optima. Cladistics 1999;15(4):415–28.
[48] Carpenter JM. Random cladistics. Cladistics 1992;8:147–53.
[49] Kluge AG. Sophisticated falsification and research cycles: consequences for differential character weighting in phylogenetic systematics. Zool Scr 1997(1998);26:349–360.
[50] Le Quesne WJ. Frequency distributions of lengths of possible networks from a data matrix. Cladistics 1989;5:395–407.
[51] Hillis DM. Discriminating between phylogenetic signal and random noise in DNA sequences. In: Miyamoto MM, Cracraft J, editors. Phylogenetic Analysis of DNA Sequences. Oxford: Oxford University Press; 1991. p. 278–94.
[52] Huelsenbeck JP. Tree-length distribution skewness: an indicator of phylogenetic information. Syst Zool 1991;40(3):257–70.
[53] Archie JW. A randomization test for phylogenetic information in systematic data. Syst Zool 1989;38(3):239–52.
[54] Faith DP, Cranston PS. Could a cladogram this short have arisen by chance alone?: On permutation tests for cladistic structure. Cladistics 1991;7:1–28.
[55] Sokal RR, Rohlf FJ. Biometry. New York: W.H. Freeman and Company; 1995.
[56] Lyons-Weiler J, Hoelzer GA, Tausch RJ. Relative apparent synapomorphy analysis (RASA) I: the statistical measurement of phylogenetic signal. Mol Biol Evol 1996;13(6):749–57.
[57] Lyons-Weiler J, Hoelzer G. Null model selection, compositional bias, character state bias, and the limits of phylogenetic information. Mol Biol Evol 1999;16(10):1400–5.
[58] Faith DP. Cladistic permutation tests for monophyly and nonmonophyly. Syst Zool 1991;40(3):366–75.
[59] Nelson G, Platnick N. Three-taxon statements: a more precise use of parsimony? Cladistics 1991;7:351–66.
[60] Nelson G, Platnick NI. Systematics and biogeography: cladistics and vicariance. New York: Columbia University Press; 1981.
[61] Lyons-Weiler J. RASA 2.3 for Macintosh and Manual. In: http://test1.bio.psu.edu/LW/rasatext.html; 1999.
[62] Kluge AG, Wolf AJ. Cladistics: what's in a word? Cladistics 1993;9:183–99.
[63] Siddall ME, Kluge AG. Probabilism and phylogenetic inference. Cladistics 1997;13(4):313–36.
[64] Nei M. Relative efficiencies of different tree-making methods for molecular data. In: Miyamoto MM, Cracraft J, editors. Phylogenetic analysis of DNA sequences. Oxford: Oxford University Press; 1991. p. 90–128.
[65] Hillis DM, Bull JJ. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst Biol 1993;42(2):182–92.
[66] Felsenstein J, Kishino H. Is there something wrong with the bootstrap on phylogenies? A reply to hillis and bull. Syst Biol 1993;42(2):193–200.
[67] Zharkikh A, Li W-H. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. Mol Biol Evol 1992;9(6):1119–47.
[68] Li W-H, Zharkikh A. Statistical tests of DNA phylogenies. Syst Biol 1995;44(1):49–63.
[69] Efron B, Halloran E, Holmes S. Bootstrap confidence levels for phylogenetic trees. Proc Natl Acad Sci USA 1996;93:13429–34.

[70] Grant T. Testing methods: the evaluation of discovery operations in evolutionary biology. Cladistics 2002;18:94–111.
[71] Zander RH. A conditional probability of reconstruction measure for internal cladogram branches. Syst Biol 2001;50(3):425–37.
[72] Zander RH. Minimal values for reliability of bootstrap and jackknife proportions, decay index and Bayesian posterior probability. PhyloInformatics 2004;2:1–13.
[73] Douady CJ, Delsuc F, Boucher Y, Doolittle WF, Douzery EJP. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol Biol Evol 2003;20(2):248–54.
[74] Simmons MP, Pickett KM, Miya M. How meaningful are Bayesian support values? Mol Biol Evol 2004;21(1):188–99.
[75] Kluge AG. Distinguishing ''or'' from ''and'' and the case for historical identification. Cladistics 2002;18(6):585–93.
[76] Goloboff PA. Homoplasy and the choice among cladograms. Cladistics 1991;7:215–32.
[77] Goloboff PA. Random data, homoplasy and information. Cladistics 1991;7:395–406.
[78] Källersjö M, Farris JS, Kluge AG, Bult C. Skewness and permutation. Cladistics 1992;8:275–87.
[79] Farris JS, Kallersjo M, Kluge AG, Bult C. Permutations. Cladistics 1994;10(1):65–76.
[80] Carpenter JC, Goloboff PA, Farris JS. PTP is meaningless, T-PTP is contradictory: a reply to trueman. Cladistics 1998;14:105–16.
[81] Goloboff PA. Estimating character weights during tree search. Cladistics 1993;9:83–91.
[82] Goloboff PA. Self-weighted optimization: tree searches and character state reconstructions under implied transformation costs. Cladistics 1997;13(3):225–45.
[83] Farris JS. Support weighting. Cladistics 2001;17(4):389–94.
[84] Farris JS. A successive approximations approach to character weighting. Syst Zool 1969;18(4):374–85.
[85] Popper KR. Objective knowledge an evolutionary approach. Revised ed. New York: Oxford University Press; 1979.
[86] Farris JS. The Logical Basis of Phylogenetic Analysis. In: Platnick NI, Funk VA, editors. Advances in cladistics, vol. 2: Proceedings of the second meeting of the Willi Hennig Society. New York: Columbia University Press; 1983. p. 7–36.
[87] Bremer K. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. Evolution 1988;42:795–803.
[88] Bremer K. Branch support and tree stability. Cladistics 1994;10(3):295–304.
[89] Donoghue MJ, Olmstead RG, Smith JF, Palmer JD. Phylogenetic relationships of dipsacales based on rbcL sequences. Ann MO Bot Gard 1992;79(2):333–45.
[90] Gatesy J, Arctander P. Hidden morphological support for the phylogenetic placement of *Pseudoryx nghetinhensis* with bovine bovids: a combined analysis of gross anatomical evidence and dna sequences from five genes. Syst Biol 2000;49(3):515–38.
[91] Lee MSY. Measuring support for phylogenies: the ''proportional support index''. Cladistics 1999;15(2):173–6.
[92] Swofford DL. PAUP* Phylogenetic Analysis Using Parsimony (*and other methods). Version 4. Sunderland, Massachusetts: Sinauer Associates; 2002.
[93] Eriksson T. AutoDecay Version 5.0. In. http://www.bergianska.se/personal/TorstenE: Stockholm, Bergius Foundation, Royal Swedish Academy of Sciences; 2001.
[94] Sorenson MD. TreeRot Version 2. Boston, MA: Boston University; 1999.
[95] Goloboff PA. NONA (No Name) Version 2. Tucumán, Argentina: Published by the author; 1999.
[96] Goloboff P, Farris S, Nixon K. TNT (Tree analysis using New Technology) Beta Version xxx. Tucumán, Argentina: Published by the authors; 2000.
[97] Lee MSY. Tree robustness and clade significance. Syst Biol 2000;49(4):829–36.
[98] Templeton A. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. Evolution 1983;37:221–44.
[99] Gatesy J. Linked branch support and tree stability. Syst Biol 2000;49(4):800–7.
[100] Faith DP, Ballard JWO. Length differences and topology-dependent tests: a response to Kallersjo et al. Cladistics 1994;10(1):57–64.
[101] Farris JS. Names and origins. Cladistics 1996;12(3):263–4.
[102] Baker RH, DeSalle R. Multiple sources of character information and the phylogeny of Hawaiian Drosophilids. Syst Biol 1997;46(4):654–73.
[103] Lambkin CL, Lee MSY, Winterton SL, Yeates DK. Partitioned Bremer support and multiple trees. Cladistics 2002;18(4):436–44.
[104] Baker RH, Yu X, DeSalle R. Assessing the relative contribution of molecular and morphological characters in simultaneous analysis trees. Mol Phylogenet Evol 1998;9(3):427–36.
[105] Baker RH, Wilkinson GS, DeSalle R. Phylogenetic utility of different types of molecular data used to infer evolutionary relationships among stalk-eyed flies (Diopsidae). Syst Biol 2001;50(1):87–105.
[106] Gatesy J, Amato G, Norell M, DeSalle R, Hayashi C. Combined support for wholesale taxic atavism in Gavialine Crocodylians. Syst Biol 2003;52(3):403–22.
[107] Lee M, Hugall A. Partitioned likelihood support and the evaluation of data set conflict. Syst Biol 2003;52:15–22.
[108] Gatesy J, Baker RH. Hidden likelihood support in genomic data: can forty-five wrongs make a right? Syst Biol 2005;54(3):483–92.
[109] Kluge AG. Philosophical conjectures and their refutation. Syst Biol 2001;50(3):322–30.
[110] Farris JS. The information content of the phylogenetic system. Syst Zool 1979;28(4):483–519.
[111] Schuh RT. Biological systematics: principles and applications. Ithaca: Cornell University Press; 2000.
[112] Kluge AG. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). Syst Zool 1989;38(1):7–25.
[113] Sober E. Reconstructing the past: parsimony, evolution, and inference. Cambridge, Massachusetts: MIT Press; 1988.
[114] Frost DR, Kluge AG. A consideration of epistemology in systematic biology, with special reference to species. Cladistics 1994;10(3):259–94.
[115] Kluge AG. The science of phylogenetic systematics: explanation, prediction, and test. Cladistics 1999;15(4):429–36.
[116] Nixon KC, Carpenter JM. On simultaneous analysis. Cladistics 1996;12:221–41.
[117] Mickevich MF, Farris JS. The implications of congruence in *Menidia*. Syst Zool 1981;30(3):351–70.
[118] Eernisse DJ, Kluge AG. Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology. Mol Biol Evol 1993;10(6):1170–95.
[119] Kluge AG. Total evidence or taxonomic congruence: cladistics or consensus classification. Cladistics 1998;14(2):151–8.
[120] Olmstead RG, Sweere JA. Combining data in phylogenetic systematics: an empirical approach using three molecular data sets in the Solanaceae. Syst Biol 1994;43(4):467–81.
[121] Wheeler WC. Extinction, sampling, and molecular phylogenetics. In: Novacek MJ, Wheeler QD, editors. Extinction and Phylogeny. New York: Columbia University Press; 1992. p. 205–15.
[122] Philippe H, Douzery E. The pitfalls of molecular phylogeny based of four species, as illustrated by the Cetacea/Artiodactyla relationships. J Mamm Evol 1994;2:133–52.
[123] Day WHE, Johnson DS, Sankoff D. The computational complexity of inferring rooted phylogenies by parsimony. Math Biosci 1986;81:33–42.
[124] Felsenstein J. The number of evolutionary trees. Syst Zool 1978;27(1):27–33.
[125] Felsenstein J. Inferring phylogenies. Sunderland, MA: Sinauer; 2004.

[126] Maddison DR. The discovery and importance of multiple islands of most-parsimonious trees. Syst Zool 1991;40(3):315–28.
[127] Soltis DE, Soltis PS, Mort ME, Chase MW, Savolainen V, Hoot SB, et al. Inferring complex phylogenies using parsimony: an empirical approach using three large DNA data sets for angiosperms. Syst Biol 1988;47:32–42.
[128] Maddison DR, Maddison WP. MacClade 4: analysis of phylogeny and character evolution. Sunderland, Massachusetts: Sinauer Associates; 2000.
[129] Nixon KC. WinClada Version 1.00.08. Ithaca, NY: Published by the author; 2002.
[130] Farris JS. Hennig86 software, version 1.5 and reference: Privately printed and distributed; 1988.
[131] Goloboff PA. Character optimization and calculation of tree lengths. Cladistics 1993;9(4):433–6.
[132] Goloboff PA. Methods for faster parsimony analysis. Cladistics 1996;12(3):199–220.
[133] Sikes DS, Lewis PO. PAUPRat: PAUP* implementation of the parsimony ratchet. Program distributed by the authors. Storrs: Department of Ecology and Evolutionary Biology, University of Connecticut; 2001.
[134] Page RDM. TreeView: an application to display phylogenetic trees on personal computers. Comput Appl Biosci 1996;12:357–8.
[135] Page RDM. TreeView Version 1.6.6. Glasgow: University of Glasgow; 2001.
[136] Müller K. PRAP–computation of Bremer support for large data sets. Mol Phylogenet Evol 2004;31:780–2.
[137] Müller J, Müller K. TreeGraph: generating complex postscript trees using an extensible tree description format, program distributed by the authors. Bonn: Botanical Institute, University of Bonn; 2003.
[138] Sarkar IN, Egan MG, de la Torre JE, DeSalle R, Coruzzi G. ASAP: Automated Simultaneous Analyses Phylogenies; manuscript in preparation.