Title: A sensitivity analysis for choosing reliability functions to select cladograms

Joan Salvador Arias Becerra

Director: Daniel Rafael Miranda Esquivel

Escuela de biología
Universidad Industrial de Santander
Bucaramanga, Colombia
Enero, 2006

Sometido a: *Zoologica scripta*

Supplementary information:
<http://ciencias.uis.edu.co/labsist/salvador/relfun.htm>

Resumen en español

TITULO: ANÁLISIS DE SENSIBILIDAD PARA ESCOGER FUNCIONES DE CONFIANZA USADAS PARA SELECCIONAR CLADOGRAMAS[*]

Joan Salvador Arias Becerra[**]

Palabras clave: análisis de sensibilidad, cladística, parsimonia, pesos implícitos, sistemática

Diferentes métodos para seleccionar las funciones usadas en cladística (parsimonia lineal y pesos implícitos, que usa una función convexa de la homoplasia) son propuestos y explorados dentro del marco de un análisis de sensibilidad. Los procedimientos propuestos son eliminación de caracteres, terminales o combinado, búsquedas rápidas y partición de matrices.
Aunque los resultados parecen ser dependientes de la matriz y el procedimiento usado comportamientos generales pueden extraerse del análisis. Se encontró que la mejor medida para la selección es el número absoluto de nodos recuperados, puesto que es independiente de la resolución de los cladogramas usados. Medidas escaladas de la topología dependen de la resolución de los cladogramas, y tienden a escoger los resultados menos resueltos. Las medidas basadas en ajuste/distorsión están sesgadas hacia las funciones más fuertes de pesos implícitos, además de que no puede usarse en ciertas circunstancias. Así mismo se encontró que un número pequeño de replicas (20) produce resultados indistinguibles de los encontrados tras un gran número de replicas. Las funciones de pesos implícitos, en general, son mejores que parsimonia lineal en términos de la estabilidad de los resultados. Se rechazo el uso de partición de matrices debido a que sus resultados no difieren de los otros procedimientos y consume el doble

[*] Tesis de grado

[**] Facultad de Ciencias, Biología, dirigida por Daniel Rafael Miranda Esquivel

del tiempo.

Dados los resultados se propone que un análisis de sensibilidad de este tipo debe constar de dos etapas: una primera etapa donde se realiza una exploración amplia de procedimientos y funciones usando pocas replicas, y una segunda etapa donde se explora intensamente (muchas replicas) las funciones y procedimientos que produjeron los mejores resultados en la primera etapa con el fin de hacer una selección óptima.

Abstract

Arias, J. S. & Miranda-Esquivel, D. R. (2006) A sensitivity analysis for choosing reliability functions to select cladograms. *Zoologica scripta, 00*, 000-000.
Different methods to select among reliability functions used in cladistics (linear parsimony and implied weights) are proposed and explored under a sensitivity analysis framework. The proposed methods are character, taxon and mixed jackknife, and fast searches. Although results seems to be matrix-procedure dependent some general behaviors could be extract from the analysis. We found that the absolute number of recover nodes is the best measure for selection, as they are independent of resolution of compared cladograms. Scaled topological measures strongly depend on resolution of compared cladograms. Fit/distortion measures are biased towards the strongest functions of implied weights. Also a small number of replications (as 20) produces nearly the same results than a huge number of replicates. Implied weights outperform linear parsimony in terms of stability of results. Given those results, perspectives and analysis guidelines are proposed.
Keywords: cladistics, implied weights, parsimony, sensitivity analysis, systematics.

J. Salvador Arias, Daniel Rafael Miranda-Esquivel, Universidad Industrial de Santander, Escuela de biología, A.A. 678 Bucaramanga, Colombia. E-mail: dmiranda@uis.edu.co

Introduction

Just after Hennig (1965, 1968) showed the logic of cladistic analysis several numerical implementations were developed to assign the optimal set of character states of each node (Wagner 1961; Camin & Sokal 1965; Dayhoff 1969; Kluge & Farris 1969; Farris 1970; Fitch 1971). The number of implied homoplasic transformations (extra-steps) was calculated using the assigned states. With a function of the character homoplasy we have a measure of how well (*fit*) or how bad (*distortion*) the character is adjusted to the examined tree (nomenclature from Goloboff 1993a, 1997a; Goloboff *et al*. 2004). Each character distortion or fit is summed and provides a general score for the tree. This score is then used to discriminate among cladograms. So, these homoplasy functions are “optimality criteria” (*sensu* Swofford *et al*. 1996).

These homoplasy functions could be infinite. But the implemented ones and by far the most used, could be generalized as the distortion of *i*-character calculated as

$$d(i) = h(i) / (a * h(i) + k) \quad (1)$$

where $h(i)$ is the homoplasy of *i*-character, k is the Goloboff’s (1993a, 1995) constant of concavity, and a is a constant that modifies homoplasy. When $a = 0$, $d(i)$ is equal to homoplasy, that is traditional parsimony, hereafter referred as *linear parsimony*. With $a = 1$, the distortion is the inverse of the fit of the Goloboff’s (1993a, 1995) concave function (see Goloboff 1997a, 1997b), hereafter referred as *convex parsimony*. We use the term *reliability function* is used to refer to any homoplasy function derived from equation (1).

Not all of the homoplasy functions are used as optimality criteria or count directly the extra steps. Some functions coupled with an optimality criterion can be used to measure character fit, and then a new analysis is performed using each character fit as the weight of that character (e.g. Farris 1969, 2001; Carpenter 1988; Kjer *et al*. 2001, 2002). Here we focused on functions used as optimality criteria.

Different functions and weighting schemes might change the results of an analysis (e.g. Farris 1969; Platnick *et al*. 1991; Goloboff 1993a, 1995, 1997a; Turner &

Zandee 1995; Wheeler 1995; Kluge 1997; Ramírez 2003) and choosing a function based on its pragmatic qualities is not feasible. Each function has been developed to minimize homoplasy (Farris 1983; Goloboff 1993a, 1995; Kluge 1997). So, it is necessary to have an external criterion in order to choose among functions (Wheeler 1995; Giribet & Wheeler 1999; Giribet *et al*. 2002; implied by Goloboff 1997a). Accuracy can not be used as an external criterion because "The Real" phylogeny remains unknown. Therefore the *stability* of results is a direct criterion to make a function selection (Wheeler 1995; Goloboff 1997a; Ramírez 1999, 2003; Giribet & Wheeler 1999; Farris 2001; Giribet *et al*. 2002; *precision* in Wheeler's terminology). Here we mean with stability the property of how the *actual results* remain equal under the same function and data when conditions of search change, for example, using a slightly different matrix. Stability could be measured in two ways, using topologies (Nelson 1979; Wheeler 1995; Goloboff 1997a; Ramírez 1999, 2003; Goloboff & Farris 2001; Lopardo 2005) or character fit (e.g. Kluge 1989; Farris *et al*. 1994; Wheeler 1995; Allard & Carpenter 1996; Allard *et al*. 1999; Wheeler *et al*. 2001; Aagesen *et al*. 2005).

The first comparison among different functions with real data sets was Goloboff's (1997a) work. He used character jackknife resampling and node count to compare stability between results from linear, convex parsimony, and Farris' (1969) successive weighting. Ramírez (1999, 2003) proposed to use character jackknife coupled with scaled node count as a way to choose among homoplasy functions. Ramírez (2003) argued that node count could be increased simply by over resolution, and proposed scaled measures. He also proposed taxon jackknife (Ramírez 1999) but suggested to use it with caution because he thought that optimal homoplasy functions are matrix dependent.

We present several procedures that, using stability of results, permit the selection of a function of reliability. We also explore their performance using topology and fit measures of stability under different conditions of analysis. To accomplish this, we use several real data sets.

Materials and methods

The procedures used to select among functions and weighting schemes were based on the concept of stability of partial searches. A partial search is a shortcut to find near optimum trees with little computational effort (Farris *et al*. 1996; Farris 1997; Goloboff & Farris 2001). Partial searches might be coupled with resampling of the matrix (character or taxon removal) followed by a fast search (Farris *et al*. 1996; Farris 1997), or performing directly a fast search without resamping the matrix (Goloboff & Farris 2001). If partial searches are done several times, the stability of results can be used to discriminate among functions (Goloboff 1997a; Ramírez 1999, 2003; Farris 2001). The stability is calculated with respect to a reference cladogram or a taxonomic scheme (e.g. Wheeler 1995; Goloboff 1997a; Ramírez 1999, 2003; Lopardo 2005), or with respect to the results from the same partial searches (as in consensus of Farris *et al*. 1996; Goloboff & Farris 2001). A flow chart with the basic steps of procedures is illustrated in figure 1.
Also, we explored the performance of the different procedures. To facilitate visualization and quick comparison of results, in some cases we scaled the measures with respect to linear parsimony. We used linear parsimony only as a reference point, without giving it more or less importance than any of the other functions. The use only reflects that it is, by far, the most used function. To show the deviation of results from a selected reference result, we use the χ^2 statistic. Our purpose is only descriptive (to show the dispersion) and we did not embrace any particular statistical property to the data.

Data matrices

We used several real data matrices extracted from literature (Appendix 1). An artificial data set has the advantage to know whether a given method recovers “The True” tree or not. But we saw the problem in other way. With real data, we could address the typical caveats of a current phylogenetic analysis. Our data

selection was simply intended to capture different data types and matrix sizes.

Measuring stability

Reference cladograms. Although a reference cladogram has been used in stability analyses, this did not necessarily imply that the reference cladogram was “the real one” (*contra* Grant 2002; Rydin & Källersjö 2002; Grant & Kluge 2003). The only assumption required is that the reference cladogram represents an optimal solution to the problem at hand.

Here we use as the *reference cladogram* the result of an explicit search for the optimal cladograms (even if the search is not intensive). A cladogram from a partial search (with resampling or fast searches see below) is called the *approximate* cladogram (fig. 2).

Even accepting that a reference cladogram is not a claim for the “real” phylogeny, it is possible to think that its use is not wise. For example, using convex functions with high precision calculation could over-estimate the difference among trees (Goloboff 1993a, 1995; Turner & Zandee 1995; Kluge 1997; Goloboff *et al*. 2004), therefore the use of the optimal tree could be problematic. To avoid this problem we also compared each cladogram from a partial search with a cladogram extracted from another replicate selected at random from the same partial search. In those cases both compared trees are the result from a partial search. We refer to these as analyses *without a reference cladogram* (fig. 2).

Topology measures. One of the simplest ways to compare cladograms is to perform a raw count of the number of nodes shared by the two cladograms (Nelson 1979; Goloboff 1997a; Ramírez 2003). Here we refer to this count as the “number of common nodes” or NC. Different functions usually produce different degrees of resolution (Carpenter 1988; Goloboff 1993a, 1997a; Ramírez 2003; Goloboff *et al*. 2004), then the result could be biased toward functions that produce the most resolved consensus (Ramírez 2003). To avoid bias we scaled the number of common nodes, by the number of nodes on the reference tree (Nelson 1979;

Ramírez 2003). That is the “index of common nodes” or IC (PC of Ramirez 2003). However, the approximate tree, although resolved (with a high IC or NC), could have many ‘spurious’ groups (i.e. groups absent in the target tree). So we also used the number of nodes in the approximate tree to scale results, this is the inverse of Ramírez’s (2003) PE value, hereafter IE. A third measure, the “index of ‘similarity’ among trees” or IA, is IC * IE. The three scaled measures varied from 1.0, the best, to 0.0 the worst. In the analyses without a reference cladogram, as both trees compared are from a partial search, the label (and values) of IC and IE were freely interchangeable.

Fit measures. The comparison among cladograms can be viewed as a fit adjustment problem (Farris *et al*. 1994; Wheeler 1995; Allard & Carpenter 1996). These measures were developed in the context of data partitions, and here we used several modifications of the Mickevich-Farris’ index (see Kluge 1989; Farris *et al*. 1994). As each modification was procedure specific, we describe these in the procedure section (see below).

Procedure description

The sets of trees used to measure the stability of the results of each data set were produced using six modifications of three general methods.

Jackknife methods. In cladistics the jackknife procedure follows Farris *et al*. (1996). At each replicate, data arrays were deleted from the original matrix forming a resampled matrix. The *approximate* cladogram for stability measures was drawn from a fast search on the resampled matrix. We used three forms of resampling. In *character jackknife* (**cjac**; Farris *et al*. 1996; Goloboff 1997a; Ramírez 2003) only characters were susceptible of deleting. In *taxon jackknife* (**tjac**; Ramírez 1999), only taxa were eliminated. In *mixed jackknife* (**mjac**), both taxa and characters could be deleted.

For taxon jackknife, we used the compatibility rules of Goloboff & Pol (2002) to calculate the number of shared nodes. For fit measures, we used

$Q = (D_a – D_r) / D_r$

where D_a is the sum of all character distortion from the approximate cladogram, and D_r is the sum of all character distortion from the reference cladogram. Distortion of the approximate cladogram was measured using all characters (for character and mixed jackknife), but not all taxa in the case of taxon and mixed jackknife (i.e. ignoring excluded terminals). In analyses without a reference cladogram, the distortion of each compared tree was used to scale the Q-value, so two values were scored, and they were freely interchangeable.

Partition methods. Random partitions were used to measure incongruence between data sets (Farris *et al*. 1994; Allard & Carpenter 1996). At each replicate the complete matrix was divided into two equally probable partitions (**p2**) that were analyzed separately.

When comparing topologies using a reference cladogram, a common node was a node found in the target cladogram and in the two strict consensus trees of each partition. If no reference cladogram was used instead of selecting other cladogram at random, the comparison was between the consensus of each partition of the same replicate.

For fit measures, the Mickevich-Farris index was used in the most familiar way. When a reference tree was used the measure was

$Q = (D_r – (D_1 + D_2)) / D_r$

where D_1 and D_2 were the distortions of each partition. In analyses without a reference cladogram, we calculate Q as in jackknife procedures, except that distortions were calculated to the partitioned data and both partitions were from the same replicate.

Fast search methods. The result from a fast search uses a consensus of the trees found with few replicates of a tree-search, keeping one tree from each replicate, irrespectively of their optimality. Fast searches provide rough approximations of the consensus of the optimal cladograms (Goloboff & Farris 2001; implied in Farris *et al*. 1996). The fast searches used here were based on 20 replicates of Wagner-Dayoff cladograms (i.e. RAS + SWAP): The Wagner procedure (Wagner 1961;

Kluge & Farris 1969; Farris 1970), improved by a random addition sequence and branch swapping, as in Dayoff (1969). A single tree collapsed with rule 1 (Swofford & Begle 1993) was retained in each replication, so the consensus is based on 20 trees.

To make comparisons, we directly used the *strict consensus* (**gf-s**) or a *majority rule consensus* with a 75% cutoff value (**gf-75**), stopping in the first consensus (step 3) of Goloboff and Farris (2001) method. The topology measures were equal to those used in jackknife methods and using Wagner-Dayoff trees for fit measures.

Availability. The described procedures, as TNT (Goloboff *et al*. 2004) macros, and the programs that perform the measurements could be downloaded from the LSB web page <http://ciencias.uis.edu.co/labsist/software/>.

Searches

Our search strategies were developed to minimize the time used in the partial searches. If we want to use this kind of analysis to select among functions, it would be preferable to do it before performing a complete analysis so that they would represent a small fraction of the research effort.

Reliability functions. We used linear parsimony ($a = 0$; $k = 1$; in equation 1), and convex parsimony using a k-value of 10 as the upper limit ($a = 1$; $0 < k <= 10$).

Approximate cladograms. Independently of the procedure used for the approximate cladogram we used 20 replicates of Wagner-Dayoff trees (TBR swap), keeping only one tree per replication. The strict consensus of the 20 trees found was the approximate cladogram. A majority rule consensus of 75% was used with **gf-75** procedure. For fit measures, the best tree among the 20 found was used to calculate the distortion.

Search for reference cladogram. To measure the effect of search depth we explored three different sources for reference cladograms. Intensive searches (10 replicates of ratchet with 200 iterations), typical searches (100 and 500 Wagner-

Dayoff searches, with final branch swapping), and fast searches (200 replicates of ratchet with 10 iterations, and a strict consensus from 20 Wagner-Dayoff trees). We only keep one tree in each Wagner-Dayoff replicate. We used the strict consensus among the best trees found (except for 20 Wagner-Dayoff trees, were all 20 trees were used independently of their distortion) to evaluate the topological measures, and the distortion of any of the best trees for fit measures. Measures derived from the comparison with results from 10 ratchet replicates with 200 iterations were used as reference ("expected" value for χ^2). The comparisons of search depth were only performed on 100 replicate procedures. Otherwise, the target cladogram used was derived from a typical search of 100 Wagner-Dayoff (with TBR swap) independent cladograms.

Number of procedure replicates. For the matrices DRO, FON, GIG, GUI, SCH, and TAB, on all procedures, we modify the *number of partial search replicates*, using 20, 100, 200, and 1000. Ideally, the function selection phase would be done quickly, and then few replications were preferable over a great number. Results from 1000 replicates were used as reference ("expected" value for χ^2). In other cases, only 100 replicates were used.

In analyses without a reference cladogram IC and IE, and the two MF indexes, are interchangeable. We used this property to measure the overall difference among cladograms of a resampling cycle, and then estimate the number of replications needed for a data set. When the differences were in average, less than 0.05 for IC and IE, and less than 0.01 for MF we considered the results to be stable.

Effects of support. To evaluate the effects on clade support we changed the strictness of fast searches (as **gf-s** and **gf-75**). Although fast methods are not suitable to measure support, strictness of cut value in consensus removed possible 'spurious' groups (Goloboff and Farris 2001). A more specific way to measure the effects of support is to change the cut value in jackknife procedures. We used 0.05, 0.15, 0.25 (Goloboff 1997a), 0.36 (Farris *et al*. 1996; Ramírez 2003), and 0.45 for matrices DRO, FON, GIG, GUI, SCH, and TAB. The comparisons of the deletion probability were only performed on 100 replicate procedures. Otherwise, the

probability used was 0.36.

Fast results. Given our results, we performed in all matrices, including Zilla (Rice *et al*. 1997) a fast approximation (20 replicates) with jackknife and our modification of Goloboff-Farris procedure. As the reference tree, we used the result from Goloboff-Farris method as modified here (see above).

Results

In our different runs, the most marked behavior is that, according to the procedure and measurement used, results could change for the same matrix (table 1). Linear parsimony nearly always produces the least resolved trees and found fewer nodes than convex parsimony (as in Goloboff 1997a). We remark that we only show the *general* results. Values of measures across functions could be found at <http://ciencias.uis.edu.co/labsist/salvador/relfun.htm>.

Different measures

When NC is used, the most preferred functions are the milder ones (*k*=9-10, 34% of times; fig. 3, table 1), other functions (from 1-8) are selected about the 8-10% of times. Linear parsimony is not selected using the direct node count. For IC the preferred function is linear parsimony (54%); it agrees with the fact that linear parsimony produces the least resolved trees. IE values are not quite informative, as their values are nearly equal between different functions. If the value of IE for linear parsimony is taken as a scaled reference, the values for the other functions are, in average, 98.48% (*s*=8.35%) of the linear parsimony value. As a consequence of this, IA is only a scaled form of IC (R^2=0.9124).

For the MF index, the most preferred function is the strongest function (*k*=1, 66%). It seems that the MF value increases, as functions become weaker. In any way, MF values are very small for **cjac**, and **gf-s** procedures, with Q-values usually below 10^{-3}, as there are (relatively) small or no differences in fit between the

procedure trees. Those small values made the function discrimination very difficult because calculations needed to be based on highly precision values. When **cjac** and **gf-s** procedures are not taken into account, the preference for convex parsimony with *k*=1 rises to 86% of the times. Because of this bias, we reject MF as a measure for function selection.

Reference tree

Scaled measures (IC) are particularly sensible when the reference tree is very conservative (i.e. unresolved), specifically for reference trees from a fast search as in Goloboff & Farris (2001) (Table 2). In contrast with IC the direct node count (NC) remains constant irrespectively of the deepness of the reference tree search (Table 2). So most resolved results do not increase the number of shared nodes and the same results could be obtained using fast searches saving time during the selection phase. Using a reference tree from 100 Wagner-Dayoff trees, the selected function using IC is the same on analyses without a reference tree in 41.17% of cases (table 1). But when functions are distinct, (and ignoring cases when one of the functions is linear parsimony), the difference of k-values between functions is 4.80 (s=2.73). The number of recovered nodes (NC) is proportionally similar with and without a reference tree (R^2=0.898). Excluding **p2** procedure that counts slightly different the number of nodes with and without a reference tree, the correlation is higher (R^2=0.9720). The same function is selected in 41.49% of cases (table 1), and the difference of k-values between functions is 2.44 (s=1.85). From these results we conclude that IC is biased by cladogram resolution.

Number of replicates

The same results were obtained regardless of the number of replicates used (table 3), and χ^2 shows that dispersion is small. In other context (a fast estimation of the consensus tree) Goloboff & Farris (2001) found equivalent results. Without the

reference tree, values of IC-IE could be interchanged within our limit of 5% of difference as few as 20 replicates (table 4). The same holds for MF measure (table 4). So it seems that IC-IE and MF are valuable to found the dispersion of results. Form these results we conclude that a broad range of functions explored using fast searches is better than a narrow set of functions explored with a deep search.

Effect of support

The same function is selected with NC in the 58.82% of cases (34) when using of the 75% majority rule consensus in each replicate (**gf-75**) instead of the strict consensus (**gf-s**) for each replicate (appendix 2). When functions are distinct the difference in *k*-value is small (NC, 1.38 *s*=1.86). It seems that consensus strength does not affect (other than the node values) the selected function.
Results from character jackknife are rather constant than the other two jackknife methods (taxon and mixed) with respect to the “0% jackknife” (i.e. gf procedure). Nevertheless the selected function remains nearly constant for different cut values under all jackknife procedures for the matrices explored by us (appendix 3). Among the five cut values, the average of standard deviation of *k*-values is 1.15 for NC (35 cases, *s*=1.04).

Fast searches

As in the case of the number of replicates and the different kinds of reference trees, using fast searches combined with fast reference trees (if used) produce the same results, in terms of NC, than a “typical” number of replicates (here 100) and the “typical” reference tree (table 5). In terms of selected function, in average, the function value is slightly different, but usually the difference of the *k*-values is of one unit. In some cases the value has 5 units of difference (compare tables 1 and 5). The difference is independent of matrix size.
It is important to remark that fast search on Zilla using **cjac** without a reference tree

is the only case where NC selects linear parsimony slightly outperforming results from *k*-value of 10 (NC value for linear parsimony=174.7; NC value for *k*-10=173.8). It seems that fast searches found an adequate neighborhood for function selection. With small and middle size matrices, a more stringent approach could follow the fast search, as computer time is not a problem. For huge matrices (like Zilla), this approach is the only way to avoid time consuming procedures.

Discussion

What measure would we use?

Scaled node measures are dependent on the resolution of the cladogram used to scale the value. As cladogram's resolution decreases, the value of the scaled measure increases. However, direct node count remains constant regardless the use or not of a reference cladogram, and the goodness of the reference cladogram used, and then the measure is independent of the resolution (Table 2). Fit measures depend on the function. Functions with low *k*-value produce better results than milder ones (Table 1). Those results allow us to claim that raw node count could be the best option when topology measures are taken into account. Ramírez (2003) argued than it is possible that the number of nodes might bias direct measurements. But this bias is not removed by scaling the direct node counts. Partial searches usually found fewer nodes than the complete search. As a result, the proportional values for each node found is greater in the least resolved solutions. The same applies to the IE. Again, the scaled values are high (see also Lopardo, 2005). In the cases where the resolution is nearly equal between linear parsimony and convex parsimony, IC and NC show the same pattern, but in those cases arguing for a scaled measure is not necessary. Scaled measures came from an intuitive idea that an increase in the number of nodes might increase the *possibility* of finding more common node between two cladograms irrespectively of their stability (Ramírez 2003: 25). But NC selects the function that maximizes the

number of stable nodes. A well-resolved consensus (i.e fewer possible solutions) could be highly unstable, and then few nodes were shared between partial and complete search (or within partial cladograms). NC selects the function that found more stable groups, and *consequently* a resolved final solution given the data at hand. With many stable groups, resolution is a consequence of the number of supported groups, not its cause.

It is usually recalled that topological measures do not take into account the group support (e.g. Farris *et al*. 1994; Allard & Carpenter 1996) at least for partitions. But this does not seem to be the case. The procedure used here to count the amount of nodes is the same used for jackknife procedures, a topological method to measure support (see Farris *et al*. 1996; Farris 1997; Goloboff *et al*. 2003; Ramírez 2005).

Ramírez (2003) thought that fit measures, if implemented in some way, could produce a better approximation to the problem of function selection, as it is used to select among cladograms. Here our implementations are dependant on the function used (for taxon and mixed jackknife, character jackknife with high cut values), the stronger the function the lesser the difference in fit between trees (a similar result is found in a context of partitions by Aagesen *et al*. 2005). Tree topologies, no matter how different are, have a similar fit for strong functions, like convex parsimony with *k*=1. In the other case (for character jackknife and Goloboff-Farris trees), our fit measures do not discriminate among functions. As our results show, in small matrices nearly all search methods found the most parsimonious solution or only slightly less-than-parsimonious one, so the differences (if any) are small. For huge matrices, the difference of fit could be more marked. But as the results are scaled with respect of the distortion of one of the trees and as a consequence of the size of the matrix, the total distortion is also enormous. Then the difference, even if it is a large number, is proportionally very small. Maybe a better approximation could be the RILD measure (Wheeler & Hayashi 1998). Aagesen *et al*. (2005) found in partitions that this measure is more sensible to extreme weighting, so it is possible that the bias found here to strong convex

functions would be present in RILD. Even under this measure it is highly probable that values that produce **cjac** and **gf-s** would be small. And the problem with very small values (with a great number of significant decimals) is that they require a computationally demanding search, that we think to be preferable for the final answer search rather than to data exploration.

Anyway, fit measures have some particular problems. Approaches based on fit measures exclude, necessarily, any weighting scheme. Distortion depends on the set of weighted characters and it is not directly comparable among trees derived from different weight schema (Goloboff 1993a; Swofford *et al*. 1996). Also, the fit measures for methods that use a form of consensus tree as the reference tree are forbidden, as distortion of a consensus tree is meaningless (Nixon & Carpenter 1996). Examples of such methods are jackknife (Farris *et al*. 1996) and double consensus (Goloboff & Farris 2001) that could be used as an approximate answer for very large matrices.

So, what is the best procedure?

This is a difficult and tricky question. We do not intend to provide a direct answer, rather we think that is the work with different procedures that provides a particular answer for each different data set (see for example Wheeler 1995; Rydin & Källersjö 2002; Aagesen *et al*. 2005).

We found that the values chosen by **gf-s** are somewhat the same for **cjac**, moreover although (as expected) jackknife values change with the cut-strength, the chosen function remains equal (appendix 3). The same holds for **tjac** and **mjac**. So the problem of randomization, that the results could change with the strength of resampling (Ramírez, 2005; implied in Goloboff et al., 2003; Miller, 2004), does not affect the function selection. The other procedure implemented by us, a partitioned analysis, appears to be very similar to a strong character jackknife (cuts above 45%). In this case, jackknife procedure appears to be a better option based only on a pragmatic criterion. Partitioned analysis requires two searches, one for each

partition, in each replicate. On the other hand jackknife only requires one search in each replication, making it faster.

Ramírez (1999) criticizes taxon jackknife as a mean to select reliability functions. He argues that the function depends on the taxa number. If matrices used to select among functions are different (in terms of number of terminals) from the complete matrix, it is not clear why to use the chosen function in the whole matrix. The procedures presented here are intended to provide stable groupings with resolved cladograms. The amount of stable groups is somewhat measured by perturbing the data matrix several times. Different functions could provide stable results to different forms of perturbation. Taxon and mixed jackknife measures how the function performs with different taxa sets. The chosen function in this case is the one that produces somewhat similar results independently of the data set used (as in Rydin and Källersjö, 2003). They are not really “different matrices”, because the overall data is based on the same research, several terminals and characters are equal, and the groups found are the same groups in both matrices. Matrices are different when different data (character and terminals) are used. For example, using some function in a molecular data set, does not imply that the use of this function on data sets for the same taxa but with different character sources, say a morphological matrix, another molecule data set, or one of total evidence. They are clearly different matrices.

We remark that the different procedures found different aspects of result’s stability (e.g. Goloboff et al., 2003; Ramírez, 2005). Maybe the best approach is a multi-procedure approximation. As the number of replications and the quality of the reference tree do not affect the results, several functions and procedures could be used in an initial exploration without severe time load. This is not a plea for “anything goes” perspective. Rather the global exploration could serve as an easy way to choose promissory functions, and then a more stringent search could be limited to these functions, for example using better reference trees, more replications, more detailed functions, and a careful examination of the groups found. But if the selected function is matrix- and procedure-dependant someone

could ask,

What happens if selected functions differ?

Although we found different results in different procedures, a particular trend was found. Using convex parsimony is nearly always better than using linear parsimony. Results were more stable and more resolved. It is possible than in some matrices linear parsimony behaves better, but in this cases it is only slightly better than convex parsimony (e.g. Zilla). So although even if the 'exact' function could not be found, at least we could learn what kind of functions *do not produce* stable results.

We advocate that only one of the functions would be chosen. A matrix may be stable (or unstable) to different alterations, like adding characters, adding taxa or recoding characters. A researcher should examine whether the results are stable to the alterations that the matrix is likely to suffer in the future. For example, in a matrix containing all species of a very well sampled genus, new taxa are unlikely to be added, and thus is more meaningful to examine stability to character addition (through character deletion), instead of stability to taxon addition. Otherwise the source of instability would came, probably, from taxa (as showed by Rydin and Källersjö, 2003) and character sampling, as for example the phylogeny of a kingdom using species as terminals as is done in molecular analyses.

So final selection depends on the desired kind of stability. This is not an arbitrary choice, because whatever the procedure selected (and then the chosen function), the selection is based on the different interactions of evidence. And results could be contrasted with the ones found with other procedures. Here, researcher selection of a procedure falls in the same category of selection of taxa and characters. All those decisions are though decisions, but need to be done, and if done carefully, they never decrease the quality of the research work. Selection of one kind of procedure does not imply that an exploration of the other procedures is forbidden.

Conclusions

Our exploration of the different procedures does not provide evidence for the preference of some specific function of reliability over others functions. In general, linear parsimony is outperformed by convex functions of parsimony.

A desirable finding of our analysis shows than it is possible a broad exploration of different functions without an expensive time payoff. This is possible with a small number of replicates in any of the procedures proposed here, the small number of replicates (20 replicates) produce nearly the same results than huge number of replicates (200 or 1000). Even if one is worried about 20 as a small number, results from 100 and 1000 replicates are indistinguishable.

In the same respect, as direct node count (NC) does not difference with the reference cladogram, it has a direct consequence on time, over a broad exploration. If a reference tree will be used, then the search for this tree would not need to be a high time consuming activity, the only requirement is the use of an heuristic that provides a rough approximation to the answer. Goloboff & Farris's (2001) procedure, and the slightly variation used here, might be a perfect "fast and clean" alternative. Also, it shows that the amount of common nodes between replicates and the reference tree is not a matter of the cladogram resolution. NC is the most versatile of the measures used here, it could be used in any case and can discriminate among different functions. NC is easy to interpret, as a shared node between two cladograms implies than the elements of that node are common, including terminals and the potential synapomorphies (or, at least, preliminary state sets) for the node. Under these results we propose the following suggestions to perform a sensitivity analysis for selecting among reliability functions used in cladistic analysis.

Broad exploration. We encourage an initial broad exploration of different functions and procedures. It could be done in a reasonable time using few replicates (e.g. 20 replicates). It allow exploration of several parameters, as jackknife cut values. If a

reference cladogram is used, we recommend a fast search to find it.
Measures. Given their performance we recommend the use of NC as a measure for selection among functions. It remains constant independently of the reference tree and is not biased by a particular function.
Pick a function. As a result of the broad exploration, high stable areas could be found. It is important to perform a deeper exploration of these functions (more replicates, better reference trees, and maybe fractional values). In addition, with character jackknife, or fast searches, it is possible to recycle these trees for support measures (jackknife and Bremer support).
Finally, we want to encourage readers in two ways. First, we hope that this paper encourage these kinds of sensitivity analyses, not only the ones proposed here but their own procedures and modifications. Second to remark that the problem in cladistics (and in other methodologies) is the cladogram selection. Linear parsimony is only one of many possibilities, but any method that grouping by synapomorphies and not excluding evidence, is perfectly valid.

Acknowledgements

This work was supported by the grant 1102-05-13563 Colciencias (Consejo Nacional de Ciencia y Técnica, Colombia). P. Goloboff and M. Ramírez read the manuscript and provided many valuable comments and suggestions. Programs and data sets are available at Laboratorio de Sistemática y Biogeografía UIS website <http://ciencias.uis.edu.co/labsist/>. The authors acknowledge the effort from “The Bloodshed software group” <http://www.bloodshed.net> for providing a free C++ developer environment. S. Gavassa helped with manuscript’s draft. All our crew at Laboratorio de Sistemática y Biogeografía provided feedback and encouragement.

References

Aagesen, L., Pettersen, G. & Seberg, O. (2005). Sequence length variation, indel

costs, and congruence in sensitivity analysis. *Cladistics, 21*, 15-30.

Allard, M. W. & Carpenter, J. M. (1996). On weighting and congruence. *Cladistics, 12*, 183-198.

Allard, M. W., Farris, J. S. & Carpenter, J. M. (1999). Congruence among mammalian mitochondrial genes. *Cladistics, 15*, 75-84.

Camin, J. H. & Sokal, R. R. (1965). A method for deducing branching sequences in phylogeny. *Evolution, 19*, 311-326.

Carpenter, J. M. (1988). Choosing among multiple equally parsimonious cladograms. *Cladistics, 4*, 291-296.

Cerdeño, E. (1995). Cladistic analysis of the family Rhinocerotidae. *American Museum novitates, 3143*, 1-25.

Dayhoff, M. O. (1969). Computer analysis of protein evolution. *Scientific American, 221*, 87-95.

Eernisse, D. J. 1997. Arthropod and annelid relationships re-examined. In R. A. Fortey & R. H. Thomas (Eds.) *Arthropod relationships* (pp. 43-56) London: Chapman and Hall.

Erséus, C., Källersjö, M., Ekman, M. & Hovmöller, R. (2001). 18s rDNA phylogeny of the Tubificidae (Clitellata) and its constituent taxa: dismissal of the Neididae. *Molecular phylogenetics and evolution, 22*, 414-422.

Farris, J. S. (1969). A successive approximations approach to character weighting. *Systematic zoology, 18*, 374-385.

Farris, J. S. (1970). Methods for computing Wagner trees. *Systematic zoology, 19*, 83-92.

Farris, J. S. 1983. The logical basis of phylogenetic analysis. In N. I. Platnick & V. A. Funk (Eds.) *Advances in cladistics, volume 2* (pp. 7-36). New York: Columbia University Press.

Farris, J. S. (1997). The future of phylogenetic reconstruction. *Zoologica scripta, 26*, 303-311.

Farris, J. S. (2001). Support weighting. *Cladistics, 17*, 389-394.

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., Källersjö, M., Kluge, A. G. & Bult, C. (1994). Testing significance of incongruence. *Cladistics, 10*, 315-319.
Fitch, W. M. (1971). Towards defining the course of evolution: minimum change for a specific tree topology. *Systematic zoology, 20*, 406-416.
Fontal-Cazalla, F. M., Buffington, M. L., Nordlander, G., Liljeblad, J., Ros-Farré, P., Nieves-Aldrey, J. L., Pujade-Villar, J. & Ronquist, F. (2002). Phylogeny of the Eucolinae (Hymenoptera: Cynipoidea: Figitidae). *Cladistics, 18*, 154-199.
Giribet, G. & Wheeler, W. C. (1999). On gaps. *Molecular phylogenetics and evolution, 13*, 132-143.
Giribet, G., DeSalle, R. & Wheeler, W. C. 2002. 'Pluralism' and the aims of phylogenetic research. In R. DeSalle, G. Giribet & W. C. Wheeler, (Eds.) *Molecular systematics and evolution: theory and practice* (pp. 141-146). Basel: Birkhäuser.
Goloboff, P. A. (1993a). Estimating character weights during tree search. *Cladistics, 9*, 83-91.
Goloboff, P. A. (1993b). A reanalysis of Mygalomorph spider families (Araneae). *American Museum novitates, 3056*, 1-32.
Goloboff, P. A. (1995). Parsimony and weighting: a reply to Turner and Zandee. *Cladistics, 11*, 91-104.
Goloboff, P. A. (1997a). Self-weighted optimization: tree searches and character state reconstructions under implied transformation cost. *Cladistics, 13*, 225-245.
Goloboff, P. A. (1997b). *Principios básicos de cladística*. Buenos Aires: Sociedad Argentina de Botánica.
Goloboff, P. A., & Farris, J. S. (2001). Methods for quick consensus estimation. *Cladistics, 17*, S26-S34.
Goloboff, P. A., Farris, J. S., Källersjö, M., Oxelman, B., Ramírez, M. J. & Szumik, C. (2003) Improvements to resampling measures of group support. *Cladistics, 19*, 324-332.
Goloboff, P. A., Farris, J. S. & Nixon, K.C. (2004). *TNT* [program and documentation]. Available via http://www.zmuc.dk/public/phylogeny/TNT/

Goloboff, P. A., Pol, D. (2002). Semi-strict supertrees. *Cladistics, 18*, 514-525.
Grant. T. (2002). Testing methods: the evaluation of discovery operations in evolutionary biology. *Cladistics, 18*, 94-111.
Grant, T. & Kluge, A. G. (2003). Data exploration in phylogenetic inference: scientific, heuristic, or neither. *Cladistics, 19*, 379-418.
Grimaldi, D. A. (1990). Phylogenetic revised classification of genera in the Drosophilidae (Diptera). *Bulletin of American Museum of Natural History, 197*, 1-139.
Guilbert, E. (2001) Phylogeny and evolution of exaggerated traits among the Tingidae (Heteroptera, Cimicomorpha). *Zoologica scripta, 30*, 313-324.
Gustafsson, M. H. G., Pepper, A. S. –R., Albert, V. A. & Källersjö, M. (2001). Molecular phylogeny of the Barnadesioideae (Asteraceae) *Nordic Journal of Botany, 21*, 149-160.
Hennig, W. (1965). Phylogenetic systematics. *Annual review of entomology, 10*, 97-116.
Hennig, W. (1968). *Elementos de una sistemática filogenética*. Buenos Aires: Eudeba.
Horovitz, I. (1999). A phylogenetic study of living and fossil Platyrrhines. *American Museum novitates, 3269*, 1-40.
Kjer, K. M., Blahnik, R. J. & Holzanthal, R. W. (2001). Phylogeny of Trichoptera (Caddisflies): characterization of signal and noise within multiple datasets. *Systematic biology, 50*, 781-816.
Kjer, K. M., Blahnik, R. J. & Holzenthal, R. W. (2002). Phylogeny of caddisflies (insecta, Trichoptera). *Zoologica scripta, 31*, 83-91.
Kluge, A. G. (1989). A cencern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). *Systematic zoology, 38*, 7-25.
Kluge, A. G. (1997). Sophisticated falsification and research cycles: consequences for differential character weighting in phylogenetic systematics. *Zoologica scripta, 26*, 349-360.
Kluge, A. G. & Farris, J. S. (1969). Quantitative phyletics and the evolution of

anurans. *Systematic zoology, 18*, 1-32.
Lopardo, L. (2005). Phylogenetic revision of the spider genus *Negayan* (Araneae, Anyphaenidae, Amaurobioidinae). *Zoologica scripta, 34*, 245-277.
Miller, J. A. (2003). Assessing progress in systematics with continuous jackknife function analysis. *Systematic biology, 52*, 55-65.
Nelson, G. (1979). Cladistic analysis and synthesis: principles and definitions, with a historical note on Adanson's Familles des plantes (1763-1764). *Systematic zoology, 28*, 1-21.
Nixon, K. C. & Carpenter, J. M. (1996). On consensus, collapsibility, and clade concordance. *Cladistics, 12*, 305-321.
Pinto-Sánchez, N. R., Miranda-Esquivel, D. R. & Muñoz de Hoyos, P. (2005). Phylogenetic analysis of *Gigantodax* (Diptera: Simuliidae). *Insect systematics and evolution, 36*, 219-240.
Platnick, N. I., Coddington, J. A., Forster, R. R. & Griswold, C. E. (1991). Spinneret morphology and the phylogeny of haplogyne spiders (Araneae, Araneomorpha). *American Museum novitates, 3016*, 1-73.
Ramírez, M. J. (1999). *Revisión filogenética de los géneros de arañas de la subfamilia Amaurobioidinae (Anyphaenidae)*. Doctoral thesis dissertation. Buenos Aires: Universidad de Buenos Aires.
Ramírez, M. J. (2003). The spider subfamily Amaurobioidinae (Araneae, Anyphaenidae): a phylogenetic revision at the generic level. *Bulletin of American Museum of Natural History, 277*, 1-262.
Rice, K. A., Donoghue, M. J. & Olmstead, R. G. (1997) Analyzing large data sets: *rbc*L 500 revisited. *Systematic biology, 46*, 554-563.
Ruiz-Trillo, I., Riutost, M., Littlewood, D. T. J., Herniou, E. A. & Bagun, J. (1999). Acoel flatworms: earliest extant bilateralian metazoans, not members of the Platyhelminthes. *Science, 283*, 1919-1923.
Rydin, C., Källersjö, M. (2002). Taxon sampling and seed plant phylogeny. *Cladistics, 18*, 485-513.
Rydin, C., Källersjö, M. & Friis, E. M. (2002). Seed plant relationships and the

systematic position of Gnetales based on nuclear and chloroplast DNA: conflicting data, rooting problems, and the monophyly of conifers. *International journal of plant sciences, 163*, 197-214.

Schuh, R. T. (1984). Revision of the Phylinae (Hemiptera, Miridae) of the Indo-Pacific. *Bulletin of American Museum of Natural History, 177*, 1-476.

Swofford, D. L. & Begle, D. P. (1993). User's manual for PAUP, version 3.1. Washington: Smithsonian institution.

Swofford, D. L., Olsen, G. L., Wadell, P. J., & Hillis, D. M. 1996. Phylogenetic inferemce. In D. M. Hillis, C. Moritz, and B. K. Mabvle (Eds.) *Molecular systematics* (pp. 407-514). Sunderland: Sinauer.

Turner, H. & Zandee, R. (1995). The behaviour of Goloboff's tree fitness measure *F*. *Cladistics, 11*, 57-72.

Wagner, W. H., Jr. (1961). Problems in the classification of ferns. In *Recent advances in botany, volume 1* (pp. 841-844) Toronto: University of Toronto Press.

Wheeler, W. C. (1995). Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data. *Systematic biology, 44*, 321-331.

Wheeler, W. C. & Hayashi, C. Y. (1998). The phylogeny of the extant Chelicerate orders. *Cladistics, 14*, 173-192.

Wheeler, W. C., Whiting, M., Wheeler, Q. D. & Carpenter, J. M. (2001). The phylogeny of the extant hexapod orders. *Cladistics, 17*, 113-169.

Whiting, M. F., Carpenter, J. M., Wheeler, Q. D. & Wheeler, W. C. (1994). The Strepsiptera problem: phylogeny of the holometabolous insect orders inferred from 18s and 28s ribosomal DNA sequences and morphology. *Systematic biology, 46*, 1-68.

Wikström, N., Kenrick, P. & Chase, M. (1999). Epiphytism and terrestrialization in tropical *Huperzia* (Lycopodioaceae). *Plant systematics and evolution, 218*, 221-243.

Appendix 1. The number of taxa, characters (chas) and informative characters (ichars) for the matrices used in this study. We use morfological (morf), molecular (mol) or combined (ev tot) matrices. TreeBASE is the study accession number for treeBASE <http://treebase.org>

Matrix	type	Taxa	chars	ichars	TreeBASE	Reference
HOR	morf	33	86	86		Horovitz (1999)
GOL	morf	41	71	70		Goloboff (1993b)
GUS	mol	41	509	61	S660	Gustafsson *et al*. (2001)
FON	morf	45	148	145		Fontal-Cazalla *et al*. (2002)
CER	morf	46	72	71		Cerdeño (1995)
GUI	morf	55	76	68		Guilbert (2001)
WIK	mol	63	1272	262	S420	Wikström *et al*. (1999)
TAB	morf	65	96	90		Coscarón & Miranda-Esquivel (unpublished)
GIG	Morf	66	71	71		Pinto-Sanchez *et al*. (2005)
KEH	mol	67	1871	272	S659	Erseus *et al*. (2001)
SCH	morf	76	75	57		Schuh (1984)
RUI	mol	78	2555	1456	S655	Ruiz-Trillo *et al*. (1999)
WHI	ev tot	82	1590	696	S325	Whiting *et al*. (1994)
DAS	morf	85	60	55		Gonzalez *et al*. (unpublished)
EER	mol	110	3008	1346		Eernise (1997)
RYD	mol	119	5923	2194	S709	Rydin *et al*. (2002)
DRO	morf	159	217	206		Grimaldi (1990)
Zilla	mol	500	759	759		Rice *et al*. (1997)

Appendix 2. Comparison between **gf-s** and **gf-75** procedures. The results without and with a reference tree are market with "no" and "yes" respectively. See text and appendix 1 for acronyms.

Matrix	Measure	gf75-no	gfs-no	gf75-yes	gfs-yes
CER	NC	10	7	9	7
DAS	NC	10	10	10	10
DRO	NC	7	9	7	9
EER	NC	8	10	7	10
EKH	NC	6	6	6	6
FON	NC	2	6	2	6
GIG	NC	9	9	9	9
GOL	NC	6	6	6	6
GUI	NC	10	10	10	10
GUS	NC	6	6	6	6
HOR	NC	7	7	7	7
RUI	NC	4	7	10	7
RYD	NC	1	1	4	1
SCH	NC	10	4	4	4
TAB	NC	10	5	10	5
WHI	NC	1	1	1	1
WIK	NC	1	1	1	1

Appendix 3. Comparison between several cut values for three different jackknife proedures. The results without and with a reference tree are market with "no" and "yes" respectively. See text and table one for acronims. "**gf-s**" results are included to allow comparisons. "**p2**" are results from partition that could be roughly equivalent to cut of 50%.

		Measure				cjac			
Matrix	**Ref**	**/cut value**	**gf-s**	**0.05**	**0.15**	**0.25**	**0.36**	**0.45**	**p2**
DRO	no	NC	9	9	9	9	8	6	10
DRO	yes	NC	9	9	9	8	8	5	5
FON	no	NC	6	5	4	3	2	2	3
FON	yes	NC	6	3	3	3	3	3	3
GIG	no	NC	10	10	10	9	9	1	2
GIG	yes	NC	9	10	10	10	10	8	10
GUI	no	NC	10	10	10	10	9	10	10
GUI	yes	NC	10	10	10	4	9	9	9
SCH	no	NC	3	10	10	9	9	10	10
SCH	yes	NC	4	4	5	7	7	7	7
TAB	no	NC	4	4	10	9	9	7	10
TAB	yes	NC	5	4	6	6	5	7	5
						mjac			
			gf-s	**0.05**	**0.15**	**0.25**	**0.36**	**0.45**	
DRO	no	NC	9	3	3	5	5	4	
DRO	yes	NC	9	4	3	5	5	5	
FON	no	NC	6	3	3	3	3	2	
FON	yes	NC	6	3	3	3	3	2	
GIG	no	NC	10	10	7	8	9	8	
GIG	yes	NC	9	10	10	8	9	10	
GUI	no	NC	10	2	3	3	7	9	
GUI	yes	NC	10	9	9	9	8	9	
SCH	no	NC	3	9	8	9	9	9	
SCH	yes	NC	4	5	5	7	7	7	
TAB	no	NC	4	3	4	4	5	4	
TAB	yes	NC	5	4	4	4	5	5	
						tjac			
			gf-s	**0.05**	**0.15**	**0.25**	**0.36**	**0.45**	
DRO	no	NC	9	9	9	9	2	2	
DRO	yes	NC	9	9	8	4	4	4	
FON	no	NC	6	3	4	3	3	3	
FON	yes	NC	6	3	3	3	3	3	
GIG	no	NC	10	10	10	10	9	8	
GIG	yes	NC	9	10	10	10	10	10	
GUI	no	NC	10	0	0	0	8	10	
GUI	no	MF	1	0	1	1	1	1	

GUI	yes	NC	10	10	10	10	10	9
SCH	no	NC	3	3	5	5	9	9
SCH	yes	NC	4	4	5	5	5	5
TAB	no	NC	4	4	4	3	3	3
TAB	yes	NC	5	4	4	4	4	4

Tables

Table 1. The selected function according to the procedure and measure used. The first and second columns are results without (wo) and with (w) a reference tree, respectively. See text and appendix 1 for acronyms.

Matrix	Measure	gf-s		cjac		tjac		mjac		p2n	
		wo	w	wo	w	wo	w	wo	w	wo	w
CER	NC	7	7	4	8	5	8	3	3	2	5
	IC	1	0	0	0	0	0	0	0	0	0
	MF	1	1	3	4	1	1	1	1	1	1
DAS	NC	10	10	2	2	10	10	2	10	2	2
	IC	1	10	0	0	0	0	0	0	0	0
	MF	1	1	5	3	1	1	1	1	1	1
DRO	NC	9	9	8	8	2	4	5	5	10	5
	IC	2	9	0	8	0	4	0	4	0	4
	MF	3	3	10	3	2	1	3	1	1	0
EER	NC	10	10	10	10	9	9	9	9	10	10
	IC	2	10	0	10	9	9	9	9	0	10
	MF	0	3	0	6	1	1	1	1	1	9
EKH	NC	6	6	8	10	6	6	7	8	10	10
	IC	3	6	0	0	0	6	0	0	0	0
	MF	1	1	7	1	1	1	1	1	1	7
FON	NC	6	6	2	3	3	3	3	3	3	3
	IC	9	2	3	3	0	3	0	0	6	0
	MF	1	1	10	10	1	1	1	1	1	2
GIG	NC	10	9	9	10	9	10	9	9	2	10
	IC	3	0	1	0	1	0	0	0	0	0
	MF	4	10	4	1	2	1	2	1	1	1
GOL	NC	3	6	10	10	3	7	5	6	10	10
	IC	3	6	0	10	0	7	0	6	0	10
	MF	0	0	1	1	1	1	1	1	1	0
GUI	NC	10	10	9	9	8	10	7	8	10	9
	IC	1	0	0	0	0	0	0	0	0	0
	MF	1	0	0	1	1	1	1	1	1	1
GUS	NC	6	6	10	9	5	5	8	10	10	9
	IC	3	0	0	0	0	0	0	0	0	0
	MF	1	0	1	1	1	1	1	1	1	5
HOR	NC	7	7	2	1	4	2	3	2	4	1
	IC	2	7	6	4	0	10	0	9	0	4
	MF	0	0	3	3	1	1	1	1	1	1
RUI	NC	7	7	7	6	4	4	5	5	9	5
	IC	5	7	0	0	0	0	0	0	0	0

	MF	0	0	0	4	1	1	1	1	1	5
RYD	NC	1	1	9	10	6	6	7	9	10	9
	IC	2	1	0	3	6	3	10	3	0	3
	MF	0	0	0	5	1	1	1	1	1	2
SCH	NC	3	4	9	7	9	4	8	7	10	7
	IC	3	10	9	10	4	10	8	10	0	10
	MF	0	0	1	1	2	1	2	1	1	1
TAB	NC	4	5	9	5	3	4	5	5	10	5
	IC	0	0	0	0	0	0	0	0	0	0
	MF	1	2	4	1	1	1	1	1	1	1
WHI	NC	1	1	2	2	1	1	2	1	2	2
	IC	1	1	0	2	3	1	8	1	0	2
	MF	1	0	7	10	1	1	1	1	1	1
WIK	NC	1	1	1	1	1	1	5	1	6	1
	IC	10	1	0	2	0	1	0	3	8	3
	MF	1	1	0	3	1	1	1	1	1	1

Table 2. The effect of the reference tree's search. Results from IC are variable specifically when comparing results from a reference tree from a fast search (*gf20*). Each entry is the average of 10 different runs (17 matrices, 11 functions, 6 procedures, 5 different reference trees). χ^2 values and probabilities are only intended to measure dispersion and are not intended for statistical inference. NC used with direct values, IC and MF multiplied by 100. Values used as "expected" are the results of the reference tree search with 10 ratchet replicates of 200 iterations. NA is used because the MF-index is impossible to use with **gf-75**. *gf20*=fast search, 20 replicates; *wd100*=Wagner-Dayoff trees, 100 replicates; *wd500*=Wagner-Dayoff trees, 500 replicates; *rat200x10*= 200 replicates of 10 iterations of ratchet.

Measure	**NC**			**IC**			**MF**		
Split	χ^2	**cases**	**p**	χ^2	**cases**	**p**	χ^2	**cases**	**p**
Total	230.983	3916	1.000	17481.129	3916	0.000	11.132	3005	1.000
Procedures									
Cjac	31.084	715	1.000	4849.082	715	0.000	3.088	704	1.000
Tjac	56.267	715	1.000	3687.706	715	0.000	0.382	704	1.000
Mjac	39.320	715	1.000	3225.080	715	0.000	0.492	704	1.000
gf-s	10.424	528	1.000	181.569	528	1.000	5.486	189	1.000
gf75	56.791	528	1.000	272.197	528	1.000	NA	NA	NA
p2n	37.097	715	1.000	5265.495	715	0.000	1.684	704	1.000
Reference tree									
gf20	142.137	748	1.000	16540.167	748	0.000	4.173	748	1.000
wd100	7.740	1122	1.000	634.822	1122	1.000	3.948	821	1.000
wd500	2.054	990	1.000	170.458	990	1.000	1.507	722	1.000
rat200x10	79.053	1056	1.000	135.682	1056	1.000	1.504	714	1.000

Table 3. The effect of the number of replicates for each procedure. In all the measures, the number of replicates does not change the measure values. Each case represents the average of 10 different runs (17 matrices, 11 functions, 6 procedures, 4 different number of replicates, with and without a reference tree). χ^2 values and probabilities are only intended to measure dispersion not for statistical inference. NC used with direct values, IC and MF multiplied by 100. Values used as "expected" are the results of 1000 replicates. Excluded cases due to division by zero.

Measure	**NC**			**IC**			**MF**		
Split	χ^2	**cases**	**p**	χ^2	**cases**	**p**	χ^2	**cases**	**p**
Total	9.511	1980	1.000	26.324	1980	1.000	449.769	1800	1.000
Procedures									
Cjac	4.525	396	1.000	7.332	396	1.000	2.604	396	1.000
Tjac	1.545	396	1.000	5.511	396	1.000	7.426	396	1.000
Mjac	2.494	396	1.000	10.398	396	1.000	8.295	396	1.000
gf-s	0.603	396	1.000	1.433	396	1.000	3.186	216	1.000
p2n	0.345	396	1.000	1.651	396	1.000	428.259	396	0.120
Number of replicates									
20	8.532	660	1.000	23.548	660	1.000	206.293	600	1.000
100	0.667	660	1.000	1.862	660	1.000	241.524	600	1.000
200	0.312	660	1.000	0.914	660	1.000	1.952	600	1.000

Table 4. The effect of the number of replicates on the difference between IC and IE, and on MF when any of the compared trees is used as reference, and no-reference tree is used. At least with 20 replicates measures could be freely interchanged. Each case represents the average of 10 different runs (17 matrices (for 100 replicates, 6 for the other), 11 functions, 6 procedures, 4 different number of replicates, without a reference tree). “dif”, shows the average difference between IC and IE and MF, “s” the standard deviation, and “max” the maximum difference found.

Split	**IC-IE dif**	**s**	**Max**	**cases**	**MF dif**	**s**	**Max**	**Cases**
Total	0.004	0.004	0.041	1925	0.002	0.004	0.058	1320
Procedures								
Cjac	0.004	0.004	0.030	385	0.000	0.000	0.001	264
Tjac	0.004	0.004	0.041	385	0.001	0.001	0.012	264
Mjac	0.004	0.004	0.029	385	0.001	0.001	0.007	264
gf-s	0.002	0.002	0.015	385	0.000	0.000	0.000	264
p2n	0.005	0.006	0.040	385	0.006	0.008	0.058	264
Number of replicates								
20	0.008	0.007	0.041	330	0.003	0.008	0.058	330
100	0.004	0.003	0.024	935	0.001	0.003	0.017	330
200	0.003	0.002	0.011	330	0.001	0.002	0.011	330
1000	0.001	0.001	0.006	330	0.001	0.001	0.006	330

Table 5. Results from fast searches. The selected function according to the procedure using NC. The first and second column are results without and with a reference tree respectively. See text and appendix 1 for acronyms.

Matrix	**cjac**		**gf-s**		**mjac**		**tjac**	
	wo	**w**	**wo**	**w**	**wo**	**w**	**wo**	**w**
CER	2	10	8	8	3	5	5	7
DAS	2	3	10	10	3	8	10	8
DRO	6	7	9	9	5	3	5	3
EER	10	9	10	9	8	9	6	9
EKH	10	7	8	5	10	5	9	5
FON	2	2	5	5	5	2	3	3
GIG	9	9	10	10	6	9	10	9
GOL	8	10	3	6	10	10	4	6
GUI	8	9	10	10	10	5	9	9
GUS	10	5	6	5	10	5	5	5
HOR	2	3	7	7	4	8	7	8
RUI	8	8	7	8	5	4	8	4
RYD	10	10	1	1	5	6	7	6
SCH	10	10	3	4	10	10	8	10
TAB	9	8	5	5	7	8	3	8
WHI	2	2	1	1	1	1	1	1
WIK	6	1	1	1	10	1	1	1
ZIL	0	10	9	9	10	10	10	9

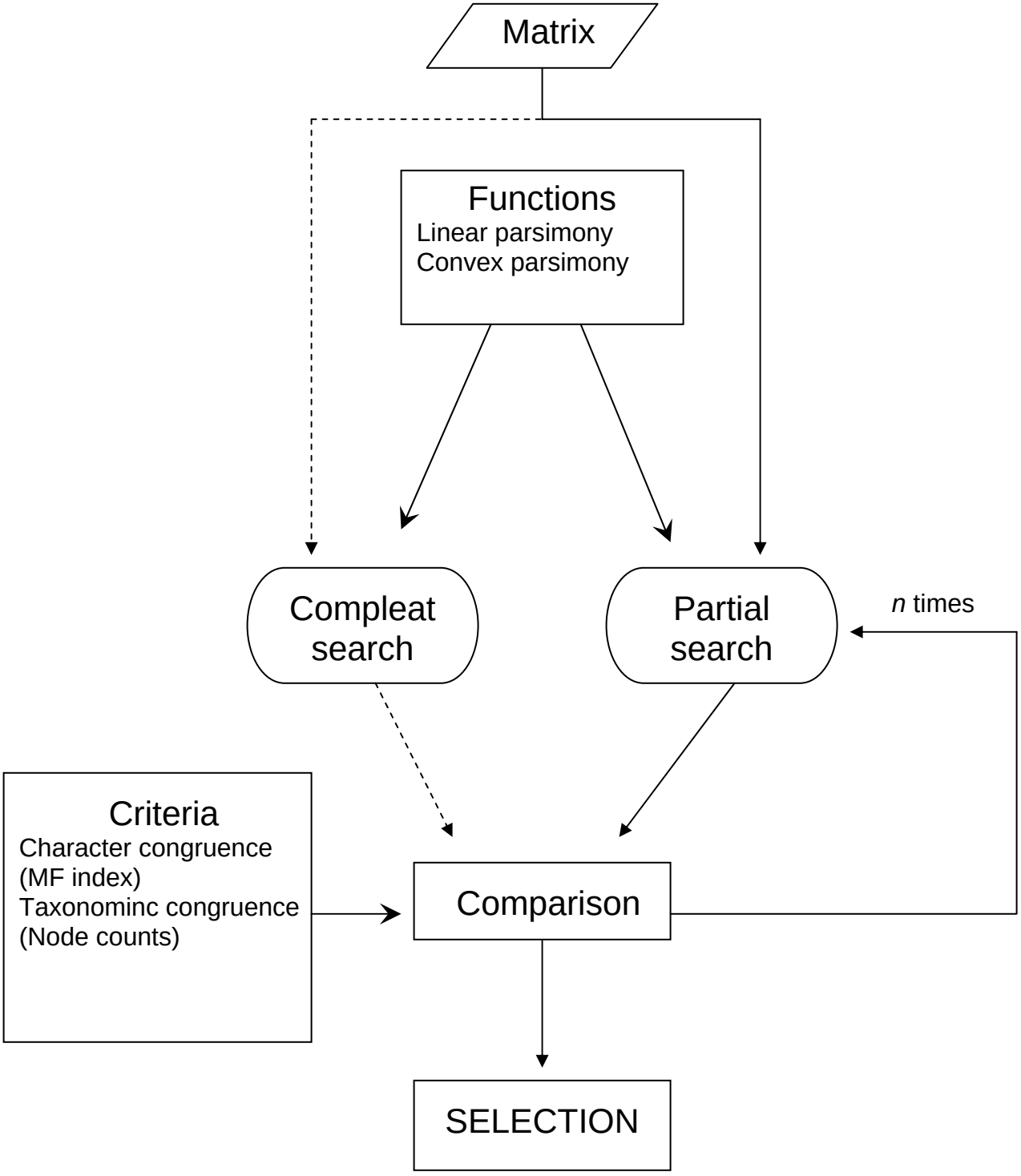
Figure caption

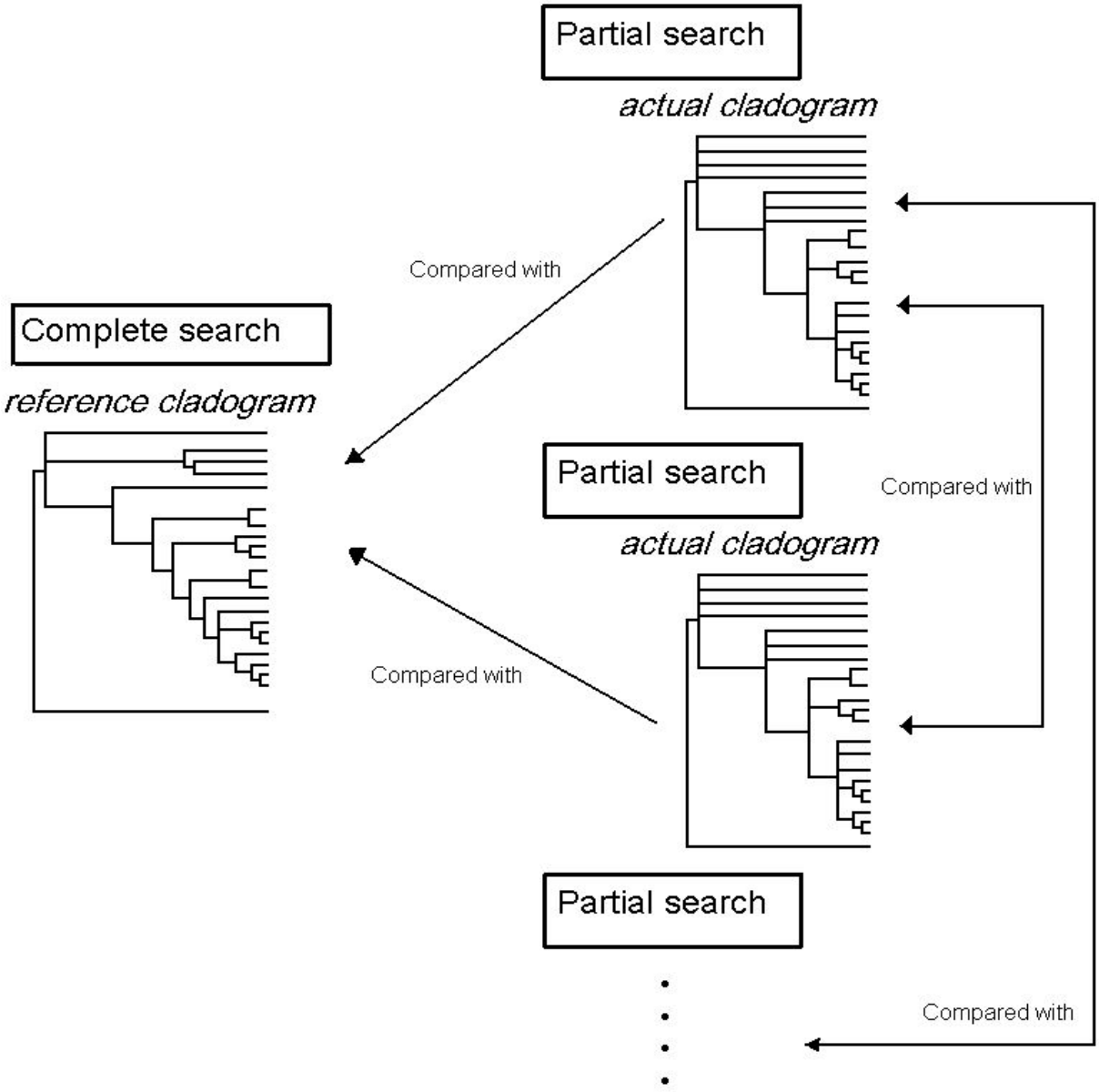
Figure 1. A flowchart showing the basic steps of the procedures described here. Broken line indicates optional steps. An elliptic box indicates specific operations. Squared box shows the criterions used during the operation performed.

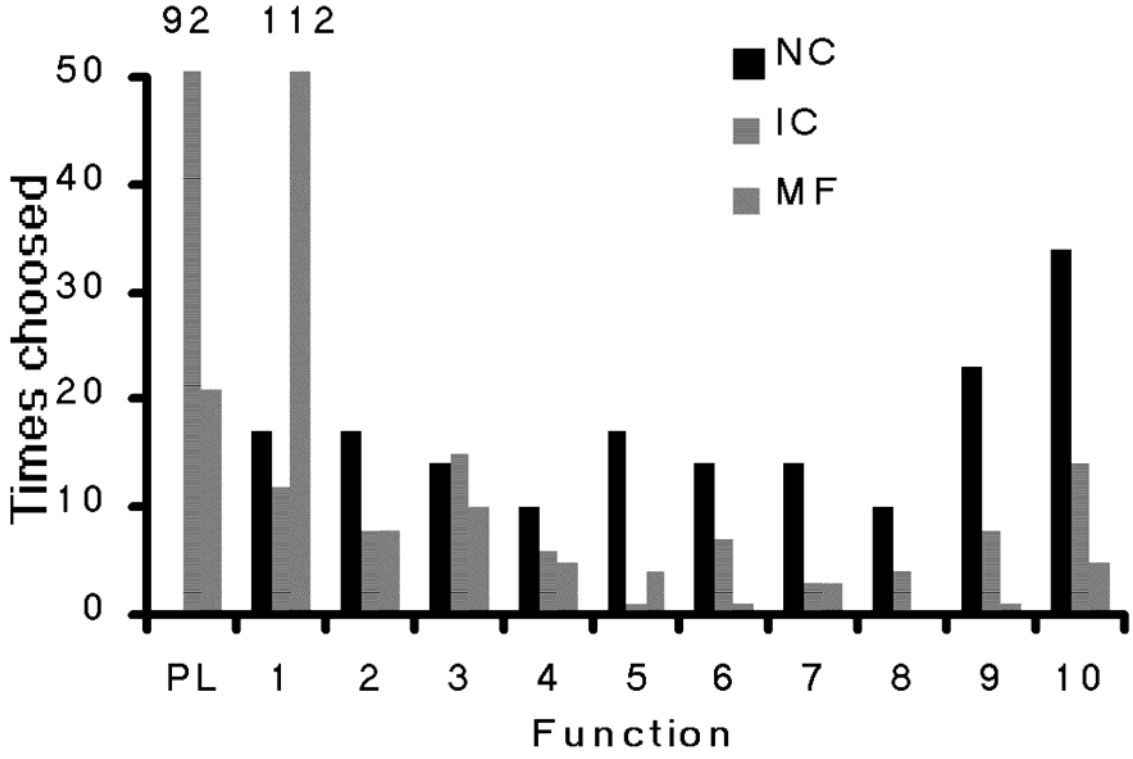
Figure 2. Comparison with and without a reference tree. The reference tree is the solution from a compleat search, whereas an approximate tree is the solution of a partial search. When a reference cladogram is used (left) each approximate tree is compared with the reference cladogram. In analyses without a reference tree (right) each approximate tree is compared with other (randomly selected) approximate cladogram obtained from the same kind of partial search, as both trees are from partial search all used measures could be scaled with any of the two trees (so the arrows are two sided).

Figure 3. Selected function (100 replicates, all procedures, all matrices except Zilla) according to the measure used. “PL” is used to represent linear parsimony.

Figures


Matrix
Functions
Linear parsimony
Convex parsimony
Compleat search
Partial search
n times
Criteria
Character congruence (MF index)
Taxonominc congruence (Node counts)
Comparison
SELECTION


Partial search
actual cladogram
Compared with
Complete search
reference cladogram
Partial search
Compared with
actual cladogram
Compared with
Partial search
Compared with


92
112
NC
IC
MF
Times choosed
50
40
30
20
10
0
PL
1
2
3
4
5
6
7
8
9
10
Function