Blackwell Publishing

Cladistics 20 (2004) 64–75

Cladistics

www.blackwell-synergy.com

Measures of stratigraphic fit to phylogeny and their sensitivity to tree size, tree shape, and scale

Diego Pol[1,*], Mark A. Norell[1] and Mark E. Siddall[2]

[1]*Division of Paleontology, American Museum of Natural History, Central Park West at 79th street, New York, NY 10024*
[2]*Division of Invertebrates, American Museum of Natural History, Central Park West at 79th street, New York, NY 10024, USA*

Accepted 17 November 2003

Abstract

Measures of stratigraphic fit to phylogeny are analyzed to test how they are affected by the shape and size of the phylogenetic trees and by the number of stratigraphic intervals encompassed. Monte Carlo randomizations are used to investigate the sensitivity of three commonly used measures (SCI, GER and MSM*) approximating their distribution of possible values under certain conditions. All are shown to vary in different ways as parameters are varied, although MSM* seems to be the most invariant in the analyzed parameter space. These results suggest that the raw metrics should not be used for comparing the fit of different taxonomic groups or competing phylogenetic trees of the same group that differ in tree size or balance. Tree balance also affects the distributions used in significance tests based on randomization and therefore their results should not be interpreted in terms of the amount of conflict implied by a phylogenetic tree.

Introduction

Phylogenetic trees that contain fossil taxa possess two distinct sources of information on the timing of their evolutionary history. A given phylogenetic tree implies a certain temporal order of the successive branching events. Fossil terminal taxa, instead, provide temporal information based on chronostratigraphy. If these two sources of temporal information are known independently, they can be compared in order to measure their congruence. Several measures have been proposed, attempting to describe the agreement between chronostratigraphy and phylogenetic trees (Norell and Novacek, 1992; Huelsenbeck, 1994; Benton and Hitchin, 1996; Siddall, 1998; Wills, 1999; Pol and Norell, 2001). These measures have been used for two main purposes, a comparison of the congruence between phylogeny and stratigraphy in several taxonomic groups (e.g. Benton and Hitchin, 1996; Benton et al., 1999, 2000), and a comparison of stratigraphic congruence in competing phylogenetic trees. The latter use was regarded either as a descriptive statistic (e.g. Brochu and Norell, 2000) or as an auxiliary optimality criterion for selecting the preferred phylogenetic hypothesis (e.g. Huelsenbeck, 1994).

In order to evaluate how to measure the fit of stratigraphy to a phylogenetic tree we must first determine what is being measured and what are the properties of different measures. We will only briefly consider the first of these, because it has been the main focus of prior work (Hitchin and Benton, 1997a,b; Siddall, 1998; Wagner and Sidor, 2000; Pol and Norell, 2001). The second aspect, concerning the properties and biases of measures, is particularly relevant for inferring the suitability of these measures for different purposes. This aspect has only been briefly discussed in the literature (Siddall, 1996, 1997; Hitchin and Benton, 1997a,b; Wills, 1999) and these considerations focused on the influence of tree size and shape and were based on both theoretical approaches and empiric data. However, no broad consensus has emerged (Huelsenbeck, 1994; Hitchin

*Corresponding author.
E-mail address: dpol@amnh.org

and Benton, 1997a,b; Siddall, 1997, 1998; Pol and Norell, 2001). Here, the sensitivity of three commonly used measures of phylogenetic-stratigraphic congruence are analyzed through a theoretical approach.

Measures of conflict

Several different measures have been proposed to measure the conflict between the temporal information contained in the fossil record and a phylogenetic tree. Some of them (e.g. SRC, SCI, MSM and GER; Norell and Novacek, 1992; Huelsenbeck, 1994; Siddall, 1998; Wills, 1999) purport to measure how well the first appearances in the fossil record are reflected in the branching order of taxa in the phylogenetic tree. Two of these congruence metrics are based on the measurement of ghost ranges or implied gaps (MSM and GER), while others just measure the number of mismatches between phylogeny and stratigraphy (SRC and SCI). Other metrics based on ghost ranges have been proposed in order to measure the completeness of the fossil record implied by the phylogenetic tree, in terms of how much is inferred to be missing respect to how much is actually represented in the fossil record (Z, RCI and IG; Norell, 1992; Benton and Storrs, 1994; Smith and Littlewood, 1994). We will focus here on those measures designed for measuring congruence rather than the completeness metrics.

The Spearman Rank Coefficient (SRC; Norell and Novacek, 1992) has been criticized on several grounds and suffers from serious limitations because it is only applicable to fully pectinate trees or to trees reduced to their pectinate components (Norell, 1993). Therefore, this measure is currently in disuse and it will not be considered in this study.

Huelsenbeck (1994) proposed the Stratigraphic Consistency Index (SCI) as a measure of the proportion of stratigraphically consistent nodes:

$$SCI = \frac{C}{N},$$

where N is the number of internal nodes (excluding the root) and C the number of stratigraphically consistent nodes (i.e. a node in which the oldest first occurrence above the node is equal or younger than the oldest first occurrence of the sister taxon of that node). SCI was subsequently used for measuring the conflict between stratigraphy and topology in several related studies (Benton and Hitchin, 1996; Benton et al., 1999, 2000).

Two measures based on the magnitude of ghost lineages (Norell, 1992) were proposed. Siddall (1998) suggested the use of a new measure, the Manhattan Stratigraphic Measure (MSM), which was subsequently modified (MSM*) in order to correct for some undesirable properties observed in some special cases (Pol and Norell, 2001). This measure is based on the optimization of an age character using Sankoff parsimony (Sankoff et al., 1976). The step matrix of the age character is set up assigning each stratigraphic age of first appearance to a different character state and the cost of transformation between states determined by the absolute temporal difference among the recorded stratigraphic ages. In contrast to the original MSM formulation, MSM* holds transformations between states to be irreversible to an older age (Pol and Norell, 2001). As a consequence of this modification, the length of the age character on the phylogenetic tree quantifies the sum of ghost lineages. Therefore, the MSM* is calculated in an analogous way as the consistency index of the age character:

$$MSM^* = \frac{L_m}{L_0},$$

where L_m is the minimum length (amount of ghost lineages) for the dataset on any possible topology and L_0 is the actual length obtained optimizing the age character on the independently derived phylogenetic tree.

Wills (1999) proposed another ghost lineage-based measure, the Gap Excess Ratio (GER), formulated originally in the following terms:

$$GER = 1 - \frac{MIG - G_{\min}}{G_{\max} - G_{\min}},$$

where $G_{\min}$ is minimum possible sum of ghost ranges for any given distribution of origination dates, $G_{\max}$ is the maximum possible sum of ghost ranges for a given distribution of origination dates, and MIG is sum of ghost ranges implied by a phylogeny. Although the GER was not formulated based on the optimization of a Sankoff age character, it can be conveniently formulated in an analogous way as the retention index of the age character for comparisons with the MSM* and for an easy implementation in current software packages:

$$GER = \frac{L_m - L_0}{L_M - L_m}$$

where L_M is the maximum length for the age character on any topology (i.e. the tree length on a completely unresolved topology), L_m is the minimum length for the age character on any topology, and L_0 is the actual length obtained optimizing the age character on the independently derived phylogenetic tree.

These two measures, the modified MSM (MSM*) and the GER are based on the age character that measures the extent and amount of ghost lineages but, as in the *ci* and *ri* of regular characters, these two indices are not mutually exclusive and they inform on two different aspects of the same pattern of incongruence between two sources of hierarchical information (see Farris, 1989). A simple example can show what is being measured by these two indices. In Fig. 1, the only

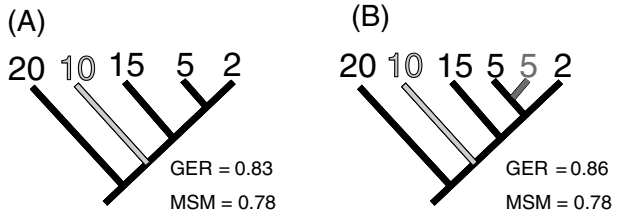

Fig. 1. Two hypothetical examples of phylogenetic trees labeled with the first appearance ages (Ma) of terminal taxa. Both have the same amount of conflict with stratigraphic ages (i.e. presence of a basal lately recorded taxa (10 Ma). However, these differ on the presence of an additional taxon (5 Ma) that adds no conflict since it has the same age as its sister group but increases the maximum possible amount of conflict.

conflict in the temporal information of both trees is the same, the presence of a basal taxon that has a late appearance in the fossil record (10 Ma). The only difference between the two trees is the presence of an additional taxon in tree B (that adds no conflict since it is the same age as its sister taxon). The MSM* remains equal in both cases since it measures the raw amount of conflict, which is the same in these two trees (given by the late appearing taxon at age 10). The GER, instead, is higher in the second case since it measures the amount of conflict respect to the maximum possible amount of conflict, which is larger in tree B.

These formulations of MSM* and GER should be used assuming the absence of ancestors among the terminal taxa in the phylogenetic tree. If ancestors are assumed to be present among terminal taxa, these indices are not simply calculated based on ages of first appearance and can yield different results (Norell, 1992, 1996; Wagner, 1996; Wills, 1999; Wagner, 2000; Wagner and Sidor, 2000). Here, the behavior of these metrics is explored assuming the absence of ancestors since they have been calculated in such way in empirical studies (e.g. Benton and Storrs, 1994; Benton and Simms, 1995; Weishampel, 1996; Benton and Hitchin, 1996, 1997; Hitchin and Benton, 1997a, b; Benton et al., 1999, 2000; Brochu and Norell, 2000, 2001; O'Leary, 2002).

Previous approaches

Properties of SCI

Despite its common use, this measure suffers from several flaws and tree shape biases (Siddall, 1996, 1997; Hitchin and Benton, 1997a,b; Wills, 1999; Wagner and Sidor, 2000).

Siddall (1996) initially criticized the SCI as being biased by tree shape, as evidenced by its sensitivity to tree size. He noted simple cases in which two trees with no conflict in age distribution yielded lower SCI values for balanced trees in comparison to pectinate trees. He also found a positive correlation between SCI and imbalance and a negative correlation between SCI and tree size on a sample of 14 empirical datasets. He inferred that SCI could only be used to compare the stratigraphic fit of trees of equal size and shape. Hitchin and Benton (1997b) did not corroborate these correlations in a compilation of 357 cladograms (except for echinoderms, in which a positive correlation with imbalance was found). Siddall (1997) criticized several aspects of Hitchin and Benton's (1997b) empirical test and explored the distribution of mean SCI values in random assignments of ages in pectinate and balanced trees with different numbers of taxa. The logic behind this procedure was that random assignments of ages would produce meaningless distributions in both balanced and pectinate trees and should therefore lead to equally poor SCI values if this measure is not affected by tree shape. The results of these randomizations suggested a marked sensitivity of the SCI to tree shape irrespective of the number of taxa and number of ages in the analyzed cases. This shape effect (favouring pectinate trees) became more severe as the information content of the trees increased (more taxa or ages).

Wills (1999) noted that the possible range of SCI values was constrained by tree balance, tree size, and also by the distribution of stratigraphic ages among terminal taxa. Furthermore, he suggested that SCI values are not strictly comparable either among trees derived from different datasets (due to potential biases in tree shape, tree size and distribution of ages), or for competing phylogenetic trees coming from the same dataset (due to tree shape biases).

Furthermore, Wagner (2000) and Wagner and Sidor (2000) also noted the reasons for the tree shape bias of SCI and pointed out numerous other parameters that affected SCI values in a series of simulations (e.g. sampling intensity, sampling heterogeneity, phylogenetic accuracy, taxonomic practice, cladogenesis models, and speciation and extinction intensities).

In sum, there is no current agreement on the sensitivity of SCI to some parameters such as tree shape and size, while there are indications that a wide range of other relevant parameters can affect SCI. As noted above, statements on SCI performance differ widely, ranging from considering it biased by all possible factors (Wills, 1999), to considering it unbiased and the best available measure of stratigraphic fit to phylogenetic trees (Hitchin and Benton, 1997b).

*Properties of GER and MSM**

Wills (1999) introduced the GER as an improved measure of stratigraphic fit stating that, in contrast to

previous measures, it controlled for the distribution of temporal range data. Despite this advantage, Wills acknowledged that GER was sensitive to differences in tree balance. Although Wills (1999) devised the GER to improve certain limitations of SCI and other indices, a thorough examination of GER properties is lacking.

Furthermore, some known properties of the retention index, such as its sensitivity to the number of autapomorphic states in a character (Naylor and Kraus, 1995), demands an exploration of this issue on the GER, since it could affect its suitability for comparing data with different age distributions.

Siddall (1998) proposed the MSM as a solution to the problems found with other indices. Employing a randomization procedure similar to that employed for the SCI (Siddall, 1997), he noted that raw MSM values (but not significance) are influenced by tree size but are not influenced by tree shape. Siddall's original measure has been modified in order to avoid spurious performance (Pol and Norell, 2001), and an examination of the modified MSM (i.e. MSM*) sensitivity to tree shape, size and distribution of ages is needed. Additionally, since both GER and MSM* are claimed to be an improvement over previous measures (Pol and Norell, 2001; Siddall, 1998; Wills, 1999), it is necessary to thoroughly test these assertions and to evaluate their relative performance over a wide range of conditions.

Sensitivity analysis

The randomization procedure employed by Siddall (1997) provided an interesting theoretical framework within which to simultaneously explore the sensitivity of different measures to several parameters. This procedure was used here in order to obtain an approximation of the distribution of all possible metric values for a given set of conditions. For a given tree shape (and set of fossil taxa with ages of first appearance) there is a finite universe of possible values that a metric can have, that correspond to each of the possible arrangements of taxa across the unlabeled topology. The effects that tree balance or other parameters have in these distributions have been poorly explored at the time of writing and this Monte Carlo randomization procedure is an efficient way of obtaining an approximation of them (since an exact evaluation of all possible arrangements would be extremely time consuming, even for moderate number of taxa).

In addition, this procedure also provides information on how the significance test could be affected by tree shape or other parameters (since the random data permutation procedure is identical and these distributions are used to assess significance).

Previous randomizations conducted by Siddall (1997, 1998) have proven extremely valuable, although they represent only a small portion of the parameter space. Thus, sensitivity to several parameters remains almost unexplored. Sensitivity of SCI, GER and MSM* to different parameters is analyzed here for a wide range of values. The performance of these indices is evaluated relative to changes in three parameters that can be directly measured in empirical datasets and have been suspected to influence them: tree size, tree shape and number of possible ages of first appearance among terminal taxa.

For the first parameter, three different values have been analyzed: 8, 16 and 32 taxa. These values were chosen because a fully balanced tree can only be obtained with $n_{tax} = 2^x$ (x being any positive integer larger than two). Then, for each of these values, two other parameters (tree shape and number of possible ages) were varied to represent the axes of a bidimensional tree space. Along the asymptote, the shape of the trees was varied, starting with a fully pectinate tree at the bottom, up to a fully balanced tree at the top of the axis (Fig. 2). Tree shape was ordered according to the Imbalance Index (Heard, 1992). Along the abscissa, the number of possible ages was increased from two to 35 different ages (Fig. 2). The ages were randomly assigned to the terminal taxa for each particular combination of tree size, shape and number of possible ages. This procedure was repeated 100-fold for each of the 2500 possible combinations of parameter values (i.e. each point in the three tree spaces).

Finally, the SCI, GER and MSM* values were calculated for each of the 250,000 trees. These measures were calculated using only the age of first appearance (i.e. assuming the absence of ancestors), as is usually done by researchers who previously used these metrics (e.g. Benton and Storrs, 1994; Benton and Simms, 1995; Weishampel, 1996; Benton and Hitchin, 1996, 1997; Hitchin and Benton, 1997a, b; Benton et al., 1999, 2000; Brochu and Norell, 2000, 2001; O'Leary, 2002). If ancestors are identified among the terminal taxa (Wagner, 1995, 1996, 1998, 2000), some of these metrics would have to be calculated in a different way and they can have different properties and yield different results (Wills, 1999; Wagner and Sidor, 2000). Therefore, the analyses conducted here are only relevant to the behavior and properties of these metrics when ancestors are assumed to be absent among the terminal taxa (an assumption made in most empirical studies that used these metrics).

The mean value of each metric for the hundred replicates of each point in the tree space was calculated and plotted with colors corresponding to the magnitude of the mean. As previously noted, if a measure is not sensitive to variations in these parameters, random assignments of possible ages should not produce, on average, significantly different means across the tree space (Siddall, 1997). To display the results obtained in

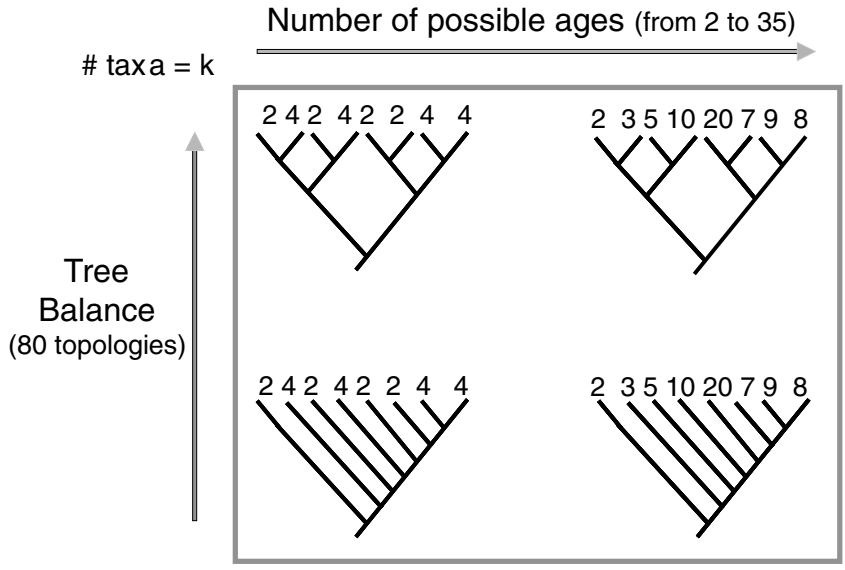

Fig. 2. Example of a tree space of eight taxa, showing tree shape change along the asymptote and different number of ages along the abscissa. Ages were randomly assigned to terminal taxa in each of the possible trees.

these analyses, mean values were codified in a gradual scale ranging from green (index value equal to 0) to blue (index value equal to 1). The amount of color shifts within each tree space represent the sensitivity of an index to tree shape and number of ages. Color shift among the three different tree spaces of 8, 16 and 32 taxa represent the sensitivity of an index to the number of taxa.

Monte Carlo randomization procedures and imbalance ranking was implemented using the macro language of Nona ver. 1.9 (Goloboff, 1993) and the calculation of the SCI, GER and MSM* was derived using the free software package Ghosts 2.4 (Wills, 1999; available at http://paleo.gly.bris.ac.uk/cladestrat/Gho2.html).

In addition to a general sensitivity analysis, other randomization procedures were implemented to highlight the particular behavior of these measures in a more detailed analysis of the influence of tree shape. These procedures were implemented using the macro language of Nona (Goloboff, 1993) and SPA (Goloboff, 1996). A calculation of indices was implemented using macro files for these programs available at <http://research.amnh.org/~dpol/strat.html>.

Results

Because our results are based on randomly assigned data (where ages are distributed randomly across taxa) and hence lack any definitive structure, it would be inappropriate for a measure to yield high values (indicating a tight congruence between stratigraphy and phylogeny) with these data. However, here we are unconcerned with how well a given data set performs, what is of concern is how differently these methods perform under different parameter sets. In these analyses we found that all measures display some sensitivity to parameter variation; however, each behaves in different ways.

SCI

The SCI is the index which is most sensitive to variations in tree shape and the number of possible ages in the randomizations we carried out (Fig. 3). For the three explored tree sizes, mean SCI values of a random assignment of ages decrease with the number of possible ages, corroborating the results of Siddall (1997). Wagner and Sidor (2000) showed a similar relationship in a simulation study. This trend is observed in all analyzed tree shapes, although it was most apparent in pectinate trees with many taxa. Interestingly, in these randomizations mean SCI values decreased consistently with tree imbalance, in contrast to previous reports (Siddall, 1997). Tree shape sensitivity was slightly greater in larger trees. Finally, the previously expected decrease in mean SCI values for increasing number of taxa only occurs in highly imbalanced trees (Figs 3 and 4).

GER

The GER values in these randomizations were also found to be sensitive to the explored parameters, though their pattern was very different from that of SCI. The main difference related to their sensitivity to the number of different ages among terminal taxa. Here, GER mean values increased with increasing numbers of ages, in

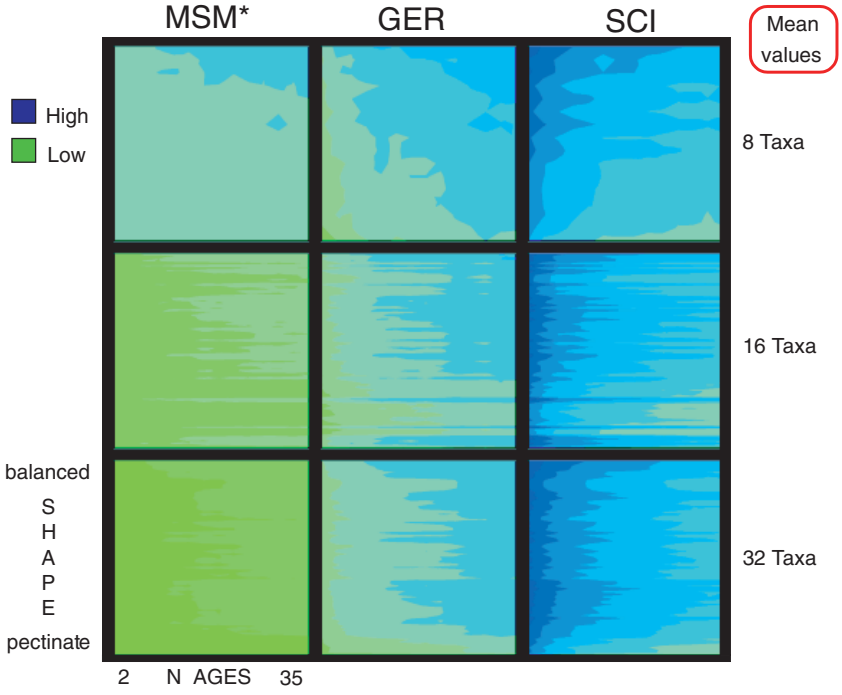

Fig. 3. Color-coded mean values of MSM*, GER, and SCI for the random assignments of ages across taxa. Each box represents a tree space of fixed number of terminals (8, 16, and 32 taxa).

contrast to the decreasing pattern present in SCI values (Fig. 3). Additionally, this trend in mean GER values was more apparent in balanced trees with smaller numbers of taxa (Fig. 4). In our results, the tree shape effect seems to be present across the range of parameter values analyzed here, favoring mean GER values on balanced trees. This difference seems to be greater in larger trees (Fig. 4).

As was the case with SCI, larger pectinate trees consistently depicted lower mean GER values, while larger balanced trees had only slightly lower GER values.

*MSM**

In relation to the parameters under investigation, MSM* behaved most appropriately in relation to unstructured data. Where it showed a slight bias in these randomizations, the MSM* seemed to be affected by tree shape and number of ages in the same way as GER but not nearly to the same degree. This is not surprising, since both are based on a calculation of the number and extent of ghost lineages. Notably, within the range of this randomization analysis, the MSM* was less sensitive to tree shape and to the number of possible ages than was GER (Fig. 3). Except for the pattern seen in small trees (eight taxa), there were minor differences in mean MSM* values when both tree shape and number of ages were varied (Figs 3 and 4). Mean MSM* values quickly decreased with increasing tree size (i.e. number of taxa), as was expected with random data (Siddall, 1998). This trend was present across all tree shapes and all possible number of ages (Fig. 4). However, in large taxon cases, the variations found in the MSM* parameter spaces were very small.

Discussion

As stated above, these indices are mainly used in two different contexts: the comparison of stratigraphic fit of trees with different sets of taxa, and the evaluation of stratigraphic conflict among competing phylogenetic hypothesis of a given dataset.

Comparing trees of different taxa

These comparisons between tress of different sets of taxa have been made in order to test specific hypotheses regarding the fossil record. Some studies have tested whether there is an overall general congruence between stratigraphic and phylogenetic data or if the stratigraphic fit varies between different groups of organisms (Benton and Hitchin, 1996, 1997). Others have tested if the fossil record of more recent geological periods is significantly better than that of older geological periods (e.g. Benton et al., 2000), or if marine fossils tend to

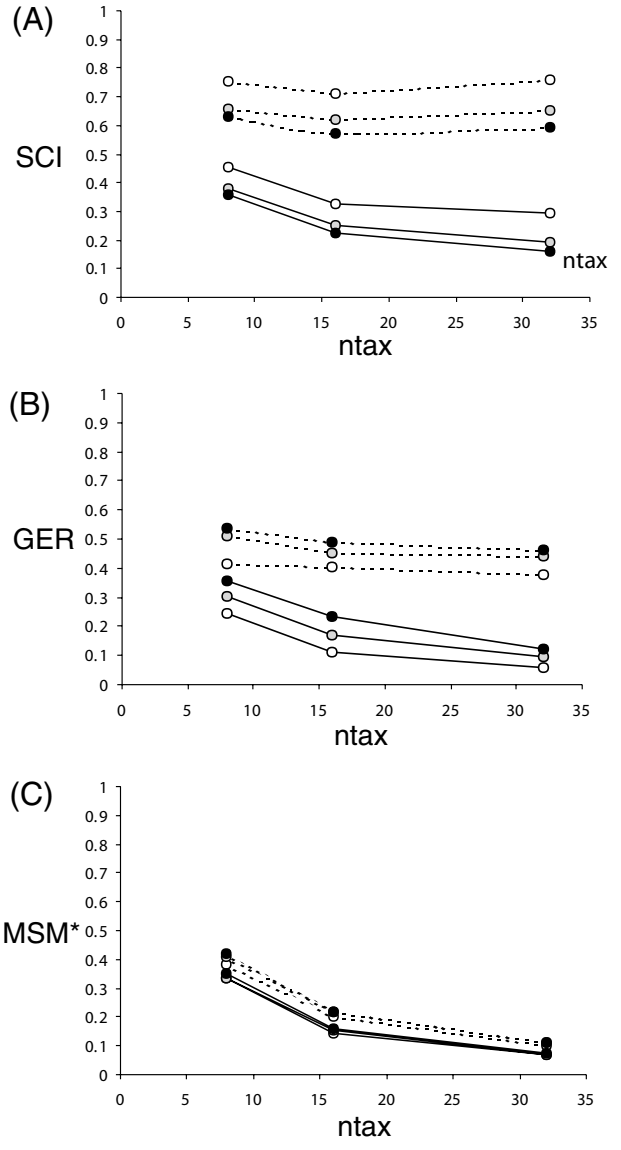

Fig. 4. Sensitivity of mean measure values to variations in number of taxa on fully balanced (dashed line) and fully pectinate trees (solid line) for different number of possible ages (white dots: 4 ages; gray dots: 8 ages; black dots: 16 ages). (A) Mean values of SCI (B) Mean values of GER, and (C) Mean values of MSM*.

have a better stratigraphic fit than terrestrial fossils (e.g. Benton and Hitchin, 1996; Benton and Simms, 1995), or if trees derived from morphology, fit the stratigraphic record better than those based on molecular data (Benton, 1998), or whether the addition of newly discovered taxa increased the congruence between stratigraphic and phylogenetic data (Benton and Storrs, 1994, 1996; Benton, 2001; Weishampel, 1996).

In these studies, cladograms can have different tree shapes, number of taxa, and number of different ages of first appearance. Thus, if raw index *values* are intended to be used for this purpose, we would ideally need a measure that is insensitive to at least these three parameters, as we would be interested in measuring differences in stratigraphic fit rather than differences in tree size, shape or distribution of fossils in geological time (or other parameters; see Wagner and Sidor, 2000).

Our randomization results show that none of the available measures is totally insensitive to some of the explored parameters. These problems are analogous to those noted in homoplasy measures (Farris, 1989; Klassen et al., 1991; Meier et al., 1991; Sanderson and Donoghue, 1989). Furthermore, it seems that in the case of stratigraphic indices, most of these measures cannot control even one of these three parameters (Figs 3 and 4). Thus, our results suggest that comparisons of raw values of these indices are not appropriate for trees differing in tree size, shape, or distribution of ages since these parameters markedly affect their distributions. However, comparisons of raw values are commonly found in the literature (Benton and Hitchin, 1996; but see Benton et al., 1999). If these sorts of studies were conducted, large samples of cladograms may still be compared if they have, on average, a similar number of taxa and tree shape (e.g. Benton et al., 2000). This might alleviate the biasing factors explored here, although these comparisons can still be affected by other parameters (Wagner and Sidor, 2000).

The observed influence of number of taxa for measures of raw amount of conflict (i.e. MSM*) is expected in the face of random distribution of ages, since meaningless information is progressively introduced as the number of taxa increases. However, the magnitude of a statistic (the amount of conflict in this case) and its significance should not be confounded (Siddall, 1998).

Several authors have proposed a randomization procedure analogous to that of the PTP (Archie, 1989; Faith and Cranston, 1991) to obtain significance values for stratigraphic fit measures. In contrast to raw index values, these are directly comparable among trees of different sizes, shapes, and with different age distributions (Huelsenbeck, 1994; Siddall, 1998; Wills, 1999). The aim of this test was simply to determine if the congruence implied by a given tree was significantly different from a random distribution of ages across taxa, rather than if it differed from a realistic biological expectation (Wills, 1999). Holding possible influential parameters constant, such as tree size, shape, and distribution of ages, the significance test intends to yield comparable results for different types of data. Using the *P*-value of a given measure might be a possible solution for studies comparing trees of different sets of taxa (Benton et al., 1999; Wills, 1999). However, the results of such comparisons clearly do not indicate which group has a better stratigraphic fit (in absolute values or according to some evolution-

ary/preservational expectation), but rather, which group has a stratigraphic fit most different from a random distribution of ages.

A final caveat of comparisons of *P*-values is that the three parameters held constant in this test are not the only possible factors that could affect these indices (see Wagner, 2000 and Wagner and Sidor, 2000 for an analysis of the effect on SCI of other parameters that commonly vary among datasets).

Comparing competing trees

These measures can also be used for comparing the stratigraphic fit of alternative phylogenetic trees of the same taxonomic group as a descriptive statistic (e.g. Brochu and Norell, 2000, 2001), or as an optimality criterion (Huelsenbeck, 1994). Despite the divergent philosophical perspectives of these approaches, if raw index values were used for any of these purposes, we would need a measure that was only insensitive to tree shape (since the two other parameters remain fixed in alternative phylogenetic trees of a particular taxonomic group). It is therefore particularly important to know the extent and cause of tree shape sensitivity on a given measure before using it for direct comparisons of competing phylogenetic trees (as a descriptive statistic and especially as an optimality criterion). A more detailed examination of the tree shape effect is given here for each of the three measures.

SCI and tree shape. This measure was thought to favor pectinate trees (Siddall, 1996, 1997; Hitchin and Benton, 1997b). The results presented here, however, indicate that mean SCI values of random distributions are lower for pectinate trees than for balanced trees (Figs 3 and 4). Previous authors correctly suggested that when each terminal taxon has a different age of first appearance, only pectinate trees can reach SCI values of 1, while fully balanced trees will always yield SCI values of 0.5 (Siddall, 1996; Wills, 1999; Wagner and Sidor, 2000). However, under these conditions, certain pectinate trees can also reach SCI values of 0 (Wills, 1999). Therefore, although it seems clear that the maximum SCI is constrained by the topology, the open question is whether the SCI is bound to be higher in pectinate trees? To explore this, the distribution of SCI values for all possible fully pectinate and fully balanced trees was calculated for two eight-taxon trees. These distributions depend upon the number of taxa and the particular distribution of ages across taxa (i.e. number of ages and proportion of taxa with equal age of first appearance). In Fig. 5, the distribution of SCI values are given for fully pectinate and balanced topologies of eight taxa with two different distributions of ages. It is evident from these histograms that most pectinate trees have lower SCI values than most balanced trees (Fig. 5), in accordance with expectations provided by previous authors (Siddall, 1996, 1997; Wagner, 1998; Wills, 1999).

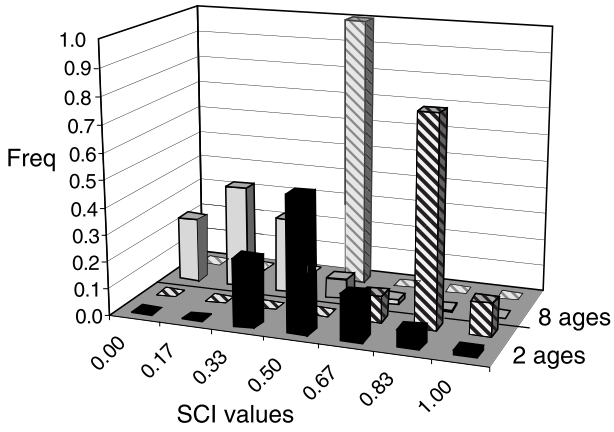

Fig. 5. Frequency histogram of SCI values for all possible distributions of ages on a fully pectinate (solid bars) and a fully balanced (dashed bars) tree of eight taxa. Front bars represent frequencies when only two different ages are assigned among terminal taxa. Back bars represent frequencies when each taxon has a different age of first appearance.

Minimal SCI values will always be lower in pectinate topologies (except for the trivial case in which all taxa have the same stratigraphic age; SCI = 1 for every possible tree). The number of taxa and the distribution of ages determine the specific values of these minima. In general, depending on the distribution of ages, the minimum SCI values of balanced topologies will vary between 0.5 and 1 (contra Benton et al., 2000), while pectinate topologies will vary between 0 and 1. Maximum SCI values, instead, will vary between 0.5 and 1 for balanced trees, and will always be 1 for pectinate trees.

MSM, GER, and tree shape*. GER is more sensitive to tree shape than MSM* under the explored conditions in the randomizations (Figs 3 and 4), even though both measures are based on the same marker of conflict (the length of the age character (L_0) that measures the extension of ghost lineages). MSM* depends on two parameters (L_m and L_0) while GER depends on three parameters (L_M, L_m, and L_0). Since L_M (i.e. g) and L_m (i.e. m) are topology independent, L_0 (i.e. s) must be the parameter sensitive to variations in tree shape. Therefore, both MSM* and GER should be affected by the tree shape sensitivity of L_0. To understand the difference in sensitivity of MSM* and GER to variations in tree shape, we analyzed the frequency distribution of L_0 values for balanced and pectinate trees in random assignments of ages across taxa. This is shown in Fig. 6, for fully pectinate and balanced trees of 32 taxa (for two different numbers of possible ages). The results clearly

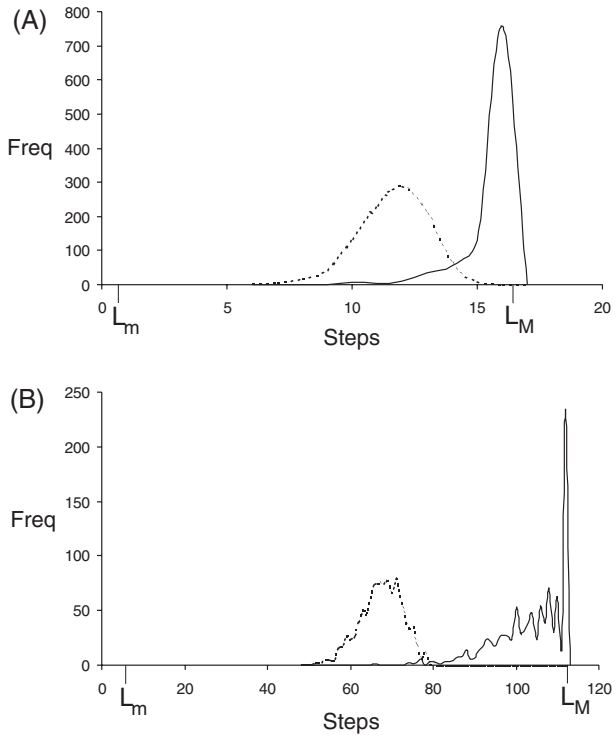

Fig. 6. Frequency of number of steps (L_0) of the age character obtained from 1000 replicates of random assignments of ages on a fully pectinate (solid line) and a fully balanced (dashed line) tree of 32 taxa. (A) Frequency distribution when only two possible ages were randomly assigned. (B) Frequency distribution when eight possible ages were randomly assigned. The minimum (L_m) and maximum (L_M) possible lengths of the age character are indicated on the abscissa.

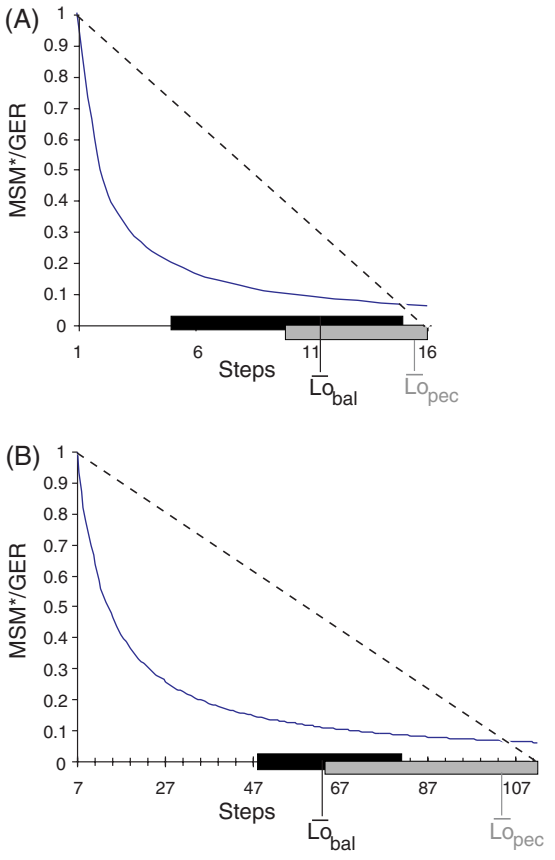

Fig. 7. MSM* and GER values as a function of the number of steps of the age character (L_0) for a tree of 32 taxa. (A) when there are only two different ages. (B) when there are eight different ages. On both graphs, black bars on the abscissa represent the range of L_0 obtained in the randomization procedure described in Fig. 6 for fully balanced topologies. Gray bars represent the range of L_0 obtained in the randomization procedure described in Fig. 6 for fully pectinate topologies. Mean length values of the age character obtained in the randomizations are also indicated for fully balanced (($\bar{L}_{0_{bal}}$)) and fully pectinate trees (($\bar{L}_{0_{pec}}$)).

indicate that, irrespective of the number of states in the age character, L_0 has a strikingly different distribution for pectinate and balanced trees (Fig. 6).

If the distributions of L_0 are so drastically influenced by tree shape, why is MSM* unaffected while GER is so sensitive to tree shape? Considering the range and mean L_0 for each tree shape distribution obtained in the randomizations, it is evident that the reason for the difference in sensitivity between MSM* and GER is a scaling artifact.

GER decreases linearly with L_0, therefore a given difference in length will have the same effect in GER values irrespective of the magnitude of those lengths (Fig. 7). In contrast, MSM* follows a concave-up decreasing function. Here, a given difference in length will drastically influence MSM* values for relatively small L_0 values, but it will only slightly affect MSM* values if the L_0 are large enough (Fig. 7). In the random distributions analyzed here, the range and mean of L_0 is indeed very different for pectinate and balanced trees. However, L_0 values are usually large and differences between mean L_0 values for pectinate and balanced trees are not as evident in MSM* as in GER values (Fig. 7). It is therefore expected that GER distributions are different for pectinate and balanced topologies, reflecting the distribution of L_0, while the two MSM* distributions are similar, only slightly favoring balanced trees (Fig. 8).

It must be noted that the influence of tree shape on distributions of character length for random assignments of states is a particular property of irreversible characters. In reversible characters, the length distributions in pectinate and balanced trees do not differ (Fig. 9), and as far as we have tested, this equality holds, irrespective of the number of states and transformation costs among character states.

In summary, it is clear that any measure based on L_0 will be sensitive to the tree shape parameter. Further-

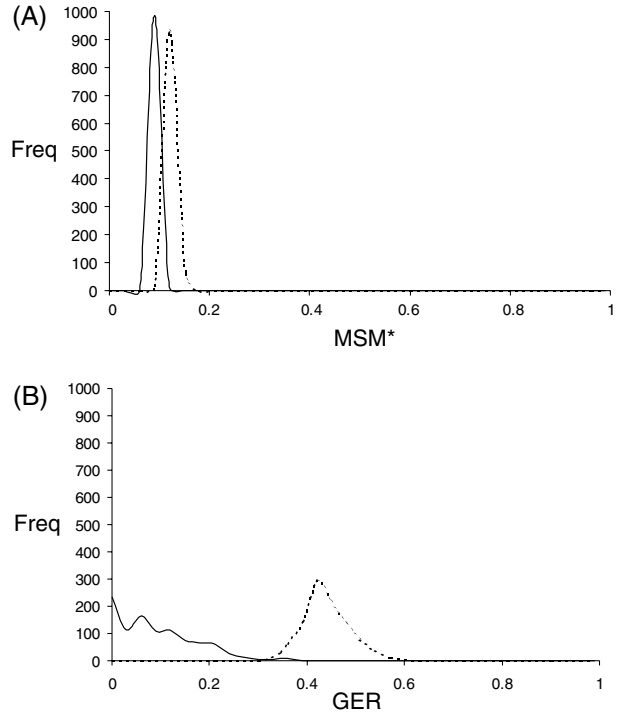

Fig. 8. Frequency of measure values in the randomizations presented in Fig. 6. (A) Frequency distribution of MSM* values for fully pectinate (solid lines) and fully balanced (dashed lines) topologies. (B) Frequency distribution of GER values for fully pectinate (solid lines) and fully balanced (dashed lines) topologies.

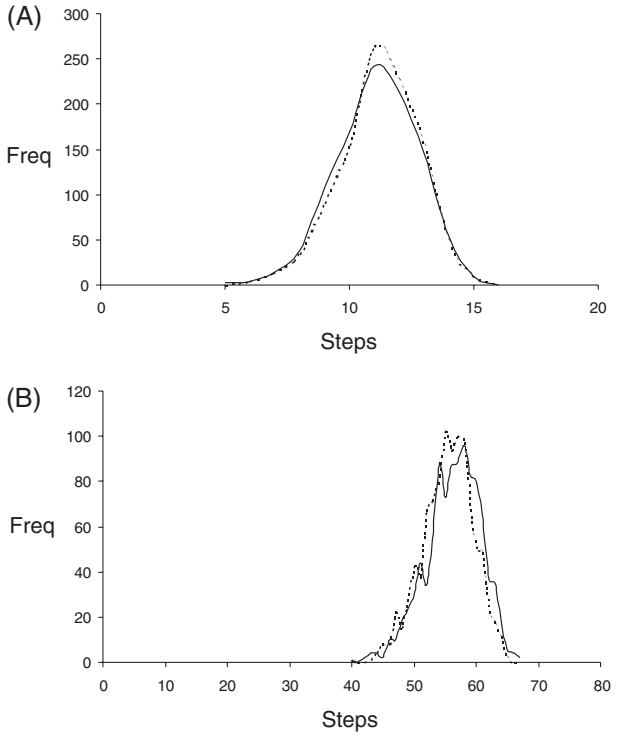

Fig. 9. Frequency of number of steps of a reversible character obtained from a thousand replicates of random assignments of ages on a fully pectinate (solid line) and a fully balanced (dashed line) tree of 32 taxa. (A) Frequency distribution when only two possible ages were randomly assigned. (B) Frequency distribution when eight possible ages were randomly assigned.

more, in some cases (e.g. when each taxon has a different age), L_0 cannot reach either the L_m or L_M on a balanced tree. Therefore, minimum and maximum values of GER and MSM* will not be attainable for balanced trees but they will for pectinate trees. In such cases the discriminatory power of both MSM* and GER for balanced trees will be lower than for pectinate trees, although this difference would not be as drastic as in the SCI.

Summary of tree shape and conflict. It is clear that tree shape has an effect on all of the analyzed measures of stratigraphic fit. Both the frequency distribution and the possible range of variation of a given measure are determined by tree shape. In general, tree balance reduces the range of possible values and therefore reduces the discriminatory power of all measures. This effect is most noticeable in SCI, which shows no discriminatory power for balanced trees under some conditions (Fig. 5).

The difference in behavior of these measures certainly flows from the fact that comparisons between stratigraphic data and phylogenetic trees measure the conflict between two different patterns of signal. One of them, stratigraphy, is necessarily linear, while phylogeny is hierarchical, with nodes arranged in a variety of hierarchical structures (i.e. tree shape). Therefore, we should not expect stratigraphic data to identically fit all kinds of tree shapes (at least as measured by these metrics).

If amount of conflict is the focus of interest, it would be advisable to compare competing trees with MSM* or GER (since they seem to be much less affected by tree shape than SCI and they consider the temporal magnitude of the conflict), although we should be aware that pectinate trees are potentially more suitable for minimizing the conflict. This does not mean that a given measure will always prefer a more asymmetrical tree. However, if the conflict is the minimum allowed by the topology, all of these measures will favor the pectinate topology under most conditions.

The use of a significance test for comparisons of conflict between competing trees would not be advisable since it could favor a tree that implies a greater amount of conflict (measured by L_0 or number of consistent nodes) than a less conflictive tree, just because the nodes of the first one are nonlinearly arranged (e.g. balanced tree shape). This is most evident in the case shown in Fig. 5, in which the SCI of all balanced trees are not significantly different from random.

Conclusions

Our results show that the three analyzed measures have different properties and behave differently with variation in tree size, tree shape, and number of possible ages. Although SCI and GER are more sensitive to these parameters, all three measures show variation and therefore the use of their raw values should not be used for comparing the stratigraphic fit of different taxonomic groups (if they differ in these parameters).

Regarding their use for comparing the stratigraphic fit of competing phylogenetic trees of the same group, we show that these measures can have different minima, maxima, and discriminatory power for different topological shapes. Therefore, it is not expected that the stratigraphic data would have the potential to identically fit all kinds of tree shapes. In general, SCI seems to be much more affected by these problems than GER and MSM*. Differences between the GER and MSM* are based on how they scale the same marker of conflict (i.e. the sum of ghost lineages).

Accepting or correcting the effects of these parameters depends on what are we interested in measuring. Measures that consider just the amount of conflict seem to be unavoidably constrained by tree shape. Modifications of these measures that control the effect of tree shape are possible, although the resulting measures would not be focused on the amount of conflict. Finally, regarding the use of a significance test of these measures, we underscore the drastic difference of the question being asked (see above) and the interpretation of obtained results.

Acknowledgments

We would like to thank P. Makovicky, M. Wills, and P. Wagner for useful comments and discussions on this subject. M. Benton and an anonymous reviewer provided insights that improved the quality of this manuscript.

References

Archie, J.W., 1989. A randomization test for phylogenetic information in systematic data. Syst. Zool. 38, 239–252.

Benton, M.J., 1998. Molecular and morphological phylogenies of mammals: congruence with stratigraphic data. Mol. Phyl. Evol. 9, 398–407.

Benton, M.J., 2001. Finding the tree of life: matching phylogenetic trees to the fossil record through the 20th century. Proc. Roy. Soc. Lond. Ser. B, 268, 2123–2130.

Benton, M.J., Hitchin, R., 1996. Testing the quality of the fossil record by groups and by major habitats. Hist. Biol. 12, 111–157.

Benton, M.J., Hitchin, R., 1997. Congruence between phylogenetic and stratigraphic data on the history of life. Proc. Roy. Soc. Lond. B, 264, 885–890.

Benton, M.J., Simms, M.J., 1995. Testing the marine and continental fossil records. Geology, 23, 601–604.

Benton, M.J., Storrs, G.W., 1994. Testing the quality of the fossil record: Paleontological knowledge is improving. Geology, 22, 111–114.

Benton, M.J., Storrs, G.W., 1996. Diversity in the past: comparing cladistic phylogenies and stratigraphy. In: Hochberg, M., Clobert, J. Barbault, R. (Eds.), Phylogeny and Biodiversity. Oxford University Press, Oxford, UK, pp. 19–24.

Benton, M.J., Wills, P.M., Hitchin, R., 1999. Assesing congruence between cladistic and stratigraphic data. Syst. Biol. 48, 580–591.

Benton, M.J., Wills, P.M., Hitchin, R., 2000. Quality of the fossil record through time. Nature, 403, 534–537.

Brochu, C.A., Norell, M.A., 2000. Temporal congruence and the origin of birds. J. Vert. Paleont. 20, 197–200.

Brochu, C.A., Norell, M.A., 2001. Time and Trees: a quantitative assessment of temporal congruence in the bird origins debate. In: Gauthier, J.A., Gall, L.F. (Eds.), New Perspectives on the Origin and Early Evolution of Birds. Peabody Museum of Natural History, Yale University, New Haven, CT, pp. 511–533.

Faith, D.P., Cranston, P.S., 1991. Could a cladogram this short have arisen by chance alone? On permutation tests for cladistic structure. Cladistics, 7, 1–28.

Farris, J.S., 1989. The retention index and homoplasy excess. Syst. Zool. 38, 406–407.

Goloboff, P., 1993. NONA (a bastard son of Pee-Wee). Version. 1.9. Program and documentation. Published by the author, Tucumán, Argentina.

Goloboff, P.A., 1996. SPA. (S)ankoff (P) arsimony (A) nalysis, version 1.1 (32 bit version). Program and documentation. Published by the author, Tucumán, Argentina.

Heard, S.B., 1992. Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. Evolution, 46, 1818–1816.

Hitchin, R., Benton, M.J., 1997a. Congruence between parsimony and stratigraphy: Comparisons of three indices. Paleobiology, 23, 20–32.

Hitchin, R., Benton, M.J., 1997b. Stratigraphic indices and tree balance. Syst. Biol. 46, 563–569.

Huelsenbeck, J.P., 1994. Comparing the stratigraphic record to estimates of phylogeny. Paleobiology, 20, 470–483.

Klassen, G.J., Mooi, R.D., Locke, A., 1991. Consistency indices and random data. Syst. Zool. 40, 446–457.

Meier, R., Kores, P., Darwin, S., 1991. Homoplasy slope ratio: a better measurement of observed homoplasy in cladistics. Syst. Zool. 40, 74–88.

Naylor, G., Kraus, F., 1995. The relationship between s and m and the retention index. Syst. Biol. 44, 559–562.

Norell, M.A., 1992. Taxic origin and temporal diversity: The effect of phylogeny. In: Novacek, M.J., Wheeler, Q.D. (Eds.), Extinction and Phylogeny. Columbia University Press, New York, pp. 89–118.

Norell, M.A., 1993. Tree based approaches to understanding history: comments on ranks, rules and the quality of the fossil record. Am. J. Sci. 293, 407–417.

Norell, M.A., 1996. Ghost taxa, ancestors, and assumptions: A comment on Wagner. Paleobiology, 22, 453–455.

Norell, M.A., Novacek, M.J., 1992. The fossil record and evolution: comparing cladistics and paleontological evidence for vertebrate history. Science, 255, 1690–1693.

O'Leary, M., 2002. The phylogenetic position of cetaceans: Further combined data analyses, comparisons with the stratigraphic record and a discussion of character optimization. Am. Zool. 41, 487–506.

Pol, D., Norell, M.A., 2001. Comments on the Manhattan Stratigraphic Measure. Cladistics, 17, 285–289.
Sankoff, D., Cedergren, R.J., Lapalme, G., 1976. Frequency of insertion-deletion, transversion, and transition in the evolution of 5S ribosomal RNA. J. Mol. Evol. 7, 133–149.
Sanderson, M.J., Donoghue, M.J., 1989. Patterns of variation in levels of homoplasy. Evolution, 43, 1781–1795.
Siddall, M.E., 1996. Stratigraphic consistency and the shape of things. Syst. Biol. 45, 111–115.
Siddall, M.E., 1997. Stratigraphic indices in the balance: a reply to Hitchin and Benton. Syst. Biol. 46, 569–573.
Siddall, M.E., 1998. Stratigraphic fit to phylogenies: a proposed solution. Cladistics, 14, 201–208.
Smith, A.B., Littlewood, D.T.J., 1994. Paleontological data and molecular phylogenetic analysis. Paleobiology, 20, 259–273.
Wagner, P.J. 1995. Stratigraphic tests of cladistic hypotheses. Paleobiology, 21, 153–178.
Wagner, P.J., 1996. Testing the underlying patterns of active trends. Evolution, 50, 990–1017.
Wagner, P.J., 1998. A likelihood approach for evaluating estimates of phylogenetic relationships among fossil taxa. Paleobiology, 24, 430–449.
Wagner, P.J., 2000. The quality of the fossil record and the accuracy of estimated phylogenies. Systematic Biology, 49, 21–42.
Wagner, P.J. and Sidor, C.A., 2000. Age rank/clade rank metrics—sampling, taxonomy, and the meaning of 'stratigraphic consistency'. Syst. Biol. 49, 463–479.
Weishampel, D.B., 1996. Fossils, phylogeny, and discovery: a cladistic study of the history of tree topologies and ghost lineage durations. J. Vert. Paleont. 16, 191–197.
Wills, M.A., 1999. Congruence between stratigraphy and phylogeny: randomization tests and the gap excess ratio. Syst. Biol. 48, 559–580.