Department of Systematic Zoology
Evolutionary Biology Centre
Uppsala University

TAXON SAMPLING IN PHYLOGENETIC ANALYSIS:

PROBLEMS AND STRATEGIES REVIEWED

JOHAN A. A. NYLANDER

Introductory Research Essay No. 1.

Uppsala 2001

CONTENTS

INTRODUCTION

The amount of data available for phylogenetic analysis has increased significantly during the past decade. Particularly, the advent of new techniques such as the polymerase chain reaction (PCR) has paved the way for a substantial input of molecular data. This has not only led to an increase in the number of phylogenetic analyses, but also to the discovery of new areas of analysis-related problems. One such area was discovered when it became clear that hypotheses of relationships could sometimes be (radically) different depending on what taxa are included (or excluded) in a phylogenetic analysis. For example, if one were to examine the higher relationships in Mammalia using a single representative of each order, the inferred relationship between the orders could change solely depending on which specific taxa that were chosen. Thus, when replacing one perissodactyl ungulate (say, the Sumatra rhino) with another (say, the Przewalski's horse) in an analysis, this could change the position of the Perissodactyla in the tree, regardless of how well the particular group (in this case Perissodactyla) was previously defined. Furthermore, given that the systematist has a choice of which taxa to sample, there is a risk of an investigator bias when deciding which terminal taxa to include or exclude in an analysis (Hillis, 1998). This is truly a reason for concern and as a result, the issue of taxon sampling (which taxa to include or exclude in a phylogenetic analysis) has emerged as one of the most important issues in contemporary systematics (see Hillis, 1998 for an introduction).

Taxon sampling has most commonly been discussed in the context of parsimony reconstruction, and although this review will focus primarily on cases related to parsimony, the taxon-sampling effect (a change in hypothesis caused by a change in representatives sampled) is not restricted to a particular method for tree reconstruction.

The theoretical basis for some of the taxon-sampling effects are now well understood, especially concerning the analysis of molecular data. However, the outcome of an analysis based on a specific sample is still difficult to predict. Consequently, the effects are often difficult to interpret and ascribe to specific underlying causes (sometimes even impossible [Siddall, 1998]). The problems associated with taxon sampling are not unique to the issue of sampling taxa *per se* but coincides with other major reasons for concern in phylogenetic inference. The general problem could be described as follows: *given a sample of taxa and a method of analysis, certain properties of the data could render the reconstruction of the "true" tree* (often the conceived tree for the taxa, see discussion below) *difficult or even impossible*. For the immediate connection to taxon sampling, one could add; *Furthermore, a different sample of taxa could have facilitated the reconstruction, or conversely, even made it more difficult*. The study of taxon sampling is thus rather complex and involves exploration of many sources of error affecting the analysis. Poe (1998*a*: 18) described the task as investigating "what combinations of number of taxa, number of characters, completeness, homoplasy, branch lengths, etc., are optimal for accurate estimation of phylogeny, and how relaxation of the optimality of any of these factors affects the estimate."

WAYS ASTRAY - PROBLEMS IN PHYLOGENETIC RECONSTRUCTION

There are generally two, fundamentally different, reasons why an analysis may lead to erroneous conclusions ("impossible" above). Either we are unable, due to flaws in our methodology, to detect the "true-tree pattern", or the data at hand might actually show a tree different from the "true" or preconceived tree. The first case is related to the fact that methods of inference are sometimes misled by data (e.g., Felsenstein, 1978*a*; Siddall, 1998). The second reason has to do with the disconnection between the levels of inference (see discussion about character trees and taxon trees below). There are several reasons why phylogenetic reconstruction is made difficult. The general problem in phylogenetic analyses is that some characters disagree in the way they indicate relationships. Phylogenetic reconstruction methods run into problems when there is much disagreement in the data. Then character states shared by common ancestry (synapomorphies, Hennig, 1966) are mixed with those that are not. The latter category of characters is called homoplastic characters (non-homologous, Wiley, 1981). They are depicted on a phylogenetic tree as having multiple origins (parallelisms and convergences) or being lost, once gained (reversals). Each character change, homoplastic or not, can be plotted on the branches of a tree. In the same sense as the character changes define groups or clades, they can estimate the lengths of the branches. The fact that character evolution is heterogeneous (the amount of anagenetic change differ among lineages) will often cause branches to differ in length. Furthermore, lineages will often have different branching frequencies (amount of cladogenesis) that will effect the branching pattern. Thus together with the branch lengths, the branching pattern will define the overall shape of a tree. Tree shape has immediate consequences on tree reconstruction.

Tree Shape

It has been shown that the success of phylogenetic reconstruction from a sample of taxa is dependent on the shape of the underlying tree (e.g., Felsenstein, 1978*a*; Rohlf *et al*., 1990; Shao and Sokal, 1990; Debry, 1992; Heijerman, 1993; Mooers, 1995; Huelsenbeck and Kirkpatrick, 1996). The term "tree shape" has been used for describing the symmetry, or balance, of trees but herein I will use shape to include also the heterogeneity of relative lengths of internal and terminal branches (stemminess, Rohlf *et al*., 1990). A tree with long internal branches and short terminal branches is said to have high stemminess (see Fig. 1b, d and e.g., Smith, 1994). Conversely, a tree having short internal branches and long terminal branches (also called star-like, Hillis, 1998 or starburst trees, Cunningham *et al*., 1998) have low stemminess (Fig. 1c, e). A maximally asymmetric or unbalanced tree is one that has a pectinate or comb-like shape (Fig. 1a-c). A balanced tree is one where all terminal branches have the same number of nodes counted from the root; it is thus dichotomously symmetrical (Fig. 1d, e).

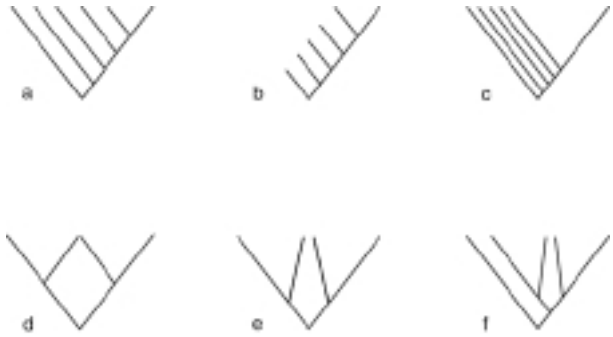

FIGURE 1. Six tree topologies showing different balance, stemminess and heterogeneity in branch lengths. All components together determine the overall shape of the tree. **a** A maximally unbalanced tree with heterogeneity in branch lengths and a constant rate of character evolution (generated by data following e.g. a “molecular clock”). The probability of change in a character is (generally) highest for the most basal branch. **b** An unbalanced tree with equal branch lengths (i.e. the probability for a character to change is the same for all branches). The tree in Fig. 1b has higher stemminess (long internal branches relative to the terminal branches) than tree Fig. 1a. **c** A maximally unbalanced tree with low stemminess (short internal branches relative to the terminal branches). **d** A maximally balanced tree with equal branch lengths. **e** A symmetrical tree with heterogeneity in branch lengths. The tree in Fig 1d has higher stemminess than the tree in Fig 1e. **f** Example of a tree where addition of an outgroup (the most basal branch) has made the reconstruction inconsistent (see also Fig. 2).

If branch-length heterogeneity is high, there is an increasing risk that homoplastic changes on long disjunct branches will coincide (thus erroneously be taken as synapomorphies) and outnumber the changes on the short branches separating them. This results in the long branches being placed together in the reconstructed tree. Furthermore, this means that correct assessment of changes on a short branch (where the number of changes is low) is more critical for correct topological reconstruction than correct assessment of changes on a long branch. An unbalanced tree will generally feature more heterogeneity in branch lengths than a balanced tree. Consequently, unbalanced trees will more often be erroneously reconstructed than balanced (Rohlf *et al*., 1990; Shao and Sokal, 1990; DeBry, 1992; Heijerman, 1993; Mooers, 1995; Huelsenbeck and Kirkpatrick, 1996). The extreme cases of low stemminess where the critical, internal branches are short are some of the most difficult to reconstruct (e.g., Felsenstein, 1978*a*; Hendy and Penny 1989, Huelsenbeck and Hillis, 1993; Hillis, 1998; Halanych, 1998). The case where independent character changes on long branches are falsely interpreted as synapomorphies has received special attention in the literature under the name long-branch attraction (long edges attract, Hendy and Penny 1989: 305). Long branch attraction is currently one of the most debated subjects in theoretical systematics (see e.g. recent discussion of the "Strepsiptera problem" by Carmean and Crespi, 1995; Huelsenbeck, 1997; Whiting *et al*., 1997; Huelsenbeck, 1998; Hwang *et al*., 1998; Siddall and Whiting, 1998; Whiting, 1998). Felsenstein (1978*a*) described circumstances where unequal rates of change caused parsimony (and compatibility methods) to be statistically inconsistent, i.e. converge to an incorrect tree as characters are added to an analysis. Felsenstein described the behavior of the methods as being "positively misleading". The part of the parameter space in which those methods are misleading has later been called the Felsenstein zone [of inconsistency] (Huelsenbeck and Hillis, 1993: 253). Felsenstein's original example contained a four-taxon tree with low stemminess but long-branch attraction has been shown to exist, at least theoretically, for more taxa (Hendy and Penny, 1989; Zharkikh and Li, 1993; Kim, 1996), and for methods other than parsimony and compatibility methods (DeBry, 1992; Huelsenbeck and Hillis, 1993; Gaut and Lewis, 1995; Huelsenbeck, 1995; Yang, 1996). Felsenstein (1978*a*) suggested that when rates are equal (constant) or low, parsimony would always converge to the correct tree. Hendy

and Penny (1989) showed that this was indeed true for four taxa (as in Felsenstein's example) but not for more. They further explained that it was the juxtaposition of long and short branches that caused parsimony to be inconsistent, not unequal rates *per se*.

Character Trees *vs.* Taxon Trees

Unfortunately for the practicing systematist, the true tree of one set of characters (e.g., a particular gene) does not necessarily match the true tree for the taxa (e.g. species) the data were meant to represent (e.g., Fitch, 1970; Avise *et al*., 1983; Pamilo and Nei, 1988). Consequently, there are situations were incorrect inferences are made based on the particular sample at hand even though the inferred tree is correct. Several mechanisms could cause discrepancies between a character tree and a taxon tree (often discussed in the terms gene tree and species tree) such as horizontal transfer, lineage sorting and gene duplication coupled with extinction (reviewed by e.g. Doyle, 1992; Avise, 1994). Inferring relationships at one hierarchical or organizational level using trees from another level is not always straightforward because of the disconnection that exists between the levels (see recent discussions by Maddison, 1997; Doyle, 1997). For instance, single specimens or "semaphoronts" (Hennig, 1966) often serve as representatives for higher levels in phylogenetic analyses. One could easily imagine a situation where, depending on the choice of semaphoronts, the outcome of an analysis would change. Thus, the choice of semaphoronts, and thoughts on what exactly they are meant to represent, is an important step in phylogenetic analysis.

SAMPLING ISSUES

Adding Taxa or Characters?

As for other fields of science, the assumption that more data leads to a stronger or better-corroborated hypothesis is also held among phylogenetic systematists. This assumption is often, explicitly or implicitly, taken as homoplasy or misleading "noise" present in smaller data sets will be overridden by the "true signal" when more data are added (e.g., Cummings *et al*., 1995; Nixon and Carpenter, 1996: 228 in footnote, see also Mindell and Thacker, 1996; Lecointre *et al*., 1994). However, when it comes to characters, it has been shown that under special conditions the opposite is true: the more data we add, the more certain we are to find the wrong topology (e.g., Felsenstein, 1978*a*; Hendy and Penny, 1989). This special condition equals adding characters in the Felsenstein zone. Swofford *et al*. (1996: 427) stated that "[when in the Felsenstein zone] the only hope of getting the correct tree is by sampling few enough characters that we may be lucky enough to obtain more of the character patterns favoring the true tree than of the more probable character patterns favoring the wrong tree." This reasoning supposes that the data added is of the same quality as the data first used. Thus, if the data analyzed was a fast-evolving ("fast-clock") gene, adding another fast gene will guarantee an inconsistent result. On the other hand, if one added a "slow'" gene, or hundred slow genes? One could, at least imagine, a situation where we are in the Felsenstein Zone after analyzing one set of data, then by adding an amount of "good" data we will get the correct answer (see also Huelsenbeck *et al*., 1996). However, regarding the way systematists of today gather data (see e.g. Cummings *et al*., 1995) and that genes may not easily be divided into slow or fast (or "good" or "bad"), "character sampling out of the Felsenstein zone" might not be a realistic alternative. Expanding the term data to include taxa is then an alternative approach. Graybeal (1998: 9) tried to answer if it was

"better to add taxa or characters to a difficult phylogenetic problem". In her study, she started with a tree in the Felsenstein zone, and examined whether the addition of characters or taxa had the greatest influence on phylogenetic accuracy. Not surprisingly, since taxa were added judiciously to intercept long branches, the addition of taxa was most important. This was also true even when the total amount of characters was held constant (i.e., the data with the greatest number of taxa had the least number of characters per taxon). It would be interesting to know how general the conclusions are, i.e. how commonly the suggestion applies to real phylogenetic problems. Given that an investigator is in a situation where time and funding are limited, and that there is a trade-off between gathering taxa or characters, were should she put the effort? Although very few investigations has yet included calculations on the actual effort spent on the data (measured in time and money), there are some papers dealing with the issue of what is the most important factor in phylogenetic analysis. However, surveying the literature to seek guidance in such matters, one might quickly become disoriented. Different authors reach apparently different conclusions. Russo *et al*. (1996: 525) in their evaluation study of different reconstruction methods writes: "The most important factor in constructing reliable phylogenetic trees seems to be the number of amino acids or nucleotides used". Givnitsch and Sytsma (1997: 320) concludes from their simulation study that "[the] probability of correct phylogenetic inference increases with the number of variable (or informative) characters and their consistency index and decreases with the number of taxa." On the other hand, Wheeler (1992: 205) reaches the conclusion that "the number of taxa used [...] is the most important factor in cladogram accuracy." However ambiguous this situation seems to be, behind the statements lies no actual disagreement. Each conclusion is well founded within each individual study, and hence, instead of being contradictory it merely catches the very essence of the difficulty in generalizing on this matter. There are some general statements that can be made such as "most accurate reconstruction are based on a large amount of characters - if the data is of high quality". However true such generalizations might be, they are perhaps not very useful for the practicing systematist. Even if we tried to make a fair generalization such as: "add taxa to break up long branches causing erroneous reconstruction - add characters otherwise", this might not get us much further. We are still left with difficult questions such as how to identify whether our results are right or wrong, in the first place. One other relevant issue for adding taxa or characters that sometimes is mentioned is the concern for the ratio between the numbers of taxa and characters. There is certainly a ratio (or value) where the amount of characters is too low to allow tree-reconstruction methods to discriminate between different tree topologies (e.g., Erdös *et al*., 1997; Kim, 1998). However, one aspect the researcher should consider is what to expect, or "demand", from the analysis. A large number of taxa could be grouped perfectly well with only one two-state character. Of course the expected resolution is not high, but might be satisfactory for some purposes. A completely different issue is of course whether a clade based on one character change could be considered well supported (see e.g., Felsenstein, 1985).

Adding Taxa *and* Characters?

Adding both taxa and characters means that the overall size of the data matrix increases. The question is then to what extent the properties of the data change with overall size of the matrix, or perhaps more direct, does the chance of finding the correct topology change with the overall size of the matrix? Farris (1997: 304) wrote "it is sometimes suggested that large matrices do not need to be analyzed: analyses of a few selected taxa will do as well. This idea has a pitfall [...]". He referred to the situation where crucial taxa are absent from the analysis, and therefore erroneous results could be reached. Farris showed that by simply adding taxa to an analysis, "including all relevant information", inconsistencies caused by incomplete taxon sampling could be avoided (see also Gauthier *et al*., 1988; Donoghue *et al*., 1989; Farris *et al*., 1996). The crucial step is obviously to sample the "relevant information". If we focus on taxa, which taxa do we need to include in order to get the correct results?

Should we sample as many as possible, just in case, to try to avoid the pitfalls? Furthermore, could there be pitfalls associated with that strategy as well? Naylor and Brown (1998: 71) regarded it as a question of method used, they stated "More data are better than fewer data only when the inference model accommodates, in an unbiased way, the evolutionary forces that have shaped character-state distributions. Any disparities (biases) that exist between a model (implied or explicit) and the evolutionary process will be magnified with increasing amounts of data." Hillis (1996) simulated character data on a 228-taxon tree and compared the success of point-estimation methods (stepwise addition under the parsimony criterion, neighbor joining under the minimum-evolution criterion and UPGMA) in reconstructing that tree. Perhaps contrary to what many would have expected, the methods did well (except UPGMA) despite the quick-and-dirty methods, and the relatively "few" characters (5000 nucleotide positions) used. Hillis suggestion to explain the success was that homoplasy in the data where able to be distributed over the many branches of the tree, thus making "covarying patterns of homoplasy between any two taxa [...] relatively rare." Hillis suggested also that "adding large numbers of additional taxa to phylogenetic analyses may increase the accuracy of the estimated trees and at the same time reduce the need for computationally complex methods of analysis." Purvis and Quicke (1997) proposed that Hillis's study avoided inconsistency by adding taxa to shorten branches (hence avoiding long-branch attraction). They did further simulations with higher evolutionary rates than Hillis had used. According to their results, stepwise addition continued to do well and performance started to decrease only when "rates rose by a factor of 20, by which time the sequences showed virtually no obvious homology". On the other hand Kim (1996: 372), based on his simulation studies concluded, "if the evolutionary question of interest does not require a large number of taxa, it seems best to use fewer taxa because larger trees are more likely to contain inconsistent branches." In short, Kim showed that when the number of taxa increases, so does the probability of inconsistently estimating an internal branch. This occurs because the number of internal branches increases with every addition of a taxon. Hillis (1998) suggested that the seemingly contradictory advice given on large phylogenies stemmed from differences in how researchers evaluate "accuracy", "consistency" and "cladogram success". Kim (1998*a*; 1998*b*) goes in to depth in order to clarify the differences between those expressions. He gives an excellent overview on the complexity of the problem of assessing performance measures for phylogenetic reconstruction methods, and for the difficulties in generalizing from one size or level of analysis to another. He holds the position that whether or not the addition of taxa has a general positive effect on accuracy, is still an open question. After all we cannot tell if e.g. the accuracy found in Hillis's 228-taxon example "is due to taxon sampling or whether such easily estimated [large] trees are common" (Kim, 1998*b*: 26). This possibility must be further explored before we can make general suggestions on sampling large numbers of taxa.

Besides the theoretical considerations about relative success in reconstructing large phylogenies, the size of the data matrices has further implications. Larger numbers of both characters and taxa means that calculations get more demanding and time consuming. An increase in the number of taxa means a dramatic increase in the number of possible trees (Felsenstein, 1978*b*), thus creating potential problems for tree-reconstruction methods (e.g., Penny *et al*., 1992; Swofford *et al*., 1996). Some reconstruction methods, due to the complexity involved in calculations, are very limited in the number of taxa that could be analyzed. A method such as maximum likelihood is by today's algorithms and computers limited to perhaps 100 taxa, while others, such as parsimony jackknifing (Farris *et al*., 1996) allows inclusion of many more (see Källersjö *et al*., 1998 for an analysis of 2538 taxa). Obviously, as more efficient ways to analyze data are developed, the limits on capacity are pushed ahead. Soltis *et al*. (1998) described a promising feature of large data sets. They suggested that a possible solution to the computational problems associated with large numbers of taxa was to increase the numbers of characters. Increasing the number of characters has the potential of increasing "signal" in the data, thus making the reconstruction easier. In their single example, a data set of 190 angiosperm taxa, they found that combining several genes into a single matrix (total size 4733 bp) reduced the

computing time in comparison with analyses based on individual genes. Although some researchers approach phylogenetic problems in the most reduced ways (e.g., Lake, 1987; Graur *et al*., 1996), clearly, reconstruction methods of the future will have to be able to analyze large data sets.

Deleting Taxa and/or Characters?

Deleting data, whether it is characters or taxa, has always been a controversial issue in systematics. The fact that different characters (individual characters, classes of characters or even sources of characters) have different "quality" when it comes to inferring phylogenetic trees, has led researchers to seek to sort "bad" data from "good". In so doing, the question arises how to treat the "bad" characters. Is it legitimate to exclude those characters? One group of authors strongly holds the position that, in essence, "all characters are evidence, and evidence shouldn't be excluded" (e.g., Kluge, 1989; 1997*a*; 1998; Kluge and Wolf, 1993; Nixon and Carpenter, 1996; Wenzel, 1997), and even that deleting data is an "unacceptable *ad hoc* protection of an hypothesis from a legitimate test" (Wenzel, 1997: 31). The argument goes that since we (usually) have no knowledge of the "true" tree, there is no objective method for making *a priori* decisions on what data should be considered reliable or not for inferring that tree. Other authors (e.g., Swofford *et al*., 1996) hold the position that subjectivity in systematics never can be entirely avoided. And in order to make the best possible inferences, we should be able to exclude "unreliable" data from our analysis; "the benefits of excluding clearly unreliable regions [i.e. data] - *however subjectively determined* - outweigh the dangers [my emphasis]." (Swofford *et al*., 1996: 500). The same reasoning as for characters is equally valid for taxa. All taxa could be viewed as data in that they consist of evidence of relationships. Nixon and Carpenter (1996: 233) advocate the total-evidence approach (Kluge, 1989) "where possible all relevant taxa should be used". This viewpoint is nothing that anyone would dispute. "However" they continue, "the addition of terminals may present other non-trivial difficulties, and researchers must make decisions about inclusion based on fiscal and time constraints, and presumed relevance to the outcome of the analysis... Thus, one might be able to justify exclusion of some proportion of the species of a clade, assuming that inclusion of those terminals would not change the outcome." If one is willing to take another philosophical standpoint, their last statement could be changed and held as an argument that "one might be able to justify exclusion of some proportion of the species of a clade, assuming that inclusion of those terminals *would* change the outcome [to the worse]". A taxon could then be excluded before or after tree reconstruction. The *a priori* approach would be based on detailed knowledge of the taxa at hand, which could predict the influence of the inclusion/exclusion of individual taxa (see e.g., Lyons-Weiler *et al*., 1996; Lyons-Weiler and Hoelzer, 1997; Lyons-Weiler *et al*., 1998). The *a posteriori* approach necessarily involves comparisons with an expected result. For instance, the monophyly of a particular clade could be one expected outcome (Nixon and Carpenter's statement above). We might have a situation where inclusion of a particular member of a "well-defined clade" would break up that clade and introduce error into the analysis. Obviously this is a situation where we have "bad" data since the clade was "well-defined" given some other set of data. A rational step would be to gather "better" data, but given that "researchers must make decisions about inclusion based on fiscal and time constraints, and presumed relevance to the outcome of the analysis", exclusion of taxa (i.e. data) may be warranted.

Not Adding

We have an idea about the circumstances, under which the addition of characters does not help our analysis (e.g., adding characters in the Felsenstein zone). But are there situations where

adding taxa is not preferable (or even possible)? Several authors (e.g., Hendy and Penny 1989; Smith, 1994; Kim, 1996; Milinkovitch *et al*., 1996; Graybeal, 1998; Poe, 1998*a*; 1998*b*; Poe and Swofford, 1999) have discussed the situation where the addition of a taxon could cause a phylogenetic problem to become inconsistent. Each added taxon is a potential long branch or has the potential to create heterogeneity in branch lengths. Poe (1998*b*: 1089) explains: "Lineages that previously were not spuriously attracted to each other could become "long" in a relative sense by virtue of the shortening of another branch on which the added taxa connect." (see Fig. 2). An often-used example (e.g., Hendy and Penny, 1989) is when a single taxon is added to root a tree (see Fig. 1f and Fig. 2).

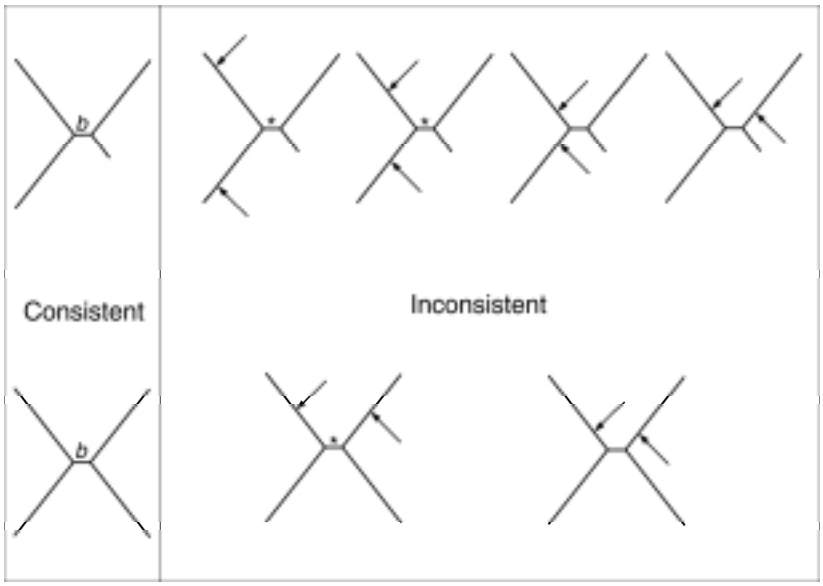

FIGURE 2. Examples of where addition of taxa to a four-taxon tree creates an inconsistent phylogenetical problem (after Poe and Swofford, 1999). Note that taxa are specifically added in order to break up the long branches in the two trees located in the left box. Arrows on the trees in the right box indicate the position where taxa are added. Trees with the internal node marked with an asterisk (*), becomes inconsistent when the internal branch *b* is much shorter than the longest terminal branches (a factor of ten in the examples of Poe and Swofford, 1999).

Is the risk of creating an inconsistent problem a reason for not adding more taxa to an analysis? Could the pitfalls in the two strategies "add taxa to avoid inconsistencies" and "don't add, You might create inconsistent situations" be held against each other? Poe (1998*b*) suggested the possibility that the conditions under which adding taxa damages preexisting relationships in a tree may be rare. He found that although adding taxa could cause a decrease in accuracy (correct reconstruction of a clade), the decrease did not generally affect original relationships. Kim (1996) argued for being cautious, and of course, all problems in phylogenetic analyses are surely not helped just by adding taxa. One obvious situation is when appropriate taxa simply don't exist. Either there are no taxa to include (e.g. due to extinction) or the potential additional taxon lacks the desired properties to make the reconstruction consistent. This even though they are placed on a long branch (see Fig. 2 and Hendy and Penny, 1989; Kim, 1996; Poe and Swofford, 1999). There are other issues than consistency that could affect the decisions of whether to add taxa or not. Each added taxon could be seen as a test of a previous set of hypotheses of relationships. A relationship withstanding such a test would be a better-corroborated hypothesis (Farris, 1983; Steppan, 1993; Lecointre *et al*., 1993; Smith, 1994; Kluge

1997*a*). Adding many outgroup taxa allows putative ingroup taxa a number of alternative attachment positions. This might prevent unintentionally constraining taxa to the ingroup because of insufficient taxon sampling (Steppan, 1993). This idea has been practiced when, for instance, multiple outgroups are added in order to test the monophyly of an ingroup (e.g., Kooistra *et al*., 1993; Nylander *et al*., 1999). Another example is when exploring the relative support (according to some index) for a particular group (e.g., Lecointre *et al*., 1993; Siddall 1995; Milinkovitch *et al*., 1996). However, there is no reason to think that all taxa constitute equally good tests. For example, when testing for monophyly, taxa should become decreasingly relevant as they become more distantly related to the group being tested. On the other hand, in a situation where a very distant taxon, say a monocot plant, ends up within the tested group, say a group of vertebrate animals, that result could be as interesting as any other. Not as a test result for monophyly but it could ultimately be viewed as a "test" of the data and/or the methods used, and are definitely worth further investigations.

Representing Higher Taxa

A special case of the taxon-sampling problem is how to represent supraspecific taxa in analyses (Arnold, 1981; Yeates, 1995; Bininda-Emonds *et al*., 1998; Wiens, 1998). If one is interested in relationships among larger groups (as in the hypothetical example given in the introduction), one might be forced to analyze a smaller number of terminals rather than including all species from all of the supraspecific taxa. The supraspecific taxon, or the clade (assuming a monophyletic taxon, see Bininda-Emonds *et al*., 1998), could then be collapsed down to a single terminal, to be used in a consecutive analysis. The issue is then how to best represent, or retain all the information from all its constituent species. Bininda-Emonds *et al*. (1998) discussed three main strategies, which they called the Ancestral, Democratic and the Exemplar method (See Fig. 3a-c). In the Ancestral method (Fig. 3a), the ancestral states for the taxon in question is inferred, either from fossil data, ontogenetic evidence, or based on previous phylogenetic analysis. These ancestral states in then taken as the groundplan for the taxon and compared with similar groundplan characters for other higher taxa. In the Democratic method (Fig. 3b) the representative state is chosen based on character state frequencies within the higher taxon. Either, the most commonly occurring state among the terminals within the supraspecific taxon is taken as the state of the higher taxon ("Common equals primitive", see Estabrook, 1977). In this way characters variable within a supraspecific taxon need not be scored as polymorphic. An alternative way is to use only those characters with the same state in all members of the supraspecific taxon (also called the "fixed-only" method, Wiens, 1998). The overall number of characters that lends itself to be used in this coding method is, however, smaller, the exact number depends on the size of the group and on the rate of character evolution within the group. The Exemplar method (Fig 3c), the most used method in molecular systematics by far (Bininda-Emonds *et al*., 1998; Hillis, 1998), uses specific taxa or even individuals as representatives for the higher taxon. One or several exemplars of each higher taxon could be used in an analysis.

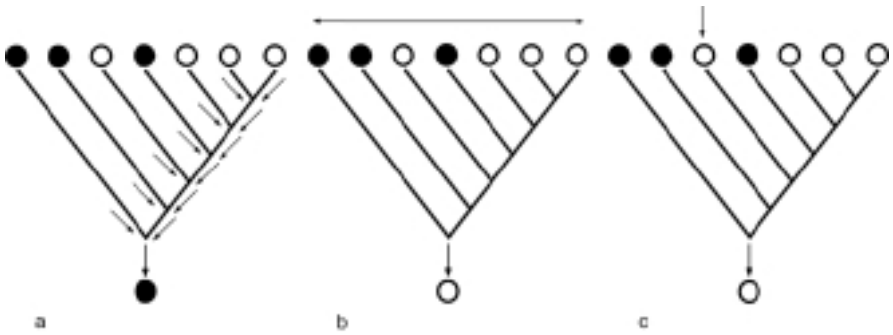

FIGURE 3. Three examples of strategies for representing higher taxa. **a** The Ancestral method, where the state(s) for the ingroup node is inferred. The figure shows a parsimony reconstruction of the ancestral state for the whole clade (the "supraspecific taxon"). Only one character is shown, with states represented by open (○) and filled (●) circles. Parsimony gives the correct assessment for the ancestral condition (filled circle, ●). **b** The Democratic method. Here practiced using the principle of "Common equals primitive". This method fails to derive the ancestral state in this example since the tree is maximally unbalanced. **c** The Exemplar method. The sensitivity of choosing a single terminal is here demonstrated by choosing a taxon possessing character state "open circle" (○). This results in an incorrect assessment of the ancestral state.

With real data, i.e. data that invariably contains homoplasy, the methods could differ substantially in performance depending on exactly how they are conducted. Of the different ways to deduce the states for a taxon using the Democratic method, the principle of "Common equals primitive" performs best (Bininda-Emonds, 1998; Wiens, 1998). This method succeeds in a variety of cases but fails if the tree is pectinate or comb like (see Fig. 3b and Watrous and Wheeler, 1983). The Democratic method, however, performs worse than both the Ancestral and the Exemplar method. The Exemplar method seems to be extremely sensitive to the specific species sampled, especially if single terminals are used. The outcome of an analysis becomes even more unpredictable when branch lengths on the real tree are long, or when representatives are sampled from a large group. These two last cases are in essence the same, both affecting the success of finding the correct tree (see below and Hillis, 1998; Wiens, 1998; Poe, 1998*a*; Rannala *et al*., 1998). Including only the most basal taxa (Yeates, 1995) could be a good strategy, since they often retain a large proportion of primitive or plesiomorphic states. However, which ones that are considered basal are dependent on the previous analysis, and in particular, the choice of outgroups in that analysis. Furthermore, the basal taxa could have undergone substantial character evolution (evolved autapomorphic changes), thus sharing a large number of character states with other taxa different from its ingroup relatives or its sister taxon. The Ancestral method seems to perform more consistently than using the Exemplar method with single representatives (e.g., Bininda-Emonds, 1998; Wiens, 1998). However, there are conceptual issues to be dealt with regarding how to infer the ancestral states in the best way possible. A best estimate of groundplan characters must arguably be when all relevant information has been considered. According to the philosophy of total evidence (Kluge, 1989) this is best accomplished when adding as much information as possible (here, information is taken as both taxa and characters). If e.g. we would have based our groundplan inferences on previous phylogenetic analyses containing only a few taxa, we might not have derived the most appropriate groundplan states. Furthermore, the monophyly of the collapsed taxon could never be tested in a consecutive analysis. If one is uncertain of the status of the chosen exemplars (if they are good representatives or not), or even if the status of the monophyly of the supraspecific taxon is in question, a better alternative seems then to include several exemplars. In that way both the monophyly and the placement of the ingroup node could potentially be tested.

MEASURING AND PREDICTING TAXON SAMPLING SENSITIVITY

Poe (1998*a*: 19) wrote that "the sensitivity of a data set to sampling is the degree to which adding or removing taxa changes the estimate of phylogeny". This statement could be extended to include not only changes in tree topology *per se*, but also changes in other inferences based on the data. Could sensitivity of a data set to taxon sampling be predicted? We might expect that "more homoplastic or weakly supported trees would be more sensitive to the effects of taxonomic sampling" (Poe, 1998*a*: 19) and that well supported groups are less sensitive to species composition than unsupported and weakly supported groups (e.g., Winnepenninckx *et al*., 1995). Swofford *et al*. (1996: 498) argued that the exclusion of a problematic taxon from an analysis would "frequently make a disproportionate change in a measure of tree quality [such as the estimated homoplasy of the parsimony tree]". "However", they continued, "such measures are correlated with the number of taxa in an analysis, so one must confirm that the change in a given statistic is significantly greater than would be predicted by the removal of an average taxon". Swofford *et al*. pointed out one of the critical problems: how to separate the effect of sampling from other factors influencing the reconstruction. Another task is how to define and measure tree quality. Several approaches have been used to explore the effects of taxon sampling, and some are briefly reviewed below.

Using the Jackknife

Lanyon (1985) proposed using the jackknife-resampling scheme (Tukey, 1958) in order to detect internal inconsistencies in a data set. He used a first-order jackknife where for each pseudoreplicate one taxon is dropped and a tree is calculated. This is repeated a (large) number of times and the resulting trees from each pseudoreplicate are saved and a strict consensus is finally calculated. This gives an indication of how sensitive a particular data set is to the exclusion of taxa. Furthermore, Lanyon even considered the consensus as preferable to the most parsimonious tree. Siddall (1995) elaborated further on using the first-order jackknife for evaluating sampling sensitivity. He concluded that Lanyon's proposition of preferring the consensus was flawed (the strict consensus being too reductionistic, see example in Siddall, 1995: 40) but that the jackknife was still useful for investigating clade stability. Siddall proposed the Jackknife Monophyly Index (JMI) for measuring the stability of particular clades to taxon removal. The JMI summarizes the proportion of occurrence for a particular clade among the jackknife pseudoreplicates, in a sense comparable with the bootstrap proportions (Felsenstein, 1985) when resampling characters. Furthermore, Siddall examined for each pseudoreplicate the tree length, ensemble retention index (RI, Farris, 1989) and number of most parsimonious trees for the remaining taxa. This is equal to a sensitivity analysis on those parameters investigated, and opens up further possibilities for critical evaluation of the data. Siddall categorized taxa which substantially increased the number of most parsimonious trees when removed, as "critical", and those that reduced the number as "problematical". A researcher, when having such information, could e.g. critically reexamine the problematic taxa. Some researchers would even consider the possibility of removing certain taxa from the analysis in order to facilitate, or stabilize the analysis. Siddall pointed out that tree lengths are not strictly comparable since data sets differ for each pseudoreplicate. Thus, he focused the discussion on the number of trees and left the RI values uncommented. However, none of the parameters investigated by Siddall is independent of the character matrix, and the same reasoning applied by Siddall to the number of trees could be applied to both tree lengths and RI (Nylander, in prep/unpublished data).

Using the Bootstrap

Siddall (1995, see also Felsenstein, 1985) mentioned using the nonparametric bootstrap (Efron, 1979) for resampling taxa in order to evaluate sampling sensitivity but concluded that the approach was problematic since there simply is no logical connection between the resampling scheme used and what it is meant to estimate. Bootstrapping characters seems to be more justifiable (Felsenstein, 1985; but see e.g., Kluge and Wolfe, 1993) and character bootstrap values have been used to evaluate taxon-sampling effects. Lecointre *et al*. (1993) measured the impact of taxon sampling on bootstrap values for clades in a tree. They used an approach in which subsets of a larger set of taxa were randomly sampled, and then bootstrapped. The corresponding bootstrap values for particular groups were then compared to the corresponding values obtained from the complete data set. They showed that by sampling few representatives from a (presumed) monophyletic group, very different bootstrap values for the monophyly of this group were obtained. Values could vary as much as from 45% (hence not present in a bootstrap majority-rule consensus) to 100% depending on which taxa were chosen. Furthermore, they found no tendency of stabilization of the fluctuation in BP for particular nodes over samples, as more taxa were sampled. This has the important implication that relationships that are highly supported under one set of taxa can receive low support in another, and vice versa. Furthermore, since the variation in bootstrap values was most pronounced with few taxa sampled, this led them to conclude (p. 210) that "phylogenetic studies based on 4 (and maybe even 16) species are insufficient (*whatever the statistical properties of the sample*) and that their conclusions are far from convincing [my emphasis]". For clarification, by "statistical properties" they meant bootstrap values. However, in their approach Lecointre *et al*. were constrained in having the complete tree to compare with, or to use "presumed monophyletic groups". However well corroborated their example trees are, there are pitfalls in extending the conclusions from their data. There is the trivial possibility that preconceived groups are truly not monophyletic (see their discussion, p. 222, and Bininda-Emonds *et al*., 1998) or that other trees than the one reconstructed from the complete data set is the correct one (Poe, 1998*a*; see also Naylor and Brown; 1998). Moreover, their results need not necessarily be a definite judgement of phylogenetic analyses based on low numbers of taxa. Obviously there are at least some phylogenetic problems that could be "well-supported" with only a few taxa, although Lecointre *et al*'s study did not support that notion. Furthermore, their statement about the "unconvincing" results with few taxa is in apparent contrast to e.g. Poe (1998*a*: 18), who suggested that "subsampling of taxa is probably not an important problem for most phylogenetic analyses using up to 20 taxa" (although Poe did not use an approach based on the bootstrap, see below). Anyway, Lecointre *et al*. highlight the very important question how to judge if a group is reliable or well supported. According to them, these concepts should include some notion of robustness. Hence it follows that a group should be regarded as well supported if it apart from having a high bootstrap value (or any other measure of branch support), also is robust to inclusion or exclusion of taxa (see also Siddall, 1995). Milinkovitch *et al*. (1996) also analyzed how taxon sampling affected bootstrap values for individual nodes. Particularly, they focused on the impact of outgroup choice on bootstrap values for three alternative hypotheses (the difference between the three was merely due to alternative positions of the root). They conducted a sensitivity analysis for seven parameters frequently varied in phylogenetic analyses, such as transition/transversion weighting and weighting of codon positions. They found that, by changing outgroups, conflicting rootings could both have high bootstrap values. They suggested (under the assumption that congruence among many outgroups supporting the same hypothesis is a reliable indicator of proper rooting) that if confidence should be put on bootstrap values for rooting hypotheses, either "proper weighting" should have been applied, or the taxa included should "span extensively and homogeneously enough the variability of the group under scrutiny". However, examining Milinkovitch *et al*.'s results shows that the effect of changing outgroup is generally low (except when excluding "informative" taxa or both informative and "redundant" taxa, see Milinkovitch *et al*., 1996: 1822, and

below). When results changed from favoring one rooting to favoring another, the change was most often from moderate (70-80%) values for the first hypothesis to low (50-60%) for the alternative. Using the same data as Milinkovitch *et al*., Milinkovitch and Lyons-Weiler (1998) found that not only could a shift in outgroup result in alternative rootings but also a change in ingroup topology. That conflicting rootings can simultaneously have high bootstrap values is evident from the study of Milinkovitch *et al*., interestingly would be to know if this also is the case for the alternative groupings found within the ingroup. Unfortunately Milinkovitch and Lyons-Weiler's paper does not address this question. However scrutinizing the results of Milinkovitch *et al*. could give some clues. While Milinkovitch *et al*. recorded bootstrap values for alternative rootings, they also recorded values for what they called "nonsensical trees". Those trees were the ones supporting other rootings than the tree originally defined. An additional rooting hypothesis necessarily meant a change in topology more than just the change in the placement of the root (Milinkovitch *et al*., 1998). No nonsensical trees though, receive higher bootstrap values than around 30% (p. 1827, their Figure 5.). Thus their results seem to suggest that if an alternative topology is supported by a change in outgroup composition, it is generally poorly supported by the bootstrap. There are other instances where high bootstrap values result for supposedly spurious groupings (e.g., Naylor and Brown 1998), but those are presumably caused by other factors than the choice of outgroup.

Using Tree Metrics

Poe (1998*a*) tested 29 data sets taken from the literature for their sensitivity to taxon sampling. Besides exploring how sensitive data sets were in general, he explicitly wanted to find out if number of taxa, number of informative characters, total support index (Bremer, 1994) or tree symmetry could be taken as predictive indicators of sensitivity to sampling. As a measure of sensitivity, he used the difference between the tree length of a culled matrix on the most parsimonious tree of the original data (with the culled taxa pruned) and the tree length of the most parsimonious tree of the culled matrix (see further Poe, 1998*a*: 19) A high tree-length difference would indicate a higher sensitivity to taxon sampling for that matrix. He found that sensitivity was generally low for the range of parameter values he investigated, and that smaller data sets were usually more resistant to the effects of sampling. Furthermore, by doing a multivariate regression analysis he found the total number of taxa to be a significant predictor of sensitivity to taxon sampling. None of the other variables were found to be significantly correlated with sampling sensitivity (although retention index and number of informative characters were both nearly so). This last result has the implication that e.g. having highly supported trees or highly congruent characters can not be taken as guarantees for robustness against changes in taxon composition. Since Poe only selected data sets that strictly met certain conditions (such as parsimony analysis would yield one, fully resolved tree, and that all or all except one of the known extant members of the study group were included), this puts limitations on how his results could be generalized. On the other hand, he had the opportunity to draw some very interesting conclusions based on his material. He found that the relationship between the percentage of a clade sampled and sensitivity to sampling could be described with a second order equation. This equation could further be used to derive a predictive regression equation for the taxon-sampling sensitivity for a particular fraction of taxa sampled and the potential number of taxa that could be sampled. This regression equation could then e.g. be used as a "sampling correction". One could, given that we knew the fraction of the clade we were examining, calculate how close the obtained tree is to the tree that would have been obtained given a complete sample. However as Poe clearly emphasized, the utility of this predictive equation is still limited to a restrictive range of variable values, and that this should be thoroughly considered before applying it to any data.

Using Data Evaluation Prior to Tree Reconstruction

If one could evaluate the data at hand, and pinpoint problematic taxa or even remove them, before making the tree, then perhaps, one could avoid the problems in inferring the topology; at least those connected with the problematic taxa or explore the "relevance" or "use" of particular taxa for a certain problem prior to analysis. Milinkovitch *et al*. (1996) suggested one approach for accomplishing this. They applied a method involving the calculation of uncorrected (p) distances between taxa, in order to explore which taxa that potentially could be relevant for analysis or not. The rational is that if taxa added to a data set are closely related to the ones already included, they will tend to either interrupt long branches at a position that will make the long branch only a little bit shorter, or they will intercept short branches where the likelihood of multiple substitutions already is low. Thus, there is little to gain by including such taxa. Hence, they categorized those as "redundant" (see also Fig. 2 and Kim, 1996, who demonstrated that inconsistency could be intensified by adding taxa close to the tip of a long branch, hence adding redundant taxa). Consequently, taxa that would be expected to behave in an opposite way were called "informative" (see Milinkovitch *et al*., 1996: 1820, for details on their approach). Measuring pairwise distances between taxa to assess their "closeness" could be a reasonable first step to data exploration. However, since there are numerous pitfalls associated with distance analysis (see e.g., Farris, 1981; 1983) one cannot be too precautious in proceeding from such results. For example, even if a taxon is indicated as distant using one measurement of similarity, this does not necessarily mean that this taxon is more "relevant" or "problematic" than other taxa (see further review by Swofford *et al*., 1996). Measures of sequence divergence are often used in order to investigate the "level of saturation" between sequences. A sequence is said to be "saturated" when a proportion of the sequence positions (the characters) has undergone multiple substitutions. Often an overall (uncorrected) sequence divergence around 20% is regarded as saturation (e.g., Friedlander *et al*., 1994; Meyer, 1994). When a high level of saturation has been found in a data matrix, this is generally regarded as a reason for applying "*ad hoc* weighting" (*sensu* Allard and Carpenter, 1996), or to remove taxa that are saturated from the analysis (e.g., Aguinaldo *et al*., 1997). This is based on the "accepted" wisdom that parsimony requires low rates of change to perform well (e.g., Felsenstein, 1978*a*; but see Hendy and Penny, 1989). However, saturation levels should not be viewed alone. That is, one should be very careful to view the values as being a direct indicator of quality or reliability of the data. Yang (1998: 132) explains: "At any rate, pairwise sequence divergence is not a good indicator of the information content in the data, as the accuracy depends on not only the amount of evolution, but also on how many branches the tree has and how the substitutions are distributed among the branches in the tree." He further suggested (1998: 132) that "a 30-40% overall uncorrected sequence divergence may be considered as a starting point for concerns about saturation". Yang's result further support the findings by e.g. Hillis (1996; 1998) and emphasizes the importance of dense taxon sampling.

Recently, an interesting method for measuring phylogenetic signal has been developed. This method, called relative apparent synapomorphy analysis (RASA, Lyons-Weiler *et al*., 1996), works by determining whether a measure of the rate of increase of "cladistic distance" among pairs of taxa as a function of "phenetic distance" is greater than a null equiprobable rate of increase (or the rate of increase expected by chance alone, Lyons-Weiler *et al*., 1996). The cladistic distance or amount of apparent synapomorphy, between a pair of taxa is the number of times a taxon pair shares a character state to the exclusion of another taxon, all characters (and taxa) considered. As a contrast, the phenetic distance is the number of variable characters for which two taxa share a character state. These two measures of similarity can be graphically viewed in a regression plot and the observed rate of increase (the slope of the regression) can be compared with a null slope, representing the relationship between the cladistic and phenetic distance that would be present if those two measures were randomly distributed among pairs of taxa. The amount of phylogenetic signal is estimated by the difference between the observed slope and the null (Lyons-Weiler *et al*., 1996; Lyons-Weiler and Hoelzer, 1997). Thus, by measuring the relationship between the two distances and by comparing all possible triplets

of taxa, the method avoids the problems with trying to infer the patristic distance (or tree distance) between two taxa. Furthermore it circumvents problems associated with tree reconstruction by not having to rely on a tree for calculation of phylogenetic signal, contrary to previously proposed methods such as the PTP (Archie, 1989; Faith and Cranston, 1991) or the g_1 statistic (Hillis, 1991) (Lyons-Weiler *et al*., 1996; Lyons-Weiler and Hoelzer, 1997). RASA could be applied to specifically investigate properties of particular taxa. For example, long-branch taxa could introduce error by distorting hierarchical structure in a matrix. This distortion is detectable with RASA and putative outliers and problematic taxa could be identified in a so-called taxon variance plot (Lyons-Weiler and Hoelzer, 1997). Hence, taxa that could cause trees to be incorrectly reconstructed can be distinguished before the reconstruction. In a similar manner, RASA has been used for selecting suitable outgroups for rooting trees (Lyons-Weiler *et al*., 1998).

When choosing an optimal outgroup there are a number of issues that must be taken into account. For example, some taxa might be too far from the ingroup, thus making the inference of the ingroup node difficult (e.g., Maddison *et al*., 1984; Wheeler, 1990; Smith, 1994; Lyons-Weiler *et al*., 1998). Finding an optimal outgroup for a clade necessarily means searching among many putative outgroup taxa or combinations of taxa (Lyons-Weiler *et al*., 1998). Applying a rooted RASA has been suggested in order to more objectively and efficiently select the outgroup(s) that maximizes phylogenetical signal for the analysis. Lyons-Weiler *et al*. (1998) suggest that the ultimate outgroup would be a taxon with "both the highest amount of plesiomorphy, and that have also converged the least on ingroup synapomorphies." This applies for every investigated set of data, and means that one outgroup taxa could be optimal for a set of data but suboptimal when more data are added. This suggests that we should adapt a dynamic view on choosing outgroup taxa. Combining rooted RASA and taxon variance plots could be a sensitive way to seek guidance in such dynamic situations (Lyons-Weiler *et al*., 1998). However, there are limitations in the rooted RASA approach that could be of potential importance. An increase in the number of outgroup taxa may eventually lead to a total loss of signal as measured by RASA (Lyons-Weiler *et al*., 1998). This is an effect of the approach taken and could be viewed as a flaw of the methodology. Clearly, even if RASA indicates no signal, a consistent and well-supported tree might anyway be possible to reconstruct. Therefore, it seems that rooted RASA is most suitable when ascribing few taxa (perhaps only two, see also Lyons-Weiler *et al*., 1998) to the outgroup. On the other hand, more than two taxa are not needed, at least theoretically, to root a tree. Further use of the RASA approach will decide its applicability and the use of this, at least for now, promising method. In the context of the methods mentioned above, there are other issues that one has to consider. That is what to do with the information given by e.g. a taxon variance plot. Even if a problematic taxon has been identified, one is left with the decision of what to do with it (see also below). One option is to simply exclude the data, then at least some kinds of problems are avoided. On the other hand, this might lead down a "slippery-slope" to complete subjectivity (occasionally called the Theriot-effect[1]).

[1] Theriot *et al*. (1995: 4) writes in the farsical journal The Annals of Improbable Research: "We added or discarded characters until we achieved the results we believed, then stopped." This humorous description of a sampling strategy was named the "Theriot effect" by Hillis (1998). To put this in the context of taxonomic sampling, "characters" could be replaced by "taxa" in the cited sentence above.

DEALING WITH THE PROBLEM

The solution to the sampling dilemma could be a general one, such as "add characters", "add taxa judiciously", "apply a particular weighting scheme", "use a particular method of inference", *et cetera*. Or, are we faced with the possibility that there is no general solution to the problem? Each addition of a taxon to an analysis creates new analytical circumstances, which we have to face. Since conditions change, should we adapt our methods to each particular situation? Huelsenbeck (1998: 533) writes "No one method of phylogenetic analysis can be expected to perform best for all possible data sets. It is important, then, to keep in mind the limitations of any method of analysis and proceed cautiously when the data appear to indicate that the method may be providing misleading results." Of course, this immediately raises the question on how to accomplish this; how to "proceed cautiously", and also how to know if we are faced with a difficult situation in the first place. Furthermore, is adapting the methods to the situation even methodologically sound - from a scientific point of view (e.g., see Siddall and Kluge 1997; Kluge, 1997*a*)? The alternative is to use one method that performs well under most circumstances, but then again, how can we tell what is good enough?

Data Manipulations - Weighting

One way to try to diminish the influence of homoplastic changes in phylogenetic analysis has been to manipulate the data in some way. This could be done by "correcting distances for unseen changes" or apply character or character-state weighting[2]. By giving a lower weight to a character (or a certain character change) known or presumed to be homoplastic, this will lower its influence on the analysis. Hence, if we could down-weight, or even exclude, all homoplastic characters from an analysis, we would be certain to get a "correct" answer. However, judging from what systematists have been doing the last ten decades or so, this is not an easy task. Furthermore, the question is if weighting is really necessary (or even defendable, Kluge, 1997*a*; 1997*b*). Albert *et al*. (1993) concludes, "Both equal and differential weighting should give the same result if sufficient numbers of terminal taxa permit the detection of historically misleading character-state changes" (see also Hillis, 1996; Milinkovitch *et al*., 1996; Purvis and Quicke, 1997; Hillis, 1998; Håstad and Björklund, 1998; Källersjö *et al*., 1999; Björklund, 2000; Sennblad and Bremer, 2000). Of course, we are left with what exactly is a "sufficient number". Furthermore, there might not be a sufficient number to sample, for other reasons than that taxa simply do not exist (see above). The question of what to do when faced with this latter situation is not trivial. If taxa can't be added because they don't exist, or if the question prevents exclusion of critical taxa, there are not many ways to go. Manipulation of data or changing the method of reconstruction seems, today, to be the only alternatives. However, none of those are guaranteed to lead to a better answer.

Changing the Method

It has been conjectured (e.g., Kuhner and Felsenstein, 1994; Huelsenbeck, 1997; Poe 1998*b*) that an incorrect inference made by, e.g. parsimony in the Felsenstein zone, could be alleviated by using a method that could "handle" such situations (e.g. maximum likelihood). Those methods less sensitive to long-branch attraction will tend to separate the long branches that were grouped together by parsimony. This would mean that given our current sample of taxa, the correct inference could be

[2] How character- or character state weighting is accomplished will not be discussed here. The reader is referred to extensive reviews by e.g. Simon *et al*., 1994 and Swofford *et al*., 1996.

secured just by choosing a "correct" method for this situation. This is true, as far as it has explicitly been demonstrated in several simulation studies (e.g., DeBry, 1992; Huelsenbeck and Hillis, 1993; Zharkikh and Li, 1993; Kuhner and Felsenstein, 1994; Tateno *et al*., 1994; Hillis *et al*., 1994; Cunningham *et al*., 1998). However, Siddall (1998) pointed out that this reasoning is not straightforward. He described a situation where a model tree contained two long branches together, and where maximum likelihood were found to separate them in the inferred tree (hence the term "long-branch repulsion", Siddall 1998: 209). Thus, in a similar situation with real data, the investigator is still left with the question "are these taxa artificially grouped because of long-branch attraction, or do they really belong together?" This question seems not to have a definite answer, at least not just by choosing between the methods of inference available today (e.g., Huelsenbeck, 1998; Siddall, 1998). Long-branch attraction and long-branch repulsion clearly appear under particular conditions, but how common are these conditions in nature? If the answer is that they are uncommon, the use of one method that has desirable properties in most cases could be legitimate. A desirable property could be one as trivial (or non-trivial!) as the ability to analyze large data sets in a reasonable time frame. After all, one of the main arguments against model-based methods has been (at least among the proponents of the very same methods) that they are intractable or computer intensive (e.g., Felsenstein, 1979; Strimmer, 1997; Lewis, 1998; Page and Holmes, 1998). This issue could be reduced down to merely a matter of opinion (but see Siddall and Kluge 1997) since one could argue that "I rather use a method that will make me sure in the limited case, and stick to that, just in case". Of course this reasoning could apply to all methods if they could be shown to misbehave under certain conditions (Siddall, 1998). Clearly, there is a need for new and refined methods to be developed in order to solve the most difficult questions (e.g., Flook and Rowell, 1997). As already has been foreseen (e.g., Lyons-Weiler *et al*., 1998; Page and Holmes, 1998) the combination of different methods is something we will see more of in the future. Already today, combinations of simple and more time demanding methods are used in calculations. Computer programs such as PAUP* (Swofford, 2000) has the option to use fast methods (e.g., Neighbor-joining or parsimony) to approximate parameters to be input in consecutive rounds of calculations using more complex methods, such as maximum likelihood. A possible extension to this could be to apply parsimony and maximum likelihood to different regions in the tree, and perhaps applying maximum likelihood only when long-branch attraction is suspected.

Sensitivity Analysis

A commonly seen practice in contemporary systematics has been to apply different methods of analysis to the same phylogenetic problem. Each method used could emphasize different aspects of patterns in character distributions, or imply different models of evolutionary change. The reason for the use of different methods has sometimes not been explicitly stated, but the underlying idea seems to be "if the result is the same despite the use of different methods, this increases the confidence in the results" (e.g., Zink and Avise, 1990; Härlid *et al*., 1997). This notion has been disputed on grounds that fundamentally different approaches cannot be seen as having equal input for summing up an added credibility. Thus, some methods may be seen as contributing nothing to support an evolutionary hypothesis, even if the resulting topologies are the same (Siddall and Kluge, 1997). However, increasing the accuracy of inference by applying several methods has received support based on simulation studies by Kim (1993). Being on the safe side, one can conclude that if the same topology is recovered whatever method applied, the data analyzed are at least robust to change in methods. More unclear is what to do if the results do differ, and which result or method to choose in these cases.

Sensitivity analysis is a somewhat different approach in which the sensitivity of an analysis to variation in one or more parameter values is assessed (e.g., Wheeler, 1995). Instead of changing the reconstructing method (i.e. changing from e.g. UPGMA to neighbor joining), parameters such as differential transition/transversion weights, gap costs used in alignment or different substitution models

could be varied. Sensibility analysis is an excellent framework when investigating how parameter values changeable by the researcher influences the results. However, in order to arbitrate between different results some optimality criterion must be used such as congruence or convergence to a specific tree topology. Wheeler (1995) argued that without having a way of objectively measuring the accuracy of reconstruction, congruence was the most reasonable way to choose between results. Wheeler used both taxonomic congruence (using consensus methods, Nelson, 1979; Bremer, 1990) and character congruence (using the Mickevitch-Farris Index, Mickevitch and Farris, 1981) for his sensitivity analysis. And when seeking the parameter values that maximized congruence, the approaches gave different results. Assessing congruence is a difficult and sometimes controversial issue (see e.g., Miyamoto and Fitch, 1995; Huelsenbeck *et al*., 1996; Cunningham, 1997*a*) and the question of how to most accurately apply measures of congruence in sensitivity analysis is surely in need for further investigations.

Sensitivity analysis is a reasonable framework as long as one realizes that the number of parameters affecting the outcome of an analysis is very large (e.g., Felsenstein, 1978*a*; 1981). Sometimes statements like "we analyzed a number of parameters that are most often used in phylogenetic analysis, and the ones we argue are most relevant to analysis" are encountered in the literature (e.g., Milinkovitch *et al*., 1996; Cunningham, 1997*b*). It is true that some parameters are more important for an outcome of an analysis than others but the choice of which parameters to investigate must be thoroughly scrutinized. Those parameters not considered in the sensitivity analysis should at least not *a priori* be taken as unimportant.

Known-tree Investigations

An important area for future study is to investigate the nature of homoplasy. Are there general properties that could be deduced from those characters that lead us to infer wrong topologies? To answer such questions we somehow have to be able to sort the wrong trees from the true trees. Using known phylogenies is a way to accomplish this. Trees and data are easily generated by computer simulations or perhaps more sophisticated, using biological simulations (Hillis *et al*., 1992; Hillis and Huelsenbeck, 1994). A drawback of the simulation approach is the limitations on how the results could be generalized to other data. The known phylogeny is then limited to serve as a first test of predictions made from other simulations or theory (see also Hillis, 1995; Poe, 1998*b*). Furthermore, experimentally generated genealogies are still few and consist of a low number of terminals (nine terminals in the case of the T7-biophage data from Hillis *et al*., 1992). A low number of terminals might not be preferable when testing the effects of taxon sampling. An alternative is then to use "well-established phylogenies" and real data (e.g., Graybeal, 1994; Philippe *et al*., 1996; Russo *et al*., 1996; Naylor and Brown, 1998), well-established phylogenies being those that are generally accepted or considered well corroborated[3]. Using such phylogenies, individual characters or taxa that disagree with the tree can easily be identified. Thereafter, an examination of the outliers could reveal e.g. structural or chemical properties of characters that will make them prone to homoplasy (Wheeler and Honeycutt, 1988; Naylor and Brown, 1997; 1998). This kind of information could possibly be used in later phylogenetic analyses. Needless to say, if the "well-established" phylogenies turn out to be wrong, they may have led us to the wrong conclusions. Naylor and Brown (1998) investigated a vertebrate data set where the mitochondrial DNA tree differed from their preconceived higher-taxon tree. By examining

[3] It could be worthwhile considering the way systematists validate character evidence, and to which extent a hypothesis is considered "well corroborated". Naylor and Brown (1998: 71) write "Despite a very large sample - 12,234 protein-coding sites, the maximum obtainable from metazoan mtDNA - an erroneous yet robust topology resulted - a result contradicted by *a wealth of other data* [my emphasis]". The analysis by Naylor and Brown is based on one of the largest data sets (in respect to number of characters) assembled for vertebrates so far.

sites that performed particularly bad in tree reconstruction, they found that nucleotide sites modally coding for the hydrophobic amino acids leucine, isoleucine, and valine, contained a nonrandom, misleading signal. Furthermore, none of the models of molecular evolution applied in the tree reconstruction could compensate this "misleading" signal, despite a data size of 12234 nucleotide positions. Although much can be said about different genes, classes of characters and even individual taxa in Naylor and Brown's study, it is not excluded (which they also appreciate) that another species sample could have given another tree (perhaps even the "true" tree). Hence, this could eventually have led to different considerations about the data. Thus, the possibility of adding more taxa needs to be explored before we can make general statements about the utility of particular classes of characters for phylogeny reconstruction or the use of different reconstruction methods.

JUDICIOUS SAMPLING - THE IMPORTANCE OF GOOD DESIGN

Purvis and Quicke (1997: 50) stressed the "importance of good design" in phylogenetic analyses. This phrase catches the very essence of the whole issue of taxon sampling. Taxa could be chosen in such a way as to ease the reconstruction of a sample tree. Seeking an absolute answer to how to accomplish this for every case is, as has been evident from the discussions above, like groping about in the dark. However, there are occasionally some gleams of light that could lead the way. A first important step is to recognize the limitations of our expectations on analyses. It has been claimed that our preconceived ideas of relationships put severe constraints on what is achievable in phylogenetic analysis (e.g., Lecointre, 1994). More specifically, a researcher could view an outcome from an analysis as so implausible that she rejects the notion of proceeding, hence preconceived ideas "constrain" the analysis. This requires us to separate what the data actually show us, from what we should or could expect from an analysis. However, assuming that we actually have a fairly good idea or notion of how taxa are related (e.g., Graybeal, 1998), we are in a situation where we can influence the outcome. A most powerful way to accomplish this is to sample taxa strategically. Then we have some chance of influencing the seemingly most important factor for successes in phylogenetic reconstruction *viz*. tree shape. Generally, cutting long branches and diminishing imbalance of the tree will lead us to more consistent reconstruction. Hillis (1998) discussed alternative sampling strategies that could be applied when initiating a phylogenetic study. The method he thought most researchers would use (and presumably are using) is a combination of selecting a few representatives of different subgroups that supposedly represents the overall diversity of the group, and sampling taxa strategically to e.g. intercept long branches in the initial (or expected) tree. Besides the idiosyncrasies mentioned in the section above, the strategy of sampling representatives (the Exemplar method) has the potential to bias an analysis toward more symmetrical or asymmetrical trees, depending on the actual shape of the underlying tree (Huelsenbeck and Kirkpatrick, 1996). Preliminary analysis of the group in question, coupled with tests for differences in diversification rates (e.g., Slowinski and Guyer, 1989) or symmetry could provide guidance about the desired direction for continued sampling (Sanderson, 1996). It seems reasonable to approach sampling in an iterative and recursive way. One option is to focus primarily on increasing the number of taxa, and after potential problems (e.g. long branches) has been identified (and perhaps removed), more characters could be added (e.g., Graybeal, 1998; Lecointre *et al*., 1993; Lecointre *et al*., 1994; see also Poe and Swofford, 1999). What action that is the most preferable (e.g. adding taxa or characters) after an initial round of analysis must be judged after evaluating the information received. How to improve an analysis given preliminary information is still relatively unexplored area of research, and should be an important task for future studies (Kim, 1998).

Giving recommendations on how to sample outgroup taxa specifically seems to be more difficult. Although the issue of outgroup sampling is very similar to the general taxon-sampling problem, there is a difference in that the number of potential outgroup taxa are much higher than the

number of ingroup taxa. Thus, we are always faced with the decision of which outgroup taxa to discard. And although researchers aim for achieving the same task (correct reconstruction of a tree with a correct placement of the root) advice on outgroup selection given in the literature can be contradictory. Smith (1994) recommended adding several taxa to a monophyletic clade (preferably the sistergroup to the ingroup). In this way the long branch leading to the ingroup could (potentially) be shortened and tree balance increased. Lyons-Weiler *et al*. (1998), on the other hand found that inclusion of more than one taxon from a monophyletic group tends to decrease signal (measured by RASA), and that inclusion of all outgroup taxa may occasionally decrease the plesiomorphy content. According to them, a possible explanation was that when more outgroups are added from a monophyletic group, the probability will also increase of sampling apomorphies in the outgroup that will have converged on ingroup synapomorphy. Their conclusions and proposal for a recommended sampling strategy (although they admitted it to be a poor one) was contrary to Smith, to use only two taxa from a paraphyletic outgroup, including (preferably) the sister taxon. Lyons-Weiler *et al*. (1998) (see also Lyons-Weiler *et al*., 1996; Lyons-Weiler and Hoelzer, 1997, Milinkovitch and Lyons-Weiler, 1998) strongly advocates that plesiomorphy content of the putative outgroup taxa should be evaluated before including them into analysis. They also extend this reasoning as to include all taxa in the ingroup as well (i.e. investigate the effect on phylogenetic signal). If this could be shown being a rational and feasible procedure, indeed data evaluation prior to tree reconstruction could serve as a powerful tool in taxon sampling.

CONCLUSIONS

As long as time, funding and technology set the boundaries for phylogenetic analyses, systematists are forced to make choices of what to include in them. Seemingly paradoxical, a common problem today is not to extract data from the organisms in the first place, but rather to sieve out the non-relevant information from the vast amount of data already available. How to accomplish this, and finding efficient ways to evaluate data prior to phylogeny reconstruction, is perhaps the most urgent field of research regarding taxon sampling. Furthermore, a particular choice of data might provide an incorrect and misleading answer. Yet, the incorrect answer might receive high credence based on our methods to measure phylogenetic confidence. Therefore, recognizing the analytical circumstances under which taxon sampling have immediate effects becomes crucial. Unfortunately, the wealth of investigations published so far offers no general solution to all the problems associated with taxon sampling, but, we might derive a few rules of thumb: If there is a choice of which taxa to add, add taxa in order to increase the symmetry of the inferred tree. If there are particularly long branches in the data set, either add taxa leading to those branches (in order to shorten them), and/or use a method of analysis that tries to reduce the effect of long branches or multiple substitutions. A third, yet controversial, alternative is to consider excluding the aberrant taxa from the analysis. When using an exemplar method, try to use as many terminals as feasible. If forced to choose a reduced data set, sample across the known variation, and sample more densely around taxa that are particularly aberrant or those that supposedly would add long branches to the tree. When sampling outgroup taxa, it seems preferable to start with the sister taxon, and then further taxa should be added in order to both cut branches and increase the balance of the predicted tree. Distant outgroup taxa contribute less in terms of character state and rooting information, and could even be introducing errors in to the analysis. In cases where ingroup monophyly is questionable, outgroup taxa should be regarded as "tests", or potential "falsifiers", and added accordingly.

As is common for many rules of thumb, there exists a number of counter examples in which they fail to apply. Understanding the nature of these exceptions helps us to make the better decision when sampling data for a phylogenetic analysis. Investigating if a phylogenetic analysis is sensitive to

taxon sampling must be seen as a fundamental part when giving credibility to its results. This can only be achieved, eventually, by increasing the number of taxa and characters in our analyses.

ACKNOWLEDGMENTS

I wish to thank Fredrik Ronquist for comments and patience.

REFERENCES

AGUINALDO, A. M. A., J. M. TURBEVILLE, L. S. LINFORD, M. C. RIVERA, J. R. GAREY, R. A. RAFF, AND J. A. LAKE. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*, 387: 489-493.

ALBERT, V. A., M. W. CHASE AND B. D. MISCHLER. 1993. Character-state weighting for cladistic analysis of protein-coding DNA sequences. *Annals of the Missouri Botanical Garden*, 80: 752-766.

ALLARD, M. W. AND CARPENTER, J. M. 1996. On weighting and congruence. *Cladistics*, 12: 183-198.

ARCHIE, J. W. 1989. A randomization test for phylogenetic information in systematic data. *Systematic Zoology*, 38: 239-252.

ARNOLD, E. N. 1981. Estimating phylogenies at low taxonomic levels. *Zetischrift für Zoologische Systematik und Evolutionsforschung*, 19: 1-35.

AVISE, J. C., J. F. SHAPIRO, S. W. DANIEL, C. F. AQUADRO AND R. A. LANSMAN. 1983. Mitochondrial DNA differentiation during the speciation process in *Peromyscus*. *Molecular Biology and Evolution*, 1: 38-56.

BININDA-EMONDS, O. R. P., H. N. BRYANT AND A. P. RUSSELL. 1998. Supraspecific taxa as terminals in cladistic analysis: implicit assumptions of monophyly and a comparison of methods. *Biological Journal of the Linnean Society*, 64: 101-133.

BJÖRKLUND, M. 2000. Are third positions really that bad? A test using vertebrate Cytochrome *b*. *Cladistics*, 15: 191-197.

BREMER, K. 1990. Combinable component consensus. *Cladistics*, 6: 369-372.

BREMER, K. 1994. Branch support and tree stability. *Cladistics*, 10: 295-304.

CARMEAN, D. AND B. CRESPI. 1995. Do long branches attract flies? *Nature*, 373: 666.

CUMMINGS, M. P., S. P. OTTO AND J. WAKELEY. 1995. Sampling properties of DNA sequence data in phylogenetic analysis. *Molecular Biology and Evolution*, 12: 814-822.

CUNNINGHAM, C. W. 1997*a*. Can three incongruence tests predict when data should be combined? *Molecular Biology and Evolution*, 14: 482-496.

CUNNINGHAM, C. W. 1997*b*. Is congruence between data partitions a reliable predictor of phylogenetic accuracy? Empirically testing an iterative procedure for choosing among phylogenetic methods. *Systematic Biology*, 46: 464-478.

CUNNINGHAM, C. W., K. E. OMLAND AND T. H. OAKLEY. 1998. Reconstructing ancestral character states: a critical reappraisal. *Trends in Ecology and Evolution*, 13: 361-366.

CUNNINGHAM, C. W., H. ZHU AND D. M. HILLIS. 1998. Best-fit maximum-likelihood models for phylogenetic inference: empirical tests with known phylogenies. *Evolution*, 52: 978-987.

DEBRY, R. W. 1992. The consistency of several phylogeny-inference methods under varying evolutionary rates. *Molecular Biology and Evolution*, 9: 537-551.

DONOGHUE, M. J., J. DOYLE, J. GAUTHIER, A. G. KLUGE AND T. ROWE. 1989. The importance of fossils in phylogeny reconstruction. *Annual Review of Ecology and Systematics*, 20: 431-460.

DONOGHUE, M.J. AND P.D. CANTINO. 1984. The logic and limitations of the outgroup substitution approach to cladistic analyses. *Systematic Botany*, 9: 192-202.

DOYLE, J. J. 1992. Gene trees and species trees: Molecular systematics as one-character taxonomy. *Systematic Botany*, 17: 144-163.

DOYLE, J. J. 1997. Trees within trees: genes and species, molecules and morphology. *Systematic Biology*, 46: 537-553.

EFRON, B. 1979. Bootstrapping methods: another look at the jackknife. *Annals of Statistics*, 7: 1-26

ERDÖS, P. L., M. STEEL, L. SZEKELEY, AND T. WARNOW. 1997. Constructing big trees from short sequences. *Proceedings of International Congress on Automata, Languages, and Programming*.

ESTABROOK, G. F. 1977. Does common equal primitive? *Systematic Botany*, 2: 36-42.

FAITH, D. P. AND P. S. CRANSTON. 1991. Could a cladogram this short have arisen by chance alone? On permutation tests for cladistic structure. *Cladistics*, 7: 1-28.

FARRIS, J. S. 1981. Distance data in phylogenetic analysis. *In* V. A. Funk and D. R. Brooks (Eds.), *Advances in cladistics: proceedings of the first meeting of the Willi Hennig Society* (pp. 3-23). Bronx: New York Botanical Garden.

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

FARRIS, J. S. 1989. The retention index and the rescaled consistency index. *Cladistics*, 5: 417-419.

FARRIS, J. S. 1998. The future of phylogeny reconstruction. *Zoologica Scripta*, 26: 303-311.

FARRIS, J. S., V. A. ALBERT, M. KÄLLERSJÖ, D. LIPSCOMB AND A. G. KLUGE. 1996. Parsimony jackknifing outperforms neighbor-joining. *Cladistics*, 12: 99-124.

FELSENSTEIN, J. 1978*a*. Cases in which parsimony and compatibility methods will be positively misleading. *Systematic Zoology*, 27: 401-410.

FELSENSTEIN, J. 1978*b*. The number of evolutionary trees. *Systematic Zoology*, 27: 27-33.

FELSENSTEIN, J. 1979. Alternative methods of phylogenetic inference and their interrelationship. *Systematic Zoology*, 28: 49-62.

FELSENSTEIN, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39: 783-791.

FITCH, W. M. 1970. Distinguishing homologous from analogous proteins. *Systematic Zoology*, 19: 99-113.

FLOOK, P. K. AND C.H. ROWELL. 1997. The effectiveness of mitochondrial rRNA gene sequences for the reconstruction of the phylogeny of an insect order (Orthoptera). *Molecular Phylogenetics and Evolution*, 8: 177-92.

FRIEDLANDER, T. P., J. C. REGIER AND C. MITTER. 1994. Phylogenetic information content of five nuclear gene sequences in animals: initial assessment of character sets from concordance and divergence studies. *Systematic Biology*, 43: 511-525.

GATESY, J., R. DE SALLE AND W. WHEELER. 1993. Alignment-ambiguous nucleotide sites and the exclusion of systematic data. *Molecular Phylogenetics and Evolution*, 2: 152-157.

GAUT, B. S. AND P. O. LEWIS. 1995. Success of maximum likelihood in the four-taxon case. *Molecular Biology and Evolution*, 12: 152-162.

GAUTHIER, J., A. G. KLUGE AND T. ROWE. 1988. Amniote phylogeny and the importance of fossils. *Cladistics*, 4: 105-205.

GIVNISH, T. J. AND K. J. SYTSMA. 1997. Consistency, characters, and the likelihood of correct phylogenetic inference. *Molecular Phylogenetics and Evolution*, 7: 320-330.

GRAUR, D., L. DURET, AND M. GOUY. 1996. Phylogenetic position of the order Lagomorpha (rabbits, hares and allies). *Nature*, 369: 363-364.

GRAYBEAL, A. 1994. Evaluating the phylogenetic utility of genes: a search for genes informative about deep divergences among vertebrates. *Systematic Biology*, 43: 174-193.

GRAYBEAL, A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? *Systematic Biology*, 47: 9-17.

HALANYCH, K. M. 1998. Lagomorphs misplaced by more characters and fewer taxa. *Systematic Biology*, 47: 138-146.

HENNIG, W. 1966. *Phylogenetic systematics*. Urbana: University of Illinois Press.

HENDY, M. D. AND D. PENNY. 1989. A framework for the quantitative study of evolutionary trees. *Systematic Zoology*, 38: 297-309.

HEIJERMAN, T. H. 1993. Adequacy of numerical taxonomic methods: further experiments using simulated data. *Zeitshrift für Systematik und Evolutions- forschnung*, 31: 81-97.

HILLIS, D. M. 1991. Discriminating between phylogenetic signal and random noise in DNA sequences. *In* M. M. Miyamoto and J. Cracraft (Eds.), *Phylogenetic analysis of DNA sequences* (pp. 278-294). New York: Oxford University Press.

HILLIS, D.M. 1995. Approaches for assessing phylogenetic accuracy. *Systematic Biology*, 44: 3-16.

HILLIS, D.M. 1996. Inferring complex phylogenies. *Nature*, 383: 130-131.

HILLIS, D.M. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Systematic Biology*, 47: 3-8.

HILLIS, D.M. AND J. P. HUELSENBECK. 1994. To the tree of truth: biological and numerical simulations of phylogeny. *In* D.M. Fambrough (Ed.), *Molecular evolution of physical processes* (pp. 55-67). New York: Rockefeller University Press.

HILLIS, D. M., J. P. HUELSENBECK AND D. L. SWOFFORD. 1994. Hobgoblin of phylogenetics? *Nature*, 369: 363-364.

HILLIS, D.M., J.J. BULL, M.E. WHITE, M.R. BADGETT, AND I.J. MOLINEUX. 1992. Experimental phylogenetics: generation of a known phylogeny. *Science*, 255: 589-592.

HUELSENBECK, J. P. 1995. Performance of phylogenetic methods in simulation. *Systematic Biology*, 44: 17-48.

HUELSENBECK, J. P. 1997. Is the Felsenstein zone a flytrap? *Systematic Biology*, 46: 69-74.

HUELSENBECK, J. P. 1998. Systematic bias in phylogenetic analysis: is the Strepsiptera problem solved? *Systematic Biology*, 47: 519-537.

HUELSENBECK, J. AND D. M. HILLIS. 1993. Success of phylogenetic methods in the four-taxon case. *Systematic Biology*, 42: 247-264.

HUELSENBECK, J. P. AND M. KIRKPATRICK. 1996. Do phylogenetic methods produce trees with biased shapes? *Evolution*, 50: 1418-1424.

HUELSENBECK, J. P., J. J. BULL AND C. W. CUNNINGHAM. 1996. Combining data in phylogenetic analysis. *Trends in Ecology and Evolution*, 11: 152-158.

HWANG, U. W., W. KIM, D. TAUTZ AND M. FRIEDRICH. 1998. Molecular phylogenetics at the Felsenstein zone: approaching the Strepsiptera problem using 5.8S and 28S rDNA sequences. *Molecular Phylogenies and Evolution*, 9: 470-480.

HÅSTAD, O. AND M. BJÖRKLUND. 1998. Nucleotide substitution models and estimation of phylogeny. *Molecular Biology and Evolution*, 15: 1381-1389.

HÄRLID, A., A. JANKE AND U. ARNARSON. 1997. The mtDNA sequence of the Ostrich and the divergence between paleognathous and neognathous birds. *Molecular Biology and Evolution*, 14: 754-761.

KIM, J. 1993. Improving the accuracy of phylogenetic estimation by combining different methods. *Systematic Biology*, 42: 331-340.

KIM, J. 1996. General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing number of taxa. *Systematic Biology*, 45: 363-374.

KIM, J. 1998*a*. Large-scale phylogenies and measuring the performance of phylogenetic estimators. *Systematic Biology*, 47: 43-60.

KIM, J. 1998*b*. What do we know about the performance of estimators for large phylogenies? *Trends in Ecology and Evolution*, 12: 25-26

KLUGE, A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). *Systematic Zoology*, 38: 7-25.

KLUGE, A. G. 1997*a*. Testability and the refutation and corroboration of cladistic hypotheses. *Cladistics*, 13: 81-96.

KLUGE, A. G. 1997*b*. Sophisticated falsification and research cycles: consequences for differential character weighting in phylogenetic systematics. *Zoologica Scripta*, 26: 349-360.

KLUGE, A. G. 1998. Total evidence or taxonomic congruence: cladistics or consensus classification. *Cladistics*, 14: 151-158.

KLUGE, A. G. AND A. J. WOLF. 1993. Cladistics: what's in a word? *Cladistics*, 9: 183-199.

KOOISTRA, W. H. C. F., J. L. OLSEN, W. T. STAM AND C. VAN DEN HOEK. 1993. Problems relating to species sampling in phylogenetic studies: an example of non-monophyly in *Cladophoropsis* and *Struvea* (Siphonocladales, Chlorphyta). *Phycologica*, 32: 419-428.

KUHNER, M. K. AND J. FELSENSTEIN. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, 11: 459-468.

KÄLLERSJÖ, M., V. A. ALBERT, AND J. S. FARRIS. 1999. Homoplasy increases phylogenetic structure. *Cladistics*, 15: 91-93.

KÄLLERSJÖ, M., J. S. FARRIS, M. W. CHASE, B. BREMER, M. F. FAY, C. J. HUMPHRIES, G. PETERSEN, O. SEBERG, AND K. BREMER. 1998. Simultaneous parsimony jackknife analysis of 2538 *rbc*L

DNA sequences reveals support for major clades of green plants, land plants, seed plants and flowering plants. *Plant Sytematics and Evolution*, 213: 259-287.

LANYON, S. M. 1985. Detecting internal inconsistencies in distance data. *Systematic Zoology*, 34: 397-403.

LAKE, J. A. 1987. Rate-independent technique for analysis of nucleotide sequences: Evolutionary parsimony. *Molecular Biology and Evolution*, 4: 167-191.

LÊ, H. L. V., LECOINTRE, G., AND PERASSO, R. 1993. A 28S rRNA-based phylogeny of the Gnathostomes: First steps in the analysis of conflict and congruence with morphologically based cladograms. *Molecular Phylogenetics and Evolution*, 2: 31-51

LECOINTRE, G. 1994. Historical and heuristical aspects of systematic Ichthyology. *Cybium*, 18: 339-430.

LECOINTRE, G., H. PHILIPPE, H. L. V. LÊ AND H. LE GUYADER. 1993. Species sampling has a major impact on phylogenetic inference. *Molecular Phylogenetics and Evolution*, 2: 205-224.

LECOINTRE, G., H. PHILIPPE, H. L. V. LÊ AND H. LE GUYADER. 1994. How many nucleotides are required to resolve a phylogenetic problem? The use of a new statistical method applicable to available sequences. *Molecular Phylogenetics and Evolution*, 3: 292-309.

LEWIS, P. O. 1998. A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. *Molecular Biology and Evolution*, 15: 277-283.

LYONS-WEILER, J. AND G. A. HOELZER. 1997. Escaping from the Felsenstein zone by detecting long branches in phylogenetic data. *Molecular Phylogenetics and Evolution*, 8: 375-384.

LYONS-WEILER, J., G. A. HOELZER AND R. J. TAUSCH. 1996. Relative apparent synapomorphy analysis (RASA) I: the statistical measurement of phylogenetic signal. *Molecular Biology and Evolution*, 13: 749-757.

LYONS-WEILER, J., G. A. HOELZER AND R. J. TAUSCH. 1998. Optimal outgroup analysis. *Biological Journal of the Linnean Society*, 64: 493-511.

MADDISON, W. P. 1997 Gene trees within species trees. *Systematic Biology*, 46: 523-536.

MADDISON, W. P., M. J. DONOGHUE AND D. R. MADDISON. 1984. Outgroup analysis and parsimony. *Systematic Zoology*, 33: 83-103.

MEYER, A. 1994. Shortcomings of the cytochrome *b* gene as a molecular marker. *Trends in Ecology and Evolution*, 9: 278-280.

MICKEVITCH, M. F. AND J. S. FARRIS. 1981. The implications of congruence in *Menidia*. *Systematic Zoology*, 30: 351-370.

MILINKOVITCH, M. C., R. G. LEDUC, J. ADACHI, F. FARNIR, M. GEORGES AND M. HASEGAWA. 1996. Effects on character weighting and species sampling on phylogeny reconstruction: a case study on DNA sequence data in Cetaceans. *Genetics*, 144: 1817-1833.

MINDELL, D. P. AND C. E. THACKER. 1996. Rates of molecular evolution: phylogenetic issues and applications. *Annual Review of Ecology and Systematics*, 27: 279-303.

MIYAMOTO, M. M. AND W. M. FITCH. 1995. Testing phylogenies and phylogenetic congruence. *Systematic Biology*, 44: 64-76.

MOOERS, A. 1995. Tree balance and tree completeness. *Evolution* 49: 379-384.

NAYLOR, G. J. P. AND W. M. BROWN. 1997. Structural biology and phylogeny estimation. *Nature*, 388: 527-528.

NAYLOR, G. J. P. AND W. M. BROWN. 1998. Amphioxus mitochondrial DNA, chordate phylogeny and the limits of inference based on comparisons of sequences. *Systematic Biology*, 47: 61-76.

NELSON, G. J. 1979. Cladistic analysis and synthesis: Principles and definitions, with a historical note on Adamson's Familles des Plantes (1763-1764). *Systematic Zoology*, 28: 1-21.

NIXON, K. C. AND J. M. CARPENTER. 1996. On simultaneous analysis. *Cladistics*, 12: 221-241.

NYLANDER, J. A. A., C. ERSÉUS AND M. KÄLLERSJÖ. 1999. A test of monophyly of the gutless Phallodrilinae (Oligochaeta, Tubificidae) and the use of a 573 bp region of the mitochondrial cytochrome oxidase I gene in analysis of annelid phylogeny. *Zoologica Scripta*, 28: 305-313.

PAGE, R. D. M. AND E. C. HOLMES. 1998. *Molecular evolution: a phylogenetic approach*. Oxford: Blackwell Science.

PAMILO, P. AND M. NEI. 1988. Relationships between gene trees and species trees. *Molecular Biology and Evolution*, 5: 568-583.

PENNY, D., M. D. HENDY AND M. STEEL. 1992. Progress with methods for constructing evolutionary trees. *Trends in Ecology and Evolution*, 7: 73-79.

PHILIPPE, H., G. LECOINTRE, LE, H. L. V. AND H. LE GUYADER. 1996. A critical study of homoplasy in molecular data with the use of a morphologically based cladogram, and its consequences for character weighting. *Molecular Biology and Evolution*, 13: 1174-1186.

POE, S. 1998*a*. Sensitivity of phylogeny estimation to taxon sampling. *Systematic Biology*, 47: 18-31.

POE, S. 1998*b*. The effect of taxonomic sampling on accuracy of phylogeny estimation: test case of a known phylogeny. *Molecular Biology and Evolution*, 15: 1086-1090.

POE, S. AND D. L. SWOFFORD. 1999. Taxon sampling revisited. *Nature*, 398: 299-300.

PURVIS, A. AND D. L. J. QUICKE. 1997. Building phylogenies: are big easy? *Trends in Ecology and Evolution*, 12: 49-50.

RAMBAUT, A. AND GRASSLY, N. C. 1997. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in Biosciences*, 13: 235-238.

RANNALA, B., J. P. HUELSENBECK, Z. YANG, AND R. NIELSEN. 1998. Taxon sampling and the accuracy of large phylogenies. *Systematic Biology*, 47: 702-710.

ROHLF, F. J., W. S. CHANG, R. R. SOKAL, AND J. KIM. 1990. Accuracy of estimated phylogenies: effects of tree topology and evolutionary model. *Evolution*, 44: 1671-1684.

RUSSO, C. A. M., N. TAKEZAKI AND M. NEI. 1996. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. *Molecular Biology and Evolution*, 13: 525-536.

SANDERSON, M. J. 1996. How many taxa must be sampled to identify the root node of a large clade? *Systematic Biology*, 45: 168-173.

SENNBLAD, B. AND B. BREMER. 2000. Is there a justification for differential *a priori* weighting in coding sequences? A case study from *rbc*L and Apocyanaceae *s.l*. *Systematic Biology*, 49: 101-113.

SHAO, K.-T. AND R. R. SOKAL. 1990. Tree balance. *Systematic Zoology*, 39: 266-276.

SIDDALL, M. E. 1995. Another monophyly index: revisiting the jackknife. *Cladistics*, 11: 33-56.

SIDDALL, M. E. 1998. Success of parsimony in the four-taxon case: long-branch repulsion by likelihood in the Farris zone. *Cladistics*, 14: 209-220.

SIDDALL, M. E. AND KLUGE, A. G. 1997. Probabilism and Phylogenetic Inference. *Cladistics*, 13: 313-336.

SIDDALL, M. E. AND M. F. WHITING. 1999. Long branch abstractions. *Cladistics*, 15: 9-24.

SILLÉN-TULLBERG, B. 1993. The effect of biased inclusion of taxa on the correlation between discrete characters in phylogenetic trees. *Evolution*, 47: 1182-1191.

SIMON, C., FRATI, F., BECKENBACH, A., CRESPI, B., LIU, H. AND FLOOK, P. 1994. Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. *Annals of the Entomological Society of America*, 87: 651-701.

SLOWINSKI, J. B. AND C. GUYER. 1989. Testing the stochasticity of patterns of organismal diversity: an improved null model. *American Naturalist*, 134: 907-921.

SMITH, A. B. 1994. Rooting molecular trees: problems and strategies. *Biological Journal of the Linnean Society*, 51: 279-292.

SOLTIS, D. E., P. S. SOLTIS, M. E. MORT, M. W. CHASE, V. SAVOLAINEN, S. B. HOOT, AND C. M. MORTON. 1998. Inferring complex phylogenies using parsimony: an empirical approach using three large DNA data sets for Angiosperms. *Systematic Biology*, 47: 32-42.

STEEL, M. A., P. J. LOCKHART AND D. PENNY. 1993. Confidence in evolutionary trees from biological sequence data. *Nature*, 364: 440-442.

STEPPAN, S. 1993. Phylogenetic relationships among the Phyllotini (Rodentia: Sigmodontinae) using morphological characters. *Journal of Mammalian Evolution*, 1: 187-213.

STRIMMER, K. S. 1997. *Maximum likelihood methods in molecular phylogenetics*. Dissertation der Fakultät für Biologie der Ludwig-Maximilians-Universität, München

SWOFFORD, D. L. 2000. PAUP*. *Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4*. Sinauer Associates, Sunderland, Massachusetts.

SWOFFORD, D. L., OLSEN, G. J., WADDELL, P. J. AND HILLIS, D. M. 1996. Phylogenetic inference. *In* D. M. Hillis, C. Moritz, and B. K. Marble (Eds.), *Molecular systematics*. Second Edition (pp. 407-514). Sinauer Associates, Sunderland, Massachusetts.

TATENO, Y., N. TAKEZAKI AND M. NEI. 1994. Relative efficiency of the maximum-likelihood, neighbor joining, and maximum parsimony methods when substitution rate varies with site. *Molecular Biology and Evolution*, 11: 261-277.

THERIOT, E. C, A. E. BOGAN, AND E. E. SPAMER. 1995. The taxonomy of Barney: Evidence of convergence in hominid evolution. *Annals of Improbable Research*, 1: 3-7.

TUKEY, J. W. 1958. Bias and confidence in not quite large samples. *The Annals of Mathematical Statistics*, 29: 614.

WENZEL, J. W. 1997. When is a phylogenetic test good enough? *In* P. Grandcolas (Ed.) The origin of biodiversity in insects: phylogenetic tests of evolutionary scenarios. *Memoires du Museum National d'Histoire Naturelle* 173: 31-45

WHEELER, W. C. 1990K. Extinction, sampling, and molecular phylogenetics. *In* M. J. Novacek and Q. D. Wheeler (Eds), *Extinction and phylogeny* (pp. 205-215). New York: Columbia University Press.

WHEELER, W. C. 1990*b*. Nucleic acid sequence phylogeny and random outgroups. *Cladistics*, 6:363-367.

WHEELER, W. C. 1995. Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data. *Systematic Biology*, 44: 321-331.

WHEELER, W. C. AND R. L. HONEYCUTT. 1988. Paired sequence difference in ribosomal RNAs: evolutionary and phylogenetic implication. *Molecular Biology and Evolution*, 5: 90-96.

WHITING, M. F. 1998. Long-branch distraction and the Strepsiptera. *Systematic Biology*, 47: 134-138.

WHITING, M. F., J. C. CARPENTER, Q. D. WHEELER, AND W. C. WHEELER. 1997. The Strepsiptera problem: phylogeny of the holometabolous insect orders inferred from 18S and 28S ribosomal sequences and morphology. *Systematic Biology*, 46: 1-68.

WIENS, J. J. 1998. The accuracy of methods for coding and sampling higher-level taxa for phylogenetic analysis: a simulation study. *Systematic Biology*, 47: 397-413.

WILEY, E. O. 1981. *Phylogenetics: the theory and practice of phylogenetic systematics*. New York: Wiley Interscience.

WINNEPENNINCKX, B., T. BACKELJAU AND R. DE WACHTER. 1995. Phylogeny of protostome worms derived from 18S rRNA sequences. *Molecular Biology and Evolution*, 12: 641-649.

YANG, Z. 1996. Phylogenetic analysis using parsimony and maximum likelihood methods. *Journal of Molecular Evolution*, 42: 294-307.

YANG, Z. 1998. On the best evolutionary rate for phylogenetic analysis. *Systematic Biology*, 47: 125-133.

ZHARKIKH, A. AND W. -H. LI. 1993. Inconsistency of the maximum-parsimony method: the case of five taxa with a molecular clock. *Systematic Biology*, 42: 113-125.

ZINK, R. M. AND J. C. AVISE. 1990. Patterns of mitochondrial DNA and allozyme evolution in the avian genus *Ammodramus*. *Systematic Zoology*, 39: 148-161.