Molecular Systematics and Evolution: Theory and Practice
ed. by R. DeSalle, G. Giribet and W. Wheeler

'Pluralism' and the aims of phylogenetic research

Gonzalo Giribet[1], Rob DeSalle[2] and Ward C. Wheeler[2]

[1] *Department of Organismic and Evolutionary Biology, Museum of Comparative Zoology, Harvard University, Cambridge, MA 02138, USA*
[2] *Division of Invertebrate Zoology, American Museum of Natural History, New York, NY 10024, USA*

Summary. In science, and particularly in the field of phylogenetic systematics, investigators may choose among different methods to analyze their data. These methods include neighbor-joining (or other genetic distance approaches), maximum-likelihood, and cladistic parsimony, among others. These distinct methods of analysis differ considerably in how they process information from the observed data. However, many published molecular analyses utilize trees generated under more than one of these methods, which we will call a 'pluralistic' approach. Here, we explore the statistical, philosophical and operational aspects of the pluralistic approach. We suggest that the pluralistic approach is misguided from all three perspectives and we propose an alternative, logically consistent, strategy as an aim of phylogenetic research.

Pluralism and 'statistical support'

Some authors have advocated pluralism as a severe test for phylogenetic hypotheses, in the sense that hypotheses may be more robust if they are supported simultaneously by different methods of data analysis. This is based on an analogy with statistical analysis of multiple samples. Any result found significant in multiple samples is thought to be more convincing than a single test. This is true. It is, however, also misleading. When multiple samples are used in statistics, they are assumed to follow the same distributional model, hence the sample error is decreased with each additional datum. The data sets are multiple, not the analytical procedures. When multiple analytical techniques are used, the "significance" attached to results can be highly variable. Results that are "significant" via a standard Fisher-type approach may not be so convincing in Likelihood or Bayesian analyses.

The outcome of a pluralistic approach

There are two possible outcomes when hypotheses are generated using different methods—either they agree (they are congruent) or they disagree (they are incongruent). When three methods are used (such as neighbor joining, maximum-likelihood and parsimony) there are five different competing outcomes: they can all agree, all can be different, or there are three ways that any two can agree. In addition each of the three major kinds of methods could also give any

number of hypotheses dependent on the parameters/models used. When the results of different analyses are incongruent two routes are usually taken: (1) some method is used to settle on a consensus result, very much resembling a taxonomic congruence approach, or (2) criteria are established that result in the choice of one hypothesis over another.

Suppose that 'a favorite method' yields a different hypothesis from that of other methods because it is able to account for something that the alternative methods do not. For example, gaps are usually treated as missing data in phylogenetic analyses. A situation could arise where the best hypothesis of a method that uses gap information contradicts all the other methods, precisely because they differ in the way of treating this certain set of characters. In addition, the alternative methods could converge on a hypothesis that differs from the one based on the unique information of gap characters. In such a case, pluralism is agnostic with respect to defending a chosen method *versus* competing ones, because the different assumptions of the various methods are indeed what make one 'superior' to the others. We ask why try everything if we are going to choose the hypothesis generated by 'our favorite method' anyway? This is why, as investigators, we should be compelled to choose a single method based on philosophical criteria. In addition, an increasing number of empirical and philosophical papers have been published that discuss the different methods of phylogenetic analysis in terms of hypothesis testing (for parsimony [Kluge, 1997] and for maximum-likelihood [Huelsenbeck and Bull, 1996; Huelsenbeck and Crandall, 1997; Huelsenbeck and Rannala, 1997; Cunningham et al., 1998]), accuracy (Hillis et al., 1994; Hillis, 1995; Swofford et al., 1996) and consistency (Felsenstein, 1978; Hillis, 1996; Kim, 1996; Swofford et al., 1996; Huelsenbeck, 1997, 1998; Siddall, 1998). None of these discussions allow for a pluralistic approach.

An example: the phylogeny of the arachnid order opiliones

The arachnid order Opiliones (daddy-long-legs or harvestmen) has been classically divided into three suborders: Cyphophthalmi, Palpatores and Laniatores. The internal phylogeny of the Opiliones was studied by Giribet et al. (1999) using sequence data from the 18S rDNA and 28S rDNA loci and morphology. In this analysis, two alternative hypotheses were supported by the molecular data, the 'suborder Palpatores' (= Eupnoi and Dyspnoi) was monophyletic (Cyphophythalmi ((Eupnoi + Dyspnoi) Laniatores)), or was paraphyletic (Cyphophthalmi (Eupnoi (Dyspnoi + Laniatores))) (Fig. 1). The first hypothesis (Palpatores monophyletic) will be referred to as topology '*A*', while the hypothesis supporting Palpatores paraphyly will be referred to as topology '*B*'. The analyses of the molecular data sets were consistent with either topologies *A* or *B*, depending on the cost matrix used, while the morphology and the combined analyses (molecular + morphology) supported topology *B*.

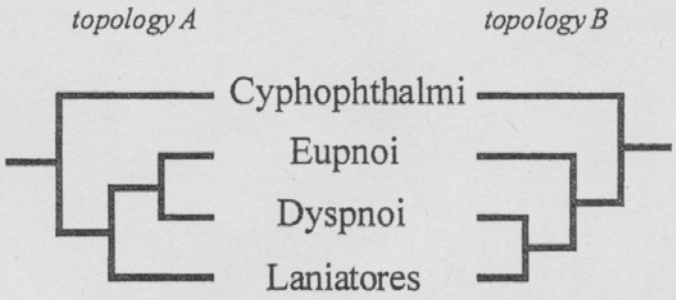

Figure 1. Two alternative hypotheses for the phylogeny of the arachnid order Opiliones, involving the arrangement among the three suborders Cyphophthalmi, 'Palpatores' and Laniatores. Palpatores (Eupnoi and Dyspnoi) result in either a monophyletic (topology *A*) or paraphyletic (topology *B*) phylogeny, depending on the method used or depending on the gap information (Giribet et al., 1999; Giribet and Wheeler, 1999b).

The 18S data set was reanalyzed (manual alignments were replaced with computer-generated alignments, using MALIGN [Wheeler and Gladstein, 1995]), and a congruence analysis was conducted for alignments generated under different insertion/deletion regimes (Giribet and Wheeler, 1999b). In this reanalysis, topology *A* was obtained when gaps were coded as 'missing data', as well as for the analyses at gap costs = 1 and 2 (with gaps treated as a character state). Topology *B* was obtained for gap costs = 4 and 8 (when gaps were coded as a character state). Using congruence with the morphological data set as a measure to decide the best-supported hypothesis (using the ILD metrics; Mickevich and Farris, 1981; W.C. Wheeler, 1995), all the analyses (at four gap regimes) were more congruent (lower ILD) when gaps were coded as a character state than when they were coded as missing data, regardless of the topology they supported. Maximum-likelihood analyses (using several models of DNA substitution, from the simplest model to models including among-site rate variation with gamma distribution) of the four alignments were consistent with topology *A*.

If an external criterion had not been available to judge these trees, one would have chosen topology *A* over topology *B* because it is the topology obtained by both methods, parsimony (for low gap costs) and maximum-likelihood. If data exploration for higher gap cost regimes had not been undertaken, topology *A* would again have been favored. But, if one appeals to congruence among data sets as the external criterion for hypothesis testing (i.e., W.C. Wheeler, 1995; Wheeler and Hayashi, 1998), topology *B* is favored.

At this point, our decision for choosing among topologies *A* and *B* is purely philosophical (whether to use parsimony or maximum-likelihood, whether to code gaps as missing data or as character states, whether to combine data sets or not, etc.), and does not have anything to do with the number of methods that yield any of the solutions.

Congruence among methods, or incongruence within a single method?

Congruence among different methods (as shown in the example presented above) may mean nothing if we apply similar 'models' to evaluate the data under each method. Suppose that we impose a simple Kimura 2-parameter (Kimura, 1980) model with transversions weighted twice as much as transitions for the maximum-likelihood and distance calculations, and we use an identical step-matrix for the parsimony calculations (gaps as missing data; tv/ts ratio = 2). In the absence of autapomorphic changes, we should obtain nearly identical solutions under the three criteria. However, if sequences are evolving in a very different manner than the designated model, the three methods could recover identical, but probably non-optimal solutions. In this case, pluralism tells us nothing about the best-supported hypothesis, but rather how similar the conditions explored in each method are. Or imagine another situation where we use parsimony and alphabetical order as criteria to formulate a phylogenetic hypothesis, and that for a particular data set, inferences using both criteria agree. Does agreement in this case really mean anything? Does congruence between results mean that alphabetical order would therefore be a defensible approach for systematic analysis? Certainly not!

A sensitivity analysis framework

The term 'sensitivity analysis' derives from the idea that trees generated from molecular data are sensitive to the parameters (indels, tv/ts ratio, etc.) used to generate the alignments (Fitch and Smith, 1983; W.C. Wheeler, 1995), the primary homology hypothesis. These authors recommended studying alignments generated under different models. In principle, we could obtain as many hypotheses as starting alignments. If we accept that, those nodes recovered under a wide range of parameters are more supported than those nodes obtained for only particular parameter sets (W.C. Wheeler, 1995). Then, alignments generated under different regimes of parameters should be studied, or our results are in peril of being based on very narrowly chosen parameter regimes.

Cunningham et al. (1998) also acknowledged that choosing among multiple models of DNA sequence evolution (in a maximum-likelihood or distance analysis framework) remains a major problem in phylogenetic reconstruction. The choice of appropriate models is especially important when there is large variation among branch lengths. This approach extends the idea of 'sensitivity analysis' to a model-testing framework. The next step in deciding on one model over another (for maximum likelihood) or a particular parameter set (for parsimony) should be based on a non-arbitrary decision, an external criterion.

Specifically, in the field of cladistic parsimony, character-based congruence can be used as an external criterion to choose a tree generated under a particular parameter set, if more than one data partition exists (see W.C. Wheeler,

1995; Whiting et al., 1997; Wheeler and Hayashi, 1998; Giribet and Wheeler, 1999b). This idea is based on the notion that the parameter set that minimizes incongruence among the different partitions would be preferred, in the same way that the shortest tree (the one that minimizes homoplasy) is to be preferred for a particular data set. Thus, the ILD metric (Mickevich and Farris, 1981) can be used for this purpose. In a maximum-likelihood framework, data exploration can be done by imposing different models, while the external optimality criterion can be defined as the likelihood itself by testing different models via the likelihood-ratio test (Huelsenbeck and Bull, 1996; Huelsenbeck and Crandall, 1997; Huelsenbeck and Rannala, 1997; Cunningham et al., 1998). This approach is accomplished by adding new sets of parameters in a hierarchical manner, and the likelihood-ratio test is performed to determine whether the more complex model can be favored over simpler models (Goldman, 1993; see Cunningham et al., 1998). The idea here is that if the log likelihood increases with parameter addition, there is a better fit to the model which is preferred by the optimality criterion. Both of these strategies, character congruence and the likelihood-ratio test, constitute external criteria by which we can choose among competing hypotheses, a property that should be present in all logically consistent methods of data analysis.

Understanding the behavior of data within a phylogenetic method as an alternative to pluralism

Incongruence of hypotheses generated under different parameters/models for a given method is well known (W.C. Wheeler, 1995; Cunningham et al., 1998). This is why data exploration (i.e., 'sensitivity analysis' [Fitch and Smith, 1983; W.C. Wheeler, 1995]) as a phylogenetic tool to study the behavior of the data is a reasonable test for robustness of phylogenetic hypotheses. We also propose here that data exploration is a preferable alternative to pluralism, specifically because sensitivity analysis within a method using different parameters/models often results in contradictory hypotheses that need to be examined. In addition, criteria can be defined that allow hypothesis testing (in a sensitivity analysis framework). In a parsimony framework, character congruence (Mickevich and Farris, 1981) is used as an optimality criterion in the case of multiple partitions (W.C. Wheeler, 1995; Whiting et al., 1997; Wheeler and Hayashi, 1998). In the case of maximum-likelihood, the likelihood-ratio test is used as an optimality criterion (Huelsenbeck and Bull, 1996; Huelsenbeck and Crandall, 1997; Huelsenbeck and Rannala, 1997; Cunningham et al., 1998). The criteria in these two different methods allow for data exploration and hypothesis testing, two extremely important (but often ignored) issues in phylogenetic systematics. However, they apply to two philosophically irreconcilable methods, and cannot be combined in a pluralistic framework.

The bottom line is that when inferences from different methods are congruent (for comparable parameters/models), but data exploration within a single

method (using different parameters/models) yields incongruence, the pluralistic approach hides inherent conflict within the data set. Even more important, when the results of different methods are incongruent, there is no criterion to choose one hypothesis over another, other than the justification of an analytical method on first philosophical principles.

Conclusions

Our intention here is not to preach a particular method of data analysis, but to advocate data exploration (through different parameters, models, and weighting schemes) and more fundamentally for philosophical consistency. Certainly, data sets may be extremely sensitive to parameters or models, which can be stable to different methods under certain assumptions. Since there are no criteria for choosing parameters/models *a priori*, the hypothesis testing issue (i.e., congruence in parsimony, likelihood-ratio test in maximum-likelihood) is extremely important. But more importantly, we stress that there is no justification for pluralism in phylogenetic systematics. Investigators should choose a method based on a philosophical rationale and carry on with their choice to the end. Non-pluralism, however, does not necessarily require distrust of other justifiable positions.

Acknowledgements
We thank Miguel Angel Arnedo, Dan Janies, Pete Makovicky and Lorenzo Prendini for discussion and suggestions.