Prior Agreement: Arbitration or Arbitrary?

MARK E. SIDDALL

Museum of Zoology, University of Michigan, Ann Arbor, Michigan 48109, USA;
Email: msiddall@umich.edu

Contemporary systematists are deeply divided on the issue of combining or partitioning data sets in phylogenetic analyses (Mickevich and Farris, 1981; Miyamoto, 1985; Bull et al., 1993; Chippindale and Wiens, 1994; Huelsenbeck et al., 1994; de Queiroz et al., 1995; Allard and Carpenter, 1996; Ballard, 1996; Huelsenbeck et al., 1996a, 1996b). Intermediate between these positions is the prior agreement (Chippindale and Wiens, 1994) approach, in which one examines properties of the data a priori with the aim of making some determination of data combinability (Bull et al., 1993; Huelsenbeck et al., 1994; Huelsenbeck et al., 1996a, 1996b), i.e., sometimes combine, sometimes do not combine. Motivations for adopting one framework or the other stem from philosophical positions that are as deeply held as are the points of view on combining or compartmentalizing (Kluge, 1989; Kluge and Wolf, 1993; Miyamoto and Fitch, 1995). For example, if one is to accept maximum likelihood as an appropriate methodology for estimating phylogenetic trees it is not clear to me what form of reasoning could justify the combined analysis of morphological and molecular data together under an HKY85 model. Thus, if the “HKY85+G5 model is the best fitting model” (Huelsenbeck, 1997:72) for nucleotide changes, it cannot also be a realistic probabilistic model for the evolution of a femur. Nor, however, should one cast all morphological data to oblivion (although some do take this extreme position; Hedges and Maxson, 1996). Similarly, if one is to take parsimony as the appropriate methodology for inferring phylogenetic trees and accept falsification and explanatory power (Farris, 1983, 1986; Mickevich and Platnick, 1989; Frost and Kluge, 1994) as its logical underpinnings, keeping data sets separate is anathema to explaining all of the data. Although Huelsenbeck et al. (1994:288) suggested that prior agreement represents a "philosophical framework for the treatment of potentially diverse data," they did not indicate what their philosophical underpinnings were. To the contrary, prior agreement is merely a methodological framework and is arbitrary in its application.

Although I disagree with Miyamoto and Fitch’s (1995) compartmentalist position, clearly they recognized the full implications of adopting a process partition foundation in their approach. That is, if the notion of process partitions is driving the decision to keep data sets separate, whether or not these partitions are incongruent is immaterial. Huelsenbeck et al. (1994; see also Huelsenbeck et al., 1996a) posited that only if data sets can be shown to be “significantly incongruent” should they be kept separate (the statement that data sets should be kept separate unless found not to be significantly incongruent differs in intent but is equivalent in implication). The argument offered by Bull et al. (1993:385) was that data sets “should not automatically be assumed to be homogeneous.” The methodological statement then reduces to the dangers of assuming homogeneity of data in an analysis. However, a necessary implication of analyzing the data sets independently thereafter is that the individual data sets are themselves assumed to be internally homogeneous, contrary to Bull et al.’s (1993) own admonishment that this is a dangerous assumption.

Consider six taxa (I–VI) and two data sets (A, B) each with four character types (Table 1). Data set A has \( h \) characters supporting (II, III, IV, V, VI), \( i \) characters supporting (II, III, IV), \( j \) characters supporting (II, III), and \( k \) characters supporting (V, VI).
Table 1. Incongruent data sets A and B. Bold values are those for the incongruent character in data set B.

<table>
<thead>
<tr>
<th>Taxon</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>0... 0...</td>
<td>0... 0...</td>
</tr>
<tr>
<td>II</td>
<td>1... 1...</td>
<td>1... 1...</td>
</tr>
<tr>
<td>III</td>
<td>1... 1...</td>
<td>0... 0...</td>
</tr>
<tr>
<td>IV</td>
<td>1... 1...</td>
<td>0... 0...</td>
</tr>
<tr>
<td>V</td>
<td>1... 0...</td>
<td>1... 0...</td>
</tr>
<tr>
<td>VI</td>
<td>1... 0...</td>
<td>1... 1...</td>
</tr>
</tbody>
</table>

* Number of characters with this pattern.

Data set B has $x$ characters supporting (II, III, IV, V, VI), $y$ characters supporting (II, III), $z$ characters supporting (V, VI), and only one character supporting (IV, V, VI). No matter how large or small one makes the numbers $h, i, j,$ and $k$ or the numbers $x, y,$ and $z,$ the lone incongruent character in data set B, which supports a contradicting clade (IV, V, VI), will always be sufficient to render the data sets significantly incongruent (using the incongruence length difference test; Farris et al., 1994). Huelsenbeck et al. (1994) would have us not combine these data sets together, even if the single incongruent character were but one of millions in data set B and billions in data set A and not withstanding that otherwise there is considerable mutual support for three clades by both data sets. Of course, data sets A and B are contrived and the example is trivial. Nonetheless, the principle can be extended to real situations.

Siddall et al. (1997) examined the relationships of the protistan phyla Ciliophora, Apicomplexa, and Dinoflagellida and particularly the placement of the genus Perkinsus therein with 18S ribosomal DNA (rDNA) and actin sequences. A different tree was found using rDNA data (Fig. 1a) than was found using actin data (Fig. 1b). The incongruence length difference test of Farris et al. (1994; implemented in ARNIE.EXE; Siddall, 1996) indicated significant heterogeneity ($P = 0.001$). This result would have us, under the prior agreement criterion, refrain from combining these data sets. However, Siddall et al. (1997) showed that the first and second codon positions in the actin gene were not incongruent with the rDNA data ($P = 0.643$), whereas these data taken together (Fig. 1a) were incongruent with the third positions in the actin gene (Fig. 1c).

The conclusion here, then, is that although prior agreement would have had us not combine the actin and rDNA data at all, the decision to act thus would have been arbitrary because most of the actin data are congruent. That is, one manner of partitioning would have been chosen over some other possible manner of partitioning that is equally or more incongruent than the one chosen without there being any particularly good reason for choosing one over the other. (By arbitrary, I mean “selected at random or without reason,” i.e., choice of one manner of partitioning over a conceivable multitude of possible partitions; moreover, the decision to not combine, if made on the basis of arbitrary choice, is then itself arbitrary.) Were this the extent of it, prior agreement could simply assert that the third position represents a nonarbitrary partition and is the cause of the incongruence. However, suppose we had a way to arbitrarily set off 75 characters, for example taking the characters in third position serial homology in the actin gene, calculating for each character the difference between its consistency on the tree suggested by the third positions and that on the tree suggested by the rest of the data, and then ranking the characters according to this difference. Arbitrarily removing only these 75 positions (less than one-sixth of the actin data) with the greatest difference in consistency on the two trees renders the rest of the third position characters congruent (Fig. 1a).
Figure 1. Single most-parsimonious trees resulting from parsimony analyses of genera in three protistan phyla. (a) 18S rDNA, or 18S rDNA + actin, or 18S rDNA + first and second positions of actin, or 18S rDNA + first and second positions + all but 75 third positions of actin, or 18S rDNA + first and second positions + all but 50 third positions of actin. (b) Actin alone. (c) Third positions of actin sequences alone. (d) Subset of 75 third positions of actin alone. (e) Strict consensus of a subset of 50 third positions of actin alone.
TABLE 2. Incongruent data sets C and D.

<table>
<thead>
<tr>
<th>Taxon</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>000000</td>
<td>0000</td>
</tr>
<tr>
<td>II</td>
<td>010101</td>
<td>0111</td>
</tr>
<tr>
<td>III</td>
<td>010101</td>
<td>1011</td>
</tr>
<tr>
<td>IV</td>
<td>100100</td>
<td>1100</td>
</tr>
<tr>
<td>V</td>
<td>010010</td>
<td>1100</td>
</tr>
<tr>
<td>VI</td>
<td>010010</td>
<td>0010</td>
</tr>
</tbody>
</table>

with the rDNA + actin first position + actin second position data set ($P = 0.581$). These 75 third position characters (Fig. 1d) were significantly incongruent with the rest of the data ($P = 0.001$). However, there is no discernible connection among these 75 characters. That is, their involvement was not linked to any particular amino acid(s) and they were randomly located across the actin sequence. Lest the incongruence now be attributed to these 75 characters, consider a ranking of their individual incongruence with the rest of the data according to the skew of the frequency distribution of the sum of tree lengths found from random partitions. Now, arbitrarily discarding only the 50 most incongruent (least skewed) characters renders the remaining 25 characters congruent with the rest of the actin + rDNA data set (Fig. 1a). These 50 third-position characters, of course, are significantly incongruent (Fig. 1e) with the rest of the data ($P = 0.001$). There may yet be even fewer than 50 characters that are contributing this heterogeneity, but I ran out of arbitrary criteria for finding them.

What is a phylogeneticist to do when confronted with a situation in which there is more than one way to identify incongruence? Data sets C and D (Table 2) are incongruent ($P = 0.02$). But if one removes the first, second, or third character in D, the data sets are congruent ($P > 0.05$). Which character is causing incongruence? In fact, incongruence frequently is a more complex issue that cannot be so readily teased apart as Huelsenbeck et al. (1994; Huelsenbeck et al., 1996a) would have us believe. Returning to the alveolate phyla in Figure 1, if the incongruence in actin data (Fig. 1b) were merely reflective of the third positions alone (Fig. 1c), we would expect third positions to reveal a poorer match to the tree in Figure 1a. In fact, both of these trees (Figs. 1b, 1c) match the whole-data tree equally well (9 of 11 possible nodes are correlated [sensu Page, 1994] as determined by an exhaustive search with TreeMap; Page, 1995), they just match in different ways.

Incongruence and combinatoriality are different things. Finding two chosen partitions to be heterogeneous carries no more meaning than finding that two other arbitrary partitions are heterogeneous. When data are partitioned, the investigator runs the risk of discarding otherwise corroborating information. The proportion of characters (as made extreme in the contrived single-character example) contributing to incongruence might in fact be vanishingly small and of marginal consequence in light of otherwise overwhelming agreement or mutual corroboration in the two data sets. Prior agreement cannot be saved from arbitrariness, and phylogeneticists must decide either to combine or to not combine according to the merits of the competing philosophies. This approach may not be as comforting as the prior agreement approach, but it is rational.

ACKNOWLEDGMENTS

I thank A. Kluge, M. Sorenson, and C. Thacker for comments on earlier drafts and Paul Chippindale for his thoughtful review.

REFERENCES


Received 13 November 1996; accepted 15 August 1997
Associate Editor: David Cannatella