THE LIMITS OF AMINO ACID SEQUENCE DATA IN ANGIOSPERM PHYLOGENETIC RECONSTRUCTION

Kåre Bremer
Swedish Museum of Natural History, Department of Phanerogamic Botany, P.O. Box 50007, S-104 05 Stockholm, SWEDEN

Abstract.—Amino acid sequence data are available for ribulose biphosphate carboxylase, plastocyanin, cytochrome c, and ferredoxin for a number of angiosperm families. Cladistic analysis of the data, including evaluation of all equally or almost equally parsimonious cladograms, shows that much homoplasly (parallelisms and reversals) is present and that few or no well supported monophyletic groups of families can be demonstrated. In one analysis of nine angiosperm families and 40 variable amino acid positions from three proteins, the most parsimonious cladograms were 151 steps long and contained 63 parallelisms and reversals (consistency index = 0.583). In another analysis of six families and 53 variable amino acid positions from four proteins, the most parsimonious cladogram was 161 steps long and contained 50 parallelisms and reversals (consistency index = 0.689). Single changes in both data matrices could yield most parsimonious cladograms with quite different topologies and without common monophyletic groups. Presently, amino acid sequence data are not comprehensive enough for phylogenetic reconstruction among angiosperms. More informative positions are needed, either from sequencing longer parts of the proteins or from sequencing more proteins from the same taxa.

Received June 22, 1987. Accepted January 26, 1988

Very little is known with certainty about the phylogeny of the angiosperms (e.g., Heywood, 1977; Dahlgren and Bremer, 1985). Often, morphological data are not evaluated in enough detail, and the vast amount of information is difficult to synthesize in a single, consistent, and parsimonious hypothesis of angiosperm phylogeny. Homoplasly (i.e., parallelisms and reversals) is common in morphological data at higher taxonomic levels.

Amino acid sequences from a number of proteins are used for inferring interrelationships between an increasing number of investigated angiosperm families. If these data are reliable, a much needed general framework of angiosperm phylogeny could be generated. Most protein sequence data used in phylogenetic reconstruction come from the small subunit of ribulose biphosphate carboxylase (RBC-SSU; Martin et al., 1983; Martin and Dowd, 1984a, 1984b, 1984c), plastocyanin (Boulter et al., 1979), and cytochrome c (Boulter et al., 1972). Sometimes the phylogenetic trees generated from sequences in these different proteins are inconsistent, and there is uncertainty about the stability of the data. The investigators are generally very cautious in drawing conclusions about phylogeny from their analyses (Boulter et al., 1979; Martin et al., 1985).

A major problem is that the number of variable amino acid positions currently available in a single protein (15–30) is generally too small for a well supported hypothesis about the interrelationships of the taxa investigated (Peacock, 1981). The combination of data from several species into familial sequences (Martin et al., 1983) and the subsequent combination of familial sequences from several proteins (Martin et al., 1985) yield data sets with a larger number of informative amino acid positions. Such data sets may become large enough to say something about the interrelationships of the angiosperm families included.

Martin et al. (1985) used sequences of five macromolecules, RBC-SSU, plastocyanin, cytochrome c, ferredoxin, and 5S rRNA, to construct phylogenetic trees for 11 angiosperm families. Since data from all five macromolecules are available for only three of the families, they used data sets based on combinations of three or four macromolecules. The trees from the different analyses were compared, and inconsistencies were discussed. One analysis included sequences from RBC-SSU, plastocyanin, and cytochrome c in nine families. One most parsimonious tree and 12 trees that were one or two steps longer were found. In all 13 trees, Chenopodiaceae and Polygonaceae appeared as sister groups, as did Fabaceae
and Brassicaceae. Martin et al. (1985 p. 399) concluded that these two sister-group relationships “appear to be very reliable.”

Despite the detailed analyses by Martin et al. (1985), uncertainty about the conclusions remains, because the extent of homoplasy in the data was not estimated. I have investigated the number of parallelisms and reversals necessary for explaining the data, and I discuss the equally or almost equally parsimonious cladograms (trees) possible in relation to the homoplasy involved. I do not consider the methods used by earlier authors to construct trees from amino acid sequence data (see Peacock, 1981). The methods differed and were much constrained by the algorithms and computer power available at the time (Boulter et al., 1972, 1979; Martin et al., 1983, 1985; Martin and Dowd, 1986a). Essentially, I use the same parsimony analysis as Martin et al. (1983, 1985; based on the algorithm of Hendy and Penny [1982]), but in addition to constructing cladograms, I seek to determine the extent of homoplasy in the amino acid sequences, how many cladograms are possible, and what reliable phylogenetic information is actually present in those cladograms.

**MATERIALS AND METHODS**

Two data matrices were analyzed. The first one (Fig. 1) includes the same nine angiosperm families discussed by Martin et al. (1985 fig. 1c) in their conclusions (Apiaceae, Brassicaceae, Caprifoliaceae, Solanaceae, Poaceae, Chenopodiaceae, Fabaceae, Asteraceae, and Polygonaceae). The sequences included are from RBC-SSU, plastocyanin, and cytochrome c and comprise 64 variable nucleotide positions, inferred from 40 amino acid positions. The second data matrix (Fig. 2) is limited to those six families for which there are sequence data for four proteins (RBC-SSU, plastocyanin, cytochrome c, and ferredoxin). The six families are Apiaceae, Brassicaceae, Caprifoliaceae, Poaceae, Chenopodiaceae, and Fabaceae, and the sequences comprise 82 variable nucleotide positions, inferred from 53 amino acid positions.

The data for each family generally come from several species (one species in some cases), the proteins of which have been sequenced. By performing preliminary analyses, Martin et al. (1983) established inferred familial sequences from the species sequences. Martin et al. (1983) converted the amino acid sequences to inferred nucleotide sequences before analysis, using the method outlined by Penny et al. (1980) and Fitch and Farris (1974). I used the family nucleotide sequences converted from RBC-SSU, plastocyanin, and cytochrome c amino acid sequences in the accession publication to Martin et al. (1983). For ferredoxin, each family is represented by one species, except the Fabaceae, which is represented by *Leucaena glauca* and *Medicago sativa*. Positions with different amino acids in these two
species have been coded to allow for this variation, generally leading to question marks (indicating unknown states in the data matrix). I converted ferredoxin amino acid sequences from Ramshaw (1982) and Nakano et al. (1981) to nucleotide sequences using the method of Penny et al. (1980) and Fitch and Farris (1974).

Naturally, complete nucleotide sequences cannot be deduced from the amino acid sequences. For example, the occurrence of most amino acids is not dependent on the third nucleotide position, which then is unknown. Converting amino acid sequences to inferred nucleotide sequences is done to estimate how many nucleotide changes are needed to accomplish the changes at each amino acid position. Since each amino acid is coded by a triplet, any particular amino acid change may depend on one, two, or three nucleotide changes. The converting procedure, combined with the practice of using the inferred variable nucleotide positions, is simply a form of weighting amino acid changes once, twice, or three times. I have followed this practice, and the columns (characters) in the data matrix thus consist of inferred nucleotide positions. For example, in the first data set (Fig. 1), amino acid changes at the second position in RCC-SSU may be explained by changes in one nucleotide position only, and are thus represented by a single column (weighted once) in the data matrix, whereas amino acid changes at the seventh position in RCC-SSU require changes in two nucleotide positions, and they are represented by two columns (weighted twice). Otherwise no weighting has been applied. All uninformative positions were excluded, i.e., synapomorphies for the whole group investigated (invariant positions) and autapomorphies (positions with a different amino acid in a single family only). The data matrices were analyzed with Swofford’s (1985) PAUP package, using the branch-and-bound algorithm of Hendy and Penny (1982) to generate cladograms minimizing the number of nucleotide changes (steps) necessary for explaining the data in a phylogenetic context. The cladograms were rooted with the single monocotyledonous family Poaceae as a hypothetical ancestor. This was done to make the cladograms more readable; naturally, there is no assumption that the dicotyledons arose from grasses or monocotyledons. The only assumption involved is that the dicotyledonous families included are more closely related to each other than any of them is to the Poaceae. This may be wrong, but alternative rootings will not affect the particular conclusions drawn from this study.

All most parsimonious cladograms, as well as the cladograms only a few steps longer, were generated (this is possible using the “bsave” option in PAUP). The cladograms were combined into strict-consensus trees, showing only those monophyletic groups present in all cladograms combined. A consistency index (Kluge and Farris, 1969) was computed for the most parsimonious cladograms. The consistency index is a measurement of how the data fit to the

<table>
<thead>
<tr>
<th>Apiaceae</th>
<th>AAUGCCCAACACAGCCGUGGCGAACGAGCAAGUUGAUCUGUCAYAGGAGAGAAGAARYCAAYACCGUUG</th>
<th>Brassicaceae</th>
<th>CAAGGAAGAAGUGAGGUGAUAUUAGGAGAAGAAGAGUUGGACACACAGAGAAAGAGGAGGGAGTCAGUGYGGTAAUGGACUACU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caprifoliaceae</td>
<td>AAUGGCUCUCAGCGUGGAGAUGAACACAGCCGACACUAUUAAAUUAYCUCAAUGCCCAAYUGGCRYYRAGGUGUAACAC</td>
<td>Poaceae</td>
<td>CGUCCUACCAACACAGCCGUGGACACACAGCCGACACUAUUAAAUUAYCUCAAUGCCCAAYUGGCRYYRAGGUGUAACAC</td>
</tr>
<tr>
<td>Chenopodiaceae</td>
<td>CGUCCUACCAACACAGCCGUGGACACACAGCCGACACUAUUAAAUUAYCUCAAUGCCCAAYUGGCRYYRAGGUGUAACAC</td>
<td>Fabaceae</td>
<td>CAAGGAAGAAGUGAGGUGGACACACAGCCGUGGACACACAGCCGACACUAUUAAAUUAYCUCAAUGCCCAAYUGGCRYYRAGGUGUAACAC</td>
</tr>
</tbody>
</table>
Fig. 3. One of the two most parsimonious cladograms (151 steps) derivable from the data matrix in Figure 1. Each symbol represents one nucleotide change (step). Bars represent synapomorphies, double lines represent parallelisms, and crosses represent reversals.

Fig. 4. One of the two most parsimonious cladograms (151 steps) derivable from the data matrix in Figure 1. Each symbol represents one nucleotide change (step). Bars represent synapomorphies, double lines represent parallelisms, and crosses represent reversals.

cladogram. The maximum value is 1, which corresponds to a complete fit; no homoplasy is involved, and changes in any particular character (position) appear only once on the cladogram, so that no parallelisms or reversals are needed. For example, the minimum number of nucleotide changes necessary to explain the data set in Figure 1 is 88, and if a cladogram of 88 steps is found, the consistency index is 1. A cladogram of, say, 100 steps has a consistency index of \( \frac{88}{100} = 0.88 \).

The minimum numbers of nucleotide changes necessary to explain the data sets in Figures 1 and 2 are 88 and 111, respectively. The reason the number of nucleotide changes is greater than the number of nucleotide positions or columns (64 and 82 in Figures 1 and 2, respectively) is simply that some of the columns contain more than two nucleotides. Naturally, such multistate characters (columns) were treated as unordered, i.e., with no assumptions about the order of nucleotide change. No attempt was made to adjust for the different probabilities of transitions versus transversions in nucleotide change; there is currently no way to do this when both types of changes are involved in the same (multistate) character.

RESULTS

The first data matrix, involving nine families and three proteins, yielded two most parsimonious cladograms (Figs. 3, 4), 151 steps long and with a consistency index of \( \frac{88}{151} = 0.583 \). The consistency index indicates that there is considerable homoplasy in the data, about the same amount as in many morphological data sets. Ideally, the data matrix requires 88 steps and 151 - 88 = 63 steps must be attributed to parallelisms and reversals. There are three more cladograms 152 steps long, nine more cladograms 153 steps long, and 24 more
cladograms 154 steps long. The four strict-consensus trees for the cladograms 151, 151–152, 151–153, and 151–154 steps long are shown in Figure 5. The last consensus tree is totally collapsed and shows that there is no monophyletic group common to all cladograms that are up to three steps longer than the most parsimonious ones.

The second data matrix involving six families and four proteins yielded a single most parsimonious cladogram (Fig. 6) 161 steps long and with a consistency index of $111/161 = 0.689$. This is a somewhat better value than in the first analysis. Ideally, the second data matrix requires 111 steps and $161 - 111 = 50$ steps must be attributed to parallelisms and reversals. There is one more cladogram 162 steps long (Fig. 7), and there are three more cladograms 163 steps long. The strict-consensus trees for the two shortest cladograms, as well as for all five cladograms, are given in Figure 8. The second consensus tree shows that there is no monophyletic group common to the five cladograms 161–163 steps long.

**Discussion**

The most parsimonious cladograms represent the best phylogenetic hypotheses. Nevertheless, it is desirable to explore the stability of the data behind the cladograms. How many changes are needed in the data matrix to yield different cladograms, how different are the cladograms, and what monophyletic groups are common to all of them? Fully resolved (without polytomies), most parsimonious cladograms may serve as bold phylogenetic hypotheses and a framework for character evaluation and further research. Such cladograms have a mix-
parts that are not well supported, we may wish to restrict the phylogenetic hypothesis based on sequence data to the well-supported monophyletic groups present in all cladograms generated and shown by the strict-consensus trees. With morphological data, consensus trees are not enough. Parsimonious cladograms are necessary as a reference scheme for character interpretation.

Considerable homoplasy is present in the amino acid sequence data, and many parallelisms and reversals are needed to fit the data to a cladogram. This is no surprise; Peacock (1981) showed that parallelisms in amino acid substitutions in plant proteins are common. However, it is more remarkable that we need to postulate so few extra parallelisms (or reversals) to fit the data to cladograms with topologies quite different from those of the most parsimonious ones. In the first analysis, there are 63 parallelisms and reversals in the most parsimonious cladogram, and only three more are needed to arrive at 38 cladograms without a single monophyletic group common to all of them.

In the second analysis, the situation is similar: there are 50 parallelisms and reversals, and only two more need be added to arrive at five cladograms without a single monophyletic group common to all of them.

A single change in one nucleotide position in one taxon may be enough to yield a most parsimonious cladogram with a different topology. The two data matrices contain $9 \times$
64 = 576 and 6 × 82 = 492 data entries, respectively. Two or three changes in data matrices of these sizes must be considered as very limited, yet they may be enough to result in several cladograms with completely different sets of monophyletic groups. Examination of the amino acid sequences in the different species from one family shows that they differ often in several positions (e.g., Martin et al., 1983). For example, the question marks in the bottom (Fabaceae) row of Figure 2 indicate such differences between *Leucaena glauca* and *Medicago sativa* in the Fabaceae. There are also positions variable within a single species. A different sampling of species or the addition of data from more species could easily result in changes in the familial sequences.

Furthermore, the most parsimonious cladograms from the two analyses are inconsistent. In the second analysis, Chenopodiaceae and Apiaceae are sister groups, whereas in the first analysis, Chenopodiaceae appears (together with Polygonaceae, which was excluded from the second analysis) as the sister group to the remaining dicotyledonous families in the analysis. We need to accept a cladogram at least two steps longer in either analysis in order to achieve consistent results. The inconsistency between the results of the two analyses is caused by the reduction of taxa from nine families to six families, rather than by the addition of the ferredoxin sequences in the second analysis. A third analysis with six families and without the ferredoxin sequences (Fig. 2, excluding the last 32 columns) yields the same result as the second analysis (see Fig. 8), which involved six families and included ferredoxin sequences. Thus, homoplasy in the data leads to inconsistent results when analyses are made with different combinations of taxa.

Felsenstein (1985) proposed the application of the bootstrap statistical method for inferring statistical confidence limits on the monophyletic groups suggested by a cladistic analysis. Similar methods are suggested by Sneath (1986). I have not performed a bootstrap analysis of the two data matrices, but the instability of the data, as outlined above, indicates clearly that no alleged monophyletic group shown in Figures 3–8 would be statistically significant.

I conclude that, with protein sequence data, it is not enough to compute the most parsimonious cladograms and to accept these as phylogenetic hypotheses without examining cladograms a number of steps longer. Not only the shortest cladograms, but also those with an increasing number of steps should be combined into strict-consensus trees as is done in Figures 5 and 8. Only those groups present in the consensus trees may be hypothesized to be monophyletic with any confidence. There is no easy way to determine how many extra steps should be allowed and, subsequently, which one of the increasingly collapsed consensus trees should be accepted as a base for well supported, cautious, phylogenetic hypotheses. This depends on the size of the data matrix and the type of data involved. In the two analyses performed here, it does not seem overcautious to allow three extra steps.

In addition to using unweighted data, Boulter et al. (1979) and Martin et al. (1983, 1985) weighted the amino acid positions in an effort to reduce the effects of homoplasy. Positions that were variable within single species and positions with a high ratio of observed/expected incompatibilities with the majority were less weighted. The latter type of weighting approaches the compatibility method and deviates from a preferred parsimony analysis (Farris, 1983). In any case, weighting did not yield significantly different results according to Martin et al. (1983, 1985).

It is important to note that Martin et al. (1985) used majority-rule consensus trees (Margush and McMorris, 1981; Penny et al., 1982), rather than strict-consensus trees. Their rationale for doing this is obscure. Majority-rule consensus trees contain those monophyletic groups that appear in a majority of the cladograms included. Majority-rule consensus trees are generally fully resolved into dichotomies (if the cladograms are fully resolved) and are incompatible with a number of the cladograms. Strict-consensus trees are generally partly collapsed into polytomies and are compatible with all the cladograms; they contain those monophyletic groups that appear in all cladograms (not only in the majority). In my opinion, majority-rule consensus trees are insufficient as bases for phylogenetic hypotheses.
They are incompatible with equally parsimonious cladograms, i.e., equally probable phylogenetic hypotheses, which are rejected on the very doubtful ground that they do not conform to the majority in the set of equally probable ones.

Martin et al. (1985) discussed the problem of whether a single protein with about 40 amino acid positions sequenced would be enough for obtaining meaningful results. An analysis of monocotyledons by Martin and Dowd (1986a) was based on sequences of only one protein, RBC-SSU. Martin and Down submitted that this may turn out to be sufficient when more species from each family have been sampled. They also stressed that, in an analysis of RBC-SSU sequences of the order Myrtales, the results conformed well with the expectations from taxonomy (Martin and Dowd, 1986b). However, in most cases, 40 positions (about half of which may be informative [variable]) from a single protein are not enough for phylogenetic reconstruction. On the contrary, the two analyses performed here, with 40 and 53 informative amino acid positions from three and four proteins, respectively, did not yield a single reasonably well-supported monophyletic group.

As seen in Figures 3–8, Brassicaceae and Fabaceae always come out as sister groups, except in the two totally collapsed consensus trees. This alleged and (from morphology) unexpected relationship was also pointed out by Martin et al. (1985 p. 399), who stated that it appears to be “very reliable.” It should be clear from the discussion above that I do not consider this hypothesis to be well supported. This is disappointing, because precisely this kind of family interrelationship is needed for progress in understanding angiosperm phylogeny. Despite the fact that very few families are included, amino acid sequence data may serve as a very important piece of phylogenetic information, if it can be shown clearly that, among a number of families, two are more closely related to each other than either is to the others. However, at present, amino acid sequence data are not comprehensive enough. A similar situation is encountered in reconstruction of mammal phylogenies based on amino acid sequence data, as shown by Wyss et al. (1987).

We need more informative positions, either from sequencing longer parts of the proteins or from sequencing more proteins from the same taxa. In addition, cladistic analysis that includes evaluation of all equally and almost equally parsimonious cladograms is necessary.

ACKNOWLEDGMENTS

I thank B. Bremer, M. Donoghue, and one anonymous reviewer for valuable comments on earlier versions of the manuscript and D. Butoi and M. Källersjö for preparing the illustrations.

LITERATURE CITED


——. 1984b. The study of plant phylogeny using amino acid sequences of ribulose-1,5-biphosphate carboxylase. IV. Proteaceae and Fabaceae and the
The complete amino acid sequence of parsley (Petroselinum sativum) ferredoxin. J. Biochem. 90:1725–1730.


Corresponding Editor: K. J. Niklas