Data Decisiveness, Data Quality, and Incongruence in Phylogenetic Analysis: An Example From the Monocotyledons Using Mitochondrial atp A Sequences

JERROLD I. DAVIS,† MARK P. SIMMONS,† DENNIS W. STEVENSON,‡ AND JONATHAN F. WENDEL§

†L. H. Bailey Hortorium, Cornell University, Ithaca, New York 14853, USA
‡New York Botanical Garden, Bronx, New York 10458, USA
§Department of Botany, Iowa State University, Ames, Iowa 50011, USA

Abstract.—We examined three parallel data sets with respect to qualities relevant to phylogenetic analysis of 20 exemplar monocotyledons and related dicotyledons. The three data sets represent restriction-site variation in the inverted repeat region of the chloroplast genome, and nucleotide sequence variation in the chloroplast-encoded gene rbcL and in the mitochondrial-encoded gene atp A, the latter of which encodes the α-subunit of mitochondrial ATP synthase. The plant mitochondrial genome has been little used in plant systematics, in part because nucleotide sequence evolution in enzyme-encoding genes of this genome is relatively slow. The three data sets were examined in separate and combined analyses, with a focus on patterns of congruence, homoplasy, and data decisiveness. Data decisiveness (described by P. Goloboff) is a measure of robustness of support for most parsimonious trees by a data set in terms of the degree to which those trees are shorter than the average length of all possible trees. Because indecisive data sets require relatively fewer additional steps than decisive ones to be optimized on nonparsimonious trees, they will have a lesser tendency to be incongruent with other data sets. One consequence of this relationship between decisiveness and character incongruence is that if incongruence is used as a criterion of noncombinability, decisive data sets, which provide robust support for relationships, are more likely to be assessed as noncombinable with other data sets than are indecisive data sets, which provide weak support for relationships. For the sampling of taxa in this study, the atp A data set has about half as many cladistically informative nucleotides as the rbcL data set per site examined, and is less homoplastic and more decisive. The rbcL data set, which is the least decisive of the three, exhibits the lowest levels of character incongruence. Whatever the molecular evolutionary cause of this phenomenon, it seems likely that the poorer performance of rbcL than atp A, in terms of data decisiveness, is due to both its higher overall level of homoplasy and the fact that it is performing especially poorly at nonsynonymous sites. [atp A; character incongruence; cladistics; data decisiveness; monocotyledons; plant mitochondrial DNA; rbcL; restriction sites.]

The success of phylogenetic analysis depends upon the discovery of character sets that provide accurate information regarding phylogenetic relationships among natural lineages. It would be fortunate if character sets could be evaluated directly in terms of the accuracy with which they reconstruct relationships, but this cannot be accomplished unless phylogenetic relationships already are known. Indeed, the laboratory generation of lineages of known phylogenetic structure has facilitated this sort of evaluation (Hillis et al., 1992; Hillis, 1995; Cunningham et al., 1997), but in these cases branch lengths and other factors have been controlled, and it is difficult to extrapolate the findings to natural phylogenies. A related approach, also dependent upon prior knowledge of actual relationships, is to determine the ability of a particular character set (or method of analysis) to resolve relationships correctly among a set of natural lineages that are few enough in number, and among which relationships are trivial enough, that prior notions of relationship verge on certainty (e.g., Huelsenbeck and Hillis, 1993). However, when natural lineages are sampled in any nontrivial depth, and particularly where credible and conflicting hypotheses of relationship already exist, the veracity of results obtained by a particular character set remains
unknown. In such cases, a character set can be evaluated in terms of its degree of incongruence with other, presumably independent, character sets (Mickevich and Farris, 1981; Kluge, 1989; Swofford, 1991; Caputo et al., 1992; Farris et al., 1994; de Queiroz et al., 1995; Miyamoto and Fitch, 1995; Nixon and Carpenter, 1996a). Although incongruence analysis does not provide direct evidence regarding the veracity of different character sets, it can identify patterns of conflict among them, and thereby help to focus additional studies that may clarify the underlying causes of incongruence.

Apart from the matter of veracity, robustness of support for a particular set of relationships by a data set is itself a quality of interest. Systematists frequently calculate a variety of measures of internal consistency of data sets, including the consistency index (CI; Kluge and Farris, 1969) and retention index (RI; Farris, 1989), and they often report the robustness of support by data sets for individual clades, as gauged by metrics such as bootstrap frequencies (Felsenstein, 1985) and Bremer support (Bremer, 1988, 1994). There is considerable disagreement on the question of whether high levels of data consistency and robustness of support for particular clades should be interpreted as indicators of veracity (e.g., Carpenter, 1992; Hillis and Bull, 1993; Kluge and Wolf, 1993; Sanderson, 1995; Givnish and Sytsma, 1997). Without engaging in that debate, we simply note that apart from other considerations, a character set that provides robust support for a limited set of relationships is valuable for the same reasons that strong hypotheses are preferable to weak ones (Popper, 1959, 1963; Farris, 1983). A character set with these attributes can refute others, and can be refuted by others, while a character set that is consistent with a large number of mutually conflicting hypotheses is less subject to refutation, and therefore less useful for scientific inference (Nixon and Carpenter, 1996b).

Here we provide a comparative analysis of intrinsic attributes and incongruence relationships among three molecular character sets that represent variation within a major angiosperm lineage, the monocotyledons, or class Liliopsida of Cronquist (1981). The first of the three character sets is a previously published set of mapped restriction sites from the inverted repeat (IR) region of the chloroplast genome (Davis, 1995). The other two are nucleotide sequence character sets, one of them the familiar chloroplast-encoded gene rbcL, the other the mitochondrion-encoded gene atpA.

Among the intrinsic attributes that we examine for these character sets is data decisiveness (Goloboff, 1991), an expression of overall robustness of support by a data set for its most parsimonious trees. As suggested by the preceding comments regarding strong and weak hypotheses, we believe that decisiveness is a useful criterion of character quality. In particular, we suggest that decisive data sets are more likely than indecisive ones to be incongruent with other data sets, and that any interpretation of the results of incongruence analysis as evidence that data sets should not be combined may have the undesirable effect of favoring the use of less decisive data sets over those that are more decisive.

**DATA DECISIVENESS**

Every cladistic data matrix resolves one or more most parsimonious cladograms. Two different matrices may resolve the same set of most parsimonious trees, yet still differ in the degree of robustness with which they favor those trees, relative to other possible relationships. Goloboff (1991) discussed this matter, indicated that a matrix that provides robust support for a set of relationships should be recognized as decisive, in a general sense, and described an index of data decisiveness (DD) that measures this quality. For any data matrix,

\[
DD = \frac{S - S}{S - M}
\]
where $S$ is the mean length of all possible cladograms, $S$ is the length of the most parsimonious cladogram or cladograms, and $M$ is the observed variation in the matrix, this being the sum of the minimum possible number of steps for each character. The denominator of this equation, the difference between $S$ and $M$, delimits the range of possible lengths of most parsimonious trees, which can be no longer than $S$ (a limit that is reached when all possible trees are equal in length), and no shorter than $M$ (a limit that is reached when there is no homoplasy). The numerator, the difference between $S$ and $S$, describes the position of the most parsimonious tree or trees along the axis that is delimited by the denominator. When $S$ approaches $M$, the magnitude of the numerator approaches that of the denominator, and $DD$ approaches 1.0; the matrix is consistent with relatively few trees, and it is decisive. When $S$ approaches $S$, the numerator approaches zero, and $DD$ approaches 0; the matrix discriminates only weakly among possible trees, and it is indecisive. In the extreme case, all possible trees are equal in length (see Goloboff, 1991, for a discussion of this possibility), and the matrix is completely indecisive. Real data sets of nontrivial size lie between these extremes; there is some inconsistency among characters, and some trees are more parsimonious than others.

In the computation of $DD$ for data sets with more than a few taxa, the lengths of all possible trees cannot be determined, and the mean length of all possible trees is estimated by determining the lengths of a random sample of arbitrarily resolved trees. This procedure also is utilized in the measure of skewness in the distribution of tree lengths (Hillis, 1991; Hillis and Huelsenbeck, 1992).

As noted by Goloboff (1991), and as should be evident from the foregoing comments, $DD$ is related to data consistency, but it reflects other qualities of a data matrix as well, and therefore it cannot be predicted entirely from CI or RI. This is evident if a hypothetical data set is considered, in which there are numerous characters, of which all except one (character X) are mutually consistent. The data set determines a single most parsimonious tree, and character X is the only character that is not consistent with that tree. State 1 of character X occurs in two taxa that are not sister taxa in the most parsimonious tree (Fig. 1a); there are two transformations from state 0 to state 1 in this character (or one transformation from 0 to 1, and one from 1 to 0), and the CI of this character is 0.5. Because character X is homoplastic with respect to this tree, the ensemble CI and RI of the data set are less than 1.0. Any alternative tree in which the two taxa with state 1 of character X are united as a monophyletic group has only one step for this character, and consequently a CI and RI of 1.0 for this character, but any such tree requires additional steps in other characters, is less parsimonious overall, and has a lower ensemble CI. This data set is now compared to a second one that is identical to it except that one of the occurrences of state 1 in character X is in a different taxon that is situated a greater distance from the other taxon with that state in the most parsimonious tree (Fig. 1b). As in the first data set, character X has two steps, and it is the only homoplastic character, so the ensemble CI and RI of this data set are identical to those of the first. However, any rearrangement of the tree that places the two taxa with

![Figure 1](http://sysbio.oxfordjournals.org/)
state 1 together, and causes the number of steps in character X to drop from 2 to 1, requires a more drastic modification of the most parsimonious tree than in the first example, and therefore involves a greater increase in the number of steps in other characters in the data set. Thus, the two data sets are equally homoplastic, and resolve the same most parsimonious tree, but the second data set provides more robust support for this tree, and is more decisive.

**Mitochondrial atp A and Plant Systematics**

The mitochondrial genome has been employed extensively in systematic studies of animals, fungi, and other taxa (e.g., Avise, 1994; Hafner et al., 1994; Freeman and Zink, 1995; Crandall and Fitzpatrick, 1996; Skupski et al., 1997; Waters and White, 1997; Wilcox et al., 1997). In contrast, its use in plant systematics has been minimal. Although the potential utility of the mitochondrial genome in plant systematics has been discussed (Crozier, 1990; Palmer, 1992), and some analyses using characters of this genome have been conducted, (e.g., Dong and Wagner, 1993; Hiesel et al., 1994; Senda et al., 1995; Guo et al., 1996; Malek et al., 1996; Pesole et al., 1996; Bowe and DePamphilis, 1997), these studies have been relatively few. The chloroplast genome, as is well known to plant systematists, is structurally conservative; in contrast, plant mitochondrial genomes are complex and varied in structure (Siculella and Palmer, 1988; Palmer, 1992; Palmer et al., 1992, and citations therein). Against the backdrop of a structurally dynamic genome, however, nucleotide sequences of protein-encoding genes in the plant mitochondrial genome evolve slowly (Wolfe et al., 1987, 1989; Palmer and Herbon, 1988; Eyre-Walker and Gaut, 1997), and this quality suggests that they may provide useful evidence of higher level relationships.

A few complete plant mitochondrial sequences are available for comparative studies (e.g., *Marchantia polymorpha* [Oda et al., 1992], *Prototheca wickerhamii* [Wolff et al., 1994]). Mitochondrial genomes include ribosomal RNA genes, transfer RNA genes, ribosomal protein genes, genes coding for enzymes involved in respiration and oxidative phosphorylation, plus a variety of open reading frames. Palmer (1992) identified five protein-encoding genes of the mitochondrial genome, plus the ribosomal RNA genes, as particularly promising for plant systematics. Three qualities that recommend the five protein-encoding genes are their relatively large sizes (each 800 bp or larger), their lack of introns (among angiosperms), and the conservative nature of nucleotide sequence evolution in their coding sequences. These features suggested to us that one or more of the genes identified by Palmer as promising would be useful in the analysis of large-scale phylogenetic relationships within angiosperms, not as a solitary source of data, but as complements to other character sets.

Of the five protein-encoding genes identified by Palmer, three are between 1,500 and 1,600 bp in length (the other two are smaller), and from these two we selected *atp A* (sometimes referred to as *atp1*), which encodes the α-subunit of mitochondrial ATP synthase, and which has been the subject of an analysis of evolutionary rate heterogeneity (Eyre-Walker and Gaut, 1997). This gene, which is similar in length to the widely employed *rbc L*, is conservative in size and sequence within the angiosperms. For example, sequence identity between a species of the rosid dicot *Oenothera* (Onagraceae) and the commelinid monocot *Zea* (Poaceae) is 92.2% (Isaac et al., 1985; Schuster and Brennicke, 1986), and the length of *atp A* is 1,524 bp in *Zea*, and 1,533 bp in *Oenothera*. Alignment of these two sequences is a trivial matter, for the nine additional nucleotides in *Oenothera* occur just beyond the 3' terminus of the corresponding sequence of *Zea*; hence, the 1,524 nucleotides in *Zea* align, without gaps, with the first 1,524 in
Oenothera. Other available atpA sequences also are quite similar (see Senda et al., 1993), and this pattern holds, in general, for the data presented here, with the notable exception of four taxa with deletions in one region of the gene.

MONOCOT SYSTEMATICS

The monocotyledons, with about 50,000–65,000 species (Cronquist, 1981; Takhtajan, 1997), and including such diverse groups as the palms, grasses, and orchids, are widely recognized as a monophyletic group within the angiosperms, but many aspects of phylogenetic structure within the monocots remain obscure. Although there is strong evidence for the monophyly of several lineages within the monocots (e.g., Poales, Zingiberales), there are many dramatic differences among the higher level groupings circumscribed in comprehensive taxonomic accounts (e.g., Cronquist, 1981; Dahlgren et al., 1985; Thorne, 1992; Takhtajan, 1997). Formal cladistic analyses of the monocots, based on variation in morphology (Stevenson and Loconte, 1995), 18S ribosomal DNA sequences (Bharathan and Zimmer, 1995; also sampled by Soltis et al., 1997), restriction-site variation of the chloroplast inverted repeat (IR) region (Davis, 1995), and nucleotide sequences of the chloroplast-encoded genes rbcL (Chase et al., 1993; Duvall et al., 1993a) and rps4 (Nadot et al., 1995), have yielded conflicting results. In a preliminary combined analysis of morphological characters and rbcL, Chase et al. (1995) emphasized the need for further sampling of taxa and characters, particularly the latter.

MATERIALS AND METHODS

Data Collection

The restriction-site data used in this study are a subset of those published by Davis (1995), which included 89 cladistically informative sites and 2 informative length variants for a set of 53 accessions. Neither of the length variants is cladistically informative among the 20 taxa in the present analysis, so this portion of the present data set consists entirely of restriction sites. Ten of the 42 sites that are informative among the present taxon sample are in the ORF2280 region, which is deleted in Oryza (Shinozaki et al., 1986; Hiratsuka et al., 1989; Shimada and Sugiuara, 1991; Downie et al., 1994; Davis, 1995), and they are scored missing in that taxon.

The rbcL sequences were obtained from public domain sources (Table 1). The portion used in the present analysis consists of 1,398 nucleotide positions, corresponding to positions 31 (relative to the first position of the start codon in Oryza) through 1,428. Sequences that did not include this entire portion of the gene were scored as missing for unreported positions. Three other changes were made to the published sequences: The Piper sequence appears to be missing one nucleotide, and it was aligned with others by inserting a gap of one nucleotide (scored missing) at position 681; three nucleotides scored as subset polymorphisms or ambiguities in the Acorus sequence were rescored as missing; and the last 17 nucleotides at the 3’ end of the Gymnostachys sequence, which appeared to be anomalous, were rescored as missing.

The atpA data set includes two previously published sequences (both by Eyre-Walker and Gaut, 1997) and 18 sequences determined for the present study (Table 1); the new sequences were generated from the same DNA accessions used by Davis (1995). The portion of atpA examined in this study is 1,266 nucleotides in length (except in four accessions that have deletions), corresponding to positions 98 through 1,363 of the coding region in Oryza (Kadowaki et al., 1990); site 98 is the 305th nucleotide in the published Oryza sequence, Genbank accession X51422, which includes the coding region of the gene and flanking regions. The 18 new sequences were obtained using standard manual and automated sequencing procedures, prin-
TABLE 1. Sources of \textit{rbcL} and \textit{atpA} sequence data used in analysis of 18 monocot and two dicot taxa. Classification of dicots (Magnoliopsida) at and above the family level is according to Cronquist (1981), and that of monocots (Liliopsida) is according to Dahlgren et al. (1985), except that superorders are designated by the suffix -anae. Three-letter codes in capital letters are listed with names of subclasses and superorders. Three-letter codes in lower case, listed after some family names, refer to orders within monocot superorders sampled more than once: ara = Arales, asp = Asparagales, bro = Bromeliales, com = Commelinales, dio = Dioscoreales, hae = Haemodorales, lil = Liliales, mel = Melanthiales, poa = Poales, vel = Velloziales. Genbank accession numbers are listed for sequences deposited there. Species sampled for \textit{atpA} are the same as those sampled for restriction sites (see Davis, 1995).

<table>
<thead>
<tr>
<th>Class</th>
<th>Subclass or superorder</th>
<th>Family</th>
<th>Genus</th>
<th>\textit{rbcL} Accession</th>
<th>\textit{atpA} Accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAGNOLIOPSIDA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Magnoliidae (MAG)</td>
<td></td>
<td>Piperaeae</td>
<td>\textit{Piper} L.</td>
<td>L12660</td>
<td>AF039243</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Aristolochiaceae</td>
<td>\textit{Saruma} Oliver</td>
<td>L12664</td>
<td>AF039242</td>
</tr>
<tr>
<td>LILIOPSIDA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lilianae (LIL)</td>
<td></td>
<td>Taccaceae (dio)</td>
<td>\textit{Tacc} Forster &amp; G. Forster</td>
<td>Chase et al., 1993$^a$</td>
<td>AF039252</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Trilliaceae (dio)</td>
<td>\textit{Trillium} L.</td>
<td>D28167</td>
<td>AF039253</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Smilacaceae (dio)</td>
<td>\textit{Smilax} L.</td>
<td>D28333</td>
<td>AF039251</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Xanthorrhoeaceae (asp)</td>
<td>\textit{Xanthorrhoea} Smith</td>
<td>Z73710</td>
<td>AF039250</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hypoxidaceae (asp)</td>
<td>\textit{Curculigo} Gaertner</td>
<td>Z73701</td>
<td>AF039249</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Melanthiaceae (mel)</td>
<td>\textit{Veratrum} L.</td>
<td>D28168</td>
<td>AF039255</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Alstroemeriaceae (lil)</td>
<td>\textit{Alstroemeria} L.</td>
<td>Chase et al., 1993$^a$</td>
<td>AF039254</td>
</tr>
<tr>
<td>Aranae (ARA)</td>
<td></td>
<td>Araceae (ara)</td>
<td>\textit{Acorus} L.</td>
<td>M91625</td>
<td>AF039256</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gymnostachys R. Br.</td>
<td>\textit{Gymnostachys} R. Br.</td>
<td>M91629</td>
<td>AF039244</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Symplocarpus Salisb.</td>
<td>\textit{Symplocarpus Salisb.}</td>
<td>L10247</td>
<td>AF039245</td>
</tr>
<tr>
<td>Bromelianae (BRO)</td>
<td></td>
<td>Velloziaceae (vel)</td>
<td>\textit{Vellozia} Vand.</td>
<td>L19970</td>
<td>AF039247</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bromeliaceae (bro)</td>
<td>\textit{Catopsis} Griseb.</td>
<td>L19976</td>
<td>AF039257</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Haemodoraceae (hae)</td>
<td>\textit{Anigozanthos} Labill.</td>
<td>Chase et al., 1993$^a$</td>
<td>AF039246</td>
</tr>
<tr>
<td>Zingiberanae (ZIN)</td>
<td></td>
<td>Cannaceae</td>
<td>\textit{Canna} L.</td>
<td>L05445</td>
<td>AF039259</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cannaceae</td>
<td>\textit{Canna} L.</td>
<td>L05445</td>
<td>AF039259</td>
</tr>
<tr>
<td>Commelinanae (COM)</td>
<td></td>
<td>Eriocaulaceae (com)</td>
<td>\textit{Eriocaulon} L.</td>
<td>L10252</td>
<td>AF039258</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Flagellariaceae (poa)</td>
<td>\textit{Flagellaria} L.</td>
<td>L12678</td>
<td>AF039248</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poaceae (poa)</td>
<td>\textit{Oryza} L.</td>
<td>D00207</td>
<td>X51422</td>
</tr>
<tr>
<td>Arecaeanae (ARE)</td>
<td></td>
<td>Arecaceae</td>
<td>\textit{Nypa} Steck</td>
<td>M81813</td>
<td>U58833</td>
</tr>
</tbody>
</table>

$^a$ \textit{rbcL} sequences used by Chase et al. (1993) but not available in Genbank; they were provided by V. Albert.
TABLE 2. Primer set used for amplification and sequencing of mitochondrial atpA. All primers were designed by B. Gaut, except two designed by the present authors specifically for Acorus (and named as such). Primers with names beginning with F prime toward the 3' end of the gene, and those beginning with B prime toward the 5' end.

<table>
<thead>
<tr>
<th>Name</th>
<th>Nucleotide position</th>
<th>Primer sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>77-97</td>
<td>5'-AAGTGGATGAGATCGGTGCAG-3'</td>
</tr>
<tr>
<td>F2</td>
<td>283-302</td>
<td>5'-TCTATTTGATGTTCTCTGC-3'</td>
</tr>
<tr>
<td>F3</td>
<td>466-485</td>
<td>5'-GATAGCTGTGTCCTATAGG-3'</td>
</tr>
<tr>
<td>F3(Acorus)</td>
<td>466-485</td>
<td>5'-GATA(AG)TCTAGTTCCATAGG-3'</td>
</tr>
<tr>
<td>F5</td>
<td>778-797</td>
<td>5'-TTCCGGGATAATGGAATGCAG-3'</td>
</tr>
<tr>
<td>F6</td>
<td>1069-1088</td>
<td>5'-ACAGAGCTTTTTATCCGGG-3'</td>
</tr>
<tr>
<td>B1</td>
<td>1379-1364</td>
<td>5'-GCATTGCATACAGAG-3'</td>
</tr>
<tr>
<td>B2</td>
<td>1287-1267</td>
<td>5'-TCTGTAAGCGCTTACCCACCTC-3'</td>
</tr>
<tr>
<td>B3</td>
<td>1091-1073</td>
<td>5'-ATCCGGCGATAAAAGAGCT-3'</td>
</tr>
<tr>
<td>B4</td>
<td>939-920</td>
<td>5'-AGCGGCTTCCTTCTAAGAGAC-3'</td>
</tr>
<tr>
<td>B4(Acorus)</td>
<td>939-920</td>
<td>5'-TGCTGTCCTTTCTAATAGAC-3'</td>
</tr>
<tr>
<td>B5</td>
<td>486-467</td>
<td>5'-GCCATAGGAAACCAGCTAT-3'</td>
</tr>
<tr>
<td>B6</td>
<td>302-284</td>
<td>5'-GCAAGGACATCCACATAG-3'</td>
</tr>
</tbody>
</table>

Data Analysis

Sixteen of the 20 atpA sequences were identical in length, including that of Oryza (which is used as a standard for designating nucleotide positions). All four of the taxa with anomalous lengths exhibited deletions, relative to Oryza, all in the region between sites 585 and 603, inclusive. The length of the deletion is 3 bp in three of these taxa (Acorus, Catopsis, and Eriocaulon) and 6 bp in the fourth (Canna). Alignment of the deletions was determined to be ambiguous, and all nucleotides in the region (9 contiguous nucleotide positions, numbers 585 through 593 in the three taxa with 3-bp deletions, and 17 positions, 587 through 603, in Canna) were scored missing in these taxa. Although the precise placement of each of the three 3-bp deletions was ambiguous, it was possible for all three to be aligned in the same set of positions. Therefore, the 3-bp deletions were hypothesized to be homologous, and the deletion was included as an additional binary character in the matrix, with all three taxa with 3-bp deletions scored as deleted, and all other taxa except Canna scored as undeleted. Some of the potential alignment positions of the 3-bp deletions fall within the range of potential alignment positions for the 6-bp deletion in Canna, so it is possible that Canna shares a 3-bp deletion with one or more of the other three deleted taxa, plus a second 3-bp deletion. However, some of the potential alignments of the 6-bp deletion do not overlap some of those of the 3-bp deletions. Also, the 6-bp deletion may have arisen in a single event, in which case it would not be homologous with a 3-bp deletion, unless the latter was itself a transformed state of the 6-bp deletion. For these reasons, homology of the 6-bp deletion in Canna with the 3-bp deletions in the other three accessions was considered doubtful, and Canna was scored unknown for the 3-bp deletion character. To determine if this scoring influenced the results, each of the four cladistic analyses that included the atpA data (see later discussion) was conducted two additional times, once with Canna scored deleted, and once with it scored undeleted. Each of these eight analyses yielded the same trees that were
obtained when the data were analyzed with Canna scored as unknown, and these alternative analyses are not discussed further.

MEGA (Kumar et al., 1993) was used to calculate sequence variability along the length of \textit{atpA}, using nonoverlapping 25-bp windows, and to calculate Kimura two-parameter distances (Kimura, 1980) for \textit{atpA} and \textit{rbcL}.

It is widely recognized that the ensemble CI of a data set is higher if autapomorphies are included in the calculation, because each autapomorphic character has a CI of 1.0 on any tree. Autapomorphies also influence incongruence indices that are based on relative lengths of data sets on different trees, for the difference in length between a most parsimonious tree and another tree is proportionally less when each is increased by the same absolute number of autapomorphic steps (e.g., 50 steps vs. 60 steps with autapomorphies excluded, a 20% increase, and 100 steps vs. 110 steps with 50 autapomorphies included, a 10% increase). Another consideration relative to CI and incongruence analysis is that the removal of uninformative characters does not eliminate all autapomorphies from a matrix, because autapomorphic states (i.e., those that occur in just one taxon) may be present in informative multistate characters. For example, among a set of 21 taxa scored for a particular nucleotide sequence character there may be 10 with state A, 10 with state T, and one with state G. This character is informative, because each of two states occurs in two or more taxa, but the solitary occurrence of G is uninformative, and like any other autapomorphy it adds one step to any tree. Because an informative multistate character can have uninformative states, but an informative binary character cannot, the removal of uninformative characters eliminates all uninformative variation from a matrix of binary characters, but uninformative variation may remain in a matrix consisting wholly or partly of multistate characters, even after all uninformative characters have been removed. The effect of this distinction may be substantial when matrices composed predominantly or entirely of binary characters (e.g., the restriction-site character set in this study) are compared to those composed of multistate characters (e.g., nucleotide sequence character sets). The predictable bias in such comparisons is for a nucleotide sequence data set to have a higher apparent CI than other data sets, and to exhibit a lower apparent degree of character incongruence with other data sets. This bias can be removed by rescoring every uninformative state that occurs in an informative multistate character (i.e., every cell with a unique state in any particular column of a data matrix) as missing. This procedure decreases the length of every possible tree by one step for each cell that is modified, but does not affect the results of cladistic analysis if the affected characters are analyzed as nonadditive (i.e., unordered), and if no other form of transition weighting is conducted. For the present analysis, uninformative characters were removed with the \textit{mop} command of DADA (Nixon, 1995), and uninformative states of informative characters were rescored as missing with the \textit{mex} command.

The three data sets (restriction sites, \textit{rbcL}, and \textit{atpA}) were analyzed separately and in all combinations of two or three as a combined matrix. Cladistic analyses were conducted with NONA (Goloboff, 1993), using the default settings \textit{amb} (clades resolved only if support is unambiguous) and \textit{poly} (polytomies allowed), and using \textit{Piper} as the outgroup for purposes of rooting (Nixon and Carpenter, 1994). Each tree search involved 1,000 subssearches, with each subsearch comprising construction of a Wagner tree using a random taxon entry sequence, followed by tbr swapping with up to 10 most parsimonious trees retained from each replicate that yielded most parsimonious trees (\textit{hold}/10 mult*1000); after all replicate searches had been conducted, shortest trees retained from the
subsearches were tbr swapped to completion \((\text{max} \times 10)\).

Support for various clades was assessed by bootstrap analysis (Felsenstein, 1985), using two methods to assess the occurrence of clades within bootstrap replicates. Most current implementations, such as the one in PAUP (Swofford, 1993), use an approach (here termed the frequency-within-replicates bootstrap) in which the score assigned to each clade in each bootstrap replicate reflects its frequency of occurrence among most parsimonious trees resolved by that replicate. An alternative method, the strict-consensus bootstrap, assigns the score of 1.0 (i.e., 100%) to every clade that occurs in all most parsimonious trees resolved by that replicate, and 0 in all other cases, including occurrence in some but not all most parsimonious trees. If implementations of the two methods involve equally effective tree searches, strict-consensus bootstrap scores must be less than or equal to frequency-within-replicates bootstrap scores for all clades. The strict-consensus bootstrap was conducted with CLADOS (Nixon, 1993), running NONA as a daughter process for tree searches. One thousand bootstrap replicates were conducted with the same ambiguity and polytomy settings as in the basic analyses, and with each replicate comprising ten random-taxon-entry sequences with tbr swapping to up to 10 trees, followed by tbr swapping of the shortest trees obtained from the 10 sub-analyses, with a total of up to 200 shortest trees saved \((\text{hold200 hold/10 mult} \times 10 \text{ max} \times 10)\). The frequency-within-replicates bootstrap was conducted with PAUP; 1,000 bootstrap replicates were conducted using 10 random-taxon-entry sequences per replicate, each followed by tbr swapping to completion with zero-length branches collapsed and all minimal length trees saved.

Data decisiveness \((\text{DD} ; \text{Goloboff, 1991})\) was calculated for each of the three data sets and for all four combinations of two or more of them, using \(M\) and \(S\) as obtained from NONA, and the average length of 100,000 randomly resolved trees, an approximation of \(S\), as obtained from PAUP. Another aspect of the general quality of decisiveness was examined by determining the number of synapomorphies in the \(\text{atpA}\) and \(\text{rbcL}\) data sets that support the monophyly of selected pairs of terminals, and comparing these to the number of synapomorphies that support alternative relationships. Among the various groupings resolved by simultaneous analysis of all the data in the present analysis, there are five pairs of monocot terminals that also were resolved by separate analysis of \(\text{rbcL}\) and \(\text{atpA}\). Three of these five pairs (\(\text{Gymnostachys} + \text{Symlocarpus}, \text{Trillium} + \text{Veratrum},\) and \(\text{Curculigo} + \text{Xanthorrhoea}\)) were selected for analysis of relative degrees of support for them and alternative groupings. In each case the number of character-state transformations that are unambiguously optimized as synapomorphies of the pair of taxa was determined separately for \(\text{atpA}\) and \(\text{rbcL}\) with the NONA command \(\text{apo-}\). For each of the three pairs of taxa, support for alternative relationships was evaluated separately for the \(\text{atpA}\) and \(\text{rbcL}\) data sets by conducting a series of analyses in which one of the two taxa in the pair (\(\text{Gymnostachys}, \text{Trillium},\) and \(\text{Curculigo}\), respectively) was constrained to be resolved as sister of each of the 18 other terminals in the analysis. Thus, in 18 separate analyses using the \(\text{atpA}\) data, and in 18 separate analyses using the \(\text{rbcL}\) data, \(\text{Gymnostachys}\) was constrained to be sister of each of the 18 other taxa in the analysis except \(\text{Symlocarpus}\), and corresponding analyses were conducted using \(\text{Trillium}\) and \(\text{Curculigo}\), in each case with 18 taxa other than the one with which it was resolved by unconstrained analysis. The constrained analyses were conducted with NONA, using the same settings as in the basic cladistic analyses, and following each analysis the number of unambiguous synapomorphies of the constrained group common to all most parsimonious trees obtained by the con-
strained analysis was determined with the command *apo*.

The Mickevich–Farris incongruence index (\(I_{MF}\); Mickevich and Farris, 1981; Kluge, 1989) and lengths of various data sets as optimized on trees obtained from other data sets (related to the Miyamoto incongruence index; Kluge, 1989; Swoford, 1991) were determined with tree lengths obtained from NONA, and by optimizing data on trees with CLADOS. For cases in which analysis of a matrix yielded more than one most parsimonious tree, each of the other matrices was mapped on all most parsimonious trees, and incongruence was calculated using the tree or trees that required the fewest steps (see Swoford, 1991). Incongruence among data sets was tested using the random partition test of Farris et al. (1994), as implemented in DADA, using Hennig86 (Farris, 1988) as a daughter process for tree searches. For each test, 1,000 random partitions were generated, and tree searches were conducted using the Hennig86 command *mhennig* (trees constructed in multiple passes, each pass involving the construction of several trees with a random-taxon-entry sequence, followed by branch swapping and retention of up to one tree obtained from each initial tree) followed by *bb* (branch swapping on all trees saved from previous procedure, with up to 100 shortest trees saved). For each random partition, five replicate searches of this sort were conducted for each partition, each initiated with a different random taxon entry sequence.

**Results**

*Attributes of the Data Sets*

Removal of uninformative characters from the combined matrix of all three character sets leaves 394 informative characters, comprising 42 restriction sites (23.3% of the 180 mapped sites in the IR region, and 47.2% of the 89 sites that are informative among the 53 accessions analyzed by Davis, 1995); 242 *rbcL* nucleotides (17.3% of the 1,398 nucleotides sampled); and 110 *atpA* characters, including the 3-bp deletion and 109 nucleotides (8.6% of the 1,266 nucleotides sampled). The percent contribution of each of the three character sets to the total character set in the analysis is 10.7% restriction sites, 61.4% *rbcL*, and 27.9% *atpA*. Within the matrix of 20 terminals scored for 394 informative characters, the total number of cells scored missing was 254 (3.2% of 7,880 cells in the matrix), apportioned as 60 restriction-site characters (7.1% of 840 cells), 162 *rbcL* characters (3.3% of 4,840 cells), and 32 *atpA* characters (1.5% of 2,200). Among the 242 informative *rbcL* nucleotide positions there were 72 uninformative states (an average of 0.30 uninformative states per informative character), and among the 110 *atpA* nucleotide positions there were 22 uninformative states (an average of 0.20 uninformative states per informative character); informative restriction-site characters, being binary, cannot have uninformative states. Rescoring these autapomorphies of informative characters as missing added 94 such cells to the data set.

Kimura two-parameter distances are uniformly greater for *rbcL* than for *atpA*, though the ratio of distances between the same two taxa varies considerably (e.g., distance between the two magnoliid dicots, *Piper* and *Saumura*, 0.0635 vs. 0.0016, for *rbcL* and *atpA*, respectively, a ratio of 39.7:1; distance between the two representatives of Poales, *Flagellaria* and *Oryza*, 0.0978 vs. 0.0160, a ratio of 6.1:1; and distance between *Piper* and *Flagellaria*, 0.0880 vs. 0.0505, a ratio of 1.7:1).

Of the three separate character sets, the restriction-site data set, with the fewest informative characters (42), has the highest CI (0.49), RI (0.59), and DD (0.53) when analyzed individually, and the *rbcL* data set, with the greatest number of informative characters (242), has the lowest CI (0.37), RI (0.35), and DD (0.26; Table 3). Among the three pairwise combinations of data sets, the matrix of
TABLE 3. Attributes of three character sets (atp A, rbc L, restriction sites [RS]) and combinations of them, representing variation among 18 monocot and two dicot taxa. All numbers were calculated after removal of autapomorphies, including those of cladistically informative characters. MPT = most parsimonious trees; CI = consistency index; RI = retention index; DD = data decisiveness.

<table>
<thead>
<tr>
<th>Character set</th>
<th>Number of informative characters</th>
<th>Number of MPT</th>
<th>Length of MPT</th>
<th>CI</th>
<th>RI</th>
<th>DD</th>
</tr>
</thead>
<tbody>
<tr>
<td>RS</td>
<td>42</td>
<td>23 (5)</td>
<td>85</td>
<td>0.49</td>
<td>0.59</td>
<td>0.53</td>
</tr>
<tr>
<td>atp A</td>
<td>110</td>
<td>1 (17)</td>
<td>260</td>
<td>0.47</td>
<td>0.53</td>
<td>0.48</td>
</tr>
<tr>
<td>rbc L</td>
<td>242</td>
<td>1 (17)</td>
<td>748</td>
<td>0.37</td>
<td>0.35</td>
<td>0.26</td>
</tr>
<tr>
<td>RS + atp A</td>
<td>152</td>
<td>1 (17)</td>
<td>354</td>
<td>0.46</td>
<td>0.53</td>
<td>0.47</td>
</tr>
<tr>
<td>RS + rbc L</td>
<td>284</td>
<td>4 (12)</td>
<td>842</td>
<td>0.38</td>
<td>0.37</td>
<td>0.28</td>
</tr>
<tr>
<td>atp A + rbc L</td>
<td>352</td>
<td>1 (17)</td>
<td>1,013</td>
<td>0.40</td>
<td>0.40</td>
<td>0.32</td>
</tr>
<tr>
<td>RS + atp A + rbc L</td>
<td>394</td>
<td>1 (17)</td>
<td>1,107</td>
<td>0.40</td>
<td>0.41</td>
<td>0.33</td>
</tr>
</tbody>
</table>

* Number in parentheses is number of nodes resolved in the consensus tree.

restriction-site and atp A characters has the highest CI, RI, and DD, and the matrix of restriction-site and rbc L characters has the lowest.

Of the seven cladistic analyses conducted on the three separate data sets and various combinations thereof, five resolved solitary trees and two resolved multiple trees (Table 3). Analysis of the restriction-site data resolved 23 most parsimonious trees (with five nodes resolved in the strict consensus tree), and simultaneous analysis of the restriction-site and rbc L data resolved four trees (12 nodes in consensus tree). Although the other five analyses resolved solitary trees, different trees were resolved by the different data sets (see below). Each of the solitary trees resolved by those five analyses has 17 nodes.

Patterns of nucleotide sequence variation were compared by optimizing the atp A and rbc L sequence data (with autapomorphies of informative and uninformative characters included, and with the deletion character excluded from the

TABLE 4. Attributes of two nucleotide sequence character sets (rbc L and atp A; atp A excluding the deletion character) representing variation among 18 monocot and 2 dicot accessions, as optimized on the most parsimonious tree obtained by simultaneous analysis of rbc L, atp A, and restriction-site data. Summary data are provided for each gene and for each codon position within each gene. All numbers are calculated with autapomorphies of informative and uninformative characters included, except that consistency index is calculated with autapomorphies included (CI_{+A}) and excluded (CI_{-A}); RI = retention index.

<table>
<thead>
<tr>
<th>Codon position</th>
<th>No. sites</th>
<th>No. variable sites (% observed)</th>
<th>No. Cladistically informative sites observed</th>
<th>No. steps (mean/variable character)</th>
<th>CI_{+A}</th>
<th>CI_{-A}</th>
<th>RI</th>
</tr>
</thead>
<tbody>
<tr>
<td>atp A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First position</td>
<td>422</td>
<td>51 (12.1)</td>
<td>16 (3.8)</td>
<td>76 (1.49)</td>
<td>0.57</td>
<td>0.46</td>
<td>0.51</td>
</tr>
<tr>
<td>Second position</td>
<td>422</td>
<td>31 (7.3)</td>
<td>11 (2.6)</td>
<td>59 (1.90)</td>
<td>0.57</td>
<td>0.30</td>
<td>0.45</td>
</tr>
<tr>
<td>Third position</td>
<td>422</td>
<td>184 (43.6)</td>
<td>82 (19.4)</td>
<td>317 (1.72)</td>
<td>0.70</td>
<td>0.49</td>
<td>0.53</td>
</tr>
<tr>
<td>Total</td>
<td>1266</td>
<td>266 (21.0)</td>
<td>109 (8.6)</td>
<td>452 (1.70)</td>
<td>0.69</td>
<td>0.46</td>
<td>0.52</td>
</tr>
<tr>
<td>rbc L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First position</td>
<td>466</td>
<td>88 (18.9)</td>
<td>33 (7.1)</td>
<td>179 (2.03)</td>
<td>0.59</td>
<td>0.36</td>
<td>0.25</td>
</tr>
<tr>
<td>Second position</td>
<td>466</td>
<td>49 (10.5)</td>
<td>21 (4.5)</td>
<td>89 (1.82)</td>
<td>0.61</td>
<td>0.39</td>
<td>0.30</td>
</tr>
<tr>
<td>Third position</td>
<td>466</td>
<td>308 (66.1)</td>
<td>188 (40.3)</td>
<td>778 (2.53)</td>
<td>0.53</td>
<td>0.37</td>
<td>0.36</td>
</tr>
<tr>
<td>Total</td>
<td>1398</td>
<td>445 (31.8)</td>
<td>242 (17.3)</td>
<td>1046 (2.35)</td>
<td>0.55</td>
<td>0.37</td>
<td>0.34</td>
</tr>
</tbody>
</table>
Nucleotide sequence variation in \textit{rbcL} is 1.5 times as great as that of \textit{atpA} as measured by the percentage of observed sites that vary (31.8% vs. 21.0%), and 2.0 times as great as measured by the percentage of sites that are cladistically informative (17.3% vs. 8.6%). There are, on average, 1.4 times as many steps per variable \textit{rbcL} character as there are steps per variable \textit{atpA} character (2.35 vs. 1.70). The \textit{rbcL} data have a lower DD than the \textit{atpA} data (Table 3), and a lower CI and RI than \textit{atpA}, when each is optimized on its own most-parsimonious tree (Table 3), and when each is optimized on the most-parsimonious tree obtained by simultaneous analysis of all data (Table 4).

Of the three codon positions, third-position nucleotides are more variable than those of first or second positions, in both \textit{atpA} and \textit{rbcL}, as measured by the percentage of sites that are variable and percentage of sites that are cladistically informative (Table 4). In a third measure of variation, number of steps per variable character, \textit{rbcL} exhibits a greater amount of overall variation than \textit{atpA} (2.35 vs. 1.70), and a greater amount of variation at first- and third-codon positions, but at second-codon position nucleotides \textit{atpA} is slightly more variable than \textit{rbcL} by this measure (1.90 vs. 1.82 steps per character; Table 4). This pattern is mirrored in CI as apportioned by codon position (\textit{atpA} with a higher overall CI than that of \textit{rbcL}, and also at first- and third-codon position nucleotides; \textit{rbcL} with a higher CI at second-codon position nucleotides), but \textit{atpA} has a higher overall RI than \textit{rbcL} as well as a higher RI at all three codon positions.

The general association of higher levels of variation with higher levels of homoplasy (i.e., lower CI and RI), as observed in comparisons between \textit{atpA} and \textit{rbcL}, is not evident in comparisons among nucleotide positions within the two genes. For example, third-position nucleotides exhibit greater amounts of variation than first-position or second-position nucleotides in all (\textit{rbcL}) or most (\textit{atpA}) measures of variation examined, but third-position nucleotides also have higher retention indices than first- or second-position nucleotides in both \textit{rbcL} and \textit{atpA}, and third-position nucleotides have a higher ensemble CI than first- or second-position nucleotides in \textit{atpA} (Table 4).

In the unconstrained cladistic analysis of just the \textit{atpA} data there are five unambiguous \textit{atpA} synapomorphies for the clade that consists of \textit{Gymnostachys} and \textit{Symplocarpus}. Among most-parsimonious trees obtained with \textit{Gymnostachys} constrained to be resolved with one of the 18 other taxa there are two synapomorphies of \textit{Gymnostachys} with \textit{Acorus}, one each for \textit{Gymnostachys} with \textit{Veratrum} and \textit{Gymnostachys} with \textit{Oryza}, and none for \textit{Gymnostachys} with any of the other 15 taxa, for a total of four unambiguous synapomorphies, and a mean of 0.22, among the 18 analyses conducted with the position of \textit{Gymnostachys} constrained. Thus, there are 22.5 times as many synapomorphies of \textit{Gymnostachys} and its sister taxon in the unconstrained analysis than the mean among the 18 constrained analyses (Table 5). The corresponding numbers obtained by analysis of the \textit{rbcL} data are 12 synapomorphies of \textit{Gymnostachys} and \textit{Symplocarpus} (unconstrained analysis), and a range of 0 to 4 synapomorphies, with a mean of 1.44, in analyses with \textit{Gymnostachys} constrained as sister of each of the other 18 terminals, or 8.31 times as many synapomorphies in the unconstrained analysis as the mean among the constrained analyses. For all three taxon pairs tested for alternative support in this manner (Table 5), the number of \textit{rbcL} synapomorphies is greater than the number for \textit{atpA} in the unconstrained analyses, the mean number of \textit{rbcL} synapomorphies is greater than the mean number of \textit{atpA} synapomorphies among the constrained analyses, and the ratio of these numbers is greater for \textit{atpA} than for \textit{rbcL} (Table 5).
TABLE 5. Relative levels of support for monophyly of three pairs of monocot terminals resolved independently by analysis of \textit{atpA} and \textit{rbcL} data sets and for alternative relationships. For each taxon pair, the first-named terminal was constrained to be resolved as sister of each of the remaining 18 terminals in the sample.

<table>
<thead>
<tr>
<th>Taxon pair</th>
<th>No. of synapomorphies, unconstrained</th>
<th>Range (mean) synapomorphies, constrained\textsuperscript{a} analyses</th>
<th>Synapomorphy ratio, unconstrained/mean constrained</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{Gymnostachys, Sympliocarpus}</td>
<td>5 \ 12</td>
<td>0–2 (0.22) \ 0–4 (1.44)</td>
<td>22.50 \ 8.31</td>
</tr>
<tr>
<td>\textit{Curculigo, Xanthorrhoea}</td>
<td>5 \ 8</td>
<td>0–8 (0.83) \ 0–8 (2.28)</td>
<td>6.00 \ 3.51</td>
</tr>
<tr>
<td>\textit{Trillium, Veratrum}</td>
<td>5 \ 7</td>
<td>0–5 (0.44) \ 0–5 (2.28)</td>
<td>11.25 \ 3.07</td>
</tr>
</tbody>
</table>

\textsuperscript{a} There were 18 constrained analyses.

\textbf{Incongruence}

The greatest amount of incongruence among the three pairwise combinations of the three data sets, as measured by the Mickevich–Farris incongruence index, is between the restriction-site and \textit{atpA} character sets, ($I_{\text{MF}} = 0.0497$; Table 6). These two data sets also exhibit the greatest amount of incongruence as measured by percent increase in steps when one data set is optimized on shortest trees obtained from another (21.2\% increase in steps in the restriction-site data when mapped on the \textit{atpA} tree, and 14.6\% increase in steps in the \textit{atpA} data when mapped on the restriction-site tree with which it is most consistent). The combined restriction-site and \textit{atpA} data also exhibit the highest level of data decisiveness (DD = 0.465) among the three pairwise combinations of data sets (Table 3). At the other extreme, the pairwise combination of \textit{atpA} and \textit{rbcL}, besides having the lowest $I_{\text{MF}}$ (0.0083), also exhibits the smallest proportional increase in steps when each is mapped on shortest trees obtained from the other (4.3\% increase for \textit{rbcL} on the \textit{atpA} tree, 2.3\% increase for \textit{atpA} on the \textit{rbcL} tree), and the lowest decisiveness among pairwise combinations (DD = 0.315).

None of the three pairwise combinations of data sets exhibits significant differences, relative to random partitions, at the level of $P < 0.05$, though incongruence between restriction sites and \textit{atpA} approaches this level ($P = 0.088$). The 1,107 steps in the tree obtained by simultaneous analysis of all three data sets are apportioned as 94 steps for the restriction-site data (9 extra steps relative to shortest trees obtained from just the restriction-site data, an increase of 10.6\%), 749 steps for \textit{rbcL} (1 extra step relative to shortest trees obtained from just the \textit{rbcL} data, an increase of 0.1\%), and 264 steps for \textit{atpA} (4 extra steps relative to shortest trees obtained from just the \textit{atpA} data, an increase of 1.5\%).

\textbf{Monocot Relationships}

Simultaneous analysis of all three data sets resolves one most-parsimonious tree (Fig. 2a). This analysis will be referred to as the combined analysis, and the results of other analyses, based on separate character sets and pairwise combinations, will be described in comparison with it. In the combined analysis the 18 monocots are resolved as a monophyletic group, and the following relationships are resolved among them: \textit{Acorus} (Aranae) is resolved as the sister of all other monocots; among the remaining monocots, the two other elements of Aranae are resolved together as sister of the remaining monocots (Alismatanae, which often are placed with Aranae other...
than *Acorus* in this position, were not sampled in the present study); a paraphyletic alliance of Lilianae plus Velloziaceae (of Bromelinae) constitute the next set of diverging lineages; and a monophyletic alliance of Arecales, Bromelinae (other than Velloziaceae), Commelinanae, and Zingiberanae (the ABCZ alliance of Davis, 1995) is nested within the lilioid alliance. Within the ABCZ alliance, *Anigozanthos* (of Bromelinae) and *Canna* (Zingiberanae) are resolved as a monophyletic group that is sister of all other elements of the group, and *Nypa* (Arecales) is sister of the remaining lineages in the group. *Catopsis* (Bromeliaceae, a third element of the Bromelinae) is resolved as sister of a monophyletic grouping of all sampled representatives of Commelinanae.

Four monocot superorders (Aranae, Lilianae, Bromelinae, Commelinanae) are represented by more than one accession in the present taxon sample. Of these four, only the latter is resolved as monophyletic. Two members of Aranae (and Arales) are resolved as a clade, but the placement of *Acorus* as sister of all other monocots precludes a monophyletic Aranae as traditionally circumscribed (e.g., Dahlgren et al., 1985). Lilianae are nonmonophyletic, in part because *Vellozia* is nested among them, but also because a group consisting of elements of four other superorders (the ABCZ alliance) is nested within the Lilianae. Within Lilianae, the two sampled elements of Asparagales are resolved as a monophyletic group, but Dioscoreales, the only other order that is represented by two or more genera, is not resolved as monophyletic; of the three sampled elements of Dioscoreales, *Tacca* is placed as sister of a large clade that includes all other Lilianae plus the ABCZ alliance, *Smilax* is resolved as the sister of *Alstroemeria* (Liliales), and *Trillium* is resolved as the sister of *Veratrum* (Melanthiales). Members of Bromelinae are placed in three disparate positions, or in two disparate positions if *Vellozia* is excluded from Bromelinae; *Anigozanthos* (Haemodorales) is resolved as sister of *Canna* (Zingiberanae), and *Catopsis* (Bromeliales) is resolved as sister of Commelinanae. Although Commelinanae are resolved as monophyletic, the placement of *Flagellaria* (Poales) as sister of a clade that consists of *Eriocaulon* (Commelinanae) and *Oryza* (Poales) is
inconsistent with the resolution of a monophyletic Poales.

Relative levels of support for the 17 clades resolved by the combined analysis range from 30% to 100% as measured by the strict-consensus bootstrap, and from 40% to 100% as measured by the frequency-within-replicates bootstrap (Fig. 2a). As expected, strict-consensus bootstrap frequencies are less than or equal to frequency-within-replicates bootstrap frequencies for all clades, with the difference between the two measures ranging from 0% to 10%. The greatest disparities between the two bootstrap numbers are in clades with the lowest bootstrap scores. For example, each of the five clades with a strict-consensus bootstrap frequency lower than 65% has a frequency-within-replicates bootstrap frequency 7–10% lower, with an average difference 8.2%, whereas all 12 clades with strict-consensus bootstrap frequencies ≥ 65% have frequency-within-replicates bootstrap frequencies 0–5% lower, with an average difference of 2.6%.

Among relationships resolved by each of the three data sets alone, the tree resolved by rbcL (Fig. 3a) is identical to the one that is resolved by the combined analysis, except that the positions of Nypa and the clade that consists of Anigozanthos + Canna are reversed, so that the former (the only representative of Areanae), rather than the latter, is resolved as sister of all other members of the ABCZ group. Thus, if Vellozia is provisionally excluded from Bromelinae, there is an ABCZ clade, and a BCZ clade within it, with Commelininae resolved as monophyletic, Bromelinae non-monophyletic, and Zingiberanae of indeterminate status, having been sampled only once.

The tree resolved by atpA (Fig. 3b) resembles the one resolved by the combined analysis in its placement of Acorus, Gymnostachys, and Symplocarpus, and therefore in resolving all other monocots as a monophyletic group. It differs, however, in the relationships that are resolved within the monocots other than Aranae. Two major clades are resolved; the first consists of Oryza, Eriocaulon, and Flagellaria (all elements sampled from the Commelininae, itself an element of the ABCZ group) plus the lilioid taxon Vellozia, while the second clade consists of all sampled Lilianae plus the remaining elements of the ABCZ group, those elements being Nypa (Areanae), all Bromelinae (other than Vellozia), and Canna (Zingiberanae). As in the combined analysis, the only order of monocots represented by more than one genus and resolved as monophyletic is Asparagales.

The consensus of 23 trees resolved by the restriction-site data (not illustrated) resolves little, because there is a large polytomy. However, it does resolve some groups that are inconsistent with those resolved by the combined analysis. For example, a clade that consists of Acorus + Gymnostachys is placed as the sister of all other monocots.

Among analyses of pairwise combinations of character sets, that of rbcL + atpA yields the same tree that is resolved by the combined analysis of all three character sets (Fig. 2a). The consensus of four trees resolved by rbcL + restriction sites (not illustrated) is less resolved than the one obtained from the combined analysis, both among the lilioid alliance and within the ABCZ alliance; however, no resolved grouping is inconsistent with any that is resolved by the combined analysis.

The single most-parsimonious tree resolved by the combined matrix of restriction sites and atpA (Fig. 2b) is similar in some respects to the one resolved by the combined analysis (Fig. 2a). For example, it resolves Acorus as sister of all other monocots, and Gymnostachys and Symplocarpus, as a clade that is the next group to diverge from the lineage that includes all remaining monocots. The results of this analysis also resemble those of the combined analysis in resolving a paraphyletic set of Lilianae and Vellozia, within which is nested the ABCZ group (i.e., Areanae, Bromelinae except Vellozia, Commelininae, and
Zingiberanae). However, relationships within both of these groups, as resolved by just the restriction-site and atpA data, differ from those resolved by the combined analysis. Within the ABCZ alliance, for example, Flagellaria and Oryza are resolved as a monophyletic Poales, with Eriocaulon (Commelinales) as their sister,
as opposed to the resolution by the combined analysis of Flagellaria as sister of Eriocaulon + Oryza.

Evolution of atpA

Of the 1,266 nucleotide positions examined, there is variation among the 20 accessions at 266 sites, and cladistically informative variation at 109 sites (Table 4). Variable positions are distributed fairly evenly along the length of the gene (Fig. 4), with the most prominent concentration of variation in the region around position 590, where the four deletions occur. Scoring of the four taxa with deletions (three with 3-bp deletions, one with a 6-bp deletion, for a total of 15 deleted nucleotides) resulted in 44 cells scored as unknown for these four taxa in this region because of uncertainty regarding the alignment of the deletions. This scoring as missing due to uncertainty in alignment of deletions does not contribute to the observed high levels of nucleotide sequence variation in this region, and if it has any effect it is only to decrease the observed amount of variation by eliminating variant nucleotides from consideration. Thus, the higher level of nucleotide diversity observed in the region around the deletions is not an artifact of ambiguity in alignment of the deletions.

On the tree resolved by the combined analysis, the 3-bp deletion in atpA is as homoplastic as it could be (RI = 0), because all three taxa that exhibit this deletion are separated by intervening taxa without the deletion (Fig. 2a). The resolution of Canna (with the 6-bp deletion) as sister of Anigozanthos (which is undeleted) represents a fourth independent derivation of a deletion. Similar results are obtained in all except one of the six other single- and combined-character matrices that were analyzed (three steps in the atpA deletion when Canna is scored undeleted or unknown for the 3-bp deletion, four steps when Canna is scored deleted; RI = 0 in either case). The exception is the analysis of restriction sites alone, in which two taxa with deletions, Catopsis and Eriocaulon, are resolved as sisters in 3 of the 23 most-parsimonious trees. This relationship is consistent with the interpretation of the 3-bp deletion in these two taxa as homologous, but an additional step in this character still is required to account for the deletion in Acorus, and another step is required if Canna is scored as deleted. In the three trees that resolve Catopsis and Eriocaulon as sister taxa, Canna and Anigozanthos also are resolved as sisters, as are Flagellaria and Oryza.

DISCUSSION

Three Character Sets and Monocot Relationships

Restriction-site variation in the chloroplast genome has been used widely for phylogenetic analysis (Palmer et al., 1988; Olmstead and Palmer, 1994). Restriction-site analyses sample variation across whole genomes or through genomic regions that are scores of thousands of base pairs in length (e.g., the IR region), and they have been instrumental in the discovery and characterization of numerous structural mutations (Palmer et al., 1988; Olmstead and Palmer, 1994). Most analyses employing restriction-site data from the chloroplast genome have been employed in the study of relationships within families or genera. A few, however, have utilized the relatively slowly evolving IR region for the analysis.
of relationships within major multifamily clades (e.g., Downie and Palmer, 1992; Manos et al., 1993), including relationships across a wide range of monocots, using representatives of the dicot subclass Magnoliidae as outgroups (Davis, 1995). The latter analysis was based upon restriction sites plus two cladistically informative length variants, one of them the ORF2280 deletion (Downie et al., 1994), which was interpreted as homoplastic (see Hahn et al., 1995, for an alternative interpretation).

The most widely used gene for phylogenetic analysis in plants is rbcL (e.g., Chase et al., 1993). Sequences from thousands of plant species now are available, and the number continues to grow. The utility of this gene for phylogenetic analyses has been demonstrated, and a well-sampled rbcL-based analysis of relationships in the monocots has been conducted (Duvall et al., 1993a). Attributes of the mitochondrial gene atpA were discussed in the introduction of this paper; previous evidence that it evolves more slowly than rbcL is corroborated by the present findings of reduced numbers of variable and cladistically informative sites, and lower genetic distances in parallel samplings of taxa.

A modest level of incongruence was observed among the data sets used in the present study, relative to those reported in a variety of other studies. Incongruence, as measured by $I_{\text{MF}}$ (Mickevich and Farris, 1981; Kluge, 1989), ranged from 0.0083 to 0.0497 among pairwise combinations of character sets in the present study (Table 6), and even the highest of these is relatively low in comparison to other published figures (e.g., 0.114 [Kluge, 1989], 0.153 [Rodman et al., 1996]), and 0.052 [Uhl et al., 1995]). Also, the observed levels of incongruence in this study, for any of the pairwise combinations tested, are insufficient for them to be recognized as significantly more incongruent than random partitions. It is unlikely that any of the three character sets (two from the same genome, all three uniparentally inherited in most angiosperms) would have different phylogenetic histories among the disparate set of taxa sampled, but incongruence can arise from sources other than different phylogenetic histories (Wendel and Doyle, 1998). Whether or not the three data sets differ significantly, they do imply different relationships. The present sampling of monocot taxa is insufficient to support strong conclusions regarding relationships, but a few comments along those lines are in order.

Evidence continues to accumulate for the placement of Acorus as the sister of all other monocots (Duvall et al., 1993a, 1993b; Chase et al., 1995; Davis, 1995; Nadot et al., 1995; and the present results), and for a group that includes the Aranae and Alismataeae (the latter not sampled here) as the next lineage to diverge from the line that includes all remaining monocots. An alternative

Table 6. Incongruence among three data subsets (rbcL, atpA, restriction sites [RS]) representing variation among 18 monocot and 2 dicot taxa. Each row of table depicts incongruence between two data sets. For each pairwise combination of data subsets, increase in tree length is indicated for each of the character types when optimized on trees resolved by the other; the column head specifies the character set that is optimized.

<table>
<thead>
<tr>
<th>Data sets</th>
<th>$I_{\text{MF}}$</th>
<th>Incongruence probability</th>
<th>Increase in number of (%) steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>RS, rbcL</td>
<td>0.0176</td>
<td>0.813 n.s.</td>
<td>9 (10.6)</td>
</tr>
<tr>
<td>RS, atpA</td>
<td>0.0497</td>
<td>0.088 n.s.</td>
<td>18 (21.2)</td>
</tr>
<tr>
<td>rbcL, atpA</td>
<td>0.0083</td>
<td>0.873 n.s.</td>
<td>32 (4.3)</td>
</tr>
</tbody>
</table>

$^a$ Mickevich-Farris incongruence index.

$^b$ Random partition test.

$^c$ For one data set when optimized on trees resolved by another data set.
hypothesis for the placement of Acorus is that it represents an independent line of divergence from among the magnoliid dicots (Bharathan and Zimmer, 1995). In the present instance, with only two magnoliid dicots sampled, it would be difficult to detect evidence for nonmonophyly of the monocots, so this must remain a standing problem.

Within the monocots there appears to be a paraphyletic group of elements assignable to Lilianae, among which are nested a monophyletic group consisting of Velloziales (Bromeliaceae), Cyclanthanae, and Pandanae (the latter two not sampled here), and a second monophyletic group consisting of Arecanae, Bromeliaceae other than Velloziales, Commelinanae, and Zingiberanae (the ABCZ clade). Within the ABCZ clade, the palms (Arecanae) have been resolved in previous studies as the sister group of a BCZ clade that includes all other elements of the group, but there is some contrary evidence in the present analysis, favoring the placement of Zingiberanae and one or more elements of a polyphyletic Bromeliaceae (i.e., Haemodorales) in this position.

One important problem in monocot phylogenetics concerns the monophyly of Poales (Dahlgren et al., 1985; Linder, 1987; Doyle et al., 1992; Linder and Rudall, 1993; Chase et al., 1995; Davis, 1995; Kellogg and Linder, 1995; Linder and Kellogg, 1995; Stevenson and Loconte, 1995). Although monophyly of a core group within Poales (Poaceae, Joinvilleaceae, and Restionaceae sensu lato) is well supported, there is conflict on the question of whether the closest relative of this group is Flagellaria (Flagellariaceae, conventionally placed within Poales), or one or more elements usually placed in Cyperales (e.g., Cyperaceae, Juncaceae) or Commelinaceae (e.g., Eriocaulaceae, Xyridaceae). On this point there have been conflicting results, even among treatments based on cladistic analysis of rbcL sequences. For example, Chase et al. (1993), in their Search I (Fig. 6a), resolved a clade that consisted of Poaceae + Elegia (Restionaceae), and another clade that consisted of Flagellaria + Cyperales as its sister group, while in their Search II (Fig. 6b) they resolved Lachnocaulon (Eriocaulaceae) as the sister of Poaceae, with Elegia and Flagellaria as the next most closely related taxa to this group. Both of these sets of relationships are inconsistent with the monophyly of Poales. In contrast, Duvall et al. (1993a, Fig. 4), also using rbcL but a different set of taxa, resolved a group that included several representatives of Poaceae, plus Elegia and Flagellaria (i.e., Poales monophyletic). A different set of relationships was presented by Hahn et al. (1995, Fig. 4), who depicted an rbcL-based tree (attributed to Duvall et al., 1993a) in which a clade consisting of Eriocaulaceae and Xyridaceae is sister of the core Poales, and Flagellaria is sister of this larger group (i.e., Poales nonmonophyletic). In the present study, the combined analysis of all three data sets resolves Oryza as more closely related to Eriocaulon than to Flagellaria (i.e., Poales not monophyletic, Fig. 2a), but the combined analysis of the restriction-site and atpA data sets (i.e., with rbcL removed) resolves a monophyletic Poales that includes Flagellaria (Fig. 2b). This question, as one element of a larger set of questions regarding the ABCZ clade, warrants continued attention.

Evolution of atpA

The present analysis demonstrates that atpA has evolved more slowly than rbcL among the 20 sampled taxa, as evidenced by the presence of two-thirds as many variable nucleotides and half as many cladistically informative nucleotides per site examined, lower genetic distances in parallel pairwise samplings of taxa, and 75% as many steps per variable site on most-parsimonious trees (Table 4). These results are consistent with previous observations of relative evolutionary rates in these two genes (Eyre-Walker and Gaut, 1997).
In the present results, lower levels of variation in \( \text{atpA} \), relative to \( \text{rbcL} \), are consistently associated with lower levels of homoplasy. This association is observed when homoplasy is calculated on the basis of trees derived from individual or combined data sets (Tables 3, 4). When calculated on the basis of trees derived from individual data sets, this pattern is evident when homoplasy is measured by RI or CI (Table 3); when calculated on the basis of trees derived from combined data sets, it is evident when homoplasy is measured by RI or by CI with autapomorphies included or excluded (Table 4).

The relationship between level of variation and amount of homoplasy is more complex when examined in terms of codon positions within the two genes. Both genes exhibit higher levels of variation in third-position sites than in first or second-position sites, as measured by number of variable characters and the number of cladistically informative characters (Table 4), but in both \( \text{atpA} \) and \( \text{rbcL} \) this elevated amount of variation in third-position sites is associated with a lower amount of homoplasy, at least as indicated by RI (Table 4). For example, there are about six times as many variable third-position sites in \( \text{atpA} \) as there are second-position sites (184, or 43.6% of all sites examined, vs. 31, or 7.3% of sites examined; Table 4), and more than seven times as many cladistically informative third-position sites as second-position sites (82, or 19.4% of sites examined, vs. 11, or 2.6% of sites examined), but the RI of third-position sites exceeds that of second-position sites (0.53 vs. 0.45). A similar pattern is observed in \( \text{rbcL} \) (Table 4). This pattern is not evident, however, when homoplasy is compared in terms of CI; third-position sites do not exhibit uniformly higher consistency indices than first- or second-position sites (Table 4).

The co-occurrence of deletions of two different sizes with high levels of nucleotide sequence variation in the region around site 590 in \( \text{atpA} \) (Fig. 4) is consistent with the recognition of a “hot spot” of increased rates of molecular evolution in this region, both in structural attributes and nucleotide substitution rate. A similar association has been observed in the chloroplast-encoded gene \( \text{ndhF} \) (Olmstead and Reeves, 1995). There are now many examples, at a variety of taxonomic levels, of convergent molecular changes in characters that at one time might have been presupposed to be unique. These include the repeated deletions of cpDNA genes and introns (Downie et al., 1991, 1994; Davis, 1995; Doyle et al., 1995), convergent inversions of minute as well as large pieces of cpDNA (Downie and Palmer, 1994; Hoot and Palmer, 1994; Kelchner and Wendel, 1996), the repeated evolution of small insertions and deletions (Clark et al., 1995; Olmstead and Reeves, 1995; Cunningham et al., 1997), and the remarkable discovery of intron homing in rDNA of fungi (Hibbett, 1996), where a particular group I intron is reported to have been precisely inserted at an identical nucleotide position in disparate groups of homobasidiomycetes. These and other examples bear witness to the potential for homoplasy in molecular structural mutations, which we take as confirmation that the apparently homoplastic deletions in \( \text{atpA} \) do indeed represent convergences and hence are nonhomologous.

Data Decisiveness, Incongruence, and Data Quality

Across the present sampling of 18 monocots and two dicots, \( \text{rbcL} \) is more variable, faster evolving, and more homoplastic than \( \text{atpA} \). The \( \text{rbcL} \) data set also is the least decisive of the three character sets. Although the restriction-site data set is less comparable to the two nucleotide sequence data sets, it has the highest CI, RI, and DD of the three data sets (Table 3), and this association of low inconsistency and high decisiveness conforms with the pattern observed in the other two character sets. Before much is made of this relationship, however, a potential
relationship between DD and number of characters should be considered. The \textit{rbcL} data set has the greatest number of informative characters, and the restriction-site data set has the fewest, so one possibility is that the observed relationship between consistency and decisiveness reflects a dependence of both of these qualities on size of data set. This possibility, for homoplasy, was considered and tested by Sanderson and Donoghue (1989, 1996) in analyses of patterns of homoplasy among published data sets, and they observed no correlation between number of characters and amount of homoplasy. Moreover, we note that in the present case the two nucleotide sequence data sets exhibit intrinsic differences in terms of the amount of variation present per base pair examined. Finally, we conducted informal analyses in which CI, RI, and DD were examined for random subsets of the informative \textit{rbcL} characters, with each sample comprising as many characters as in each of the other two data sets. These subsets consistently exhibited higher levels of inconsistency and lower levels of decisiveness than are found in the \textit{atpA} and restriction-site data sets. Thus, we regard the pattern observed among the three data sets as prima facie evidence of a real difference in the intrinsic attributes of the three character sets, rather than as an artifact of sampling.

The relationship between decisiveness and character incongruence deserves attention, for each is related to robustness of support for relationships by individual and combined data sets. This relationship can be examined in terms of the indices themselves (decisiveness as a function of the number of extra steps in nonparsimonious trees, incongruence as a function of extra steps required when data sets are combined), as well as in terms of the general qualities that these indices measure (decisiveness as robustness of support, incongruence as conflict in terms of groups that are supported). From either of these perspectives it is evident that the two qualities are related, for each is an expression of support as measured by extra steps required for alternative cladistic structures.

A significant level of incongruence was not demonstrated for any pairwise combination of the three data sets, relative to the amount observed between random partitions. Of the three pairwise combinations, however, the greatest amount of measured incongruence is between the restriction-site and \textit{atpA} character sets, and this is also the only pairwise combination in which DD of a combined data set is lower than that of either of the two constituent data sets (0.47 vs. 0.53 and 0.48; Table 3). In the other two cases, DD of a combined data set is intermediate between those of the two separate data sets. Although DD of the data set formed by combining the restriction-site and \textit{atpA} data sets is lower than of than either of its constituent parts, these two separate data sets have the highest DD of the three, and this pairwise combination also has the highest CI, RI, and DD of the three pairwise combinations. Moreover, the data set formed by combining the restriction-site and \textit{atpA} data has a lower CI and RI, and a higher DD, than the data set formed by combining all three data sets. Conversely, the two other pairwise combinations, both of which include \textit{rbcL}, consistently exhibit comparatively high levels of inconsistency and low levels of decisiveness (Table 3).

The relationship between homoplasy, decisiveness, and incongruence is complex, but some aspects of that relationship emerge from the present study. One aspect is that indecisive data sets should have a lesser tendency than decisive ones to be incongruent with other data sets. DD is a measure of the robustness of support by a data set for most parsimonious trees, relative to all possible trees, that is, the degree to which most parsimonious trees are shorter than the average length of all possible trees. A completely indecisive data set would have equivalent support for mutually conflicting relationships in such a pattern that all possible trees were
of equal length (Goloboff, 1991). The Mickevich-Farris incongruence index is an expression of the number of extra steps required by shortest trees generated by a combined data set, relative to the sum of the number of steps required by the separate data sets. Thus, if two data sets have any most parsimonious trees in common, $I_{MF} = 0$, and the trees that are most parsimonious for both data sets, when analyzed separately, will be most parsimonious for the combined data, and no additional steps will be required. Because all possible trees are of equal length for a completely indecisive data set, the set of shortest trees for a decisive data set representing the same taxa must be a subset of the trees that are shortest for the indecisive data set, and consequently no extra steps are required when the two are combined. Similarly, the Miyamoto index of incongruence (Kluge, 1989; Swofford, 1991) is zero for a completely indecisive data set and one with nonzero decisiveness, if it is calculated as recommended by Swofford (1991), on the basis of extra steps required to map one data set on those most parsimonious trees for the other with which it most compatible. Thus, whichever of these indices is used, there is no incongruence between two data sets if one of them is completely indecisive, for there is always an intersection between the sets of most parsimonious trees for the two data sets.

With two real data sets, wherein some trees are more parsimonious than others, it is expected that a relatively indecisive data set will, in general, exhibit lower levels of incongruence with other data sets than will one that is more decisive (Allard and Carpenter, 1996). Thus, two data sets can fail to be incongruent with each other either because they support the same set of relationships or because at least one of them provides only weak support for whatever relationships it does favor. The latter situation does not seem to be desirable, yet if incongruence is used as a criterion of (non)combinability (e.g., Lutzoni, 1997; Shaffer et al., 1997), a data set with a relatively weak signal may be favored for combination with others, while another data set that is more decisive is rejected. We would argue that the potential for a data set to be incongruent with others is a desirable property, for it allows the evidence in that data set to be refuted, just as it allows that data set to refute the evidence of others. It is in this light that we note that $rbcL$, as the least decisive of the three data sets that were examined, exhibits the least evidence of incongruence with the other two.

If a data set (even one that is completely indecisive) includes any cladistically informative characters, it must exhibit state transformations (i.e., synapomorphies) on internal branches of its most-parsimonious cladograms. Although these transformations are correctly construed as support for certain phylogenetic relationships, this support, as observed on any particular tree, may be counterbalanced by support for conflicting relationships. Thus, even a relatively indecisive data set has steps on internal branches of all most parsimonious trees, and a large and indecisive data set may have more steps than a smaller one that is more decisive. Branch length, per se, therefore implies little about robustness of support for relationships (Davis, 1993), and the widespread observation that small morphological character sets are not consistently “swamped” when combined with larger molecular character sets (Chippindale and Wiens, 1994; Nixon and Carpenter, 1996a) may be attributable to general differences between these classes of characters in overall levels of decisiveness.

As in the calculation of Bremer support (Bremer, 1988, 1994), the constrained analyses in the present study (Table 5) were conducted to examine attributes of most parsimonious trees, relative to those of nonparsimonious trees with particular characteristics. Bremer support of a clade is the minimum number of extra steps required by a data set (relative to the length of most parsimonious trees) for a
tree in which that clade is not resolved as monophyletic. In contrast, the constrained analyses conducted in this study facilitate the comparison of support for a group of two taxa in terms of the number of unambiguous synapomorphies of that group in a most parsimonious tree, relative to the number of unambiguous synapomorphies of a group that consists of one of these two taxa and one other taxon. The constrained analyses therefore measure support for individual clades in a manner that is akin to the measure of DD for an entire data set (i.e., number of steps in most parsimonious resolutions, relative to number of steps in alternative resolutions). However, the constrained analyses did not involve examinations of all possible alternative placements for taxa, and the present implementation of this procedure should be recognized as heuristic.

When the rbcL data set, with 445 variable characters, is mapped on its most parsimonious tree (Fig. 3a), and the atpA data set, with 266 variable characters, is mapped on its most parsimonious tree (Fig. 3b), the internal branches of the former are markedly longer than those of the latter. For each of the three pairs of taxa examined, however, the amount of support for the resolved groups, relative to the amount of support for alternative groups, was less for rbcL than for atpA. For example, the grouping of Gymnostachys and Symplocarpus was supported by 12 unambiguous transformations in rbcL characters, and 5 unambiguous transformations in atpA characters. However, for atpA this represented 22.50 times the average amount of support present for each of 18 alternative placements that were examined for Gymnostachys, and for rbcL this represented only 8.31 times the average amount of support present for the same 18 alternative placements. These results, with respect to individual clades, corroborate the lower overall decisiveness of the rbcL data set, relative to that of the atpA data set.

It is useful to recall that evidence of relationship among members of a lineage arises during periods of common ancestry, and that this evidence, once present, can only be diluted by later mutations that occur after sublineages have diverged. Rapidly evolving character sets are more likely to evolve during a given time period (i.e., they are more likely to have evolved during a particular time in the past), but they are also more likely to experience subsequent changes that conceal the original information. On the other hand, slowly evolving characters are less likely to accumulate synapomorphic transformations, but are more likely to retain whatever changes do occur. In the present instance, rbcL exhibits a greater amount of total variation than either of the other two data sets, and the overall rate of nucleotide sequence evolution appears to be greater in rbcL than in atpA, but if data decisiveness is an appropriate indicator of information quality, the information from restriction sites and atpA is of higher overall quality. Bull et al. (1993) presented evidence that slowly evolving characters are more likely to provide an accurate reconstruction of phylogenetic relationships than rapidly evolving ones, at least under some conditions. Although they argued against combining data sets representing genes that have evolved at different evolutionary rates, their examples demonstrated an increased efficacy of combined data sets (i.e., one set of slowly evolving characters, and one set of rapidly evolving characters), relative to the results obtained from just the latter (Chippindale and Wiens, 1994).

Molecular evolutionary rate appears to be one determinant of data decisiveness, but it is clear that other factors also are involved. As noted earlier, rbcL exhibits a higher level of homoplasy than does atpA, which may be explained in part by the relatively higher rate of molecular evolution in rbcL. An increased overall rate, however, does not necessarily imply greater homoplasy, because two or more genes with equivalent overall substitution rates may differ significantly in their
patterns of molecular evolution among sites. That is, evolutionary change may be concentrated in a minority of nucleotide positions (coding or noncoding), or be more equitably spread among sites (e.g., Johnson and Soltis, 1995; Steele and Vidalys, 1994). This variation in the degree of among-site rate heterogeneity is expected to lead to differing levels of overall homoplasy as well as to varying patterns of homoplasy among sites (Olmstead et al., 1998). In the present example, \textit{rbcL} not only exhibits a higher overall amount of homoplasy than does \textit{atpA}, but it is more homoplastic than \textit{atpA} at both third-position (mostly synonymous) and first- and second-position (predominantly nonsynonymous) sites (Table 4). In addition, third-position sites are not “noisier” than first- or second-position sites for either \textit{rbcL} or \textit{atpA}, an observation that runs counter to the expectation that functionally constrained first- and second-position sites would be less subject to multiple hits than third-position sites. This unexpected pattern of homoplasy by codon position has been noted in other data sets (Kim et al., 1992; Hoot et al., 1995; Kim and Jansen, 1996; Olmstead et al., 1998), and Kellogg and Juliano (1997) have discussed some of the possible molecular evolutionary constraints that underlie the phenomenon. With respect to the present study, we note that nonsynonymous sites in \textit{rbcL} appear to be especially homoplastic, both quantitatively overall, and in comparison to \textit{atpA}; the RI of third-position sites of \textit{rbcL}, for example, is nearly 50% greater than that of first-position sites (0.36 vs. 0.25), whereas more equivalent values are obtained for \textit{atpA} (0.53 vs. 0.51). The substantially higher level of homoplasy at first-position \textit{rbcL} sites, as compared to third-position sites of the same gene, suggests the existence of constraints at nonsynonymous sites, under which change can occur, but is canalized, and hence subject to parallelisms and reversals. Whatever the molecular evolutionary cause of this phenomenon, it seems likely that the poorer performance of \textit{rbcL} than \textit{atpA}, in terms of data decisiveness, is due to both its higher overall level of homoplasy and the fact that it is performing especially poorly at nonsynonymous sites.

There is a growing consensus in the systematic community in favor of simultaneous analysis of multiple character sets (e.g., Mickevich and Farris, 1981; Miyamoto, 1985; Kluge, 1989; Eernisse and Kluge, 1993; Kluge and Wolf, 1993; Chippindale and Wiens, 1994; Allard and Carpenter, 1996; Nixon and Carpenter, 1996a). Different character sets may have different qualities, and may favor different relationships, but if true relationships cannot be known, the use of multiple data sets at least allows for the examination of alternative hypotheses and patterns of congruence and incongruence, and thereby helps to focus further investigations. The accumulation of an enormous number of \textit{rbcL} sequences has facilitated important advances in plant systematics, and these sequences stand as a baseline against which additional character sets can be compared. As evaluations of this sort are conducted, it is important that objective measures of data quality be developed. Data decisiveness does not capture all relevant attributes of data quality, but it is an informative index of overall robustness of support for relationships. Among groups as diverse as the monocots (and by implication, among major lineages of dicots and higher level groupings), \textit{rbcL} may be too indecisive to resolve the details of ancient divergence events, at least when used alone. Data sets that are more decisive, however, may lack sufficient variation to resolve relationships. The twin dilemmas of molecular evolutionary rate heterogeneity among genes and among sites within genes, as discussed here, cannot be resolved easily, but at the higher phylogenetic levels at which indecisiveness limits the utility of \textit{rbcL}, simultaneous analysis of it and other data sets that are less homoplastic and more decisive, such as \textit{atpA} and restriction sites of
the chloroplast IR region, may help to resolve relationships.

ACKNOWLEDGMENTS

We thank B. Gaut for providing atpA primers, and for granting permission to publish the sequences of primers he developed; V. Albert for providing unpublished rbcL sequences; P. Goloboff and K. Nixon for valuable discussions of the matters presented here; and M. Chase, R. Olmstead, R. Small, and one anonymous referee, for valuable comments on earlier drafts of the manuscript. We acknowledge support by the National Science Foundation.

REFERENCES


ISAAC, P. G., A. BRENNICKE, S. M. DUNBAR, AND C. J. LEAYER. 1985. The mitochondrial genome of fertile maize (Zea mays L.) contains two copies of the gene encoding the α-subunit of the

Downloaded from http://sysbio.oxfordjournals.org/ by guest on November 4, 2016


Received 24 June 1997; accepted 23 September 1997

Associate Editor: D. Cannatella