Syst. Biol. 49(1):101–113, 2000

Is There a Justification for Differential a Priori Weighting in Coding Sequences? A Case Study from *rbc*L and Apocynaceae s.l.

BENGT SENNBLAD[1] AND BIRGITTA BREMER

Department of Systematic Botany, Evolutionary Biology Centre (EBC), Uppsala University, Norbyv. 18D, S-752 36 Uppsala, Sweden; E-mail: Bengt.Sennblad@systbot.uu.se

Abstract.—Functional constraints are often assumed to influence the performance of nucleotide characters in phylogenetic analysis: First and second codon positions and sites of structural importance are considered to show less homoplasy. We investigate the performance of *rbc*L characters with differential functional constraints in a cladistic analysis of the plant family Apocynaceae s.l. (Sennblad and Bremer, in prep.). Performance is measured as rescaled consistency indices (*rc*). We show there is no significant difference in performance between parsimony-informative sites constrained by function in the enzyme, and sites that are not. Furthermore, the substitutions in third-codon position performed significantly better than those in first and second. The variation of *rc* within the different classes was high, however. Consequently, there is no support for routinely applied a priori differential weighting, neither of codon positions, nor of different functional classes from the present analysis of *rbc*L data in the Apocynaceae s.l. [Apocynaceae s.l.; character performance; character weighting; codon constraints; functional constraints; protein structure; rescaled consistency index.]

When an amino acid participates in structural interactions that are vital for the function of the enzyme, this may restrict the probabilities of substitution for the corresponding nucleotides. Mutations decreasing the function of the enzyme will be selected against and will thus occur with a lower frequency. The nucleotide site is subject to functional constraints.

The gene currently most often used for phylogenetic studies in botany is the chloroplast gene *rbc*L, which has been used at various systematic levels from within genera (Xiang et al., 1993; Daugbjerg et al., 1994) to large-scale relationships of angiosperms (Chase et al., 1993) or even of green plants (Manhart, 1994; Källersjö et al., 1998). The most common application has been at familial or ordinal taxonomic levels (Donoghue et al., 1992; Kim et al., 1992; Bremer et al., 1995; Fredericq and Ramírez, 1996). Encoded in the large single-copy region of the chloroplast genome (Shinozaki et al., 1986), the gene *rbc*L codes for the large subunit of the photosynthetic enzyme ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco). This enzyme is vital for photosynthetic plants, having a dual function in their metabolism: It catalyzes the initial steps both in photosynthetic carbon fixation (carboxylation of ribulose 1,5-bisphosphate, RuBP) and in photorespiration (oxygenation of RuBP). Rubisco is thus likely to be exposed to extensive functional constraints. Albert et al. (1994) estimated that 84% of the nucleotides are subject to functional constraints, and Kellogg and Juliano (1997) found that 22% of the amino acid positions are strictly conserved.

Rubisco is one of the best investigated plant enzymes; the structure of the subunits, the association between the subunits, the construction of the active site, and the connection of these to the amino acid sequence have been studied in detail (e.g., Andersson et al., 1989; Knight et al., 1990; Schneider et al., 1990). Amino acids participating in the secondary structure of the large subunits form α-helices and β-sheets that constitute building blocks of the subunits. These "blocks" are then folded into the correct tertiary structure. Important here are the amino acids that coalesce in hydrophobic cores. Large subunits then form dimers, four of which, together with eight small subunits, form the functional enzyme. Amino acids participating in the in-

[1]Corresponding author.

terfaces of the different subunits thus play a vital role in the quaternary structure of the protein. The eight active sites of the enzyme are located at the intradimer interface of the large subunits. Amino acids binding to the substrate and to a coenzymatic Mg^{2+} ion are important in the active site (for further details, see Andersson et al., 1989; Knight et al., 1990). How do these functional constraints affect information content of the corresponding nucleotide? One assumption often advanced is that, because substitutions in nucleotides with functional constraints are likely to be subjected to strong negative selection, they may be more conservative (i.e., less likely to become fixed) than those without such constraints (Hillis et al., 1993; Miyamoto et al., 1994; Simon et al., 1994).

Similar reasoning can be applied to nucleotide substitutions in different codon positions. Substitutions in the third position are often silent, i.e., they do not give rise to an amino acid substitution. Substitutions in the first and second position, on the other hand, almost always result in amino acid changes. The first and second positions can therefore be assumed to be the more conservative and the third position the least conservative of the codon positions (Kellogg and Juliano, 1997). Conservative characters are generally assumed to be less likely to be homoplastic or become saturated with change, and thus are thought to perform better in phylogenetic analyses (e.g., Hillis et al., 1993; Miyamoto et al., 1994; Simon et al., 1994; Swofford et al., 1996). There have been a few investigations of the potential impact of codon constraints on *rbc*L substitution rates (see e.g., Albert et al. [1993], and Olmstead et al. [1998] for an analytical approach and an empirical approach, respectively). However, the actual performance of different structural sites and codon positions in *rbc*L as characters in phylogenetic analyses have scarcely been investigated.

In the Apocynaceae s.l., phylogenies based on *rbc*L data are in conflict with traditional classification. Furthermore, the variations in branch-support distribution (as well as in the branch-length distribution) of the *rbc*L tree (Sennblad and Bremer, in prep.; see Fig. 1) may reflect a variation in the substitution rate over the tree. This may be indicative of different constraints at work throughout the gene, which is of interest in the context of the above discussion of constraints on nucleotide data and motivates an evaluation of the effect of constraints on performance of *rbc*L data, based on the analysis of the Apocynaceae s.l. by Sennblad and Bremer (in prep).

The parsimony analysis of Sennblad and Bremer (in prep.) is based on characters from the coding region and the region directly downstream of the *rbc*L gene. In addition, a combined analysis of a smaller set of taxa by using *rbc*L and *ndh*F data indicates further support for some weakly supported relationships in the subfamily Apocynoideae s.l. (indicated by an arrow in Fig. 1). Furthermore, the general topology and the supported groups (bootstrap frequencies $\geq$63%) of the tree in Figure 1 are to a large degree congruent with other recent studies of the Apocynaceae s.l., including the studies by Civeyrel (1998) based on *mat*K sequence data; by Endress et al. (1996), using data from *rbc*L and *mat*K sequences and floral and pollen morphology; by Sennblad et al. (1998) on the tribe Wrightieae, using *rbc*L and morphological data; and preliminary results from an analysis of the Apocynoideae s.l. based on *rbc*L, *mat*K, and *atp*B data by Sennblad, Civeyrel, Endress, and Chase (pers. comm.).

As an objective way of estimating the differential performance of characters in parsimony analyses, we have chosen the rescaled consistency index (*rc*), the measure used in the successive weighting approach suggested by Farris (1969, 1989). In this procedure, characters are evaluated a posteriori to an initial analysis. Characters are down-weighted according to their *rc,* after which the data set is reanalyzed; this is repeated until weights are stable. The *rc* is the product of the consistency index (*ci*) and the retention index (*ri*) (Farris, 1989) and is thus related to the homoplasy and fit of the character on a cladogram. We will in this article consider *rc* as a measure of the performance of the different nucleotide characters in parsimony analysis.

We here use *rc* to evaluate the performance of *rbc*L nucleotide positions in the analysis of the plant family Apocynaceae

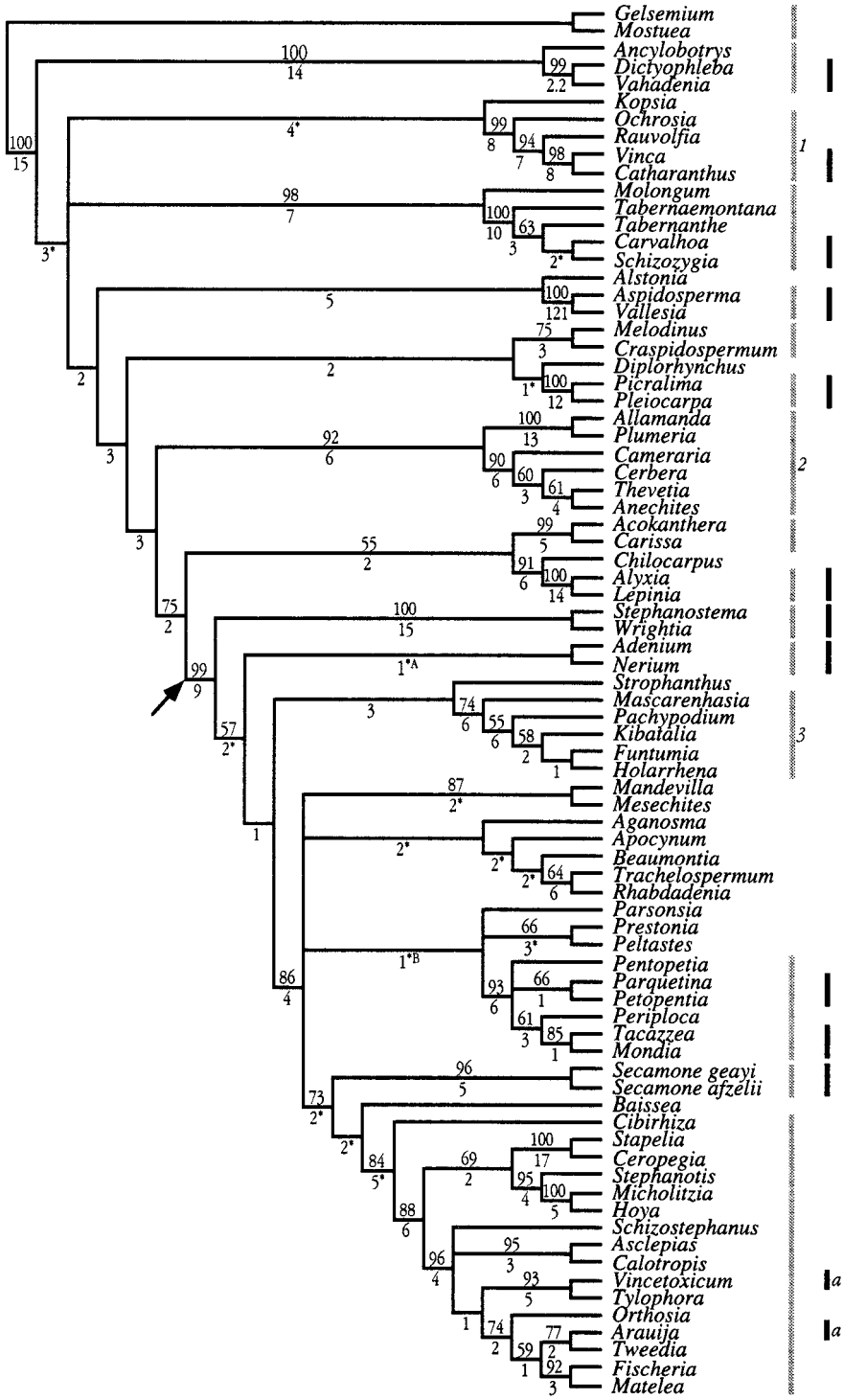


FIGURE 1. Combinable component consensus tree of the 252 most-parsimonious trees from the successive weighting analysis of the complete *rbc*L data set. Numbers below branches are ACCTRAN branch lengths; bootstrap values are indicated above branches (weights applied). Branches not present in the strict consensus tree from the unit weight analysis are indicated with an asterisk (*). For branches A and B, the frequencies of trees where the clade is resolved are 95% and 50%, respectively; all other branches have 100% frequency. The clade indicated with an arrow corresponds to the subfamily Apocynoideae s.l. Gray bars indicate clades present in the minimum evolution bootstrap analysis; numbers *1*–*3* indicate clades for which the internal resolution is incongruent with the parsimony trees. Black bars indicate clades present in the bootstrap analysis with third codon position excluded; the letter *a* indicates taxa forming a clade that is incongruent with the parsimony tree.

s.l. (Sennblad and Bremer, in prep.). The nucleotide positions are classified in two schemes. The first is based on the structural or functional role of their corresponding amino acid (Knight et al., 1990): secondary structure, tertiary structure, quaternary structure, active site, and no identified function (henceforth called "no-function"). The second classification is based on codon position: first and second positions grouped together versus third position. The performance of the nucleotides between the different classes, as measured by *rc,* is tested statistically.

Materials and Methods

We use the parsimony analysis of Sennblad and Bremer (in prep.), in which *rbc*L sequences are sampled for 77 representatives of the Apocynaceae s.l. All tribes but one and a clear majority of all subtribes of the classification of Leeuwenberg (1994) and all tribes of the traditional Asclepiadaceae according to the classification of Liede and Albers (1994) were represented in this analysis. *Gelsemium* and *Mostuea* were chosen as outgroup taxa, because they were indicated to be the closest sister group to the Apocynaceae s.l. in the recent analysis of Gentianales by Backlund et al. (in press).

The analysis of Sennblad and Bremer (in prep.) included characters corresponding to the nucleotide positions 27–1,425 of the *rbc*L gene and characters corresponding to gaps and parsimony-informative nucleotide positions occurring up to 110 nucleotides downstream of position 1,425. Only parsimony-informative characters (237 in total) were analyzed. The cladistic analysis was performed by using PAUP 3.1.1 (Swofford, 1993). An initial heuristic search (PAUP settings: HSEARCH [ADDSEQ=RANDOM NREPS=100 SWAP=TBR]; other options with default settings) with all characters given unit weight resulted in 18,969 trees 829 steps long, with *ci* = 0.379 and *ri* = 0.678. This was followed by a successive weighting analysis (Farris, 1969, 1989), using heuristic searches (PAUP settings: HSEARCH [ADDSEQ=RANDOM NREPS=10 SWAP= TBR]; other options with default settings) and characters reweighted according to their *rc,* which resulted in 252 most-parsimonious trees, 178 steps long, with *ci* = 0.663 and *ri* = 0.880. The combinable consensus tree (Bremer, 1990; sometimes called semistrict consensus tree) is shown in Fig. 1. The length of these trees with unit weight characters is 831 steps; thus these trees are not identical to any of the trees from the unit weight analysis. Branches not present in the strict consensus tree from the unit weight analysis are indicated in Figure 1.

The bootstrap frequencies (Felsenstein, 1985), *boot,* indicated in Figure 1 are the ones reported by Sennblad and Bremer (in prep.). These were calculated with 10,000 replicates on the weighted data matrix (PAUP settings: BOOTSTRAP [NREPS=10,000 METHOD= HEURISTIC CONLEVEL=50] [/ADDSEQ= RANDOM NREPS=1 SWAP=SPR NOMUL PARS]; other options with default settings). In addition to bootstrap frequencies, we have indicated branch lengths in Figure 1. These are optimized with ACCTRAN optimization on an arbitrarily chosen most-parsimonious tree. The choice of ACCTRAN optimization is by convention, and one should recognize that branch lengths may vary with optimization.

In our study of performance of *rbc*L nucleotide positions, we included parsimony-informative positions from the coding region of *rbc*L (i.e., nucleotides 27–1,425). Nucleotide substitutions were classified into five classes, according to functions of the corresponding amino acid as identified by Knight et al. (1990), and into two classes according to their codon position. As mentioned earlier, the functional classes are secondary structure, tertiary structure, quaternary structure, active site, and no-function. The first four of these are overlapping to some extent, because an amino acid may participate in multiple functions. The two codon classes are first and second codon positions grouped together (because they may be assumed to be under similar constraints) and the third position.

The variable tested is the maximum *rc* of the character (a function of the homoplasy and fit of the character; Farris, 1989). For each functional class and for the codon classes we calculated, both within functional classes and over all sites the numbers, frequencies, and mean and standard deviation of *rc* (all of parsimony-informa-

tive sites). To test for significant differences in different combinations of groups, we used the SAS program package (SAS Institute, 1990) to perform a Wilcoxon two-sample test (cf. Sokal and Rohlf, 1995). The combinations tested were (a) secondary, tertiary, and quaternary structure and active-site classes—separately and pooled—versus the no-function class, and (b) codon classes over all characters and within each functional class.

We investigated whether dependence due to character covariance might occur in the data set by performing a crude substitution pattern test, using MacClade 3.07 (Maddison and Maddison, 1992, 1997). In investigating character covariance due to structural constraints at the amino acid level, the nucleotide sequences were translated into the corresponding amino acid sequences. We then searched for exactly similar distribution patterns of at least one state between two (or more) parsimony-informative characters (note that the state itself, e.g., leucine, need not be the same in the compared characters). Character covariance attributable to codon constraints was tested in a similar way: Exactly similar patterns between one (or more) states of first, second, or third parsimony-informative position within each codon were sought. The properties of the possible character covariance found above were further evaluated through comparison with the corresponding amino acids or nucleotides.

To test for possible biases on the phylogeny, we performed some additional analyses of characters 27–1,425, using PAUP 4.0b1a (Swofford, 1998).

A minimal evolution distance analysis (Rhzetsky and Nei, 1992) using maximum likelihood–estimated distances corrected for superimposed events (PAUP settings: DSET [DIST=ML OBJECTIVE=ME NEGBRLEN= PROHIBIT]; LSET [NST=6 RMATRIX= ESTIMATE BASEFREQ=ESTIMATE RATES= GAMMA SHAPE=0.5 PINVAR=0]; [HEU RISTIC START=NJ SWAP=TBR]; other options with default settings) was performed to investigate for long-branch attraction. The resulting tree was compared with the trees from Sennblad and Bremer (in prep.) by using the Kishino–Hasegawa test (Kishino and Hasegawa, 1989), with the same settings as above. To estimate branch supports, we performed a bootstrap analysis (PAUP settings: BOOTSTRAP [NREPS= 1000] [/START=NJ SWAP=SPR MULTREES= NO]; other settings as above). All characters were initially equally weighted in both analyses. Rescaled consistency indices, *rc,* were calculated for the minimum evolution tree, and a Wilcoxon two-sample test was performed on these *rc* values in the same way as described above.

A 5% χ-square test for unequal base composition was performed by using PAUP 4.0b1 (PAUP command: BASEFREQS) and Puzzle version 4.0.1 (Strimmer and von Haesler, 1996; default settings). This test ignores correlation to phylogenetic structure.

A bootstrap analysis was performed on a subset of the positions 27–1,425 data set with all third codon positions removed (PAUP settings: BOOTSTRAP [NREPS= 10,000 SEARCH=HEURISTIC] [/ADDSEQ= RANDOM NREPS=1 SWAP=SPR MUL TREES=NO]; other options with default settings). Only equally weighted parsimony-informative characters (66) were used.

Results

The results from the investigation of the performance of characters with differential functional constraints are presented in Tables 1–2 and Figure 2. The frequencies of parsimony-informative substitutions were approximately the same in all functional classes (12–17%, Fig. 2: all codon positions). The active-site class had the lowest frequency, and the secondary structure and no-function classes had the highest frequencies. However, the *rc* values for the parsimony-informative characters in the different classes (Fig. 2) were not significantly different (Table 1); the active-site class had the lowest mean *rc* (0.318 ± 0.353) and the tertiary structure class had the highest, *rc* = 0.403 ± 0.406 (Fig. 2). As can be seen from the standard deviations of these values, the variation of *rc* within the classes is considerable and in all classes the full range of *rc,* 0–1.000, is covered.

Parsimony-informative substitutions in the third codon position were much more frequent than substitutions in the first or second position (Fig. 2), as was also found

TABLE 1. P-values from the Wilcoxon two-sample test[a] for differences in rescaled consistency index for parsimony-informative substitutions among each of the classes with identified function and the no-function class.

	Compared class				
	Secondary structure	Tertiary structure	Quaternary structure	Active site	All classes except no function
No function	0.8196	0.5033	0.8657	0.9622	0.9858

[a]All test results were not significant.

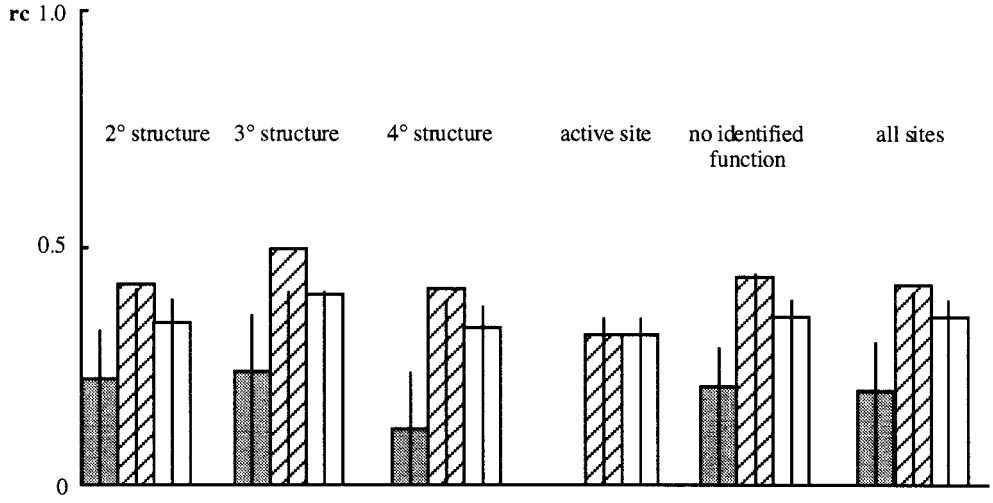
in the investigation of Olmstead et al. (1998). This could be viewed as being consistent with the premise that third-position sites are less conservative than are the first and second sites (e.g., Kellogg and Juliano, 1997). However, for parsimony-informative characters, Olmstead et al. (1998) found no significant difference in substitution rates between codon classes. Furthermore, in our study, third-position substitutions received significantly higher weights than did the first and second positions when tested over all parsimony-informative substitutions ($P < 0.001$; Fig. 2 and Table 2). This indicates that the information in the third codon position on average performs better in the present analysis than that in the two other positions. Within the different functional classes, the same general pattern can be seen; first and second position substitutions have mean weights between $rc = 0.119$ (quaternary structure class) and $rc = 0.240$ (tertiary structure class), whereas third positions have mean weights between $rc = 0.318$ (active-site class) and $rc = 0.497$ (tertiary structure class). This last-mentioned difference is significant at the 0.01 level in the classes of secondary structure and quaternary structure, and at the 0.05 level in the tertiary structure class. Because sample size may affect significance, the low number of substitutions in first and second positions of some classes may account in part for the lack of significance. In the active-site class, no substitutions at all occurred in the first and second positions; thus, this class could not be tested.

The test for character covariance found only one possible occurrence of character covariance that was attributable to structural constraints. A change in amino acid position 19 to aspartic acid coincided with a change in amino acid position 301 to leucine. Amino acid 301 is classified as quaternary structure, whereas no function has been identified for position 19. This may indicate either an unknown function for position 19, or coincident substitutions occurring, in this case, merely by chance. Assuming the latter case may be more conservative because the change is mapped as a single event on the tree. The corresponding changes at the nucleotide level occurred at different codon positions (position 1 for amino acid 19 and position 3 for amino acid 301). Five occurrences of exact matches between nucleotide changes within codons were found, all but one of which occurred at first and second position. However, the change occurring at the third position also brought about an amino acid change, meaning that all changes found were nonsilent. One of the possible character covariances in codons occurred in the quaternary structure class and two each in the tertiary structure class (including the one at third position) and the no-function class. Their fit to the present tree ranged from a perfect

TABLE 2. P-values from the Wilcoxon two-sample test for differences in rescaled consistency index among codon positions over all parsimony-informative substitutions and within each functional class. The active site contained only third-position substitutions and is thus not tested.

	Tested classification				
	Secondary structure	Tertiary structure	Quaternary structure	No function	Total
Codon 1+2 vs. 3	0.0080**	0.0235*	0.0031**	0.2438[ns]	0.0001***

***$P < 0.001$, **$P < 0.01$, *$P < 0.05$, [ns]not significant.



codon	1+2	3	all	1+2	3	all	1+2	3	all	1+2	3	all	1+2	3	all	1+2	3	all
n	46	70	116	19	33	52	18	48	66	0	7	7	11	42	53	66	152	218
f (%)	10	32	17	8.2	28	14	5.8	31	14	0	35	12	5.5	42	17	6.9	32	15
rc (‰)	225	425	345	240	497	403	119	414	333	-	318	318	209	439	391	200	423	356
SD (‰)	327	413	392	359	407	406	236	388	376	-	353	353	290	444	424	200	405	390

FIGURE 2. Statistics for parsimony-informative sites in the functional and codon position classes. Mean rescaled consistency indices and standard deviations are indicated in the histogram by bars and lines, respectively. Number (*n*) and frequency (*f*) of parsimony-informative sites, mean rescaled consistency index (*rc*), and standard deviation of the mean (*SD*) for each class are given below the histogram.

apomorphy (*ci* = 1.0), through being an apomorphic or autapomorphic state within an otherwise homoplastic multistate character, to complete homoplasy (*ci* = 0).

The minimum evolution analysis using maximum likelihood–estimated distances resulted in a single tree with a distance score of 0.62809. Although the topology of this tree is not completely congruent with the tree from Sennblad and Bremer (in prep.), the Kishino–Hasegawa test found no significant difference between the two analyses. Clades present in the minimum evolution bootstrap analysis are indicated in Figure 1. These clades are largely congruent with the tree from the analysis of Sennblad and Bremer (in prep.). However, the topology of some of these clades, e.g., *1–3* in Figure 1, differs slightly (clade *1: Ochrosia* and *Rauvolfia* as sister groups, clade *2: Cerbera* and *Thevetia* as sister groups, and clade *3: Mascarenhasia* and *Pachypodium* shifting positions). The main difference in tree topology between the optimal minimum evolution tree and the parsimony trees lies in the nesting order of the clades from the bootstrap analysis. In particular, the position of the *Ancylobotrys, Dictyophleba,* and *Vahadenia* clade is different; in the parsimony trees this clade is sister to the rest of the Apocynaceae, whereas in the minimum evolution tree it is sister to the subfamily Apocynoideae s.l. (indicated with an arrow in Fig. 1). These differences in topology (predominantly the different position of the *Ancylobotrys, Dictyophleba,* and *Vahadenia* clade) affected the optimization and *rc* of 33 characters. The mean *rc* for the different classes on the minimum evolution tree is presented in Table 3. The significance levels for differences in *rc* between different functional and codon position classes were unaffected except for codon positions in the secondary and tertiary classes, where the significance was lowered one level each, to 0.05 significance and nonsignificance, respectively.

The 5% χ-square tests, run in PAUP 4.0b1, and Puzzle 4.0.1, found no significant inequality in base composition.

The parsimony bootstrap analysis of the data set with third positions excluded was

TABLE 3. Mean rescaled consistency index for parsimony-informative substitutions in the different codon and functional classes from the minimum evolution analysis. Differences between these values and the values in Fig. 1 are indicated in parentheses.

	Functional class					
Codon class	Secondary structure	Tertiary structure	Quaternary structure	Active site	No-function	All sites
1 + 2	0.228	0.241	0.137	—	0.370	0.230
	(+0.003)	(+0.001)	(+0.018)		(+0.161)	(+0.030)
3	0.401	0.457	0.406	0.310	0.461	0.417
	(−0.024)	(−0.047)	(−0.008)	(−0.008)	(0.022)	(−0.006)
All sites	0.332	0.378	0.332	0.310	0.442	0.361
	(−0.013)	(−0.025)	(−0.001)	(−0.008)	(+0.051)	(+0.005)

largely unresolved. Twelve clades, each comprising two taxa, had a bootstrap support >50%. These clades are indicated in Figure 1. One of these clades (indicated with an *a* in Fig.1), with a bootstrap value of 51% and comprising *Arauija* and *Vincetoxicum,* is incongruent with the tree of Sennblad and Bremer (in prep.).

DISCUSSION

It is often assumed that nucleotides subjected to functional constraints are more conservative and thus less likely to show homoplasy or become saturated with change, than are nucleotides free of such constraints (Hillis and Bull, 1993; Miyamoto et al., 1994; Simon et al., 1994). However, our results show that, contrary to what could be expected, frequencies of substitutions at parsimony-informative sites were approximately the same in all classes: in total 15% of the nucleotide positions contained substitutions (Fig. 2, third row below figure). The frequency of substitutions in the active-site class was slightly lower (12%), and the frequency of the secondary structure and the no-function classes were slightly higher (17%). All of the substitutions in the active site occurred in the third codon position and did not give rise to amino acid changes. The higher frequency in the secondary structure class concurs with assumed higher substitution rates in the α-helices of Rubisco (Knight et al., 1990; Kellogg and Juliano, 1997). The study also indicates that there are no significant differences in *rc* values between any of the four function classes and the no-function class (Table 1). The mean *rc* of all classes lies relatively close to the mean *rc* over all substitutions, *rc* = 0.356 (Fig. 2). Furthermore, the standard deviations of the means indicate that the variation in *rc* within the classes is considerable (Fig. 2). These results contradict the premise that nucleotide sites under functional constraints generally behave better in parsimony analyses. There is thus no support for a routinely applied differential a priori weighting between functional positions and nonfunctional positions in parsimony analysis. Possible functions (i.e., not yet discovered) of amino acids in the no-function class may, however, bias the results above (see, e.g., the possible character covariance among amino acids in Results). Soltis and Soltis (1998) in an analysis of functional constraints in 18S rDNA suggested that a positional weighting scheme for functional constraints, based on properties of the data set at hand (cf. also certain model-based methods estimating model parameters from the investigated data), may be more appropriate. Nevertheless, the conclusion from their study was that this weighting had little effect on topology and support of the trees.

A similar analysis may be performed for codon positions. Nucleotide substitutions occurring in the third position of the codon are often silent; i.e., they do not give rise to an amino acid substitution. Such substitutions should occur with a higher frequency, because they presumably are subjected to lower selection, and thus are thought to be more susceptible to homoplastic events. On the other hand, all substitutions in the second codon position and most substitutions

in the first result in amino acid changes. However, our results show that the third codon positions have significantly higher *rc* than first and second positions. Even if the frequency of parsimony-informative third-position substitutions is much higher than for the other positions, their mean *rc* is at least 1.5 times those for the first and second positions (Fig. 2)—both for the complete data set and within the different functional classes. The difference is significant tested over all substitutions and in the secondary, tertiary, and quaternary structure classes (Table 2); small sample sizes may account for the lack of significance in the no-function class. In the active-site class only third-position changes occur, so a comparison within this class was not possible. Similar findings have been reported by Kim et al. (1992), in an analysis of the Asteraceae, in which substitutions in the second position had the highest percentage of homoplastic characters (90%), and in Olmstead (1998), wherein for *rbc*L and *ndh*F sequence data in Solanaceae, Lamiaceae, and Scrophulariaceae, third positions had the highest *ci* in four cases out of six (in all cases for *rbc*L).

The reason for the better performance of the third position is difficult to explain. A key feature may be that selection against certain amino acid changes effectively reduces the number of accepted nucleotide states. This could be expected to give a bias towards a higher frequency of multiple hits in the allowed substitutions. Thus, the disadvantages of fewer allowed states may outweigh the advantage of a lower substitution rate. Additionally, the study by Olmstead et al. (1998) on evolutionary rates indicates third-position substitution rates equal to or slightly lower than in the two other positions for parsimony-informative sites of *rbc*L. Attempting a more biological explanation could include hypothesizing the existence of a feedback-initiated DNA repair system that is activated by the presence of damaged enzymes or low concentrations of functional enzyme. Such a system would convert deleterious substitutions to allowed nucleotides, thereby effectively increasing the probability for multiple substitutions. However, several additional assumptions to the hypothesis are needed for such a system to have an effect over the necessary evolutionary time. Furthermore, the organization of the repair systems of the chloroplast is not well known, and the presence of a system described above is so far not supported.

The statistical test used, Wilcoxon two-sample test, assumes independence between data points, i.e., *rc* of nucleotides in this analysis (Sokal and Rohlf, 1995). Thus, dependence between nucleotide substitutions could be a potentially important drawback in this study. Several types of dependence occur or may occur among nucleotides: Dependence due to phylogeny is trivial, and dependence due to similar functional constraints is what is being tested. However, dependence due to character covariance between nucleotides (e.g., Wheeler and Honeycutt, 1988)—i.e., if a change at one nucleotide (or amino acid) position necessitates a change at another (separate) position—would violate the assumption of independence. The potential structural, as well as the codon, constraints in *rbc*L makes it possible that character covariance is at work in the data set.

Our test for character covariance found one possible occurrence of character covariance attributable to structural constraints and five possible occurrences attributable to codon constraints. The possible effect of character covariance was spread among the functional classes (and in two cases also among codon classes). In no case would the mean *rc* of a codon class change by >0.05 (in the tertiary class), and in all cases this change would increase the difference in mean *rc* between the codon classes. The changes in mean *rc* in functional classes would be even smaller, because of the larger sample size in these classes. Thus, we will here assume that a bias caused by character covariance in the data set will have little effect on the significance test.

Because third-position changes are so much more frequent in the data set, the topology may be largely dependent on them. Thus, if there is a problem with third positions, then the general lower weight for first and second positions in the present analysis could be due solely to incongruence with the (presumed problematic) third-position tree.

One such problem could be long-branch attraction (Felsenstein, 1978; Hendy and Penny, 1989), in which chance similarities between nonsister branches, arising because of the large number of changes occurring along the branches (i.e., long branch lengths), may bias phylogenetic analyses towards a "false" tree. Given its greater substitution rate, the third position may be hypothesized to be more sensitive to this effect than are the first and second positions. Maximum likelihood has been suggested (Felsenstein, 1978; Huelsenbeck, 1995; Yang, 1996) to be less sensitive than parsimony to long-branch attraction (assuming the model of evolution correctly reflects the actual evolutionary process), because of the possibility of including a gamma correction (Yang, 1994) for rate heterogeneity among nucleotides into the maximum likelihood model. Unfortunately, the number of taxa in this study (79) makes a maximum likelihood analysis almost impossible. We started a 6ST likelihood analysis with gamma correction included on the UNIX version of PAUP 4.0d064 (D. L. Swofford, pers. comm.), but the analysis had not finished when the search was terminated after 2 weeks.

We instead performed an alternative analysis to using maximum likelihood, which is to use a distance method such as minimum evolution (Rhzetsky and Nei, 1992) with maximum likelihood–estimated distances (including gamma correction). To investigate the support for different clades, we also performed a minimum evolution bootstrap analysis. The clades present in the bootstrap analysis are indicated in Figure 1. None of the clades from the bootstrap analysis is radically incongruent with the tree of Sennblad and Bremer (in prep.; see Fig. 1). The nesting order of these clades differs between the best tree from the minimum evolution analysis and the parsimony trees of Sennblad and Bremer (in prep.). However, the maximum likelihood differences between these trees were not significant when evaluated with the Kishino–Hasegawa test. Nevertheless, the differences in topology affected the *rc* of a total of 33 characters, spread among all functional and codon position classes. The effect of the different topology on the mean *rc* of the different classes is shown in Table 3. In most cases, the change in mean *rc* was small. However, in the no-function class the difference in mean *rc* between codon classes was reduced by 0.139 (Table 3), further supporting the nonsignificance found in this class. Furthermore, although the difference in mean *rc* was still quite large in the secondary structure and tertiary structure classes, the significance was reduced to the 0.05 level in the secondary structure class and became not significant in the tertiary structure class. Thus, the overall trend seems to be the same in the two analyses, and our results seem not to be adversely biased by potential effects from rate-heterogeneity in the data set. The minimum evolution analysis indicated a higher mean *rc* for first and second codon positions in the no-function class. Because nucleotides in this class presumably are subject to less constraint, this is consistent with the explanation based on reduction of the effective number of states, discussed above.

The other major problem that may affect characters in phylogenetic analysis is a nonphylogenetic bias working over all nucleotides. If first and second positions are less sensitive to this bias and support another incongruent phylogeny, this would explain the general lower weight for first and second positions in the present analysis. The major such bias known is uneven base composition (Hasegawa and Hashimoto, 1993; Steel et al., 1993; Lockhart et al., 1994). If there is a bias towards higher concentration of (e.g.) free guanosines and cytidines than of free adenosines and thymidines in a lineage, there might be a bias towards mutations to guanosines and cytidines in all nucleotide positions. If this occurs in several lineages, this may increase homoplastic events between these lineages. Again, a bias could be hypothesized to affect the third-position changes more because of being less constrained, and given the larger number of third positions, such changes may be driving the topology. Assuming that base composition in sequences reflects the base composition in free nucleotides, one may test the risk for this problem to occur. Using PAUP 4.0b1 and Puzzle 4.0.1 to perform a 5% χ-square test for unequal base composition among taxa, we found no significant differences. Thus,

there is little reason to assume a base composition bias.

Another possibility to test for topological incongruence between first plus second positions and third position is to do an analysis excluding third positions and investigate for incongruent groupings between this topology and the tree from the full data set (i.e., including third positions; cf. Källersjö et al., 1998). However, because of the small number of parsimony-informative first and second positions, such an analysis collects a large number of trees and could therefore not be completed (the first replicate of a 100 random addition sequence, TBR-branch swapping heuristic analysis had found 218,600 trees when interrupted and required >140 MB RAM memory). Therefore, we instead performed a bootstrap analysis, the result of which was largely unresolved; in total only 12 clades (each comprising two genera) was supported with a bootstrap support >50% (Fig.1). Comparing these groups to the tree of Sennblad and Bremer (in prep.) showed only one of the clades (indicated with *a* in Fig. 1) to be incongruent. However, this incongruent clade is contradicted by recent studies, molecular as well as morphological, supporting a sister relationship between *Tylophora* and *Vincetoxicum* (Liede, 1996; Civeyrel et al., 1998), as indicated in the tree from Sennblad and Bremer (in prep.). We believe this incongruence does not indicate problems with third positions driving the topology; rather, the low resolution of the tree may lend support to the lower performance of first and second position (cf. Källersjö et al., 1998).

Even if we naturally cannot completely exclude the possibility of errors in the tree of Sennblad and Bremer (in prep.), we feel, in light of the argument above, quite confident in the results. This is further supported by the congruence with other recent studies, partly or completely based on other types of data, mentioned earlier (e.g., Wanntorp, 1988; Judd et al., 1994; Endress et al., 1996; Civeyrel et al., 1998; Sennblad et al., 1998).

This study was performed on relatively close relationships, i.e., at the family level and below. When studying more distant relationships, the third position could be assumed to become more homoplastic as the longer branches increase the probability of multiple hits. Also, new first and second position changes may appear, which may behave in another manner, e.g., less homoplastic, than here. The behavior of functional constraints on nucleotides may thus vary with the degree of phylogenetic divergence of the sequences performed. However, a recent study using *rbc*L sequences at a higher systematic level, i.e., on green plant relationships, indicate a similar general pattern with the third codon positions performing best (Källersjö et al., 1998). This suggests that the pattern described in this study may be general.

Our results firmly contradict the usage of routinely giving a priori lower weight to third codon position nucleotides. The premise for such weighting is that the lower net substitution rate imposed by functional constraints on the first and second position will lower the probability for multiple substitutions (Irwin et al., 1991; Albert et al., 1993; Miyamoto et al., 1994; Simon et al., 1994; Swofford et al., 1996). We suggest that, contrary to this premise, the functional constraints by limiting the effective number of states may increase the net frequency of multiple hits in first and second position. This general behavior of functional constraints is also discussed in Albert et al. (1994). More importantly, the variation in performance among substitutions, i.e., *rc,* in the different codon positions was shown to be high, as evidenced by standard deviations, also when structural and functional constraints at the amino acid level were accounted for. This is also true for substitutions in functional classes. Thus, any routinely applied weighting scheme based on codon position or functional classes may be inappropriate. This is consistent with the conclusion of Olmstead et al. (1998) for weighting by codon position, based on their analysis of substitution rates in plastid genes, and of Soltis and Soltis (1998) for functional constraints in 18S rDNA. Our study of character performance does not extend to model-based methods such as maximum likelihood. However, we think it likely that the result from our study may have bearings for most phylogenetic methods that assume routinely a priori

weighting schemes based on functional and codon position classes of *rbc*L.

Acknowledgment

We thank I. Andersson, K. Bremer, M. Fishbein, R. Hirt, D. Horner, J. Maad, R. Olmstead, L.-G. Reinhammar, F. Ronquist, M. Sanderson, S. Ås, and an anonymous reviewer for valuable comments and discussions of an earlier draft of the manuscript. This study was supported by the Swedish Natural Science Research Council, NFR B-BU 1487–322 to B. B. and a travel grant from J. A. Wahlberg's memorial fund to B. S.

References

Albert, V. A., A. Backlund, K. Bremer, M. W. Chase, J. R. Manhart, B. D. Mishler, and K. C. Nixon. 1994. Functional constraints and *rbc*L evidence for land plant phylogeny. Ann. Mo. Bot. Gard. 81:534–567.

Albert, V. A., M. W. Chase, and B. D. Mishler. 1993. Character-state weighting for cladistic analysis of protein-coding DNA sequences. Ann. Mo. Bot. Gard. 80:752–766.

Andersson, I., S. Knight, G. Schneider, Y. Lindqvist, T. Lundqvist, C.-I. Brändén, and G. H. Lorimer. 1989. Crystal structure of the active site of ribulose-bisphosphate carboxylase. Nature 337:229–234.

Backlund, M., B. Oxelman, and B. Bremer. in press. Phylogenetic relationships within the Gentianales based on *ndh*F and *rbc*L sequences, with particular reference to the Loganiaceae. Am. J. Bot. (in press).

Bremer, B., K. Andreasen, and D. Olsson. in press. Subfamilial and tribal relationships in the Rubiaceae based on *rbc*L sequence data. Ann. Mo. Bot. Gard. 82:383–397.

Bremer, K. 1990. Combinable component consensus. Cladistics 6:369–372.

Chase, M. W., D. E. Soltis, R. G. Olmstead, D. Morgan, D. H. Les, B. D. Mishler, M. R. Duvall, R. A. Price, H. G. Hills, Y.-L. Qiu, K. A. Kron, J. H. Rettig, E. Conti, J. D. Palmer, J. R. Manhart, K. J. Sytsma, H. J. Michaels, W. J. Kress, K. G. Karol, W. D. Clark, M. Hedrén, B. S. Gaut, R. K. Jansen, K.-J. Kim, C. F. Wimpee, J. F. Smith, G. R. Furnier, S. H. Strauss, Q.-Y. Xiang, G. M. Plunkett, P. S. Soltis, S. M. Swensen, S. E. Williams, P. A. Gadek, C. J. Quinn, L. E. Eguiarte, E. Golenberg, G. H. Learn, Jr., S. W. Graham, S. C. H. Barrett, S. Dayanandan, and V. A. Albert. 1993. Phylogenetics of seed plants: An analysis of nucleotide sequences from the plastid gene *rbc*L. Ann. Mo. Bot. Gard. 80:528–580.

Civeyrel, L., A. Le Thomas, K. Ferguson, and M. W. Chase. 1998. Critical reexamination of palynological characters used to delimit Asclepiadaceae in comparison to the molecular phylogeny obtained from plastid *mat*K sequences. Mol. Phylogenet. Evol. 9:517–527.

Daugbjerg, N., Ø. Moestrup, and P. Arctander. 1994. Phylogeny of the genus *Pyramimonas* (Prasinophyceae, Chlorophyta) inferred from the *rbc*L gene. J. Phycol. 30:991–999.

Donoghue, M. J., R. G. Olmstead, J. F. Smith, and J. D. Palmer. 1992. Phylogenetic relationships of Dipsacales based on *rbc*L sequences. Ann. Mo. Bot. Gard. 79:333–345.

Endress, M. E., B. Sennblad, S. Nilsson, L. Civeyrel, M. W. Chase, S. Huysmans, E. Grafström, and B. Bremer. 1996. A phylogenetic analysis of Apocynaceae s.str. and some related taxa in Gentianales: A multidisciplinary approach. Opera Bot. Belg. 7:59–102.

Farris, J. S. 1969. A successive approximations approach to character weighting. Syst. Zool. 18:374–385.

Farris, J. S. 1989. The retention index and the rescaled consistency index. Cladistics 5:417–419.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–411.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–791.

Fredericq, S., and M. E. Ramírez. 1996. Systematic studies of the antarctic species of the Phyllophoraceae (Gigartinales, Rhodophyta) based on *rbc*L sequence analysis. Hydrobiologia 326/327:137–143.

Hasegawa, M., and T. Hashimoto. 1993. Ribosomal RNA trees misleading? Nature 361:23.

Hendy, M. D., and D. Penny. 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. 38:297–309.

Hillis, D. M., M. W. Allard, and M. M. Miyamoto. 1993. Analysis of DNA sequence data: Phylogenetic inference. Methods Enzymol. 224:456–487.

Hillis, D. M., and J. J. Bull. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42:182–192.

Huelsenbeck, J. P. 1995. Performance of phylogenetic methods in simulation. Syst. Biol. 44:17–48.

Irwin, D. T., T. D. Kocher, and A. C. Wilson. 1991. Evolution of the cytochrome *b* gene of mammals. J. Mol. Evol. 32:128–144.

Judd, W. S., R. W. Sanders, and M. J. Donoghue. 1994. Angiosperm family pairs: Preliminary phylogenetic analyses. Harv. Pap. Bot. 5:1–51.

Källersjö, M., J. S. Farris, M. W. Chase, B. Bremer, M. F. Fay, C. J. Humphries, G. Petersen, O. Seberg, and K. Bremer. 1998. Simultaneous parsimony jackknife analysis of 2538 *rbc*L DNA sequences reveals support for major clades of green plants, land plants, seed plant and flowering plants. Plant Syst. Evol. 213:259–287.

Kellogg, E. A., and N. D. Juliano. 1997. The structure and function of RuBisCO and their implication for systematic studies. Am. J. Bot. 84:413–428.

Kim, K.-J., R. K. Jansen, R. S. Wallace, H. J. Michaels, and J. D. Palmer. 1992. Phylogenetic implications of *rbc*L sequence variation in the Asteraceae. Ann. Mo. Bot. Gard. 79:428–445.

Kishino, H., and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. 29:170–179.

Knight, S., I. Andersson, and C.-I. Brändén. 1990. Crystallographic analysis of ribulose-1,5-bisphosphate carboxylase from spinach at 2.4 Å resolution. J. Mol. Biol. 215:113–160.

LEEUWENBERG, A. J. M. 1994. Taxa of the Apocynaceae above the genus level. Series of revisions of Apocynaceae XXXVIII. Wageningen Agric. Univ. Pap. 94: 45–60.

LIEDE, S. 1996. *Cynanchum–Rhodostegiella–Vincetoxicum–Tylophora* (*Asclepiadaceae*): New considerations on an old problem. Taxon 45:193–211.

LIEDE, S., AND F. ALBERS. 1994. Tribal disposition of genera in the *Asclepiadaceae*. Taxon 43:201–231.

LOCKHART, P. J., M. A. STEEL, M. D. HENDY, AND D. PENNY. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11:605–612.

MADDISON, W. P., AND D. R. MADDISON. 1992, 1997. MacClade: Analysis of Phylogeny and Character Evolution, version 3.07. Sinauer Associates, Sunderland, Massachusetts.

MANHART, J. R. 1994. Phylogenetic analysis of green plant *rbc*L sequences. Mol. Phylogenet. Evol. 3:114–127.

MIYAMOTO, M. M., M. W. ALLARD, R. M. ADKINS, L. L. JANECEK, AND R. L. HONEYCUTT. 1994. A congruence test of reliability using linked mitochondrial DNA sequences. Syst. Biol. 43:236–249.

OLMSTEAD, R. G., P. A. REEVES, AND A. C. YEN. 1998. Patterns of sequence evolution and implications for parsimony analysis of chloroplast DNA. Pages 164–187. *In* Molecular systematics of plants, Volume 2; DNA sequencing (P. S. Soltis, D. E. Soltis, and J. J. Doyle, eds.). Chapman and Hall, New York.

RHZETSKY, A., AND M. NEI. 1992. A simple method for estimating and testing minimum-evolution trees. Mol. Biol. Evol. 9:945–967.

SAS INSTITUTE. 1990. SAS/STAT User's guide, 6. SAS Institute Inc., Cary, North Carolina.

SCHNEIDER, G., S. KNIGHT, I. ANDERSSON, C.-I. BRÄNDÉN, Y. LINDQVIST, AND T. LUNDQVIST. 1990. Comparison of the crystal structures of L_2 and L_8S_8 Rubisco suggests a functional role for the small subunit. EMBO J. 9:2045–2050.

SENNBLAD, B., M. E. ENDRESS, AND B. BREMER. 1998. Morphology and molecular data in phylogenetic fraternity—the tribe Wrightieae (Apocynaceae) revisited. Am. J. Bot. 85:1143–1158.

SHINOZAKI, K., M. OHME, M. TANAKA, T. WAKASUGI, N. HAYASHIDA, T. MATSUBAYASHI, N. ZAITA, J. CHUNWONGSE, J. OBOKATA, K. YAMAGUCHI-SHINOZAKI, C. OHTO, K. TORAZAWA, B. Y. MENG, M. SUGITA, H. DENO, T. KAMOGASHIRA, K. YAMADA, J. KUSUDA, F. TAKAIWA, A. KATO, N. TOHDOH, H. SHIMADA, AND M. SUGIURA. 1986. The complete nucleotide sequence of the tobacco chloroplast genome: Its gene organization and expression. EMBO J. 5:2043–2049.

SIMON, C., F. FRATI, A. BECKENBACH, B. CRESPI, H. LIU, AND P. FLOOK. 1994. Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved polymerase chain reaction primers. Ann. Entomol. Soc. Am. 87:651–701.

SOKAL, R., AND F. J. ROHLF. 1995. Biometry, the principles and practice of statistics in biological research. W. H. Freeman, New York.

SOLTIS, P. S., AND D. E. SOLTIS. 1998. Molecular evolution of 18S rDNA in angiosperms: Implications for character weighting in phylogenetic analysis. Pages 188–210. *In* Molecular systematics of plants, Volume 2; DNA sequencing (P. S. Soltis, D. E. Soltis, and J. J. Doyle, eds.). Chapman and Hall, New York.

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

STRIMMER, K., AND A. VON HAESLER. 1996. Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. Mol. Biol. Evol. 13:964–969.

SWOFFORD, D. L. 1993. PAUP: Phylogenetic analysis using parsimony, 3.1.1. Center of Biodiversity, Illinois Natural History Survey, Champaign.

SWOFFORD, D. L. 1998. PAUP*. Phylogenetic analysis using parsimony (*and other methods), 4.0b1. Sinauer Associates, Sunderland, Massachusetts.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, AND D. M. HILLIS. 1996. Phylogenetic inference. Pages 407–514 *in* Molecular systematics (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer Associates, Sunderland, Massachusetts.

WANNTORP, H.-E. 1988. The genus Microloma (Asclepiadaceae). Opera Bot. 98:1–69.

WHEELER, W. C., AND R. L. HONEYCUTT. 1988. Paired sequence difference in ribosomal RNAs: Evolutionary and phylogenetic implications. Mol. Biol. Evol. 5:90–96.

XIANG, Q.-Y., D. E. SOLTIS, D. R. MORGAN, AND P. S. SOLTIS. 1993. Phylogenetic relationships of *Cornus* L. sensu lato and putative relatives inferred from *rbc*L sequence data. Ann. Mo. Bot. Gard. 80:723–734.

YANG, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Mol. Evol. 39: 306–314.

YANG, Z. 1996. Phylogenetic analysis using parsimony and likelihood methods. J Mol. Evol. 42:294–307.

Received 17 June 1998; accepted 2 April 1999
Associate Editor: M. Sanderson