Noncoding Sequences from the Slowly Evolving Chloroplast Inverted Repeat in Addition to *rbc*L Data Do Not Support Gnetalean Affinities of Angiosperms

Vadim Goremykin,† Vera Bobrova,* Jens Pahnke,† Aleksey Troitsky,* Andrew Antonov,* and William Martin†*

*A. N. Belozersky Institute of Physicochemical Biology, Moscow State University; and †Institut für Genetik, Technische Universität Braunschweig

We developed PCR primers against highly conserved regions of the rRNA operon located within the inverted repeat of the chloroplast genome and used these to amplify the region spanning from the 3′ terminus of the 23S rRNA gene to the 5′ terminus of the 5S rRNA gene. The sequence of this roughly 500-bp region, which includes the 4.5S rRNA gene and two chloroplast intergenic transcribed spacer regions (*cpITS*2 and *cpITS*3), was determined from 20 angiosperms, 7 gymnosperms, and 16 ferns (21,700 bp). Sequences for the large subunit of ribulose bisphosphate carboxylase/oxygenase (*rbc*L) from the same or confamilial genera were analyzed in both separate and combined data sets. Due to the low substitution rate in the inverted repeat region, noncoding sequences in the *cpITS* region are not saturated with substitutions, in contrast to synonymous sites in *rbc*L, which are shown to evolve roughly six times faster than noncoding *cpITS* sequences. Several length polymorphisms with very clear phylogenetic distributions were detected in the data set. Results of phylogenetic analyses provide very strong bootstrap support for monophyly of both spermatophytes and angiosperms. No support for a sister group relationship between Gnetales and angiosperms in either *cpITS* or *rbc*L data was found. Rather, weak bootstrap support for monophyly of gymnosperms studied and for a basal position for the aquatic angiosperm *Nymphaea* among angiosperms studied was observed. Noncoding sequences from the inverted repeat region of chloroplast DNA appear suitable for study of land plant evolution.

Introduction

Many questions concerning the general course of seed plant evolution, and in particular angiosperm evolution, are still not resolved (Chase et al. 1993; Martin et al. 1993; for a recent review see Crane, Friis, and Pedersen 1995). Early molecular studies of higher plant evolution involved protein sequence comparisons (Boulter et al. 1972; Martin and Jennings 1983). These were followed by nucleotide sequence analyses of rRNA (Hori, Lim, and Osawa 1985; Bobrova et al. 1987; Zimmer et al. 1989; Troitsky et al. 1991) and nuclear genes (Niesbach-Klösgen et al. 1987; Martin, Gierl, and Saedler 1989). With the advent of PCR techniques, cpDNA became the molecule of choice for plant molecular systematics (Palmer 1985; Palmer et al. 1988; Clegg and Zurawski 1992; Downie and Palmer 1992) *inter alia* due to its conservative mode of evolution. The recent widespread use of *rbc*L as a marker for evolutionary studies has had impact on plant molecular systematics (Chase et al. 1993; Baum 1994; Manhart 1994), yet further markers should be studied in order to derive a more robust picture of plant evolution. Due to the very low rate of nonsynonymous substitution in *rbc*L on the one hand and saturation of synonymous sites in *rbc*L in comparisons involving taxa that diverged during the early phases of land plant evolution on the other, *rbc*L sequences alone cannot resolve phylogeny at all taxonomic levels within higher plants (Martin et al. 1993). Additional molecular markers from cpDNA are needed.

Noncoding DNA has an advantage over coding DNA in that the number of potentially polymorphic sites per kilobase sequenced is higher (Böhle et al. 1994). In the absence of functional constraints, noncoding cpDNA in the single copy regions should undergo substitution at a rate similar to that observed at synonymous sites (Nei 1987, pp. 64–110). But in the inverted repeat (IR) region of cpDNA, the neutral substitution rate was estimated to be about threefold lower than that in the single copy regions (Wolfe et al. 1989). We reasoned that due to this lower substitution rate, noncoding regions of the IR may bear suitable markers for plant evolution.

Here we report the use of conserved primers directed against slowly evolving regions of the 5S and 23S rRNA genes for amplification of noncoding sequences from the IR region of cpDNA. Because the chloroplast 4.5S rRNA gene is flanked by two ITS regions, cpDNA possesses three internal transcribed rDNA spacers instead of two as in bacteria (Troitsky and Bobrova 1986). The PCR fragment contains two of these (*cpITS*2 and

Key words: noncoding DNA, angiosperms, gymnosperms, Gnetales, ferns, molecular phylogeny, chloroplast inverted repeat, molecular evolution.

Address for correspondence and reprints: William Martin, Institut für Genetik, Technische Universität Braunschweig, Spielmannstrasse 7, D-38023 Braunschweig, Federal Republic of Germany. E-mail: w.martin@tu-bs.de.

Mol. Biol. Evol. 13(2):383–396. 1996
 ISSN: 0737-4038

*cpITS*3), the 4.5S rRNA gene and the termini of the flanking 23S and 5S rRNA genes, respectively (roughly 500 bp per sequence). On the basis of sequences determined from 43 higher plants, we examined the utility of this region for reconstruction of plant phylogeny and compared it to *rbc*L from the same or closely related (confamilial) taxa.

Materials and Methods

Plant Material

Plant material for this study was collected from the Botanical Garden of Moscow University, from the Botanical Gardens of the Russian Academy of Sciences, from the Botanical Garden of the University of Braunschweig, and from the Botanical Garden of the University of Berlin. *Ceratopteris richardii* was a gift of Prof. L. G. Hickok. Species investigated are listed in table 1.

Molecular Methods

Plant DNA was isolated from either fresh or lyophilized leaf tissue ground in liquid nitrogen by the CTAB method (Murray and Thompson 1980) and subsequently purified by diafiltration in Microcon 30 columns (Amicon) according to the manufacturer's protocol. The diafiltration step was critical for DNA preparations from ferns and some gymnosperms. DNA was amplified using primers directed against highly conserved regions of the 23S and 5S rRNA genes flanking the 4.5S gene and spacers. The primers used were 5′ CCGGATAACTGCTGAAAGCATC 3′ and 5′ TCCTGGCGTCGAGCTATTTTTCC 3′. Each PCR reaction contained 0.4 μM of each primer, 3.5 mM $MgCl_2$, 50 mM KCl, 10 mM Tris-HCl (pH 8.3), 200 μM of each dNTP, approx 10 ng DNA, and 2 units Taq polymerase (Perkin Elmer) in a final volume of 50 μl.

Amplification was started with 3 min denaturation at 95°C and continued for 28 cycles of 50 s 95°C, 40 s 58°C, 60 s 72°C. PCR products were diluted to 400 μl and extracted once with phenol/chloroform and centrifuged. Primers and salts from aqueous supernatants were removed in Microcon 30 (Amicon) ultrafiltration devices according to the manufacturer's protocol using two additional diafiltration steps with 400 μl of 10 mM Tris (pH 8.0), 1 mM EDTA each.

Reverse-spin recovered amplification products of *Poa, Peperomia, Magnolia, Delphinium, Fagopyrum, Ephedra,* and *Cycas* were made blunt with Klenow polymerase as described (Sambrook, Fritsch, and Maniatis 1989), purified by diafiltration as above, and cloned in *Escherichia coli* nm522 using Sma I cut pBluescriptKS+ (Stratagene). Plasmids from these species were isolated from individual transformants and sequenced by the dideoxy method either with α-^{32}P dATP and T_7 DNA polymerase (Tabor and Richardson 1987) or by the automatic laser fluorescence method (Ansorge et al. 1986) with a commercially available apparatus (Pharmacia). All regions were sequenced from two independent subclones, in cases of ambiguity, a third clone was sequenced.

Aliquots of reverse-spin recovered amplification products from the other 35 species were electrophoresed against standards to determine DNA concentration. Aliquots were subjected to cycle sequencing in both directions with α-^{35}S dATP using a commercially available kit (Stratagene) according to the manufacturer's protocol except that 15 picomoles of the primers described above and 100 femtomoles of template were used. The sequencing reaction was performed for 30 cycles of 40 s 95°C, 40 s 58°C, 60 s 72°C. Sequences were resolved on 6% acrylamide gels (Sambrook, Fritsch, and Maniatis 1989).

Data Analysis

General sequence handling was performed with the GCG (version 8.0) package (Devereux, Haeberli, and Smithies 1984). The alignment was produced manually with the Vostorg package (kindly provided by A. Rhzetsky and A. Zharkikh). Programs of the Phylip (version 3.5; Felsenstein 1981, 1989) and Treecon (Van de Peer and De Wachter 1993) packages were used for tree construction. For *cpITS,* divergence was estimated using the two-parameter method of Kimura (1980) as numbers of substitutions per site (for convenience, referred to here as d_k) or with the gamma distance (Jin and Nei 1990). A gamma parameter of 1.3 was estimated from the data using the method of Ota and Nei (1994) on the basis of the Kimura distance tree. For *rbc*L sequences, sequence divergence was estimated as numbers of synonymous and nonsynonymous substitutions per site (K_s and K_a, respectively) with the methods of Li, Wu, and Luo (1985) and Nei and Gojobori (1986). Additional statistical analyses were performed with the Kaleidagraph program for MacIntosh (Abelbeck Software, Inc.).

Results

The *cpITS* Data Set

We determined 43 *cpITS* sequences from various land plants and retrieved seven others from the database for analysis. Each sequence entry spans from the 3′ 57 bp of the 23S rRNA gene to the 5′ 30 bp of the 5S rRNA gene (fig. 1). In some cases the sequence of the initial ~20 nucleotides was not readable due to proximity to the primer binding sites, a total of 21,700 bases were determined unambiguously. In order to obtain an overall impression of sequence conservation across the investigated region, we plotted the degree of sequence identity for each position in the 50 OTU alignment (fig. 1). The *cpITS* region contains highly conserved (rDNA

Table 1
Species Investigated in This Study

Species (*cpITS*)	Acc. No.[a]	Family[b]	Species (*rbc*L)	Acc. No.
Angiosperms				
Epifagus virginiana	M81884*	Orobanch'	—	—
Conopholis americana	X58863*	Orobanch'	—	—
Fagopyrum sagittatum	L41604	Polygon'	*Rheum* × *cultorum*	M77702
Nicotiana tabacum	Z00044*	Solan'	*Nicotiana tabacum*	J01450
Alnus incana	M75719*	Betul'	*Betula niger*	L01889
Alchemilla vulgaris	L41580	Ros'	*Geum chiloense*	L01921
Eryngium billardieri	L41602	Api'	*Apium graveolens*	L01885
Ferulago galbanifera	L41564	Api'	*Conium maculatum*	L11167
Pisum sativum	M37430*	Fab'	*Pisum sativum*	X03853
Delphinium elatum	L41598	Ranuncul'	*Ranunculus trichophyllus*	L08766
Caryota mitis	L41592	Arec'	*Caryota mitis*	M81811
Cryptocoryne ciliata	L41594	Ar'	*Gymnostachys anceps*	M91629
Bambusa multiplex	L41591	Po'	*Bambusa multiplex*	M91626
Poa pratensis	L41587	Po'	*Pennisetum glaucum*	L14623
Oryza sativa	X15901*	Po'	*Oryza sativa*	D00207
Molineria recurvata	L41547	Po'	*Avena sativa*	L15300
Semele androgyna	L41571	Rusc'	*Danae racemosa*	L05034
Tillandsia usneoides	L41573	Bromeli'	*Tillandsia elizabethae*	L19971
Strelitzia nicolaii	L41572	Magnoli'	*Strelitzia nicolaii*	L05461
Magnolia campbellii	L41568	Magnoli'	*Magnolia salicifolia*	L12656
Annona montana	L41582	Magnoli'	*Anonna muricata*	L12629
Eupomatia laurina	L41603	Magnoli'	*Eupomatia bennettii*	L12644
Drimys winterii	L41600	Magnoli'	*Drimys winteri*	L01905
Piper longum	L41586	Piper'	*Piper betle*	L12660
Peperomia glabrata	L41550	Piper'	*Peperomia* sp.	L12661
Nymphea coerulea	L41548	Nymphae'	*Nymphaea odorata*	M77034
Gymnosperms				
Zamia floridiana	L41556	Zami'	*Zamia inermis*	L12683
Cycas revoluta	L41596	Cycad'	*Cycas circinalis*	L12674
Ginkgo biloba	L41565	Ginkgo'	*Ginkgo biloba*	D10733
Pinus canariensis	L41585	Pin'	*Pinus radiata*	X58134
Welwitschia mirabilis	L41555	Welwitschi'	*Welwitschia mirabilis*	D10735
Gnetum gnemon	L41566	Gnet'	*Gnetum gemon*	L12680
Ephedra kokanica	L41601	Ephedr'	*Ephedra tweediana*	L12677
Ferns				
Phyllitis scolopendium	L41551	Aspleni'	*Asplenium nidus*	U05907
Polypodium aureum	L41588	Polypodi'	*Colysis sintenensis*	U05612
Davallia bullata	L41597	Davalli'	*Davallia epiphylla*	U05917
Athyrium sp.	L41583	Dryopteridi'	*Athyrium felix-femina*	U05908
Pteris cretica	L41570	Pterid'	*Pteris fauriei*	U05647
Adiantum capillus-veneris	L41579	Pterid'	*Adiantum pedatum*	U05602
Ceratopteris richardii	L41593	Pterid'	*Ceratopteris thalictroides*	U05609
Dicksonia antarctica	L41599	Dicksoni'	*Dicksonia antarctica*	U05618
Pilularia globulifera	L41584	Marsile'	*Marsilea quadrifolia*	L13480
Cyathea cooperi	L41595	Cyatha'	*Cyathea lepifera*	U05616
Azolla anabenae	L41590	Salvini'	*Salvinia cucullata*	U05649
Trichomanes radicans	L41554	Hymenophyll'	*Cephalomanes thysanostomum*	U05608
Osmunda regalis	L41549	Osmund'	*Osmunda cinnamomea*	D14882
Angiopteris palmiformis	L41581	Maratti'	*Angiopteris evecta*	L11052
Psilotum triquestrum	L41569	Psilot'	*Psilotum nudum*	L11059
Lycopods, Bryophytes				
Lycopodium bifurcatum	L41567	Lycopodi'	*Lycopodium digitatum*	L11055
Marchantia polymorpha	X04465*	Marchanti'	*Marchantia polymorpha*	X04465

[a] *cpITS* sequences that were not determined in this paper are indicated with an asterisk.
[b] Family names are abbreviated with an apostrophe for "-aceae."

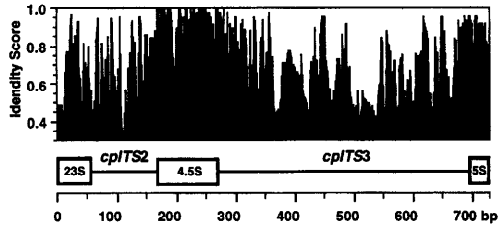

FIG. 1.—Identity profile across the 50 *cpITS* sequences and rRNA genes analyzed in this study. The PROFILE program of the Wisconsin package was applied to the prealigned sequences using a one-base window; identity of gaps was counted as dissimilarity. The drop in similarity in the 23S and 5S regions is due to the presence of undetermined bases in these regions in some sequences as a result of their proximity to the sequencing primer binding sites. An identity score of 1.0 indicates complete site conservation.

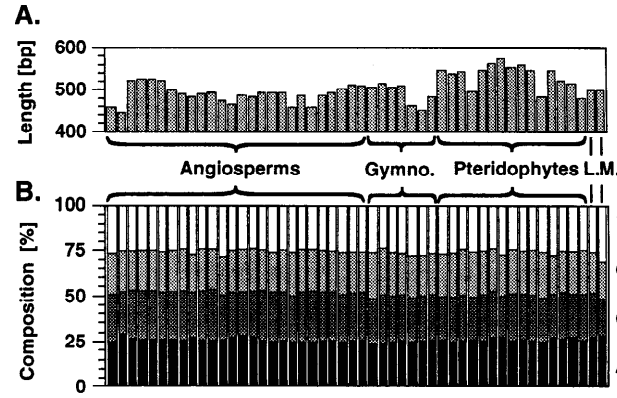

FIG. 2.—Length and base composition variation in land plant *cpITS* sequences. Species order from left to right corresponds to vertical order in table 1. *A.* Histogram of *cpITS* length across OTUs (excluding gaps). Gymno, gymnosperms; L, *Lycopodium*; M, *Marchantia*. *B.* Base composition (in %) of land plant *cpITS* sequences.

coding) and highly variable stretches. A considerable portion of variability observed is due to numerous indels present in the *cpITS*2 and *cpITS*3 regions. Despite the high degree of variation in length in both intergenic transcribed spacers, several shorter (~30 bp) regions exist within each with an identity score >0.6 that aid considerably in alignment (see also below). The total alignment of *cpITS* sequences covers 731 positions, but the average length of raw *cpITS* sequences we determined is only about 510 bp (fig. 2*A*). The shortest sequence analyzed was that of the parasitic angiosperm *Conopholis* (444 bp), the longest was 585 bp, found in the leptosporangiate fern *Ceratopteris*. In order to assess variation in G+C content, we plotted the base composition for each sequence (fig. 2*B*). Nucleotide composition in the *cpITS* region is extremely homogeneous across land plant taxa. Only the sequence from the *Marchantia* displays a slightly lower G+C content, but because *Marchantia* is the outgroup in our phylogenetic analyses, varying G+C content across ingroup OTUs should not pose problems in phylogenetic analyses.

A number of indels show a very clear phylogenetic distribution, such as an 8-bp deletion shared by the three grasses *Poa, Oryza,* and *Bambusa* at position 197 of the alignment within the highly conserved region encompassing the 4.5S rRNA gene (fig. 3). Outside of the rRNA coding regions, indels are much more abundant. In figure 4 the most highly variable region of the alignment is shown, corresponding to the region around position 500 in figure 1. Although the placement of indels and identification of homologous regions *within* ferns and *within* spermatophytes are generally clear in this highly variable region, *across* these groups assignment of unambiguous positional homology in this segment of the alignment becomes tenuous. Despite the high degree of variability, several indels within this region also show a marked phylogenetic distribution. Examples are Δ562–576 in the two cycads, Δ519–550 in angiosperms, or Δ507–513 in angiosperms surveyed except *Nymphaea*. Other indels appear autapomorphic in this taxon sample but may show an ordered phylogenetic distribution as more sequences are obtained. The general impression of positional homology in this difficult region of the alignment tended to improve as more taxa were introduced. Positions such as 506–510 in the gnetophytes *Welwitschia* and *Gnetum* (and other positions in other OTUs) are still not clear because they entail short duplications; their placement is therefore somewhat ambiguous but is not wholly random in light of the surrounding motifs found in other gymnosperms. In the region shown in figure 4, numbers of substitutions per site between distantly related taxa will be underestimated.

Due to this high degree of sequence dissimilarity in the most variable region of *cpITS*3, we examined patterns of sequence divergence prior to distance estimation or tree construction. First we determined frequencies of transitions and transversions observed in the *cpITS* data set. The numbers of transitions and transversions observed are positively correlated (fig. 5) ($R^2 = 0.83$). The 15 or so values that scatter sparsely above the majority of points plotted involve comparisons within ferns. Considering the large number of comparisons under consideration, the shape of the distribution is quite uniform. Transition-to-transversion ratios were calculated for *cpITS* sequences. The average transition/transversion ratio for all pairwise comparisons is 1.99. Largest deviations from the average ratio are observed at low values of total divergence where stochastic variation is greatest.

Positional homology in the functionally conserved rRNA coding regions of the alignment is unambiguous (fig. 3). We reasoned that if positional uncertainty in highly variable (nonconstrained) regions of the alignment introduces randomness into distance measure-

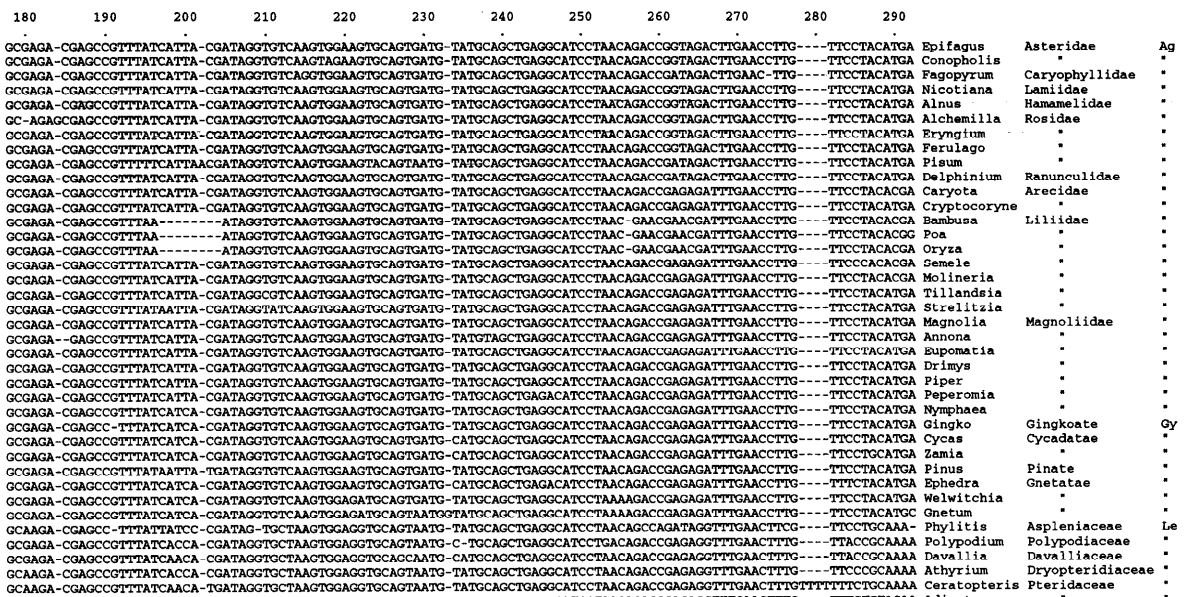

FIG. 3.—Segment of the nucleotide alignment in the region of the 4.5S rRNA gene (marked by asterisks). Indels are indicated as dashes. Nucleotide positions indicated refer to the alignment used for phylogenetic analysis, in which the terminal base of the 23S gene is arbitrarily located at position 25. Subclass (spermatophytes) and family (ferns) assignments are indicated. Ag, angiosperms; Gy, gymnosperms; Le, leptosporangiate ferns; Eu, eusporangiate fern; Ps, psilophyte; Ly, lycopsid; Br, bryophyte.

ments, then little correlation should be observed between distances estimated from the rDNA coding regions and those estimated from the noncoding regions of the alignment. We estimated the number of substitutions per site using the Kimura two parameter method (d_k) separately for rDNA and noncoding spacer regions of *cpITS* for pairwise comparisons of all OTUs. We excluded *Epifagus* and *Conopholis* in these comparisons, because we wished to compare divergence in *cpITS* regions to that in *rbc*L (see below) and these species do not possess functional *rbc*L genes (Wolfe, Morden, and Palmer 1992). Values of divergence in coding and noncoding regions of the *cpITS* region were plotted against one another (fig. 6*A*). The highest values of divergence for the noncoding region do not exceed 0.9 substitution per site even in comparisons between ferns and spermatophytes. Notably, there is a very positive correlation between sequence divergence in the coding and noncoding portions of the *cpITS* data (R^2 = 0.922). The average ratio of substitution rates in noncoding *vs.* coding *cpITS* regions for all comparisons is 2.05 ± 0.018. For more closely related sequences, i.e., for values of divergence in noncoding regions <0.3 (475 comparisons), the correlation becomes slightly weaker (R^2 = 0.81, data not shown) and the average ratio of substitution rates in noncoding *vs.* coding regions of the *cpITS* becomes 1.66 ± 0.03. This drop in correlation is probably due to the very small number of rDNA coding sites (173) in the alignment. If divergence in the *cpITS*3 region alone is plotted against divergence in rDNA coding regions for all comparisons, the correlation remains strongly positive (R^2 = 0.767, data not shown), indicating that also the most variable noncoding regions are not saturated with substitutions over most of their length, even in comparisons between spermatophytes and ferns, and that pairwise distances between noncoding regions—although high—are not randomized through saturation.

For comparison to other data widely used for the study of plant evolution, we plotted estimates of numbers of substitutions per site determined from constrained (nonsynonymous) and nonconstrained (synonymous) sites between *rbc*L sequences from the same or congeneric species; in those cases where such sequences were not available in the database, we used *rbc*L sequences from confamilial genera (see table 1). For *rbc*L, we measured divergence at synonymous and nonsynonymous sites using the method of Li, Wu, and Luo (1985)

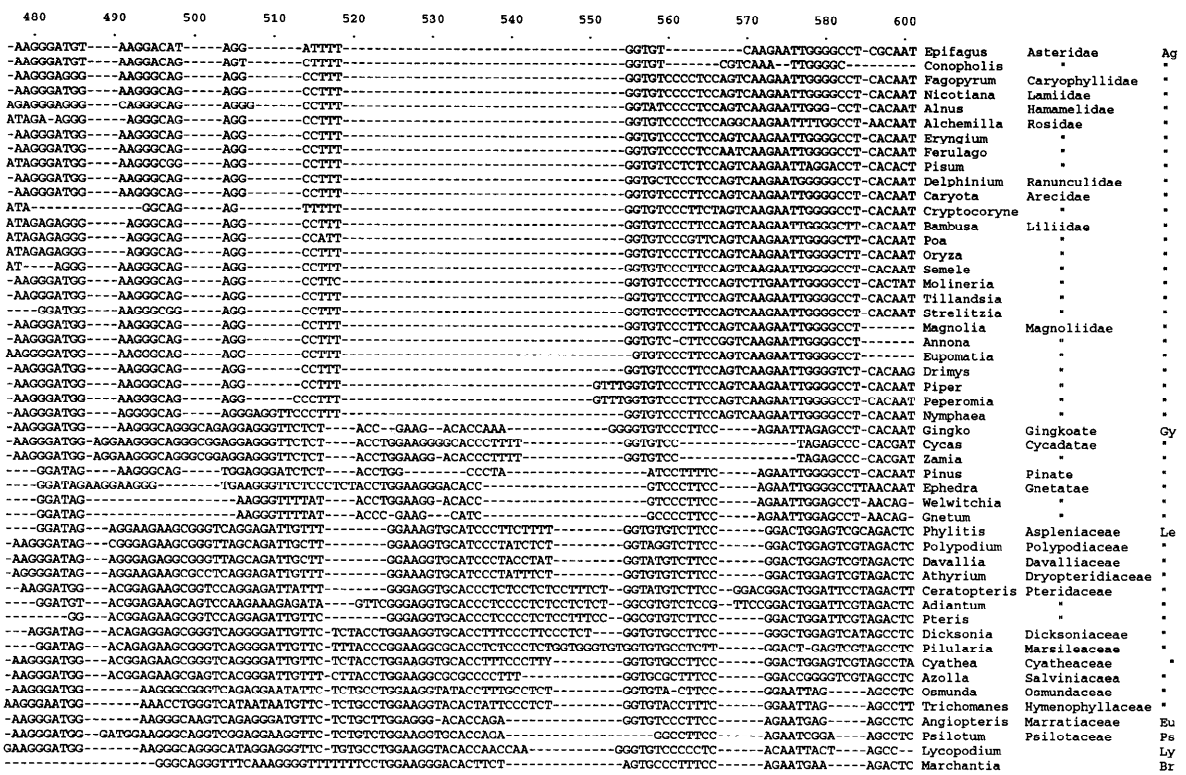

FIG. 4.—Segment of the nucleotide alignment in the region of highest variability. Designations as in the legend to figure 3.

and plotted these values against one another; similar results were obtained using the method of Nei and Gojobori (1986) (see below). Figure 6*A* shows that the divergence at nonsynonymous sites (K_a) for land plant *rbc*L sequences surveyed is very low, less than 0.08 substitutions per site in all cases. At synonymous sites, by comparison, estimates of divergence between *rbc*L sequences (K_s) are very high, greater than one substitution per site in most cases, and therefore very unreliable. The correlation between K_a and K_s is poor for the *rbc*L land plant data set ($R^2 = 0.18$). Even in comparisons at lower values of divergence for *rbc*L ($K_s < 0.6$), the correlation between K_a and K_s is poor ($R^2 = 0.16$, 266 comparisons). It is quite obvious that the rate of substitution at synonymous sites is much higher in *rbc*L than at nonsynonymous sites. For all comparisons, the average ratio of K_s/K_a is 29, for values of $K_s < 1$ (571 comparisons), the average ratio of K_s/K_a is 21; for values of $K_s < 0.6$ (266), average ratio of K_s/K_a is 16. This is not surprising but stands in sharp contrast to assertions that the rates of substitution at synonymous and nonsynonymous sites in *rbc*L may be quite similar (Chase et al. 1993). This result also indicates that a systematic error exists in the calculations of Albert et al. (1994), because they estimated substitution rates between *rbc*L sequences by dividing the total proportion of nucleotide differences between sequences by estimated divergence time. In light of the great difference between synonymous and nonsynonymous rates in *rbc*L, Albert et al.'s estimates of sequence divergence and substitution rate in *rbc*L are erroneous.

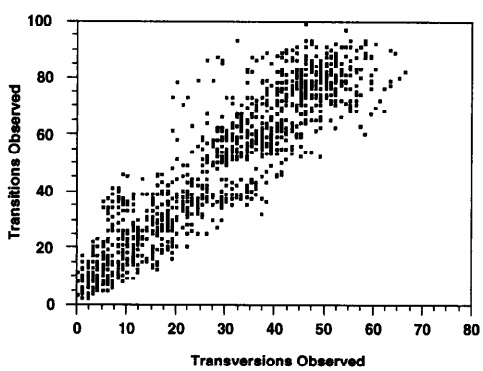

FIG. 5.—Transitions and transversions in *cpITS* sequences. Plot of numbers of observed transitions *vs.* observed transversions in pairwise comparisons of *cpITS* sequences; each point represents the plot for one pairwise comparison.

A Low Substitution Rate in Noncoding *cpITS* Regions

Sequences within the inverted repeat region of chloroplast DNA have a lower neutral substitution rate than those in the single copy regions (Wolfe, Li, and Sharp 1987). Using the data set at hand, we wished to

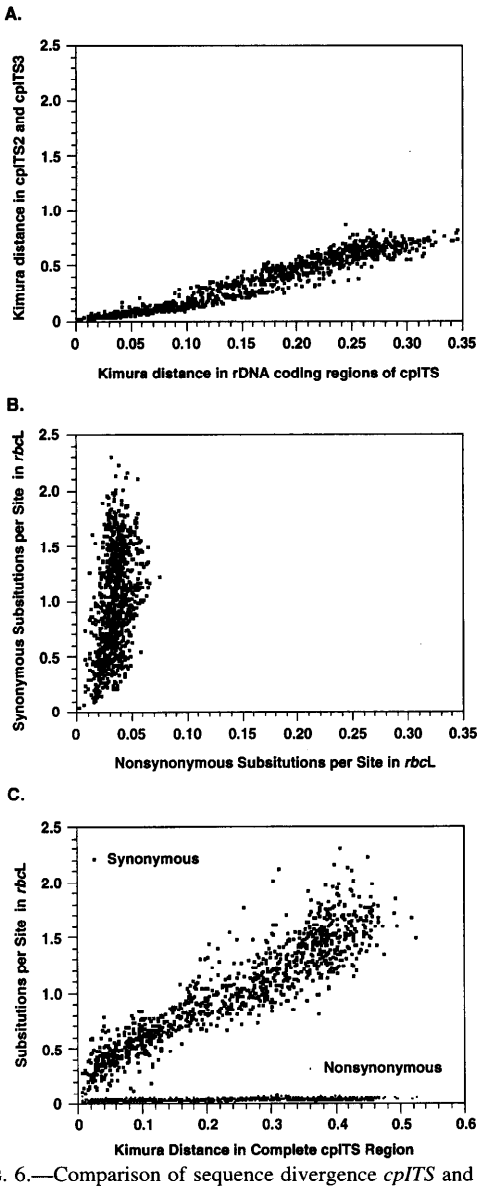

FIG. 6.—Comparison of sequence divergence *cpITS* and *rbc*L sequences. *A*. Plot of sequence divergence (estimated by the two-parameter method (Kimura 1980)) at functionally constrained (rRNA coding) and unconstrained (noncoding) positions of *cpITS* sequences. Noncoding regions correspond to combined *cpITS*2 and *cpITS*3 regions as shown in figure 1 (avg. 370 positions). Scale units are substitutions per site. Each point represents the plot of respective values for an individual pairwise comparison. *B*. Plot of sequence divergence (estimated by the method of Li, Wu, and Luo [1985]) at functionally constrained (nonsynonymous) and unconstrained (synonymous) positions of *rbc*L sequences (K_a and K_s, respectively). An average *rbc*L pair compared here has 996 nonsynonymous and and 298 synonymous sites, respectively, less than complete sequences due to missing data in PCR entries. Note that axis scales are identical to those in (*A*) for direct comparison of divergence at constrained *vs.* unconstrained positions in *rbc*L and *cpITS* and for direct comparison of overall sequence divergence in the two markers. *C*. Numbers of synonymous and nonsynonymous substitutions per site in pairwise comparisons of land plant *rbc*L sequences plotted against Kimura distance (d_k) between aligned *cpITS* sequences for the same (or confamilial) taxa (see table 1). Use of a single ordinate scale is intentional to underscore the low divergence at nonsynonymous sites in *rbc*L sequences.

Table 2
Ratio of Substitution Rates at Synonymous Sites in *rbc*L to Kimura Distance in Noncoding *cpITS* Regions

Range of $K_{s,rbcl}$	Average ratio $K_{s,rbcl}/d_{k,cpITS2/3}$	N[a]	Min	Max
<0.2	4.77	16	1.10	11.7
0.2–0.3	6.92	33	2.45	17.6
0.3–0.4	7.24	62	2.18	19.4
0.4–0.5	6.97	82	2.51	22.4
0.5–0.6	5.76	71	2.57	18.1
0.6–0.7	5.18	84	2.07	16.6
0.7–0.8	4.44	62	1.77	13.9
0.8–0.9	2.97	75	1.24	6.92
0.9–1.0	2.74	84	1.63	4.22
>1.0	2.58	510	1.67	5.04
Average[b] $K_{s,rbcl}$ (<0.8)	5.9	410		

[a] N indicates number of pairwise comparisons in the given range of $K_{s,rbcl}$. Minimum and maximum values of $K_{s,rbcl}/d_{k,cpITS2/3}$ observed for the range are indicated.

[b] For calculation of the average, values from the range of $K_{s,rbcl} > 0.8$ were excluded because saturation at synonymous sites is observed, particularly evident in the column for maximum values.

estimate the relative rates of nucleotide substitution at synonymous sites in *rbc*L and noncoding regions of *cpITS*. Values of K_s for *rbc*L were divided by values of d_k in the noncoding *cpITS* regions (*cpITS*2 and *cpITS*3 combined, designated here as *cpITS*2/3) for corresponding comparisons. This was performed for several ranges of K_s in *rbc*L (table 2). We did not perform a relative rate test prior to calculation of average rates, but the effects of the most rapidly and slowly evolving sequences in the relatively large data set probably counteract one another. Both the average ratio of substitution rates and the maximum values of same decline sharply above values of $K_s > 0.8$ substitutions per site, probably due to saturation and underestimation of divergence. In 410 comparisons for values of $K_s < 0.8$, the average ratio of numbers of substitution per site at synonymous sites in *rbc*L and *cpITS*2/3 was 5.9. Thus, although the *cpITS*2/3 region is noncoding chloroplast DNA, its rate of substitution is about six times lower than that at synonymous sites in *rbc*L. For the same 410 comparisons, the average ratio of substitution rate in *cpITS*2/3 to nonsynonymous substitution rate in *rbc*L was sightly greater than four, but with an extremely wide range, as evident from the wide variation in K_a at low values of K_s seen in figure 6*B*. The reduction in substitution rate for *cpITS*2/3 relative to K_s in *rbc*L could either be due to structural constraints imposed by rRNA transcript processing, by copy correction in the inverted repeat, or both. These results indicate that the *cpITS* region, and perhaps other noncoding regions of the inverted repeat in cpDNA, are sufficiently conserved as to be phylo-

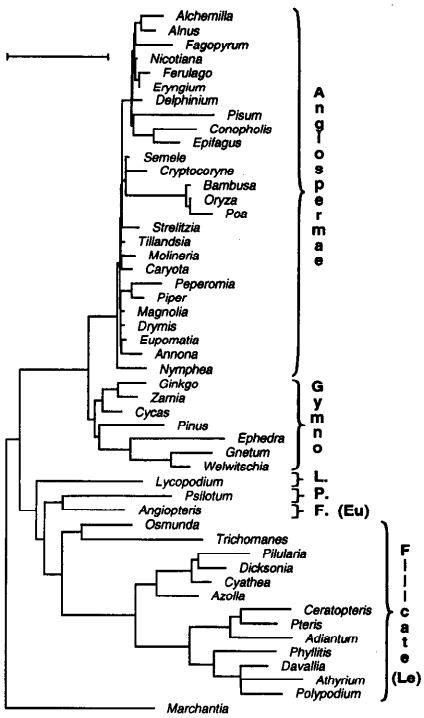

FIG. 7.—Neighbor-joining (NJ) tree (Saitou and Nei 1987) for cpITS sequences using the Kimura distance. The scale bar indicates 0.1 substitutions per site. L, Lycopodiaceae; P, Psilotaceae; F, Filicatae; Eu, eusporangiate; Le, leptosporangiate.

genetically useful in comparisons of land plant taxa, the *rbc*L sequences of which are saturated at synonymous sites.

Phylogenetic Analyses

Results in the previous section indicated that the *cpITS* region should be suitable for phylogenetic analyses: the base composition is quite constant, the degree of divergence is not too extreme (<0.3 substitutions per site in most cases for the entire region, <0.6 all cases), and transitions are twice as frequent as transversions. The Kimura two-parameter distance (d_k) performs well under these parameters (Jin and Nei 1990) and was used here to estimate sequence divergence. But because substitution rate varies considerably across sites (fig. 1), we also used the method of Jin and Nei (1990) for comparison.

Figure 7 shows the neighbor-joining (Saitou and Nei 1987) tree for *cpITS* sequences constructed from Kimura distance values and provides a general impression of the data. The most notable feature of the tree is the very low degree of divergence observed between most angiosperm taxa. Several angiosperm sequences are borne on long branches, suggesting an elevated substitution rate relative to other angiosperms (*Pisum, Bambusa, Oryza, Poa, Epifagus,* and *Conopholis*). In the case of *Pisum,* this may be due to the loss of one copy of the inverted repeat in the cpDNA (Palmer and Thompson 1981), because the presence of two copies of the inverted repeat appears to reduce the rate of nucleotide substitution in the IR region (Wolfe, Li, and Sharp 1987). For the grasses, the elevated substitution rate in cpDNA reported for Poaceae (Gaut et al. 1992) may also apply to the IR region. For *Epifagus* and *Conopholis,* the apparent elevation of substitution rate is likely due to loss of functional constraints in the cpDNA of these parasitic plants (Wolfe, Morden, and Palmer 1992). Considerably greater sequence divergence is observed in *cpITS* sequences in comparisons between ferns than between seed plants. Spermatophytes are separated from remaining taxa by a very robust branch, the length of which may be exaggerated due to the difficulties in aligning variable regions across this boundary.

The reliability of the topology was estimated by bootstrapping. The 80% bootstrap proportion consensus NJ tree for *cpITS* sequences is shown in figure 8*A*; the threshold of 80% was chosen arbitrarily. Results of bootstrapping using the Kimura distance or Jin and Nei (1990) distance are summarized in the figure. The gamma parameter of 1.3 estimated from the *cpITS* data is probably too low, but gamma parameters of 1.0 or 2.0 gave identical topologies at the 50/100 bootstrap proportion consensus level (data not shown). Using either gamma distance, only one branch was found in 80 or more replicates (a common branch for *Alchemilla* and *Alnus* in 82/100 with a gamma parameter of 2, found in 72/100 with d_k) that was not found in 80 or more replicates using d_k. Conversely, only one branch was detected in 80 or more replicates using d_k that was found in less than 80 replicates using the gamma distances (the common branch for *Ginkgo, Zamia,* and *Cycas,* 76/100). Thus, the topologies obtained were very similar with different distance estimation methods, although absolute branch lengths were slightly (~10%) greater with the gamma distances as compared to those obtained for d_k.

The position of the Gnetales relative to angiosperms and other gymnosperms is of interest, because several lines of data point to Gnetales as the sister group to angiosperms. This relationship is not resolved in figure 8*A,* which provides a conservative view of the *cpITS* gene phylogeny. As shown in figure 8*B,* the data do not support a sister group relationship between angiosperms and Gnetales, but rather provide weak support (about 50/100 replicates) for monophyly of gymnosperms surveyed. Although divergence between *cpITS* sequences is rather high for maximum parsimony analyses, we constructed bootstrap parsimony trees for the alignment to see if it provided support for sister group affinities between Gnetales and angiosperms. Using parsimony, the branch shared by *Pinus* and Gnetales in figure 8 occurred in more than 90/100 replicates.

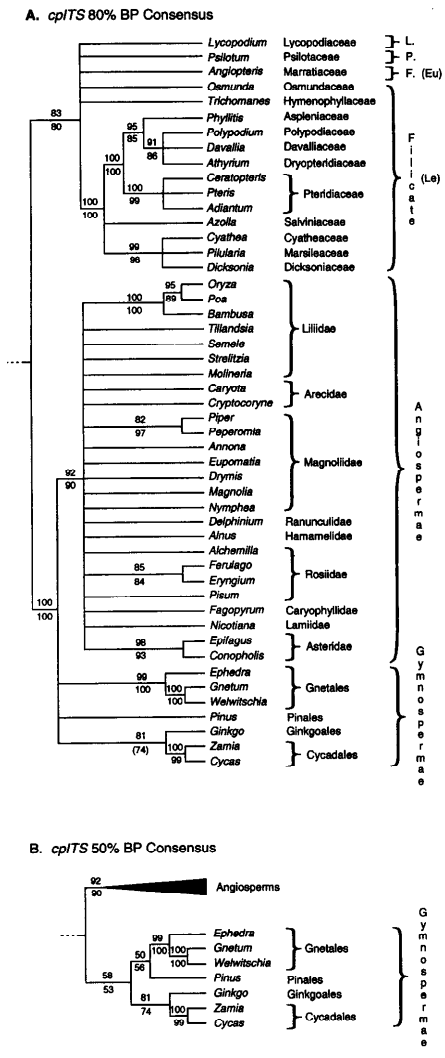

FIG. 8.—Trees derived from *cpITS* sequences. *Marchantia* was used as the outgroup. *A.* 80% bootstrap proportion consensus NJ tree for Kimura distances between *cpITS* sequences. Numbers above branches indicate the number of times the branch occurred out of 100 replicates using the Kimura distance; less frequently occurring branches are not shown. Numbers below branches indicate the number of times the branch occurred out of 100 replicates using the Jin and Nei (1990) distance with a gamma parameter of 2.0. Bootstrap values less than the consensus indicated are shown in parentheses. Abbreviations are as in the legend to figure 7. Higher taxon designations indicated are those of Ehrendorfer (1991, pp. 471–282) (spermatophytes) and Kramer (1990, pp. 49–52) (ferns). *B.* Portion of the 50% bootstrap proportion consensus NJ tree for Kimura distances between *cpITS* sequences showing the common branch for gymnosperms detected in 58/100 replicates.

Thus, *cpITS* provide no support for the view that Gnetales are the sister group of angiosperms, in contrast to reports based on *rbc*L sequences (see Discussion). We reanalyzed published *rbc*L data for the same or confamilial genera as for *cpITS*. Synonymous sites are saturated in most *rbc*L comparisons on this data set (see above).

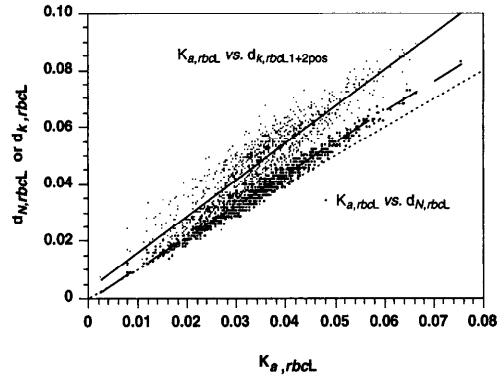

FIG. 9.—Divergence in *rbc*L sequences investigated estimated as d_k (Kimura two-parameter distance; Kimura 1980) at first and second codon positions of *rbc*L ($d_{k,rbcL1+2pos}$) or numbers of nonsynonymous substitutions per site estimated as d_N with the method of Nei and Gojobori (1986) (both ordinate) plotted against numbers of nonsynonymous substitutions per site estimated as K_a (abscissa) with the method of Li, Wu, and Luo (1985). Plots of K_a *vs.* d_N are indicated as heavy points, the corresponding linear regression ($R^2 = 0.97$) is indicated by the dashed line. Plots of K_a *vs.* d_k are indicated as light points, the corresponding linear regression ($R^2 = 0.86$) is indicated by the solid line. The dotted line indicates the expectation for identical estimates obtained with K_a and the other two methods.

Divergence between *rbc*L sequences should be estimated at synonymous and nonsynonymous sites independently (Martin, Somerville, and Loiseaux-deGoër 1992; Martin et al. 1993). But many groups currently using *rbc*L to study plant evolution use mainly the programs of PHYLIP or PAUP packages; to make our results more directly comparable to theirs, we removed third positions from the *rbc*L alignment and then estimated divergence at first and second positions with the Kimura method. Because about 75% of *rbc*L's third positions are synonymous in an average comparison, deleting third positions removes about 8%–10% of the nonsynonymous sites but also eliminates stochastic similarity from the data set in comparisons of divergent taxa. The few (about 5%) synonymous sites remaining at first positions should not distort distance estimations heavily. This distance estimation (d_k at first and second positions) neglects the effects of alternative pathways of amino acid substitution or likelihood of amino acid replacements but permits us to use a single substitution model for both individual and concatenated *cpITS* and *rbc*L sequences. Because only a small fraction of first positions in *rbc*L are synonymous, the correlation between the Kimura distance at first and second positions and K_a is quite positive ($R^2 = 0.86$) yet lower than the correlation between K_a estimated with Li et al.'s method and the same value estimated with Nei and Gojobori's method ($R^2 = 0.97$) (fig. 9). Bootstrap resampling should counteract this effect sufficiently so that first position synonymous

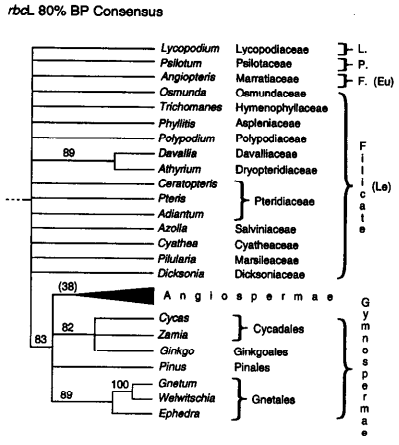

FIG. 10.—80% bootstrap proportion consensus NJ tree for Kimura distances at first and second codon positions for *rbc*L sequences (see text). *Marchantia* was used as the outgroup. Parentheses indicate that the consensus value of 38/100 for monophyly of angiosperms is below the threshold for other branches in the figure. Abbreviations are as in the legend to figure 7. Three differences in 80% consensus branches relative to figure 8 are mentioned in the text.

sites should not have serious impact on trees constructed with d_k at first and second *rbc*L positions.

The 80% bootstrap proportion NJ consensus tree for divergence at first and second *rbc*L codon positions is shown in figure 10. Within angiosperms, all groups found at the 80% bootstrap proportion consensus level for *cpITS* were also found for *rbc*L. Three additional branches within angiosperms were detected for *rbc*L in 80 or more replicates; these were *Bambusa–Oryza* (93/100), *Bambusa–Oryza–Pennisetum–Avena* (100/100), and *Piper–Annona* (91/100). With *rbc*L, angiosperms were detected as a monophyletic group in only 38/100 replicates. We found no support for the view that *rbc*L sequences suggest a sister group relationship between angiosperms and Gnetales. Because many recent reports using *rbc*L have included all codon positions, we constructed bootstrap consensus NJ trees for complete *rbc*L sequences from the same taxa. In those analyses, the group (*Zamia, Cycas, Ginkgo, Pinus*) was found in 99/100 replicates and was the sister group to angiosperms, the branch indicating a sister group relationship between these four gymnosperms and angiosperms was found in 86/100 replicates. Thus, also analysis of complete *rbc*L sequences did not support claims of sister group affinities between angiosperms and Gnetales.

Finally, we concatenated *cpITS* (complete) and *rbc*L sequences (first and second positions only) for the taxa indicated in table 1 and constructed the 80% consensus NJ tree (fig. 11). The result is based on an average of 1,486 nucleotides per comparison. Few changes in topology for the combined data set are evident relative to the *cpITS* topology in figure 8. The only differ-

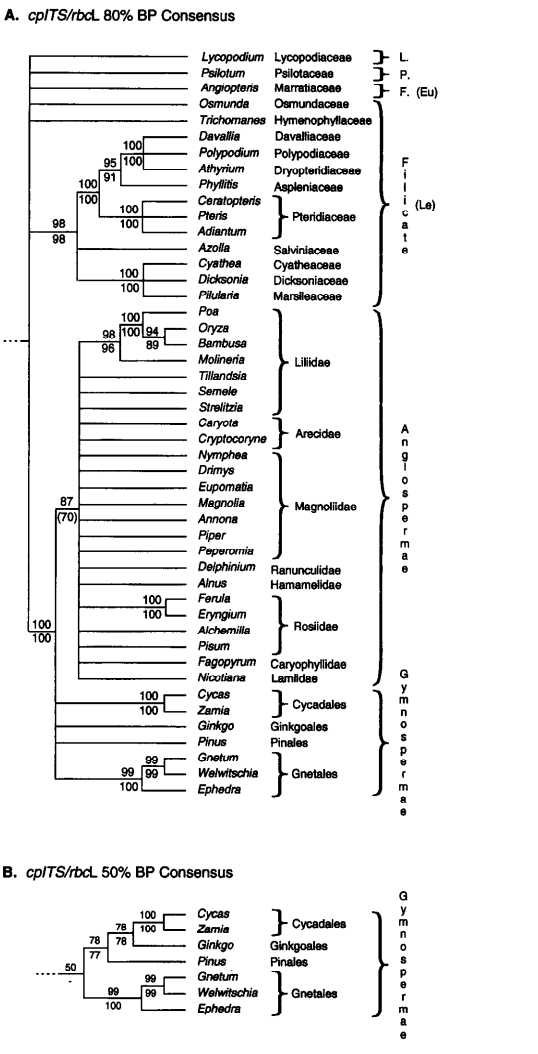

FIG. 11.—Trees derived from *cpITS* sequences concatenated with first and second positions of *rbc*L. Genus names refer to *cpITS* sequences, for genus names of concatenated *rbc*L sequences please refer to table 1. *A.* 80% consensus NJ tree for Kimura distances. Numbers above branches indicate bootstrap proportion using the Kimura distance. Numbers below branches indicate the bootstrap proportion using the Jin and Nei (1990) distance with a gamma parameter of 2.0. Bootstrap values less than the consensus indicated are shown in parentheses. Abbreviations are as in the legend to figure 7. *B.* Portion of the 50% bootstrap proportion consensus NJ tree showing the common branch for gymnosperms detected in 50/100 replicates.

ences in consensus topology are the separation of the two Piperales, *Piper* and *Peperomia,* and the lack of a common branch for ferns, *Psilotum* and *Lycopodium.* Relative to figure 10, however, quite a few differences are evident, most notably increased resolution within ferns and more robust branching within spermatophytes. Thus, the gene tree of the combined data set generally reflects the *cpITS* topology, which is not surprising because many more substitutions are observed between *cpITS* sequences than between first and second positions

of *rbc*L sequences. With the single exception of *Piper,* there are no conflicting branches for analyses of either marker alone or for the combined data set at the 80% consensus level.

Discussion

We previously addressed questions concerning the general course of angiosperm (Martin, Gierl, and Saedler 1989; Martin et al. 1993) and land plant evolution (Troitsky et al. 1991) with the help of relatively limited nucleotide sequence data sets. By employing PCR primers against conserved regions, data collection for study of plant evolution has become very simple. Recently, strong emphasis has been placed on *rbc*L sequencing (Les, Garvin, and Wimpee 1991; Chase et al. 1993; Clegg 1993), but other markers are needed that increase the amount of data available per taxon for evolutionary investigation. The conserved primers from the rRNA operon in the inverted repeat region of cpDNA used here efficiently amplify a roughly 500-bp fragment from land plants; we encountered no land plants from which we could not amplify this region. Sequence characteristics and divergence of *cpITS* are suitable for the study of land plant evolution.

Molecular Resolution within Angiosperms

The molecular phylogeny obtained with the combined *cpITS–rbc*L data set is probably more reliable than those obtained with either marker alone. The consensus tree in figure 11 contains several notable findings. Foremost, there is very strong evidence for the monophyly of angiosperms surveyed. The evidence for angiosperm monophyly, however, is not contained within the *rbc*L data set (cf. fig. 10) but rather in the *cpITS* data (fig. 8). The monophyly of angiosperms is also very strongly supported by analyses of their morphological characters (for a lucid review, see Crane, Friis, and Pedersen 1995). Also, the indel Δ519–550 (fig. 4) shared by angiosperms surveyed supports monophyly of flowering plants, as does the region around Δ567–571.

Within angiosperms, no resolution at the subclass level was obtained at the 80% consensus level with either *cpITS, rbc*L, or the combined data set; this finding is also reflected in the very short internal branch lengths within angiosperms in figure 7. With regard to the most primitive angiosperms sampled, we note that in nonbootstrapped NJ trees using either the Kimura or gamma distance for *cpITS* sequences, the aquatic angiosperm *Nymphaea* was basal on the flowering plant branch (fig. 7 and data not shown). Bootstrap support for this position was, however, very weak (47/100 with either Kimura or gamma [a = 2] distance), and the branch separating *Nymphaea* from other angiosperms was not found at all in either the *rbc*L or combined data sets. But consistent with the basal position of *Nymphaea,* and perhaps more noteworthy, is a small stretch of 7 bp (positions 507–513 in fig. 4) that appears to be shared between *Nymphaea* and gymnosperms (allowing for some substitutions) but is clearly absent from all other angiosperms surveyed. The alignment in this region can be modified, but even if portions of the alignment from positions 400–600 (or even positions 300–600) are excluded, *Nymphaea* retains its basal position among angiosperms and receives increased bootstrap support (data not shown). The specific indel under consideration is therefore consistent with—but independent of—substitutions in the remainder of the alignment.

The finding that an aquatic angiosperm is weakly supported by *cpITS* data to be the earliest branching flowering plant is compatible with current views on the nature of primitive angiosperms (Endress 1994) and with the findings of Les, Garvin, and Wimpee (1991) in their study of *rbc*L genes, although their taxon sampling was quite different from ours. They found that *Ceratophyllum* was the most primitive of several aquatic angiosperms surveyed, although the use of outgroups other than the one gymnosperm *Pseudotsuga* in that analysis may have produced different results. The phylogenetic distribution of Δ507–513 in other (aquatic) angiosperms (such as *Ceratophyllum*) deserves further attention. Also, more markers need to be employed in order to increase the total number of bases for analysis. If angiosperm evolution occurred as a true radiation, similar to the Cambrian explosion of invertebrate phyla (Hervé, Chenuil, and Adoutte 1994), resolution in the basal regions of the angiosperm tree may be a very difficult molecular phylogenetic problem, and—as for invertebrates—a very large number of sites may be required (Lecointre et al. 1994).

Relationship of Gnetales to Angiosperms and Other Gymnosperms

Answers to the question of angiosperm origins are inextricably coupled to the identification of their sister group among extinct and extant taxa. A number of lines of morphological evidence point to members of the Gnetales as the possible sister group to angiosperms among extant gymnoperms (Friedmann 1990, 1994; Nixon et al. 1994; Crane, Friis, and Pedersen 1995), but molecular support for this view is extremely weak at best. Albert et al. (1994) and Doyle, Donoghue, and Zimmer (1994) conducted parsimony analyses of molecular sequences combined with morphological characters and concluded that Gnetales are the sister group of angiosperms, but if molecular data are combined with character state data, the result cannot be regarded as an independent molecular test of hypotheses concerning morphological evolution. The power of molecular data to

reconstruct evolution independently of parallelisms at the morphological level is lost if the two types of data are combined. Therefore, the conclusions of such analyses cannot be taken as molecular support *sensu strictu* for sister group status between angiosperms and Gnetales. In Doyle, Donoghue, and Zimmer (1994), trees based purely on molecular (rRNA) data are also shown, but these do not include nonspermatophyte outgroups, in the absence of which sister group relationships between Gnetales and angiosperms cannot be addressed because outgroups may have dissected the angiosperm–gymnosperm branch. Hamby and Zimmer (1992) did include *Equisetum* and *Psilotum* as outgroups in some trees and found that the data did not permit resolution of the angiosperm–Gnetales relationship. Other studies of rRNA (Rakhimova et al. 1989; Troitsky et al. 1991; Chaw et al. 1994) and *rbc*L sequences (Hasebe et al. 1992, 1993) that included outgroups suggested that no extant gymnosperm taxon is a sister taxon to angiosperms and that gymnosperms may be a monophyletic group. The latter findings are consistent with the results of our analyses on *cpITS* and *rbc*L sequences, although we only have very weak bootstrap support for the monophyly of gymnosperms sampled. We find, however, very strong support for the monophyly of Gnetales with both markers (figs. 8, 10, and 11), which is incongruent with results of parsimony analyses on morphological characters recently presented by Nixon et al. (1994), in which *Ephedra* branched below angiosperms and other Gnetales.

Phylogenetic Analysis within Ferns and Fern Allies

In the analyses of *cpITS* sequences from 16 pteridophytes (including representatives from the fern allies *Lycopodium* and *Psilotum,* as well as one eusporangiate and 13 leptosporangiate ferns), the phylogeny appears to yield better resolution than within spermatophytes, probably due to the less star-like topology of the pteridophyte tree. Resolution was considerably better with *cpITS* (figs. 7 and 8) or concatenated (fig. 11) sequences than with *rbc*L sequences alone (fig. 10). Only one internal branch was found in the 80% consensus *rbc*L tree within ferns (suggesting a close affinity between Davalliaceae and Dryopteridaceae to the exclusion of *Polypodium*). Notably, the degree of internal branch support that we found for *rbc*L was much lower than that reported by Hasebe et al. (1994), in which all positions of *rbc*L were considered. Within the fern *rbc*L sequences sampled, average divergence at synonymous sites across 101 comparisons was >1.0, suggesting that these are saturated, or nearly so, in most comparisons (by contrast, average divergence between *cpITS* sequences of ferns is 0.35). We did not sample as many taxa as Hasebe et al. (1994) did, but we could not corroborate th high bootstrap values they reported in the fern *rbc*L tree Also, we found a major discrepancy between our topologies and those of Hasebe et al. in that the commo branch shared by representatives of two families of taxonomically highly uncertain heterosporous ferns (Marsileaceae and Salviniaceae, 100/100 replicates in Haseb et al. 1994) was found in neither *rbc*L nor *cpITS* analyses. Rather, we found a very close affinity betwee Marsileaceae and representatives of tree ferns (Dickson iaceae and Cyathaceae) to the exclusion of Salviniacea (although *Azolla* possess a very large deletion encompassing the entire *cpITS*2 region). Otherwise, the topology within leptosporangiate ferns with *cpITS* sequence was largely congruent with that of Hasebe et al. (1994) including the basal position of Hymenophyllaceae, Maratiaceae, and Osmundaceae. Deeper branches withi ferns in figure 7 find low bootstrap support (figs. 8, 10 and 11). The position of *Lycopodium* in figure 7 is compatible with data from cpDNA gene rearrangement (Raubeson and Janson 1992). The inclusion of additional OTUs and outgroups might be expected to have influence on the common branches shared by *Psilotum* an *Angiopteris,* and the two primitive leptosporangiat ferns *Osmunda* and *Trichomanes,* respectively.

Conclusions

Substitutions occur in the noncoding sequences o *cpITS* regions in the inverted repeat at about one-sixt the rate of that found for synonymous sites in *rbc*L Despite this lower substitution rate, average divergenc between 16 pteridophytes and 31 spermatophytes i about 0.8 substitutions per site in the noncoding *cpIT* regions. This value is quite high but still can be esti mated with some degree of reliability (the average stan dard error across these comparisons is about 0.2). Be cause synonymous sites in *rbc*L evolve about six time faster, they are saturated in comparisons across the sper matophye–pteridophyte boundary and in most compar isons within pteridophytes, where average divergenc between noncoding regions of *cpITS* is 0.35 substitu tions per site. Within spermatophytes, *cpITS* seems t be a very useful marker even though it is quite short. I can be used to increase the number of sites available fo comparison in studies of higher plant evolution, an alignments reveal a number of indels with conspicuou phylogenetic distribution. Our phylogenetic analyse marshalled no support for the "anthophyte concept," i.e., for the view that Gnetales and angiosperms are sis ter groups and may be collectively designated antho phytes by virtue of the flower-like gnetalean reproduc tive structures (reviewed in Crane, Friis, and Pederse 1995). On the contrary, both *cpITS* and *rbc*L data sug

gest with low bootstrap support that gymnosperms surveyed (conifers, cycads, gnetales, *Ginkgo*) may constitute a monophyletic group. Previous reports on the basis of *rbc*L sequence data that gnetales may be the sister group of angiosperms entailed analyses of all *rbc*L sites and may have contained a high number of stochastically similar nucleotides. Careful analyses of further molecular data are needed before conclusions about the general course of higher plant evolution can be drawn.

Acknowledgments

This work was supported by grant Ma 1426/1-3 from the Deutsche Forschungsgemeinschaft to W.M. and by grant N 93-04-6962 from the Russian Fund of Fundamental Science to A.A. V.G. gratefully acknowledges stipends from the DAAD and DFG. We thank H. Saedler for general support and the Gesellschaft für Biotechnologische Forschung, Braunschweig, for the generous use of their computer facilities. We thank S. Bünte and C. Köhler for excellent technical assistance and Dr. B. Zimmer and K. Baeske for fern material.

LITERATURE CITED

ALBERT, V. A., A. BACKLUND, K. BREMER, M. W. CHASE, J. R. MANHARDT, 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.

ANSORGE, W., B. S. SPROAT, J. STEGEMANN, and C. SCHWAGER. 1986. A non-radioactive automated method for DNA sequence determination. J. Biochem. Biophys. Methods **13**: 315–323.

BAUM, D. 1994. *rbc*L and seed plant phylogeny. Trends Ecol. Evol. **9**:39–41.

BOBROVA, V. K., A. V. TROITSKY, A. G. PONOMAREV, and A. S. ANTONOV. 1987. Low-molecular-weight rRNA sequences and plant phylogeny reconstruction: nucleotide sequences of chloroplast 4.5S rRNAs from *Acorus calamus* (Araceae) and *Ligularia calthifolia* (Asteraceae). Plant Syst. Evol. **156**:13–27.

BÖHLE, U.-R., H. H. HILGER, R. CERFF, and W. MARTIN. 1994. Non-coding chloroplast DNA for plant molecular systematics at the infrageneric level. Pp. 391–403 *in* B. SCHIERWATER, B. STREIT, G. WAGNER, and R. DESALLE, eds. Molecular ecology and evolution: approaches and applications. Birkhäuser, Basel.

BOULTER, D., J. A. M. RAMSHAW, E. W. THOMPSON, M. RICHARDSON, and R. H. BROWN. 1972. A phylogeny of higher plants based on the amino acid sequences of cytochrome *c* and its biological implications. Proc. R. Soc. Lond. B **181**: 441–455.

CHASE, M. W., D. E. SOLTIS, R. G. OLMSTEAD et al. (42 coauthors). 1993. DNA sequence phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbc*L. Ann. Mo. Bot. Gard. **80**:528–580.

CHAW, S.-M., H.-M. SUNG, H. LONG, A. ZHARKIKH, and W.-H. LI. 1994. Phylogeny of the major subclasses of angiosperms and date of the monocot–dicot divergence. Am. J. Bot. **81**: S146.

CLEGG, M. T. 1993. Chloroplast gene sequences and the study of plant evolution. Proc. Natl. Acad. Sci. USA **90**:363–367.

CLEGG, M. T., and G. ZURAWSKI. 1992. Chloroplast DNA and the study of plant phylogeny. Pp. 1–13 *in* P. S. SOLTIS, J. J. DOYLE, and D. E. SOLTIS, eds. Molecular systematics of plants. Chapman & Hall, New York.

CRANE, P. W., E. M. FRIIS, and K. R. PEDERSEN. 1995. The origin and early diversification of angiosperms. Nature **374**: 27–33.

DEVEREUX, J., P. HAEBERLI, and O. SMITHIES. 1984. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. **12**:387–395.

DOWNIE, S. R., and J. D. PALMER. 1992. Use of chloroplast DNA rearrangements in reconstructing phylogeny. Pp. 14–35 *in* P. S. SOLTIS, J. J. DOYLE, and D. E. SOLTIS, eds. Molecular systematics of plants. Chapman & Hall, New York.

DOYLE, J. A., M. J. DONOGHUE, and E. A. ZIMMER. 1994. Integration of morphological and ribosomal RNA data on the origin of the angiosperms. Ann. Mo. Bot. Gard. **81**:419–450.

EHRENDORFER, F. 1991. Evolution und Systematik. Pp. 666–826 *in* P. SITTE, H. ZIEGLER, F. EHERENDORFER, and A. BREZINSKY, eds. Lehrbuch der Botanik. Gustav Fischer Verlag, Stuttgart.

ENDRESS, P. 1994. Floral structure and evolution of primitive angiosperms: recent advances. Pl Syst. Evol. **192**:79–97.

FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: a maximum-likelihood appoach. J. Mol. Evol. **17**:368–376.

———. 1989. PHYLIP—phylogeny inference package (version 3.2). Cladistics **5**:164–166.

FRIEDMANN, W. 1990. Double fertilization in *Ephedra,* a nonflowering seed plant: its bearing on the origin of angiosperms. Science **247**:951–954.

———. 1994. The evolution of embryogeny in seed plants and the developmental origin and early history of endosperm. Am. J. Bot. **81**:1468–1486.

GAUT, B. S., S. V. MUSE, W. D. CLARK, and M. T. CLEGG. 1992. Relative rates of nucleotide substitution at the *rbc*L locus of monocotyledonous plants. J. Mol. Evol. **35**:292–303.

HAMBY, R. K., and E. A. ZIMMER. 1992. Ribosomal RNA as a phylogenetic tool in plant systematics. Pp. 50–91 *in* P. S. SOLTIS, J. J. DOYLE, and D. E. SOLTIS, eds. Molecular systematics of plants. Chapman & Hall, New York.

HASEBE, M., M. ITO, R. KOFUJI, K. UEDA, and K. IWATSUKI. 1993. Phylogenetic relationships of ferns deduced from *rbc*L gene sequence. J. Mol. Evol. **37**:476–482.

HASEBE, M., R. KOFUJI, M. ITO, M. KATO, K. IWATSUKI, and K. UEDA. 1992. Phylogeny of the gymnosperms inferred from *rbc*L gene sequences. Bot. Mag. Tokyo **105**:673–679.

HASEBE, M., T. OMORI, M. NAKAZAWA, T. SANO, M. KATO, and K. IWATSUKI. 1994. *rbc*L gene sequences provide evidence for the evolutionary lineages of leptosporangiate ferns. Proc. Natl. Acad. Sci. USA **91**:5730–5734.

HERVÉ, P., A. CHENUIL, and A. ADOUTTE. 1994. Can the Cambrian explosion be inferred through molecular phylogeny? Development (Supplement) 15–25.

HORI, H., B.-L. LIM, and S. OSAWA. 1985. Evolution of green plants as deduced from 5S rRNA sequences. Proc. Natl. Acad. Sci. USA **82**:820–823.

JIN, L., and M. NEI. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. **7**: 82–102.

KIMURA, M. 1980. A simple method for estimating evolutionary rate of base subsitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

KRAMER, K. U. 1990. Notes on the higher level classsification of the recent ferns. Pp. 49–52 *in* K. U. KRAMER and P. S. GREEN, eds. The families and genera of vascular plants. Vol. **I**. Pteridophytes and gymnosperms. Springer Verlag, Berlin.

LECOINTRE, G., P. HERVÉ, H. L. V. LE, and H. LE GUYADER. 1994. How many nucleotides are required to resolve a phylogenetic problem? The use of a new statistical method applicable to available sequences. Mol. Phylogenet. Evol. **3**: 292–309.

LES, D. H., D. K. GARVIN, and C. F. WIMPEE. 1991. Molecular evolutionary history of ancient aquatic angiosperms. Proc. Natl. Acad. Sci. USA **88**:10119–10123.

LI, W.-H., C.-I. WU, and C.-C. LUO. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2**:150–174.

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

MARTIN, P. G., and A. C. JENNINGS. 1983. The study of plant phylogeny using amino acid sequences of ribulose-1,5-bisphosphate carboxylase. Aust. J. Bot. **31**:395–409.

MARTIN, W., A. GIERL, and H. SAEDLER. 1989. Molecular evidence for pre-Cretaceous angiosperm origins. Nature **339**:46–48.

MARTIN, W., D. LYDIATE, H. BRINKMANN, G. FORKMANN, H. SAEDLER, and R. CERFF. 1993. Molecular phylogenies in angiosperm evolution. Mol. Biol. Evol. **10**:140–162.

MARTIN, W., C. C. SOMERVILLE, and S. LOISEAUX-DE GOËR. 1992. Molecular phylogenies of plastid origins and algal evolution. J. Mol. Evol. **35**:385–403.

MURRAY, M. G., and W. F. THOMPSON. 1980. Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res. **8**: 4321–4325.

NEI, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

NEI, M., and T. GOJOBORI. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous substitutions. Mol. Biol. Evol. **3**:418–426.

NIESBACH-KLÖSGEN, U., E. BARZEN, J. BERNHARDT, W. ROHDE, ZS. SCHWARZ-SOMMER, H.-J. REIF, U. WIENAND, and H. SAEDLER. 1987. Chalcone synthase genes in plants: a tool to study evolutionary relationships. J. Mol. Evol. **26**:213–225.

NIXON, K. C., W. L. CREPET, D. STEVENSON, and E. M. FRIIS. 1994. A reevaluation of seed plant phylogeny. Ann. Mo. Bot. Gard. **81**:484–533.

OTA, T., and M. NEI. 1994. Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. J. Mol. Evol. **38**:642–643.

PALMER, J. D. 1985. Comparative organization of chloroplast genomes. Ann. Rev Genet. **19**:325–354.

PALMER, J. D., and W. F. THOMPSON. 1981. Rearrangements in the chloroplast genomes of mung bean and pea. Proc. Natl. Acad. Sci. USA **78**:5533–5537.

PALMER, J. D., R. K. JANSEN, H. J. MICHEALS, M. W. CHASE, and J. R. MANHART. 1988. Chloroplast DNA variation and plant phylogeny. Ann. Mo. Bot. Gard. **75**:1180–1206.

PICHI-SERMOLLI, R. E. G. 1958. The higher taxa of Pteridophyta and their classification. Syst. Today (Uppsala Universitets Aarsskrift) **6**:70–90.

RAKHIMOVA, G. M., A. V. TROITSKY, I. N. KLIKUNOVA, and A. S. ANTONOV. 1989. Phylogenetic analysis of partial nucleotide sequences of 18S rRNA of 14 plant species. Mol. Biol. (Moscow) **23**:830–842.

RAUBESON, L. A., and R. K. JANSON. 1992. Chloroplast DNA evidence on the acient evolutionary split in vascular land plants. Science **255**:1697–1699.

SAITOU, N., and M. NEI 1987. The neighbor-joining method: a new method for the reconstruction of phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

SAMBROOK, J., E. F. FRITSCH, and T. MANIATIS. 1989. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

TABOR, S., and C. C. RICHARDSON. 1987. DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. Proc. Natl. Acad. Sci. USA **84**:4767–4771.

TROITSKY, A. V., and V. K. BOBROVA. 1986. 23S-derived small ribosomal RNAs: their structure and evolution with regard to plant phylogenies. Pp. 137–170 *in* K. S. DUTTA, ed. DNA systematics. Vol. **II**. CRC Press, Boca Raton.

TROITSKY, A. V., Y. F. MELEKHOVETS, G. M. RAKHIMOVA, V. K. BOBROVA, K. M. VALIEJO-ROMAN, and A. S. ANTONOV. 1991. Angiosperm origin and early seed plant evolution deduced from rRNA sequence comparisons. J. Mol. Evol. **32**: 253–261.

VAN DE PEER, Y., and R. DE WACHTER. 1993. TREECON: a software package for the construction and drawing of evolutionary trees. Comput. Appl. Biosci. **9**:177–182.

WOLFE, K. H., M. GOUY, Y. W. YANG, P. SHARP, and W.-H. LI. 1989. Date of the monocot–dicot divergence estimated from chloroplast DNA sequence data. Proc. Natl. Acad. Sci. USA **86**:6201–6205.

WOLFE, K. H., W.-H. LI, and P. SHARP. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. USA **84**: 9054–9058.

WOLFE, K. H., C. W. MORDEN, and J. D. PALMER. 1992. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. Proc. Natl. Acad. Sci. USA **89**: 10648–10652.

ZIMMER, E. A., R. K. HAMBY, M. L. ARNOLD, D. A. LEBLANC, and E. L. THERIOT. 1989. Ribosomal RNA phylogenies and flowering plant evolution. Pp. 205–226 *in* B. FERNHOLM, K. BREMER, and H. JÖRNVALL, eds. The hierarchy of life. Elsevier, Amsterdam.

TAKASHI GOJOBORI, reviewing editor

Accepted October 23, 1995