The Phylogenetic Potential of Entire 26S rDNA Sequences in Plants

Robert K. Kuzoff, Jennifer A. Sweere,† Douglas E. Soltis,* Pamela S. Soltis,* and Elizabeth A. Zimmer†*

*Department of Botany, Washington State University; and †Laboratory of Molecular Systematics, National Museum of Natural History, Smithsonian Institution

18S ribosomal RNA genes are the most widely used nuclear sequences for phylogeny reconstruction at higher taxonomic levels in plants. However, due to a conservative rate of evolution, 18S rDNA alone sometimes provides too few phylogenetically informative characters to resolve relationships adequately. Previous studies using partial sequences have suggested the potential of 26S or large-subunit (LSU) rDNA for phylogeny retrieval at taxonomic levels comparable to those investigated with 18S rDNA. Here we explore the patterns of molecular evolution of entire 26S rDNA sequences and their impact on phylogeny retrieval. We present a protocol for PCR amplification and sequencing of entire (~3.4 kb) 26S rDNA sequences as single amplicons, as well as primers that can be used for amplification and sequencing. These primers proved useful in angiosperms and Gnetales and likely have broader applicability. With these protocols and primers, entire 26S rDNA sequences were generated for a diverse array of 15 seed plants, including basal eudicots, monocots, and higher eudicots, plus two representatives of Gnetales. Comparisons of sequence dissimilarity indicate that expansion segments (or divergence domains) evolve 6.4 to 10.2 times as fast as conserved core regions of 26S rDNA sequences in plants. Additional comparisons indicate that 26S rDNA evolves 1.6 to 2.2 times as fast as and provides 3.3 times as many phylogenetically informative characters as 18S rDNA; compared to the chloroplast gene *rbcL*, 26S rDNA evolves at 0.44 to 1.0 times its rate and provides 2.0 times as many phylogenetically informative characters. Expansion segment sequences analyzed here evolve 1.2 to 3.0 times faster than *rbcL*, providing 1.5 times the number of informative characters. Plant expansion segments have a pattern of evolution distinct from that found in animals, exhibiting less cryptic sequence simplicity, a lower frequency of insertion and deletion, and greater phylogenetic potential.

Introduction

Since the advent of comparative DNA sequencing in plants, the chloroplast gene *rbcL* has been the primary molecular marker used for phylogenetic inference at higher taxonomic levels. Its utility in these taxonomic levels has been well established (e.g., Chase et al. 1993; reviewed in Baum 1994; Soltis and Soltis 1998). Nonetheless, phylogenetic hypotheses based on a single gene or character may not represent true organismal relationships. Molecular systematists therefore seek additional genes for phylogeny reconstruction to test *rbcL*-based topologies, to obtain additional resolution, and to elucidate relationships at a variety of taxonomic levels. A majority of the genes or DNA regions that have been proposed as alternatives or supplements to *rbcL* (e.g., *ndhF, atp*β, *matK, trnL* intron and spacer regions, *trnK* intron and spacer regions, *rps2*, and *rps4*) come from the chloroplast genome (cpDNA; reviewed in Soltis and Soltis 1998).

The need to compare higher-level cpDNA-based topologies with phylogenetic hypotheses derived from nuclear sequences has been stressed by several authors (e.g., Doyle 1992; Chase et al. 1993; Nickrent and Soltis 1995). The most widely sequenced nuclear gene for higher-level phylogenetic inference in plants is 18S rDNA. Comparative sequencing of 18S rDNA or rRNA has been used in algae, bryophytes, ferns, fern allies, gymnosperms, and angiosperms (reviewed in Hamby and Zimmer 1992; Mishler et al. 1994; Soltis and Soltis 1998). In several instances, 18S rDNA sequence data have provided a critical test of *rbcL*-based topologies (e.g., Kron 1996; Soltis et al. 1997; Soltis and Soltis 1997). However, *rbcL* yields 1.4 times as many informative bases as 18S rDNA for the same suite of angiosperms (Nickrent and Soltis 1995). Consequently, although phylogenetic analysis of 18S rDNA sequences provides a critically needed independent data set for the assessment of higher-level relationships, in many instances analysis of these sequences alone will not provide adequate resolution (Soltis et al. 1997).

Optimally, a nuclear gene or DNA region used for phylogenetic inference would contain a large number of phylogenetically informative characters, be single-copy and present in all plants, and be easy to amplify, sequence, and align with other sequences. 26S rDNA sequences offer the potential of satisfying most, if not all, of these criteria. The lengths of reported entire 26S rDNA sequences in plants are just under 3.4 kb and range from 3,375 to 3,393 bp (Sugiura et al. 1985; Kiss, Kiss, and Solymosy 1989; Kolosha and Fodor 1990; Okumura and Shimada 1992). Based on partial 18S and 26S rRNA sequences for the same suite of land plants, Hamby and Zimmer (1992) suggested that 26S rDNA has a slightly higher rate of base substitution than does 18S rDNA. In addition, 26S rDNA can be analyzed in a manner similar to that used to analyze single-copy genes, due to concerted evolution of the sequences (Zimmer et al. 1980; Arnheim 1983).

The phylogenetic utility of entire 26S rDNA sequences in plants had been questioned because of the inferred structure and evolution of 28S rDNA sequences obtained from animals (Clark et al. 1984; Hassouna, Michot, and Bachellerie 1984; Tautz, Trick, and Dover 1986; Hancock and Dover 1988, 1990). LSU rDNA se-

Key words: 26S rDNA, LSU rDNA, expansion segments, cryptic sequence simplicity, molecular evolution, phylogenetic inference.

Address for correspondence and reprints: Robert K. Kuzoff, Department of Botany, Washington State University, Pullman, Washington 99164-4238. E-mail: kuzoff@wsunix.wsu.edu.

Mol. Biol. Evol. 15(3):251–263. 1998

Table 1
26S rDNA Sequences Analyzed

Taxon	GenBank Accession No.[a]	Length (bp)	G+C Content[b]		
			Entire 26S	Conserved Core	Expansion Segments
Fragaria × *ananassa*	Simovic/X58118	3,377	0.56245	0.51893	0.66599
Arbidopsis thaliana	Unifried and Gruendler (1990)/X52320	3,375	0.55788	0.51959	0.64824
Brassica napus	Okumura and Shimada (1992)/D10840	3,378	0.54453	0.51062	0.62475
Sinapis alba	Capesius (1991)/X57137	3,381	0.55240	0.51807	0.63327
Citrus limon	Kolosha and Fodor (1990)/X05910	3,393	0.58095	0.52846	0.70378
Lycopersicon esculentum	Kiss, Kiss, and Solymosyd (1989)/X13557	3,381	0.56674	0.51971	0.67879
Oryza sativa	Sugiura et al. (1985)/M11585	3,377	0.59336	0.53061	0.74243
Drimys winteri	AFO36491	3,378	0.55269	0.51421	0.64153
Parnassia fimbriata Banks	AFO36496	3,377	0.54131	0.51210	0.60862
Hamamelis virginiana	AFO36495	3,366	0.57397	0.52569	0.68644
Jepsonia parryi	AFO36497	3,366	0.56506	0.52187	0.66568
Lithophragma trifoliatum Eastw.	AFO36501	3,367	0.55908	0.52078	0.64852
Mitella pentandra Hook.	AFO36502	3,353	0.56061	0.52269	0.64916
Tellima grandiflora (Pursch) Dougl.	AFO36500	3,368	0.55819	0.52143	0.64392
Peltoboykinia tellimoides (Maxim.) Hara	AFO36499	3,365	0.55483	0.52315	0.62872
Saxifraga mertensiana Bong.	AFO36498	3,365	0.54800	0.51783	0.61843
Eucryphia lucida Druce	AFO36494	3,363	0.57598	0.52696	0.69048
Tragopogon dubius Scop.	AFO36493	3,364	0.56375	0.52185	0.66170
Plumbago auriculata Lam	AFO36492	3,353	0.57424	0.52356	0.69370
Gnetum gnemon	AFO36488	3,410	0.55009	0.50933	0.64083
Acorus graminae	AFO36490	3,213[c]	0.58045	0.52985	0.70632
Ephedra distachya	AFO36489	3,380	0.55148	0.50574	0.65598

Note.—Sequence lengths and G+C contents for entire 26S sequences, conserved core regions and expansion segment regions are listed (data set 1; see table 4 for family designations).

[a] References (where available) and GenBank accession numbers are given for previously published sequences. GenBank accession numbers are given for new sequences provided by this study.

[b] Average G+C content: entire 26S sequences = 0.56212; conserved core = 0.52009; expansion segments = 0.66047. Exact locations of expansion segments are provided in tables 3 and 4.

[c] The *Acorus graminae* 26S rDNA sequence is incomplete due to primer divergence.

quences are composed of conserved core regions, which are alignable across kingdoms, and divergence domains (Hassouna, Michot, and Bachellerie 1984), or expansion segments (Clark et al. 1984), which evolve more rapidly and are the loci of the majority of length mutations for this gene, presumably due to reduced functional constraints (Clark et al. 1984; Dover and Flavell 1984; Hassouna, Michot, and Bachellerie 1984; Appels and Honeycutt 1986; Flavell 1986). Expansion segments are dispersed throughout 28S rDNA in animals and exhibit (1) a higher rate of base substitution than do the conserved core areas; (2) biased base composition, which is manifested in the form of greater G + C content; (3) frequent insertion-deletion mutations; and (4) character nonindependence due to presumed compensatory mutations and sequence coevolution among remote domains (Tautz, Trick, and Dover 1986; Hancock and Dover 1988, 1990; Larson 1991; Dixon and Hillis 1993). Despite these concerns, expansion segments of 28S rDNA have been employed for phylogeny retrieval in Chytridiomycetes (Auwera and deWachter 1996), Foraminifera (Holzmann, Piller, and Pawlowski 1996), salamanders (Larson and Wilson 1989), and tiger beetles (Vogler, Welsch, and Hancock 1997).

Since Hamby and Zimmer (1988) investigated the phylogenetic potential of partial 18S and 26S sequences in plants to infer angiosperm phylogeny, partial 26S rRNA or rDNA sequences have been included in phylogenetic analyses of a diverse array of green plants (e.g., Buchheim and Chapman 1991; Bult and Zimmer 1993; Mishler et al. 1994; Ro, Keener, and McPheron 1997; Stefanovic et al. 1998). Despite the phylogenetic potential suggested by these studies, entire sequences of 26S rDNA have been generated for only seven plant species (Sugiura et al. 1985; Kiss, Kiss, and Solymosy 1989; Kolosha and Fodor 1990; Unfried and Gruendler 1990; Okumura and Shimada 1992). Each of these previously published entire 26S rDNA sequences was generated through laborious cloning techniques. Given this small sample of complete sequences, it is unclear whether LSU rDNA evolves similarly in plants and animals. The extent to which 26S rDNA sequences will be useful for phylogeny reconstruction in plants also remains unresolved. A larger sample size of sequences is needed to address these issues. In this paper we (1) provide a simple protocol for amplification of nearly entire plant 26S rDNA sequence as a single unit; (2) provide primers for the amplification and sequencing of the gene; (3) compare relative levels of base substitution for 26S rDNA, 18S rDNA, and *rbcL*; and (4) provide initial insights into the molecular evolution of 26S rDNA and the impact of these evolutionary patterns on phylogeny retrieval.

Materials and Methods

Amplification

26S rDNA sequences were amplified via PCR from total DNA extracts from 15 species (table 1) with the

Table 2
Primers Used for PCR Amplification and Sequencing of 26S rDNA

Primer	Direction	5′ to 3′ Sequence	Position in *Oryza*	Designers
N-nc26S1	Forward	CGACCCCAGGTCAGGCG	4–21	C. Bult and E. Zimmer
N-nc26S2	Forward	GAGTCGGGTTGTTTGGGA	266–283	C. Bult and E. Zimmer
N-nc26S3	Forward	AGGGAAGCGGATGGGGGC	417–434	C. Bult and E. Zimmer
N-nc26S4	Forward	TTGAAACACGGACCAA	645–662	C. Bult and E. Zimmer
N-nc26S5	Forward	CGTGCAAATCGTTCGTCT	877–894	C. Bult and E. Zimmer
N-nc26S6	Forward	TGGTAAGCAGAACTGGCG	1,125–1,142	C. Bult and E. Zimmer
N-nc26S7	Forward	GATGAGTAGGAGGGCGCG	1,360–1,377	C. Bult and E. Zimmer
N-nc26S8	Forward	ACGTTAGGAAGTCCGGAG	1,629–1,646	C. Bult and E. Zimmer
N-nc26S9	Forward	AATGTAGGCAAGGGAAGT	1,879–1,896	C. Bult and E. Zimmer
N-nc26S10	Forward	TAAAACAAAGCATTGCGA	2,130–2,147	C. Bult and E. Zimmer
N-nc26S11	Forward	AATCAGCGGGGAAAGAAG	2,372–2,389	C. Bult and E. Zimmer
N-nc26S12	Forward	GTCCTAAGATGAGCTCAA	2,642–2,659	C. Bult and E. Zimmer
N-nc26S13	Forward	CCTATCATTGTGAAGCAG	2,875–2,892	C. Bult and E. Zimmer
N-nc26S14	Forward	TTATGACTGAACGCCTCT	3,094–3,111	C. Bult and E. Zimmer
N-nc26S15	Forward	TGCCACGATCCACTGAGA	3,333–3,350	C. Bult and E. Zimmer
268rev	Reverse	GCATTCCCAAACAACCCGAC	Compl. 268–287	D. Nickrent and D. Soltis
641rev	Reverse	TTGGTCCGTGTTTCAAGACG	Compl. 641–660	D. Nickrent and D. Soltis
950rev	Reverse	GCTATCCTGAGGGAAACTTC	Compl. 950–969	D. Nickrent and D. Soltis
1229rev	Reverse	ACTTCCATGACCACCGTCCT	Compl. 1,229–1,248	D. Nickrent and D. Soltis
1499rev	Reverse	ACCCATGTGCAAGTGCCGTT	Compl. 1,499–1,518	D. Nickrent and D. Soltis
1839rev	Reverse	TTCACCTTGGAGACCTGATG	Compl. 1,839–1,858	D. Nickrent and D. Soltis
2134rev	Reverse	GGACCATCGCAATGCTTTGT	Compl. 2,134–2,153	D. Nickrent and D. Soltis
2426rev	Reverse	MCTACACCTCTCAAGTCAT	Compl. 2,426–2,444	D. Nickrent and D. Soltis
2782rev	Reverse	GGTAACTTTTCTGACACCTC	Compl. 2,782–2,801	D. Nickrent and D. Soltis
3058rev	Reverse	TTCGCGCCACTGGCTTTTCA	Compl. 3,058–3,077	D. Nickrent and D. Soltis
3331rev	Reverse	ATCTCAGTGGATCGTGGCAG	Compl. 3,331–3,350	D. Nickrent and D. Soltis

NOTE.—Primers designed by C. Bult and E. Zimmer in 1993 and by D. Nickrent and D. Soltis in 1993 have previously not been published. Compl. = sequence complementary to positions indicated.

forward primer N-nc26S1 and the reverse primer 3331R (table 2). Amplification reactions followed the general methodology of Bult, Kallersjo, and Suh (1992) with the following minor modifications. PCR reactions contained the following: 10 μl 10 × Stratagene *Taq* Extender Additive buffer (supplied with Stratagene *Taq* Extender Additive); 16 μl of 10 mM dNTPs; 5 μl of 10 μM N-nc26S1 (forward primer); 5 μl of 10 μM 3331rev (reverse primer); 5 μl dimethylsulfoxide (DMSO); 0.7 μl Promega *Taq* polymerase; 0.7 μl Stratagene *Taq* Extender Additive; 10 μl diluted total DNA extract (concentration from 10 to 20 ng DNA/μl TE); and 47.6 μl deionized water (to a total volume of 100 μl); the PCR reactions were covered with three or four drops of mineral oil. PCR amplifications were carried out in a Perkin-Elmer 480 thermocycler as follows: (1) a hot start at 94°C for 3 min; (2) 30 amplification cycles of 94°C for 1 min, 55°C for 1 min, 72°C for 3.5 min; (3) a terminal extension phase at 72°C for 5 min; and (4) an indefinite terminal hold at 4°C.

The double-stranded (ds) PCR products were subsequently purified via precipitation with 90 μl of 20% PEG 8000/2.5 M NaCl at 37°C for 15 min (Soltis and Soltis 1997). Precipitated ds products were centrifuged for 15 min at 14,000 rpm and 4°C. The pellets were washed with 200 μl 80% ethanol (prechilled to 4°C) and centrifuged for 7 min and then washed with chilled 95% ethanol and centrifuged for 7 min at 4°C. The ethanol was decanted, and the pellets were dried in a Sorvall Speed-Vac for 20 min using low heat. Products were resuspended in 25 μl dH_2O at 37°C for approximately 25 min. One microliter of each sample was electrophoresed in a 0.7% agarose mini-gel for quantification.

DNA Sequencing

Nearly complete sequences were generated for 14 of the 15 amplified 26S rDNA regions (terminal sequencing primers were internal, located at the 5′ and 3′ ends of 26S rDNA). For *Acorus graminae*, 164 bp was not obtained, apparently due to sequence divergence in the primer-binding sites for N-nc26S1 and 3331rev. Automated sequencing of the 26S rDNA was conducted on an ABI 373A automated sequencer following the general protocol described by Soltis and Soltis (1997) for 18S rDNA. Cycle sequencing reactions contained the following: 50–60 ng purified ds PCR product (template); 4.8 μl PRISM Ready Reaction Dye Deoxy Terminator Cycle enzyme, dNTP, and buffer mixture (Applied Biosystems, Foster City, Calif.); 0.5 μl DMSO; 0.5 μl sequencing primer (1.6 mM); and dH_2O to a total volume of 10 μl. Reactions were conducted as follows: (1) a warm start at 96°C for 3 min; (2) 25 amplification cycles of 96°C for 30 s, 50°C for 15 s, 60°C for 4 min; and (3) an indefinite hold at 4°C (for additional details, see Soltis and Soltis 1997). Sequencing primers for 26S rDNA were designed by C. Bult, D. Nickrent, and the authors (for primer sequences, see table 2; for primer locations, see table 2 and fig. 1). Sequence chromatogram output files were initially aligned and edited base by base with Sequencher™, version 2.1.1 (Gene Codes Corporation, Inc. 1994). Sequences were subsequently exported to PAUP, version 4.0d54 (Swofford, personal

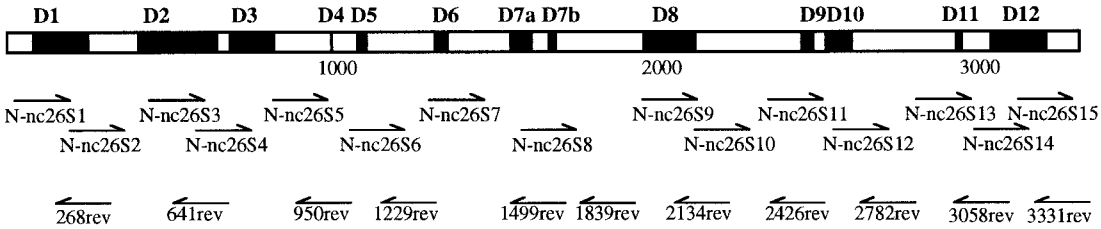

FIG. 1.—Location of primers used in PCR amplification and sequencing of 26S rDNA. Relative locations of the 12 expansion segments (D1–D12) are shown as filled boxes in the diagram of the 26S RNA gene. Below this figure are 26 primers used to sequence 26S rDNA. N-nc26S1 and 3331rev were also used in PCR amplification (see table 2 for primer sequences and their exact locations).

communication; used with permission), and aligned by eye with the seven complete 26S rDNA sequences available in GenBank (Bilofsky et al. 1986; alignment available from the authors on request).

Location of Expansion Segments

Expansion segments were located using the coordinates for expansion segment positions in the sequence of *Oryza sativa* (see table 3; used with permission from J. Hancock). Consensus motifs for the 10 bases preceding and following each expansion segment in the 22 sequences analyzed were constructed from aligned sequences using Sequencher, version 2.1.1. Standard ambiguity codes were used to designate positions within each motif exhibiting nucleotide variation among the aligned sequences. Sizes and exact locations of each expansion segment were ascertained through inspection of the aligned sequences (table 3).

Estimation of Sequence Dissimilarity in *rbcL*, 18S rDNA, and 26S rDNA

Rates of evolution of the expansion segments, conserved core regions, and entire 26S rDNA sequences were computed from 22 genera representing 14 families

Table 3
Locations and Lengths and Positions of 12 Expansion Segments (ESs) (D1–D12) (a) in 26S rDNA Sequence of *Oryza sativa*[a] and (b) in Aligned 26S Sequence[b]

(a)

Expansion Segment	Position in *Oryza*	Length	Before ES in *Oryza*	After ES in *Oryza*
D1	115–264	150	GAGATGCCCA	ACGAGTCGGG
D2	425–637	213	GGAGGGAAGC	ACCCGTCTTG
D3	679–785	107	ACATGCGTGC	AGCACGCCTG
D4	978–985	8	GCTGGAGCCC	TTCTATCGGG
D5	1,066–1,104	39	ATAGGTAGGA	ATCTCCAAGT
D6	1,137–1,163	27	GTAAGCAGAA	TGGTTACGGT
D7a.....	1,533–1,577	45	CGATCCTAAG	AAAGGGAATC
D7b	1,620–1,646	27	ACGTGGCGGT	ACGCCGGCGG
D8	1,943–2,082	140	GCTCTGAGGG	CAGCCGACTC
D9	2,453–2,476	24	GGATAAGTGG	CCACTACTTT
D10	2,504–2,581	78	ATTTTACTTA	GACATTGTCA
D11	2,984–2,987	4	CCCTACTGAT	GTGCCGCGAT
D12	3,142–3,271	130	AAGCGGCGCC	GAATCCTTTG

(b)

Expansion Segment	Position in Alignment	Length	Before ES in Seed Plants	After ES in Seed Plants
D1	111–270	160	GADNNGCCCA	DYGAGTCGGG
D2	433–679	247	RGAGGGMAVM	RCCCGTCTTG
D3	721–839	119	ACATGYGTGC	AGCAYRCCTG
D4	1,033–1,043	11	GCTRGAGCYB	TTYTATCGGG
D5	1,126–1,165	40	ATAGGTARGA	WGCTCCAAGY
D6	1,199–1,226	28	GTAAGCAGAA	KRGYTACGGT
D7a.....	1,603–1,649	47	CGATCCTAAG	AAAGGGWATC
D7b	1,693–1,718	26	AYRYGGYGGY	ACRTCGGCGR
D8	2,020–2,178	159	GCTCTGAGGR	CARYCRRCTC
D9	2,554–2,581	27	GDATAAGTGG	CCACTACTTT
D10	2,609–2,687	79	ATTTTACTTA	GACAKWGTCA
D11	3,100–3,104	5	CCTASTGAT	GYRYCGYRRT
D12	3,261–3,406	146	AVGCGRHGCV	GAATCCTTTG

[a] Also listed are the 10 bases immediately before and after each expansion segment in the *O. sativa* 26S rDNA sequence. Length of rice 26S rDNA sequence = 3,377 bp; conserved core = 2,385 bp; expansion segments = 992 bp.

[b] Also listed are the consensus sequences for the 10 bases before and after each expansion segment in the alignment. Length of matrix of aligned sequences = 3,484 bp; core regions = 2,390 bp; expansion segments = 1,094 bp.

Table 4
Sources of Sequences Used to Compare Rates of Evolution and Phylogenetic Utility of 26S rDNA, 18S rDNA, and *rbcL* (data set 2)

Family	26S Taxon	18S Taxon	*rbcL*
Ephedraceae[a]	*Ephedra distachya*	*Ephedra sinica*	*Ephedra tweediana*
Gnetaceae[a]	*Gnetum gnemon*	*Gnetum gnemon*	*Gnetum gnemon*
Araceae[b,c]	*Acorus graminae*	*Acorus calamus*	*Acorus calamus*
Poaceae[b,c]	*Oryza sativa*	*Oryza sativa*	*Oryza sativa*
Winteraceae[c]	*Drimys winteri*	*Drimys winteri*	*Drimys winteri*
Plumbaginaceae[c,d]	*Plumbago auriculata*	*Plumbago auriculata*	*Plumbago auriculata*
Brassicaceae[c]	*Brassica napus*	*Brassica campestris*	*Brassica campestris*
Rutaceae[c]	*Citrus limon*	*Citrus aurantium*	*Poncirus trifoliata*
Asteraceae[c,d]	*Tragopogon dubius*	*Tragopogon dubius*	*Tragopogon porrifolius*
Solanaceae[d]	*Lycopersicon esculentum*	*Brunfelsia pauciflora*	*Lycopersicon esculentum*
Eucryphiaceae[c,e]	*Eucryphia lucida*	*Eucryphia lucida*	*Eucryphia lucida*
Hamamelidaceae[c,e]	*Hamamelis virginiana*	*Hamamelis virginiana*	*Hamamelis mollis*
Saxifragaceae[e]	*Jepsonia parryi*	*Boykinia intermedia*	*Jepsonia parryi*
Saxifragaceae[e]	*Parnassia fimbriata*	*Parnassia fimbriata*	*Parnassia fimbriata*
Saxifragaceae[c,e]	*Saxifraga mertensiana*	*Saxifraga mertensiana*	*Saxifraga mertensiana*
Saxifragaceae[e]	*Peltoboykinia tellimoides*	*Peltoboykinia tellimoides*	*Peltoboykinia tellimoides*
Saxifragaceae[e]	*Tellima grandiflora*	*Tellima grandiflora*	*Tellima grandiflora*
Saxifragaceae[e]	*Lithophragma trifoliatum*	*Heuchera micrantha*	*Heuchera micrantha*

NOTE.—Rows contain either identical species or place-holder species from which each sequence was obtained. Rows containing taxa used to represent the Gnetales,[a] monocots,[b] angiosperms,[c] expanded Asteridae[d] (based on Chase et al. 1993), and Saxifragaceae *sensu lato*[e] (data set 3; sampling based on Morgan and Soltis 1993) are also indicated.

of seed plants (data set 1, see table 1). Comparisons of levels of base substitution among expansion segments, the conserved core regions, and entire 26S rDNA sequences were calculated by comparing sequence dissimilarity under three models of sequence evolution from taxa representing five taxonomic groups: (1) Saxifragaceae *sensu stricto* (Morgan and Soltis 1993); (2) Asteridae *sensu lato* (Olmstead et al. 1992; Chase et al. 1993); (3) the monocots; (4) the Gnetales; and (5) the angiosperms (see table 4). Sequence dissimilarity for each suite of sequences was calculated with PAUP for the Jukes-Cantor (Jukes and Cantor 1969), Kimura (1980) two-parameter, and LogDet/paralinear (Steel 1994; Lake 1994; Lockhardt et al. 1994) models of sequence evolution (reviewed in Swofford et al. 1996). Average sequence dissimilarity values were calculated for each of the five clades noted above and employed to compare inferred rates of sequence evolution.

Similarly, sequences of entire 26S rDNA, 26S rDNA expansion segments, 26S conserved core regions, entire 18S rDNA, and *rbcL* obtained from 22 genera, representing 13 families of seed plants, were used to compare levels of base substitution among these genes (data set 2; see table 4). One taxon, *Fragaria* × *ananassa*, has not been sequenced for all three genes; hence, only 13 of the 14 families for which 26S rDNA sequences were generated are included in this comparison. In six instances, different species represent the same genus (e.g., 26S rDNA and 18S rDNA sequences were generated for *Tragopogon dubius*, but the only available *rbcL* sequence was for *T. porrifolius*; see table 4); in three instances, closely related genera served as placeholders for a family (e.g., 18S rDNA and *rbcL* sequences were obtained for *Heuchera micrantha*, but the 26S rDNA sequence was from the closely related *Lithophragma trifoliatum*; see table 4).

Phylogenetic Analyses

Phylogenetic analyses of seven entire 26S rDNA sequences retrieved from GenBank (table 1) and 15 newly generated entire 26S rDNA sequences (data set 1, table 1) were conducted using PAUP. Heuristic searches were conducted on three data sets: (1) expansion segments alone; (2) conserved core areas alone; and (3) entire 26S rDNA sequences. *Gnetum gnemon* and *Ephedra distachya* were used as outgroups, following previous phylogenetic studies (Chase et al. 1993; Doyle, Donoghue, and Zimmer 1994; Soltis et al. 1997). To assess the nonrandom structure of each of the three 26S rDNA partitions (expansion segments, conserved core regions, and total 26S rDNA sequences) as well as the 18S rDNA and *rbcL* sequences for a suite of 22 genera representing 13 families (see table 4), skewness tests (Hillis and Huelsenbeck 1992) were conducted using 10,000 randomly selected trees. Heuristic searches were conducted using random taxon addition with 100 replications, TBR branch swapping, MULPARS, collapsing of branches having a maximum length of zero to yield polytomies, and ACCTRAN character state optimization. Bootstrap analyses were conducted on each data set, with 100 replicates and sampling limited to nonexcluded, nonignored (parsimony-informative) characters. To compare the strength of phylogenetic inference based on analysis of 26S, 18S, and *rbcL* sequences for the suite of 22 genera representing 13 families (data set 2; see table 4), the percentage of bootstrap values above 50 (B_{50} values, see Sanderson and Donoghue 1996) was calculated.

We also assessed the phylogenetic potential of conserved core regions and expansion segments within a well-defined clade. To accomplish this, we used a second data set of eight taxa representing Saxifragaceae *sensu stricto* (data set 3, table 4; Morgan and Soltis

1993); *Eucryphia lucida* was used as an outgroup. Heuristic searches and bootstrap analyses were conducted as above. To assess the impact of character homoplasy on phylogenetic inference for each of the five data sets, characters were reweighted based on their rescaled consistency indices and reanalyzed with parsimony. To estimate G+C bias, nucleotide homogeneity tests were conducted on each data set using PAUP. To assess the impact of sequence G+C bias on phylogenetic inference, LogDet/paralinear distance trees were inferred and compared with the results of parsimony analyses (reviewed in Swofford et al. 1996). To assess variation in levels of base substitution among sites, the number of steps per four consecutive bases was estimated with MacClade 3.0 (Maddison and Maddison 1992) using one of the four most-parsimonious trees. To assess the impact of variation in rates of base substitution among sites within 26S rDNA on phylogenetic inference, trees were also inferred using maximum-likelihood estimation corrected with an estimated shape parameter of the gamma distribution and compared with the results of parsimony analysis (Yang 1996; Lewis 1998; reviewed in Swofford et al. 1996). The shape parameter of the gamma distribution and the transition-to-transversion ratio were estimated simultaneously using PAUP. The characters were divided into three rate categories, estimated by PAUP from the shape of the gamma distribution. These rate categories were then employed in the maximum-likelihood estimation of the phylogeny.

Sequence Simplicity

Cryptic sequence simplicity can be defined as repetitive sequence motifs that are shuffled among themselves by repeated slippage events (Tautz, Trick, and Dover 1986). If present, cryptic sequence simplicity violates the assumption that characters at different sites are evolving independently. Sequence simplicity was assessed for the 15 newly generated 26S rDNA sequences using SIMPLE34 (Hancock and Armstrong 1994). Analyses were conducted on an SGI-UNIX computer at the Washington State University VADMS Center using the conditions described in Bult, Sweere, and Zimmer (1995), except that the default setting of a 64-base sliding window was employed here. The relationship between the sequence simplicity score for an actual sequence and a randomly generated sequence having the same base composition is expressed quantitatively as the relative simplicity factor (RSF). Sequences showing an RSF greater than one and a raw simplicity factor greater than three exhibit significant cryptic sequence simplicity (for details, see Hancock and Armstrong 1994). Sequences having a significant RSF score were reanalyzed using a second order Markov rule to assess whether elevated G+C composition significantly affected RSF values (Hancock and Armstrong 1994).

To assess internal sequence similarity, dot plot comparisons for the 15 new 26S rDNA sequences were conducted using the Genetics Computer Group package, release version 8.0 (Genetics Computer Group 1994). Dot plots were generated using a sliding window of nine bases with an offset of three bases and a stringency of 78%, which yielded the best signal-to-noise ratio. Dot plot analyses were compared with those obtained for *Escherichia coli* as a negative control and *Homo sapiens* and *Oryza sativa* as positive controls to aid in dot plot interpretation (for details see Hancock and Dover 1988; Bult, Sweere, and Zimmer 1995). The presence of cryptic sequence simplicity was assessed through visual inspection of dot plots. SIMPLE34 profiles were plotted along each axis of the dot plot profiles to assess the extent of agreement between the SIMPLE34 and dot plot analyses for each taxon (data not shown).

Results and Discussion

Patterns of Plant 26S rDNA Evolution

The 15 new 26S rDNA sequences were easily aligned by eye with the seven previously available entire 26S sequences. Site variability was greatest in the expansion segments (fig. 2), which evolve 6.4 to 10.2 times as fast as conserved core regions (table 5). The number of steps on a most-parsimonious tree was assessed using windows of four consecutive bases. In the conserved core regions, the number of steps did not exceed 20. In contrast, values in the expansion segments frequently surpassed 20 steps per four consecutive nucleotides (fig. 2). Although expansion segments exhibit a higher rate of base substitution and may be under lower functional constraint than the conserved core regions, it is unlikely that they lack a functional role altogether, as some have suggested (Gerbi et al. 1987). Expansion segments exhibit a rate of base substitution lower than those of nuclear noncoding regions or neutral bases (Larson and Wilson 1989), show conservation of secondary structure among highly divergent organisms (Hancock and Dover 1990), and are present in the functional ribosomes of some eukaryotes (Hassouna, Michot, and Bachellerie 1984). Larson (1991) suggested that this difference in rates of base substitution between conserved core regions and expansion segments could be exploited for phylogenetic inference at different taxonomic levels; conserved core regions could be used at higher taxonomic levels, and expansion segments could be employed among more closely related taxa. Larson indicated that sites in expansion segments might become saturated and should be excluded from investigations of taxa with a common ancestor older than 200 MYA, which is consistent with our phylogenetic results (see below).

The average G+C content of the entire 26S rDNA is 56.2%, which deviates slightly from expectations given equal nucleotide frequencies ($\chi^2 = 65.65$; df = 63; $P = 0.3850$); that of the conserved core regions is 52.0%. These G+C values are consistent with the findings of Bult, Sweere, and Zimmer (1995) and expectations given equal nucleotide frequencies ($\chi^2 = 10.94$; df = 63; $P = 1.000$). The expansion segments have an average G+C content of 66.0%, which is slightly higher than the values obtained by Bult, Sweere, and Zimmer (1995) and is significantly higher than would be expected if nucleotides were equally abundant ($\chi^2 = 131.81$; df = 63; $P = 8.9 \times 10^{-7}$). The significantly

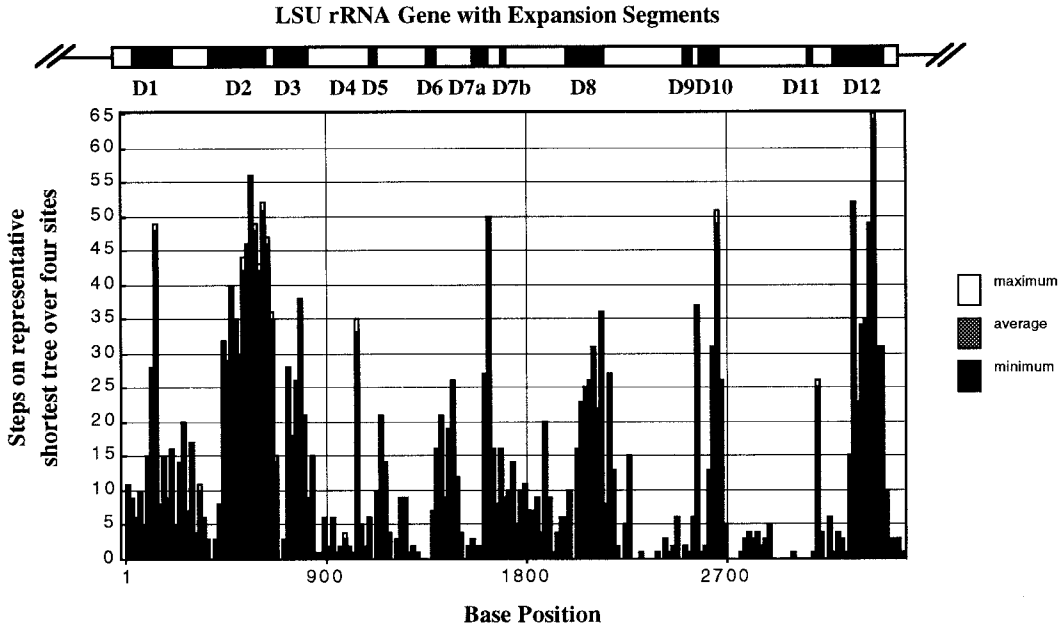

FIG. 2.—Site variation across 26S rDNA sequence. Above is an illustration of the relative locations of the 12 expansion segments (D1–D12) of 26S rDNA. Variation in base substitution rates over the length of 26S rDNA for the 22 sequences of 26S rDNA analyzed in this study were calculated using a window size of four consecutive bases and graphed using MacClade 3.0 (Maddison and Maddison 1992).

higher G+C content in the 26S rDNA expansion segment regions poses a potential problem for phylogeny retrieval if the algorithm used to construct a tree assumes equally abundant nucleotides (for example, maximum-likelihood estimation using a Jukes-Cantor model of sequence evolution; see Swofford et al. 1996).

Dot plot analyses suggest that 4 of the 15 new 26S rDNA sequences showed marginal evidence of cryptic sequence simplicity and molecular coevolution among the divergence domains (data not shown): *Peltoboykinia tellimoides*, *Mitella pentandra*, *L. trifoliatum*, and *A. graminae*. SIMPLE34 analyses showed that 5 of the 15 new 26S sequences exhibited some degree of significant tri- and tetranucleotide motif repetition: *G. gnemon*, *Plumbago auriculata*, *Saxifraga mertensiana*, *L. trifoliatum*, and *T. dubius*. However, qualitative comparisons of dot plots and SIMPLE 34 output for plant 26S rDNA sequences suggest that the amount and intensity of detected internal sequence similarity was considerably less than that observed in animal 28S rDNA. This observation was also evident in the results reported by Bult, Sweere, and Zimmer (1995) and Hancock and Dover (1988). Sequence simplicity was confined primarily to expansion segments, also consistent with the findings of these authors.

Cryptic sequence simplicity results from compensatory slippage, which obscures phylogenetic signal and impairs homology assessment among LSU rDNA sequences, particularly at higher taxonomic levels (Bult, Sweere, and Zimmer 1995). Our findings suggest that although cryptic sequence simplicity is strongest in the same relative position in LSU rDNA of plants and animals, it is not present at the same levels. This observation is partially explicable in terms of the length of LSU rDNA in these taxonomic categories. The 26S rDNA sequences of plants are much shorter than the 28S rDNA sequences of animals, averaging 3.4 kb, whereas in animals, the average length is 4.5 kb, ranging from 3,519 to 5,025 bp (Hancock and Dover 1988). Recent investigation of the evolution of expansion segments in crustaceans demonstrated a positive correlation between expansion segment length and cryptic sequence simplicity (Nunn et al. 1996). Expansion segments in animal 28S rDNA show significantly more character nonindependence in the form of cryptic sequence simplicity (Tautz, Trick, and Dover 1986; Hancock and Dover 1988), compensatory slippage (Hancock and Dover 1990), and compensatory mutations (Dixon and Hillis 1993). Our findings suggest that the shorter expansion segments of plant rDNA appear to undergo less compensatory slippage and exhibit fewer length mutations and, consequently, may retain greater phylogenetic signal at higher taxonomic levels than in animal 28S rDNA. The extent to which the major clades of life differ in their patterns of cryptic sequence simplicity should be explored on a larger scale through analysis of LSU rDNA across all kingdoms.

Phylogenetic Informativeness of Plant 26S rDNA Sequences

The three partitions of the entire 26S data set, containing the 7 previously available sequences and the 15 newly generated sequences (data set 1), were also assessed for phylogenetic informativeness. Three indicators were used in these analyses: (1) skewness values to indicate the nonrandom structuring of the data, (2) homoplasy indices to indicate the relative amounts of character homoplasy, and (3) their B_{50} value to provide a measure of the robustness of inferences made from each data set. Analysis of the expansion segments yielded

Table 5
Comparison of Dissimilarity (a) Within 26S rDNA Sequences, (b) Between 26S and 18S rDNA Sequences, and (c) Between 26S rDNA and *rbcL* Sequences Calculated Under Jukes-Cantor, Kimura Two-Parameter, and LogDet/Paralinear Models of Sequence Evolution

(a)

Taxonomic Unit	Model of Evolution	ES/26S	ES/CC	26S/CC
Saxifragaceae *s.l.*	Jukes-Cantor	2.64857	7.53662	2.84554
	Kimura two-parameter	2.67233	7.62458	2.85316
	LogDet/paralinear	2.91028	8.51393	2.9254
Asteridae *s.l.*	Jukes-Cantor	2.6695	7.5918	2.84391
	Kimura two-parameter	2.69465	7.68648	2.8525
	LogDet/paralinear	3.07316	9.09221	2.95859
Monocots	Jukes-Cantor	2.67763	6.43957	2.40495
	Kimura two-parameter	2.68235	6.45037	2.40475
	LogDet/paralinear	3.6704	7.44229	2.4265
Gnetales	Jukes-Cantor	2.9115	9.23612	3.17224
	Kimura two-parameter	2.94459	9.36215	3.17944
	LogDet/paralinear	3.1854	10.21447	3.20666
Angiosperms	Jukes-Cantor	2.64211	6.68645	2.53073
	Kimura two-parameter	2.66264	6.75017	2.53514
	LogDet/paralinear	2.96611	7.63928	2.57554

(b)

Taxonomic Unit	Model of Evolution	ES/18S	26S/18S	CC/18S
Saxifragaceae *s.l.*	Jukes-Cantor	5.76081	2.17506	0.76438
	Kimura two-parameter	5.82783	2.1808	0.76435
	LogDet/paralinear	6.54885	2.25025	0.76919
Asteridae *s.l.*	Jukes-Cantor	4.48755	1.68105	0.59111
	Kimura two-parameter	4.54109	1.68522	0.59079
	LogDet/paralinear	5.39515	1.75557	0.59338
Monocots	Jukes-Cantor	4.15891	1.5532	0.64584
	Kimura two-parameter	4.16667	1.55337	0.64596
	LogDet/paralinear	4.85514	1.583	0.65237
Gnetales	Jukes-Cantor	5.53506	1.90107	0.59928
	Kimura two-parameter	5.59336	1.89954	0.59744
	LogDet/paralinear	6.02956	1.89288	0.5903
Angiosperms	Jukes-Cantor	5.4683	2.06967	0.81782
	Kimura two-parameter	5.52707	2.07578	0.8188
	LogDet/paralinear	5.29067	2.11597	0.82157

(c)

Taxanomic Unit	Model of Evolution	ES/*rbcL*	26S/*rbcL*	CC/*rbcL*
Saxifragaceae *s.l.*	Jukes-Cantor	2.64733	0.99953	0.35126
	Kimura two-parameter	2.66985	0.99907	0.35016
	LogDet/paralinear	2.95745	1.01621	0.34737
Asteridae *s.l.*	Jukes-Cantor	1.20123	0.44998	0.15823
	Kimura two-parameter	1.21165	0.44965	0.15763
	LogDet/paralinear	1.40351	0.4567	0.15436
Monocots	Jukes-Cantor	1.50978	0.56385	0.23445
	Kimura two-parameter	1.49294	0.55658	0.23145
	LogDet/paralinear	1.71199	0.55819	0.23003
Gnetales	Jukes-Cantor	2.29771	0.56385	0.24877
	Kimura two-parameter	2.31169	0.78506	0.24692
	LogDet/paralinear	2.52873	0.79385	0.24756
Angiosperms	Jukes-Cantor	1.79425	0.6791	0.27771
	Kimura two-parameter	1.80297	0.67714	0.2671
	LogDet/paralinear	2.02293	0.68202	0.26481

NOTE.—26S = entire sequence, CC = conserved core region, ES = expansion segments.

five most-parsimonious trees on two islands, a skewness value of $g_1 = -1.0924$ ($P < 0.01$), a homoplasy index of 0.492, and a B_{50} value of 57.9%. Analysis of the conserved core regions for the same suite of taxa yielded 15 most-parsimonious trees on one island, a skewness value of $g_1 = -1.545$ ($P < 0.01$), a homoplasy index of 0.488, and a B_{50} value of 37.8%. Analysis of entire 26S sequences yielded three most-parsimonious trees, a skewness value of $g_1 = -1.205$ ($P < 0.01$), a homoplasy index of 0.494, and a B_{50} value of 57.9%.

The skewness values indicate that entire 26S rDNA sequences, the conserved core regions, and the expansion segment regions all contain significant nonrandom structure that likely reflects phylogenetic signal. Of the three partitions, the strict consensus trees resulting from analysis of entire 26S sequences yielded the greatest res-

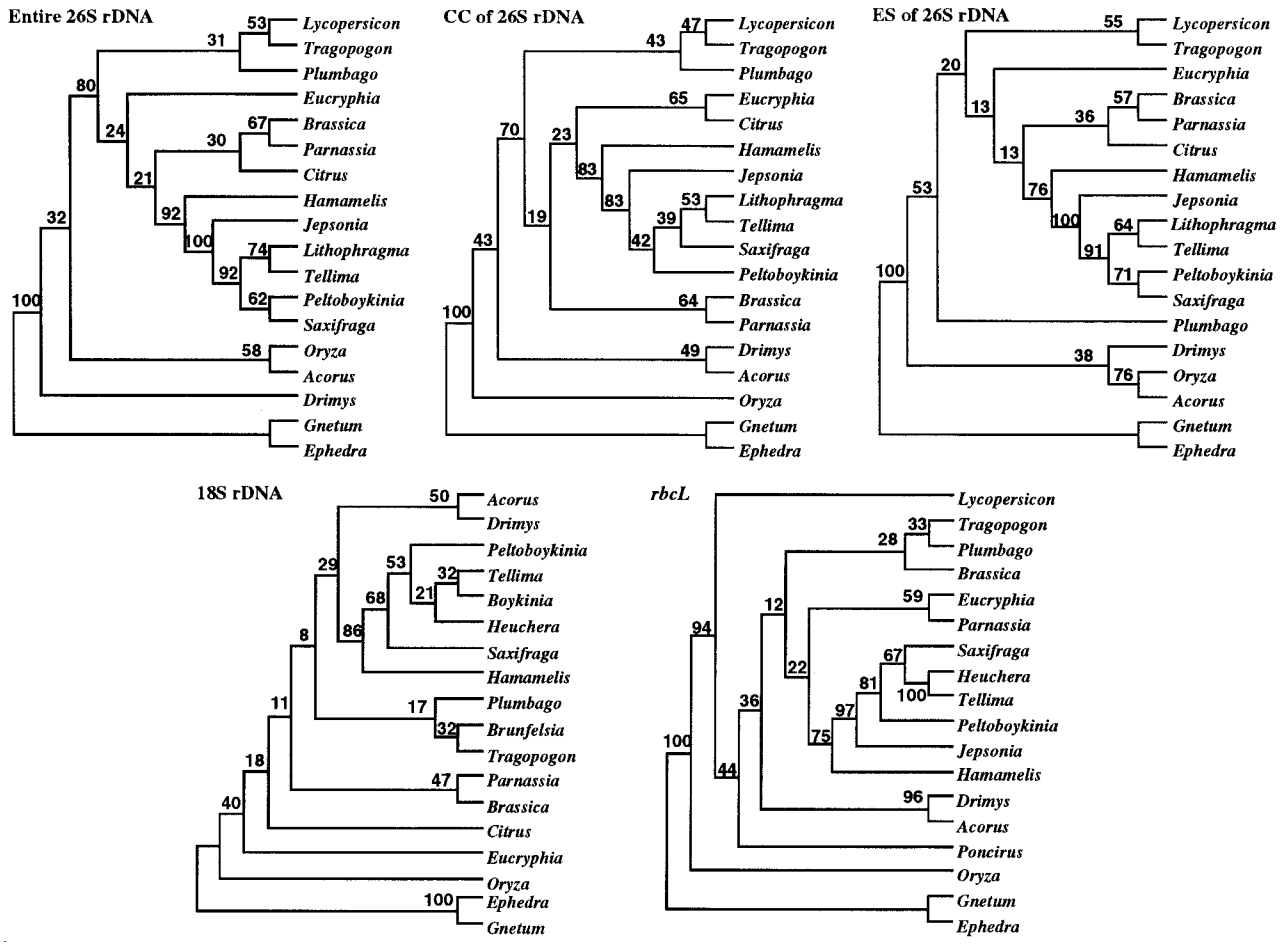

FIG. 3.—Comparison of consensus trees from bootstrap analysis of entire 26S rDNA, conserved core regions of 26S rDNA (CC), expansion segments of 26S rDNA (ES), 18S rDNA, and *rbcL* sequences. Numbers labeled above each node are bootstrap values. The number of phylogenetically informative characters calculated by PAUP for each data set was 556 for 26S, 158 for CC, 428 for ES, 163 for 18S, and 291 for *rbcL*.

olution, identified the largest number of groups believed to be monophyletic on the basis of previous large-scale analyses of angiosperms (e.g., Chase et al. 1993; Soltis et al. 1997), and showed the greatest average support for those groups. Analysis of the expansion segments alone yielded results that were more similar to those previous large-scale analyses than did analysis of conserved core regions. The topology based on analysis of the conserved core regions exhibited the lowest B_{50} value, but also had the lowest homoplasy index. This result is consistent with the findings of Sanderson and Donoghue (1996) that homoplasy indices are not inversely related to average bootstrap values.

Entire 26S rDNA sequences for the suite of 22 genera representing 13 families (data set 2; see table 4) exhibit levels of base substitution 1.6 to 2.2 times as high as that of entire 18S rDNA sequences. Because of their greater length (about 3.4 kb vs. 1.8 kb), the 26S rDNA sequences provided 3.3 times as many phylogenetically informative characters as 18S rDNA. Entire 26S rDNA sequences exhibit levels of base substitution 0.4 to 1.0 times that of *rbcL*, but, again due to its greater length (3.4 kb vs. 1.4 kb), 26S rDNA provided 2.0 times the number of phylogenetically informative characters (table 5). Conserved core regions exhibit a level of base substitution that is 0.59 to 0.81 times that of entire 18S rDNA sequences. Due to their greater length (2,390 vs. 1,850 aligned bases), the conserved core regions of 26S rDNA provide approximately the same number of phylogenetically informative characters as 18S rDNA sequences. From these comparisons of evolutionary rates, we infer that 26S rDNA sequences have a rate of base substitution useful for phylogeny retrieval at taxonomic levels comparable to those for which 18S rDNA or *rbcL* sequences have been used. Furthermore, the use of 26S rDNA sequences would increase the amount of phylogenetically informative characters fourfold when added to 18S rDNA sequence data.

B_{50} values obtained from analysis of 18S and *rbcL* sequences as well as the three partitions of the 26S sequences for 22 genera representing 13 families (data set 2, table 4) were as follows: entire 26S rDNA, 66.7%; conserved core regions alone, 46.7%; expansion segments, 73.3%; entire 18S rDNA, 33.3%; and *rbcL*, 60.0% (fig. 3). For the suite of eight taxa centered on Saxifragaceae *sensu stricto* (data set 3, table 4), B_{50} values were as follows: entire 26S rDNA, 100%; conserved core regions, 60%; expansion segments alone, 100%;

entire 18S rDNA, 40.0%; *rbcL*, 100%. The pattern in these values indicates that when taxon sampling is very sparse, as in our sampling across seed plants (data set 2, table 4), or when it is more concentrated, as in our sampling of the Saxifragaceae *sensu stricto* (data set 3, table 4), B_{50} values for entire 26S rDNA sequences and expansion segments are equal to or greater than those of entire 18S rDNA or *rbcL* sequences. These results suggest that, when analyzed phylogenetically, sequences of entire 26S rDNA, and even expansion segments alone, provide inferences that are at least as well resolved and supported as those based on *rbcL* and 18S rDNA for a similar suite of taxa. B_{50} values for the conserved core regions of 26S suggest that, when analyzed phylogenetically, they will provide inferences at least as robust as those drawn from entire 18S rDNA sequences.

Topologies of consensus trees obtained from bootstrap analysis, using parsimony as an optimizing criterion, are similar for the entire 26S rDNA, 26S conserved core regions, and 26S expansion segments (fig. 3). All of these topologies correctly place *Oryza*, *Acorus*, and *Drimys* near the base of the angiosperms. Additionally, they all identify the Saxifragaceae *sensu stricto* and the Asteridae *sensu lato* as monophyletic groups. Furthermore, analysis of all partitions of 26S rDNA place the Asteridae *sensu lato* clade as sister to a large clade containing all the members of the expanded Rosidae, consistent with the findings of Chase et al. (1993) and Soltis et al. (1997). Bult, Sweere, and Zimmer (1995) and others expressed concern that pronounced compensatory slippage could render 26S rDNA sequence variation unfit for use in phylogeny retrieval at higher taxonomic levels. However, analysis of entire 26S rDNA provided more resolution, greater internal support, and a topology that was more consistent with previous cladistic analysis than did analysis of conserved core regions alone. Hence, our findings suggest that expansion segments should not be excluded from phylogenetic analyses of 26S rDNA at higher taxonomic levels in the angiosperms. Since the major diversification of angiosperms is believed to have begun around 130 MYA (Crane, Friis, and Pedersen 1995), our finding is consistent with an estimate by Larson and Wilson (1989) that expansion segments would be a useful source of phylogenetic information for evolutionary events occurring within the last 100–200 Myr.

When compared with the results of previous large-scale cladistic analyses (Chase et al. 1993; Doyle, Donoghue, and Zimmer 1994; Soltis et al. 1997), our topologies based on phylogenetic analysis of 26S rDNA across angiosperms and gnetophytes appear to be at least as accurate as those based on either 18S rDNA or *rbcL*. However, within 26S rDNA sequences, unequal rates of base substitution between the expansion segments and conserved core regions could have a negative impact on phylogeny retrieval based on parsimony analysis of 26S rDNA sequences (Wakeley 1996; Yang 1996), but might be compensated for with an appropriate weighting scheme (Farris 1969; Williams and Fitch 1990) or with maximum-likelihood estimation corrected for the shape of the gamma distribution (Yang 1996; Lewis 1998). Consequently, combining 26S rDNA sequences with 18S rDNA and *rbcL* sequences for inferences at higher taxonomic levels will increase the number of phylogenetically informative characters and likely provide greater resolution and support.

For the suite of taxa representing the Saxifragaceae *sensu stricto* (table 4), pooled sequences from the 12 expansion segments have an aligned length of 1,094 bp, exhibit levels of base substitution 2.6 to 3.0 times that of *rbcL*, and provide 1.5 times the number of phylogenetically informative characters. Comparison of topologies resulting from unweighted parsimony analyses of 26S rDNA expansion segments, conserved core regions, and *rbcL* indicate that the underlying phylogenetic signal is highly similar (fig. 4). Topologies are entirely concordant, with the exception of the placement of *Saxifraga*. However, very little resolution is seen in the Saxifragaceae *sensu stricto* in the topology derived from analysis of the conserved core regions alone. This suggests that conserved core regions could safely be excluded from phylogenetic investigations below the family level, saving the researcher considerable sequencing effort. Changing reconstruction methods and optimality criteria in analyses of *rbcL* or the expansion segments alone had little impact on the resultant topology. This suggests that GC bias, among-site rate variation, and homoplastic characters are not present in expansion segments at sufficient levels to affect phylogeny retrieval adversely. Hence, pooled sequence data from the expansion segments might be used profitably at levels comparable to those investigated with *matK* cpDNA sequences, which were shown to evolve approximately three times faster than *rbcL* sequences in the Saxifragaceae *sensu stricto* (Johnson and Soltis 1994). Thus, 26S expansion segments may provide a marker from the nuclear genome useful for phylogeny retrieval at the intrafamilial level.

Conclusions

We conclude that plant 26S rDNA sequences contain significant phylogenetic signal in both conserved core regions and expansion segments. These sequences are easily aligned, evolve 1.6 to 2.2 times as fast as 18S rDNA, and yield 3.3 times the number of phylogenetically informative characters. When compared with *rbcL*, 26S rDNA evolves 0.4 to 1.0 times as fast, but produces 2.0 times as many phylogenetically informative characters due to its greater length. Hence, entire 26S sequences may be useful for phylogeny retrieval at taxonomic levels comparable to those for which *rbcL* has been used. Conserved core regions evolve 0.59 to 0.82 times as fast as entire 18S rDNA sequences and provide approximately the same number of phylogenetically informative characters, and therefore should be appropriate for phylogeny retrieval at taxonomic levels similar to those investigated with 18S rDNA. Expansion segments evolve at a rate 6.4 to 10.2 times as fast as the conserved core regions of 26S rDNA; they may need to be appropriately weighted, or perhaps excluded, from phylogenetic analyses at much higher taxonomic levels

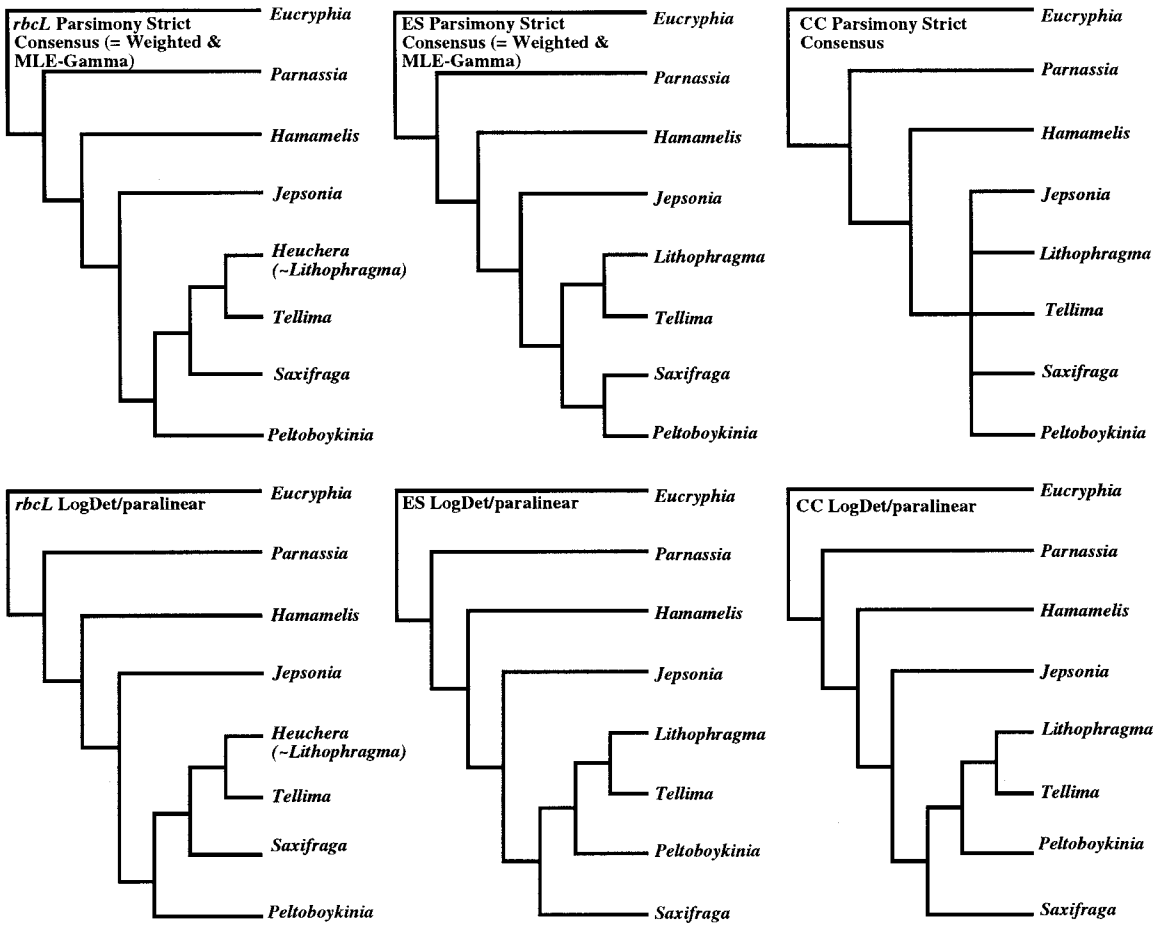

FIG. 4.—Comparison of phylogenetic inferences based on *rbcL*, 26S rDNA expansion segments (ES), and conserved core regions (CC) when analyzed with unweighted parsimony, weighted parsimony (Weighted), maximum-likelihood estimation using an estimated shape parameter for the gamma distribution and three rate categories (MLE-Gamma) (Yang 1996), or the log-determinant/paralinear distance method (LogDet/paralinear) (Lake 1994; Lockhardt et al. 1994; Steel 1994). Topologies based on *rbcL* are identical when analyzed with any of the methods of phylogenetic inference. Topologies based on unweighted or weighted parsimony, as well as MLE-Gamma analysis of ES, were identical, but discordant with those based on *rbcL* regarding the placement of *Saxifraga.* LogDet/paralinear analysis of ES and CC regions produced identical topologies that differed from those based on *rbcL* regarding the placement of *Saxifraga.*

(e.g., among green plants as a whole), but they appear to be informative at or below the interfamilial level in angiosperms. For the Saxifragaceae *sensu stricto*, 26S rDNA expansion segments evolve from 2.6 to 3.0 times as fast as *rbcL* and provide nuclear characters that are useful at taxonomic levels similar to those addressed with *matK* sequences. Finally, LSU rDNA expansion segments in plants have a pattern of evolution that is distinct from that found in animals, exhibiting less cryptic sequence simplicity and a lower frequency of insertion and deletion mutations, and, in general, may have greater phylogenetic potential than has been suggested for animal 28S rDNA.

Acknowledgments

The authors thank D. Swofford for technical assistance and permission to use PAUP* for data analysis; P. Lewis for technical assistance and advice on maximum-likelihood analysis; C. Bult for technical assistance with expansion segment analysis and primer design; D. Nickrent for technical assistance with data analysis and primer design; S. Thompson and S. Johns and the VADMS Center of Washington State University for technical assistance with the Genetics Computer Group Package and SIMPLE 34 analyses; J. Hancock for permission to use unpublished coordinates for expansion segments in *Oryza sativa*; two anonymous reviewers for comments on an earlier version of this paper; the Smithsonian Institution's Laboratory of Molecular Systematics in the National Museum of Natural History, where much of this work was completed; the Mellon Foundation for a grant to E.A.Z., P.S.S., and D.E.S.; the Higinbotham Trust Fund for an award to R.K.K.; and the National Science Foundation for a doctoral dissertation improvement grant to D.E.S., R.K.K., and L. Hufford.

LITERATURE CITED

APPELS, R., and R. L. HONEYCUTT. 1986. rDNA: evolution over a billion years. Pp. 81–135 *in* S. K. DUTTA, ed. DNA systematics, Vol. 2. Plants. CRC Press, Boca Raton, Fla.

ARNHEIM, N. 1983. Concerted evolution of multigene families. Pp. 38–61 *in* M. NEI and R. K. KOEHN, eds. Evolution of genes and gene proteins. Sinauer, Sunderland, Mass.

AUWERA, G. V. D., and R. DEWACHTER. 1996. Large-subunit rRNA sequence of the Chytridiomycete *Blastocladiella emersonii* and implications for the evolution of zoosporic fungi. J. Mol. Evol. **43**:476–483.

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

BILOFSKY, H. S., C. BURKS, J. W. FICKETT, W. B. GOAD, F. I. LEWITTER, W. P. RINDONE, C. D. SWINDELL, and C. S. TUNG. 1986. The GenBank genetic sequence data bank. Nucleic Acids Res. **14**:1–4.

BUCHHEIM, M. A., and R. L. CHAPMAN. 1991. Phylogeny of the colonial green flagellates: a study of 18S and 26S rRNA sequence data. BioSystems **25**:85–100.

BULT, C. J., M. KALLERSJO, and Y. SUH. 1992. Amplification and sequencing of 16/18S rDNA from gel-purified total plant DNA. Plant Mol. Biol. Rep. **10**:273–284.

BULT, C. J., J. A. SWEERE, and E. A. ZIMMER. 1995. Cryptic sequence simplicity, nucleotide composition bias, and molecular coevolution in the large subunit of ribosomal DNA in plants: implications for phylogenetic analyses. Ann. Mo. Bot. Gard. **82**:235–246.

BULT, C. J., and E. A. ZIMMER. 1993. Nuclear ribosomal RNA sequences for inferring tribal relationships within Onagraceae. Syst. Bot. **18**:48–63.

CHASE, M. W., D. E. SOLTIS, R. G. OLMSTEAD et al. (42 coauthors). 1993. Phylogeneticts of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcL*. Ann. Mo. Bot. Gard. **80**:628–680.

CLARK, G. B., B. W. TAGUE, V. C. WARE, and S. A. GERBI. 1984. *Xenopus laevis* 28S ribosomal RNA: a secondary structure and its evolutionary and functional implications. Nucleic Acids Res. **12**:6197–6220.

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

DIXON, M. T., and D. M. HILLIS. 1993. Ribosomal RNA secondary structure—compensatory mutations and implications for phylogenetic analysis. Mol. Biol. Evol. **10**:256–267.

DOVER, G. A., and R. B. FLAVELL. 1984. Molecular coevolution: DNA divergence and the maintenance of function. Cell **38**:622–623.

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

DOYLE, J. J. 1992. Gene trees and species trees: molecular systematics as one-character taxonomy. Syst. Bot. **17**:144–163.

———. 1993. DNA, phylogeny, and the flowering of plant systematics. BioScience **43**:380–389.

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

FLAVELL, R. B. 1986. Structure and control of expression of ribosomal RNA genes. Oxf. Surv. Plant Mol. Cell Biol. **3**: 252–274.

GENE CODES CORPORATION, INC. 1994. Sequencher™ 2.1 reference, advanced, user friendly software tools for DNA sequencing. Gene Codes Corporation, Inc., Ann Arbor, Mich.

GENETICS COMPUTER GROUP. 1994. Program manual for the GCG package. Version 8. Genetics Computer Group, Madison, Wis.

GERBI, S. A., C. JEPPESEN, B. STEBBINS-BOAZ, and M. ARES JR. 1987. Evolution of eukaryotic rRNA: constraints imposed by RNA interactions. Cold Spring Harb. Symp. Quant. Biol. **L11**:709–719.

HAMBY, K. R., and E. A. ZIMMER. 1988. Ribosomal RNA sequences for inferring phylogeny within the grass family (Poaceae). Plant Syst. Evol. **160**:29–37.

———. 1992. Ribosomal RNA as a phylogenetic tool in plant systematics. Pp. 50–91 *in* P. SOLTIS, D. SOLTIS, and J. DOYLE, eds. Molecular systematics of plants. Chapman and Hall, New York.

HANCOCK, J. M., and J. S. ARMSTRONG. 1994. SIMPLE34: an improved and enhanced implementation for VAX and SUN computers of the SIMPLE algorithm for the analysis of clustered repetitive sequences. CABIOS **10**:67–70.

HANCOCK, J. M., and G. A. DOVER. 1988. Molecular coevolution among cryptically simple expansion segments of eukaryotic 26S/28S rRNAs. Mol. Biol. Evol. **5**:377–391.

———. 1990. 'Compensatory slippage' in the evolution of ribosomal RNA genes. Nucleic Acids Res. **18**:377–391.

HASSOUNA, N., B. MICHOT, and J. BACHELLERIE. 1984. The complete nucleotide sequence of mouse 28S rRNA gene: implications for the process of size increase of the large subunit rRNA in higher eukaryotes. Nucleic Acids Res. **12**: 3563–3574.

HILLIS, D. M., and M. T. DIXON. 1991. Ribosomal DNA: molecular evolution and phylogenetic inference. Q. Rev. Biol. **66**:411–453.

HILLIS, D. M., and J. P. HUELSENBECK. 1992. Signal, noise, and reliability in molecular phylogenetic analyses. J. Hered. **83**:189–195.

HOLZMANN, M., W. PILLER, and J. PAWLOWSKI. 1996. Sequence variation in the large-subunit ribosomal RNA gene of *Ammonia* (Foraminifera, Protozoa) and their evolutionary implications. J. Mol. Evol. **43**:145–151.

JOHNSON, L. A., and D. E. SOLTIS. 1994. *mat*K DNA sequences and phylogenetic reconstruction in Saxifragaceae s. s. Syst. Bot. **19**:143–156.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. N. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

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

KISS, T., M. KISS, and F. SOLYMOSY. 1989. Nucleotide sequence of a 25S rRNA gene from tomato. Nucleic Acids Res. **17**:796.

KOLOSHA, V. O., and I. I. FODOR. 1990. Nucleotide sequence of *Citrus limon* 26S rRNA gene and secondary structure model of its RNA. Plant Mol. Biol. **14**:147–161.

KRON, K. 1996. Phylogenetic relationships of Empetraceae, Epacridaceae, and Ericaceae: evidence from nuclear ribosomal 18S sequence data. Ann. Bot. **77**:293–303.

LAKE, J. 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc. Natl. Acad. Sci. USA **91**:1455–1459.

LARSON, A. 1991. Evolutionary analysis of length variable sequences: divergence domains of ribosomal RNA. Pp. 221–247 *in* M. MIYAMOTO and J. CRACRAFT, eds. Phylogenetic analysis of DNA sequences. Oxford University Press, New York.

LARSON, A., and A. C. WILSON. 1989. Patterns of ribosomal RNA evolution in salamanders. Mol. Biol. Evol. **6**:131–154.

LEWIS, P. O. 1998. Alternatives to parsimony for inferring phylogeny using nucleotide sequence data. *In* D. E. SOLTIS, P. S. SOLTIS, and J. J. DOYLE, eds. Molecular systematics of plants II. Chapman and Hall, New York (in press).

LOCKHARDT, P. J., M. A. STEELE, 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. MacClade. Version 3.0. Sinauer, Sunderland, Mass.

MISHLER, B. D., L. A. LEWIS, M. A. BUCHHEIM, K. S. RENZAGLIA, D. J. GARBARY, C. F. DELWICHE, F. W. ZECHMAN, T. S. KANTZ, and R. L. CHAPMAN. 1994. Phylogenetic relationships of the ''green algae'' and ''bryophytes'' Ann. Mo. Bot. Gard. **81**:451–483.

MORGAN, D. R., and D. E. SOLTIS. 1993. Phylogenetic relationships among Saxifragaceae *sensu lato* based on *rbcL* sequence data. Ann. Mo. Bot. Gard. **80**:631–660.

NICKRENT, D. L., and D. E. SOLTIS. 1995. A comparison of angiosperm phylogenies from nuclear 18S rDNA and *rbcL* sequences. Ann. Mo. Bot. Gard. **82**:208–234.

NUNN, G. B., B. F. THEISEN, B. CHRISTENSEN, and P. ARCTANDER. 1996. Simplicity-correlated size growth of the nuclear 28S ribosomal RNA D3 expansion segment in the crustacean order Isopoda. J. Mol. Evol. **42**:221–223.

OKUMURA, S., and H. SHIMADA. 1992. Nucleotide sequence of genes for 5.8S and 25S rRNA from rapeseed (*Brassica napus).* Nucleic Acids Res. **20**:3510.

OLMSTEAD, R. G., H. J. MICHAELS, K. M. SCOTT, and J. PALMER. 1992. Monophyly of the Asteridae and identification of their major lineages as inferred from DNA sequences of *rbc*L. Ann. Mo. Bot. Gard. **79**:249–265.

OLMSTEAD, R. G., and J. A. SWEERE. 1994. Combining data in phylogenetic systematics: an empirical approach using three molecular data sets in the Solanaceae. Syst. Biol. **43**: 467–481.

RO, K., C. S. KEENER, and B. A. MCPHERON. 1997. Molecular phylogenetic study of the Ranunculaceae: utility of the nuclear 26S ribosomal DNA in inferring intrafamilial relationships. Mol. Phylogenet. Evol. **8**:117–127.

SANDERSON, M. J., and M. J. DONOGHUE. 1996. The relationship between homoplasy and confidence in a phylogenetic tree. Pp. 67–89 *in* M. J. SANDERSON and L. HUFFORD, eds. Homoplasy and the evolutionary process. Academic Press, New York.

SOLTIS, D. E., D. R. MORGAN, A. GRABLE, P. S. SOLTIS, and R. K. KUZOFF. 1993. Molecular systematics of Saxifragaceae *sensu stricto.* Am. J. Bot. **80**:1056–1081.

SOLTIS, D. E., and P. S. SOLTIS. 1997. Phylogenetic relationships among Saxifragaceae sensu lato: a comparison of topologies based in 18S rDNA and *rbcL* sequences. Am. J. Bot. **84**:504–522.

———. 1998. Choosing an approach and an appropriate gene for phylogenetic analysis. *In* D. E. SOLTIS, P. S. SOLTIS, and J. J. DOYLE, eds. Molecular systematics of plants II. Chapman and Hall, New York (in press).

SOLTIS, D. E., P. S. SOLTIS, D. L. NICKRENT et al. (16 coauthors). 1997. Angiosperm phylogeny inferred from 18S ribosomal DNA sequences. Ann. Mo. Bot. Gard. **84**:1–49.

STEEL, M. 1994. Recovering a tree from the Markov leaf colourations it generates under a Markov model. Appl. Math. Lett. **7**:19–23.

STEFANOVIC, S., M. JAGER, J. DEUTSCH, J. BROUTIN, and M. MASSELOT. 1998. Phylogenetic relationships of conifers inferred from partial 28S rDNA gene sequences. Am. J. Bot. (in press).

SUGIURA, M., Y. IIDA, K. OONO, and F. TAKAIWA. 1985. The complete nucleotide sequence of a rice 25S rRNA gene. Gene **37**:255–259.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, and D. M. HILLIS. 1996. Phylogenetic inference. Pp. 407–514 *in* D. M. HILLIS, C. MORITZ, and B. K. MABLE, eds. Molecular systematics. 2nd edition. Sinauer, Sunderland, Mass.

TAUTZ, D., M. TRICK, and G. A. DOVER. 1986. Cryptic simplicity in DNA is a major source of genetic variation. Nature **322**:652–656.

UNFRIED, I., and P. GRUENDLER. 1990. Nucleotide sequence of the 5.8S and 25S rRNA genes and of the internal transcribed spacers from Arabidopsis thaliana. Nucleic Acids Res. **18**:4011.

VOGLER, A. P., A. WELSCH, and J. M. HANCOCK. 1997. Phylogenetic analysis of slippage-like sequence variation in the V4 rRNA expansion segment in tiger beetles (Cicindelidae). Mol. Biol. Evol. **14**:6–19.

WAKELEY, J. 1996. The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. Trends Ecol. Evol. **11**:158–163.

WILLIAMS, P. L., and W. M. FITCH. 1990. Phylogenetic determination using dynamically weighted parsimony method. Methods Enzymol. **183**:615–626.

YANG, Z. 1996. Among site variation and its impact on phylogenetic analysis. Trends Evol. Ecol. **11**:367–372.

ZIMMER, E. A., S. L. MARTIN, S. M. BEVERLEY, Y. W. KAN, and A. C. WILSON. 1980. Rapid duplications and loss of genes coding for alpha chains of hemoglobin. Proc. Natl. Acad. Sci. USA **77**:2158–2162.

BARBARA A. SCHAAL, reviewing editor

Accepted December 5, 1997