Error, Bias, and Long-Branch Attraction in Data for Two Chloroplast Photosystem Genes in Seed Plants

M. J. Sanderson, M. F. Wojciechowski,*† J.-M. Hu,* T. Sher Khan,* and S. G. Brady‡*

*Section of Evolution and Ecology, University of California at Davis; †University/Jepson Herbaria and Museum of Paleontology, University of California at Berkeley; and ‡Center for Population Biology, University of California at Davis

Sequences of two chloroplast photosystem genes, *psaA* and *psbB*, together comprising about 3,500 bp, were obtained for all five major groups of extant seed plants and several outgroups among other vascular plants. Strongly supported, but significantly conflicting, phylogenetic signals were obtained in parsimony analyses from partitions of the data into first and second codon positions versus third positions. In the former, both genes agreed on a monophyletic gymnosperms, with Gnetales closely related to certain conifers. In the latter, Gnetales are inferred to be the sister group of all other seed plants, with gymnosperms paraphyletic. None of the data supported the modern ''anthophyte hypothesis,'' which places Gnetales as the sister group of flowering plants. A series of simulation studies were undertaken to examine the error rate for parsimony inference. Three kinds of errors were examined: random error, systematic bias (both properties of finite data sets), and statistical inconsistency owing to long-branch attraction (an asymptotic property). Parsimony reconstructions were extremely biased for third-position data for *psbB.* Regardless of the true underlying tree, a tree in which Gnetales are sister to all other seed plants was likely to be reconstructed for these data. None of the combinations of genes or partitions permits the anthophyte tree to be reconstructed with high probability. Simulations of progressively larger data sets indicate the existence of long-branch attraction (statistical inconsistency) for third-position *psbB* data if either the anthophyte tree or the gymnosperm tree is correct. This is also true for the anthophyte tree using either *psaA* third positions or *psbB* first and second positions. A factor contributing to bias and inconsistency is extremely short branches at the base of the seed plant radiation, coupled with extremely high rates in Gnetales and nonseed plant outgroups.

Introduction

Given a finite amount of data, all phylogenetic methods can be misled. Mistaken inferences about relationships can be more or less random, or, if only certain incorrect topologies are preferred, they can be biased in the context of the underlying process of molecular evolution for those data. At worst, this bias can persist as more and more character data are added, a phenomenon known as statistical inconsistency. Maximum parsimony (MP) can be inconsistent in a simple four-taxon tree in which two long terminal branches are separated by a short interior branch (Felsenstein 1978). This is the source of the term ''long-branch attraction'' (LBA; Hendy and Penny 1989), for which length is understood to mean the expected number of substitutions, a function of rate and time. Maximum-likelihood (ML) and distance methods can also be statistically inconsistent when the assumed model of evolution is incorrect (Chang 1996). Because consistency is an asymptotic property (i.e., one emerging with an infinite number of characters) that is not directly detectable in real data sets, the term ''long-branch attraction'' has colloquially been applied to bias (a property of finite data sets) when branch length heterogeneity is suspected. In this paper, the term ''long-branch attraction'' will be used to refer to conditions under which bias in finite data sets and/or statistical inconsistency arises due to a combination of long and short branches. Unfortunately, despite the evocativeness of the term ''long-branch attraction,'' very little is known about the mathematical conditions under which either bias or statistical inconsistency occurs in phylogenies of more than a few taxa (Hendy and Penny 1989; Huelsenbeck 1995; Kim 1998).

Not surprisingly, teasing apart random error, bias, and inconsistency has proven to be difficult for real data sets (Huelsenbeck 1998). Several methods have been proposed to identify LBA in real data (e.g., Lyons-Weiler and Hoelzer 1997). One method is to use an algorithm that is putatively statistically consistent, such as ML rather than MP (Huelsenbeck 1997). However, in finite (real) data sets, even if ML is consistent and MP is not, the real issue is the extent of bias, rather than the asymptotic behavior as more data are obtained. ML estimates are not guaranteed to be unbiased (Lehmann 1983). Under some model conditions, likelihood methods are not as efficient as nonparametric methods such as parsimony (Huelsenbeck 1998; Siddall 1998). An alternative approach is to use Monte Carlo simulation to examine whether the tree-building method is biased under model conditions that appear appropriate for the given data. Cases that have been studied in this way include the putative attraction of Diptera and Strepsiptera (Huelsenbeck 1998) and of mammals and birds (Huelsenbeck, Hillis, and Jones 1996) and basal relationships of carabid beetles (Maddison, Baker, and Ober 1999) and of yucca moths (Pellmyr and Leebens-Mack 1999). In each of these studies, simulation showed that long branches were sometimes long enough to cause spurious relationships to be reconstructed with high probability.

In this paper, we investigate a possible pattern of LBA in the phylogeny of extant seed plants. Recent molecular work has identified a striking case of conflict

Abbreviations: LBA, long branch attraction; MP, maximum parsimony; ML, maximum likelihood; NJ, neighbor joining.

Key words: statistical consistency, maximum likelihood, parsimony.

Address for correspondence and reprints: Michael J. Sanderson, Section of Evolution and Ecology, One Shields Avenue, University of California, Davis, California 95616. E-mail: mjsanderson@ucdavis.edu.

Mol. Biol. Evol. 17(5):782–797. 2000
© 2000 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

between morphological and molecular data and among different partitions of molecular data in several independent data sets. Living seed plants comprise five groups: angiosperms and four gymnosperm groups—conifers, cycads, Gnetales, and *Ginkgo.* Morphological cladistic analyses had produced a consensus that angiosperms and Gnetales were sister groups to the exclusion of the other seed plants, rendering the extant gymnosperms paraphyletic (Crane 1985; Doyle and Donoghue 1986, 1992; Loconte and Stevenson 1990; Nixon et al. 1994; Rothwell and Serbet 1994; Doyle 1996). Although well supported at the morphological level in bootstrap analyses (J. A. Doyle, personal communication), this ''anthophyte hypothesis'' has not been supported in molecular studies. Instead, recent studies from chloroplast, mitochondrial, and nuclear genomes are converging on a radically different alternative, in which the gymnosperms are monophyletic and the Gnetales are nested within them, sometimes within the conifers themselves (Frohlich and Parker 1999; Graham and Olmstead 1999; Hansen et al. 1999; Ross et al. 1999; Soltis et al. 1999; Winter et al. 1999; Bowe, Coat, and dePamphilis 2000; Chaw et al. 2000). This latter hypothesis would require reevaluation of homologies of conifer and gnetalean reproductive structures, because the angiospermlike ''flowers'' of Gnetales might have been derived from the highly specialized strobili (cones) of conifers. On the other hand, putative morphological synapomorphies exist that might unite conifers and Gnetales, such as features of wood anatomy (Carlquist 1996), and thus the molecular data are in a position to play an important role in developing an understanding the evolution of this dominant and diverse group of land plants.

Although an apparent consensus is emerging from these molecular studies on seed plant relationships, some interesting internal inconsistencies in the molecular data have yet to be fully explored. For example, different but strongly supported phylogenetic inferences emerge from different subpartitions of the data (i.e., first and second vs. third codon positions) and/or different weighting schemes applied to these partitions. The concept of saturation—the assumption that silent substitutions in third codon positions in protein-coding genes are occurring so rapidly that they essentially become randomized (i.e., saturated) compared with slower-evolving first and second codon positions, especially for relatively deep divergences—is often invoked when particular methods or genes fail to produce the expected results or to explain unexpected results. For this reason, third codon positions have commonly been regarded as less reliable and phylogenetically uninformative, if not potentially misleading (Meyer 1994). Some of the recent seed plant studies using both chloroplast and mitochondrial protein-coding genes have presumed that silent substitutions (in third positions) are in fact saturated and should be downweighted (Hansen et al. 1999; Chaw et al. 2000). However, the theoretical justification for downweighting saturated sites, or excluding them altogether from phylogenetic analyses (Meyer 1994; Swofford et al. 1996), is shaky and has been questioned by recent work showing that third codon positions sometimes contain most of the phylogenetic signal in the data (Yoder, Vilgalys, and Ruvolo 1996; Björklund 1999; Källersjö, Albert, and Farris 1999).

Yang (1998) recently studied the effect of overall substitution rate on phylogenetic accuracy and concluded that rates well above those that are conventionally viewed as saturated (ca. 20–30% or greater; Meyer 1994) actually improve accuracy, and that the decline in accuracy that emerges at even higher rates is quite slow. Although Felsenstein (1978) implicated high rates of evolution in LBA problems, these were coupled with rate heterogeneity across branches. The magnitude of saturation needed to induce LBA was dependent on the pattern of rate variation across lineages (or, in the case of constant rates, branch lengths in different lineages; Hendy and Penny 1989). In more complex trees, theoretically grounded guidelines are not available to indicate when saturation really occurs. Because studies of deep phylogeny typically hinge on data that represent mixtures of tempos and modes of evolution even within a single gene (e.g., Chase et al. 1993), decisions about character weighting often have a significant impact on results. Studies of the performance of these heterogeneously evolving genes are greatly needed.

In this paper, we analyze the phylogeny and molecular evolution of two highly conserved chloroplast photosystem genes that demonstrate precisely this pattern of mixed rates. At the amino acid level, these are among the most conserved genes in photosynthetic organisms. At the DNA level, third positions are evolving very rapidly and appear to be saturated in pairwise comparisons across land plants (ca. one substitution per site). Variation in rate across these genes is significant, as is variation across different lineages. We use simulation analysis to study whether and how rate heterogeneity and saturation can lead to errors in phylogeny reconstruction. In particular, we focus on whether third-position data are statistically inconsistent owing to saturation. The main goal of the paper is to assess whether LBA can explain the disparate but strongly supported phylogenetic results that are obtained when third-position data are included in analyses of seed plant relationships, or if it is necessary to seek other explanations for conflict between these data partitions. We also use likelihood methods to characterize variation in rates among lineages (''lineage effects'') and to assess whether this rate variation may be a significant factor affecting phylogenetic inferences in these data.

Materials and Methods

Taxa, Genes, Primers, and Sequencing

Taxa were sampled from all five extant seed plant groups, including angiosperms (*Oryza, Zea, Nicotiana, Pisum, Chloranthus,* and *Drimys*), Gnetales (*Ephedra* and *Welwitschia*), conifers (*Pinus, Araucaria, Torreya,* and *Sequoia*), cycads (*Encephalartos* and *Cycas*), and *Ginkgo.* Vascular plant outgroups included representatives of homosporous eusporangiate ferns (*Angiopteris*) and leptosporangiate ferns (*Adiantum* and *Asplenium*), heterosporous ferns (*Marsilea*), lycopsids (*Huperzia*),

Table 1
Taxon Names, Voucher Information, and GenBank Accession Numbers for Plants Used in this Study

Taxon	Voucher Information[a]	GenBank Accession No.	
		psaA	*psbB*
Adiantum capillus-veneris	UCDBC-B90.225	AF180022	AF222698
Angiopteris evecta	UCDBC-MJS/MFW LP2	AF180020	AF222699
Araucaria araucana	CULT-MJS/MFW LP4	AF180018	AF222701
Asplenium nidus	UCDBC-B94.274	AF180021	—[b]
Chloranthus spicatus	UCDBC-BAA.719	—[b]	AF222709
Cycas taiwaniana	CULT-B73.098	AF180015	AF222697
Drimys winteri	CULT-MJS/MFW LP10	AF180016	AF222708
Encephalartos lebobombensis	CULT-B80.014	AF180011	AF222700
Ephedra tweedyani	UCDA-MJS/MFW LP8	AF180017	AF222702
Equisetum palustre	CULT-MJS/MFW LP3	AF180019	AF222696
Ginkgo biloba	CULT-MJS/MFW LP11	AF223226	AF222705
Huperzia squarrosum	UCDBC-B91.598	AF180024	AF222703
Marchantia polymorpha	CCP-GB	X04465	X04465
Marsilea botrycarpa	UCDBC-B97.438	AF180014	—[b]
Nicotiana tabacum	CCP-GB	Z00044	Z00044
Oryza sativa	CCP-GB	X15901	X15901
Pinus thunbergii	CCP-GB	D17510	D17510
Pisum sativum	Wojciechowski 398	AF223227	AF222710
Psilotum nudum	UCDBC	AF180023	AF222707
Sequoia sempervirens	UCDA	AF180012	—[b]
Torreya californica	UCDA	AF180025	AF222706
Welwitschia mirabilis	UCDBC-MJS/MFW LP15	AF180013	AF222704
Zea mays	CCP-GB	X86563	X86563

[a] Source and voucher number. Source abbreviations: UCDBC, University of California–Davis Botanical Conservatory; UCDA, University of California–Davis Arboretum; CULT, cultivated at UCD; CCP-GB, data from complete chloroplast genomes deposited in GenBank (collection information available from GenBank).
[b] No sequence obtained for this taxon and gene.

sphenopsids (*Equisetum*), and *Psilotum.* A nonvascular plant, the liverwort *Marchantia,* was used to root the tree. Except for *Marchantia, Pinus,* and several angiosperms, all sequences reported here are new. Taxon names, voucher information, and GenBank accession numbers are listed in table 1.

Table 2
Primers Used for Both PCR and DNA Sequencing of the Chloroplast DNA *psaA* and *psbB* Loci

Loci	
psaA	
psaA1	ATTCGTTCGCCGGAACCAGA
psaA1a	CTATTCGTTCGCCGGAACCAGA
psaA552	AGCTGCTCCAAAATTRGCTTGG
psaA700	GATCCTAAAGAGATACCACTTCC
psaA997	GCTCATAAAGGTCCATTTACRGG
psaA1020R	CCYGTAAATGGACCTTTATGAGC
psaA1392	TGATACYATGAGTGCTTTAGGAC
psaA1415R	GTCCTAAAGCACTCATRGTATCA
psaA1854R	CCCCAAACATCNGACTGCATTTTCC
psaA2R	GTTGTGGCAATTCCACCCAGAA
psbB	
psbB1	GGGTTTGCCTTGGTATCGTGTTCAT
psbB471	TGTAACRGGTTTGTATGGTCCTGG
psbB495R	CCAGGACCATCAAACCYGTTACA
psbB993	GGACAATGGAGATGGAATAGC
psbB1014R	GCTATTCCATCTCCATTGTCC
psbB4R	AGCCCCATGCCAAATGTGTC

Note.—Primers are forward primers (i.e., primers that match the coding strand) except for those with ''R's'' in their names, which are reverse primers that match the noncoding strand (i.e., anneal to the coding strand). All primer sequences are shown 5′ to 3′. Numbers represent approximate nucleotide positions downstream from the 5′ end (except for 2R and 4R labels). Ambiguous nucleotides follow the IUBMB code.

Sequences were obtained from two chloroplast genes, *psaA* and *psbB,* encoding thylakoid membrane-bound structural proteins functioning in the chloroplast photosystems I and II, respectively (Ort and Yocum 1996). Nondegenerate primers for PCR were designed to match conserved sequences at the 5′ and 3′ ends of each gene by comparisons of complete, aligned coding sequences of the *psaA* and *psbB* genes obtained from complete chloroplast genomes of *Nicotiana, Zea, Oryza, Pinus,* and *Marchantia.* Internal primers for sequencing were designed in the same way (some with slight degeneracy) but supplemented by comparisons of new sequences from other taxa as they became available during the study. Primers were designed to provide redundancy (i.e., sequence overlap) in both the forward and the reverse directions. Primer sequences for both genes and their positions are shown in table 2. All primers were obtained from Operon Technologies (Alameda, Calif.). Double-stranded DNA copies of the genes were amplified by standard PCR methods using the following thermal cycler conditions: 3 min at 94°C; 35–40 cycles of 45 s at 94°C, 45 s at 52–58°C, and 3 min at 72°C; followed by a final 7-min incubation at 72°C. The PCR products were purified by ultrafiltration (Millipore Ultrafree-MC tubes) and sequenced directly by automated fluorescent dye sequencing methods on an ABI model 377 at the University of California–Davis Division of Biological Sciences Sequencing Facility. Sequences from different primer reactions were analyzed and assembled into ''contigs'' using Sequencher, version 3.0 (GeneCodes Corp., Ann Arbor, Mich.). DNA sequences were translated and checked against the highly con-

served amino acid sequences available for these proteins from reference taxa listed above. Because of the extreme conservation in these genes, sequence alignment was not an issue. Only a few small indels were observed across all vascular plants. These were not included as separate characters in the data sets.

Phylogenetic Analysis

Trees were reconstructed using MP, ML, and neighbor-joining (NJ) algorithms, using PAUP*, version 4.0b2 (Swofford 1999), for nucleotide data and MOLPHY, version 2.3b3 (Adachi and Hasegawa 1996), for amino acid translations. The most extensive analyses used the nucleotide data. Two partitions of the nucleotide data were established, comprising (1) first and second codon positions and (2) third positions. According to the genetic code, 30% of third-position changes in codons can cause amino acid changes (are nonsynonymous; Li 1997), but the extremely low rate of nonsynonymous changes observed in these amino acid sequences (about 10–20 times as low as that for third positions—see *Results*) means that, to a very good approximation, third-position changes are entirely dominated by synonymous changes, and they will be interpreted as such throughout. Partition homogeneity tests (Farris et al. 1995) were conducted using parsimony to test whether significant conflicting signals were present between partitions for each gene. Separate and combined analyses for these codon partitions were conducted for each gene. Some analyses were conducted on data sets formed by concatenating the two genes. Confidence limits were estimated by bootstrapping (Felsenstein 1985).

MP searches used heuristic search strategies, consisting of ASIS addition sequences followed by tree bisection- reconnection (TBR) branch swapping (Swofford et al. 1996). ML searches used the HKY85+Γ substitution model. Because ML using this model required about 3,500 times as much computation time as MP, a series of heuristic procedures were used to permit ML bootstrap runs to be implemented. The transition/transversion ratio and the shape parameter of the gamma distribution were first estimated on MP trees separately for each gene and codon position. These parameters were then fixed in the ML searches using the PREVIOUS command, and ML heuristic searches were run with ASIS addition sequences and NNI branch swapping. Although more exhaustive search procedures such as TBR swapping could be used to search for optimal trees, running times were prohibitive when repeated across 100 bootstrap replicates. Another analysis used the SITERATES facility in PAUP* to permit a different rate of evolution in the two codon partitions. This facility estimates different fixed rates in the different partitions and should help handle any heterogeneity between partitions. However, it does not simultaneously allow rate variation across sites within the partitions (e.g., with a gamma distribution). Bootstrap runs were implemented with the same options as described above but were done with data sets consisting of the two genes concatenated together.

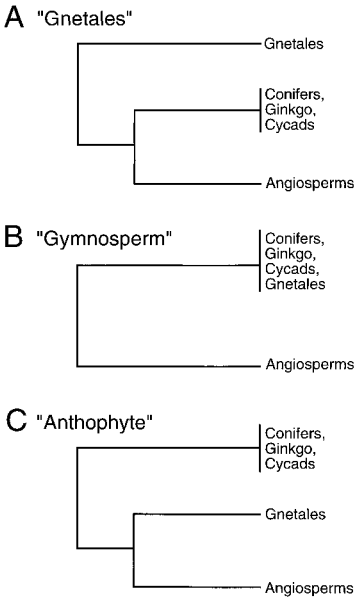

FIG. 1.—The three main hypotheses of seed plant relationships discussed in this paper, showing major groups: angiosperms (Ag); conifers, *Ginkgo*, cycads (CGC), and Gnetales (Gn). *A,* The ''Gnetales hypothesis.'' *B,* The ''gymnosperm hypothesis.'' *C,* the ''anthophyte hypothesis.'' Throughout the paper, these will be referred to as trees (or hypotheses) A, B, and C.

MP analyses of the amino acid data used step matrices based on the inferred minimum number of substitutions required to transform one amino acid into another. ML and NJ analyses of the amino acid data used the Dayhoff substitution model (Dayhoff, Schwartz, and Orcutt 1978). ML searches used the ''quick'' heuristic strategy in MOLPHY. Finally, because of shifts in base composition across taxa, some neighbor-joining analyses were performed with the log-det distance transformation, which may be relatively robust to composition changes across the tree (Lockhart et al. 1994).

Estimates of divergence (expected numbers of substitutions) along branches were obtained by ML for each gene and codon position partition separately on each of a set of three trees which formed the basis for subsequent analyses of bias, error, and LBA. The three trees were chosen so as to satisfy three major hypotheses about seed plant relationships that have been discussed in the morphological and molecular literature on seed plant relationships (fig. 1). Two of these three major hypotheses are consistent with trees obtained with different partitions of the present data sets. The third can be obtained easily by imposing one phylogenetic constraint on the present data. Thus, tree A consisted of one MP tree estimated from third positions only for a given gene. For both genes, this tree has Gnetales as the sister group of the other seed plants. When discussing seed plant relationships per se in the context of tree A, we use the term ''Gnetales hypothesis.'' Tree B consisted of one MP tree estimated from first and second positions only. For both genes, this tree has a monophyletic gymnosperms as sister to angiosperms (hence, the ''gymnosperm hypothesis''; note that this arrangement is not found on the bootstrap tree for *psaA*). Tree C is one MP

tree found under the constraint tree corresponding to the anthophyte hypothesis of seed plant relationships (described above), obtained from a data set combining both codon partitions. Model parameters, including shape parameter and transition/transversion (ti/tv) ratio (unlike for the ML searches), were all estimated from the data under an HKY85+Γ model.

Analysis of Reconstruction Error, Bias, and Long-Branch Attraction

We studied three properties of parsimony reconstruction in the context of these data sets: error, bias, and statistical consistency. The first two are properties associated with finite data sets. Error refers to the overall probability of obtaining an incorrect tree, averaged over replicate samples of the same size as the original data set. Conceptually, we can think of these as samples from genes of the same size evolving according to the same model as that governing the actual gene. This immediately suggests a Monte Carlo simulation strategy for estimating error (Manly 1997). Bias refers to a preference of the tree algorithm to erroneously reconstruct one or a few particular trees, rather than a set of trees randomly distributed around the true tree. Bias can also be studied by Monte Carlo simulation. Statistical consistency refers to the convergence of the reconstructed tree to the true tree as character data are added. Simulation can be used in this problem too, but there are limits to the inferences that can be made, because one cannot simulate an infinite sample of characters (unless simulation is used in conjunction with analytical results). Instead, progressively larger samples are simulated to identify likely asymptotic behavior.

A simulation protocol was set up to identify the error rate for recovering each of the three main hypotheses of seed plant relationships under model conditions estimated from the data. The three relationships tested were the Gnetales, gymnosperm, and anthophyte hypotheses (fig. 1), ignoring relationships outside of seed plants or within angiosperms or the other taxa. For example, we might be interested in estimating the probability of rejecting the anthophyte hypothesis if it indeed is true—the type I error. To test this sort of hypothesis, the tempo and mode of evolution of a codon partition of one of the genes, say, third positions in *psaA*, was estimated on tree C, which represents the null hypothesis. Parameter estimates of the substitution model (composition, ti/tv ratio, shape parameter) were obtained via ML, along with the estimates of overall branch lengths for each branch. Then, 1,000 new data sets were generated using this tree as input, along with estimated branch lengths and parameters, using the program Seq-Gen, version 1.1 (Rambaut and Grassly 1997). Finally, the MP solution for each replicate data set was obtained and checked to see how often the three hypotheses of seed plant relationships entailed by trees A, B, and C (fig. 1) were found. Type II error rates (the probability of mistakenly accepting the null hypothesis when it is false) were obtained in the same fashion. For example, in the case of the anthophyte hypothesis, the type II error rate is the probability of incorrectly reconstructing the anthophyte hypothesis given that one of the two other hypotheses (A or B) is correct.

Together, type I and type II errors can be summarized in a three-by-three matrix in which the *ij*th element gives the probability of obtaining the seed plant relationships entailed by hypothesis *i* given that tree *j* is in fact correct (where *i, j* is one of A, B, or C in fig. 1). The *ii*th (diagonal) element is just 1–type I error on hypothesis *i*. The *ij*th (off-diagonal) elements give the type II error for hypothesis *i* given the alternative hypothesis, *j*. Under ideal circumstances, hypothesis *i* should be supported when Tree *i* is in fact true. Therefore, the *ii*th (diagonal) elements of the matrices should be high (low type I error rate) and the *ij*th (off-diagonal) elements should be low, indicating low type II error rate. On the other hand, if some off-diagonal elements are high, then trees other than the true tree have a high chance of (mistakenly) being inferred. Bias is indicated by a skew toward increasing probability of some of the off-diagonal elements but not all of them.

The simulations just described all entail generating data sets with the same number of characters as that observed with the real data. This permits assessments of sampling error and bias. In addition, simulations were undertaken to make inferences about the statistical consistency of tree reconstruction with the given data sets. Selected tests along the lines described above for type I error were repeated with progressively larger numbers of characters. This was easily implemented by changing a single parameter in Seq-Gen and repeating the experimental protocol described above. Since statistical consistency is an asymptotic property, it was not strictly possible to assess it by the finite-sized simulation implemented here, because it is always possible that a seemingly monotonic trend in one direction or the other might reverse itself at some point in simulations larger than those performed (Kim 1998). However, we assume that the results from analyses of large numbers of characters are suggestive.

Experiments were also performed to test the effects of saturation in some data sets. Overall rates of sequence evolution were modified by multiplying the treewide rate estimated from the data by a constant (another Seq-Gen option). This permitted the type I and type II error rates to be estimated as a function of increasingly higher rates of substitution above that observed in the real data, to address the question of what levels of saturation are needed to obtain a high error rate in the context of the given topology and relative branch lengths.

Results

DNA Sequences

Primers described in table 2 permitted approximately 2,155 bp of *psaA* (out of 2,265 bp) (716/755 codons) and 1,415 bp of *psbB* (out of 1,530 bp) (471/510 codons) to be sequenced, with high overlap on one or both strands. We were unable to obtain good sequence for two ferns and one seed plant in the *psbB* data set (*Asplenium, Marsilea,* and *Sequoia*), but two

Table 3
Summary Statistics on Sequence Data from *psaA* and *psbB*

	PSA*A* (2,155 bp, 22 taxa)		PSB*B* (1,415 bp, 20 taxa)	
STATISTIC	1st + 2nd Positions	3rd Position	1st + 2nd Positions	3rd Position
G + C fraction	0.472	0.315	0.513	0.281
Base frequency homogeneity test across taxa				
χ^2	11.445	276.67	4.53	203.96
df	63	63	57	57
P	1.000	0.000	1.000	0.000
Ti/tv[a]				
Tree A	3.122	5.033	1.774	4.468
Tree B	3.013	5.127	1.816	4.612
Tree C	3.130	5.071	1.772	4.488
Estimated gamma shape parameter[a]				
Tree A	0.128	1.367	0.128	1.258
Tree B	0.125	1.293	0.155	1.248
Tree C	0.127	1.356	0.128	1.248

NOTE.—See figure 1 for trees. GC content and base frequency tests are independent of topology.

[a] Transition/transversion ratio and shape parameter of gamma distribution estimated using maximum likelihood under an HKY85+Γ model of substitution for trees indicated.

ferns and three other vascular nonseed plants were still obtained for use as outgroups. The sequence for *Chloranthus*, an angiosperm, was not obtained for *psaA*. This yielded a data set of 22 taxa for *psaA* and 20 taxa for *psbB*, with 19 taxa shared in a combined data set. For alignments, see *Supplementary Materials*. Table 3 shows basic summary statistics for these sequences. As with most chloroplast genes, AT content is quite high. Moreover, base compositional heterogeneity among taxa appears to be significant for both genes in the third-position data ($P = 0.0$ in a χ^2 test of heterogeneity), but not in first and second positions ($P = 1.0$). Transition/transversion ratios and shapes of the pattern of rate variation across sites are also different between the third-position partition and the first- and second-position partition. Rate variation is discussed further below.

Phylogeny

Nucleotide Data

A partition homogeneity test indicated significantly different signals in third positions versus first and second positions in both the *psaA* and the *psbB* data sets ($P < 0.024$ and $P < 0.002$, respectively). On the other hand, no significant difference was found between genes for the entire data sets ($P = 0.77$), for first and second positions alone ($P = 0.31$), or for third positions alone ($P = 0.85$). Thus, there is much more conflict within each gene between codon partitions than there is between the two genes within the same partition. This is reflected in the very different phylogenies reconstructed from the two codon partitions and in the high support for these conflicting results (figs. 2 and 3). However, results from analyses combining all codon positions are identical to the third-position trees because of the larger

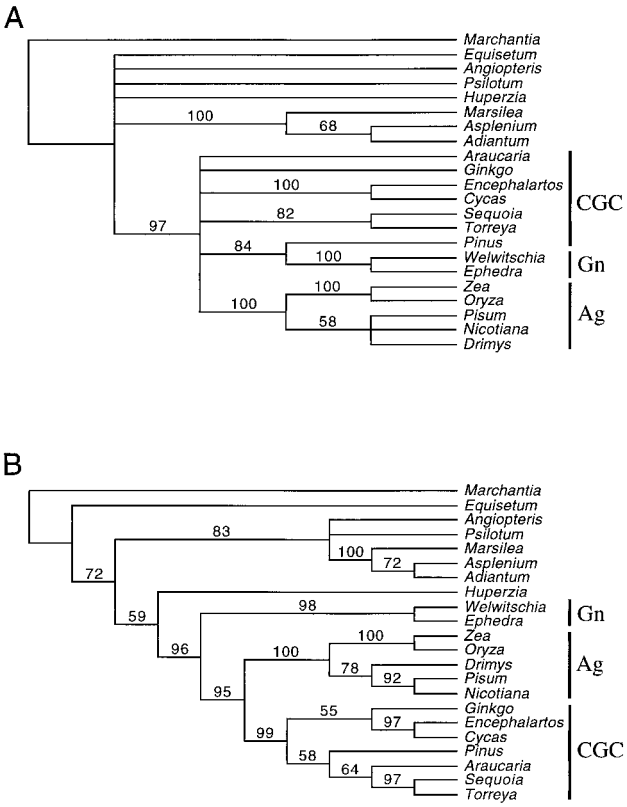

FIG. 2.—Bootstrap majority rule tree from maximum-parsimony analysis of *psaA* sequences. Numbers over branches are bootstrap support values. Phylogenies are based on analyses of (*A*) first and second positions or (*B*) third positions only. In this and subsequent figures, brackets identify major groups of seed plants.

number of parsimony-informative sites at third positions. With respect to the three trees corresponding to the three basic hypotheses about seed plant relationships (fig. 1), Kishino-Hasegawa tests indicate that the two suboptimal trees were generally significantly worse than the optimal third tree (which is either tree A or tree B) with respect to the parsimony length (within data partitions). With respect to likelihood scores, all of the *psaA* first- and second-position trees were indistinguishable, as were the *psbB* third-position trees, but the other two combinations showed significant differences (table 4).

In MP analyses, the monophyly of angiosperms and of seed plants was supported in all partitions and both genes. However, the relationships within seed plants and among the other vascular plants varied. For the first and second positions, both genes agreed on a placement of Gnetales as the sister group to *Pinus*, one of the conifers included in the analysis. In *psbB*, gymnosperms were a clade (93% support), and the Gnetales were nested within a now-paraphyletic conifers (fig. 3*A*). In *psaA*, this sister group relationship with *Pinus* was supported at the 84% level, but seed plant relationships overall were not well resolved, and there was not strong support for gymnosperms as a clade (fig. 2*A*). For the third-position data, both genes agreed on the unusual placement of Gnetales as the sister group of all other seed plants, with the latter supported at the 95% and

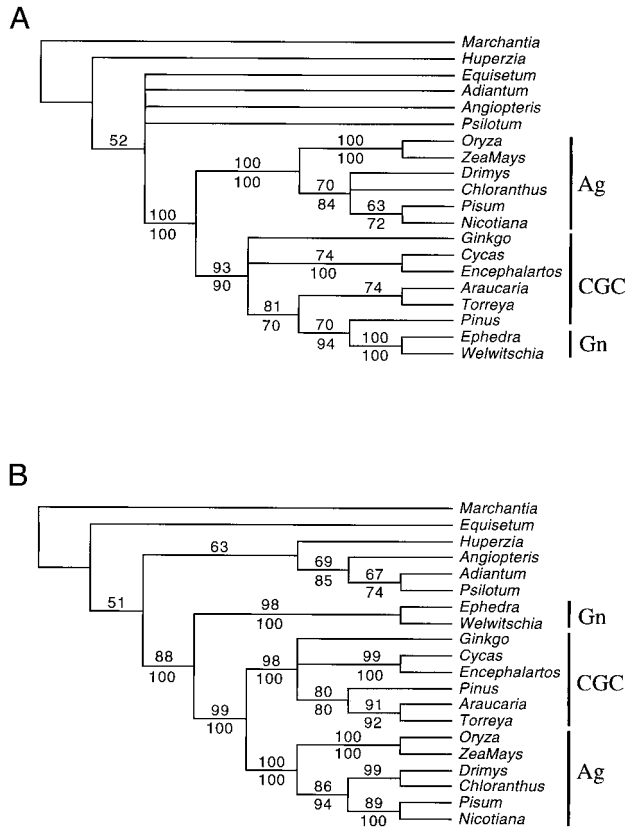

FIG. 3.—Bootstrap majority rule tree from maximum-parsimony analysis of *psbB* sequences. Numbers over branches are bootstrap support values for *psbB* data alone, while numbers below branches refer to bootstrap support values for nodes also present (>50% bootstrap support) in MP analyses of concatenated *psaA* and *psbB* data. Phylogenies are based on analyses of (*A*) first and second positions or (*B*) third positions only.

99% bootstrap levels for *psaA* and *psbB*, respectively (figs. 2*B* and 3*B*). The clade consisting of conifers, *Ginkgo*, and cycads was supported at the 99% and 98% levels, but relationships within that clade were not so clear (conifers are not even monophyletic in *psaA*).

MP trees obtained after concatenating the *psaA* and *psbB* sequences together (but maintaining the position partitions separate) were essentially identical to the *psbB* tree with some changes in support for clades (bootstrap values are shown superimposed on the *psbB* tree of fig. 3). In data from first and second positions, the support for the monophyly of conifers was lowered, but the support for a placement of Gnetales with *Pinus* was improved to 94%. In the third-position data, support was generally improved throughout the tree, reflecting the strong agreement between the two genes. Thus, one partition placed the Gnetales as the sister group to all other seed plants with a 100% bootstrap level, whereas the other partition nested it within conifers at the 94% level, a striking conflict.

ML results were reported as bootstrap majority rule trees (figs. 4 and 5). For *psaA*, little resolution of relationships within seed plants was apparent, although in both partitions there was some support for Gnetales being nested with conifers (71% bootstrap support for first and second positions; 81% for third positions). For *psbB* data, first and second positions showed modest support (75%) for Gnetales nested within conifers, and contradictory but modest support for third-position data for Gnetales as sister group to all other seed plants (75%). In the concatenated data, the optimal trees (this time with TBR swapping) both in first and second positions and in third positions had Gnetales nested within conifers (trees not shown). Results from runs assuming fixed but different rates in the two codon partitions (SITERATES option, also with concatenated data sets) supported the Gnetales hypothesis (tree not shown). The bootstrap tree indicated a paraphyletic gymnosperms, with Gnetales as the sister group of other seed plants (100% support) and cycads as the sister group of angiosperms (65% support only), with the latter being a result not seen in other experiments with these data.

Because the third-position partitions showed significant compositional changes across the trees, neighbor-joining analyses with the log-det transformation were performed. For comparative purposes, we also

Table 4
Kishino-Hasegawa Tests Comparing Three Different Hypotheses of Seed Plant Relationships Under the Null Hypothesis of No Difference Between the Two Trees

	Positions 1 and 2				Position 3			
	MP[b]		ML[c]		MP		ML	
Tree[a]	Length	*P*	−ln *L*	*P*	Length	*P*	−ln *L*	*P*
psaA								
A	675	<0.05*	5,652	0.15	2,608	—	10,397	—
B	656	—	5,629	—	2,712	<0.01*	10,481	<0.01*
C	680	<0.01*	5,657	0.07	2,615	0.16	10,400	0.30
psbB								
A	390	<0.01*	3,572	<0.01*	1,471	—	6,098	—
B	366	—	3,521	—	1,547	<0.01*	6,108	0.18
C	389	<0.01*	3,569	<0.05*	1,488	<0.01*	6,101	0.43

[a] See figure 1.
[b] Maximum parsimony; Kishino-Hasegawa test using parsimony scores.
[c] Maximum likelihood; Kishino-Hasegawa test using likelihood scores under an HKY85+Γ model of substitution. *L* = likelihood.
* 0.05 significance level.

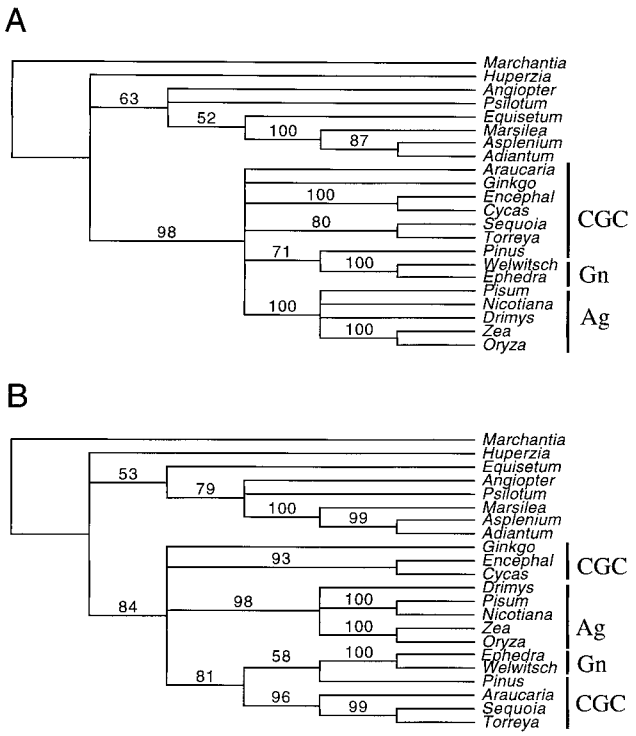

FIG. 4.—Bootstrap majority rule tree for maximum-likelihood analyses of *psaA* sequences. Numbers over branches are bootstrap support values. Phylogenies are based on analyses of (*A*) first and second positions or (*B*) third positions only.

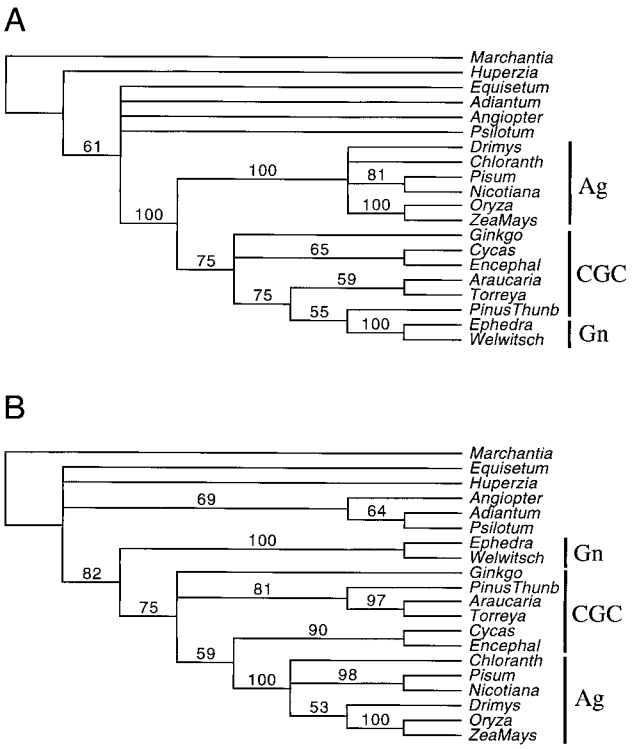

FIG. 5.—Bootstrap majority rule tree for maximum-likelihood analyses of *psbB* sequences. Numbers over branches are bootstrap support values. Phylogenies are based on analyses of (*A*) first and second positions or (*B*) third positions only.

conducted neighbor-joining analyses with a more standard Kimura two-parameter (K2P) model. Rate variation across sites was specified using parameters estimated by ML (table 3). These neighbor-joining runs produced nearly the same trees as the MP runs for the same gene partition data set. The only significant difference was that the *psaA* first- and second-position data had a paraphyletic gymnosperms, but still placed Gnetales as the sister group of *Pinus* (trees not shown). Neighbor-joining analyses with just the third-position data and the log-det transformation were quite unresolved and poorly supported at the level of seed plants. This may be due to two factors. First, log-det requires the estimation of many more parameters in the transformation matrix than do simpler models of substitution. Estimation of more parameters generally leads to more error variance in the estimation of each (Zharkikh 1994). Second, log-det distance methods do not permit specification of site-to-site variation in rates, which seems to be pervasive in these data (table 3 and fig. 6), because of theoretical problems arising with nonstationary base composition and site-to-site rate variation (Bakke and von Haeseler 1999).

Amino Acid Data

Results of analyses using amino acid translations of the nucleotide sequences were largely the same as those obtained with the first- and second-position nucleotide data (trees not shown). A bootstrap analysis of the concatenated data using weighted parsimony with an amino acid step-matrix reconstructed the same relationships as with the nucleotide data, but with slightly lower bootstrap support for some clades. The monophyly of gymnosperms was only supported at the 64% level, and that of *Pinus*-Gnetales was supported at the 87% level. A neighbor-joining tree based on Dayhoff distance matrix calculations (in MOLPHY) also had a gymnosperm clade and a *Pinus*-Gnetales clade, but conifers included *Ginkgo* and cycads and were thus paraphyletic. Finally, an ML tree of the amino acid data also indicated that gymnosperms were a clade, with conifers plus Gnetales as a clade nested within it (trees not shown).

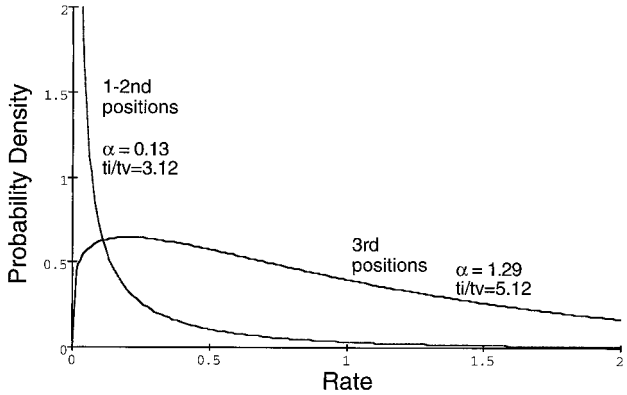

FIG. 6.—Shapes of gamma distributions of rate variation across sites for codon partitions in *psaA* for tree A (see table 3). Note that the rate scale is normalized so that the mean is 1.0. The mean absolute rates of first and second positions versus third positions are actually different by roughly an order of magnitude. The graph for *psbB* is very similar (see parameter estimates in table 3).

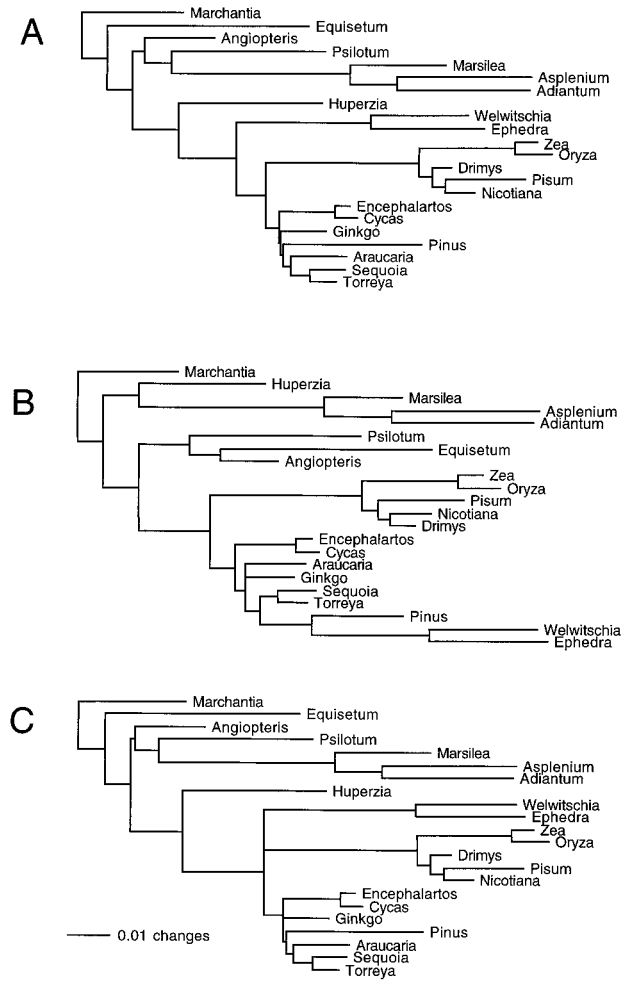

FIG. 7.—Branch lengths estimated for *psaA* (first and second positions only) using maximum likelihood given the three model trees depicted in figure 1 for hypotheses of seed plant relationships. Panels are labeled A–C to correspond to the hypotheses of figure 1.

FIG. 8.—Branch lengths estimated for *psaA* (third positions only).

Rates, Error, Bias, and Long-Branch Attraction

Rate Parameters and Branch Lengths

Transition/transversion ratios were higher for third positions than for first and second positions but did not differ greatly between genes (table 3). Rate variation at the nucleotide level was characterized by the shape parameter of a gamma distribution fit to the data via ML. In both genes, the shape of the distribution was highly left-skewed for first- and second- position data (indicating most sites are highly conserved), whereas the shape for third-position data was modal (fig. 6). Most pairwise distances between taxa (Kimura two-parameter corrected) fell in the range of 0.03–0.07 substitutions per site for both genes at first and second positions, but ranged from 0.5 to 1.0 at third positions, suggesting a rate 10–20 times as high.

Branch lengths (expected numbers of substitutions per site) estimated by ML exhibited striking differences across lineages (figs. 7–10). Many of the patterns in these lineage effects were largely independent of which of the three phylogenetic hypotheses were used in their estimation and seemed to be correlated between genes and partitions. For third-position data (e.g., fig. 8), several outgroups had exceptionally long branches (*Equisetum* and *Adiantum*), as did the Gnetales (*Ephedra* and *Welwitschia,* as well as their stem lineage) among seed plants. The stem lineage subtending angiosperms was relatively long compared to that of cycads, *Ginkgo,* and conifers, which in fact had quite short branches in all three trees. This pattern was also found in first and second positions, although branch lengths for *Equisetum* were shorter.

Other branches were exceptionally short. A notable example occurs in the C trees (the anthophyte hypothesis). For both genes and both position partitions, there was a trichotomy at the base of seed plants involving angiosperms, Gnetales, and a clade consisting of the other three seed plant groups (conifers, cycads, and *Ginkgo*). This situation reflects estimation of a zero-length branch supporting the anthophytes (Gnetales plus angiosperms). Other short branches were typically found toward the base of the gymnosperm clade in the B trees and toward the base of the conifer/cycad/*Ginkgo* clade in the A trees, perhaps reflecting the very low rates for these latter three seed plant taxa.

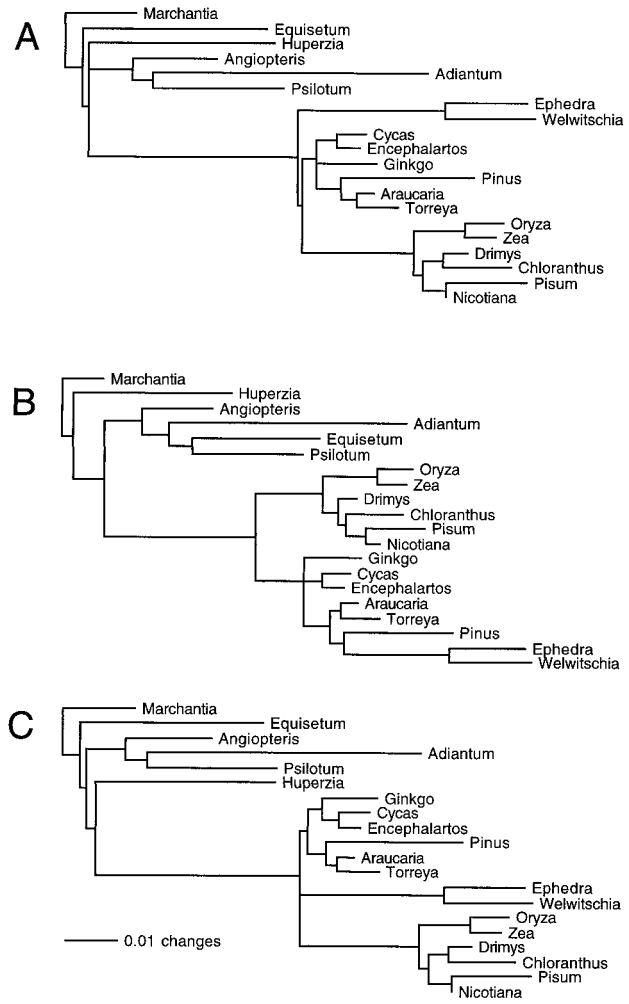

FIG. 9.—Branch lengths estimated for *psbB* (first and second positions only) using maximum likelihood given the three model trees depicted in figure 1 for hypotheses of seed plant relationships. Panels are labeled A–C to correspond to the hypotheses of figure 1.

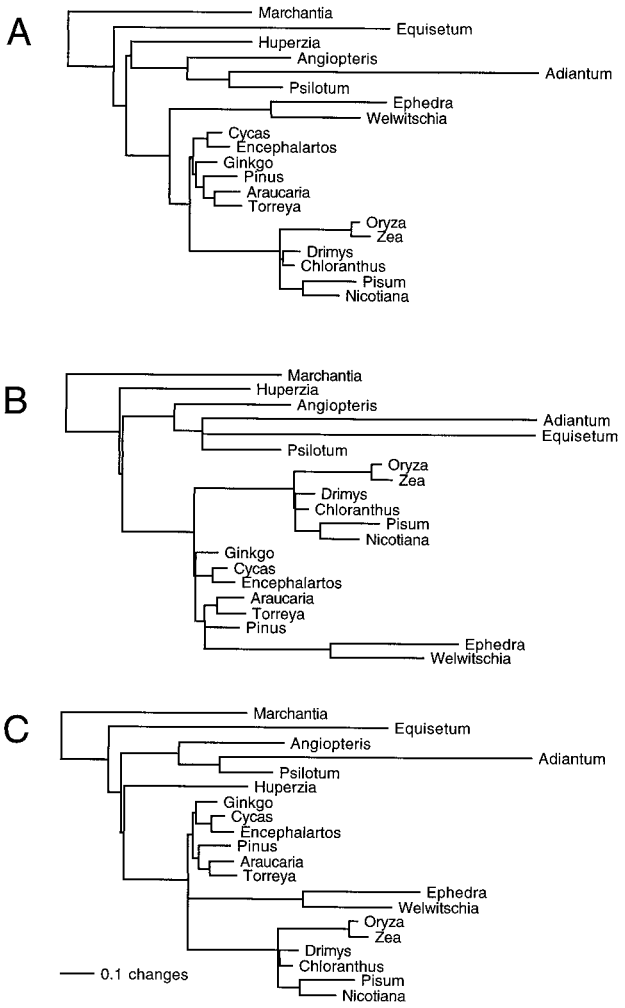

FIG. 10.—Branch lengths estimated for *psbB* (third positions only).

Simulations of Error Rates and Bias

Table 5 shows results from simulation experiments of the error rates of tree reconstructions. Results differed importantly between the two genes. For *psaA*, the type I and type II error rates were uniformly low if either tree A or tree B was correct for either position partition. For example, if tree A were correct, tree A would be estimated by MP with high probability (0.887 and 0.926, respectively, for the two codon partitions). The same holds for tree B. However, if tree C were the true tree and third-position data were used, chances are 0.666 that tree A would be mistakenly reconstructed (note that tree A actually was reconstructed from those data). This indicates a bias in the reconstruction. If first and second positions were used instead and tree C were correct, chances are about even that any of the three trees might be reconstructed (probabilities ranging from 0.273–0.386). Hence, for *psaA*, tree C is not easy to estimate if it is true; the other trees are.

The situation was more complex for *psbB*. Approximately the same high error rate was associated with tree C, as was seen for *psaA*, but there are also problems with tree B for third-position data. If tree B is correct, there is at least a 71.9% chance that a sequence evolving with the tempo and mode of the third-position partition would falsely cause an inference of tree A (which was indeed found with those data). Thus, there is a clear bias in data from *psbB* third positions toward reconstructing tree A regardless of which of the three trees is true.

The bias toward tree A when tree B is correct does not occur with third-position data for *psaA*. Simulation of progressively higher rates of substitution over and above the observed rates indicated that rates would have to be about eight times as high as the already-saturated rates observed (0.5–1.0 substitutions per site) for this bias to appear, meaning that, effectively, this gene's level of saturation is nowhere close to what is necessary to generate positively misleading levels (fig. 11).

Simulations on Statistical Consistency

The probability of correct tree reconstructions should decrease with additional character data if a method is truly statistically inconsistent. Figure 12 shows the

Table 5
Estimated Error Rates for Parsimony Inference Given Specified Tree Topologies

	Gnetales Hypothesis (tree A)	Gymnosperm Hypothesis (tree B)	Anthophyte Hypothesis (tree C)
psaA, 1st and 2nd position			
Tree A	0.887	0.022	0.090
Tree B	0.000	0.971	0.000
Tree C	0.308	0.273	0.386
psaA, 3rd positions			
Tree A	0.926	0.022	0.051
Tree B	0.142	0.840	0.017
Tree C	0.666	0.168	0.165
psbB, 1st and 2nd positions			
Tree A	0.577	0.233	0.126
Tree B	0.000	0.989	0.000
Tree C	0.409	0.252	0.228
psbB 3rd positions			
Tree A	0.945	0.013	0.041
Tree B	0.719	0.151	0.125
Tree C	0.668	0.149	0.182

NOTE.—Given the tree indicated in the leftmost column, these values are the probabilities of obtaining trees that satisfy each of the three hypotheses of relationships among seed plants shown in figure 1. Diagonal elements are 1–type I error probabilities; off-diagonal elements are type II error probabilities.

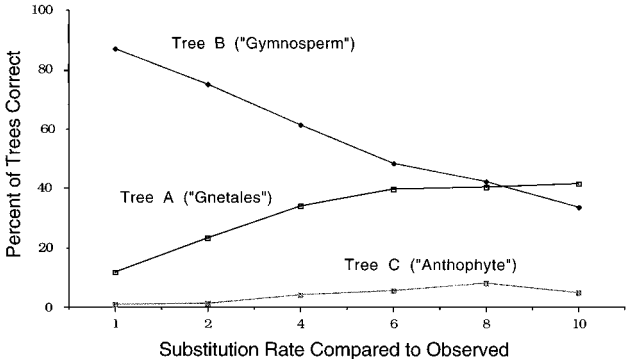

FIG. 11.—Graph of putative saturation effects for *psaA* third-position data. Each line represents the probability of obtaining a hypothesis given that the hypothesis is true (1–type I error probability) for the three hypotheses of seed plant phylogeny (fig. 1), with an increasing rate of substitution, up to 10 times that observed for the *psaA* data set.

behavior of these probabilities in simulations of increasingly large data sets for hypotheses (trees) A, B, and C. Once again, the patterns differ between the two genes. For first and second positions in *psaA*, increasing the number of characters merely improves the probability of obtaining the correct tree and lowers the probability of obtaining an incorrect tree (lowers type I error). Even for tree C, which is reconstructed accurately with fairly low probability in a data set of the size of the real data set ($P = 0.39$; table 5), the accuracy improves steadily with increasing numbers of characters (although it may approach an upper limit below 1). For *psaA* third positions, on the other hand, if hypothesis C is correct, reconstruction using MP is inconsistent because the probability of obtaining C decreases with increasing numbers of characters. Precisely the same pattern holds for first and second positions of *psbB*; reconstruction under hypothesis C is inconsistent. Finally, for third positions of *psbB*, reconstruction under both hypothesis B and hypothesis C is inconsistent. The most likely mistaken tree for all cases of inconsistency described here is tree A.

Discussion

Phylogeny

The strongest conclusion from this study is that in parsimony analyses, there was a striking conflict between signals from different partitions of the two chloroplast genes but not within the same partitions between genes. Trees based on first- and second-position data were well supported but conflicted dramatically with well-supported trees based on third positions (or based on a combination of both partitions). The former implies the gymnosperm hypothesis; the latter, the Gnetales hypothesis (fig. 1). Neither partition supported the modern anthophyte hypothesis that Gnetales are the sister group of angiosperms (Doyle 1996). High bootstrap values for the conflicting trees and significant partition homogeneity tests mean that it is not simply the case that one of the partitions is random noise and the other has signal. Both partitions have strong signals, and they conflict. This suggests the existence of statistical bias in one or both of the partitions.

Results from likelihood analyses were the same for *psbB*, but the conflict between partitions was much less evident for *psaA*. Both partitions supported Gnetales with conifers, although neither partition was sufficiently well supported to suggest the monophyly of gymnosperms. A hypothesis of gymnosperm monophyly and a sister group relationship between Gnetales and conifers does not require a radical reinterpretation of morphological evolution in seed plants. However, a placement of Gnetales within conifers, which is the typical signal emerging from first and second positions in our data (figs. 2*A*, 3*A*, 4*A*, 4*B*, and 5*A*), requires both a major modification of reproductive morphology and the reacquisition of one copy of a large inverted repeat in the chloroplast genome, which conifers have lost but angiosperms, Gnetales, and all other seed plants have retained (Raubeson and Jansen 1992).

The two different signals we observed in the photosystem genes are each corroborated by other studies. Data from *rbc*L had long supported the third-position tree (Hasebe et al. 1992; Albert et al. 1994), and a recent reanalysis of these data only supported the tree based on first and second positions when third positions were downweighted (Chaw et al. 2000). Studies of 18S rDNA (Chaw et al. 1997; Soltis et al. 1999) and chloroplast ITS sequences (Goremykin et al. 1996) had suggested the monophyly of gymnosperms, but this was widely questioned. Recently, several studies have once again argued for this result, including the remarkable relationship of Gnetales to conifers. Hansen et al. (1999), using a 10-kb piece of the chloroplast genome of *Gnetum* to compare with complete genomes of *Pinus* and several

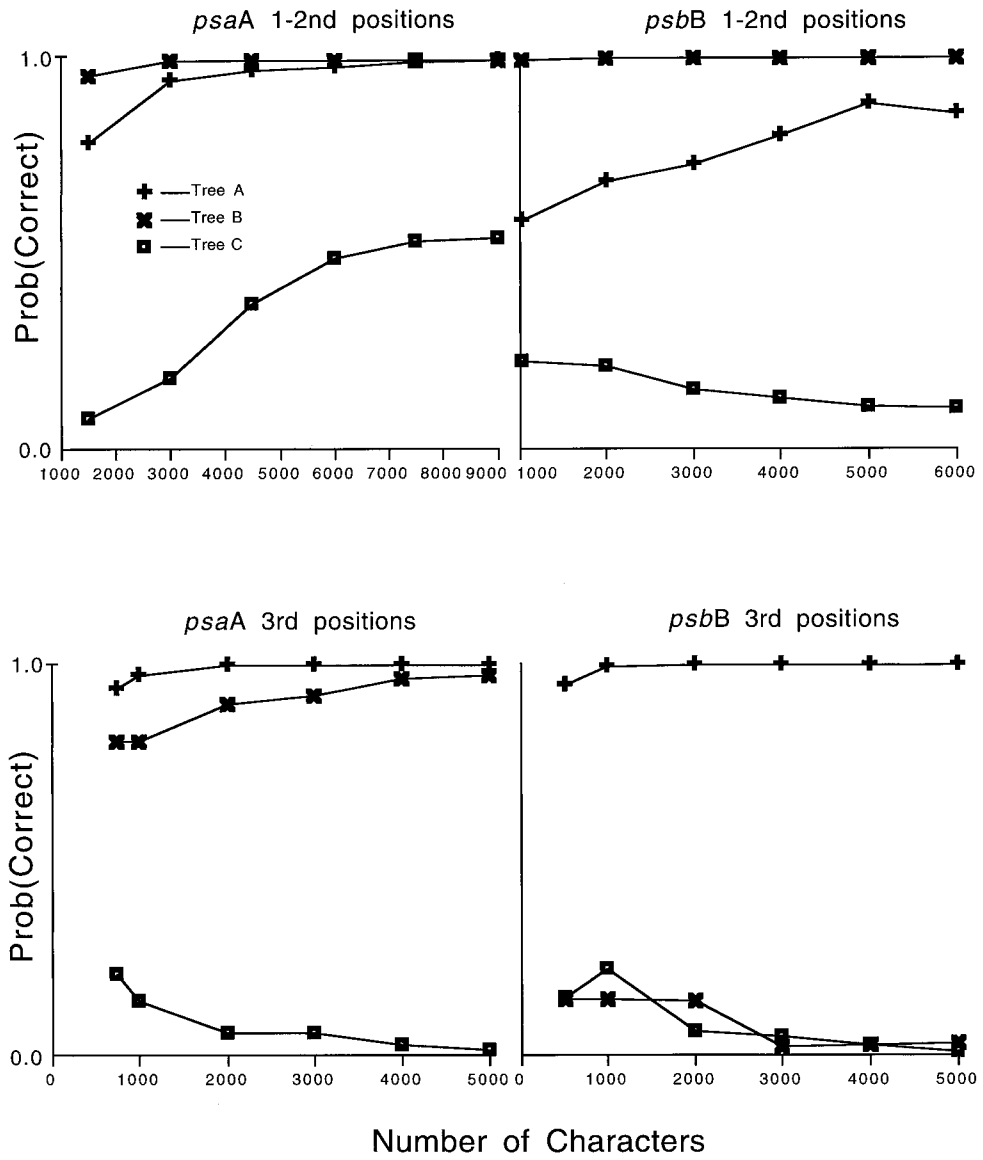

FIG. 12.—Simulations to identify statistical inconsistency. Each line shows the probability of obtaining a hypothesis given that the hypothesis is true (1–type I error probability).

angiosperms, found high support for a monophyletic gymnosperms. Winter et al. (1999), using multiple paralogs from the MADS-box family of nuclear genes, found the same result in each clade of orthologs (some with high support, some not). Bowe, Coat, and dePamphilis (2000) sequenced two mitochondrial protein-coding genes and inferred a gymnosperm clade with Gnetales nested among conifers. Chaw et al. (2000) inferred the same result with a different mitochondrial gene and in a reanalysis of 18S rDNA and *rbc*L (see also Ross et al. 1999, but note that they also found support for the anthophyte hypothesis with 26S data alone). Finally, M. W. Frohlich (personal communication) inferred a monophyletic gymnosperms in trees from phylogenetic analysis of the nuclear gene LEAFY.

That most of these results from recent studies are in agreement is significant, but it should be noted that almost all authors have downweighted or excluded third positions in protein-coding data, either explicitly at the nucleotide level (Hansen et al. 1999; Chaw et al. 2000; M. W. Frohlich, personal communication) or implicitly by working with only amino acid sequences (Winter et al. 1999). We wonder if third-position signals are as strong in these other genes as they are in these two photosystem genes.

Rates of Evolution

Properties of phylogenetic inference methods are in part determined by patterns in the tempo and mode of evolution of the sequences under study. Part of the explanation for the conflicting signals found in these photosystem genes may lie in the tremendous variation in rates seen across partitions, across sites within genes,

and across lineages. The amino acid sequences of photosystem genes are extremely conservative because their gene products function in the fundamental light-harvesting systems in plants. Rates of substitution at third positions are 10–20-fold higher than they are at first and second positions. Third-position pairwise distances (Kimura two-parameter corrected) range up to about one substitution per site--high enough to be interpreted as saturated, although the implications of saturation are not always obvious (Yang 1998; Björkland 1999). For example, the relative performance of reconstruction methods when the data are in fact considered saturated at third codon positions has yet to be adequately explored. In fact, the very definition of saturated may be contingent on the methods being used to reconstruct the phylogeny.

Rates also vary widely from site to site within both genes. Replacement substitutions are extremely variable in rate, ranging from many sites that are essentially invariant to a few that evolve fairly rapidly. Silent rates are more modal with somewhat less variation across each gene. Rates are even more variable across lineages, and variation is correlated between genes and, to a lesser extent, between partitions. Lineages with high rates, such as Gnetales and some ferns, tend to have high rates for both *psaA* and *psbB*, and both silent and replacement rates tend to be high. The same pattern holds for lineages with low rates. Seed plants other than angiosperms and Gnetales tend to have low rates at all positions.

The cause of this extensive rate variation among plant taxa is uncertain. Presumably, genes that are as functionally constrained and highly conserved as photosystem genes should be subject to intense purifying selection at nonsynonymous sites. However, the rate of nonneutral substitution is affected not only by the strength of selection, but also by demographic factors such as population size (Gillespie 1999). Long-lived seed plants such as conifers, cycads, and *Ginkgo* have lower rates of amino acid substitution than most angiosperms or Gnetales that were sampled. Evidently, some combination of selection intensity, population size, and perhaps the dynamics of fluctuations in each of these has caused these dramatic rate differences.

Lineage differences in rates of evolution of silent sites, on the other hand, are usually attributed to organism-wide factors such as generation time, metabolic rate (in animals at least), and differences in DNA repair mechanisms, or clade-specific factors like diversification rate (Bousquet et al. 1992). Little is known about generation times in plants, and even less is known about these other factors (Gaut 1998). Nonetheless, most extant gymnosperms are likely to have long generation times. Although it might seem difficult to attribute the high rates observed in Gnetales to short generation times because extant Gnetales are all long-lived perennials, the fossil record of Gnetales includes a much more diverse array of species, many of which might have had quite short life cycles (Stewart and Rothwell 1993). This could explain at least the high rates along the stem lineage to Gnetales. However, *Gnetum*, *Welwitschia*, and *Ephedra* also have high rates, despite their presumably long generation times.

Long-Branch Attraction

The detection of LBA is problematic. One method proposed to identify statistical inconsistency in parsimony reconstruction is to use an estimation procedure that is putatively consistent, such as ML (Huelsenbeck 1997). However, there are three problems with this approach. First, in finite data sets, every method, including ML, is subject to sampling error and bias. Second, ML has been shown to be inconsistent anyway, at least in some cases of model misspecification (Chang 1996). Third, computational limitations prevented us from assessing the strength of the signal in ML analyses (via bootstrapping) to quite the same degree as was possible with MP. We were limited to NNI branch swapping in the ML runs rather than the more exhaustive TBR swapping we undertook in the MP runs. There were some hints that ML results could differ from MP results just as would be expected if ML were more robust than MP to LBA. However, these results were dependent on the precise substitution model used and were inconsistent between genes. The third-position *psaA* ML results suggested moderate support for nesting of Gnetales with conifers (unlike under parsimony), but *psbB* showed no evidence of this. Moreover, the ML runs using SITERATES options that allow fixed but different rates in the different codon positions supported the Gnetales hypothesis, just as parsimony runs on the combined data sets did. We suspect that the SITERATES model is just not a very good one in the face of the tremendous rate variation across sites observed in these genes (fig. 6). Perhaps what is needed is a model permitting different gamma-distributed patterns of rate variation in different partitions.

Thus, we concentrated on a simulation approach. Both bias and statistical inconsistency were evident in third-position data for *psbB* if either the anthophyte tree (tree C) or the gymnosperm tree (tree B) was correct. In either of those cases, the tempo and mode of evolution of *psbB* was such that it was most likely that the Gnetales tree (tree A) would be reconstructed in a data set the size of the one used. Moreover, matters would become even worse if more data of the same type were gathered, which is the hallmark of statistical inconsistency. The statistical bias and inconsistency arise because of a combination of extremely short branches at the base of seed plants, long branches within Gnetales, and long branches in some outgroups. Evidently, given a near trichotomy at the base of seed plants, the *psbB* third-position data are such that the long-branched Gnetales tend to be attracted to a more basal position near the ferns.

On the other hand, there is less evidence for these same kinds of errors for *psaA* third-position data. For these data, the bias involves only the anthophyte tree, and once again it is closely associated with a trichotomy at the base of seed plants. Unlike in the case of *psbB* data, the gymnosperm tree (tree B) can be reconstructed

accurately even with third-position data. This apparent contradiction is puzzling. One explanation might be that the estimates of branch lengths by ML are themselves biased and that branch lengths are actually much longer than they appear in the *psaA* phylograms (figs. 7 and 8). If so, the parameters used in simulations might be too low to force LBA to occur. The saturation experiments described in figure 11 test for this. These experiments show that branch lengths would have to be underestimated by nearly an order of magnitude to mask a long-branch problem that is really in the data. Given this, it is doubtful that the likelihood approach used to estimate branch lengths is that highly biased (Zharkikh 1994). Note that there is a short but finite branch supporting the gymnosperm clade in tree B of *psaA* (fig. 8), unlike the case for *psbB*, in which its length is nearly 0 (fig. 10).

Assessments of statistical error must be interpreted cautiously. Type I and type II errors are conditional statements of the form, ''if a given hypothesis is correct, what is the probability that this experiment would mistakenly reject it?'' Such statements do not assign probabilities to whether or not the hypotheses really are correct. Hence, if the Gnetales hypothesis (tree A) were indeed true, our new understanding of the high rate of error using third-position data for *psbB* would not make it any less true. It would just mean that these data have little power to discriminate among the three hypotheses, and therefore it would lessen one's confidence in the inference. Estimates of error in this study paint a mixed picture. The worst case is that of *psbB* third-position data, which will always point to the Gnetales hypothesis regardless of the true tree. At the other extreme are the first- and second-position data for both genes, which have extremely low error rates and which both point to the Gymnosperm hypothesis as the true tree. The only real conflict is provided by the *psaA* third-position data. They have a low error rate if either tree A or tree B is correct, but they point to the Gnetales hypothesis rather than the gymnosperm hypothesis, as suggested by the first-second position data. One conservative conclusion is that if the anthophyte hypothesis is correct, none of these data would be particularly good at demonstrating it. Overall, however, perhaps the fairest conclusion would be that there is a slight edge in terms of weight of evidence for the hypothesis that extant gymnosperms form a clade but that some explanation for the *psaA* third-position results is still wanting. Further statistical analyses along the lines outlined here for other genes may be illuminating.

A few other cases of strongly conflicting signal between codon partitions have been described in the literature. Stanger-Hall and Cunningham (1998), for example, investigated conflicting relationships implied by mitochondrial *Cyt b* and *COII* data for lemurs, following up on the work of Yoder, Vilgalys, and Ruvolo (1996) and Adkins and Honeycutt (1994). They argued that it was the first- and second-position data that were more misleading than the third-position data, owing to selective constraints on amino acid evolution, but that specifics of the reconstruction algorithm (e.g., details of substitution model) interacted in a complex way with the data partitions, a result we found as well. Considered together, these results and ours suggest that rejection of third-position data a priori in any phylogenetic analysis on account of presumed saturation is not a good strategy. Sometimes, this might cause more reliable data to be excluded in favor of less reliable data. More generally, this strategy may mask interesting cases of conflict between data partitions and consequently weaken our understanding of how data and algorithms interact in efforts to reconstruct credible phylogenies.

Supplementary Material

Sequence alignments are available from the following web site: http://loco.ucdavis.edu/sandlab/sp2k.htm.

Acknowledgments

We thank J. Palmer, C. dePamphilis, and M. Frohlich for sharing unpublished manuscripts; T. Metcalf, E. Sandoval, and the staff of the UCD Botanical Conservatory and the UCD Arboretum for graciously providing plant material and expertise; and the Green Plant Phylogeny Research Coordinating Group for providing financial support. M. J. Donoghue, J. A. Doyle, and K. P. Steele provided useful insights. M.J.S. and T.S.K. are supported by NSF grant DEB-9726856 to M.J.S. and J. A. Doyle.

LITERATURE CITED

ADACHI, J., and M. HASEGAWA. 1996. MOLPHY. Computer program published by the authors. Institute of Statistical Mathematics, Tokyo.

ADKINS, R. M., and R. L. HONEYCUTT. 1994. Evolution of primate cytochrome c oxidase subunit II gene. J. Mol. Evol. **38**:215–231.

ALBERT, V. A., A. BACKLUND, K. BREMER, M. W. CHASE, J. R. MANHART, B. D. MISHLER, and K. C. NIXON. 1994. Functional constraints and *rbc*L evidence for land plant phylogeny. Ann. Mo. Bot. Gard. **81**:534–567.

BAKKE, E., and A. VON HAESELER. 1999. Distance measures in terms of substitution process. Theor. Popul. Biol. **55**:166–175.

BJÖRKLAND, M. 1999. Are third positions really that bad? A test using vertebrate cytochrome *b*. Cladistics **15**:191–197.

BOUSQUET, J., S. H. STRAUSS, A. H. DOERKSEN, and R. A. PRICE. 1992. Extensive variation in evolutionary rate of *rbc*L gene sequences among seed plants. Proc. Natl. Acad. Sci. USA **89**:7844–7848.

BOWE, L. M., G. COAT, and C. W. DEPAMPHILIS. 2000. Phylogeny of seed plants based on all three plant genomic compartments: extant gymnosperms are monophyletic and Gnetales are derived conifers. Proc. Natl. Acad. Sci. USA (in press).

CARLQUIST, S. 1996. Wood, bark, and stem anatomy of Gnetales: a summary. Int. J. Plant Sci. **157**(Suppl.):S58–S76.

CARMEAN, D., and B. CRESPI. 1995. Do long branches attract flies? Nature **373**:666.

CHANG, J. T. 1996. Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. Math. Biosci. **134**:189–215.

CHASE, M. W., D. E. SOLTIS, R. G. OLMSTEAD et al. (39 coauthors). 1993. 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., C. L. PARKINSON, Y. CHENG, T. M. VINCENT, and J. D. PALMER. 2000. Seed plant phylogeny inferred from all three plant genomes: monophyly of extant gymnosperms and origin of Gnetales from conifers. Proc. Natl. Acad. Sci. USA (in press).

CHAW, S.-M., A. ZHARKIKH, H.-M. SUNG, T.-C. LAU, and W.-H. LI. 1997. Molecular phylogeny of extant gymnosperms and seed plant evolution: analysis of nuclear 18S rRNA sequences. Mol. Biol. Evol. **14**:56–68.

CRANE, P. R. 1985. Phylogenetic analysis of seed plants and the origin of angiosperms. Ann. Mo. Bot. Gard. **72**:716–793.

DAYHOFF, M. O., R. M. SCHWARTZ, and B. C. ORCUTT. 1978. A model of evolutionary change in proteins. Pp. 345–352 *in* M. O. DAYHOFF, ed. Atlas of protein sequence and structure. Vol. 5, Suppl. 3. National Biomedical Research Foundation, Washington, D.C.

DOYLE, J. A. 1996. Seed plant phylogeny and the relationships of Gnetales. Int. J. Plant Sci. **157**(Suppl. 6):S3–S39.

DOYLE, J. A., and M. J. DONOGHUE. 1986. Seed plant phylogeny and the origin of angiosperms: an experimental cladistic approach. Bot. Rev. **52**:321–431.

———. 1992. Fossils and seed plant phylogeny reanalyzed. Brittonia **44**:89–106.

FARRIS, J. S., M. KÄLLERSJÖ, A. G. KLUGE, and C. BULT. 1995. Constructing a significance test for incongruence. Syst. Biol. **44**:570–572.

FELSENSTEIN, J. 1978. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. Biol. J. Linn. Soc. **16**:183–196.

———. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783–791.

FROHLICH, M. W., and D. S. PARKER. 1999. Seed plant phylogeny: evidence from Floricaula/Leafy. XVIth International Botanical Congress [abstract].

GAUT, B. 1998. Molecular clocks and nucleotide substitution rates in higher plants. Pp. 93–120 *in* M. K. HECHT, ed. Evolutionary biology. Vol. 30. Plenum Press, New York.

GILLESPIE, J. H. 1999. The role of population size in molecular evolution. Theor. Popl. Biol. **55**:145–156.

GOREMYKIN, V., V. BOBROVA, J. PAHNKE, A. TROITSKY, A. ANTONOV, and W. MARTIN. 1996. Noncoding sequences from the slowly evolving chloroplast inverted repeat in addition to *rbc*L data do not support gnetalean affinities of angiosperms. Mol. Biol. Evol. **13**:383–396.

GRAHAM, S. W., and R. G. OLMSTEAD. 1999. A phylogeny of basal angiosperms inferred from 17 chloroplast genes. XVIth International Botanical Congress [abstract].

HANSEN, A., S. HANSMANN, T. SAMIGULLIN, A. ANTONOV, and W. MARTIN. 1999. *Gnetum* and the angiosperms: molecular evidence that their shared morphological characters are convergent rather than homologous. Mol. Biol. Evol. **16**:1006–1009.

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

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

HUELSENBECK, J. P. 1995. The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor-joining. Mol. Biol. Evol. **12**:843–849.

———. 1997. Is the Felsenstein Zone a fly trap? Syst. Biol. **46**:69–74.

———. 1998. Systematic bias in phylogenetic analysis: is the Strepsiptera problem solved? Syst. Biol. **47**:519–537.

HUELSENBECK, J. P., D. M. HILLIS, and R. JONES. 1996. Parametric bootstrapping in molecular phylogenetics: applications and performance. Pp. 19–45 *in* J. D. FERRARIS and S. R. PALUMBI, eds. Molecular zoology: advances, strategies and protocols. Wiley-Liss, New York.

KÄLLERSJÖ, M., V. A. ALBERT, and J. S. FARRIS. 1999. Homoplasy increases phylogenetic structure. Cladistics **15**:91–93.

KIM, J. 1998. Large-scale phylogenies and measuring the performance of phylogenetic estimators. Syst. Biol. **47**:43–60.

LEHMANN, E. L. 1983. Theory of point estimation. Wiley, New York.

LI, W.-H. 1997. Molecular evolution. Sinauer, Sunderland, Mass.

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

LOCONTE, H., and D. W. STEVENSON. 1990. Cladistics of the Spermatophyta. Brittonia **42**:197–211.

LYONS-WEILER, J., and G. A. HOELZER. 1997. Escaping from the Felsenstein Zone by detecting long branches in phylogenetic data. Mol. Phylogenet. Evol. **8**:375–384.

MADDISON, D. R., M. D. BAKER, and K. A. OBER. 1999. Phylogeny of carabid beetles inferred from 18S ribosomal DNA (Coleoptera: Carabidae). Syst. Entomol. **24**:103–138.

MANLY, B. F. J. 1997. Randomization, bootstrap and Monte Carlo methods in biology. Chapman and Hall, New York.

MEYER, A. 1994. Shortcomings of the cytochrome *b* gene as a molecular marker. TREE **9**:278–280.

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.

ORT, D. R., and C. F. YOCUM. 1996. Oxygenic photosynthesis: the light reactions. Kluwer, Boston.

PELLMYR, O., and J. LEEBENS-MACK. 1999. Forty million years of mutualism: evidence for Eocene origin of the yucca-yucca moth association. Proc. Natl. Acad. Sci. USA **96**:9178–9183.

RAMBAUT, A., and N. C. GRASSLY. 1997. Seq-Gen: an application for the monte-carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. **13**:235–238.

RAUBESON, L. A., and R. K. JANSEN 1992. A rare chloroplast DNA structural mutation is shared by all conifers. Biochem. Syst. Ecol. **20**:17–24.

ROSS, V. A., M. J. ZANIS, P. S. SOLTIS, and D. SOLTIS. 1999. Phylogenetic relationships among extant seed plant lineages inferred from 26S rDNA sequences. XVIth International Botanical Congress [abstract].

ROTHWELL, G. R., and R. SERBET. 1994. Lignophyte phylogeny and the evolution of spermatophytes: a numerical cladistic analysis. Syst. Bot. **19**:443–482.

SIDDALL, M. E. 1998. Success of parsimony in the four-taxon case: long branch repulsion by likelihood in the Farris zone. Cladistics **14**:209–220.

SOLTIS, P. S., D. E. SOLTIS, P. G. WOLF, D. L. NICKRENT, S.-M. CHAW, and R. L. CHAPMAN. 1999. The phylogeny of land plants inferred from 18s rDNA sequences: pushing the limits of rDNA signal? Mol. Biol. Evol. **16**:1774–1784.

STANGER-HALL, K., and C. W. CUNNINGHAM. 1998. Support for a monophyletic lemuriformes: overcoming incongruence between data partitions. Mol. Biol. Evol. **15**:1572–1577.

STEWART, W. N., and G. W. ROTHWELL. 1993 Paleobotany and the evolution of plants. 2nd edition. Cambridge University Press, New York.

SWOFFORD, D. S. 1999. PAUP* 4.0. Phylogenetic analysis using parsimony (*and other methods). Version 4b2. Sinauer, Sunderland, Mass.

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

WINTER, K.-U., A. BECKER, T. MUNSTER, J. T. KIM, H. SAEDLER, and G. THEISSEN. 1999. MADS-box genes reveal that gnetophytes are more closely related to conifers than to flowering plants. Proc. Natl. Acad. Sci. USA **96**:7342–7347.

YANG, Z. 1998. On the best evolutionary rate for phylogenetic analysis. Syst. Biol. **47**:125–133.

YODER, A. D., R. VILGALYS, and M. RUVOLO. 1996. Molecular evolutionary dynamics of cytochrome beta in strepsirrhine primates: the phylogenetic significance of third-position transversions. Mol. Biol. Evol. **13**:1339–1350.

ZHARKIKH, A. 1994. Estimation of evolutionary distances between nucleotide sequences. J. Mol. Evol. **39**:315–329.

PAMELA SOLTIS, reviewing editor

Accepted Accepted January 25, 2000