Letter to the Editor

Usefulness of RNA Polymerase II Coding Sequences for Estimation of Green Plant Phylogeny

Amy L. Denton, * Betty L. McConaughy, † and Benjamin D. Hall*‡

*Department of Botany and †Department of Genetics, University of Washington

Phylogenetic reconstruction based on chloroplast (cp) and nuclear ribosomal (r)DNA sequences have contributed greatly to our understanding of green plant evolution. Many relationships remain unresolved, however, especially among basal angiosperm taxa. Here, both plastid and ribosomal gene analysis have generated inconclusive, and occasionally contradictory, hypotheses of deep branching topology (Chase et al. 1993; Soltis et al. 1997). Such incongruities underscore the need for sequence data from additional independent loci. The pilot study we report here was motivated by an interest in studying the early diversification of land plants. Toward this end, we PCR-amplified and sequenced the genes encoding the second largest subunit of RNA polymerase II (RPB2) from nine taxa representing tricolpate, monocolpate, and eudicot angiosperm lineages, one gymnosperm, and one liverwort species to determine the feasibility of recovering the RPB2 gene from any green plant using a single set of PCR primers. We also performed a preliminary evaluation of the potential phylogenetic utility of RPB2 sequences.

Other than the widely studied large- and small-subunit rDNA genes, protein-encoding genes are the main potential source of comparative nuclear DNA sequence. In plants, many of these genes are members of repeated families, complicating phylogenetic comparisons with issues of paralogy and within-family concerted evolution. These problems were noted in phylogenetic studies with the genes for phytochrome (Mathews, Lavin, and Sharrock 1995), small heat shock proteins (Waters 1995), and cytosolic phosphoglucose isomerase (Gottlieb and Ford 1996). Currently, the nuclear gene with the broadest phylogenetic applicability in plants is alcohol dehydrogenase (Adh), but separate gene duplication events within angiosperm lineages may limit the utility of universal angiosperm Adh primers (Morton, Gaut, and Clegg 1996; Sang, Donoghue, and Zhang 1997).

The identification of nonrepeated nuclear genes that evolve at rates suitable for interfamilial phylogenetic assessment is made difficult by the paucity of comparative sequence data currently available for plant nuclear genomes. In principle, promising nuclear genes for such studies are those that encode components of the multiprotein assemblies involved in the apparatuses for replication and transcription in the cell nucleus. These proteins are constrained because each enzyme subunit contacts many other proteins, and also because the holoenzyme complex must be able to copy many different nuclear sequences with fidelity. For different modes of DNA replication, as for the transcription of different classes of genes, there are separate and distinct multi-subunit enzymes (as RNA polymerases I, II, and III), the major subunits of which diverged from one another in the earliest eukaryotes (Pühler et al. 1989).

RNA polymerase II catalyzes messenger RNA synthesis in eukaryotic cell nuclei. It has already proven valuable for evolutionary research at the broadest taxonomic levels, because it is ubiquitous and contains motifs that are conserved across the primary kingdoms of life (Allison et al. 1985; Iwabe et al. 1991). Gene sequences of RPB1 and RPB2 encoding the two largest subunits of RNA polymerase II have been applied recently to questions of protist evolution (Klenk et al. 1995), plastid introduction, and diversification within eukaryotes (Stiller and Hall 1997), and the origin of eukaryotes and major eukaryotic lineage divergence (Iwabe et al. 1991; Sidow and Thomas 1993).

The nuclear RNA polymerase II core enzyme of plants consists of two large subunits, 205 and 140 kDa, and as many as nine smaller polypeptides (Guilfoyle 1983). The second largest subunit (RPB2) forms a part of the catalytic core that is believed to function in nucleotide binding and RNA chain elongation (Woychik and Young 1990). Within the 3,564 bp of exon sequence in the Arabidopsis thaliana RPB2 gene (the only complete genomic plant RPB2 sequence known) are interspersed 24 introns ranging in size from 75 to 776 nt (Larkin and Guilfoyle 1993). The high degree of sequence identity within several conserved RPB2 motifs suggested to us the possibility of using PCR to recover the same section of this gene from many different taxa for sequencing. Particularly advantageous for phylogenetic analysis is the occurrence between pairs of priming sites in RPB2 of readily alignable sequences that are (nonsynonymously) quite variable in codon content.

Both empirical observations and theoretical considerations indicate the presence of a single RPB2 gene per haploid genome in major eukaryotic lineages (Thuriaux and Sentenac 1992; Sidow and Thomas 1993; unpublished data). RPB2 is encoded by a single gene in Arabidopsis (Larkin and Guilfoyle 1993) and tomato (Warrilow and Symons 1996). In plants, a multicopy status has been observed only for the RPB1 gene of Glycine max, a species believed to have undergone ancient polyploidization (Kumar and Hymowitz 1989). In addition, direct experimental evidence based on Southern blots, restriction fragment length polymorphisms, and a Mendelian test cross supports the existence of a single RPB2...
Fig. 1—Locations and sequences of PCR primers used for amplification and sequencing of RPB2. Black boxes indicate coding regions conserved throughout eukaryotes. Lines indicate less conserved coding regions. Asterisks indicate intron positions in the Arabidopsis thaliana reference sequence (GenBank accession numbers X52494, Z19120). Arrows indicate approximate position and direction of primers. RPB2 region 6–11 sequences determined in this study were assigned GenBank accession numbers AF020839–AF020844, AF041852, AF043426–AF043427.

locus in Rhododendron and related genera (Denton 1997).

Consistent with expression of a single orthologous gene, no evidence of multiple RPB2 copies was observed in any of the species sampled. Minipreparations of cloned DNA were either sequenced manually using standard protocols or cycle-sequenced and analyzed with an Applied Biosystems 377 automated sequencer.

A 1,668-bp coding region of RPB2 encoding 556 amino acids was recovered from Spinacia, Hordeum, Magnolia, Aristolochia, Peperomia, Nymphaea, Chloranthus, Ginkgo, and Marchantia. Published Arabidopsis and Lycopersicon sequences were added, and unambiguous alignment of the RPB2 sequences was accomplished by eye (alignment available from the authors upon request). The 11 aligned RPB2 sequences indicate a remarkable degree of constancy in the architectural framework of this protein. Within the taxa examined, only one indel was found between conserved regions 6 and 11, a single amino acid indel shared among eudicots. The genomic DNA sequences that we determined,
those of *Aristolochia*, *Ginkgo*, and *Marchantia*, represent 400 Myr of land plant evolutionary divergence. Each of these has the same number of introns and the same intron positions in regions 6–11 as does *Arabidopsis RPB2* (fig. 1). While intron placement is conserved, intron length and sequence is highly variable. *Arabidopsis* introns are generally much shorter than those of other plant taxa. Intron 13 is the most extreme example, ranging from 80 bp in *Arabidopsis* to 1629 bp in *Ginkgo*.

To detect phylogenetic signal in the *RPB2* sequences, maximum-parsimony and maximum-likelihood trees were constructed using the aligned amino acid data (fig. 2). Protein data were chosen due to third-position nucleotide saturation (based on average Jukes-Cantor pairwise divergence value of 0.82 calculated for *RPB2* nucleotide data sets using PHYLIP 3.5; Felsenstein 1993). Both methods consistently recovered the eudicot (*Spinacia*, *Arabidopsis*, *Lycopersicon*) and angiosperm clades. The monophyly of each of these groups was evaluated in a separate K-H (Kishino and Hasegawa 1989) paired-sites test, conducted under the parsimony criterion using a test version of PAUP (version 4.0 d060, written by D. Swofford). Constraint trees specifying the monophyly of either the angiosperms or the eudicots were used to search for the shortest trees not compatible with these constraints. Three trees violating the monophyly of each group were found, and K-H tests rejected these alternative hypotheses of nonmonophyly as significantly worse for both angiosperms (*P* = 0.045 for all three) and eudicots (*P* = 0.04, 0.04, 0.08), strengthening the argument for the phylogenetic utility of *RPB2*. Branching topologies among the magnoliid and monocot taxa were not resolved in this preliminary analysis of *RPB2* sequences, a result that is not surprising given the limited sampling. Analyses of large sequence data sets for other genes have not been able to unequivocally resolve relationships among the woody magnoliid, monocot, and diverse paleoherb lineages (Chase et al. 1993; Soltis et al. 1997; Rice, Donoghue, and Olmstead 1997). In order to understand the ancient divisions within basal angiosperms, a coordination of efforts with several independent data sets, increased taxon sampling, and improved computational methods may be required. It is our hope that *RPB2* sequences can contribute toward this goal.

While our sampling of *RPB2* sequence data is currently limited, the taxa chosen encompass a range of plant families, and for each species examined, the PCR primers specifically amplified a fragment of the *RPB2* gene. In addition to the nine new green plant sequences reported here, the plant-specific 6–11 primers have been used in our laboratory to recover partial *RPB2* 6–11 sequences from species in the Gnetales, Pinaceae, Taxa-
ceae, Cupressaceae, Ceratophyllaceae, Musaceae, Dioscoreaceae, Cyperaceae, Austrobaileyaceae, Illiciaceae, Winteraceae, Fagaceae, Crassulaceae, Caprifoliaceae, Ericales s.l., Scrophulariaceae, and Asteraceae. These same oligonucleotides work well to prime dideoxynucleotide sequencing reactions. Alignment of \( RPB2 \) sequences across diverse plant species is straightforward. The rate of evolutionary change in the coding sequences of \( RPB2 \) appears well-suited for broad-scale phylogenetic studies. In addition to using amino acid sequences to resolve deeper relationships, analyses of exonic nucleotide substitution and changes within \( RPB2 \) intron sequences will make it possible to address phylogenetic questions at a number of levels.

**Acknowledgments**

We thank two anonymous reviewers for comments on the manuscript, A. Mullennix for the *Marchantia RPB2* 10–11 sequence, J. Stiller for collaboration on primer design, D. Ewing for permission to raid the greenhouse, and C. Abair for help with the figures.

**LITERATURE CITED**


Kumar, P. S., and T. Hymowitz. 1989. Where are the diploid (2n=2x=20) genome donors of *Glycine* Willd. (Leguminosae, Papilionoideae)? Euphytica **40**:221–226.


Barbara A. Schaal, reviewing editor

Accepted April 21, 1998.