BASAL ANGIOSPERM PHYLOGENY INFERRED FROM DUPLICATE 
PHYTOCHROMES A AND C

Sarah Mathews1 and Michael J. Donoghue

Division of Biological Sciences, University of Missouri—Columbia, Columbia, Missouri 65211, U.S.A.; and Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut 06520-8106, U.S.A.

We have extended our studies of angiosperm phylogeny based on a pair of duplicated phytochrome (PHY) genes, PHYA and PHYC. Phylogenetic analyses of sequences from 52 species yield unrooted gene networks in which all of the elements resolved in our previous analysis of 26 species appear. Amborella still emerges as the sister group of all other angiosperms. However, we cannot reject alternative rootings in which water lilies, either alone or in combination with Amborella, are basal. Austrobaileya + Illicium diverges next from the remaining angiosperms. Eudicots and monocots form rather well-supported clades, as do Magnoliidae, Laurales, Piperales, and winteroids, but relationships among these major lineages remain uncertain, as do the positions of Chloranthaceae and Ceratophyllum. Magnoliidae may be directly linked with Laurales and Piperales with winteroids, but support for these relationships is not strong. Within eudicots, a basal split between ranunculids (Ranunculales, Papaverales) and the rest of the eudicots is supported, though the position of Nelumbo is equivocal. These same relationships are obtained in combined analyses of PHYA and PHYC (species as terminals) when Ceratophyllum is excluded. However, when Ceratophyllum is included, Austrobaileya + Illicium and then Chloranthaceae diverge from the remaining angiosperms before Ceratophyllum + water lilies. Rooted species trees inferred from duplicate gene networks by minimizing gene duplications and losses are highly congruent with the gene subtrees and with the results of recent analyses of other genes, even when Ceratophyllum is included. More attention must be paid to the methods for obtaining rooted species trees from data sets that include duplicate genes, especially if we are to fully implement the search for species trees that simultaneously minimize the multiple possible causes of conflict among gene trees.

Keywords: angiosperm phylogeny, duplicate genes, duplicate gene rooting, phytochrome genes, uninode coding.

Introduction

In two previous analyses we explored the use of a gene duplication in the phytochrome gene family, leading to PHYA and PHYC, in order to root the phylogeny of angiosperms (Donoghue and Mathews 1998; Mathews and Donoghue 1999). Our first analysis of data from just 12 species resolved highly congruent gene subtrees, demonstrating the potential of the approach, but it lacked many potentially basal lineages. Sampling was improved in our subsequent analysis, and we obtained a more convincingly rooted species tree. However, many major clades within angiosperms were represented by just one (e.g., Chloranthaceae) or a few (e.g., monocots and eudicots) species. Here we report results of similar analyses of PHYA and PHYC sequences from nearly twice as many species.

Background

Phylogenetic analyses of phytochrome genes in green plants indicate that the phytochrome gene pair PHYA and PHYC diverged prior to the origin of angiosperms (Mathews and Sharrock 1997; Donoghue and Mathews 1998). PHYA and PHYC are found in most angiosperms examined (but see Howe et al. 1998; Lavin et al. 1998) and form well-supported clades (bootstrap values >90%), indicating that they are evolving independently (Mathews and Sharrock 1997). A single-gene lineage related to this pair is known from other seed plants (Mathews and Sharrock 1997; Mathews and Donoghue 1999); this gene lineage has diversified in conifers (Clapham et al. 1999; S. Mathews, unpublished data) and perhaps in other seed plants.

In most angiosperms, phytochrome genes occur as a small family comprising three to five members. The completely characterized phytochrome family of Arabidopsis has five members, PHYA through PHYE (Sharrock and Quail 1989; Clack et al. 1994). Typical phytochrome coding sequences range from ca. 3.3 to ca. 4.1 kb and are interrupted by three introns, a structure shared by most land plant PHY (Cowl et al. 1994; Quail 1994; Wada et al. 1997). Amino acid sequences inferred from the five Arabidopsis PHY are 46%–80% identical; PHYA and PHYC are 52% identical (Sharrock and Quail 1989; Clack et al. 1994). The five encoded photoreceptors have both distinct and overlapping roles in photomorphogenesis. Phytochrome A (phyA) controls hypocotyl elongation, cotyledon expansion, and seed germination in response to prolonged far-red irradiation; phyA also mediates responses to single pulses of far-red light or very low doses of red light (Whitelam and Devlin 1997). Overexpression of phyC in transgenic Arabi-
Our current analyses include both PHYA and PHYC sequences from 48 taxa, plus PHYA only from *Tacca* and PHYC only from *Illicium, Lemma*, and *Plea* (table 1). From Cabombaceae, we sampled PHYA from *Brasenia* but PHYC from *Cabomba*. Phytochrome loci were sampled using primers and protocols described previously (Mathews and Donoghue 1999); stepdown gene amplification protocols (Hecker and Roux 1996) with starting annealing temperatures of 60°C to 65°C were most effective. Fragments of exon 1 (1.2–1.4 kb) or of exons 1 and 2 and the intervening intron (1.6–1.8 kb) were amplified, cloned, and sequenced.

Each cloned sequence was aligned with all other available phytochrome sequences from land plants. The alignment was edited, and final data matrices of 3255 nucleotide sites from all four exons and of 1303 nucleotide sites from exon 1 were constructed using Se-Al (Rambaut 1996). Three alignments were analyzed. In the first two alignments, sequences are terminals, whereas in the third, species are terminals. The first (3255 nucleotide sites) was a comparison of all duplicated and unduplicated sequences to infer a gene phylogeny. From this we determined whether newly obtained data altered the interpretation (of previous studies) that PHYA and PHYC are monophyletic gene lineages and that they diverged along the branch to angiosperms (e.g., Mathews et al. 1995; Mathews and Sharrock 1997). The second alignment (1303 nucleotide sites) included only PHYA and PHYC sequences obtained from the species listed in table 1, which were analyzed to obtain an unrooted gene network of two gene subtrees. We used GeneTree (Page 1998) to identify optimal species trees compatible with most parsimonious gene networks inferred in analysis of the second alignment. Ten random starting trees were rearranged by alternating subtree pruning and regrafting and nearest neighbor interchange to find species trees into which the gene networks fit with the fewest number of duplications and losses (see “Discussion”). Equally good solutions were retained, and the steepest descent option was in effect.

If PHYA and PHYC are evolving independently, they can be combined to obtain a species phylogeny. In the third alignment (2606 nucleotide sites), species are terminals, and the data from each gene are concatenated. The PHYA from *Brasenia* and the PHYC *Cabomba* were combined in a single terminal, Cabombaceae, in this alignment. The root of the species phylogeny was positioned according to evidence from the gene networks inferred in analysis of the second alignment.

 Parsimony analyses using PAUP* version 4.0b3a or 4.0b4a (Swofford 2000) comprised heuristic searches of 100 replicates of random taxon addition with tree bisection and reconnection (TBR) branch swapping. We used two character-weighting schemes, one in which characters were equally weighted and one in which third codon positions were assigned a weight of 0.75. Bootstrap values (Felsenstein 1985) were estimated in 100 replicates of heuristic searches with 10 replicates of random taxon addition and TBR branch swapping. We conducted analyses with and without the sequences from *Ceratophyllum*.
Fig. 1 One of 10 most parsimonious networks of PHYA and PHYC from 52 angiosperms (1303 nucleotide sites, 838 parsimony informative) that maximizes identical components in the gene subtrees; weighted parsimony analysis (100 random taxon addition replicates with TBR swapping in PAUP* 4.0 [Swofford 2000]) yielded networks of 10,875.25 steps (retention index [RI] = 0.54; consistency index [CI] = 0.15, excluding autapomorphies). Bootstrap percentages (from 100 replicates with the same search parameters, but using 10 random addition replicates) are shown above branches. Identical components in the PHYA and PHYC subtrees are labeled A through BB.
Table 1

<table>
<thead>
<tr>
<th>Clade</th>
<th>Species</th>
<th>Voucher</th>
<th>GenBank accession numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>PHYA</td>
</tr>
<tr>
<td>Amborellaceae</td>
<td>Amborella trichopoda Baill.</td>
<td>M. P. Simmons 1846 LH</td>
<td>AF190062 AF1900063</td>
</tr>
<tr>
<td></td>
<td>Ambrotea baileyana C. T. White</td>
<td>SYM 388 A</td>
<td>AF190068 AF1900069</td>
</tr>
<tr>
<td>Austrobaileyaceae</td>
<td>Austrobaileya scandens C. T. White</td>
<td>SYM 388 A</td>
<td>AF190068 AF1900069</td>
</tr>
<tr>
<td>Canellaceae</td>
<td>Canella winterana Gaertner</td>
<td>A-6087-FTG</td>
<td>AF190074 AF1900075</td>
</tr>
<tr>
<td>Winteraceae</td>
<td>Drimys winteri J. R. &amp; G. Forst.</td>
<td>SYM 397 A</td>
<td>AF190080 AF190081</td>
</tr>
<tr>
<td></td>
<td>Pseudowintera axillaris Dandy</td>
<td>SYM 412 A</td>
<td>AF276737 AF276738</td>
</tr>
<tr>
<td>Chloranthaceae</td>
<td>Chloranthus spicatus Mak.</td>
<td>A. Bernhard 176-96</td>
<td>AF190076 AF190077</td>
</tr>
<tr>
<td></td>
<td>Hedyosum Sw. sp.</td>
<td>SYM 419 A</td>
<td>AF190072 AF190073</td>
</tr>
<tr>
<td></td>
<td>Sarcandra glabra (Thunb.) Nakai</td>
<td>SYM 419 A</td>
<td>AF190072 AF190073</td>
</tr>
<tr>
<td>Illiciaceae</td>
<td>Illicium oligandrum Merr. and Chun</td>
<td>SYM 450 A</td>
<td>nd</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AF276729</td>
</tr>
<tr>
<td>Ceratophyllales</td>
<td>Ceratophyllum denserum L.</td>
<td>SYM s.n. MONT</td>
<td>AF276716 AF276717</td>
</tr>
<tr>
<td>Laurales</td>
<td>Calycanthus floridus L.</td>
<td>SYM 404 A</td>
<td>AF190072 AF190073</td>
</tr>
<tr>
<td></td>
<td>Hedycarya angustifolia A. Cunn.</td>
<td>NSW 410584 A</td>
<td>AF190084 AF190085</td>
</tr>
<tr>
<td></td>
<td>Hernandia L. sp.</td>
<td>Alverson &amp; Rubio 2238 A</td>
<td>AF190086 AF190087</td>
</tr>
<tr>
<td></td>
<td>Idiospermum australiense S. T. Blake</td>
<td>NSW 410587 A</td>
<td>AF190089 AF190090</td>
</tr>
<tr>
<td>Magnoliaceae</td>
<td>Annona L. sp.</td>
<td>SYM 40586 A</td>
<td>AF190082 AF190083</td>
</tr>
<tr>
<td></td>
<td>Degeneria vitiensis L. W.</td>
<td>John Miller 1189-63</td>
<td>AF190078 AF190079</td>
</tr>
<tr>
<td></td>
<td>Eupomatia laurina Hook.</td>
<td>SYM 460 A</td>
<td>AF190094 AF190095</td>
</tr>
<tr>
<td></td>
<td>Magnolia × soulangiana</td>
<td>SYM 460 A</td>
<td>AF190094 AF190095</td>
</tr>
<tr>
<td></td>
<td>Rapistrum laurina Hook.</td>
<td>SYM 460 A</td>
<td>AF190094 AF190095</td>
</tr>
<tr>
<td>Piperales</td>
<td>Aristolochia grandiflora Sw.</td>
<td>SYM 399 A</td>
<td>AF276712 AF276713</td>
</tr>
<tr>
<td></td>
<td>Houttuynia cordata Thunb.</td>
<td>A. Bernhard 174-96</td>
<td>AF276726 AF276748</td>
</tr>
<tr>
<td></td>
<td>Lactoris fernandeziana Phil.</td>
<td>MWC 1014 K</td>
<td>AF190091 AF190092</td>
</tr>
<tr>
<td></td>
<td>Piper nigrum L.</td>
<td>SYM 401 A</td>
<td>AF190100 AF190101</td>
</tr>
<tr>
<td></td>
<td>Saraurus cernuus L.</td>
<td>A. Bernhard 176-96</td>
<td>AF190106 AF190107</td>
</tr>
<tr>
<td></td>
<td>Saruma henryi Oliver</td>
<td>A. Bernhard 176-96</td>
<td>AF190104 AF190105</td>
</tr>
<tr>
<td>Monocots:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alismataceae</td>
<td>Lemna gibba L.</td>
<td>SYM 40586 A</td>
<td>AF190082 AF190083</td>
</tr>
<tr>
<td></td>
<td>Sagittaria L. sp.</td>
<td>SYM 40586 A</td>
<td>AF190082 AF190083</td>
</tr>
<tr>
<td>Asparagaceae</td>
<td>Asparagus falcata W. J. Hahn 6881 WIS</td>
<td>SYM 40586 A</td>
<td>AF190082 AF190083</td>
</tr>
<tr>
<td>Dioscoreales</td>
<td>Dioscorea elephantipes Engl.</td>
<td>SYM 40586 A</td>
<td>AF190082 AF190083</td>
</tr>
<tr>
<td></td>
<td>Tacca chantrieri Andre</td>
<td>SYM 40586 A</td>
<td>AF190082 AF190083</td>
</tr>
<tr>
<td>Liliales</td>
<td>Lilium superbum L.</td>
<td>SYM 40586 A</td>
<td>AF190082 AF190083</td>
</tr>
<tr>
<td></td>
<td>Pleea tenuifolia Michaux</td>
<td>SYM 40586 A</td>
<td>AF190082 AF190083</td>
</tr>
<tr>
<td></td>
<td>Smilax rotundifolia L.</td>
<td>SYM 40586 A</td>
<td>AF190082 AF190083</td>
</tr>
<tr>
<td>Poales</td>
<td>Oryza sativa L.</td>
<td>SYM 40586 A</td>
<td>AF190082 AF190083</td>
</tr>
<tr>
<td></td>
<td>Sorghum bicolor Moench</td>
<td>SYM 40586 A</td>
<td>AF190082 AF190083</td>
</tr>
<tr>
<td>Eudicots:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buxaceae</td>
<td>Pachysandra Michaux sp.</td>
<td>SYM 40586 A</td>
<td>AF190082 AF190083</td>
</tr>
<tr>
<td>Trochodendraceae</td>
<td>Tetracentron sinense Oliver</td>
<td>SYM 40586 A</td>
<td>AF190082 AF190083</td>
</tr>
<tr>
<td></td>
<td>Trochodendron aralioides</td>
<td>SYM 40586 A</td>
<td>AF190082 AF190083</td>
</tr>
<tr>
<td></td>
<td>Siebold &amp; Zucc.</td>
<td>SYM 40586 A</td>
<td>AF190082 AF190083</td>
</tr>
<tr>
<td>Proteales</td>
<td>Nelumbo nucifera Gaertner</td>
<td>SYM 40586 A</td>
<td>AF190082 AF190083</td>
</tr>
</tbody>
</table>
Table 1

(Continued)

<table>
<thead>
<tr>
<th>Clade</th>
<th>Species</th>
<th>Voucher</th>
<th>GenBank accession numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranunculales</td>
<td>Akebia quinata (Houtt.) Dezne.</td>
<td>S. Mathews 446 A</td>
<td>AF276710 AF276711</td>
</tr>
<tr>
<td></td>
<td>Aquilegia L. sp.</td>
<td>nv</td>
<td>AF190066 AF190067</td>
</tr>
<tr>
<td></td>
<td>Coptis trifolia (L.) Salisb.</td>
<td>S. Mathews 459 A</td>
<td>AF276718 AF276719</td>
</tr>
<tr>
<td></td>
<td>Hypericum imberbe Sm.</td>
<td>M. W. Chase 528 K</td>
<td>AF276727 AF276728</td>
</tr>
<tr>
<td></td>
<td>Lardizabala bternata Ruiz &amp; Pav.</td>
<td>S. Mathews 405 A</td>
<td>AF276730 AF276731</td>
</tr>
<tr>
<td></td>
<td>Pteridophyllum racemosum</td>
<td>M. W. Chase 531 K</td>
<td>AF276739 AF276740</td>
</tr>
<tr>
<td></td>
<td>Saxifragales</td>
<td>Xanthorrhiza simplicissima Marshall</td>
<td>AF276750 AF276751</td>
</tr>
<tr>
<td></td>
<td>Brassicales</td>
<td>Arabidopsis thaliana (L.) Heynh.</td>
<td>Published data L21154 X17343</td>
</tr>
<tr>
<td></td>
<td>Solanales</td>
<td>Solanum lycopersicon L.</td>
<td>Published data AJ001915 AF178568</td>
</tr>
</tbody>
</table>

Note. Clade arranged according to APG (1998); nd = not determined; nv = no voucher; BGSC = Botanic Garden of Smith College, courtesy of R. Nicholson; HUBL = Harvard University Biological Laboratories, courtesy of A. MacDonald or P. F. Stevens. Alignments of these sequences are available in TreeBase (http://phylogeny.harvard.edu/treebase) under accession number SN534.

Results

Analyses in Which Sequences Are Terminals

Analysis of the 49 PHYA and the 51 PHYC sequences resulted in eight or two most parsimonious duplicate gene networks in unweighted or weighted parsimony analyses, respectively. The consensus trees from searches with and without weights are nearly identical, differing only with respect to the position of the Piperales in the PHYC subtree. In the consensus from the unweighted search, Piperales are united in a polytomy with eudicots and monocots. In the consensus of the weighted search, they are united with winteroids. Of the 10 shortest networks from both searches, one of the two inferred in the weighted search maximizes identical components in the two gene subtrees and is depicted in figure 1. All of the elements resolved in our analyses of 26 angiosperms appear in this tree. Amborella diverges first from the remaining angiosperms, followed by Nymphaeaceae + Ceratophyllaceae and then by Austrobaileya + Illiciaceae in the PHYC subtree or by Nymphaeaceae and then by Austrobaileya + Ceratophyllaceae in the PHYA subtree. As before, within the remaining angiosperms, a magnolid clade is resolved in which Magnoliaceae (P) are sister to Laurales (M) and winteroids (J) are sister to Piperales (H), as are clades of monocots (F) and eudicots (G). However, a few elements are differently placed. For example, in the PHYA subtree, monocots (F) are sister to the remaining angiosperms. This conflicts with their position in the PHYC subtree and with our earlier results, which placed them as sister to the eudicots. Chloranthaceae (E) are also differently resolved, as sister to Magnoliaceae + Laurales (D) in the PHYC subtree or as sister to the magnoliids + eudicots in the PHYA subtree. In our analysis of 26 angiosperms, Chloranthaceae was represented by Chloranthus only and was sister to the monocots (Mathews and Donoghue 1999).
In general, nodes that were well supported (bootstrap values above 70%) in the analysis of 26 angiosperms are also supported in analyses of the larger phytochrome data set. For example, Nymphaeales (C), Magnoliidae (P), winteroids (J), and Piperales (H) are supported by bootstrap values of 87%–100% and monocots (F) by values of ≥73% (fig. 1). In some cases, additional sampling resulted in better resolution, as was the case with eudicots (G), which are now supported by bootstrap values of 92% and 99%, respectively, in the PHYC and PHYA subtrees (compared with 81% and 96%, respectively, in the analysis of 26 species). Laurales (M) are better supported by the PHYA data in the larger analysis, with a bootstrap value of 76% (compared with 57% in the analysis of 26 species), whereas support by the PHYC data remains high (100% in both analyses). The larger analysis supports the monophyly of the Chloranthaceae (E), with bootstrap values of 99% and 100%, respectively, in the PHYC and PHYA subtrees. The placement of the newly sampled Illicium PHYC with Austrobaileya PHYC is supported by a bootstrap value of 98%. In some cases, however, additional sampling resulted in decreased resolution. Specifically, the position of Amborella remains well supported in the PHYA subtree (bootstrap value of 88%, compared with 92% in the analysis of 26 species) but is less well supported in the PHYC subtree (bootstrap value of 57%, compared with 83% in the analysis of 26 species).

Exclusion of the sequences of Ceratophyllum has a marked effect on resolution of the early divergences in each subtree (fig. 2), but it has a limited impact elsewhere in the subtrees. In both subtrees, the position of Amborella is relatively well supported (82% and 74%, respectively, for the remaining angiosperms in the PHYA and PHYC subtrees). In the PHYC subtree, Nymphaeales diverge next from the remaining angiosperms, followed by Austrobaileya + Illicium (76% and 72%, respectively, for the remaining angiosperms). In the PHYA subtree, Austrobaileya diverges next from the remaining angiosperms, followed by Nymphaeales (65% and 51%, respectively, for the remaining angiosperms). When Ceratophyllum is excluded, the PHYC subtree differs in placing Chloranthaceae (E) as sister to the remaining angiosperms. The PHYA subtree differs when Ceratophyllum is excluded because it places Chloranthaceae (E) as sister to the eudicots (G) and because Piperales (H), winteroids (J), and magnoliids (P) are paraphyletic with respect to eudicots + Chloranthaceae (fig. 2).

In both analyses, relationships within the eudicot clade are well supported in the PHYA subtree (nine of 11 clades with ≥70% bootstrap values). The subtree diverges into two principal clades, one of ranunculids (sensu Hoot et al. 1999), in which Papaverales (V) are united with Ranunculaceae (S) + Lardizabalaceae (U), and one in which Nelumbo, Trochodendrales (W), and Pachysandra are paraphyletic with respect to the core eudicots (X) (figs. 1, 2). In the PHYC subtrees, the ranunculids do not form a clade, and Papaverales (V) are united with Lardizabalaceae (U). The latter relationship receives moderate (62%) or little (<50%) support in analyses that include or exclude (respectively) Ceratophyllum.

To assess the degree of support in the phytochrome data for alternative rootings, we compared networks rooted near Amborella with those resulting from nine searches with different constraints enforced. Individual constraint trees enforced rootings of both subtrees near Nymphaeales + Amborella, Nymphaeales alone, Austrobaileya, Magnoliidae, Laurales, Piperales, Chloranthaceae, winteroids, or monocots. In paired tests, all of the networks resulting from constrained searches are significantly worse ($P < 0.0001$–0.0089 using the Templeton test) than the shortest networks found in unconstrained searches, except for the networks constrained so that each subtree was rooted near Nymphaeales or near Nymphaeales + Amborella. Thus, all rootings other than those near Amborella or Nymphaeales, or near the two together, are rejected by the phytochrome data.

None of the conflicting nodes resolved by the PHYA and PHYC subtrees is supported above 50% in bootstrap analyses (figs. 1, 2). However, paired tests revealed that the PHYA data reject the PHYC tree and that the PHYC data reject the PHYA tree, regardless of whether Ceratophyllum is included in the analyses ($P < 0.0001$ for both pairwise comparisons using the Templeton test). Furthermore, partition homogeneity tests reject the hypothesis that the single-gene data sets are homogeneous ($P = 0.02$). The latter result may reflect the fact that PHYA apparently is evolving significantly faster than PHYC in most of the species sampled (S. Mathews, unpublished data).

We used constraint analyses to determine whether this conflict could be attributed to specific nodes of the subtrees. Specifically, we constrained three searches of the PHYA data such that in each, the resulting tree matched the PHYC subtree with respect to the placement of one of the following: (1) monocots, (2) Chloranthaceae, or (3) Ceratophyllum. Likewise, we constrained three searches of the PHYC data such that in each, the resulting tree matched the PHYA subtree with respect to placement of one of the same three taxa. The resulting sets of trees were compared with the PHYA and PHYC subtrees, respectively, from unconstrained analyses. None of the trees resulting from constraint analyses is rejected by the PHYA or PHYC data ($P > 0.1$ for all pairwise comparisons using the Templeton test).

Although we cannot attribute the conflict that we observed to any of the single nodes that we tested, the evidence that we obtained supporting significant conflict in more general tests might serve as an argument against combining data from the genes (e.g., Bull et al. 1993). Nonetheless, we conducted a combined analysis in order to compare the results with those obtained in gene tree parsimony analyses.

Analyses in Which Species Are Terminals

When Ceratophyllum is included, both weighted and unweighted parsimony analyses of the combined data resolved the same set of three most parsimonious trees, the consensus of which is depicted in figure 3. Basal relationships are poorly supported by bootstrap values (<50%) and differ in two ways from those resolved in the gene subtrees that included Ceratophyllum. Austrobaileya + Illicium, rather than Nymphaeales, diverge from the remaining angiosperms after Amborella, and Chloranthaceae diverge next. Nymphaeales + Ceratophyllum are sister to the remaining angiosperms. This branching order is surprising because Nymphaeales branched before Austrobaileya in all PHYC subtrees and in some PHYA subtrees. Moreover, none of the subtrees resolved Chloranthaceae as diverging before Nymphaeales (e.g., figs. 1, 2). The remaining angiosperms are resolved as they were in the PHYC
Fig. 2  Strict consensus of two most parsimonious networks of PHYA and PHYC from 51 angiosperms (Ceratophyllum excluded; 1303 nucleotide sites, 834 parsimony informative); weighted parsimony analysis (100 random taxon addition replicates with TBR swapping in PAUP* 4.0 [Swofford 2000]) yielded most parsimonious networks of 10,575.00 steps (retention index [RI] = 0.54; consistency index [CI] = 0.16, excluding autapomorphies). Bootstrap percentages (from 100 replicates with the same search parameters, but using 10 random addition replicates) are shown above branches. Identical components in the PHYA and PHYC subtrees are labeled A through DD.
Fig. 3  Strict consensus of three most parsimonious trees from combined analysis of PHYA and PHYC, with species as terminals (2606 nucleotide sites; 1408 parsimony informative), rooted near *Amborella*, based on the duplicate gene analysis (Figs. 1, 2); weighted parsimony analysis (100 random taxon addition replicates with TBR swapping in PAUP* 4.0 [Swofford 2000]) yielded trees of 11,821.75 steps (retention index RI = 0.41; consistency index CI = 0.23, excluding autapomorphies). Bootstrap percentages (from 500 replicates with the same search parameters but using 10 random addition replicates) are shown above branches. Components identical to those found in both PHYA and PHYC subtrees are labeled A through DD.

Overall, this result contrasts with results from our analyses of 26 angiosperms, in which combining the data led to better support for the earliest divergences as well as for some of the major clades (Mathews and Donoghue 1999). Constraint analyses reveal that the tree from the combined data is rejected by the PHYA data ($P < 0.0001$ using the Templeton test) but not by the PHYC data ($P > 0.3247$).

As in our analyses of the sequences as terminals, excluding Ceratophyllum from analyses of the combined data resulted in better resolution at the base of the tree (Fig. 4). After Amborella, Nymphaeales diverge first from the remaining angio-
Fig. 4 Bootstrap consensus (50% majority rule) from combined analysis of PHYA and PHYC, with sequences as terminals (2606 nucleotide sites; 1399 parsimony informative) when Ceratophyllum is excluded, rooted near Amborella, based on the duplicate gene analysis (figs. 1, 2); weighted parsimony analysis (100 random taxon addition replicates with TBR swapping in PAUP* 4.0 [Swofford 2000]) yielded trees of 11,495.50 steps (retention index [RI] = 0.42; consistency index [CI] = 0.23, excluding autapomorphies). Bootstrap percentages (from 100 replicates with the same search parameters but using 10 random addition replicates) are shown above branches. Components identical to those found in both PHYA and PHYC subtrees are labeled A through DD.

In our analyses of 26 angiosperms, the PHYA and PHYC subtrees without Ceratophyllum conflict with respect to the position of Austrobaileya (fig. 2), but the combined data (fig. 4) support the separation of Nymphaeales first (bootstrap value of 78% for the remaining angiosperms), followed by Austrobaileya + Illicium (bootstrap value of 68% for the remaining angiosperms). In general, clades within the remaining angiosperms are supported at similar levels in analyses with and without Ceratophyllum. For example, eudicots (G), Piperales (H), winteroids (J), Magnoliidae (P), Laurales (M), and Chloranthaceae (E) are supported by bootstrap values ≥95%. However, a few clades are better supported when Ceratophyllum is excluded, including monocots (F), Magnoliidae + Laurales...
Fig. 5 Strict consensus of 64 most parsimonious rooted species trees resulting from analysis of the gene network in figure 1 by using gene tree parsimony implemented with GeneTree (Page 1998), minimizing duplications and losses (cost = 14 duplications and 52 losses).
tions gave rise to paralogues that were not sampled or were lost (e.g., Page 1994). However, a number of conflicts between the PHYA and PHYC subtrees are not well supported (see above) and likely result from causes other than undetected paralogy.

Discussion

The analyses presented here of PHYA and PHYC from many more taxa uphold most of the conclusions of our earlier phytochrome analyses and provide some additional insights. *Amborella* still appears to be the sister group of all other angiosperms. Notably, however, we cannot reject alternative rootings in which water lilies, either alone or in combination with *Amborella*, are basal. *Austrobaileya* still appears to be the sister group of the remaining angiosperms and is united with the newly sampled *Illicium*. Within the remaining angiosperms, eudicots and monocots form rather well-supported clades, as do Magnoliolae, Laurales, Piperales, and winteroids. We find some support for a clade within which Laurales are united with Magnoliolae and Piperales with winteroids. Interestingly, *Aristolochia* and *Saruma* are not united in the phytochrome trees, and in our larger analysis, at least in PHYA subtrees, monocots diverge first from the remaining angiosperms, followed by Chloranthaceae. Within eudicots we find support for a basal split between ranunculids (Ranunculales, Papaverales) and the rest, though the position of *Nelumbo* remains equivocal. *Acorus* is not sister to the rest of the monocots, as it is in some plastid phylogenies (Duvall et al. 1993).

The use of duplicated phytochrome genes to infer angiosperm phylogeny is an important complement to recent analyses based on nuclear ribosomal and organellar DNA sequences. We have sampled protein coding nuclear DNA sequences, and our analyses do not depend on the inclusion of divergent outgroup sequences. The relationships at the base of the angiosperms that we describe above are basically the same as those resolved in analyses (with outgroups) of data combined from plastid, mitochondrial, and nuclear ribosomal genes (Parkinson et al. 1999; Qiu et al. 1999). Moreover, many of the relationships among the remaining angiosperms that we find are more equivocal. Unfortunately, as in Qiu et al. (1999), the positions of eudicots, magnoliolae, Chloranthaceae, and monocots are not well supported (bootstrap values <50%). Similarly, the position of *Ceratophyllum* remains equivocal. Our results imply that it diverged early rather than with monocots (as in Qiu et al. 1999), but neither position is well supported.

Duplicated gene rooting, which seldom has been employed when outgroups are available (but see Sang et al. 1997; Telford and Holland 1997), is likely to prove useful elsewhere. Single-copy nuclear genes appear to be rare (Page 2000), and gene duplications may have occurred along many branches of interest. However, to make the most of gene duplications, more attention to analytical issues is needed. Our previous discussion of duplicate gene rooting distinguished between a “reciprocal outgroups” and a “minimum events” interpretation (Donoghue and Mathews 1998). Under the reciprocal outgroups view (see Doolittle and Brown 1994), sequences of one of the gene copies are viewed as outgroups for the other, and vice versa, and a rooted species tree is derived by consensus of the two gene subtrees. In contrast, under the minimum events view, the best-rooted species tree is the one that minimizes additional duplications and losses, lineage sorting, and lateral transfer events in the gene tree (e.g., Donoghue and Mathews 1998, fig. 2; see discussion of gene tree parsimony in Slowinski and Page 1999). In theory we prefer the minimum events interpretation, but we know of no algorithm to infer reconciled trees that deals simultaneously with the full range of events that might lead to incongruent gene trees and which also takes into account the underlying support for the gene tree. This is a difficult optimization problem, even if relative weights could be assigned beforehand to the different processes (Maddison 1997). In the meantime, to partially implement the logic of the minimum events approach, we have used Page’s (1998) GeneTree program to identify species trees that minimize gene duplications and losses.

Weston (1994) provided yet another interpretation of duplicate gene rooting. This hinges on assessing the polarity of individual characters by determining the distribution of states across paralogous forms of a gene. If one nucleotide, for example, A, is universally present at a particular site in one form of the gene but there is variation at that site in the second form of the gene (e.g., A is found in some species and C in others), then the nucleotide present in both gene forms (A) is considered ancestral and the alternative state (C) derived. A polarized character implies that the root does not lie near species with the derived condition. Therefore, by assessing the polarity of each of the characters that vary across paralogous genes, it may be possible to infer a rooted species tree.

Regardless of interpretation, duplicate gene rooting allows the inference of a rooted species tree without inclusion of sequences from outgroups. Previously (Donoghue and Mathews 1998; Mathews and Donoghue 1999) we argued that this could be useful if outgroup sequences were highly diverged from ingroup sequences that their inclusion would create long branch artifacts. In effect, a duplication occurring along the branch to the ingroup would bisect the long branch connecting the ingroup with outgroups (fig. 6A, 6B). However, as we also noted (Donoghue and Mathews 1998), the impact of this section will depend on where along the branch the duplication occurred and on how rapidly the genes have evolved (fig. 6B–6F). If sequences evolve in a clocklike manner, the branch between two paralogues will always be shorter than the branch between a paralogue and an outgroup, though the difference may be slight depending on the time elapsed before the duplication (fig. 6B, 6C). In contrast, if evolution is not clocklike, there may be cases in which the branch that separates paralogues would be longer than the branch between one or both of the paralogues and the outgroup sequences (fig. 6D). Even when rates of evolution differ between paralogues, these rates may be less diverged from one another than from outgroup sequences (fig. 6E, 6F). Clearly, the impact of excluding outgroups from analyses of duplicate genes is a complex function of when the duplication occurred and the rates of evolution. It is unclear at this time which of the patterns depicted in figure 6 will be encountered most often (cf. Li and Gojobori 1983; Goodman et al. 1987; Iwabe et al. 1996; Cronn et al. 1999). But as the phytochrome analyses show, even when rates of divergence are high, duplicate gene rooting can provide valuable insight.

We also noted previously the need for more attention to be
Fig. 6 The effect of duplication time and rates of nucleotide evolution on branch lengths between a pair of duplicated genes and outgroup sequences. Hatch marks indicate numbers of changes along branches (OG = outgroup; 1 = paralogue 1; 2 = paralogue 2). a, Rooted tree of one outgroup sequence and three sequences of one of the duplicated genes. b–f, Rooted trees of one outgroup sequence and six paralogous ingroup sequences resulting from a late duplication when evolution is clocklike (b), an early duplication when evolution is clocklike (c), an early duplication when evolution of the outgroup is slow relative to the paralogues (d), an early duplication when evolution of one paralogue is slow relative to the other paralogue and the outgroup (e), and a late duplication when the rate of one paralogue is high relative to the other paralogue and the outgroup (f).

paid to the treatment of unduplicated outgroup sequences in analyses of duplicated genes to obtain a rooted species tree (Donoghue and Mathews 1998). Simmons et al. (2000) proposed “uninode coding” to combine unduplicated and duplicated sequences in a matrix in which species are terminals. After an initial analysis of all gene copies to establish the phylogenetic position of gene duplications, hypothetical ancestral sequences are reconstructed, each corresponding to ancestors at inferred duplications. A matrix is then constructed in which each species is scored for each form of the gene. A single duplication yields three forms of a gene—the unduplicated form and the two duplicated forms. Species with an unduplicated form are coded as having the hypothetical ancestral sequence for the two duplicated forms of the gene (fig. 7A). Species with two forms are coded as having the hypothetical ancestral sequence for the unduplicated form of the gene. A single binary character is added to reflect the duplication event, and the matrix is analyzed to obtain a species tree.

Uninode coding is designed to resolve the issue of including outgroups with unduplicated sequences when duplicate genes are combined for analysis. But in some cases it will not be appropriate, and in others it may not be the best of the available approaches. Simmons et al. (2000) noted that uninode coding should not be used if duplicate loci are subject to concerted evolution. Similarly, it seems inappropriate when other processes, such as duplication and loss, lineage sorting, and lateral transfer, have caused genuine and strong conflicts between duplicate gene trees. In these cases, which may be common in multigene families (Morton et al. 1996; Clegg et al. 1997), approaches that do not involve data combination may be more appropriate (see de Quieroz et al. 1995). In other cases, analyses that exclude outgroups will be useful for comparison with results from analyses in which artifacts are suspected to result from long branches (Felsenstein 1978; Kim 1996) that lead to outgroups.

In cases in which uninode coding does seem appropriate, there are other concerns. First, because it is a sequential procedure, final estimation of the species phylogeny may be based on erroneous assumptions (e.g., Maddison et al. 1984; Nixon and Carpenter 1993). For example, if the initial gene tree is not well supported and/or the gene family is not well sampled, a uninode analysis may be built on erroneous inferences about hypothetical ancestors. Second, the use of hypothetical ancestors increases the number of polymorphic characters and taxa in uninode matrices, which may sometimes be problematic (e.g., Nixon and Davis 1991).

There also are significant challenges involved in the construction of a uninode matrix. Simmons et al. (2000) asserted that the method could be generalized to cases in which there are multiple duplications, but their reanalysis of phytochrome data from Donoghue and Mathews (1998) failed to demonstrate this. Our analysis provided evidence of four gene duplications (Donoghue and Mathews 1998, fig. 3), as illustrated by the gene tree in figure 7B. Simmons et al. (2000) based their uninode analysis on just part of our tree, ignoring duplications leading to the two forms in seed plants and the additional forms within angiosperms (cf. Donoghue and Mathews 1998, fig. 3; Simmons et al. 2000, fig. 1b). The sequences
Fig. 7  Schematic representations of gene phylogenies within their containing species phylogenies and their corresponding uninode matrices. Filled circles at duplication points in the gene phylogenies represent hypothetical ancestors (HA) of two descendant gene lineages.  

(a) Summary of phytochrome gene phylogeny presented by Simmons et al. (2000), showing only a single gene duplication.  

(b) Summary of the phytochrome gene phylogeny presented by Donoghue and Mathews (1998), showing four gene duplications. Characters to reflect gene duplication events (Simmons et al. 2000) are omitted from the corresponding matrix shown here.

Excluded from their analysis (e.g., PHYB, PHYD, and PHYE) might bear significantly on the reconstruction of hypothetical ancestors. Furthermore, Simmons et al. (2000) did not precisely indicate how species should be coded when a series of gene duplications are nested within one another (e.g., how angiosperms should be scored for PHY2 and PHY3 in fig. 7B). Reanalysis of the data in Donoghue and Mathews (1998, fig. 3) should have included four hypothetical ancestors and eight gene forms (fig. 7B). A uninode matrix for the phytochrome sequences now available in GenBank would include at least 11 forms of the gene, since at a minimum, there are two copies in Adiantum, three in nonflowering seed plants, and three to five copies in angiosperms. In view of the uncertainty surrounding uninode coding when duplications are nested within one another, we have not presented a uninode analysis of our present data set. However, just as the uninode results of Simmons et al. (2000) were congruent with our previous analyses, our experiments with uninode coding have yielded results that are similar in all major respects to the results we report here based on combined analysis of PHYA and PHYC and on reconciled trees that minimize gene duplications and losses.

Conclusions

Our results provide insights into relationships at the base of the angiosperm tree and are an important complement to
recent analyses based on nuclear ribosomal and organellar DNA sequences. They also imply that there may be diminishing returns in much more intensive sampling of PHYA and PHYC to resolve relationships among monocots, magnoliids, Chloranthaceae, Ceratophyllum, and eudicots. Instead, we expect phytochrome sequences to provide convincing resolution of more recent divergences (e.g., Mathews and Sharrock 1996; Lavin et al. 1998; Mathews et al. 2000; Simmons et al., in press).

Our results also highlight the use of duplicate gene rooting, which is likely to prove useful elsewhere. Within the phytochrome gene family, other duplications may be useful for resolving relationships among seed plant lineages and within eudicots. But as we noted, more attention to analytical issues is needed in order to make the most of gene duplications in this way. Specifically, methods are needed to infer reconciled trees that simultaneously consider several different processes leading to incongruence (e.g., gene duplication and loss, lineage sorting) and that take into account the possibility that the underlying gene trees may be wrong or only weakly supported.

Acknowledgments

We thank Mike Crisp, Jim Doyle, Ned Friedman, Mike Sanders, and Bob Sharrock for helpful discussions; Mark Simons for sharing manuscripts; Bil Alversed, Ron Beckwith, Mark Chase, Peter Endress, Beth Farnsworth, Taylor Feild, Bill Hahn, Sara Hoot, Larry Kelly, Andrew MacDonald, Rob Nicholson, Yin-Long Qiu, Mark Simmons, Peter Stevens, Elaine Tobin, and the Royal Botanic Garden, Sydney, for DNA or tissues; Lisa Racki and Christina SooHoo for technical assistance; and Rick Ree for help with figure 6. This work was supported by NSF grant DEB-9806937.

Literature Cited


Felsenstein J 1978 Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27:401–410.


Hecker KH, KH Roux 1996 High and low annealing temperatures increase both specificity and yield in touchdown and stepdown PCR. Biotechniques 20:478–485.


Howe GT, PA Bucciaglia, WP Hackert, GR Furnier, M-M Cordonnier-Pratt, G Gardner 1998 Evidence that the phytochrome gene family in black cottonwood has one PHYA locus and two PHYB loci but lacks members of the PHYCF and PHYE subfamilies. Mol Biol Evol 15:160–175.


Mathews S, RA Sharrock 1996 The phytochrome gene family in grasses (Poaceae): a phylogeny and evidence that grasses have a
Mathews S, RC Tsai, EA Kellogg 2000 Phylogenetic structure in the
grass family (Poaceae); evidence from the nuclear gene phytochrome
Morton BR, B Gaut, and MT Clegg 1996 Evolution of alcohol de-
hydrogenase genes in the palm and grass families. Proc Natl Acad
Sci USA 93:11735–11739.
Nixon KC, JJ Davis 1991 Polymorphic taxa, missing values and clad-
Page RDM 1994 Maps between trees and cladistic analysis of his-
torical associations among genes, organisms, and areas. Syst Biol
——— 1998 GeneTree: a program for comparing gene and species
phylogenies on computers running Mac OS or Microsoft Windows
95/NT. University of Glasgow, Glasgow.
——— 2000 Extracting species trees from complex gene trees: re-
conciled trees and vertebrate phylogeny. Mol Phylogenet Evol 14:
89–106.
Parkinson CL, KL Adams, JD Palmer 1999 Multigene analyses iden-
tify the three earliest lineages in extant flowering plants. Curr Biol
Qin M, R Kuhn, S Moran, PH Quail 1997 Overexpressed phyto-
chrome C has similar photosensory specificity to phytochrome B
but a distinctive capacity to enhance primary leaf expansion. Plant
J 12:1163–1172.
Qiu Y-L, J Lee, F Bernasconi-Quadroni, DE Soltis, PS Soltis, M Zanis,
EA Zimmer, Z Chen, V Savolainen, MW Chase 1999 The earliest
angiosperms: evidence from mitochondrial, plastid and nuclear ge-
Quail PH 1994 Phytochrome genes and their expression. Pages
71–104 in RE Kendrick, GHM Kronenberg, eds. Photomorphoge-
Rambaut A 1996 Se-Al: sequence alignment editor, version 1.0 alpha
Sang T, MJ Donoghue, D Zhang 1997 Evolution of alcohol dehy-
drogenase genes in peonies (Paeonia): phylogenetic relationships of
Sharrock RA, PH Quail 1989 Novel phytochrome sequences in Ar-
abidopsis thaliana: structure, evolution and differential expression
Simmons MP, CD Bailey, KC Nixon 2000 Phylogeny reconstruction
Simmons MP, CC Clevenger, V Savolainen, RH Archer, S Mathews,
JJ Doyle In press Phylogeny of the Celastraceae inferred from phy-
Slowinski JB, RDM Page 1999 How should species phylogenies be
Soltis PS, DE Soltis, MW Chase 1999 Angiosperm phylogeny inferred
from multiple genes as a tool for comparative biology. Nature 402:
402–404.
Soltis PS, DE Soltis, M Zanis, S Kim 2000 Basal lineages of angio-
sperms: relationships and implications for floral evolution. Int J
Telford MJ, WH Holland 1997 Evolution of 28S ribosomal DNA in
chaetognaths: duplicate genes and molecular phylogeny. J Mol Evol
44:135–144.
Wada M, T Kangae, K Nozue, S Fukuda 1997 Cryptogram phyto-
in RW Scotland, DJ Siebert, DM Williams, eds. Models in phylogeny
Whitelam GC, PF Devlin 1997 Roles of different phytochromes in