Duplicate Genes and the Root of Angiosperms, with an Example Using Phytochrome Sequences

Michael J. Donoghue and Sarah Mathews

Department of Organismic and Evolutionary Biology, Harvard University Herbaria, 22 Divinity Avenue, Cambridge, Massachusetts 02138

Received August 12, 1997; revised December 16, 1997

The root of the angiosperm tree has not yet been established. Major morphological and molecular differences between angiosperms and other seed plants have introduced ambiguities and possibly spurious results. Because it is unlikely that extant species more closely related to angiosperms will be discovered, and because relevant fossils will almost certainly not yield molecular data, the use of duplicate genes for rooting purposes may provide the best hope of a solution. Simultaneous analysis of the genes resulting from a gene duplication event along the branch subtending angiosperms would yield an unrooted network, wherein two congruent gene trees should be connected by a single branch. In these circumstances the best rooted species tree is the one that corresponds to the two gene trees when the network is rooted along the connecting branch. In general, this approach can be viewed as choosing among rooted species trees by minimizing hypothesized events such as gene duplication, gene loss, lineage sorting, and lateral transfer. Of those gene families that are potentially relevant to the angiosperm problem, phytochrome genes warrant special attention. Phylogenetic analysis of a sample of complete phytochrome (PHY) sequences implies that an initial duplication event preceded (or occurred early within) the radiation of seed plants and that each of the two resulting copies duplicated again. In one of these cases, leading to the PHYA and PHYC lineages, duplication appears to have occurred before the diversification of angiosperms. Duplicate gene trees are congruent in these broad analyses, but the sample of sequences is too limited to provide much insight into the rooting question. Preliminary analyses of partial PHYA and PHYC sequences from several presumably basal angiosperm lineages are promising, but more data are needed to critically evaluate the power of these genes to resolve the angiosperm radiation.

INTRODUCTION

Over the past decade our understanding of the phylogeny of green plants has progressed a great deal (Donoghue, 1994; Crane et al., 1995; Kenrick and Crane, 1997). However, despite considerable attention, some significant issues remain unresolved. Conspicuously, the root of the angiosperm tree remains uncertain. Morphological and molecular datasets have yielded alternative, poorly supported rootings, with very different implications for angiosperm evolution (see below). Equivocal results from these studies may result from a similar cause, namely that angiosperms are very different from their closest living relatives and may have radiated rather quickly. Virtually all phylogenetic analyses of seed plants show an exceptionally long branch subtending modern angiosperms, indicating a large number of changes since their divergence from other extant groups (see below). In morphological analyses, this causes difficulty in interpreting the homology of key structures (Doyle, 1994), which has necessitated either the omission of characters or the use of ambiguity coding in numerical phylogenetic analyses. Such expedients may render several outgroup rootings equally parsimonious or nearly so. In molecular analyses the problem may be even more difficult because comparisons are limited to extant species. The inclusion of poorly aligned sequences from distantly related taxa could lead to false estimates of relationship, whereas their exclusion may result in poor resolution (Wheeler et al., 1995). Another problem is potentially more serious, namely the possible spurious attachment of long outgroup branches to the longest branches within angiosperms owing to convergence at homologous nucleotide sites (Felsenstein, 1978; Hendy and Penny, 1989; Wheeler, 1990; Huelsenbeck and Hillis, 1993; Kim, 1996; Huelsenbeck, 1997). When long ingroup branches are present, the rooting may be ambiguous or, in the worst case, unambiguous but wrong.

How might these problems be overcome? One approach would be to obtain additional characters and, perhaps, characters of greater relevance to the problem at hand. For example, careful genetic and developmental comparisons could elucidate the homology of morphological structures in angiosperms and their extant relatives (Doyle, 1994; Frohlich and Meyerowitz, 1997). Combined analysis of nucleotides from a variety
of relatively slowly evolving genes, each of which contributes a small amount of relevant signal, also will be useful (Soltis et al., in press; but see Phillippe et al., 1994, on the number of characters that may be needed). The discovery of structural molecular characters (e.g., rearrangements in plastid genomes: Jansen and Palmer, 1987; Manhart and Palmer, 1990; Raubeson and Jansen, 1992; or indel characters: Goremykin et al., 1996) or major shifts in function (e.g., type of intron splicing: Qiu and Palmer, 1997) may also help position the angiosperm root.

Ideally, however, one also would like to shorten the distance between angiosperms and their outgroups (Hillis, 1996; Kim, 1996). Addition of fossil taxa along the line leading to angiosperms (i.e., other angiophytes; Doyle and Donoghue, 1993) would accomplish this for morphological analyses (see Donoghue et al., 1989). Unfortunately, no fossils are widely accepted as attaching along the appropriate branch, and those that could be relevant are incomplete and may be of little use (Crane, 1987; Doyle and Donoghue, 1993; but see Doyle, 1996). Even if better fossils are discovered, they will most likely not reduce branch lengths in trees inferred from molecular data (Austin et al., 1997). Duplicate gene rooting, developed to root the entire tree of life (Iwabe et al., 1989; Gogarten et al., 1989; reviewed by Doolittle and Brown, 1994), could provide a solution in such cases. If a gene were to have duplicated along the long branch leading to angiosperms, the resulting forms of the gene would share a more recent common ancestor than either one would with the single gene present in outgroup taxa. Inclusion of both forms of the gene in a phylogenetic analysis would therefore effectively bisect the branch leading to angiosperms.

The aim of the present paper is to explore the use of duplicate genes in rooting the angiosperm tree. After providing background on the angiosperm problem itself, and on the logic of duplicate gene rooting, we develop a search image for relevant gene duplications and suggest several possibilities. As a concrete example, we present preliminary analysis from the phytochrome gene family. Although these genes do not yet provide convincing evidence on the angiosperm root, our results illustrate how simultaneous analyses of duplicate genes can increase resolution and may eventually provide more definitive answers. In the meantime, our exploratory studies highlight the logic and possible limitations of the procedure, and we hope that they will encourage the development of relevant theory and analytical tools.

THE ROOT OF ANGIOSPERMS

The root of the angiosperm tree has attracted a great deal of attention without, however, yielding a satisfactory solution (Fig. 1). Morphological analyses have been inconclusive. Donoghue and Doyle (1989) found that their shortest trees were rooted in the vicinity of Magnoliidae (Fig. 1a), but that trees only one step longer were rooted among paleoherbs, along a branch leading to water lilies (Nymphaeales) and monocots. The trees presented by Loconte and Stevenson (1991), Taylor and Hickey (1992), Doyle et al. (1994), Nixon et al. (1994), Loconte (1996), and Doyle (1996) differ both from the Donoghue and Doyle results and from one another as regards the position of the root (Figs. 1b–1g). It also should be noted that in most of these cases there are several rather different possibilities among the equally parsimonious trees or in trees that are only one or two steps longer. The Nixon et al. (1994) analysis provides a good example. Their strict consensus of 834 most parsimonious trees is poorly resolved at the base of the angiosperms, but reanalysis of their data shows that these trees fall into just two categories as regards the root. In two islands of trees (of 108 and 144 trees), Chloranthus is the sister group of the other 17 angiosperms in the analysis, followed by Ceratophyllum (Fig. 1e). In the other two tree islands (of 97 and 485 trees), Casuarina appears as the sister group of the rest of the angiosperms, followed by Betula (Fig. 1f). These trees have very different implications for the evolution of angiosperm characters (e.g., pollen morphology), as well as for diversification patterns and interpretation of the fossil record.

In these morphological studies, one likely source of ambiguity is the long branch subtending angiosperms. For example, in a tree shown by Nixon et al. (1994; their Figs. 4–6) there are 17 character changes along the angiosperm branch, 15 along the anthophyte branch (angiosperms, Gnetales, and Bennettitales), and 12 along the Ceratophyllum branch. All of the remaining branches have less than 10 changes, with an average of approximately 3. Likewise, in the morphological tree of extant seed plants featured by Doyle et al. (1994; their Fig. 9) the angiosperm branch is marked by 15 ambiguous character changes, compared with 8 such changes for genetophytes and for Piperales, 7 for gnetophytes plus angiosperms, and an average of about 2.5 for all remaining branches. Missing data also may obscure the position of the root in morphological analyses. Approximately 10% of the characters used by Nixon et al. (1994) could not be scored for any nonangiosperm group in their analysis. Similarly, in Doyle's (1996) analysis, about 13% of the characters are scored as unknown outside of angiosperms. These include perianth, stamen, carpel, and seed characters, for which homologies with structures outside of angiosperms remain highly uncertain.

Individual molecular studies show similar instability regarding the position of the root. A good example is provided by analyses of the chloroplast gene rbcL. Although the studies by Chase et al. (1993; see Fig. 1j) and Rice et al. (1997; Fig. 1n) show Ceratophyllum as the sister group to the rest of the angiosperms, support
FIG. 1. Alternative rootings of the angiosperm tree based on morphological data (a–g), molecular data (h–n), and combined analyses (o–p). Four terminal branches are shown in each case, one of which represents the rest of the taxa in the analysis (labeled ETC.). Such tree simplification requires choices as to which names to highlight (differential resolution, sensu O’Hara, 1992), but we have tried to capture relationships among major lineages implied by trees presented in the literature. The analyses represented here included many, but not all, of the same terminal taxa; overlap of taxa is sufficient to conclude that different trees could not be reconciled through simple insertion or deletion of terminals. MAGN, Magnoliales; LAUR, Laurales; CHLO, Chloranthaceae; WINT, Winteraceae; ILLI, Illiciales; CALY, Calycanthaceae; AUST, Austrobaileyae; PIPE, Piperales; MONO, monocotyledons; NYMP, Nymphaeales; ARIS, Aristolochiaceae; CERA, Ceratophyllum; CASU, Casuarina; BETU, Betula; CHRY, Chrysolepis; HAMA, Hamamelis; SCHI, Schisandraceae; EUDI, eudicots; AMBO, Amborella (circumscription of several of these groups differs somewhat among authors). A plus sign indicates that other taxa also belong in the clade.
for this rooting was not tested owing to the large size of the dataset. This arrangement does not appear in the parsimony jackknife tree of Farris et al. (1996; V. Albert, personal communication), nor is it found consistently in smaller, more thorough rbcL analyses. Qiu et al. (1993) found the position of Ceratophyllum to be equivocal in analyses focused on magnoliid taxa and dependent on the inclusion of particular outgroups. Likewise, Sytsma and Baum (1996) found this arrangement to be poorly supported by bootstrap and decay analyses in a study involving 109 sequences. Their analyses also demonstrated that the rooting based on rbcL sequences is highly sensitive to exactly which other seed plant sequences are included in the analysis (Fig. 12.6 in Sytsma and Baum, 1996; also see Nickrent and Soltis, 1995).

The position of the root also differs considerably among molecular analyses (Figs. 1h–1n). Studies based on rbcL sequences (Chase et al., 1993; Rice et al., 1997), on rbcS amino acid sequences (Martin and Dowd, 1991), and on ribosomal sequences (Hamby and Zimmer, 1992; Doyle et al., 1994; Chaw et al., 1997; Soltis et al., 1997) have produced quite varied results (also see Goremykin et al., 1996, on cpITS sequences; Martin et al., 1993, on the nuclear gene gapC). In each of these cases support for the position of the root has been rather weak (with bootstrap values generally well under 50% for major clades near the base of the tree), and many alternatives can be found among trees that are almost as parsimonious. Again, these trees have rather different implications concerning the early evolution of angiosperms and the factors involved in their diversification.

As in morphological analyses, the branch subtending angiosperms tends to be especially long in molecular trees. For example, in the rbcL analysis presented by Chase et al. (1993; their Fig. 2B), 59 changes separate the angiosperms, and 61 steps separate the gnophytes, from their inferred common ancestor. These are the longest internal branches in the entire tree. Ceratophyllum is the longest branch near the base of the tree, with 44 steps shown on its branch and 22 more along the branch subtending the rest of the angiosperms. This observation has suggested the possibility that the Ceratophyllum rooting is an artifact of long branch attraction (Donoghue, 1994). Similarly, in the Chaw et al. (1997) analysis of 18S rRNA, 45 character changes are shown on the angiosperm branch, and 14 more on the sister branch subtending the gymnosperms (which form a clade in this analysis), for a total of 59 differences separating angiosperms from other seed plants (Chaw et al., 1997; their Fig. 2). Within the rest of the seed plants, the only other difference of this magnitude is that between the gnophyte clade and its sister group, the conifers (a total of 61 steps). The two ferns and the lycopod used by Chaw et al. (1997) to root the seed plant tree are all extremely divergent from seed plants, and their attachment along the angiosperm branch (as opposed to within “gymnosperms”) may be caused by long branch attraction. Likewise, the position of the root of the angiosperms in this analysis falls on one of the longest internal angiosperm branches (32 steps), separating water lilies from all the rest.

These problems are not overcome in combined morphological and molecular analyses (Figs. 1o and 1p). Ceratophyllum is the sister group of the rest of the angiosperms in most, but not all, of the combined analyses of rbcL and morphology presented by Albert et al. (1994), with woody magnoliids variously situated near the base (Fig. 1o). In contrast, in the combined rRNA and morphology trees of Doyle et al. (1994) the angiosperms are rooted among paleoherbs, and woody magnoliids are nested well within the tree (Fig. 1p). As expected, the branch separating angiosperms from the rest of the seed plants, and especially from the gnetophytes, becomes even longer in combined analyses, with minimal positive impact on confidence in the position of the angiosperm root. Thus, in the best tree from the combination of rRNA sequences and morphology (Doyle et al., 1994) there are 40 unambiguous changes on the branch subtending angiosperms and 26 marking their sister group, the gnetophytes; most other branches are marked by less than 10 changes. Although bootstrap percentages associated with several basal branches within angiosperms increase somewhat in the combined analyses, the highest value is 65%, and the first three branches are all marked by a decay index of 1 (Doyle et al., 1994; their Fig. 16). Although these numbers indicate that the exact position of the root is unstable, Doyle et al. (1994) did find that rootings near Magnoliaceae, Chloranthaceae, or Calycanthaceae were considerably worse, requiring at least 13 extra steps.

This brief survey is intended only to highlight the fact that the position of the angiosperm root is unclear. This is not to say, however, that we know nothing about the root. Indeed, a number of possibilities have been more or less eliminated in recent years. For example, it now seems unlikely that the root lies within monocots or within eudicots, and it is probably not within core Magnoliaceae, Laurales, Piperales, etc. While this is surely progress, we remain unable to discriminate among phylogenetic hypotheses that differ a great deal in their evolutionary implications.

**DUPLICATE GENE ROOTING**

The use of duplicate genes for rooting purposes was developed by Iwabe et al. (1989) and Gogarten et al. (1989) for rooting the entire tree of life. In this case, outgroups are unavailable, and some other means of
rooting is necessary. The basic strategy is to identify genes that are present in two copies in all organisms and to include both copies of the gene (from some sample of organisms) in an unrooted analysis. The expectation is that each form of the gene will show the same relationships among the taxa, under the assumption that the evolution of each gene coincided with the species phylogeny (but see, e.g., Maddison, 1997). This expectation has been upheld in several cases examined to date (Gogarten et al., 1989; Iwabe et al., 1989; Brown and Doolittle, 1995; Lawson et al., 1996) and has yielded the widely accepted conclusion that (eu)bacteria are the sister group of archaea plus eukaryotes (Doolittle and Brown, 1994; but see Forterre et al., 1993; Hilario and Gogarten, 1993; Creti et al., 1994).

Several different explanations of the duplicate gene rooting procedure have been presented (see below), implying that more attention is needed to the underlying logic and the optimality criterion employed. Furthermore, although this approach could be used in any case involving gene duplication, there have been few such uses (but see, e.g., Ford et al., 1995; Gottlieb and Ford, 1996; Guigo et al., 1996; Telford and Holland, 1997), presumably because in most cases it has been possible to include outgroup sequences. Therefore, little is known of the behavior of the method in real cases, and little attention has been paid to how to interpret the results if there are unresolved relationships or differences between the duplicate gene trees. Also, in cases other than the universal tree there is the issue of how best to treat outgroup sequences in connection with duplicate gene rooting.

Our interpretation of the duplicate gene rooting procedure is shown in Fig. 2, with reference to three hypothetical species, A, B, and C. The aim is to choose among the set of possible rooted species trees, where each species has two copies of a particular gene, and where these have not diverged to an extent that sequence alignments are problematic. An analysis is performed including both copies of the gene from each species, which results in an unrooted gene network (Fig. 2a). The expectation (realized here) is that sequences of the two gene copies will form separate subtrees connected by a single branch. We can then ask with which of the possible rooted species trees the unrooted network is most compatible. In order to fit the gene network into the (A(B,C)) species tree (Fig. 2b), it can simply be folded along the central branch connecting the two forms of the gene. This requires only the assumption of a single duplication event prior to the diversification of the group. In contrast, fitting the gene network to the other rooted species trees is more difficult and entails additional hypotheses, such as other gene duplications, gene losses, lineage sorting and/or lateral transfer events. For example, it can be fit to the (B(C,A)) species tree by invoking the origin of polymorphism in each of the gene copies followed by congruent patterns of lineage sorting (Fig. 2c) or to the (C(B,A)) tree by imagining that the gene duplication occurred later, in species A, followed by lateral transfer.
to species B and then to species C (Fig. 2d). Many other scenarios are possible.

Viewed in this way, the optimality criterion should be clearer, namely to position the root of the species tree so as to minimize extra events in the gene trees (Maddison and Maddison, 1992, p. 53). The (A(B,C)) species tree is preferred in our example because it requires the fewest possible events/processes to be invoked. This “minimum events” formulation differs in subtle but important ways from another interpretation of the procedure, in which sequences of one of the gene copies are viewed as outgroups for the other, and vice versa (e.g., Doolittle and Brown, 1994, p. 6724, and their Fig. 5). Under this “reciprocal outgroups” interpretation one could root one of the gene trees with only one or a few copies of the other form, and vice versa, and this could be accomplished in separate analyses. In contrast, under the minimum events view it is important to simultaneously analyze both genes from each species, and the details of the congruence of the two gene trees are critical in determining the best rooted species tree. Under the reciprocal outgroups interpretation, if the two gene trees are found to be rooted in the same place, this provides evidence for a rooted species tree with the same topology, on the grounds that congruent trees from different datasets provide good evidence of relationships (Miyamoto and Fitch, 1995; but see Barrett et al., 1991). If the two gene trees happen to be rooted differently, a decision on the rooted species tree is not possible. In contrast, under minimum events, a score can potentially be calculated that measures how difficult it is to fit a two-gene network into each of the possible rooted species trees, and the species tree can be chosen that allows the minimum score. Finally, reciprocal outgroups assumes that the two forms of the gene are sister groups, that is, that they originated from a common ancestor through duplication. In contrast, minimum events does not require this assumption, because all alternative rootings of the gene network can be evaluated, in some of which one form of the gene will be paraphyletic with respect to the other (such as in Fig. 2d).

Although the minimum events interpretation of duplicate gene rooting may seem conceptually straightforward, it is difficult to fully implement in practice owing to the variety of different kinds of events that might be invoked in fitting a gene network to a set of species trees. An algorithm is needed to calculate a score that takes all of these types of events into account (along with their possibly different weights; Maddison, 1997). Progress along these lines has been made by Goodman et al. (1979), Page (1994), Guigo et al. (1996), Maddison (1997), and Page and Charleston (1997). Page and Charleston’s (1997) GeneTree program minimizes gene duplications/sorting events. Alternative approaches to disagreement among gene trees, ambiguous resolution, and missing information on genes from some species are being explored by M. Siddall (personal communication) and M. Frohlich (personal communication).

**RELEVANT GENE DUPLICATIONS**

Although duplicate gene rooting is promising in theory, its utility depends on identifying appropriate gene families for the problem at hand. Here we develop a general search image for relevant genes and evaluate several gene families in relation to the angiosperm problem.

Ideally, candidate genes would be present in two distinct forms in all members of the focal group and present in only one form outside of this group. This implies that a gene duplication occurred somewhere along the branch subtending the focal group and that the descendant gene lineages had achieved a degree of evolutionary independence and stability (i.e., gene conversion is infrequent; see Sanderson and Doyle, 1992). Functional divergence among lineages of a gene family may be a better predictor of evolutionary stability than absolute copy number. For example, despite high copy number in the MADS-box gene family, it includes at least some stable gene lineages correlated with gene function (Theissen et al., 1996). Conversely, the two or three Adh loci found in most angiosperms belies a history of repeated duplication and loss (Morton et al., 1996).

Sampling of relevant taxa both within and outside of the focal group must be complete enough for reasonable confidence in the phylogenetic position of the duplication. In some cases, it might not be clear whether a particular duplication arose along the branch immediately subtending the group of interest or along an earlier branch. This uncertainty does not invalidate the rooting procedure, though in such cases the distances between the two forms of the gene might be great and it would be best to include genes from taxa representing branches closer to the duplication event. Cases in which there are additional gene copies in the ingroup might also be acceptable if it could safely be assumed (based on previous phylogenetic analyses and the restricted distribution of some forms of the gene) that the extra duplications occurred well within the group and therefore would not complicate the rooting issue. Likewise, the absence of a gene copy in some lineages might be tolerated if one were confident that this resulted from a loss well within the focal group. Otherwise, gene absence could imply that a duplication occurred within the focal group, which would confound the analysis.

Finally, sequences of the two forms of the gene must not have diverged to the extent that they cannot be confidently aligned and included in the same analysis. Ideally, the two forms of the gene within the focal group will be less diverged from one another than either one is from the form present in outgroups. This should often be the case, but the degree of divergence will depend (1)
on whether the duplication occurred more recently or in the more distant past along the subtending branch of interest and (2) on whether rates of evolution have been more or less constant or variable among the relevant branches. Of special concern is the possibility that rates of evolution are significantly elevated after duplication events (perhaps owing to divergence in function and/or structure; e.g., Goodman et al., 1987), in which case average branch lengths might actually be increased by including both copies, even if the duplication were quite recent. Even in such cases, however, inclusion of both copies in an analysis can provide insights that might be missed by considering one copy at a time (see below).

Bearing these general considerations in mind, we consider several gene families for their potential usefulness with respect to the angiosperm rooting problem. Many, if not most, plant nuclear genes are members of small to rather large gene families. However, additional data are needed for most of these to determine whether gene duplication events might have occurred along the branch subtending angiosperms. Moreover, gene lineages in some of the better characterized plant gene families are nonindependent or relatively short-lived. For example, rbcS loci are homogenized by frequent gene conversion (Meagher et al., 1989; Dean et al., 1989), and, despite the relative conservation of copy number (Gottlieb, 1982; Morton et al., 1996), Adh loci apparently have been duplicated and lost rather frequently within angiosperms (e.g., in grasses (Morton et al., 1996) and in Paeonia (Sang et al., 1997)). In other gene families, at least some forms may have diverged prior to or very early in angiosperm evolution (e.g., actin, chalcone synthase, chlorophyll a/b binding protein), but these have since diversified and are now found in rather high copy number in many angiosperm taxa (McDowell et al., 1997; Demmin et al., 1989; Durbin et al., 1995; Helariutta et al., 1996). In such cases, mistaken orthology is a serious issue, but if gene lineages can be distinguished, recently duplicated forms may prove useful in rooting clades within angiosperms (as Adh duplications have helped root Paeonia; Sang et al., 1997).

Several other gene families show greater potential. MADS-box genes occur in at least seven major forms in angiosperms, and functional conservatism apparently is higher within these gene lineages than among them (Theissen et al., 1996). However, gene relationships are still quite uncertain (Purugganan et al., 1995; Munster et al., 1997), copy number is high in some gene lineages, and the genes are relatively small (~570 bp). Furthermore, relationships of the most promising duplicate pair, DEF and GLO (AP3 and PI of Purugganan et al., 1995), to genes outside angiosperms remain unknown. The small heat-shock protein genes (sHSPs, ~800 bp) comprise five distinct subfamilies (possibly more) within angiosperms (Waters, 1995a,b), suggesting that duplications preceded the origin of angiosperms. However, homologs of three of the angiosperm lineages occur in the moss Funaria, implying that two gene duplications occurred very early in the evolution of green plants, and the other two duplications are not likely to have occurred along the branch immediately subtending angiosperms (E. Waters, personal communication). Two distinct clades of legumin genes occur in angiosperms, one comprising methionine-rich (MetR) legumins, detected only in mono- or dicotyledonous taxa so far, and the other comprising methionine-poor (MetP) legumins, which also are known from eudicots. Single legumins are found in nonangiosperms (Fischer et al., 1996), suggesting the possibility of a duplication along the angiosperm branch. Although copy number in legumins and other seed plants is still uncertain, further attention to legumins clearly is warranted, especially because these are larger (~1700 bp) and may show more phylogenetically relevant variation than MADS-box or sHSP genes.

The phytochrome (PHY) gene family appears especially promising from the standpoint of the angiosperm problem. Genes in this family form four independent lineages in two pairs within angiosperms. PHYA and PHYC comprise one pair, and PHYB (including PHYD of Arabidopsis) and PHYE comprise another (Mathews et al., 1997). So far, only single full-length phytochrome genes have been sequenced from individual conifers. These, along with all known partial sequences from other seed plants, are unified in phylogenetic analyses with either PHYA/C or PHYB/E, suggesting that each pair occurs as a single copy outside the angiosperms (Mathews and Sharrock, 1997; S. Mathews, unpublished results). Thus, two duplications in the phytochrome gene family may have occurred along the branch leading to angiosperms. However, since PHYE apparently is lacking from certain potentially basal angiosperm taxa, the duplication leading to PHYB and PHYE may have occurred after the angiosperms started to diversify (Mathews and Sharrock, 1996; Mathews, 1997). In the PHYA and PHYC lineages, there is some evidence of duplication and loss; however, single PHYA loci have been detected in most angiosperms, and no taxa are known to have more than one PHYC gene (Mathews et al., 1995; Mathews and Sharrock 1996; Howe et al., 1998; Lavin et al., in press; S. Mathews, unpublished results). Phytochrome coding sequences are organized into one large exon (Exon I, with between 2035 and 2164 bp), wherein sequence conservation is quite high, and three smaller exons (totaling approximately 1440 bp); this organization appears to be highly conserved among all land plants (Quail, 1994). Complete PHYA and PHYC coding regions typically yield about 3200 alignable nucleotide sites.

**Preliminary Phytochrome Analyses**

To explore the potential of duplicate genes for rooting the angiosperm tree, we conducted a series of prelimi-
nary analyses of phytochrome sequences. Data matrices are available upon request from the second author or from TreeBASE (http://www.herbaria.harvard.edu/treebase). Figure 3 shows the result of a parsimony analysis of available complete phytochrome sequences (3264 bp), along with one partial sequence (1104 bp from *Pseudotsuga*), from land plants, including two mosses (*Physcomitrella, Ceratodon*), a lycopod (*Selaginella*), a fern (*Adiantum*), the whisk-fern (*Psilotum*), three conifers (*Picea, Pseudotsuga, Pinus*), and several angiosperms (a *PHYA* sequence from *Lathyrus* was not included, as two other legumes were in the dataset). Although *PHY* sequences are highly diverged, and there is considerable homoplasy, bootstrap support is high for many clades (15 clades at 100%). The four phytochrome genes that occur in most angiosperms are homologous with *PHYA, PHYB, PHYC,* or *PHYE* of *Arabidopsis*; a fifth form, *Arabidopsis* PHYD, likely resulted from a duplication within eudicots. No additional gene lineages are detected in phylogenetic analyses including all available partial sequences from angiosperms (Mathews and Sharrock, 1997).

Figure 3 indicates that an initial gene duplication, leading to *PHYA/C* and *PHYB/E* forms of the gene, occurred after the divergence of ferns from seed plants and before the divergence of conifers from angiosperms.

**FIG. 3.** Phytochrome phylogeny inferred from full-length sequences of land plants in GenBank (3264 nucleotides, 2190 informative sites). Heuristic parsimony analyses (100 random replicates using PAUP* 4d55; Swofford, 1997) yielded a single tree of 15,344 steps; CI = 0.350, RI = 0.51, and RC = 0.19 (excluding autapomorphies). Numbers above branches are inferred character changes (branch lengths) under ACCTRAN optimization; numbers below branches are bootstrap percentages from 100 replicates; circles on branches represent inferred duplication events (see text).
Conifers and angiosperms have both forms, and seedless plants do not. This is not to say that PHY diversification is entirely absent outside of seed plants (cf. Wada et al., 1997), but no independently evolving PHY lineages homologous with the major forms found in seed plants have been found in seedless plants (reviewed in Mathews and Sharrock, 1997). Analyses including partial phytochrome sequences (570 bp fragments) in GenBank from additional seed and seedless plants, including more ferns and conifers, gnetalian genera, and a cycad, corroborate this conclusion (not shown; cf. Kolukisaoglu et al., 1995).

Figure 3 also suggests that the duplication leading to PHYA and PHYC occurred after the separation of conifers and angiosperms but before the separation of monocots from eudicots. Among the partial sequences from additional seed plants, none, including one from Ephedra, are more closely related to either PHYA or PHYC (not shown; cf. Kolukisaoglu et al., 1995). In a preliminary survey of monosulcate angiosperms (S. Mathews, unpublished results), PHYA and PHYC were detected in every clade examined, indicating that the duplication probably occurred before the radiation of angiosperms. Within angiosperms, the better sampled PHYA sequences show concordance with many other lines of evidence in uniting the four grasses (Avena, Oryza, Sorghum, Zea), the two legumes (Glycine, Pisum), and the two Solanaceae (Nicotiana, Solanum), and in linking the umbelifer, Petroselinum, with the Solanaceae in an asterid clade.

Unfortunately, analysis of the available complete sequences is virtually uninformative regarding the angiosperm root because there are too few relevant taxa. To further evaluate the potential of the PHYA-PHYC gene pair to resolve basal relationships in angiosperms, we conducted an analysis of 12 species (representing a number of possibly basal branches within angiosperms) from which we obtained approximately 1 kb of sequence of both the PHYA and the PHYC genes. The single most parsimonious unrooted network resulting from this analysis is shown in Fig. 4. Although bootstrap support is weak for most clades, the match between the PHYA and the PHYC subtrees is impressive. The eight identical components, lettered A–H in Fig. 4, include Magnoliidae, Lauralidae, Magnoliidae plus Lauralidae, eudicots, Austrobaileyidae and Chloranthus with eudicots, and placement of the two monocots at the base. Indeed, the PHYA and PHYC subtrees differ only with respect to the placement of Candia and the monophyly versus paraphyly of Austrobaileyidae and Chloranthus. Under the reciprocal outgroups interpretation, the strict consensus of the two subtrees is rooted along the Sorghum branch, followed by Sagittaria, then by a trichotomy comprising Candia, the Magnoliidae-Lauralidae clade, and the eudicot-Austrobaileyidae-Chloranthus clade. Simultaneous analyses of PHYA and PHYC sequences, but also including a variety of distant outgroup sequences, produced nearly identical results (PHYA subtrees were identical, while PHYC subtrees differed only in uniting Sorghum with Sagittaria). Although the sample of taxa is still far too limited, these analyses do not support rooting the angiosperm tree near or within Magnoliidae, or near Austrobaileyidae, Chloranthus, or eudicots, as suggested by some previous analyses (see Fig. 1). The results are consistent with rooting within or near monocots and with a paleoherb rooting (sensu Doyle and Donoghue, 1993).

We compared these results of simultaneous analysis of the two genes with separate analyses of PHYA and of PHYC that included nonangiosperm PHY sequences. Outgroup rooting of the PHYA tree resulted in a single tree that is identical to the PHYA portion of the combined tree (Fig. 4). However, two trees resulted when PHYC trees were rooted with outgroups, one of

**FIG. 4.** Phylogeny of PHYA and PHYC sequences from 12 angiosperm taxa (1011 nucleotides, 541 informative sites). Heuristic parsimony analyses (100 random replicates with TBR swapping in PAUP+4d59; Swofford, 1997) yielded a single tree of 3412 steps; CI = 0.34, RI = 0.66 (excluding autapomorphies). Numbers above branches are inferred branch lengths under ACCTRAN; numbers below branches are bootstrap percentages from 100 replicates. Identical components are labeled A–H.
which is rooted along the Austrobaileya branch and differs topologically in several ways from the PHYC portion of the combined tree (not shown). In contrast to the simultaneous analyses described above, which root angiosperms in the vicinity of monocts, the consensus of the outgroup-rooted trees is unresolved with respect to the root. This illustrates that simultaneous analysis of both forms of a gene can yield greater resolution than when single forms are analyzed with outgroup sequences.

Finally, we used GeneTree (Page and Charleston, 1997) to identify species trees requiring the minimum number of duplication and sorting events for the PHYC network. GeneTree does not fully implement the minimum events optimality criterion discussed above, but it is the closest available approximation. Seven trees were found to require 4 duplications and 9 losses; the consensus of these trees is identical to the consensus of the two gene subtrees described above. For comparison we also tried fitting the gene network to two other rooted species trees, obtained by rooting the PHYC subtree along the Annona plus Magnolia branch (cf. Fig. 1a) and along the Austrobaileya branch (cf. Fig. 1m). The Annona plus Magnolia rooting requires 11 duplications and 42 sorting events, and the Austrobaileya rooting requires 11 duplications and 45 sorting events. The monocot-rooted tree entails far fewer costs than either of these alternatives.

CONCLUSIONS

Our principal aim has been to highlight the idea and the potential of using duplicate genes to root the angiosperm tree. Phytochrome genes are promising, but complete sequences are available for too few of the relevant taxa, and partial sequences are inadequate to confidently resolve relationships. We anticipate clearer results when complete sequences from more of the appropriate taxa are included. In particular, we expect that the placement of the very divergent Sorghum branch at the base of the tree will change as taxa are added. In general, we are encouraged that simultaneous analysis of duplicate genes (phytochromes and others) will contribute significantly to the resolution of the angiosperm problem.

The rooting problem exemplified by angiosperms is hardly unique. Indeed, a great distance between a focal group and its closest living relatives is characteristic of a number of the major radiations that have attracted the attention of evolutionary biologists (e.g., Phillippe et al., 1994). It may not be possible to overcome this difficulty by adding more taxa (as suggested by Hillis, 1996, and others), simply because relevant taxa are not, and presumably will never be, available. We believe that the use of duplicate genes for rooting purposes could have a significant (and perhaps more immediate) impact in resolving such problems.

ACKNOWLEDGMENTS

We are grateful to David Baum, Jim Doyle, Mike Frohlich, Toby Kellogg, Rick Ree, Mike Sanderson, Tao Sang, and Mark Siddall for helpful discussion of these issues, to Dave Swoford for allowing us to use a test version of PAUP*, and to Rick Ree for assistance with the figures.

REFERENCES


DUPLICATE GENES AND THE ROOT OF ANGIOSPERMS


