Evolution of the phytochrome gene family in land plants and its utility for phylogenetic analyses of flowering plants
by Sarah L Mathews

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Biological Sciences
Montana State University

Abstract:
The phytochrome nuclear gene family encodes photoreceptor proteins that mediate diverse developmental responses to red and far red light throughout the life of a plant. From studies of the angiosperm Arabidopsis thaliana, the family has been modeled as comprising five loci, PHYA-PHYE. In gymnosperms, two loci have been detected, while in other nonangiosperm groups just one locus has been observed. A polymerase chain reaction (PCR) based strategy to sample plant DNAs was developed in order to test the gene family model inferred from Arabidopsis in other plant species and' to evaluate the utility of phytochrome DNA sequence data for phylogenetic studies. Results presented here indicate that the Arabidopsis model is not completely appropriate for all angiosperm groups because additional PHY loci related to PHYA and PHYB have evolved independently several times in dicot angiosperms, and monocot angiosperms may lack orthologs of PHYD and PHYE. Nonetheless, for studies of organismal evolution, the phytochrome gene family is potentially useful because "the loci occur as single copy sequences, and the data suggest that the various loci are evolving independently. In two plant families, dicotyledonous Fabaceae (legumes) and monocotyledonous Poaceae (grasses), phytochrome data provided phylogenetic resolution. In addition to nucleotide substitutions, phylogenetically informative insertions and deletions characterize the phytochrome data sets. Furthermore, together with data obtained from public databases, the data detected in this study suggest that the differential distribution of phytochrome loci among flowering plant groups may be phylogenetically informative. The presence of a legume-specific locus most closely related to PHYA may be informative once its phylogenetic distribution is known; likewise, the apparent absence of PHYD and PHYE from monocots and some dicot plant groups potentially resolves relationships among major angiosperm lineages.

EVOLUTION OF THE PHYTOCHROME GENE FAMILY IN LAND PLANTS AND ITS UTILITY FOR PHYLOGENETIC ANALYSES OF FLOWERING PLANTS

by

Sarah L. Mathews

A thesis submitted in partial fulfillment
of the requirements for the degree

of

Doctor of Philosophy

in

Biological Sciences

Montana State University
Bozeman, Montana

August 1995

APPROVAL

of a thesis submitted by

Sarah L. Mathews

This thesis has been read by each member of the thesis committee and has been found to be satisfactory regarding content, English usage, format, citation, bibliographic style, and consistency, and is ready for submission to the College of Graduate Studies.

July 14, 1995
Date

Robert A Shaw
Chairperson, Graduate Committee

Approved for the Major Department

July 14, 1995
Date

E R Hyre
Head, Major Department

Approved for the College of Graduate Studies

8/26/95
Date

R L Brown
Graduate Dean

EVOLUTION OF THE PHYTOCHROME GENE FAMILY IN LAND PLANTS AND ITS UTILITY FOR PHYLOGENETIC ANALYSES OF FLOWERING PLANTS

Sarah Mathews

Advisor: Robert A. Sharrock, Ph.D.

Montana State University
1995

Abstract

The phytochrome nuclear gene family encodes photoreceptor proteins that mediate diverse developmental responses to red and far red light throughout the life of a plant. From studies of the angiosperm *Arabidopsis thaliana*, the family has been modeled as comprising five loci, *PHYA-PHYE*. In gymnosperms, two loci have been detected, while in other nonangiosperm groups just one locus has been observed. A polymerase chain reaction (PCR) based strategy to sample plant DNAs was developed in order to test the gene family model inferred from *Arabidopsis* in other plant species and to evaluate the utility of phytochrome DNA sequence data for phylogenetic studies. Results presented here indicate that the *Arabidopsis* model is not completely appropriate for all

angiosperm groups because additional *PHY* loci related to *PHYA* and *PHYB* have evolved independently several times in dicot angiosperms, and monocot angiosperms may lack orthologs of *PHYD* and *PHYE*. Nonetheless, for studies of organismal evolution, the phytochrome gene family is potentially useful because the loci occur as single copy sequences, and the data suggest that the various loci are evolving independently. In two plant families, dicotyledonous Fabaceae (legumes) and monocotyledonous Poaceae (grasses), phytochrome data provided phylogenetic resolution. In addition to nucleotide substitutions, phylogenetically informative insertions and deletions characterize the phytochrome data sets. Furthermore, together with data obtained from public databases, the data detected in this study suggest that the differential distribution of phytochrome loci among flowering plant groups may be phylogenetically informative. The presence of a legume-specific locus most closely related to *PHYA* may be informative once its phylogenetic distribution is known; likewise, the apparent absence of *PHYD* and *PHYE* from monocots and some dicot plant groups potentially resolves relationships among major angiosperm lineages.

STATEMENT OF PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a doctoral degree at Montana State University, I agree that the Library shall make it available to borrowers under rules of the Library. I further agree that copying of this thesis is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for extensive copying or reproduction of this thesis should be referred to University Microfilms International, 300 North Zeeb Road, Ann Arbor, Michigan 48106, to whom I have granted "the exclusive right to reproduce and distribute my dissertation in and from microform along with the non-exclusive right to reproduce and distribute my abstract in any format in whole or in part."

Signature Sarah L. Mathews

Date 18 August 1995

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The phytochrome nuclear gene family encodes photoreceptor proteins that mediate diverse developmental responses to red and far red light throughout the life of a plant. From studies of the angiosperm *Arabidopsis thaliana*, the family has been modeled as comprising five loci, *PHYA-PHYE*. In gymnosperms, two loci have been detected, while in other nonangiosperm groups just one locus has been observed. A polymerase chain reaction (PCR) based strategy to sample plant DNAs was developed in order to test the gene family model inferred from *Arabidopsis* in other plant species and to evaluate the utility of phytochrome DNA sequence data for phylogenetic studies. Results presented here indicate that the *Arabidopsis* model is not completely appropriate for all angiosperm groups because additional *PHY* loci related to *PHYA* and *PHYB* have evolved independently several times in dicot angiosperms, and monocot angiosperms may lack orthologs of *PHYD* and *PHYE*. Nonetheless, for studies of organismal evolution, the phytochrome gene family is potentially useful because the loci occur as single copy sequences, and the data suggest that the various loci are evolving independently. In two plant families, dicotyledonous Fabaceae (legumes) and monocotyledonous Poaceae (grasses), phytochrome data provided phylogenetic resolution. In addition to nucleotide substitutions, phylogenetically informative insertions and deletions characterize the phytochrome data sets. Furthermore, together with data obtained from public databases, the data detected in this study suggest that the differential distribution of phytochrome loci among flowering plant groups may be phylogenetically informative. The presence of a legume-specific locus most closely related to *PHYA* may be informative once its phylogenetic distribution is known; likewise, the apparent absence of *PHYD* and *PHYE* from monocots and some dicot plant groups potentially resolves relationships among major angiosperm lineages.

CHAPTER 1

INTRODUCTION

Phytochromes

The phytochromes are photoreceptors for red and far-red light in all green plants and some green algae (reviewed in Quail, 1991; Furuya, 1993). Each subunit of these large cytoplasmic receptors comprises a protein of 1100 to 1200 amino acids and a covalently attached linear tetrapyrrole chromophore. Existing in two continuously interconvertible forms, Pr, the red light-absorbing form, and Pfr, the far-red light-absorbing and biologically active form, phytochromes mediate diverse developmental responses throughout the plant's life cycle. While the mechanisms whereby phytochromes participate in cellular signalling remain unknown, regions of the polypeptide required for photosensory and regulatory activities and for dimerization have been identified (Cherry et al., 1993; Edgerton & Jones,

1992; Quail et al., 1995).

Several reports have described the presence of only a single *PHY* gene in certain nonangiosperms (Hanelt et al., 1992; Kolukisaoglu et al., 1993; Morand et al., 1993; Okamoto et al., 1993; Thümmler et al., 1992; Winands et al., 1992), while evidence of two *PHY* genes is reported for other nonangiosperms. For example, Maucher et al. (1992) refer to a putative second sequence in the fern *Dryopteris filix-mas* L., although the fragment remains uncharacterized. Two unpublished *PHY* sequence fragments from *Psilotum nudum* (L.)Griseb. (GenBank accessions X74930, X74931) differ from one another in the region of overlap; and two *PHY* cDNAs from *Pinus palustris* Mill. reportedly have been cloned and partially sequenced (Furuya, 1993), while a single *PHY* cDNA from *Gingko biloba* L. is cited in the same report. However, in angiosperms, five related sequences encoding phytochrome proteins designated *PHYA-PHYE* have been characterized from *Arabidopsis thaliana* (L.)Schur (Sharrock & Quail, 1989; Clack et al., 1994). The genes for these five phytochromes have been mapped to *Arabidopsis* chromosomes 1, 2, 4, and 5, (unpublished), and no evidence for *PHY* pseudogenes was found. Homologs of *Arabidopsis PHYA* and *PHYB* have been characterized in other angiosperms (Adam et al., 1993;

Christensen & Quail, 1989; Cordonnier-Pratt et al., 1994; Dehesh et al., 1991; Hershey et al., 1985; Heyer & Gatz, 1992a, 1992b; Kay et al., 1989; Sato, 1988; Sharrock et al., 1986), as have homologs of *PHYC* and *PHYE* (Cordonnier-Pratt et al., 1994). A putative pseudogene most similar to *PHYA* has been reported in *Pisum* (Sato, 1990), and a cDNA clone from *Zea* containing a partial *PHY* fragment has been interpreted as a pseudogene (Christensen & Quail, 1989). Overall, these studies suggest that the gene family increases in complexity from nonangiosperms to angiosperms. This suggestion is consistent with data recently submitted to GenBank in which the dicot *Piper* is represented by three distinct sequences, but additional nonangiosperm taxa are represented by just single sequences (Kolukisaoglu et al., unpublished).

Nearly all *PHY* genes that are fully characterized share high sequence identity (App. A) and structural similarity with the *Arabidopsis* loci (example in Fig. 1). Peptide fragments from the nonangiosperms *Anemia phyllitidis* (L.) Sw. and *Dryopteris filix-mas* (Maucher et al., 1992) share high sequence identity with the *Arabidopsis* phytochromes in their N-termini, as do sequence fragments from the N-termini of phytochromes from the major nonangiosperm taxa (Mathews et

al., 1995:App. 2); small internal PHY peptides from the alga *Mesotaenium caldariorium* (Lagerh.)Hansg. are highly similar to both N- and C-terminal peptides of other phytochromes (Morand et al., 1993). Two exceptional *PHY* genes have been described in nonangiosperms. The *PHY* gene from the alga *Mougeotia scalaris* Hässel (Winands et al., 1992) contains additional introns in the N-terminal coding sequence, and in the *PHY* gene from the moss *Ceratodon purpureus* (Hedw.)Brid., the conserved N-terminal region is combined with a highly divergent C-terminal coding region (Fig. 1) which encodes a putative light-regulated protein kinase (Thümmler et al., 1992). However, in another moss, *Physcomitrella patens* (Hedw.)B.S.G., the C-terminal coding region is similar to all other *PHY* genes (Kolukisaoglu et al., 1993). In angiosperms, the *PHYC* locus from *Arabidopsis* lacks the third intron found in all other fully characterized angiosperm loci (Cowl et al., 1994).

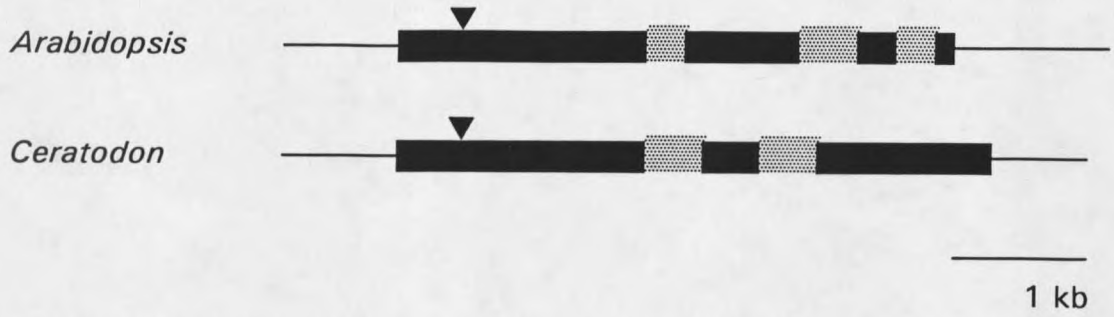

Figure 1. Phytochrome gene structure of *Arabidopsis* (Clack et al., 1994) and *Ceratodon* (Thümmler et al., 1992), from N-terminus (left) to C-terminus (right) showing untranslated regions (lines), exons (filled rectangles), introns (shaded rectangles), and the approximate site of chromophore attachment (triangle).

Phytochrome mediated responses that are characterized in green algae, mosses, and ferns include reorientation of chloroplasts (algae and ferns), rhizoid induction (algae, mosses, ferns), changes in cell membrane potential (algae and ferns) and water permeability (algae), cell elongation (ferns), reorientation of growth in protonemata in response to polarized light (mosses and ferns), germination of spores (ferns), organization of the cytoskeleton and cell cycle control (fern gametophytes), and phototropism in young leaves of some ferns (Wada & Kadota, 1989; Morand et al., 1993; Kraml, 1994; Haupt & Häder, 1994; Wada & Sugai, 1994). Neither the shade avoidance nor the deetiolation responses that are important in angiosperms have been detected in nonflowering plants (Smith, 1994).

In flowering plants, phytochrome mediated responses

include germination, seedling hypocotyl elongation, stem cell differentiation, plastid development, flavonoid pigment synthesis, and floral induction in response to photoperiod. Modulation of plant gene expression by phytochrome is well documented (Batschauer et al., 1994). In the dicot *Arabidopsis*, the *PHYA-E* genes are differentially expressed in response to the light environment (Sharrock & Quail, 1989; Somers et al., 1991; Clack et al., 1994), and unique, contrasting photosensory roles are attributed to *PHYA* and *PHYB* that cannot be accounted for by *PHYC*, *PHYD*, or *PHYE* (summarized in Quail, 1995).

Goals of this project

1. The primary model plant species used for investigations of phytochrome function in angiosperms is *Arabidopsis thaliana*. A critical consideration in evaluating the importance of the *PHY* photoreceptor family to plants in general is whether homologs of the various *PHY* genes are present in a wide variety of flowering plant species. Therefore the first goal of this project was to develop a method for detecting the various *PHY* genes in genomic DNA from diverse plant species, to clone and

sequence portions of those genes, and subsequently to estimate the relationships among the detected *PHY* genes using phylogenetic analyses.

2. The functional divergence observed among PHY loci in *Arabidopsis*, together with high sequence divergence (approximately 50% among the *PHYA*, *PHYB*, and *PHYC* loci) suggests that nonhomologous recombination is infrequent among *PHY* genes of *Arabidopsis*. If the loci are evolving independently, distinguishing orthologous genes from paralogous genes should not be difficult, thus predicting that these molecules might be useful tools for plant phylogeneticists (see MATERIALS AND METHODS). The second goal of this project was to use the phytochrome sequence data obtained to test the hypothesis that the genes are independently evolving and to ascertain the taxonomic level at which *PHY* data might be useful for phylogenetic studies of flowering plants.

CHAPTER 2

MATERIALS AND METHODS

DNA sequence detection

Total DNA was isolated from fresh, lyophilized, or dried herbarium material of taxa listed in Table 1 by standard methods (Doyle & Doyle, 1987). Aliquots were extracted once with phenol:chloroform-isoamyl alcohol (1:1 volume), and the aqueous portions were purified over sepharose CL-6B (Pharmacia, Piscataway, NJ) columns. DNAs were sampled from different subclasses of angiosperms (sensu Cronquist, 1981) and, two angiosperm plant families, Fabaceae (legumes) and Poaceae (grasses) were extensively sampled; to assess phytochrome gene and nucleotide diversity in Poaceae, DNAs were sampled from genera of the five major subfamilies Arundinoideae, Bambusoideae, Chloridoideae, Panicoideae, and Pooideae. DNAs of *Bambusa*, *Dianthus*, *Equisetum*, and *Quercus* were kindly provided by Elizabeth

Kellogg, Randy Woodson, Pamela Soltis, and Paul Manos respectively, and DNAs of *Flagellaria*, *Joinvillea*, and *Thamnochortus* were kindly provided by Jerrold Davis.

Table 1. Sources of *PHY* sequences determined in this study. Taxonomic arrangement follows Cronquist (1981) and Clayton & Renvoise (1986).

Subclass	Species	Source/Voucher
Sphenophyta	*Equisetum arvense* L.	P. Soltis (no voucher)
Pinophyta	*Gingko biloba* L.	*S. Mathews 365* MONT
	Pseudotsuga menziesii (Mirb.) Franco	*S. Mathews s.n.* MONT
Magnoliophyta		
Dicots		
MAGNOLIIDAE	*Ceratophyllum demersum* L.	*S. Mathews s.n.* MONT
	Aquilegia L. sp.	S. Mathews (no voucher)
HAMAMELIDAE	*Urtica dioca* L.	*S. Mathews 330* MONT
	Quercus turbinella Greene	*J. M. Tucker 4491* UCD
CARYOPHYLLIDAE	*Dianthus caryophyllus* L.	R. Woodson (no voucher)
	Spinacia oleracea L.	S. Mathews (no voucher)
DILLENIIDAE	*Arabidopsis thaliana* (L.) Schur	S. Mathews (no voucher)
ASTERIDAE	*Lycopersicon esculentum* Mill.	S. Mathews (no voucher)
	Antirrhinum majus L.	*S. Mathews 301* MONT
ROSIDAE	*Daucus carota* L.	S. Mathews (no voucher)
Monocots		
ALISMATIDAE	*Elodea* Michx. sp.	S. Mathews (no voucher)
ARECIDAE	*Lemna gibba* L.	Silverthorne (no voucher)
ZINGIBERIDAE	*Billbergia nutans* H. Wendl	*S. Mathews 351* MONT
LILIIDAE	*Muscari* Mill. sp.	S. Mathews (no voucher)
COMMELINIDAE		
Flagellariaceae		
	Flagellaria indica	*J. I Davis s.n.* BH
Joinvilleaceae		
	Joinvillea ascendens Gaudich.	*J. I Davis s.n.* BH
Restionaceae		
	Thamnochortus P. Bergins sp.	*J. I Davis s.n.* BH
Poaceae		
ARUNDINOIDEAE		
	Aristida purpurea Nutt.	*Lavin s.n.* MONT
	Danthonia unispicata (Thurber) Munro ex Macoun	*Lavin s.n.* MONT
	Phragmites australis (Cav.) Trin. ex Steudel	*Lavin s.n.* MONT
BAMBUSOIDEAE		
	Bambusa Schreb.sp.	*E. A. Kellogg V6 A*
	Oryza sativa L.	*Lavin s.n.* MONT
CHLORIDOIDEAE		
Bouteloua gracilis (Willd. ex H.B.K.) Lag. ex Griffiths		*Lavin s.n.* MONT
Calamovilfa longifolia (Hook.) Scribn.		*Lavin s.n.* MONT

Table 1, continued.

Eragrostis cilianensis (All.) Mosher	*Lavin s.n.* MONT
PANICOIDEAE	
Panicum capillare L.	*Lavin s.n.* MONT
Pennisetum setaceum (Forsk.)Chiov.	*Lavin s.n.* MONT
Zea mays L.	*Lavin s.n.* MONT
POOIDEAE	
Bromus inermis Leyss.	*Lavin s.n.* MONT
Poa pratensis L.	*Lavin s.n.* MONT
Stipa viridula Trin.	*Lavin s.n.* MONT

A region of the *PHY* gene that encodes a peptide including and proximal to the chromophore attachment site was amplified using PCR, resulting in a target of 270-350 bp (See Appendix A for region of amplification). Oligonucleotides with equimolar mixtures of nucleotide pairs at two-fold degenerate sites and inosines (I) at three- to four-fold degenerate sites were designed to amplify all possible target sequences in template DNAs flanked by the conserved upstream peptide HYPATDIP (5'-CAYTAYYCIGCIACIGAYA THCC-3') and downstream PFPLRYAC (5'-CRCAIGCRTAICKIARIGGRWA IGG-3'). These peptide sequences are conserved in all *Arabidopsis* phytochromes and in the amino acid sequences inferred from other fully sequenced dicot and monocot genes, and they flank a region comprising variation likely to be phylogenetically informative. Conversely, to restrict the number of phytochrome loci amplified in single experiments

locus-specific downstream oligonucleotides were paired with HYPATDIP in some experiments. These included *PHYA*-specific 5'-ACRTGIAYIGCRAAIACYTGIGC-3' at AQVFAI/VHV, *PHYB*-specific 5'-ASYTGIARICCRAAIGCYTGCAT-3' at MQAFGLQL, and *PHYC*-specific 5'-ATYTGIACICCRAAIACYTGIGT-3' at TQVFGVQI. Additionally, oligonucleotides at the upstream peptide GYDRVM (5'-GGNTAYG AYMGNGTNATG-3') and the downstream peptide KVLDMI (5'-YTTNAC NARRTCCATDAT-3') were designed to amplify a larger *PHY* fragment (ca. 600 bp) inclusive of the target sequences detected in these investigations. Used in combination with HYPATDIP and PFPLRYAC, GYDRVM and KMLDMI potentially provide the opportunity to detect loci that are mutated at HYPATDIP or PFPLRYAC; however, this strategy was not fully tested in this study. Standard PCR protocols (Perkin-Elmer, Norwalk, CT) were modified to include an initial 5 cycles in which annealing temperatures were less stringent (e.g., 45-49^{o} C).

The PCR products were converted to blunt-end fragments with T4 DNA polymerase (BRL, Gaithersburg, MD) and were ligated to *Eco*RV-cut bacteriophage M13KRV8.2. M13KRV8.2 carries an *Eco*K cassette that facilitates screening of nonrecombinants in an *E. coli* strain which is $r_k^+m_k^+$ (Waye et al., 1985). Transformation of *E. coli* with the ligation product yielded a population of M13*PHY* clones containing

amplified genomic *PHY* sequences. Individual clones were cultured, and double-stranded phage DNA was isolated from bacterial pellets by alkaline-lysis minipreparation. Inserts cut from M13 vectors using *Eco*RI and *Hind*III were resolved on 3% NuSieve (FMC, Rockland, ME), or 2% standard, agarose gels. Up to 108 clones were screened per individual DNA and in many cases, restriction enzyme digestion of PCR products or clones was used to aid in detection of less abundantly amplified or cloned sequences. Single-stranded DNAs for Sanger dideoxy sequencing (Sequenase version 2.0, USB, Cleveland, OH) were isolated from recombinants carrying putative *PHY* inserts. In most cases, sequences of both orientations were determined, and multiple PCR products from two accessions or genera were sequenced to detect possible contaminants and PCR errors.

Alignment of DNA Sequences

Peptide sequences were multiply aligned using ALIGN (Scientific & Education Software, State Line, PA) and GDE 2.2 (Steven Smith & University of Illinois) and were adjusted by eye at sites that were not accurately aligned by the computer algorithms; peptide alignments were the basis for multiple nucleotide sequence alignments. Appendices A

through F comprise the peptide and nucleotide alignments. For sequence comparisons, just gaps that could be identified as homologous were retained and were coded as single characters. All other gaps were deleted from the data matrices, as were nonhomologous 3′ and 5′ nucleotide sites. Four sequences that were included in the full length nucleotide data set (App. B) were not included in Appendix A because they were obtained later and did not significantly alter the consensus sequence. They are the *PHY* sequence from *Psilotum* (GenBank accession X74931), the *PHY* sequence from the moss *Physcomitrella* (Kolukisaoglu et al., 1993), and the *PHY* sequences from the angiosperm *Nicotiana* (GenBank accessions X66784, L10114). To assess phytochrome diversity in early land plants, DNA sequences from different nonangiosperm phyla available from GenBank were aligned with angiosperm sequences (App. C).

Phylogenetic analyses

Phylogenetic analyses of the phytochrome data were used to evaluate the relationships among newly obtained sequences and the genes characterized from *Arabidopsis*. The orthology of fully sequenced *PHY* genes from various species to

individual *PHY* loci of *Arabidopsis* commonly has been established by overall similarity (Dehesh et al., 1991; Heyer and Gatz, 1992a, 1992b; Quail, 1991; Furuya, 1993). Similarities in gene expression and regulation have been used secondarily to imply orthology (Furuya, 1993). However, overall similarity may not reflect phylogeny, and phylogenetically related loci may differ in function due to mutations in *cis*-regulatory regions (e.g., Doyle, 1991; Li & Noll, 1994). Since orthology is best determined by shared ancestry, as evidenced by homologies, cladistic analysis was used to determine the orthology of all available full length *PHY* sequences to those characterized from *Arabidopsis*. Likewise, evidence from cladistic analysis was used to assess orthology of the sequences detected in this study to the loci from *Arabidopsis*.

The assumption of phylogenetic analysis is explicit: given the evolutionary model that sequences diverged from a common ancestor by descent with modification, the goal is to discern the branching pattern among all sequences under consideration, successively grouping those together that most recently shared a common ancestor. Specifically, results presented here are phylogenies inferred from the DNA sequences by maximum parsimony analysis, with one exception (see below). Starting with a raw data matrix of sequences

aligned such that homologous nucleotide sites form columns of characters and genes (or taxa) are rows, maximum parsimony trees are constructed by optimizing the distribution of character state changes (nucleotide mutations) on a tree such that the fewest changes are required; the minimal length (ML) tree is chosen as the best explanation of the character data, and thus, the best estimate of phylogenetic relationships among the genes. Maximum parsimony algorithms ignore characters for which all taxa share the same character state because they do not provide specific evidence of relationship among subsets of the observed taxa; such characters are said to be uninformative. Conversely, if taxa A, B, and C in a matrix share the character state adenine at a nucleotide site, and D and E share a guanine at the same position, that character is informative in that it provides evidence that A, B, and C are most closely related to one another. Thus, maximum parsimony analysis discriminates between two types of similarity observed among taxa, overall similarity that reflects ancestral states and derived similarity that is due to more recent common ancestry, and to use only the latter in formulating historical hypotheses. Throughout this thesis, the terms "maximum parsimony" and "cladistic" will be used interchangeably. Indices that accompany the maximum

parsimony trees that are presented here provide an estimate of how many of the changes on the tree are due to the independent evolution of the same character state (homoplasy) and how many are inferred to be unique, and thus a sound basis for inferring homology; these include the CI (consistency index, Kluge & Farris, 1969; Farris, 1989), RI (retention index, Farris, 1989), and the RC (rescaled consistency index, Farris, 1989). CI is the minimum possible number of steps over all characters divided by the actual number inferred from the tree; CI = 1.0 if the actual number of steps equals the minimum number possible, if the characters in a data set are perfectly congruent with each other and the tree; the CI is thus an expression of the amount of homoplasy as a proportion of total change. Alternatively, the RI expresses the observed amount of homoplasy as a proportion of total possible homoplasy. The RC is the product of the CI and RI. When more than one ML tree is found for the data observed in this study, strict consensus trees that include only those components that occur in all ML trees (Sokal & Rohlf, 1981) are used to reveal consistently resolved groups.

In contrast to trees from maximum parsimony analysis, phylogenetic trees derived from matrices of pairwise distances or from maximum likelihood methods do not

discriminate between ancestral and derived similarity. However, all methods make assumptions about evolutionary change, either that substitution frequencies fit a very specific model (maximum likelihood and distance methods), or that characters are evolving independently (maximum parsimony). Criteria for comparing the accuracy of the methods (e.g., Hillis, 1995) include consistency (the correct tree is converged upon as the data become infinite), efficiency (a measure of how quickly a method converges upon the correct tree as more data are available for analysis), and robustness (the degree to which performance of the method is affected by violations of the assumed model). These criteria have been applied to results from computer simulations (e.g., Nei, 1991; Kuhner & Felsenstein, 1994; Huelsenbeck, 1995), providing information about the expected performance of the different methods under idealized conditions; the results are predictions about how the methods will perform during analysis of real data sets. A general conclusion from simulation studies and from tests of their predictions in investigations of known phylogenies, is that parsimony, maximum likelihood, and additive distance methods perform similarly, especially when the data are corrected for multiple substitutions per nucleotide site (e.g., Huelsenbeck, 1995). Parsimony methods are known to

be misleading when parallel changes among sequences exceed informative nonparallel changes (Felsenstein, 1978) because long branches attract one another. Thus, the more conservatively evolving amino acid characters were analyzed in comparisons of nonangiosperm with angiosperm phytochrome sequences and hypotheses inferred from distance analyses are included.

Maximum parsimony algorithms used for sequence comparisons were those available in PHYLIP 3.5c (Felsenstein, 1993), Hennig86 (Farris, 1988), PAUP 3.1 (Swofford, 1993), and RNA (Farris, 1994). Minimal length trees resulted from heuristic search options available in either Hennig86 (mh*, bb* with no upper limit set), PHYLIP (DNAPARS), or in PAUP (CLOSEST or RANDOM data addition sequence, HOLD option set for 5 trees when applicable, STEEPEST DESCENT, MULPARS, and TBR branch swapping options activated, with branch swapping on nonminimal trees, and MAXTREES set at 10,000 or 20,000). The PROTDIST and NEIGHBOR algorithms in PHYLIP 3.5c were used to reconstruct a phylogeny from pairwise distances among amino acid sequences from nonangiosperms and angiosperms; and the DNAML algorithm in the same program was used to infer a maximum likelihood phylogeny from the grass data. Graphical output of trees is from COMPONENT (Page, 1993a) and PAUP 3.1

(Swofford, 1993).

For the cladistic analysis of the full length sequences, trees were rooted by designating *PHY* sequences from *Physcomitrella, Selaginella* and *Adiantum capillus-veneris* L. (Okamoto et al., 1993) as the outgroups, because they are the only fully characterized *PHY* genes from nonangiosperms. For analysis of partial sequences in angiosperms, *Selaginella* was retained as an outgroup, along with the *PHY* sequences from the gymnosperms *Gingko* and *Pseudotsuga* that were determined during this analysis. For cladistic analysis of sequences from grass genera, trees were rooted by designating *PHY* sequences from *Flagellaria indica, Joinvillea ascendens* Gaudich. and *Thamnochortus* P. Bergins as outgroups; these taxa represent families inferred from morphological (Campbell and Kellogg, 1987; Linder and Rudall, 1993) and molecular data (Doyle et al., 1992) to be among the closest relatives of Poaceae. Cladograms rooted at *Muscari* Mill. allow detection of phylogenetic structure within outgroup species, but do not differ in the details of grass relationships from those rooted at *Joinvillea*. The *PHY* sequence from *Selaginella* was designated as the outgroup for the cladogram of all *PHY* sequences from grass genera because it is not likely to be more closely related to one

angiosperm *PHY* paralog than to another. Finally, analyses that addressed relationships among nonangiosperm and angiosperm *PHY* loci were rooted by designating *Mougeotia* as the outgroup.

Tests of phylogenetic accuracy

Phylogenetic hypotheses inferred from phytochrome data were evaluated for robustness using a subset of the statistical and congruence approaches that are available (e.g., Felsenstein, 1988; Hillis et al., 1993; Li & Zharkikh, 1995); support for monophyly of clades was evaluated using the bootstrap resampling, Bremer support, and total support tests described below. Congruence approaches were used to evaluate the agreement among individual *PHY* gene trees and among *PHY* gene trees and trees inferred from other data sets.

Statistical approaches

Bootstrap resampling. The use of the non-parametric bootstrap resampling technique to place confidence limits on phylogenies was proposed first by Felsenstein (1985); it is perhaps the method most frequently used by systematists to

assess the robustness of phylogenetic hypotheses. The test "involves inferring the variability in an unknown distribution from which your data were drawn by resampling from the data" (Felsenstein, 1985:784). A single bootstrap sample in a test of a phylogenetic hypothesis maintains the original set of species, but draws characters with replacement from the original matrix; consensus trees are used to show the monophyletic groups that occur in a majority of, for example, 500 or 1000 bootstrap replicates. The technique relies on several assumptions (Felsenstein, 1985) that are probably reasonable for DNA sequence data. A more serious limitation is that for proper hypothesis testing a null model should be specified in advance. However, in phylogenetic studies, the null hypothesis is the topology that has been inferred from a data sample (Felsenstein, 1985; Li & Zharkikh, 1994). The bias that this introduces to the outcome of bootstrap testing has been explored (Li & Zharkikh, 1994, 1995) and the CP (complete-and-partial) bootstrap has been developed to compensate for bias. The method is not yet available (Li & Zharkikh, 1995), thus the bootstrap values reported in this investigation are uncorrected.

Permutation methods. Maddison and Slatkin (1991) suggested that the appropriate null model for a statistical test of a "known" tree (i.e., the tree inferred from observed data) is one in which characters are randomized. The PTP (Faith & Cranston, 1991) and total support (Källersjö et al., 1992) tests compare the observed data to randomizations of those data; character states are randomly reassigned to taxa in the observed data such that congruence among character state distributions is produced by chance alone. The PTP measures character congruence (assumed to result from common ancestry) by comparing maximum parsimony trees from the randomized and observed data; minimal length trees inferred from randomized data are expected to be longer than trees inferred from observed data because randomized data sets should comprise fewer nonrandomly covarying characters. In preliminary analyses of phytochrome DNA data, all data sets were shown to be significantly structured based on results of PTP tests. However, Källersjö et al. (1992) demonstrated that the PTP can imply significant structure in ambiguous data. Thus, further permutation tests of the phytochrome data sets were performed using the total support test.

The total support test measures departure from random character covariation differently than the PTP; first the

Bremer support (Bremer, 1988) for each group in the observed tree is calculated (i.e., the number of steps that must be added to a tree before the group is lost in a strict consensus tree). Total support is the sum of group supports, which is assumed to be greater in well-structured data than in randomized data. The error rate on concluding that a data set is significantly structured is the upper tail probability $\alpha'_s = (X + 1)/(W + 1)$, where X is a number of W total randomizations yielding total support no less than that of the observed data (Källersjö et al., 1992). The phylogenetic program RNA (Farris, 1994) was used to calculate group support values. Additionally, Bremer support was investigated manually for some phylogenies by examining all trees up to ten or twenty steps longer than the minimal length tree(s).

Congruence approaches

Congruence approaches potentially address special concerns associated with inferring organismal relationships from molecular phylogenies. Various biological processes such as introgressive hybridization and lineage sorting from polymorphic ancestry may result in discordance among gene trees and/or among gene and species trees (e.g., Harrison et al., 1987; Rieseberg & Brunsfeld, 1992; Soltis et al.,

1992). Such differences also may result from lack of resolution in one of the data sets (e.g., Olmstead, 1989; Olmstead & Sweere, 1994), or from mistaken orthology (e.g., Goodman et al., 1979; Doyle, 1992). Thus, determining organismal relationships requires that evolutionary hypotheses derived from single genes be tested with further data (e.g., Pamilo and Nei, 1988; Takahata, 1989), as well as methods for reconciling differences.

Congruence approaches assume that similar patterns of relationships observed among trees derived from multiple independent data sets are evidence of both the veracity of the shared components and the accuracy of the phylogenetic method (summarized in Hillis, 1995). The debate between advocates of combining all data in a single analysis, the total evidence approach (Kluge, 1989), and summarizing topological features of trees derived from data partitioned into different types in a consensus tree (Adams, 1972; Carpenter, 1988) is ongoing (e.g., Barrett et al., 1991; de Quieroz, 1993; Chippindale & Wiens, 1994; Page, 1990, 1993b, 1994). The major argument against combining data for analysis is that subsets of characters may have been subject to different evolutionary processes (e.g., Bull et al., 1993). However, character weighting schemes can be used to incorporate assumptions about evolutionary models (e.g.,

Chippindale & Wiens, 1994). Advocates of combining data for analysis object to consensus techniques because information about individual results is lost. A distinct advantage of combining the data is that it allows character congruence (the degree to which all available characters make a unified, internally consistent statement about relationships, Swofford, 1991:314) rather than taxonomic congruence to be evaluated. Nonetheless, consensus trees are useful for expressing areas of agreement among trees and need not be viewed as phylogenetic hypotheses. Furthermore, discordance in results from separate analyses may be informative regarding nonindependence of characters (e. g., Swofford, 1991). Thus, it seems sensible to do both when possible (e.g., Doyle et al., 1994; Olmstead & Sweere, 1994).

Combining data. Phytochrome data from individual loci were not combined in broad comparisons (i.e., those comprising sequences from all angiosperm subclasses) because the main goal was to assess homology of individual sequences to *PHY* loci from *Arabidopsis*, and to assess the degree to which the phytochrome family comprises monophyletic gene lineages. However, in order to compare the degree and strength of resolution in phylogenies comprising phytochrome

sequences from grass genera, the data were analyzed both separately (with individual loci comprising individual terminals in a data set) and together (with data from all loci that were sampled combined for each genus).

Consensus analysis. As noted above, strict consensus trees were used to combine multiple minimal length trees from individual parsimony analyses. Furthermore, because the phytochrome data from grass genera were used to infer a species phylogeny, consensus techniques available in the computer package COMPONENT (Page, 1993a) were used to measure agreement among the phylogeny inferred from phytochrome data and grass phylogenies inferred from other data sets. For example, they were used to compare species phylogenies from phytochrome, *rbc*L, chloroplast DNA (cpDNA) restriction site variation, and morphological data from grasses.

Combining trees. In response to the suggestion that it is desirable to use as many genes as possible to infer an organismal phylogeny from molecular data (Pamilo & Nei, 1988; Takahata, 1989), or to combine molecular with morphological data when possible (e.g., Hillis, 1987), Baum (1992) proposed a protocol for combining the trees from

different analyses rather than combining the data. Following the method that Brooks (1981, 1990) proposed for recoding trees as single multistate characters in order to study coevolution, additive binary coding matrices are derived for single trees and subsequently combined for cladistic analysis. Doyle (1992) proposed a similar approach because he postulated that genes might behave as single characters rather than as a set of independent nucleotide characters; according to Doyle, such a set of nonindependent nucleotide characters could potentially "swamp" the signal from morphological data in a set of combined molecular and morphological data. Phylogenies inferred from phytochrome, *rbc*L, chloroplast DNA (cpDNA) restriction site variation, and morphological data from grasses were compared in this manner.

Tree mapping. The assumption of parsimony analyses that evolution is divergent is violated by convergence through such events as nonhomologous recombination among related loci. Of special concern relative to using sequences from a multigene family in phylogenetic reconstruction are potential problems related to concerted evolution (sensu Zimmer et al., 1980). For example, an analysis of *rbc*S nucleotide sequences (Meagher et al., 1989)

indicated that gene conversions among *rbc*S loci have occurred in each genus examined, leading to regions of "partial homology" (Patterson, 1987) in a locus and thus, to the possibility of mistaken orthology of genes. Gene conversion involving a complete locus is a gene loss because one gene is lost at the expense of another; loss of a gene through other nonhomologous recombination events, or through gene inactivation, also may result in inadvertent comparison of paralogous sequences. Comparison of paralogous rather than orthologous sequences potentially results in discordant gene and species phylogenies (Goodman et al., 1979; Doyle, 1992). Sanderson and Doyle (1992), however, suggest that it is possible to reconstruct a reliable organismal phylogeny from DNA sequences of multigene families in which concerted evolution is infrequent, and preliminary data indicate that nonhomologous recombination events are infrequent among phytochrome genes, (Sharrock & Quail, 1989; Dehesh et al., 1991; Heyer & Gatz, 1992a, 1992b; Clack et al., 1994; Adam et al., 1993). Nonetheless, tree mapping procedures available in COMPONENT (Page, 1993a) based on the hemoglobin phylogenetic model of Goodman et al. (1979), which evaluate whether incongruence of gene and species trees could be due to sampling error (Page, 1993b, 1994), were used to compare the phylogeny of phytochrome sequences from nonangiosperms