The phytochrome gene family in legumes (Fabaceae) : evidence for a new locus and analysis of evolutionary rates

by Elisa Jean Eshbaugh

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

Abstract:
The phytochrome nuclear gene family has been investigated elsewhere to assess its potential phylogenetic utility for plant systematics. Phylogenetic reconstruction of the phytochrome gene family from sequences sampled throughout angiosperms detected four possible loci in legumes including PHYA1, a potentially new and unique member of the multigene family. This study summarizes attempts to further characterize the phytochrome gene family in legumes by addressing the following questions: (1I) Does the phylogenetic pattern of the phytochrome sequences support designating the PHYA1 sequences to locus status? (2) What does the evidence from analysis of nucleotide substitution rates reveal about the PHYA1 sequences? The results from phylogenetic analysis and analysis of nucleotide substitution rates support the hypothesis that the PHYA1 sequences are potentially a fourth locus in legumes. In addition, the relative rate tests indicate that members of the gene family are characterized as having heterogeneous rates of nucleotide substitution. Furthermore, it is concluded that the Glycine max PHYA sequence is very likely to be unreliable for reconstructing organismal relationships given its accelerated rate of evolution. Finally, the concrete designation of PHYA1 as a functional gene awaits further evidence from the entire gene sequence.

THE PHYTOCHROME GENE FAMILY IN LEGUMES (FABACEAE): EVIDENCE FOR A NEW LOCUS AND ANALYSIS OF EVOLUTIONARY RATES

by

Elisa Jean Eshbaugh

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

of

Master of Science

in

Biological Sciences

MONTANA STATE UNIVERSITY-BOZEMAN
Bozeman, Montana

October 1998

Approval

of a thesis submitted by

Elisa Jean Eshbaugh

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

Dr. Matthew T. Lavin — Matthew T. Lavin (Signature) — 6 Nov 98 Date

Approved for the Department of Biology

Dr. Ernest R. Vyse — E R Vyse (Signature) — 11/5/98 Date

Approved for the College of Graduate Studies

Dr. Joseph J. Fedock — Joseph J. Fedock (Signature) — 11/9/98 Date

STATEMENT OF PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a master's degree at Montana State University-Bozeman, I agree that the Library shall make it available to borrowers under the rules of the Library.

If I have indicated my intention to copyright this thesis by including a copyright notice page, copying is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for permission for extended quotation from or reproduction of this thesis in whole or in parts may be granted only by the copyright holder.

Signature Risa J. Eshbaugh

Date November 5, 1998

ACKNOWLEDGMENTS

This work was supported by the Monts/EPScoR program and by funding provided by the National Science Foundation Grant to Matt Lavin (BSR-9118785). The author thanks Matt Lavin, Robert Sharrock and Rich Stout for their service as the graduate committee, Sarah Mathews and Ted Clack for technical advice in the laboratory. In addition the following individuals provided helpful comments and insights: Bonnie McCaig, Elizabeth Friar, Micheal Clegg, Sudir Kumar, Dave Russel and Steve Eshbaugh.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The phytochrome nuclear gene family has been investigated elsewhere to assess its potential phylogenetic utility for plant systematics. Phylogenetic reconstruction of the phytochrome gene family from sequences sampled throughout angiosperms detected four possible loci in legumes including *PHYA1*, a potentially new and unique member of the multigene family. This study summarizes attempts to further characterize the phytochrome gene family in legumes by addressing the following questions: (1) Does the phylogenetic pattern of the phytochrome sequences support designating the *PHYA1* sequences to locus status? (2) What does the evidence from analysis of nucleotide substitution rates reveal about the *PHYA1* sequences? The results from phylogenetic analysis and analysis of nucleotide substitution rates support the hypothesis that the *PHYA1* sequences are potentially a fourth locus in legumes. In addition, the relative rate tests indicate that members of the gene family are characterized as having heterogeneous rates of nucleotide substitution. Furthermore, it is concluded that the *Glycine max PHYA* sequence is very likely to be unreliable for reconstructing organismal relationships given its accelerated rate of evolution. Finally, the concrete designation of *PHYA1* as a functional gene awaits further evidence from the entire gene sequence.

INTRODUCTION

Understanding molecular evolution requires detailed knowledge about different evolutionary forces from a diversity of sources (Wolfe et al., 1987). Taken together, analysis of a variety of molecular data can advance areas of research such as population genetics, systematic biology and evolutionary ecology. Historically, much of the alliance of molecular evolutionary analysis and other fields of research has come from studies using data from single-copy nuclear genes or from mitochondrial and chloroplast DNA. Low-copy nuclear genes have been less thoroughly studied largely due to difficulties associated with the very qualities that make them unique, like the potential for acquiring and losing members of gene families (Durbin et al., 1995; Gaut, 1997). Furthermore, the relative paucity of studies with multigene family data is particularly true of data from plant sources (Demmin et al., 1989; Sanderson and Doyle, 1992; Sang et al., 1997).

In their study of the phytochrome gene family in grasses, Mathews and Sharrock (1996) addressed one aspect of molecular evolutionary analysis relevant to the application of gene family data to systematics, the analysis of rates of evolution. They found some evidence for rate heterogeneity in duplicated members (paralogues) of that gene family yet the same was not observed for subfamilies acquired by speciation (orthologues). Their study suggests that the phytochrome gene family is a potentially interesting group for further study.

Phytochromes are one type of photosensing pigment, or photoreceptor, that plants use in the detection of the radiant spectrum. They specifically function in the perception of

a photon of light in the red to far-red portion (ca. 600-800nm) of the spectral region and its subsequent conversion into a chemical signal (Quail, 1994; Pratt, 1995 ;Smith 1994). The phytochrome molecule is composed of two subunits that form a dimer between the carboxyl termini of the monomers (Sharrock et al., 1986). Each subunit has an open chain tetrapyrrole chromophore linked to a polypeptide of approximately 1100 to 1200 amino acids (120- 127kD) by a covalent bond. This chromophore attachment site occurs at a cysteine residue located in the N-terminal half of the polypeptide (Quail, 1994) (Fig. 1). The structural features of this molecule underlie its reversible photoconversion between red light and far-red light absorbing forms that culminate in producing a molecular switching mechanism. This regulatory switch alters the expression of several nuclear genes including those encoding the small unit of ribulose bisphosphate carboxylase (*rbcS*) and the chlorophyll *a/b* binding protein (*cab*) (Terzaghi and Cashmore, 1995). In this way plants can coordinate growth and developmental events with information acquired about relative amounts of red/far-red light (Sharrock and Quail, 1989; Smith 1994).

The multigene family that encodes phytochromes (*PHY*) are from the low-copy fraction of the nuclear genome (Mathews, 1995). Sequence analysis of the green alga, *Mesotaenium*, revealed the presence of an ancestral phytochrome sequence and suggests that a phytochrome gene lineage existed well before the emergence of land plants (Mathews and Sharrock, 1997). Furthermore, the possibility of still greater antiquity of the phytochrome gene lineage has not been discounted since sequences similar to phytochrome have been identified in studies of photoreceptors from cyanobacteria (Kehoe and Grossman, 1996). In their review of phytochrome gene family diversity, Mathews and

Sharrock (1997) state that *PHY* subfamilies may have arisen from a repeated pattern of divergence from a common ancestor following gene duplication events. In phylogenetic analysis of full-length sequences from land plants and green algae, they also find evidence to support the premise that these duplications most likely occurred around the time of major events in land plant evolution such as the emergence of seed plants and flowering plants.

The most extensively characterized phytochrome genes are known from *Arabidopsis* where low stringency southern blot analysis and DNA sequence analysis have revealed the presence of five genes, *PHYA*, *PHYB*, *PHYC*, *PHYD* and *PHYE* (Sharrock and Quail, 1989; Clack et al., 1994). Furthermore, there is evidence for functional divergence among these subfamilies (Smith, 1995). Studies of null mutants revealed that *PHYA* in *Arabidopsis* controls plant response to far-red high irradiance while the *PHYB* gene mediates a different group of plant responses to red light and the red/far-red ratio. Such diversification suggests unique functions might be assigned to other phytochrome encoding genes as well.

The suite of genes identified in *Arabidopsis* is considered representative of phytochrome diversity in flowering plants, although there may be some slight variation in the composition of subgroups among plant lineages. For example, there is some evidence for additional gene duplication events within the *PHYB/D* lineage as there have been several *PHYB/D* homologues identified in tomato, carrot, and *Arabidopsis*. Similarly, there is evidence to suggest there are multiple *PHYA* homologues in carnation, *Ceratophyllum* (an aquatic angiosperm) and legumes (Mathews and Sharrock, 1997). As

part of a study to access the phylogenetic utility of phytochrome sequence data, Mathews et al. (1995) extensively surveyed angiosperms for *PHY* sequences. Polymerase chain reaction using degenerate primers designed to amplify all possible *PHY* homologues failed to detect a homologue to *PHYC* and *PHYD* in legumes, however homologues of *Arabidopsis PHYA*, *B* and *E* were identified as well as unique *PHYA*-like, designated *PHYA1*.

The work presented here was initiated as a satellite project to the investigation of the phylogeny within the legume tribe Millettieae by Lavin et al. (1998). That study used the phytochrome gene family to elucidate problematic legume taxonomy. This paper presents an extension of the phylogenetic analysis of the phytochrome gene family in Fabaceae with emphasis on characterizing the *PHYA1* sequences. Additional evidence for the hypothesis that *PHYA1* is a discrete locus from phylogenetic analysis, percent divergence, conserved amino acid hallmarks and the analysis of evolutionary rates is discussed. Moreover, this paper diversifies the use of this less studied fraction of the nuclear genome and attempts to diversify the use of different types of data used to explore trends in molecular evolution.

MATERIALS AND METHODS

DNA Isolation

The sources for the plant tissues used in this study are given in Appendix 1. Total DNA was isolated from 0.1-2.0g of fresh or dried material generally following the CTAB method of Doyle & Doyle (1987). Following this, the aqueous samples were subjected to an additional extraction with phenol:chloroform (1:1 volume), purified over sepharose CL-6B (Pharmacia, Piscataway, NJ) columns and finally preserved in TE buffer for use as templates for amplification.

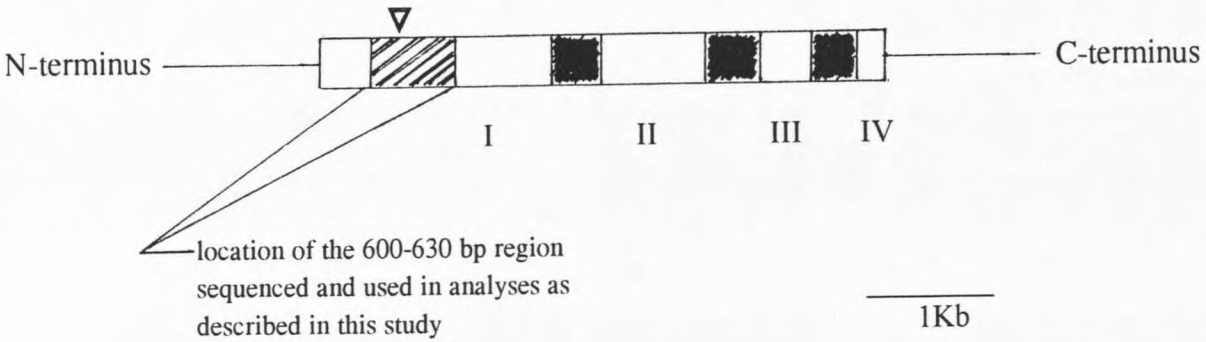
DNA Amplification

Sequences obtained from the coding region of exon I by Mathews et al. (1995) were used to identify conserved peptide sequences. These conserved sequences were then used to design degenerate PCR primers that would anneal fragments of all possible phytochrome loci (orthologs and paralogs) Mathews et al. (1995). The target region was a portion of exon I that encompasses the chromophore attachment site and a phylogenetically informative region downstream from that site (Kolukisaoglu et al., 1995) (Fig.1). Depending on the locus amplified, this target region was 600 to 630 bp and increased the size of the amplified fragment by approximately two-fold from that used for phylogenetic analysis by Mathews et al. The conserved peptide sequences from which the primers were designed are shown below in Figure 1.

Figure 1. Idealized PHYA gene structure illustrating the position of exons, introns, the approximate position of the chromophore attachment site and the region of exon I sequenced for this study. Exons (marked I-IV) are represented by open boxes, introns appear as shaded boxes. The cross-hatched area represents the approximate location of the 600-630 base pair fragment sequenced and as described in this study. "∇" indicates the approximate chromophore attachment site. The conserved peptide sequences from which the primers were designed are also shown in bold followed by identifiers that proceed the corresponding oligonucleotide sequence. Identifiers "Phy-S" and "Phy-N" specify whether the sequence is identical to the sense strand or nonsense strand of the phytochrome template DNA.

YDRMAYKFHED: Phy-S: 5′-TAYGAYAGGGTIATGGCITAYAARTTYCAYGARGA-3′

NIMDLVKCDG: Phy-N: 5′-CCRTCRCAYTTIACTAGRTCCATDATRTT-3′



(Modified from Mathews et al. 1995.)

Approximately 200 ng of purified total DNA from each taxon was amplified in a 100 µl reaction volume containing 2.5 units of *Taq* polymerase (Gibco BRL), 10 mM Tris-HCl pH 8.4 reaction buffer, 2.5 mM $MgCl_2$, 0.25 mM of each dNTP and1uM of each primer. Cycle parameters consisted of an initial denaturing at 94°C for 5 minutes, five cycles of 94°C for 1 min, 48°C for 1 min, 3 min ramp to 72°C for 1 min and 10-30 cycles of 94°C for 1 min, 55°C for 1 min, 72°C for 1 min. Upon completion of the last cycle the samples were incubated at 72°C for 10 minutes. Additional amplifications with less stringent annealing temperatures (e.g., 45-49° C) during the first five cycles were also conducted representing a modification of the protocols of Erlich (1992) and Innis et al. (1990). This approach, in concert with using degenerate primers, was designed to allow for sampling of all possible *PHY* loci (Wagner et al., 1994). Amplifications products were electrophoresed in 2% agarose gels to verify product size prior to cloning.

Cloning

Cloning was necessary to screen the pool of PCR products amplified in each individual PCR reaction. Over the course of the study, multiple PCR reactions were screened and cloned. Initially, the cloning procedure followed that of Mathews et al. (1995) wherein the amplifications products were treated with T4 ligase to generate blunt-end products for ligating into bacteriophage M13KRV8.2 (sensu Mathews et al.1995). Alternatively, some sequences were obtained by cloning with the TA vector (Invitrogen, San Diego, California) and subcloning into M13mp18 and mp19 vectors. Ligation products were then transformed into INVαF′ or JM 109 competent cells and

grown overnight on LB plates. Blue/white screening was used to identify putative phytochrome-containing recombinants. Eight colonies from each transformation were selected and cultured according to standard cloning protocols of Sambrook et al. (1989). Single-stranded DNA for sequencing was prepared by alkalinelysis minipreparation of phage DNA. Following minipreparation, putative *PHY* inserts were isolated from phage DNA by digestion with restriction enzymes and screened on 1% agarose gels.

Sequencing

Single-stranded sequencing reactions followed protocols suggested by the Sequenase version 2.0 Kit manufacturer (United States Biochemical, Cleveland, OH) using the -40 Primer. The sequences of cloned PCR products were determined by obtaining both orientations of a sequence as facilitated by directional cloning into M13mp18 and mp19 vectors, or from obtaining a consensus of several clones. Following sequencing, the reaction samples were electrophoresed in 6% acylamide gels, dried under vacuum at 80°C and exposed to autoradiograph film for 24 hours. Sequences have been deposited in GenBank and the accession numbers are given in Appendix 1.

Data Management and Phylogenetic Analysis

DNA sequences were read in manually and alignments obtained using the program ALIGN (Scientific Education Software, State Line, Pennsylvania) with a gap penalty of 10 and an extension penalty of 0.5. Minor adjustments to the alignment were made by eye to produce the best possible match upon comparison with previously sampled phytochrome

sequences. Additional sequences included in analyses that were taken from formerly published sources are referenced in Appendix 2. The program MEGA (Kumar et al., 1993) was used to generate an amino acid alignment from nucleotide data. Inspection of this alignment was the basis for initial classification of sequences into a group or putative loci based on the presence or absence of identifying amino acid hallmarks and insertions or deletions (indels) prior to analysis. The designation of *PHYA*-type sequences as *PHYA* or *PHYA1* was based on phylogenetic analysis with *Arabidopsis PHYA* (tree not shown).

Percent divergence for nucleotide sequences were calculated from the nonsynonymous proportion of the data with alignments generated in the program MEGA. CLUSTAL X, available by anonymous FTP at "ftp-igbmc.u-strasbg.fr/pub/", was used to calculate percent divergence for amino acid sequences. Phylogenetic relationships among sequences were inferred by both distance and parsimony methods. The nucleotides at the 5′ and 3′ primer sites were deleted from analyses. The Kimura two-parameter model of substitution was used to generate pairwise distances that were then subjected to cluster analysis following the neighbor-joining method (Saitou and Nei, 1986) as implemented by MEGA. Gaps and missing information were excluded. Strength of support for the major nodes within the trees were tested by bootstrap resampling of the original data matrix 1000 times.

Phylogenetic analysis of the data by maximum parsimony was obtained through PAUP 3.1.1 (Swofford, 1993). For this reconstruction, alignment gaps representing insertions or deletions were treated as single characters. Heuristic searches were conducted with all starting trees retained. For each search, the starting trees were

constructed with random stepwise addition. This search process was repeated 45 times so as to reduce the probability that any single search converged on a local optima as an artifact of the initial addition sequence. Within this reconstruction framework, tree bisection and reconnection (TBR) branch swapping algorithm was chosen to search for optimal trees and all of the most parsimonious trees were retained (MULPARS). *PHYB* sequences were used to root the reconstructions when constructing the consensus tree.

Rates of Evolution

Relative rate tests following the 1DN method of Tajima (1993) were applied to the data to test the null hypothesis of rate constancy among and within *PHYA* and *PHYA1*-type loci. This method is applicable to data for which an alignment is possible, even when phylogenetic reconstruction is inconclusive. Furthermore, the Tajima method is suitable when the substitution rate varies among different sites and does not require knowledge of the pattern of substitution rates. The neighbor-joining tree was used to guide the selection of pairs of sequences for each test. Some sequences appeared to exhibit accelerated rates based on branch length and were immediately included in the analysis. Other sequence pairs were chosen after dividing the *PHYA/PHYA1* clade into smaller groups and devising a scheme to sample from among those groups to achieve a representative cross section of the total comparisons possible. Initially all rate tests were done using *Myrospermum sousanum PHYB* as the outgroup. Significant results were tested again by calculating the relative evolutionary rate a second time using *Arabidopsis PHYA*. This measure was an effort to reduce the chance that significant results were a result of the long branch length

leading to the *PHYB* outgroup (Tourasse and Gouy, 1997). Within taxa and among loci tests were conducted on all species from which A-, A1- and E-type sequences were available. Primer sites were excluded from analysis of evolutionary rates.

RESULTS

Amino Acid Alignment

The alignment of 579 nucleotides from 43 phytochrome sequences (voucher specimens listed in Appendix 1) sampled from divergent species of legumes is shown in Appendix 2. Inspection of the amino acid alignment revealed several amino acid hallmarks as well as insertions and deletions useful for establishing the putative identity of sequences prior to further analysis (Appendix 3; Table 1). Most of the hallmarks are small features of a one to three amino acids, however the largest (not associated with an insertion or deletion) is the twelve residue hallmark that spans from position 447 through 458 of the alignment (Table 1). There are four amino acid hallmarks that are useful to distinguish *PHYA1* from *PHYA* sequences; "VK" at alignment position 307-308, "T" at position 329, "KKI" or "KKV" at position 354-357 and "NSC" or "NSS" at alignment position 383-385.

Percent Divergence

Within Groups

A matrix of percent nucleotide and amino acid divergence values were calculated from pairwise alignments to provide an initial estimate of the evolutionary relatedness of sequences prior to phylogenetic analysis (Appendix 4). A summary of the percent nucleotide and amino acid divergence is provided in Table 2. The four groups or putative

Table 1. Summary table of amino acid hallmarks and insertion/ deletion events (indels) of phytochrome sequences from legumes. Dashes indicate indels and hallmarks that are unique *PHYA1* to are underlined.

Length	Position	*PHYA*	*PHYA1*	*PHYB*	*PHYE*
3	303-305	VIA	VIA	VVA	VVS
5	307-311	ITKPG	VKKPG	SKRPD	IRRSD
1	316	L	L	I	L
7	329-335	ARFLRMK	TRFLFMK	SRFLFKQ	SRFLFKQ
1	342	V	V	C	V
4	347-350	KHVK	KHVK	SPVR	KPVK
7	354-360	DEKLPFD	DKKIPFD/DKKVPFD	DEALVQP	SEELRQP
4	363-365	TLCG	TLCG	CLVN	CLVN
1	370	A	A	A	S
1	373	S	S	G	V/G
1	376	L	L	A	T
3	383-385	DSI	NCS/NSS	GST	GST
3	392-394	VVV	VVV	VII	VIV/IIV
3	396-398	DNE/DND/DSD	DNE/DND/DSD	GND	GND
2	399-400	--	--	--	TT
16	399-415	EDGDSDA-VQPQKRK	EDGDSDA-VQPQKRK	EEGVCG------RSSM	----------
1	421	V	V	V	L
7	425-431	NTTPRFV	HTTPRFV	HTSARCI	HTSPRYV
2	435-436	LR	LR	LR	VR/LR
1	443	A	A	M	M
1	445	V	V	A	A
12	447-458	AIHVNKEIELEY	AIHVNKEIELEY	GKQLYMEIQLAS	GKQLYMEIQLAS
9	463-471	KNILRTQTL	KNILRTQTL	KRMLKTQTL	KRMLKTQTL
1	477	M	M	L	L
3	480-483	APL	APL	SPT	APF/APL

gene subfamilies were found to have fairly consistent percent divergence values in all pairwise comparisons made within groups at both the nucleotide and amino acid levels. For most of the within locus comparisons the percent nucleotide divergence values ranged from very low divergence of 0-1%, to values of 11-12% for some comparisons. The divergence values for the comparisons at amino acid level were higher overall with percentages ranging from 0-19%. The highest percent divergence values reported for these within group comparisons were from within *PHYA* and within *PHYA1*.

Table 2. Summary of percent divergence values calculated from pairwise alignments. Percent divergence of amino acid sequences appears above diagonal and percent divergence of nucleotide sequences is shown below the diagonal.

	PHYA	**PHYA1**	**PHYE**	**PHYB**	
	0-0.19	0.01-0.25	0.29-0.42	0.33-0.42	**PHYA**
		0.02-0.17	0.27-0.40	0.32-0.38	**PHYA1**
PHYA	0-0.11		0.02-0.13	0.10-0.22	**PHYE**
PHYA1	0.03-0.15	0.01-0.12		0.02-0.08	**PHYB**
PHYE	0.24-0.32	0.26-0.34	0.01-0.07		
PHYB	0.23-0.29	0.25-0.31	0.13-0.18	0.01-0.03	

Among Groups

As expected, the percent divergence values were higher for among group comparisons, with values of approximately 23-34% being fairly representative of nucleotide divergence, as compared to the values of 12% or less from within group divergence (Table 2). Notably, the percent divergence values from among E to B group comparisons were found to be much lower overall than the range considered representative; this is probably an artifact of sampling since only three *PHYB* sequences

were obtained for inclusion in analysis. Yet, more relevant to the focus of this study was the range of values calculated from pairwise comparisons among A to A1 groups for both nucleotide and amino acid sequences. Those values were found to be lower overall with percent divergence values ranging from 3-15% and 1-25%, respectively.

Taken together, the high values for percent divergence for within presumptive *PHYA* and *PHYA1* subfamilies combined with low percent divergence calculated for among *PHYA* to *PHYA1* imply the two *PHYA*-type sequences are consanguineous. High divergence within genes and low divergence among them is often correlated with sequences that are undergoing some type of homogenizing process (Ohta, 1983), although values calculated from percent divergence fail to indicate whether the relationship among the sequences analyzed here are consistent with expectations to support a hypothesis on a specific mechanism of information transfer.

Phylogenetic Reconstruction

To understand the relationships among presumptive loci, both distance and parsimony methods were used to infer the phylogeny of all sequences. Neighbor-joining methods and parsimony analysis resulted in trees with generally the same topology and consistent relationships within clades. The phylogeny inferred from the phytochrome sequences sampled in legumes support the position initially stated by Mathews et al. (1995) that four loci appear to represent phytochrome gene diversity in that plant family, three of which are comparable to loci known from *Arabidopsis* (*PHYA*, *PHYB*, *PHYE*), and one that is potentially unique in legumes, *PHYA1*.

Employing the neighbor-joining method using Kimura's two-parameter model of nucleotide substitution resulted in a tree with the topology shown in Figure 2. Bootstrap analysis of the distance-based reconstruction showed high support (100%) for the internal branches and distinguish the monophyly of the following groups: all *PHYB* sequences, the *PHYE* sequences, all *PHYA*-type sequences and finally the *PHYB* combined with the *PHYE* sequences. Although many branches within the more terminal clades were poorly supported, the same branching order and sequence relationships were consistently reconstructed under a variety of different conditions (e.g. reconstructed with the third position removed; trees not shown). As stated above, parsimony analysis resulted in a tree with similar topology to the distance tree. The strict consensus tree shown in Figure 3 was produced from 18 minimal length trees, inferred from 1254 steps, with a consistency index of 0.508 and a retention index of 0.805.

Resolution of a better supported phylogeny within the entire *PHYA* group is hindered by the inclusion of taxa from the poorly represented Mimosoideae and Caesalpinioideae subfamilies (sensu Lavin et al. 1998). Using the distance approach again, the phylogeny was inferred with the four taxa from those two legume subfamilies removed. When *Enterolobium cyclocarpum PHYA*, *Enterolobium cyclocarpum PHYA1*, *Brownea sp. PHYA* and *Gymnocladus dioica PHYA* are excluded from analysis the *PHYA1* sequences formed a monophyletic group derived from within the *PHYA* subfamily and bootstrap support was improved. Figure 4 shows *PHYA*-type sequences grouped by presumptive loci rather than taxa. For example *Millettia grandis PHYA1* clusters with other putative *PHYA1* subfamily sequences as opposed to clustering with *M. grandis*

Figure 2. Neighbor-joining tree.

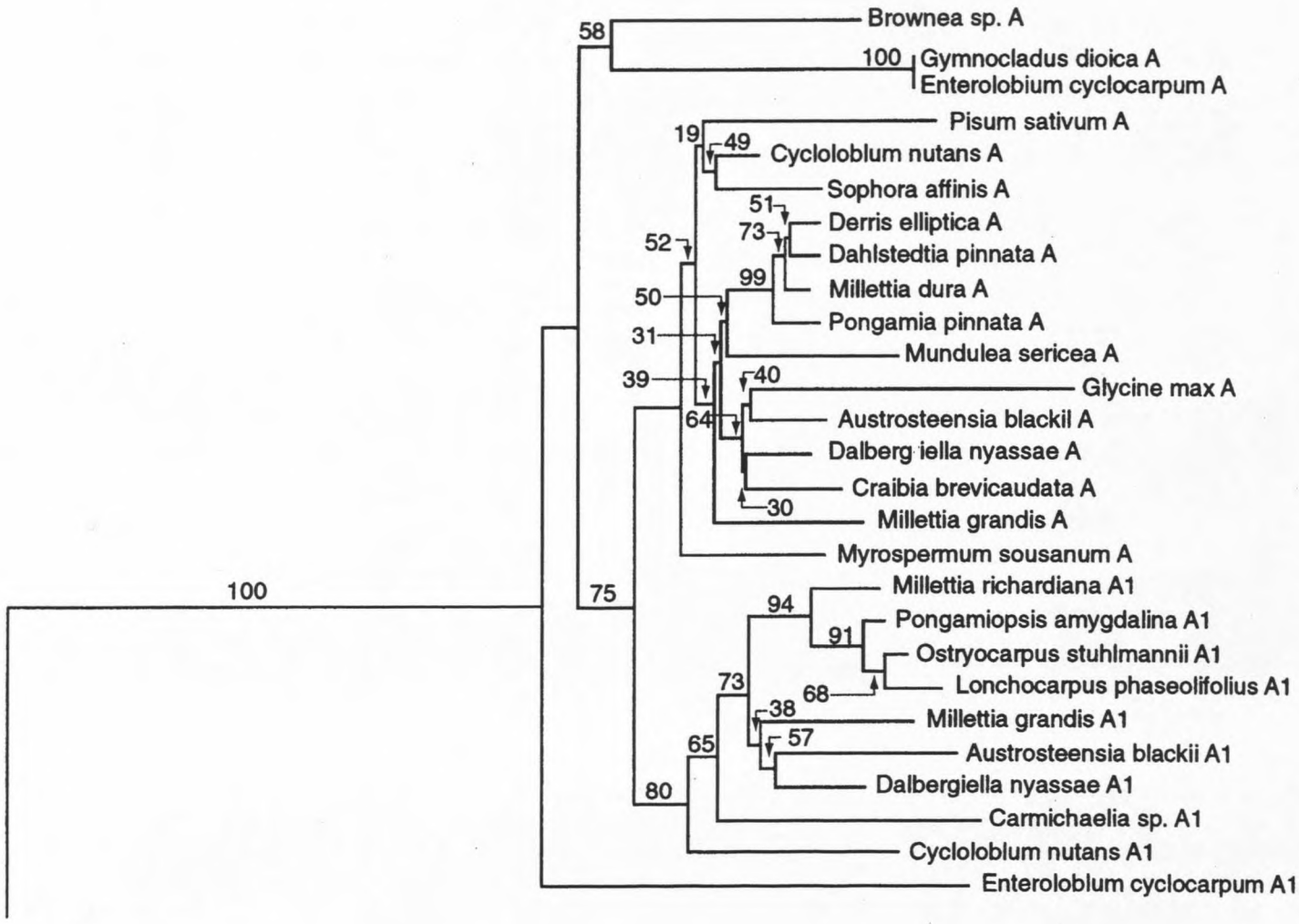


Figure 2. Continued.

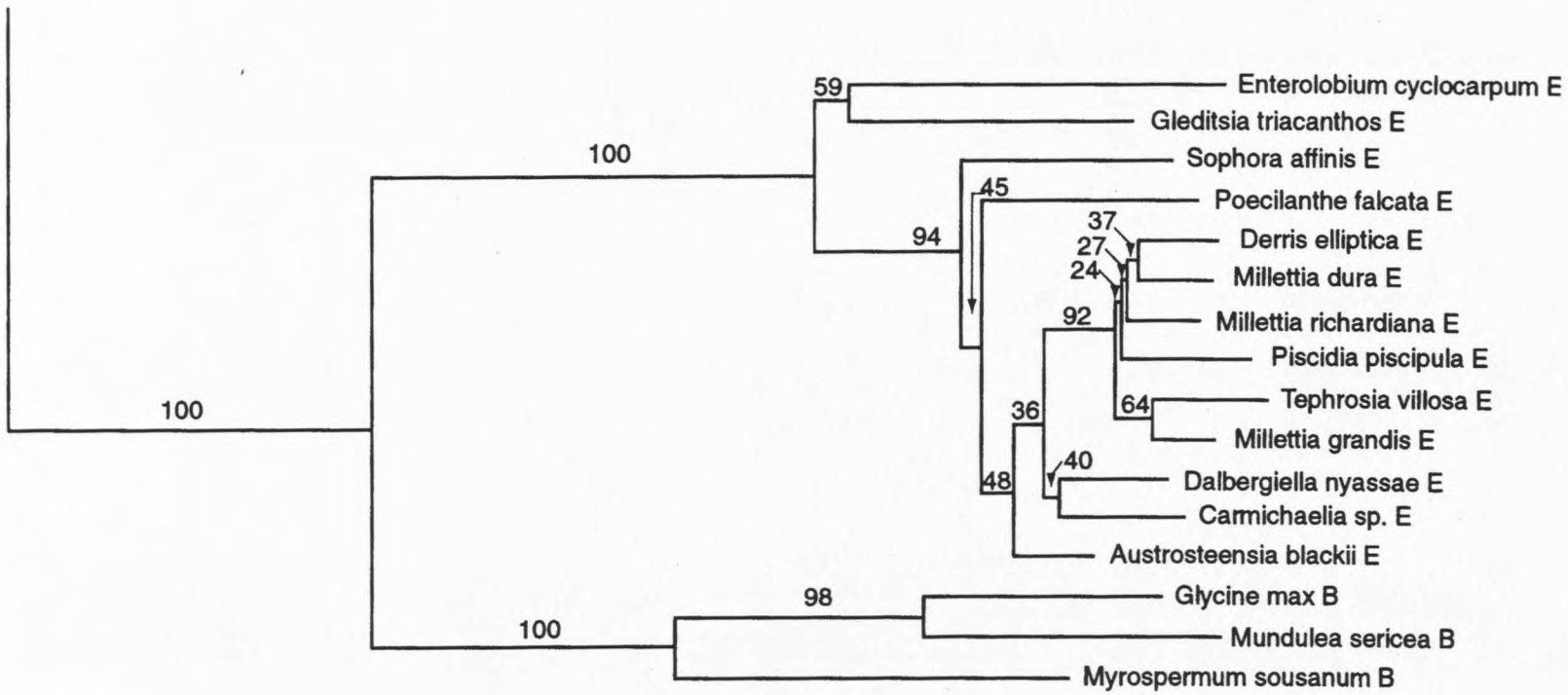


Figure 2. Phylogenetic relationship of *PHY* sequences sampled from legumes. Neighbor-joining tree constructed from genetic distances using the Kimura two-parameter model of nucleotide substitution. The number of times out of 1000 bootstrap replicates that a branch was present is noted above the branch. Branch lengths are proportional to evolutionary distances.