Spring 2008

The identification and characterization of phenylalanine ammonia-lyase gene family members in glycine max

Erin Kathleen Pringle

THE IDENTIFICATION AND CHARACTERIZATION OF PHENYLALANINE AMMONIA-LYASE GENE FAMILY MEMBERS IN *GLYCINE MAX*

by

ERIN KATHLEEN PRINGLE

A THESIS

Presented to the Faculty of the Graduate School of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE IN APPLIED AND ENVIRONMENTAL BIOLOGY

2008

Approved by

Ronald L. Frank, Advisor
Fikret Ercal
Anne Maglia
David J. Westenberg

ABSTRACT

Gene families are collections of genes with similar functions. Studying gene families is important for understanding the evolution of genes and manipulating genes. Phenylalanine ammonia-lyase (PAL) is an enzyme found in plants. It catalyzes the deamination of phenylalanine to produce cinnamic acid. Genes for PAL have been identified in many different plant species. This project used the known sequence for the PAL1 gene in *Glycine max* to find other PAL genes in *Glycine max*. The PAL1 gene sequence was used in a BLAST search to find similar sequences (ESTs). These similar sequences were assembled into contigs and compared both to each other and to PAL1. Potential gene family members were determined using this information. The new PAL gene family members, along with PAL1, were compared to PAL genes in other legumes and plants through phylogenetic analysis. A protein alignment of the sequences was used to create a DNA alignment. The DNA alignment of the gene sequences was used to generate phylogenetic trees and networks. Gene and species trees were reconciled to observe how the gene family may have evolved in legumes. Nonsynonymous and synonymous substitution rates were calculated. Finally, tissue expression was analyzed to better understand the conditions for expression of PAL genes.

Three new PAL genes were discovered. They were named PALB, PALC, and PALD. They lined up with PAL1 in exon II. Percent similarities and synonymous and nonsynonymous analysis supported the three genes as family members of the PAL gene family in *Glycine max*.

ACKNOWLEDGMENTS

I wish to thank my advisor, Ronald Frank, for introducing me to this project and offering advice, guidance, and support. I would also like to thank my graduate committee members, Fikret Ercal, Anne Maglia, and David Westenberg, for their suggestions and discussions about the material.

I would like to thank the Biological Sciences department at the Missouri University of Science and Technology for support. I would also like to thank the Missouri University of Science and Technology for providing me with the Chancellor's Fellowship.

Finally, I would like to thank my family, especially my parents and my fiancé, for providing support and encouragement.

TABLE OF CONTENTS

APPENDICES

LIST OF ILLUSTRATIONS

LIST OF TABLES

1. INTRODUCTION

1.1. *GLYCINE MAX*

Glycine max (L.) Merr. is also known as cultivated soybean. *G. max* is a diploidized tetraploid. The plant is an erect, bushy herbaceous annual that is not frost tolerant. It can reach a height of 1.5 meters. *G. max* belongs to the subgenus *Soja*. This subgenus also contains *G. soja* and *G. gracilis*. *G. soja* is a wild species of soybean. *G. soja* is thought to be the ancestor of *G. max*. *G. gracilis* is a weedy or semi-wild form of *G. max*. *G. gracilis* is thought to be a possible intermediate or hybrid between *G. soja* and *G. max* [1]. The classification for *G. max*, according to the PLANTS database at the United States Department of Agriculture [2], can be seen in Table 1.1.

Table 1.1. Classification of *Glycine max* (L.) Merr. [2]

Kingdom	*Plantae*	Plants
Subkingdon	*Tracheobionta*	Vascular plants
Superdivision	*Spermatophyta*	Seed plants
Division	*Magnoliophyta*	Flowering plants
Class	*Magnoliopsida*	Dicotyledons
Subclass	*Rosidae*	
Order	*Fabales*	
Family	*Fabaceae*	Pea family
Genus	*Glycine* Willd.	Soybean
Species	*Glycine max* (L.) Merr.	Soybean

G. max is one of the oldest cultivated crops. It is native to North and Central China. It is possible that it was first domesticated in the eastern half of China between the 17th and 11th century B.C [3]. *G. max* was introduced to the United States in 1765 [4] and Canada in 1893 [1].

Soybean is the most valuable legume crop. It has both nutritional and industrial uses. The soybean seen accounts for over 55% of all oilseed production and 80% of the edible consumption of fats and oils in the United States. Industrial applications for soybean include lubricants, emulsifiers, coatings, and biodiesel. Soybean is the principle source of biodiesel, which is also known as methyl soyate [5]. Statistics for soybeans can be found at the the National Agricultural Statistics Service. In 2007, 63,631,000 acres were planted for all purposes and 62,820,000 acres were harvested. There were 2,585,207,000 bushels produced. The price per unit was 10.40 dollars per bushel. The value of production was 26,752,197,000 dollars [6].

1.2. PHENYLALANINE AMMONIA-LYASE

Phenylalanine ammonia-lyase (PAL) is an enzyme involved in the phenylpropanoid pathway in plants. The phenylpropanoid pathway leads to the biosynthesis of many phenolic compounds. Important compounds that are eventually synthesized due to this pathway include flavonoids, phytoalexins, acetosyringone, lignin, and salicylic acid. PAL is the first enzyme in this pathway [7].

PAL catalyzes the deamination of phenylalanine to *trans*-cinnamic acid and releases ammonia [7]. PAL is responsible for shunting carbon out of primary metabolism into secondary metabolism [8]. Many different isozymes of PAL have been isolated [7]. Individual genes of PAL are differentially expressed during development [8]. PAL is regulated at the gene level by various environmental factors [7]. Some of these environmental factors include light, wounding of the plant, and microbial elicitors [8].

The first PAL gene in *G. max* has already been sequenced and described. The PAL1 gene in soybean has a coding region of 2142 basepairs. The coding region is divided between two exons: exon I and exon II. Exon I has 392 basepairs, and exon II has 1750 basepairs. There is a single intron between the two exons. This intron is made up of 1519 basepairs, and it splits the 131st codon. The PAL1 gene encodes a polypeptide that is made up of 713 amino acids. PAL1 has some similarity to PAL2 in *Phaseolus vulgaris*. For exon I, there is a 74% sequence homology at the nucleotide level, and the homology is distributed unevenly. For exon II, there is a 84% sequence homology at the nucleotide level, and the homology is distributed more evenly over the entire length of the exon. However, there are a few short fragments of limited sequence similarity. For the intron, no significant stretches of homology can be found [9].

A search at the National Center for Biotechnology Information website (discussed in Section 1.4.1) reveals that PAL has been discovered and sequenced in many different plant species. Under the *Magnoliophyta* division (flowering plants), PAL has been researched in many different species. A search in the nucleotide database for PAL gives 447 results. In *Arabidopsis thaliana*, four different PAL genes can be found in the database. Under the *Fabaceae* (pea) family, PAL has been researched in 15 different genera. These genera include *Lotus*, *Trifolium*, *Astragalus*, *Pisum*, *Glycine*, *Phaseolus*, *Stylosanthes*, *Medicago*, *Vigna*, *Sphenostylis*, *Cicer*, *Styphnolobium*, *Caragana*, *Acacia*, and *Cassia* [10].

1.3. GENE FAMILIES

Gene duplications are one major way from which new genes can evolve. Most nucleotide changes in genes that affect the fitness of the organism are deleterious. This means that genes are selectively constrained, which can be seen when looking at coding regions and non-coding regions of genes. Coding sequences tend to diverge slower than non-coding regions. Coding sequences have less mutations at places where a base change would cause a change in the amino acid. Whenever a gene is duplicated, the gene has more freedom to evolve as long as the duplicate genes continue to carry out the original function [11].

Once a gene is duplicated, the duplicate can either become eliminated or fixed in the population and preserved over time. If the duplicate gene becomes fixed and preserved, nonfunctionalization, neofunctionalization, or subfunctionalization can occur. For nonfunctionalization, the duplicate can not function due to mutations and may degrade over time. For neofunctionalization, the duplicate gains a new function. For subfunctionalization, the duplicate works with the original gene to carry out the original function. The original function becomes divided between the duplicate genes [11].

Gene duplications have helped contribute to the existence of gene families [11]. Gene families are groups of genes that share similar nucleotide sequences and produce products with similar structures or functions. Sometimes members of a gene family are grouped together because their products work together as a unit or in the same process [12]. Gene family members that share a common ancestor due to a duplication event are paralogous. Gene family members that share a common ancestor due to a speciation event are called orthologous genes. Orthologous genes are found in different genomes [11]. Gene families help with understanding how genes are related to each other. The function of a new gene can be predicted based on its similarity to known genes. Gene families can help with understanding and predicting gene expression. They can also help with identifying genes involved in diseases [12].

1.4. DATABASES AND TOOLS

1.4.1. National Center for Biotechnology Information. The National Center for Biotechnology Information (NCBI) was established in 1988. It is a division of the National Library of Medicine at the National Institutes of Health. NCBI is a national resource for molecular biology information. The overall goal of NCBI is to better understand molecular processes affecting human health and disease. NCBI creates public databases, conducts research in computational biology, develops tools for analyzing genome data, and distributes biomedical information [13].

NCBI has many different databases and software tools. GenBank is a DNA sequence database. Other databases found at NCBI are: Online Mendelian Inheritance in Man (OMIM), the Molecular Modeling Database (MMDB) of 3D protein structures, the Unique Human Gene Sequence Collection (UniGene), a Gene Map of the Human Genome as well as maps of other sequenced genomes, the Taxonomy Browser, and the Cancer Genome Anatomy project (CGAP). Entrez is a search and retrieval system for integrated access to data found at NCBI. PubMed is a web search interface that gives access to journal citations in MEDLINE. Basic Local Alignment Search Tool, or BLAST, is a program for sequence similarity searching. Other software tools found at NCBI are: Open Reading Frame Finder (ORF Finder), Electronic PCR, and Sequin and BankIt (sequence submission tools) [13].

1.4.2. Expressed Sequence Tags. Expressed sequence tags, or ESTs, are short DNA sequences that represent genes expressed in certain cells, tissues, or organs from different organisms that have been sequenced. They are usually 200 to 500 nucleotides long. ESTs can be generated by sequencing one or both ends of an expressed gene. ESTs are a quick, effective, and inexpensive way to discover new genes. These "tags" of DNA can be used to find a gene from chromosomal DNA by matching up base pairs. There can be various challenges when using ESTs to find genes. These challenges depend on genome size and the presence or absence of introns, so they vary among organisms. GenBank has a searchable database of ESTs called dbEST. This database is a collection point for ESTs. ESTs get submitted, screened, and annotated before placement in the database [14].

Since the ESTs in the database are described in detail and come from specified cells, tissues, or organs, this makes it possible to analysis of expression. The frequency of ESTs in a library should be a function of the frequency of cDNA copies of that particular gene. An abundance of mRNA for a particular gene should result in more ESTs from that gene ending up in the database. The same can also be said for tissue type, genotype, or treatment [14].

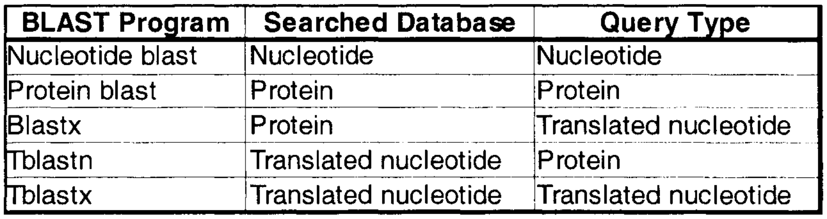
1.4.3. Contigs. There have been various definitions for contiguous sequences, or contigs, in the past. The term was originally defined by R. Staden in the 1980 paper "A new computer method for the storage and manipulation of DNA gel reading data" [15]. The given definition was as follows:

> In order to make it easier to talk about our data gained by the shotgun method of sequencing we have invented the word "contig". A contig is a set of gel readings that are related to one another by overlap of their sequences. All gel readings belong to one and only one contig, and each contig contains at least one gel reading. The gel readings in a contig can be summed to form a contiguous consensus sequence and the length of this sequence is the length of the contig. [15]

Contigs can also be defined as continuous runs of nucleotides that are longer than what any single sequencing reaction can produce. Data from multiple sequencing reactions can be compared for significant overlap and assembled into contigs. ESTs can be used to assemble contigs [16].

1.4.4. BLAST. BLAST is a tool at NCBI that calculates sequence similarity. BLAST is designed to help with finding similarity between sequences, which allows for inferring the function of new genes, predicting new members in gene families, and exploring evolutionary relationships. BLAST can be used in different ways. Different query sequences can be used with different databases. At the BLAST website, basic BLAST programs are nucleotide blast, protein blast, blastx, tblastn, and tblastx. The description of these programs can be seen in Table 1.2. Specialized BLAST programs are also available. An example of specialized BLAST is aligning two sequences with BLAST, or bl2seq [17, 18].

Table 1.2. Basic BLAST Programs



BLAST Program	Searched Database	Query Type
Nucleotide blast	Nucleotide	Nucleotide
Protein blast	Protein	Protein
Blastx	Protein	Translated nucleotide
Tblastn	Translated nucleotide	Protein
Tblastx	Translated nucleotide	Translated nucleotide

BLAST uses statistical theory to calculate a bit score and expect value (E-value). These are calculated for each alignment, and can help determine whether the similarity is due to a biological relationship or chance alone. The bit score can indicate the quality of the alignment. A higher bit score indicates a better alignment. The E-value indicates the statistical significance of a pairwise alignment. A lower E-value indicates a more significant hit. The E-value tells the chance of the similarity between the sequences occurring by chance alone [17, 18].

1.5. SEQUENCE ALIGNMENTS

An alignment can be created between two or more sequences. The sequences can be nucleotide sequences or amino acids sequences. Alignments can be used to draw conclusions about the evolutionary histories of sequences. They can be used to understand the evolutionary path for how the sequences diverged from a common ancestor. Comparing sequences can lead to a better understanding of the function of genetic sequences and the information they contain. Alignments can be an indication of how closely sequences are related to each other. Sequences that are closely related are usually easier to align. Alignments can be used to help determine the functions of new sequences and evolutionary relationships for genes, proteins, and species. Alignments can also help predict structures and functions of proteins [16].

Simple alignments can be performed between two sequences. A simple alignment is the pairwise match for all the characters of the sequences. The overall similarity between the sequences is a fractional value. An alignment score can be used to numerically represent sequence similarity. A scoring function can affect the results of a sequence alignment, so various techniques have been created to find alignments likely through evolution. Once the scoring function is selected, an algorithm can be used to find the best alignment or alignments. The Needleman and Wunsch algorithm was developed for global sequence alignments. Global sequence alignments compare two sequences over their entire lengths. The Smith-Waterman algorithm was developed for local sequence alignments. Local sequence alignments are used to find the subsequences that match the best within the two sequences. The BLAST search at the NCBI website looks through a sequence database to find the best ungapped local alignments [16].

When aligning three or more sequences, a multiple sequence alignment is usually preferable to a set of pairwise alignments. A multiple sequence alignment simultaneously aligns many sequences. One problem with methods for aligning multiple sequence is the computational complexity increases greatly with an increased number of sequences. The CLUSTAL algorithm is a multiple sequence alignment method developed to find near-optimal alignments for a larger number of sequences while allowing faster comparisons [16].

ClustalX is a commonly used multiple alignment program. CodonAlign is another alignment program that generates a DNA alignment from a corresponding protein alignment. It creates triplet gaps in the DNA alignment at the same positions the gaps in the protein alignment are found [19].

1.6. PHYLOGENETIC TREES

Taxonomy is a field of science that is used to classify life into groups. Systematics is a field of science that deals with the diversity of life and the relationships between life's components. Systematics goes beyond taxonomy to clarify new methods and theories. These can then be used to classify species based on similar traits and mechanisms of evolution [20].

Phylogenetic systematics is used to identify and understand evolutionary relationships among both living and dead organisms. It uses evolutionary theory about similarity. This theory says that similarity is due to common descent or inheritance from a common ancestor. Similarity can be studied among individuals or species. Phylogenetic systematics can establish relationships that describe a species' evolutionary history, which leads into a phylogeny. A phylogeny can describe historical relationships among lineages, organisms, or parts of organisms such as genes [20].

Phylogenetic trees are used to visually show the evolutionary relationships between a group of organisms. These trees are usually made up of nodes, branches, and a root. Nodes represent taxonomic units (taxa). These taxa can be specified by the user to be species, populations, individuals, genes, or bacterial strains. Branches are used to show the relationships between taxa based on descent and ancestry. Branches can be scaled or unscaled. Scaled branches have branches lengths that represent numbers of changes that occur along them. Unscaled branches have branch lengths that do not represent actual numbers of changes. Branches can also be used to represent time in addition to changes. A root is the common ancestor of all the taxa in the tree. However, a tree can be unrooted which means a common ancestor is not identified and an evolutionary path is not clear. An unrooted tree is used to only show the relationships between taxa [20].

Bootstrapping is a method that creates trees based on subsamples of sites in an alignment. This process is repeated multiple times. Anywhere from 100 to 2000 replicates can be done. While 1000 is a typical number of replicates, 2000 replicates are required for 95% reproducibility. The results of the process are compiled to estimate the reliability of a specific grouping. Bootstrapping a tree is used to understand the reliability of groupings within a phylogenetic tree [19].

A gene tree is a phylogenetic tree based on divergence seen within a single homologous gene. This tree represents the evolutionary history of the gene. It does not have to represent the evolutionary history of the species in which the gene is found. A species tree is a phylogenetic tree based on divergence seen in multiple genes. It is usually better to create a species tree based on analyses that use data from multiple genes. Using more data is necessary because evolution occurs at the population level of organisms and not the individual level [16].

Different methods can be used to generate phylogenetic trees. For constructing a tree, the main approaches are algorithmic and tree-searching. The algorithmic approach uses an algorithm to create a tree using the given data. The tree-searching method creates many trees, and then chooses the best tree or set of trees. Two advantages of the algorithmic approach are the faster speed and the generation of only one tree from a dataset. Neighbor Joining (NJ) and Unweighted Pair-Group Method with Arithmetic Mean (UPGMA) are two algorithmic methods. Tree-searching methods are usually slower and can generate equally good trees. There are also distance and character-based methods. NJ and UPGMA are both distance methods. Distance methods change a sequence alignment into a distance matrix. The distance matrix has pairwise differences, or distances, between the sequences. The matrix data is then used to compute branching order and branching distances. Character methods use a sequence alignment directly. These methods compare the characters at each site in the alignment. Each site has a column of characters from each sequence in the alignment. Parsimony, Maximum Likelihood, and Bayesian analysis are all character-based methods. Parsimony finds a tree or trees with the least amount of changes. This method can create trees that are equally parsimonious but have slight differences. Maximum Likelihood (ML) finds a tree that maximizes the likelihood of observing the data. It uses a model of evolution to do this. ML produces a tree where the likelihood is known. However, the ML method is significantly slower than the NJ and Parsimony methods. Bayesian analysis is a variant of the ML method. It finds a set of trees with the greatest likelihoods given the data. No bootstrapping is necessary for Bayesian analysis because the frequency of a grouping in the set of trees is nearly the same as the probability of that grouping. NJ, Parsimony, ML, and Bayesian are all accepted methods without one being clearly better or more widely used than the others. If the data and alignment are good, then the trees generated by these different methods will still be very similar. The differences represent real uncertainty [19].

PAUP*, PHYLIP (Phylogeny Inference Package), Tree-Puzzle, and MrBayes are all programs that can be used to generate phylogenetic trees. PAUP* and PHYLIP can create trees using several different methods. Tree-Puzzle can create ML trees. MrBayes can create trees using the Bayesian analysis method. TreeView is a program that can be used to draw, view, and modify phylogenetic trees. It does not actually create trees, so it uses tree files created by other programs [19].

1.7. RECONCILIATION

The process of resolving disagreement between a gene family tree and a species tree is called reconciliation. Gene duplications and losses are used to explain the differences between the trees. The resulting duplication and loss histories can be used to identify orthologs, estimate gene duplication times, and root and correct gene trees [22]. Reconciliation is done by fitting a gene tree to a species tree. A mapping between each node in the gene tree and a corresponding node in the species tree is created. The inconsistencies from the mapping are used to infer gene duplications and losses [21,22].

Notung is a program that can reconcile gene and species trees. It can identify duplications and estimate bounds on the time of duplication. Notung can also root trees. It can root unrooted trees and rearrange rooted trees with weakly supported edges. It does the rooting by minimizing gene duplications and losses. The program also has unique features compared to other reconciliation programs. Notung calculates a Duplication/Loss Score for a reconciled gene tree. The score can also be called the D/L score or D/L cost. The D/L Score is the weighted sum of losses, duplications, and conditional duplications in a reconciled gene tree. The user can specify the costs, but the default values are 1.5 for duplications, 1.0 for losses, and no cost for conditional duplications [21,22].

1.8. PHYLOGENETIC NETWORKS

Phylogenetic trees are commonly used for looking at evolutionary history. Evolutionary models that use trees can be limited in describing more complex evolutionary events. Phylogenetic networks can be used to analyze, visualize, and explore data without forcing it into a tree or tree-like model. A phylogenetic network is a network in which nodes represent taxa and edges represent evolutionary relationships of the taxa. Phylogenetic networks can then be divided into different types, with phylogenetic trees being one type of network. A split network comes from combining phylogenetic trees and then representing compatibilities seen within and between the data sets. A reticulate network shows evolutionary history when reticulation events are present. Reticulate events can include hybridization, horizontal gene transfer, and recombination. Other types of networks can also be used for specific situations. Many researchers use their own specific definitions of phylogenetic networks in studies, which can cause the definition of phylogenetic networks to be narrowed down to a certain type of network [23].

Phylogenetic networks are good to use when studying evolutionary history that may involve reticulate events such as hybridization, horizontal gene transfer, recombination, or gene duplication and loss. However, phylogenetic networks can still be useful even when these events are not present. Reticulate networks are used to explicitly represent evolutionary history, while split networks are used to implicitly represent evolutionary history. Reticulate networks have internal nodes that represent ancestral species. Nodes that have two or more parents indicate reticulation events. Split networks are able to show incompatible and ambiguous signals found in data sets. Parallel edges represent splits that are computed from the data. Nodes in split networks do not have to represent ancestral species [23].

SplitsTree4 and Spectronet are two programs that can generate phylogenetic networks. SplitsTree4 can generate various types of phylogenetic networks and trees. It can create networks or trees using methods such as split decomposition, neighbor-net, consensus network, or super networks. It also has methods to create hybridization or simple recombination networks [23]. Spectronet can generate median networks [24]. Median networks are a type of split network. They use sequence data to generate networks [23].

1.9. SYNONYMOUS/NONSYNONYMOUS SUBSTITUTIONS

The central dogma of molecular biology says that information stored in DNA is used to make RNA, and the RNA is used to make proteins. RNA is made during transcription, and proteins are made during translation. Amino acids are strung together to create proteins. The amino acid sequence determines the function of a protein. While four different nucleotides are used to make RNA and DNA, 20 different amino acids are used to make proteins. The four nucleotides can be arranged in 64 different combinations when used three at a time. A group of three nucleotides (called a codon) in RNA correspond to a specific amino acid. A codon causes the insertion of a specific amino acid into a growing amino acid sequence. Three codons that do not cause the insertion of a specific amino acid are stop codons. Out of the 20 different amino acids, 18 of them are coded for by more than one codon [16].

Substitutions, or changes, in a position of a codon can still result in the coding of the same amino acid. Synonymous substitutions are changes at the nucleotide level of coding sequences that do not cause a change in the amino acid sequence of the produced protein. Changes that occur at the nucleotide level of coding sequences and do cause a change in the amino acid sequence are called nonsynonymous substitutions. Synonymous substitutions should be observed more often than nonsynonymous substitutions since natural selection should distinguish between functioning proteins and proteins that do not function well. The nucleotides in triplet codons can be divided into three different categories. These categories are nondegenerate, twofold degenerate, and fourfold degenerate sites. Nondegenerate sites are positions in the codon in which changes always cause amino acid substitutions. Twofold degenerate sites are positions in the codon where two of the four nucleotides result in the same amino acid, but the other two nucleotides result in a different amino acid. Fourfold degenerate sites are positions in the codon where a change to any of the other nucleotides will still result in the same amino acid. Nucleotide changes accumulate fastest at fourfold degenerate sites and slowest at nondegenerate sites [16].

Synonymous Non-synonymous Analysis Program (SNAP) can be used to calculate synonymous and nonsynonymous substitution rates. It calculates rates based on nucleotide sequences that are aligned by codons. SNAP can calculate many different variables related to synonymous and nonsynonymous substitution rates. These variables can be seen in Table 1.3. The calculations are based on pairwise comparisons of the sequences [25,26].

Table 1.3. Variables Calculated by SNAP [2,3]

Variable	Description
Sd	Number of observed synonymous substitutions
Sn	Number of observed nonsynonymous substitutions
S	Number of potential synonymous substitutions (average)
N	Number of potential nonsynonymous substitutions (average)
ps	Proportion of observed synonymous substitutions (Sd/S)
pn	Proportion of observed nonsynonymous substitutions (Sn/N)
ds	Jukes-Cantor correction for multiple hits of ps
dn	Jukes-Cantor correction for multiple hits of pn
ds/dn	Ratio of synonymous to nonsynonymous substitutions

When comparing genes that are possibly in the same gene family, it can be helpful to look at the first, second, and third position changes in the codons. When assembling sequence fragments into contigs, the consensus sequences from these contigs could represent real genes or artifacts from genes. Real genes should be constructed through evolution. Gene family members should have more synonymous than nonsynonymous changes when comparing their sequences. The third position in a codon is more likely to allow synonymous substitutions. When comparing genes from the same gene family, the most differences in nucleotides should be found in the third position of the codons. To determine if two gene sequences are from the same gene family, the number of first, second, and third position differences can be recorded. If the differences for the position are about the same, then the gene sequences are probably not in the same gene family. If there are more differences in the third positions and few differences in the first and second positions, then it is likely the gene sequences are from the same gene family. This method is an alternative to using a program such as SNAP to do synonymous and nonsynonymous analyses.

2. MATERIALS AND METHODS

2.1. RETRIEVAL OF SIMILAR SEQUENCES

A Basic Local Alignment Search Tool (BLAST) search was performed at the National Center for Biotechnology Information (NCBI) website. The translated nucleotide database was searched using a protein query (tblastn search). The PAL1 protein in *Glycine max* (accession: CAA37129, GI: 18377) was used as the protein query to find similar nucleotide sequences. The non-human, non-mouse ESTs (est_others) database was selected for the search. The search was limited with an Entrez query of "glycine max[orgn]" so that only *Gylcine max* sequences would be returned by the search. The number of descriptions and Alignments was set to 250 each.

Only sequence fragments with an E-value less than 0.001 were chosen. They were transferred into a new spreadsheet. The accession numbers for all of the chosen sequence fragments were saved. These accession numbers were used for a batch Entrez nucleotide retrieval at the NCBI website. After the retrieval of the sequences, the sequences were saved to a single file in FASTA format.

2.2. ASSEMBLY AND COMPARISON OF CONTIGS

Sequencher [27] was used to assemble the retrieved fragment sequences into contiguous sequences (contigs). The FASTA file with the sequences was opened in Sequencher. The PAL1 protein coding DNA sequence was also added to the list of sequences in Sequencher. The assembly parameters were set to the following: Minimum Match Percentage was changed to 99 percent and Minimum Overlap was left as 20. The sequences were assembled into contigs automatically by Sequencher. For each contig, the accession numbers for all of its sequences members were recorded.

The open reading frame (ORF) of each contig was checked for quality in Sequencher. The ORF quality was recorded for each contig. The contigs were sorted into three groups based on ORF quality: good ORF, fair ORF, and poor ORF. ORF quality was based on how much the ORF was broken up by stop codons. One or less stop indicated a good ORF. A few stops, such as two or three, indicated a fair ORF, and many stops indicated a poor ORF.

The PAL1 protein coding DNA sequence was added into Sequencher with the assembled contigs. The PAL1 DNA sequence and all of the contigs were selected so they could be compared. The Assemble Interactively function was used to find out how similar the contigs were to the original PAL1 sequence. The Minimum Match Percentage was first set to 98 under the parameters. Any contig that showed up as a match for PAL1 was recorded along with its actual similarity percentage. The Minimum Match Percentage was then lowered to 97, and any new matches were recorded with a percent similarity. The Minimum Match Percentage was lowered in increments of one, down to a limit of 80. Each time the percentage was lowered, any new similar sequences were recorded. This comparison method was then used for each contig. Each contig was checked for similar sequences. For each contig, similar sequences and their similarity percentage were recorded down to a percentage similarity of 80. The comparison method was also repeated for each unassigned fragment sequence by selecting the PAL1 DNA sequence, all contigs, and all unassigned fragments for comparisons. For each fragment, any similar contigs or sequences were recorded along with percent similarities.

Each contig was assigned to a possible gene family member group based on percent similarity. PAL1 was also used for one gene family member group. Contigs that were at least 98% similar were grouped together. Contigs with poor ORFs were not assigned to any group. Unassigned sequences were assigned to groups later.

Contigs were assembled into consensus sequences using AssemblyLIGN. The first, second, and third positions of the contigs in the codons were compared using MacVector. The differences in the codon positions were recorded for pairwise comparisons of the contigs.

When looking at the ORFs for the contigs, all six possible reading frames were displayed in Sequencher. This allowed the best reading frame to be chosen for each contig. The contig consensus sequences were adjusted to match the best reading frame. If the first reading frame was used, no changes were made. If the second reading frame was used, the first nucleotide base was removed. If the third reading frame was used, the first two nucleotide bases were removed. If any of the other three reading frames were better, the reverse complement of the sequence was determined with MacVector and bases were removed if necessary.

2.3. MAPPING

The contigs were all mapped against the PAL1 protein coding DNA sequence. Blast 2 Sequences (Bl2Seq) at the NCBI website under BLAST tools was used to align two sequences at a time. Each contig was aligned with PAL1. The length of the contig was recorded. The starting and ending positions for the contig and PAL alignment were recorded for each contig. Alignment arrangements (plus or minus) were also recorded. All the contigs were then displayed together in a map to show how they aligned with PAL1. The contigs were mapped against PAL1 in Microsoft Excel. The cells were changed to squares in order to create a grid that was then used for mapping. Contigs were grouped together by the potential gene family membership.

The mapping method was repeated for the unassigned sequences. All the unassigned sequences were displayed together in a map to show how they aligned with PAL1.

2.4. FINALIZATION OF PAL GENE FAMILY MEMBERS

Unassigned sequences were compared to any contigs they overlapped by using a percent similarity. The unassigned sequences were then assigned to the same gene family member group if they matched any contigs found in that group. Another map was made for the how the PAL groups, including contigs and newly assigned sequences, mapped to PAL1. The resulting contigs in the same group were compared to each other again. Contigs were combined if possible, based on map overlap and similarity. Contigs that could not be compared to others based on the mapping were left out of further analyses. A lack of significant overlap between groups caused some groups to be dropped from further analysis.

Contigs and sequences assigned to a gene family member group were greater than 95% similar to at least one of the other contigs or sequences in the group. The gene family groups were at least 80% similar to at least one other gene family group.

A nucleotide consensus sequence was created for each finalized gene family member in MacVector using representative contigs. The consensus sequence for each new PAL gene family member was used to represent the gene in further analyses. The sequences were also translated into protein sequences using MacVector.

2.5. SEQUENCE ALIGNMENTS

PAL genes in other species were picked out to use for comparison. The focus was placed on legumes. The sequences can be found in the NCBI protein and nucleotide databases. The legume species that were chosen in addition to *Glycine max* are: *Pisum sativum*, *Medicago sativa*, *Cicer arietinum*, *Vigna unguiculata*, and *Phaseolus vulgaris*. *Petroselinum crispum* and *Arabidopsis thaliana* PAL sequences were also chosen as outgroup sequences for the phylogenetic analyses. The PAL sequences in *P. crispum* and *A. thaliana* were also chosen because those species have multiple PAL genes identified.

The protein sequences for all 19 PAL genes were aligned using ClustalX [37]. A complete alignment was performed by ClustalX with default settings. The protein alignment and a FASTA file of DNA sequences were used to create a DNA alignment with CodonAlign. The output files from CodonAlign had some minor errors in the files structures that had to be altered by hand. The errors were too many spaces between sequence names and their actual sequences.

2.6. PHYLOGENETIC TREE ANALYSIS

Three different phylogenetic trees were generated, each by a different method. PAUP* was used to generate a Neighbor Joining tree and a Maximum Likelihood tree. The code used to generate the NJ and ML trees came from Phylogenetic Trees Made Easy by Barry Hall [19]. The NJ tree code can be seen in figure 2.1, and the ML tree code can be seen in Figure 2.2. The sequence alignment for the 19 nucleotide sequences is not present in the figures to save space, but they were present for tree generation.

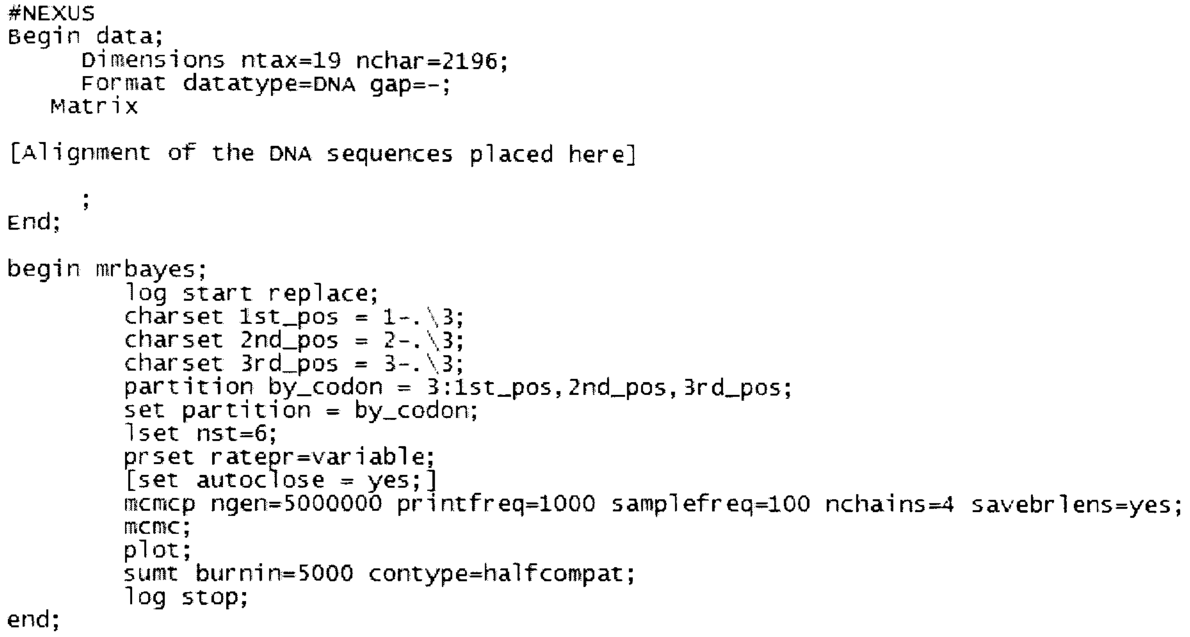
MrBayes was used to create a Bayesian tree. The code used to generate a Bayesian tree was a combination of code from Phylogenetic Trees Made Easy [19] and code and information from the MrBayes program manual [33]. The code can be seen in Figure 2.3. Once again, the DNA sequence alignment was removed from the code in the figure to save space.

```
#NEXUS
Begin data;
     Dimensions ntax=19 nchar=2196;
     Format datatype=DNA gap=-;
   Matrix

[Alignment of the DNA sequences placed here]

     ;
End;

Begin PAUP;

[This turns off all user-prompts.]
set autoclose=yes warnreset=no increase=auto;

[This specifies a distance method.]
set criterion = distance;

[This estimates the tree by the Neighbor-Joining
method with ties broken randomly.]
NJ BreakTies=Random;

[This saves the tree with branch lengths.]
SaveTrees BrLens=yes MaxDecimals=4 File=dnanjbs11000.tre
replace = yes;

[bootstrap]
log start = yes file = dnanjbs1000.log replace = yes;|
Bootstrap search = NJ nreps = 1000 conLevel = 50;
saveTrees from = 1 to=1 file=dnanjbs21000.tre
saveBootP=nodeLabels maxDecimals=1 replace=yes;
log stop;

End;
```

Figure 2.1. Neighbor Joining Phylogenetic Tree Code

```
#NEXUS
Begin data;
     Dimensions ntax=19 nchar=2196;
     Format datatype=DNA gap=-;
   Matrix

[Alignment of the DNA sequences placed here]

     ;
End;

begin paup;
        set autoclose=yes warnreset=no increase=auto;
        charset first = 1-.\3;
        charset second = 2-.\3;
        charset third = 3-.\3;
        charpartition by_codon = 1:first,2:second,3:third;

        set criterion=parsimony;
        hsearch;

        set criterion=likelihood;
        lset nst=6 rmatrix=estimate basefreq=estimate
        rates=sitespec siterates=partition:by_codon;
        lscores 1;
        lset rmatrix=prev basefreq=prev rates=sitespec
        siterates=prev;
        hsearch start=1;
        savetrees brlens=yes maxDecimals=4 file=palbook.ml.trees
        replace=yes;

end;
```

Figure 2.2. Maximum Likelihood Phylogenetic Tree Code



```
#NEXUS
Begin data;
     Dimensions ntax=19 nchar=2196;
     Format datatype=DNA gap=-;
   Matrix

[Alignment of the DNA sequences placed here]

        ;
End;

begin mrbayes;
        log start replace;
        charset 1st_pos = 1-.\3;
        charset 2nd_pos = 2-.\3;
        charset 3rd_pos = 3-.\3;
        partition by_codon = 3:1st_pos,2nd_pos,3rd_pos;
        set partition = by_codon;
        lset nst=6;
        prset ratepr=variable;
        [set autoclose = yes;]
        mcmcp ngen=5000000 printfreq=1000 samplefreq=100 nchains=4 savebrlens=yes;
        mcmc;
        plot;
        sumt burnin=5000 contype=halfcompat;
        log stop;
end;
```

Figure 2.3. Bayesian Phylogenetic Tree Code

2.7. GENE TREE AND SPECIES TREE RECONCILIATION

A species tree was created using the NCBI Taxonomy Browser. The species included in the tree were: *Glycine max*, *Petroselinum crispum*, *Arabidopsis thaliana*, *Pisum sativum*, *Medicago sativa*, *Cicer arietinum*, *Vigna unguiculata*, and *Phaseolus vulgaris*. The species tree was edited using TreeView. The tree was edited because it was a multifurcating tree and caused errors in Notung. The tree was edited according to the phylogenetic tree figures found in the paper by Wojciechowski et al [32]. The species tree is pictured in Figure 2.4. The branch lengths do not represent actual numbers of differences between the species. The species labels were changed on the tree to match the phylogenetic tree abbreviations. The abbreviations had to match so that Notung would be able to reconcile the trees. The Bayesian phylogenetic tree also had to be altered because it was a multifurcating tree. It was modified based on the NJ and ML phylogenetic trees using TreeView. Each of the three phylogenetic trees was reconciled with the species tree by Notung. Default program setting were used. The default duplication cost is 1.5 and the default loss cost is 1.0. After reconciliation, a rooting analysis was done in Notung for each tree. If necessary, the tree was rerooted by clicking on the red edge, which indicated a most parsimonious rooting.

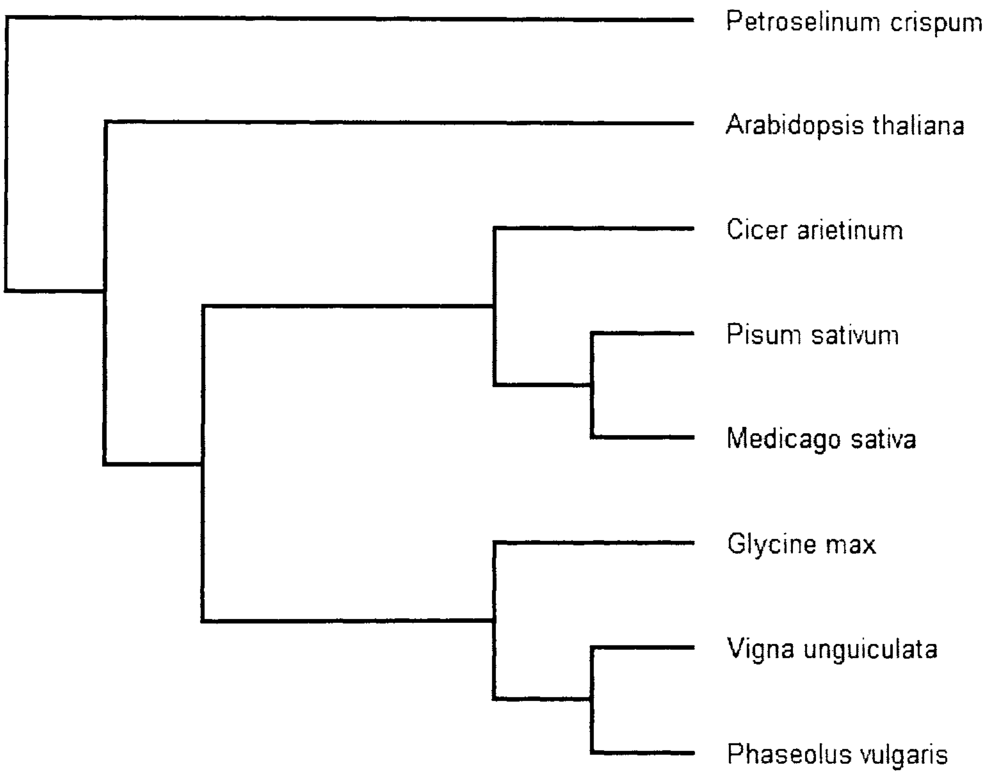


Figure 2.4. Species Tree Used in Notung

2.8. SYNONYMOUS AND NONSYNONYMOUS ANALYSIS

PAL2NAL [34] was used to create a codon alignment. The codon alignment was automatically cropped down by the program to only include the section where all 19 sequences overlapped. The protein alignment and DNA sequences in FASTA format from the Sequence Alignment section were used as input. Under option setting, the output format was changed to FASTA. Other options were left at default settings. The resulting codon alignment was copied and pasted into a text document and saved in FASTA format.

The codon alignment produced by PAL2NAL was used as input for SNAP [26]. All boxes were checked under options (default settings). The default option settings were to show an XY plot of the cumulative behavior of substitutions, neighbor joining trees based on both synonymous and nonsynonymous differences, and SNAP statistics in addition to a summary of results.

2.9. NETWORK ANALYSIS

The DNA alignment generated by CodonAlign was used in SplitsTree4 to generate phylogenetic networks. Neighbor-net, split decomposition, parsimony splits, and median networks were generated using default settings.

The same DNA alignment from CodonAlign was used to generate a median network in Spectronet. The alignment was used to create a median alignment. First, the alignment file was opened in the program. From the "characters" window (which contained the DNA alignment), splits were generated with "get splits." The splits were reduced with "make reduced splits." This reduced the number of splits so that a simpler network could be produced. Finally, a median network was generated from the reduced splits window with "make network." Default settings were used.

2.10. ANALYSIS OF EXPRESSION

Some simple analyses and calculations were done to understand possible conditions of expression for PAL1 and the new PAL sequences in *Glycine max*. A table was made that included the PAL genes in Glycine max, the accession number for each EST belonging with the gene sequence, the library for each EST, the genotype for each EST, and the tissue description for each EST. This information came from the NCBI EST database and the "Index of Soybean cDNA (EST) libraries" at Soybean Genomics Initiative [35]. For some ESTs, the genotype and library could not be determined from the two sources.

The numbers of ESTs for each genotype under each PAL gene were determined by addition. The percentage of ESTs for each genotype was determined for each gene. This was done by dividing the number of ESTs for a specific genotype by the total number of ESTs for each gene.

The numbers of ESTs for each library under each PAL gene were determined by addition. The percentage of ESTs from each library was determined for each gene. This was done by dividing the number of ESTs from a specific library by the total number of ESTs for each gene.

Each library was categorized as stressed or not stressed based on tissue description. Using that information, the number of ESTs that are from stressed libraries was determined for each gene. The percentage of stressed ESTs was determined for each gene by dividing the number of ESTs from stressed libraries by the total number of ESTs for each gene.

The tissue type for each EST was determined based on the library and tissue type description. The total number of ESTs for each tissue type was determined by addition.

3. RESULTS

3.1. RETRIEVAL OF SIMILAR SEQUENCES

The list of accession numbers for the sequences that were retrieved from the BLAST search and saved can be found in Appendix A. The sequences had an E-value < 0.001. A total of 179 sequences were retrieved from the BLAST search.

3.2. ASSEMBLY AND COMPARISON OF CONTIGS

The accession numbers of the contigs assembled by Sequencher can be seen in Table 3.1. The ORF quality of the contigs can be seen in Table 3.2. Percent similarity for contigs when compared to PAL1 and some representative contigs can be seen in Tables 3.3 – 3.6. Differences in codon positions when comparing representative sequences can be seen in Table 3.7. Initial potential PAL gene family group members can be seen in Table 3.8. There were eleven potential members initially. In some cases, "RC" may be seen after a contig name. This refers to the reverse complement of the sequence being used in that situation.

3.3. MAPPING

The map that contains the contigs mapped to PAL1 can be seen in Figure 3.1. The contigs are grouped by the potential gene family member they belong under. The map that contains the unassigned sequences mapped to PAL1 can be seen in Figure 3.2.

Table 3.1. Accession Numbers of Contigs

contig 0001	contig 0001	contig 0001	contig 0004	contig 0005	contig 0007
37995193	26047205	37996037	13311913	10709119	20075547
37996397	26057650	10843183	16346726	13311363	37997435
14125989	14990959	31466076	19938241	14516272	37994452
15287543	37997569	7692476	12772587		37996067
15287581	10237795	37997720	PAL		17518654
26056245	21993773	23735169			14989996
13477608	31561762	37995515			21887608
37995770	17519256	21601763			13480813
10237889	13479342	10709925			
16346064	7796351	10709868			
17022034	17401412	6914562			
23734096	6482967	14516273			
14258962	10237656				
38191098	17964373				

contig 0009	contig 0010	contig 0013	contig 0015	contig 0016	contig 0025
4396122	14205606	21888790	51337607	10237743	37994134
15337807	14205596	21678163	15815750	20812230	21256881
14205605	14206408	7692154	17998799	8282448	
	21600542	8283795	6951362	17153758	
	14990644	19346743	15664149	17519452	
		10237906		15813572	
		22930644		16349046	
		15203390		19935555	
		26268860		19935557	
				17998839	
				6667012	

contig 0026	contig 0029	contig 0037	contig 0040	contig 0041	contig 0045
31306218	23057120	37996285	37994913	37996181	7234039
31467226	4291177	37997633	37995839	37994190	7234197
31467171		37996200	37995872	41145961	
27424231		37994248		58016886	
37994395				58016604	
21676329				16105142	
4290589					
31309360					
21602754					

contig 0046	contig 0047	contig 0051	contig 0052	contig 0055	contig 0059
9264539	9901399	37994280	26047404	10709154	37994428
7640002	13312271	22541806	26056380	26047927	21637794
	37996801				
	5605808				

contig 0060
10237524
48575449

Table 3.2. ORF Quality

Good ORF	Fair ORF	Poor ORF
contig 0041	contig 0013	contig 0007
contig 0051	contig 0026	contig 0001
contig 0016	contig 0029	contig 0025
contig 0046	contig 0005	contig 0010
contig 0055	contig 0060	
contig 0037	contig 0040	
contig 0052		
contig 0009		
contig 0045		
contig 0059		
contig 0047		
contig 0015		
contig 0004		

Table 3.3. Percent Similarities for PAL1

contig 0004	contig 0015	contig 0041	contig 0051	contig 0013	contig 0016	contig 0046	contig 0055	contig 0007	contig 0001	contig 0037	contig 0010	contig 0029	contig 0025	contig 0026
100%	99%	95%	86%	84%	84%	84%	83%	82%	82%	82%	82%	81%	81%	80%

Table 3.4. Percent Similarities for Contig 0016

contig 0010	contig 0046	contig 0029	contig 0001	contig 0052	contig 0013	contig 0025	contig 0037	contig 0007	contig 0026	contig 0051	contig 0009	contig 0055	contig 0005	contig 0004	PAL coding	contig 0041	contig 0045	contig 0060	contig 0015
98%	98%	97%	95%	95%	94%	94%	94%	93%	93%	93%	92%	91%	88%	84%	84%	84%	82%	82%	82%

Table 3.5. Percent Similarities for Contig 0041

PAL1 coding	contig 0004	contig 0015	contig 0051	contig 0007	contig 0046
95%	94%	90%	88%	85%	85%

Table 3.6. Percent Similarities for Contig 0051

contig 0001	contig 0055	contig 0007	contig 0013	contig 0016	contig 0046	contig 0041	contig 0004	PAL1 coding	contig 0010
98%	98%	97%	96%	93%	92%	87%	86%	86%	85%

Table 3.7. Comparison of Codon Positions

Sequence		**Differences**		
First	**Second**	**1st Position**	**2nd Position**	**3rd Position**
PAL1	Contig 0013	26	23	109
PAL1	Contig 0016	35	25	177
PAL1	Contig 0041RC	11	8	37
Contig 0013	Contig 0016	11	3	43
Contig 0013	Contig 0041	29	20	112
Contig 0016	Contig 0041	31	21	115

Table 3.8. Initial Potential PAL Gene Family Members

PALB	***PALC***	***PALD***	***PALE***	***PALF***	***PALG***	***PALH***	***PALI***	***PALJ***	***PALK***	***PALL***	***PAL1***
Contig 0041	Contig 0051 Contig 0055	Contig 0016 Contig 0046	Contig 0037 Contig 0013	Contig 0052 Contig 0047	Contig 0009	Contig 0059	Contig 0040 Contig 0026	Contig 0005	Contig 0060	Contig 0029	PAL1 Contig 0004 Contig 0015

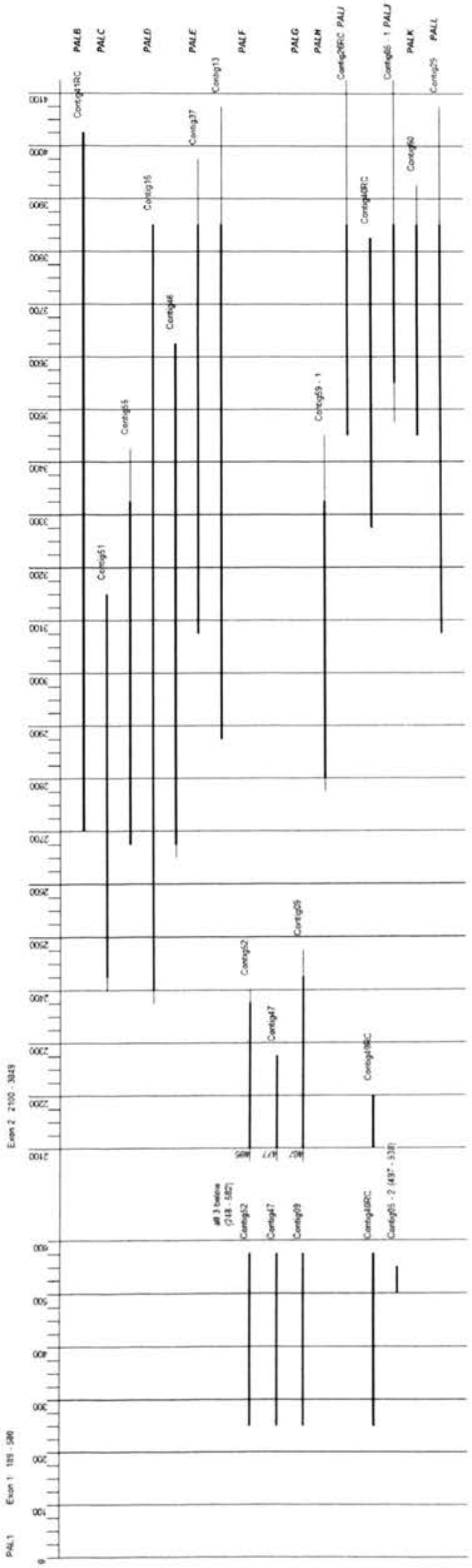


Figure 3.1. Contigs Mapped to PAL1

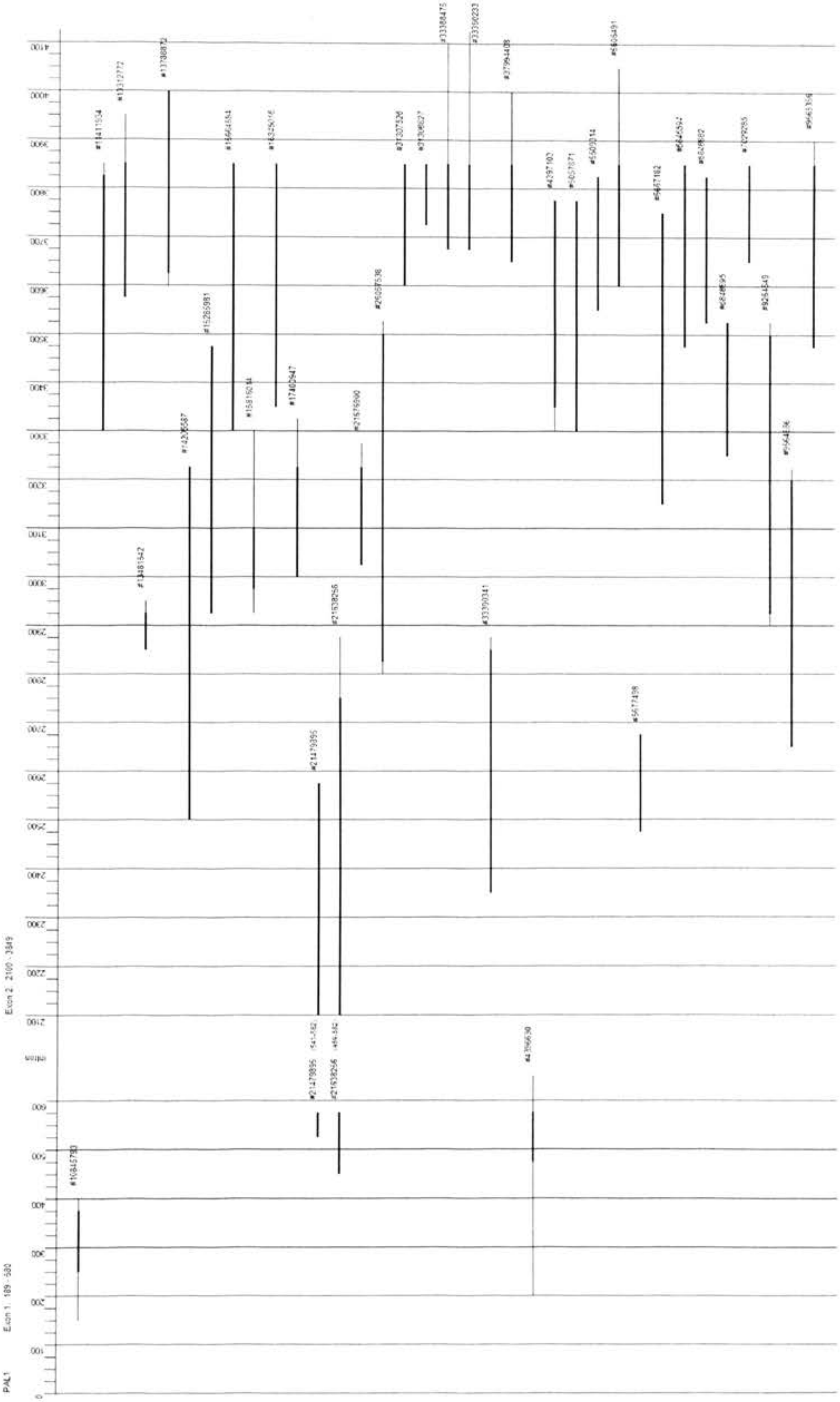


Figure 3.2. Unassigned Sequences Mapped to PAL1

3.4. FINALIZATION OF PAL GENE FAMILY MEMBERS

The finalized PAL gene family member groups can be seen in Table 3.9. A total of three new gene family members were discovered based on overlap and sequence comparisons. They were called PALB, PALC, and PALD. The nucleotide consensus sequences for these new gene family members can be seen in Appendix C.

The map that contains unassigned sequences added to PAL groups can be seen in Figure 3.3. This map shows the groups before they finalized. The consensus sequences for contigs representing the groups found in Figure 3.3 can be seen in Appendix B.

Table 3.9. Finalized New PAL Gene Family Members

PALB	*PALC*	*PALD*
contig 0041	contig 0051	contig 0016
13788872	contig 0055	contig 0046
	contig 0037	14205587
	contig 0013	16345016
	contig60	
	contig26	
	9564686	
	11411934	
	5057871	
	6667182	
	9565356	
	33390233	
	13312772	
	21676900	
	31307526	
	31308827	
	33388475	
	37994408	
	5606491	

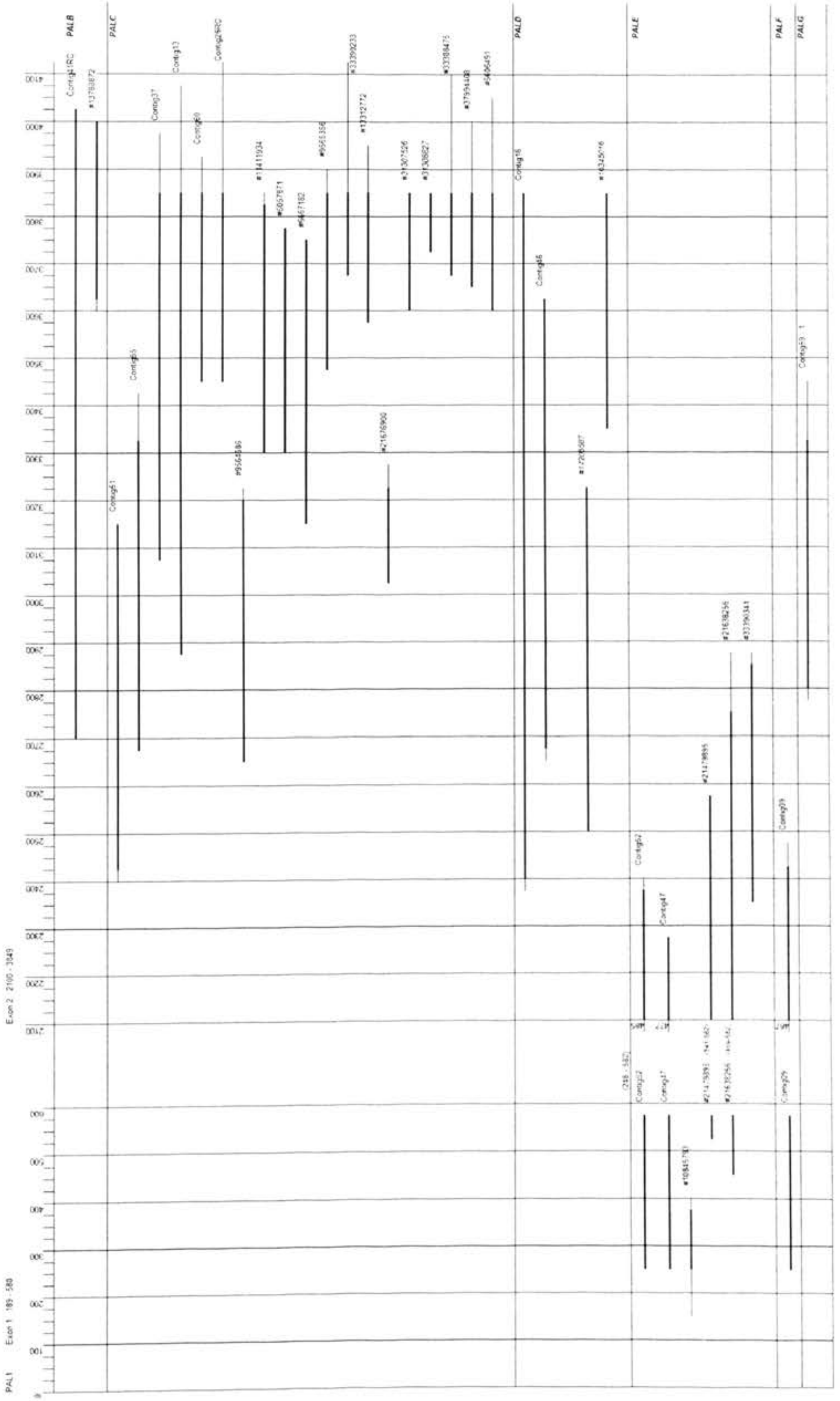


Figure 3.3. Contigs and Unassigned Sequences Mapped to PAL1

3.5. SEQUENCE ALIGNMENTS

The list of species, excluding *G. max*, that had PAL genes used in the alignments can be found in Table 3.10. PAL2 and PAL3 in *Phaseolus vulgaris* did not have nucleotide sequences in the NCBI database. The protein sequences were reverse translated to create nucleotide sequences for use in the alignments [38].

Table 3.10. Accession Numbers of PAL Genes in Alignments

Species	Protein Accession	Nucleotide Accession
Arabidopsis thaliana (1)	P35510	L33677.1
Arabidopsis thaliana (2)	P45724	L33678.1
Arabidopsis thaliana (3)	P45725	L33679.1
Arabidopsis thaliana (4)	Q9SS45	AY303130.1
Cicer arietinum	CAB60719.1	AJ250836.1
Medicago sativa	CAA41169.1	X58180.1
Petroselinum crispum (1)	P24481	Y07654.1
Petroselinum crispum (2)	CAA57056.1	X81158.1
Petroselinum crispum (3)	CAA57057.1	X81159.1
Phaseolus vulgaris (1)	P07218	M11939.1
Phaseolus vulgaris (2)	P19142	n/a
Phaseolus vulgaris (3)	P19143	n/a
Pisum sativum (1)	Q01861	D10002.1
Pisum sativum (2)	Q04593	D10003.1
Vigna unguiculata	AAD45384.1	AF165998.1

3.6. PHYLOGENETIC TREE ANALYSIS

The Neighbor Joining tree generated by PAUP* can be seen in Figure 3.4. The Maximum Likelihood tree generated by PAUP* can be seen in Figure 3.5. The Bayesian tree generated by MrBayes can be seen in Figure 3.6. All trees were viewed in TreeView. All trees are shown with *Petroselinum crispum* PAL genes used as outgroup for rooting. PAL1 from *Glycine max* is called GMax in all three of the trees.

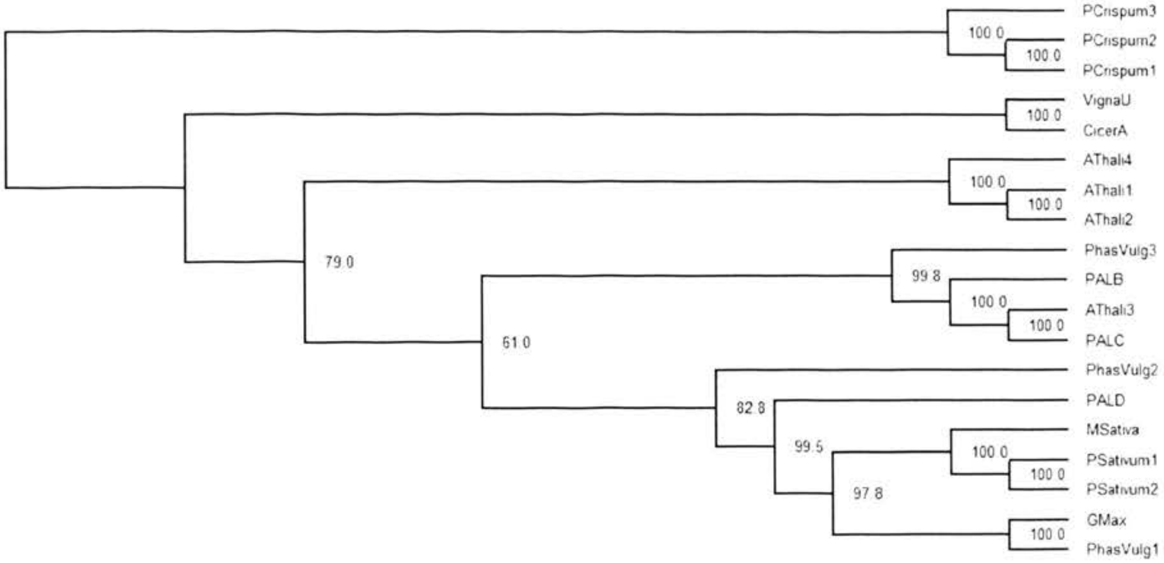


Figure 3.4. NJ Phylogenetic Tree from PAUP*

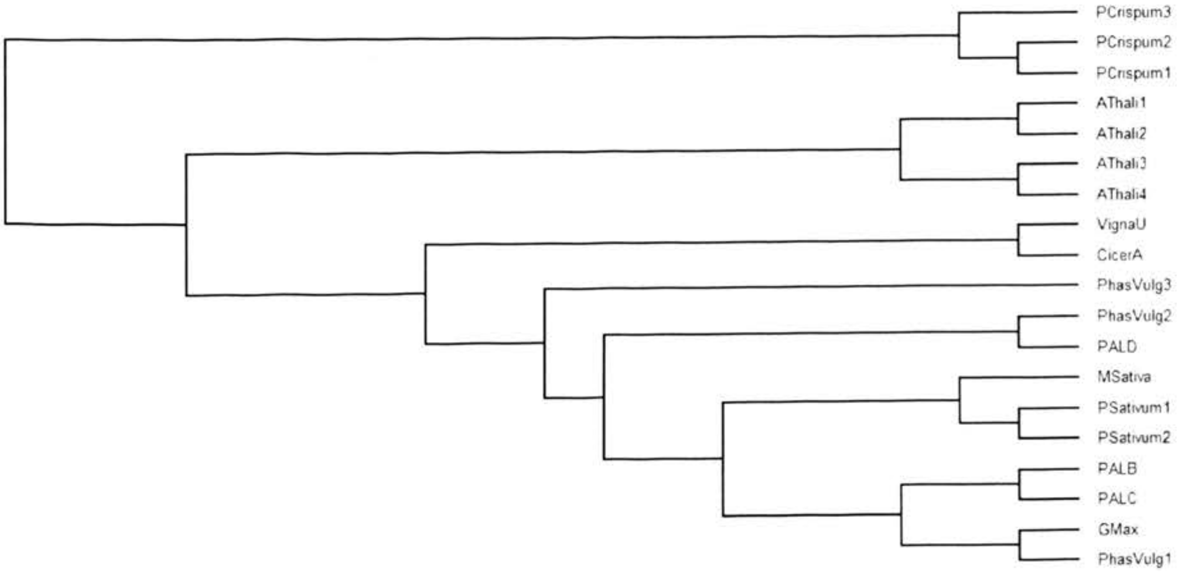


Figure 3.5. ML Phylogenetic Tree from PAUP*

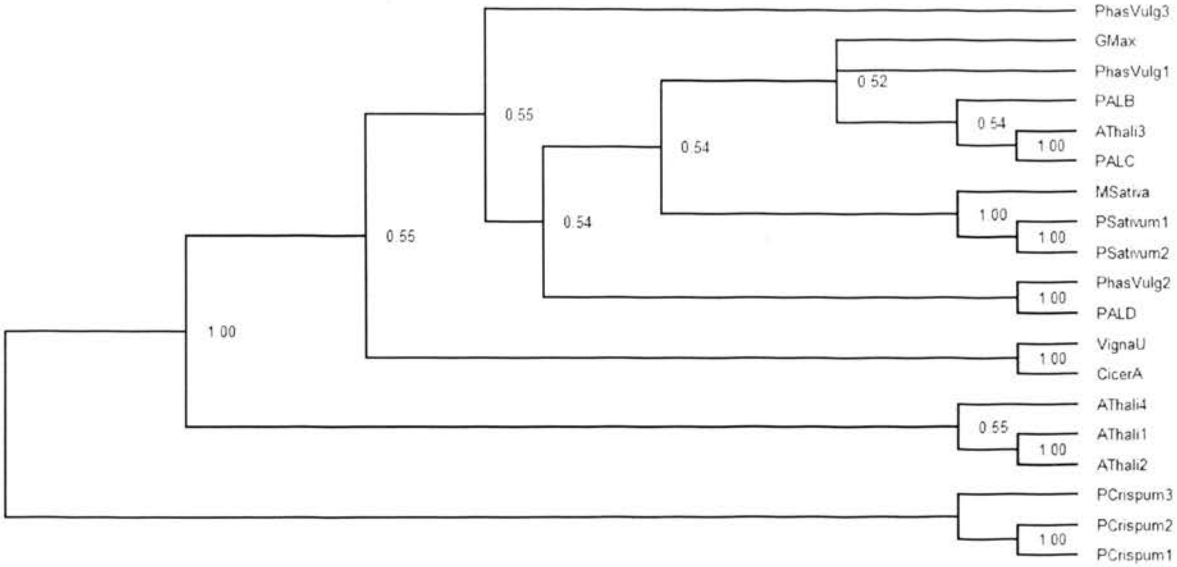


Figure 3.6. Bayesian Tree from MrBayes

3.7. GENE TREE AND SPECIES TREE RECONCILIATION

The species tree, as viewed in Notung, can be seen in Figure 3.7. The node labels are important because they show up in the reconciled trees. They do not have any specific meaning other than referring to a common ancestor. The modified Bayesian tree can be seen in Figure 3.8. The reconciled Neighbor Joining, Maximum Likelihood, and Bayesian trees can be seen in Figures 3.9 – 3.11. All reconciled trees were viewed in Notung. Duplications are indicated with a D at a node. The reconciled NJ tree had a D/L score of 40.0. It had 12 duplications and 22 losses. The reconciled ML tree had a D/L score of 27.5. It had 11 duplications and 11 losses. The reconciled Bayesian tree had a D/L score of 41.0. It had 12 duplications and 23 losses.

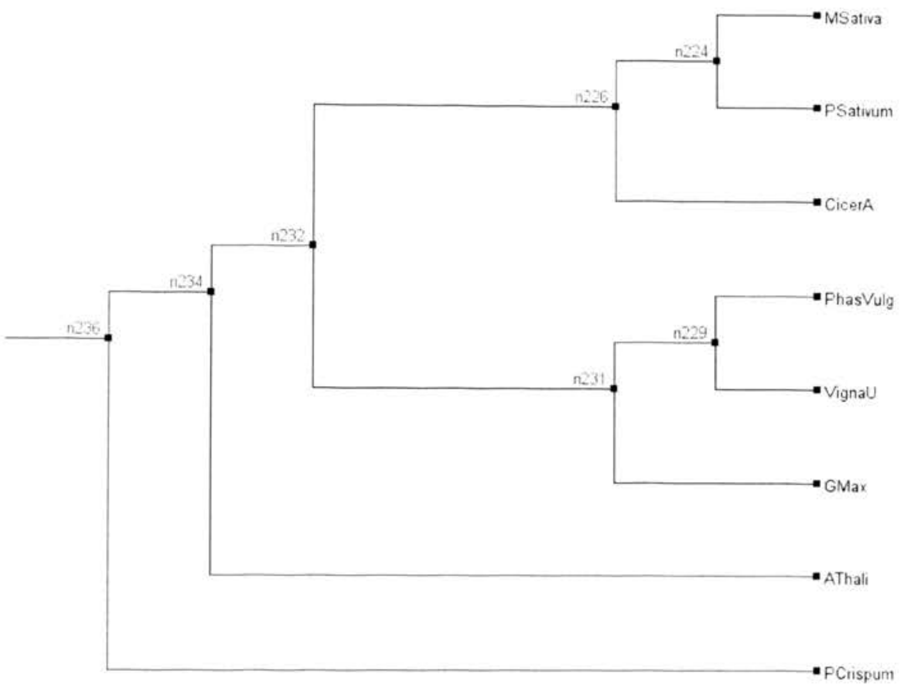


Figure 3.7. Species Tree Viewed in Notung

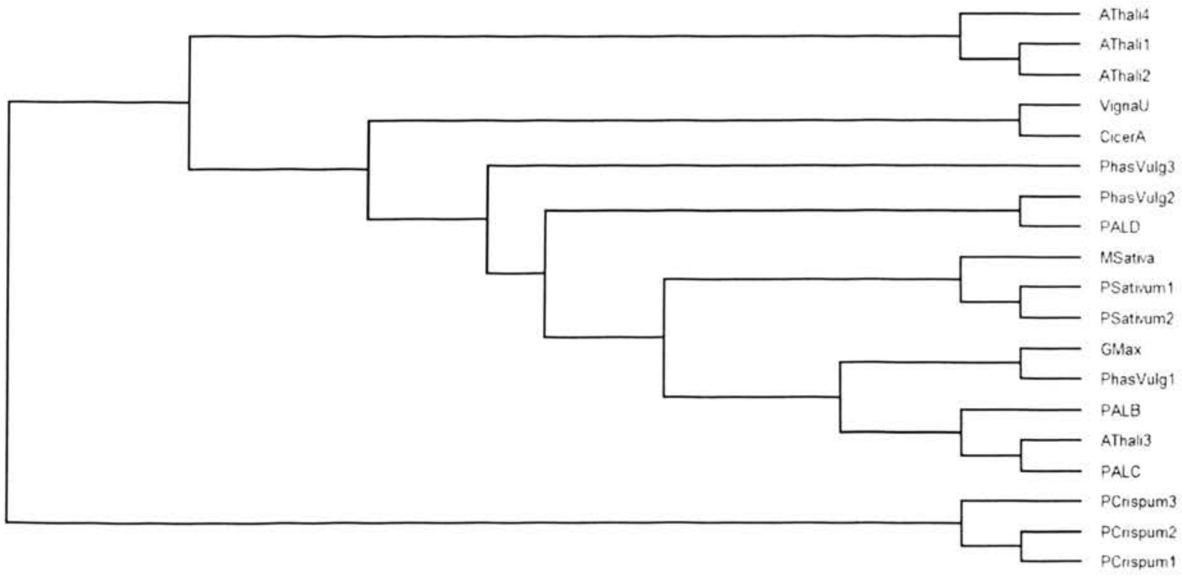


Figure 3.8. Modified Bayesian Tree

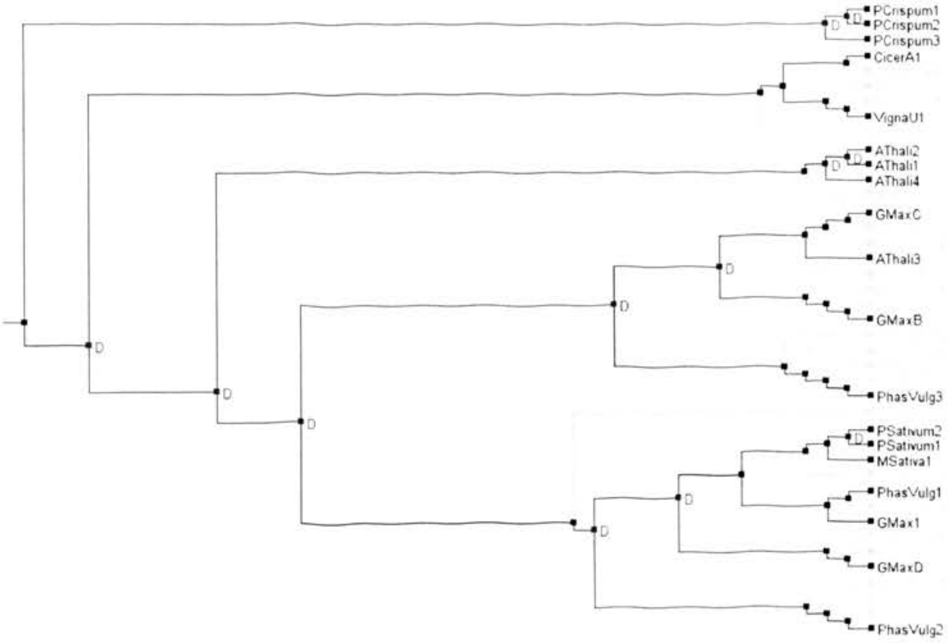


Figure 3.9. Reconciled NJ Tree

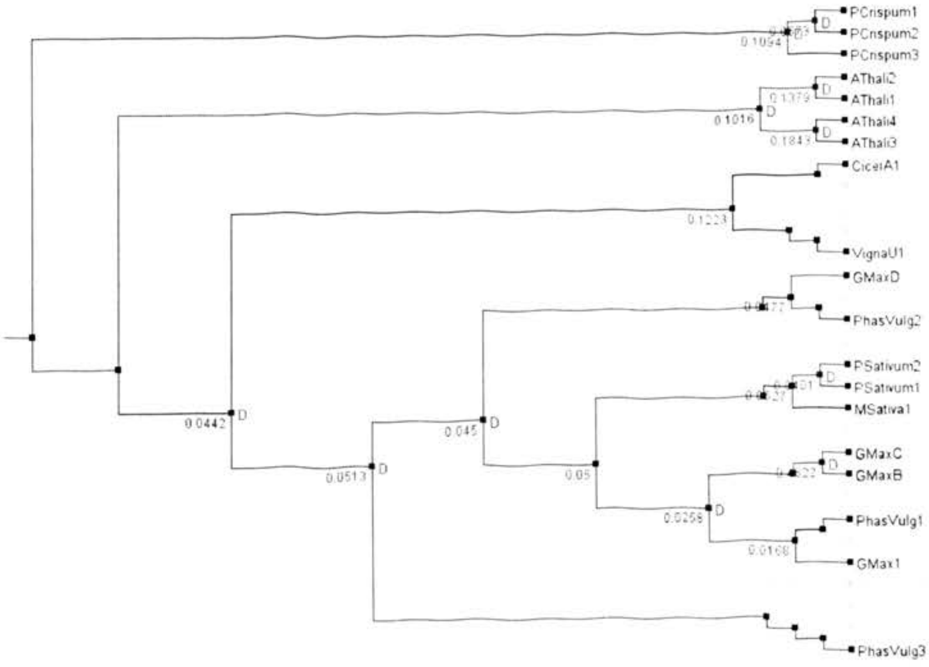


Figure 3.10. Reconciled ML Tree

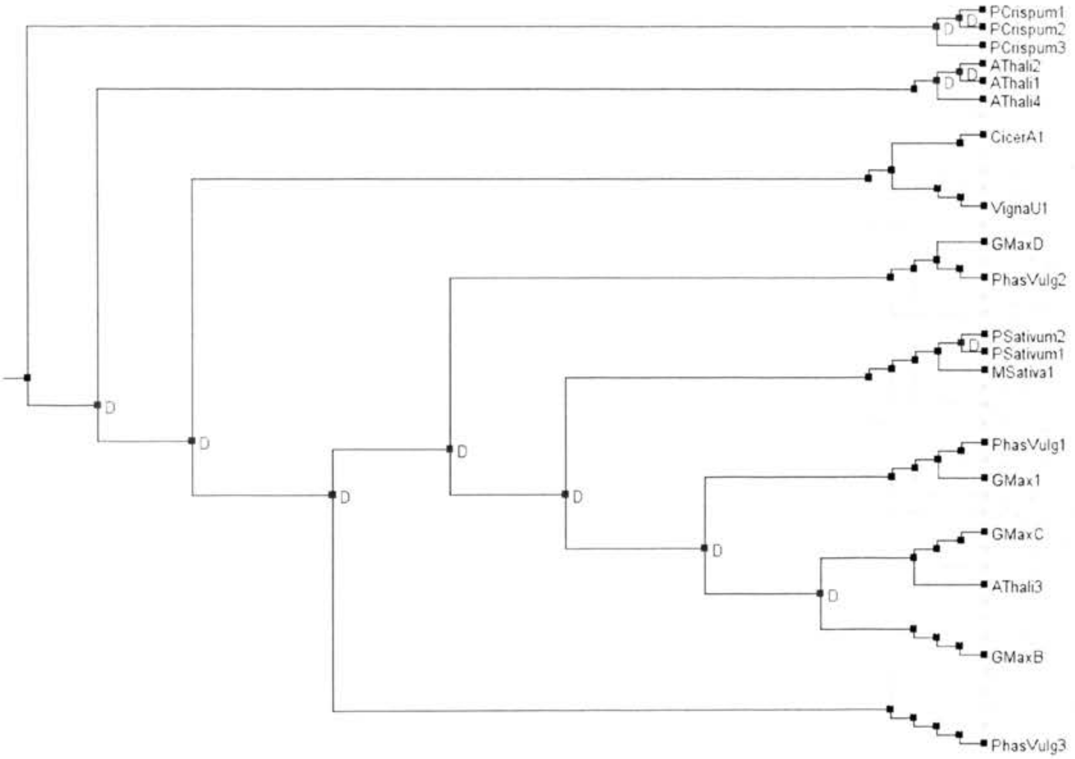


Figure 3.11. Reconciled Bayesian Tree

3.8. SYNONYMOUS AND NONSYNONYMOUS ANALYSIS

The graph generated by SNAP that shows cumulative codon behavior can be seen in Figure 3.12. It shows the cumulative behavior of the average synonymous and nonsynonymous substitutions when moving across the coding region. The Neighbor Joining tree based on synonymous distances and generated by SNAP can be seen in Figure 3.13. The Neighbor Joining tree based on nonsynonymous distances and generated by SNAP can be seen in Figure 3.14. The averages of all the pairwise comparisons can be seen in Table 3.11. Pairwise comparison results from SNAP for the gene family members in *G. max* can be seen in Table 3.12. Descriptions of the variables can be reviewed in Section 1.9.

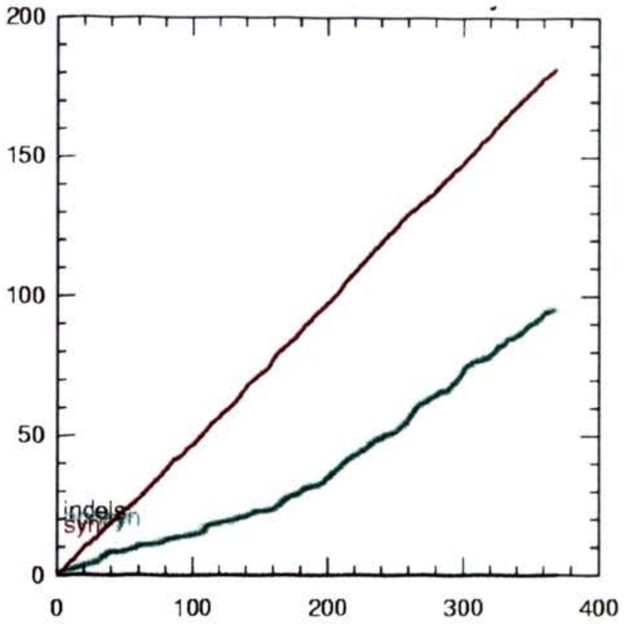


Figure 3.12. Cumulative Behavior by Codon

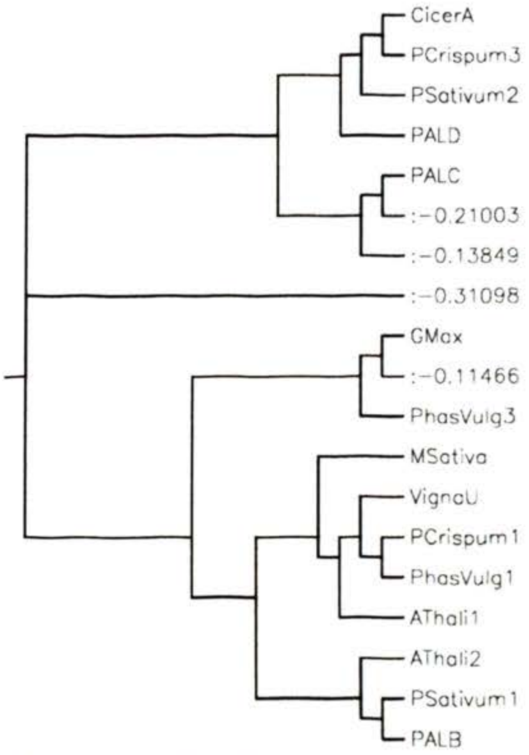


Figure 3.13. NJ Tree from SNAP Based on Synonymous Differences

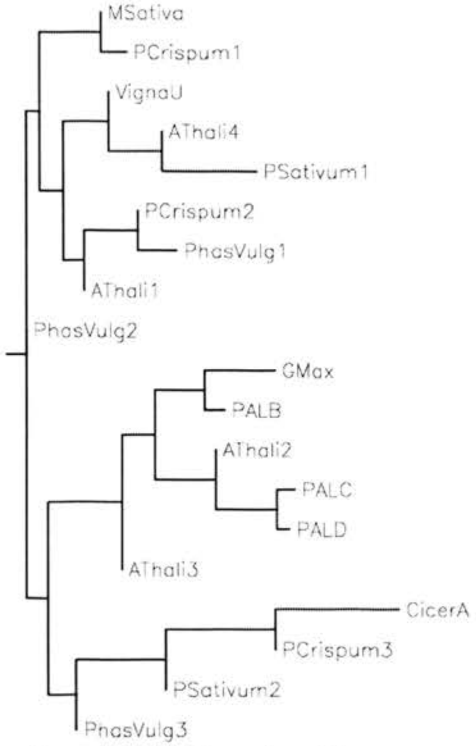


Figure 3.14. NJ Tree from SNAP Based on Nonsynonymous Differences

Table 3.11. SNAP Averages of All Pairwise Comparisons

Variable	Average
ds	1.8596
dn	0.0754
ds/dn	23.2033
ps/pn	9.7743

Table 3.12. SNAP Pairwise Comparisons of PAL Gene Family Members in *G. max*

First	Second	Sd	Sn	S	N	ds	dn	ds/dn
Gmax	PALB	38.0000	12.0000	238.5000	865.5000	0.1791	0.0140	12.7986
Gmax	PALC	119.0000	38.0000	232.0000	845.0000	0.8638	0.0464	18.6262
Gmax	PALD	126.5000	41.5000	235.5000	859.5000	0.9446	0.0499	18.9274
PALB	PALC	123.5000	38.5000	231.8333	842.1667	0.9291	0.0472	19.6983
PALB	PALD	128.0000	42.0000	235.3333	856.6667	0.9688	0.0507	19.1077
PALC	PALD	45.0000	4.0000	229.0000	836.0000	0.2279	0.0048	47.4722

3.9. NETWORK ANALYSIS

The networks generated by SplitsTree4 can be seen in Figures 3.15 – 3.19. The neighbor-net network can be seen in Figure 3.15. The split decomposition network can be seen in Figure 3.16. The parsimony splits network can be seen in Figure 3.17. The median network can be seen in Figure 3.18. A zoomed in view of the median network can be seen in Figure 3.19.

The median network generated by Spectronet can be seen in Figure 3.20. It shows the network after doing reduced splits and pruning.

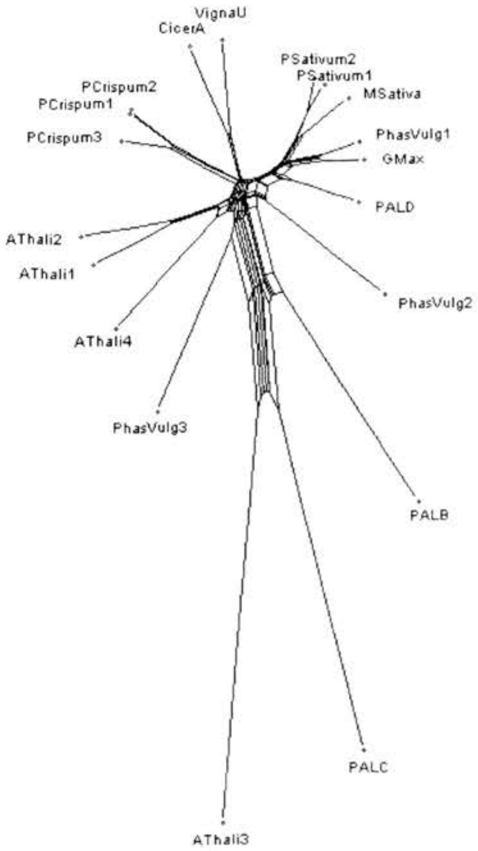


Figure 3.15. Neighbor-net Network from SplitsTree4

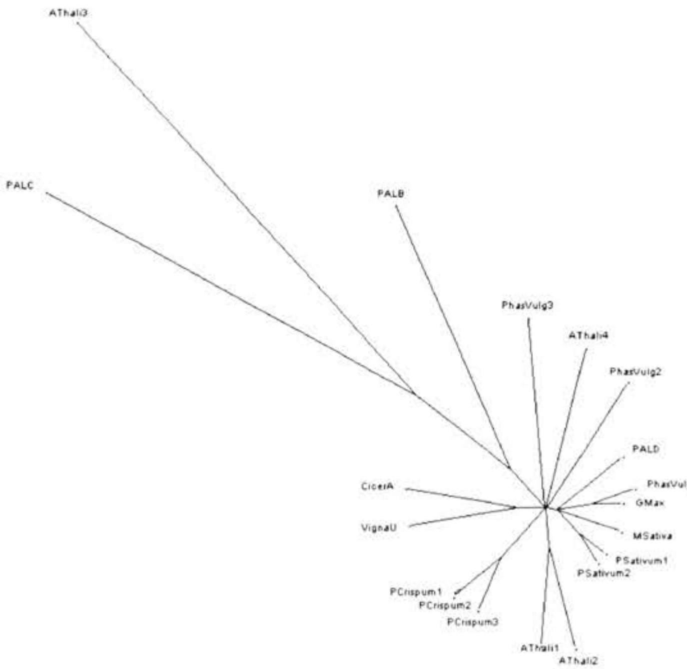


Figure 3.16. Split Decomposition Network from SplitsTree4

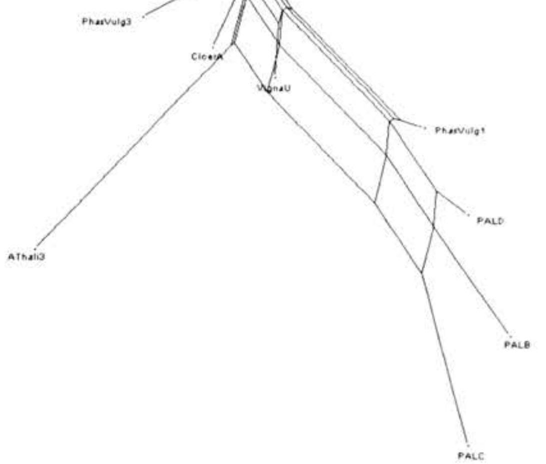


Figure 3.17. Parsimony Splits Network from SplitsTree4

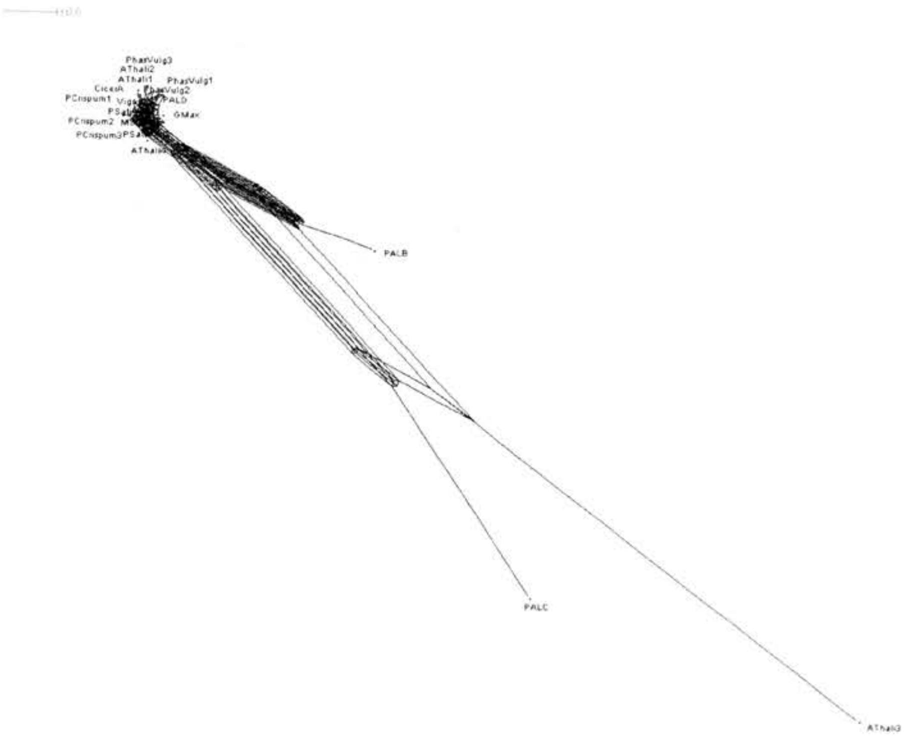


Figure 3.18. Median Network from SplitsTree4

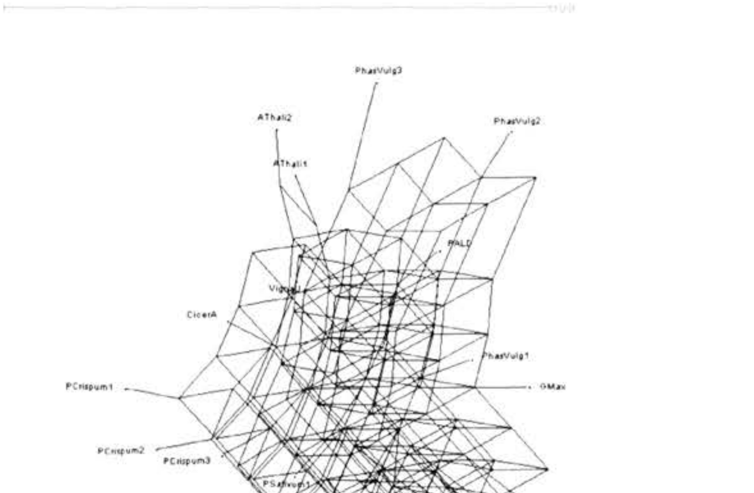


Figure 3.19. Zoomed in View of Median Network from SplitsTree4

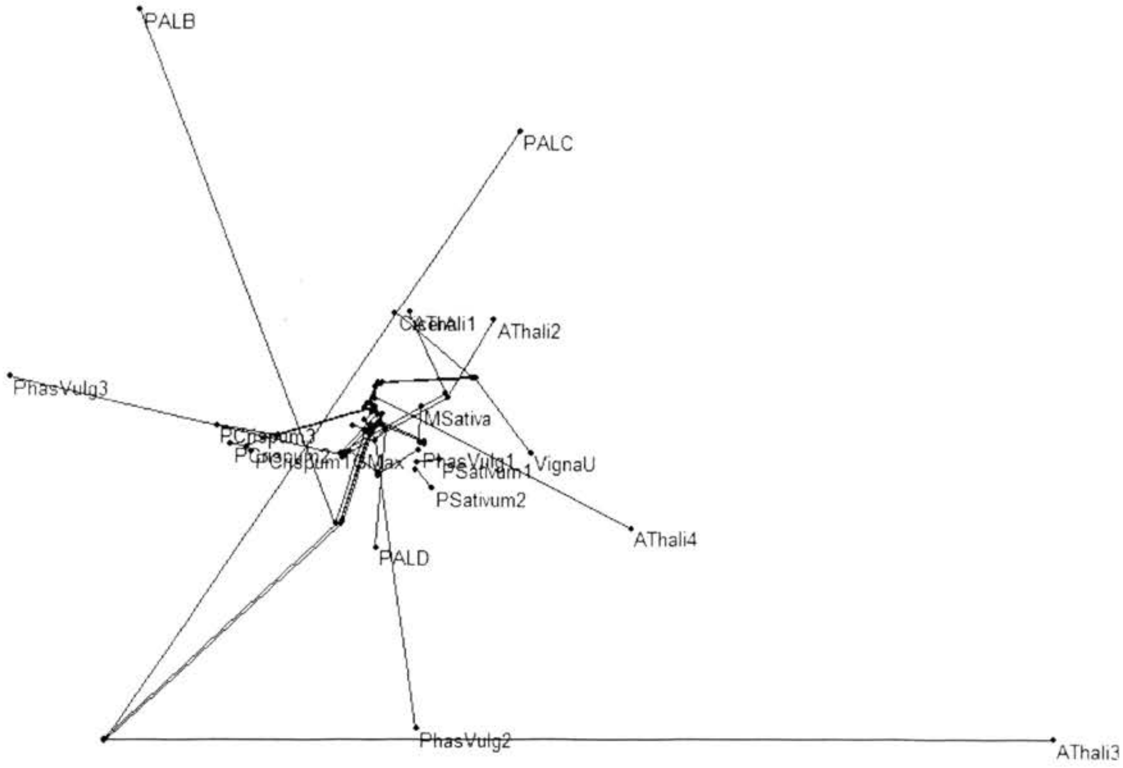


Figure 3.20. Median Network from Spectronet

3.10. ANALYSIS OF EXPRESSION

The information about the ESTs belonging to each PAL gene can be found in Appendix D. The information about the libraries and genotypes of the ESTs, and how much they are represented in each PAL gene, can also be found in Appendix D. The stress information for each library can be found in Appendix D. A stressed library means the members come from *G. max* plants under stressful conditions. The percentage of stressed ESTs for each PAL gene can be seen in Table 3.13. A stressed EST means it came from a stressed library. When just looking at *Glycine max* libraries (specified with "Gm"), the percentage of stressed libraries for each PAL gene can be seen in Table 3.14. For comparison, out of a total of 81 "Gm" libraries, 15 were considered stressed. So 17.65% of the "Gm" libraries are stressed. The tissue type for the ESTs of the PAL genes can be found in Appendix D. The number of ESTs for each tissue type in each PAL gene can be seen in Table 3.15.

Table 3.13. PAL Genes and Stress

Gene	Total ESTs	ESTs From Stressed Libraries	% ESTs from Stressed Libraries
PAL1	9	5	55.55%
PALB	7	5	71.43%
PALC	41	21	51.22%
PALD	15	8	53.33%

Table 3.14. ESTs from Stressed *Glycine max* (Gm) Libraries

Gene	Stressed Gm Libraries	Total Gm Libraries	% of Stressed Libraries
PAL1	2	5	40.00%
PALB	1	2	50.00%
PALC	4	15	26.67%
PALD	5	10	50.00%

Table 3.15. Number of ESTs in Each PAL Gene for Each Tissue Type

	Tissue Type						
Gene	Root	Flower	Stem	Leaf	Cotyledons	Embryo	Pod
PAL1	2	2	3	0	0	0	1
PALB	3	0	3	0	0	1	0
PALC	8	2	15	2	2	0	0
PALD	3	1	4	2	0	0	0

4. DISCUSSION

4.1. RETRIEVAL OF SIMILAR SEQUENCES

As stated in the introduction section, the E-value generated in a BLAST search indicates the significance of a pairwise alignment. Sequences with an E-value of 0.001 or less were chosen using the methods used in paper [36] as a guideline. However, choosing sequences with an E-value greater than 0.001 would not necessarily have affected the outcomes for contig assembly and gene family members. If any sequences were chosen from the search due to chance and not significant similarity, they would have been removed in later analyses. The matches due to chance would not assemble into contigs properly. They also would not have demonstrated patterns expected in gene family members, which would result in removal.

4.2. ASSEMBLY AND COMPARISON OF CONTIGS

The coding region of PAL1 was included in the assembly of contigs from ESTs to prevent mistaking contigs representing PAL1 for representing new PAL genes. By including PAL1, any ESTs matching PAL1 were grouped with PAL1 right away.

An acceptable open reading frame (ORF) was important when considering whether or not contigs represented new genes. A poor ORF would have many stop codons that would stop transcription. A poor ORF could indicate the assembly of ESTs that match by chance and not significant similarity. Since the coding region of PAL1 was used in the BLAST search, a contig representing a gene should have a good ORF to allow for proper transcription. However, the presence of some stop codons was accepted because ESTs are not always perfect representations of gene sequences due to errors during sequencing. The creation of consensus sequences could also cause contigs to be imperfect and include stop codons that may not exist in the real gene. Due to poor ORFs, contigs 0001, 0007, 0010, and 0025 were not used in further analyses.

Percent similarity was important when comparing contigs because contigs that are close enough in similarity probably represent the same gene. For the initial assembly of contigs, a similarity of 99% was used to place very similar ESTs together in a contig. That high similarity was used as a starting point to assemble the contigs. Later, a 95% or greater similarity was used to group contig, along with unassigned sequences, together under the same gene family member. Overlap was important when assembling the ESTs and comparing contigs. Enough overlap between two sequences was needed to determine significant similarity.

The contigs were grouped together by similarity to represent possible PAL genes. The initial new PAL genes were not meant to be final at this point. They were a way to group the contigs initially so further analyses could be done.

4.3. MAPPING

Mapping the contigs against PAL1 was important for visualizing how the contigs overlapped each other. Two contigs that do not overlap could represent different parts of the same gene. By looking at how the contigs lined up with PAL1, contigs could be found to bridge gaps between contigs that could not be compared.

Figure 3.1 showed how contigs in the initial PAL groups lined up with PAL1. Contigs from PAL groups B, C, D, E, H, J, K, and L all lined up with PAL1 in exon II. Contigs from PAL groups F, G, and I lined up with PAL1 in exon I. Viewing overlap and placement allowed further comparisons of the groups and their contig members by focusing on overlapping areas. Figure 3.2 showed how the unassigned sequences lined up with PAL1, which helped identify which groups in Figure 3.1 they might belong to based on overlap.

Visualizing how the contigs overlapped each other also allowed for a comparisons of the overlapping sections. If the overlapping sections of two contigs had a high similarity (at least 95%), then those contigs could be grouped together. This allowed groups of contigs to be combined. Unassigned sequences were assigned to contig groups based on the same method of visualizing overlap and determining similarities of the overlapping areas.

4.4. FINALIZATION OF CONTIGS

Figure 3.3 showed unassigned sequences (ESTs) that were assigned to PAL groups and mapped along with contigs. When looking at how the contigs and ESTs lined up with PAL1, it was discovered that relationships could not be determined between some contigs and sequences. PALB, PALC, and PALD lined up with PAL1 in exon II only. PALE and PALF lined up with PAL1 in the second half of exon I and the first part of exon II. PALG lined up with PAL1 in exon II, but it only had one contig as a member. PALE and PALF were removed because there was not enough information (lack of overlap and similarity) to combine them with any of the other gene family members. There was also not enough information to say they were definitely not representing the same genes as the other PAL groups. However, even though these potential genes were removed from further analyses, they could be revisited later when more EST data or more PAL gene family data is available.

Three new PAL genes were finalized due to similarity percentages, alignments, and map information. It is important to remember that PALB, PALC, and PALD are not complete PAL gene sequences. They are only partial sequences that represent most, but not all, of exon II. This can be seen in Figure 3.3.

As seen in Table 3.9, PALC had the most members. It had six contigs and thirteen ESTs. PALD had two contigs and two ESTs, and PALB had one contig and one EST.

4.5. SEQUENCE ALIGNMENTS

When choosing PAL sequences for use in alignments (seen in Table 3.10), an emphasis was placed on using PAL genes present in other species belonging to the *Fabaceae* family. The PAL genes in *Arabidopsis thaliana* and *Petroselinum crispum* were used because they had multiple gene family members. They were also used because *A. thaliana* and *P. crispum* are outside of the *Fabaceae* family.

When using ClustalX, default settings were used. Sequences were not truncated to the same length when aligning the sequences with ClustalX and CodonAlign. The default settings happened to produce a good alignment for the data, but this is not always the case for alignments. Keeping the sequences at full lengths allowed more positions to be compared. However, the lack of full PALB, PALC, and PALD sequences could potentially affect the alignment because they would be missing nucleotides for comparisons.

4.6. PHYLOGENETIC TREE ANALYSIS

Three different phylogenetic trees were generated so that they could be compared. Differences between the trees could indicate problematic or unclear areas in the data. Closeness, or relatedness, of the genes could be determined by looking for common ancestors between the genes, and how recently a common ancestor occurred. In all three trees (Figures 3.4 – 3.6), PAL1 (called Gmax) shared a most recent common ancestor with the first PAL gene in *Phaseolus vulgaris*. These can be considered sister taxa, or sister sequences. When looking at the next most recent ancestor for PAL1 in the NJ tree (Figure 3.4), PAL1 was found in the clade containing the *Medicago sativa* PAL gene and the two *Pisum sativum* PAL genes in addition to the first *P. vulgaris* gene. For the next most recent ancestor of PAL1 in the MJ tree (Figure 3.5), PAL1 was found in the clade containing PALB and PALC in addition to the first *P. vulgaris* gene. When looking at the most recent ancestor for PAL1 in the Bayesian tree (Figure 3.6), PAL1 was found in the clade containing PALB, PALC, and the third *Arabidopsis thaliana* gene in addition to the first *P. vulgaris* gene.

In the NJ (Figure 3.4) and Bayesian (Figure 3.6) trees, PALC shared a most recent common ancestor with the third *A. thaliana* gene. In both of these trees, PALC was in a clade containing PALB in addition to the third *A. thaliana* gene when looking at the second most recent ancestor. In the ML tree (Figure 3.5), PALC shared a most recent common ancestor with PALB.

In the ML (Figure 3.5) and Bayesian (Figure 3.6) trees, PALD shared its most recent common ancestor with the second *P. vulgaris* gene. In the NJ tree (Figure 3.4), PALD shared its most recent common ancestor with the clade made up of the *M. sativa* gene, the two *P. sativum* genes, PAL1, and the first *P. vulgaris* gene.

In the NJ (Figure 3.4) and the Bayesian (Figure 3.6) trees, PALB shared its most recent common ancestor with the clade of the third *P. vulgaris* gene and PALC.

Out of the three trees, the Bayesian tree (Figure 3.6) was the most difficult to generate because of combining code (Figure 2.3) from two different sources. MrBayes was also not as user friendly and required more knowledge about the program to generate results. Generating a Bayesian tree also took longer (overnight) than generating NJ or ML trees. The Bayesian tree was also multifurcating in this case (Figure 3.6) and had to be edited for further use (Figure 3.8). Creating Bayesian trees are recommended for comparison, but only if the user has the time and an efficient computer to run the analyses. The NJ (Figure 3.4) and ML (Figure 3.5) trees were easier to generate than the Bayesian tree. They are both recommended for generation so that they can be compared for differences.

4.7. GENE TREE AND SPECIES TREE RECONCILIATION

In the reconciled trees, the relationships seen in the phylogenetic trees remained the same. The reconciled NJ tree (Figure 3.9) had ten genes that were potentially lost in ancestors. The tree indicated possible lost or not yet discovered genes (numbers in parenthesis) in *G. max* (3), *P. vulgaris* (1), *A. thaliana* (4), *Vigna unguiculata* (3), and *Cicer arietinum* (1). The reconciled ML tree (Figure 3.10) had four genes that were potentially lost in ancestors. The tree indicated possible lost or not yet discovered genes in *G. max* (2), *P. vulgaris* (1), *V. unguiculata* (3), *C. arietinum* (1). The reconciled Bayesian tree (Figure 3.11) had ten genes that were potentially lost in ancestors. The tree indicated possible lost or not yet discovered genes in *G. max* (2), *P. vulgaris* (1), *V. unguiculata* (3), *C. arietinum* (1), and *A. thaliana* (6). The species that were indicated as possibly losing PAL genes may have PAL genes that have not been discovered yet. These species could be a starting point for discovering more PAL genes.

The reconciled ML tree (Figure 3.10) had the least amount of losses. The smaller amount of losses cause a lower D/L score when compared to the other two reconciled trees. When looking for the smallest D/L score, the maximum likelihood tree would be considered the best. The difference in the D/L score is probably due to the placement of the third *A. thaliana* gene. In the ML tree, the third *A. thaliana* gene is grouped together with the other *A. thaliana* genes. In the NJ (Figure 3.9) and Bayesian (Figure 3.11) trees, the third *A. thaliana* gene was grouped with the PALC gene. It is possible that the trees indicate a close relationship between the third *A. thaliana* gene and PALC because PALC is not a full sequence. If a full PALC gene sequence could be determined, that would allow for more comparison sites between the two sequences. That could cause a different relationship between the two genes.

The reconciled ML tree (Figure 3.10) showed a total of eleven potential duplications. Seven of the duplications are lineage specific. They each occurred within a specific species, and no speciation events occurred after these duplications. Two lineage specific duplications occurred in *P. crispum* and three occurred in *A. thaliana.* One lineage specific duplication occurred in *P. sativum* and one occurred in *G. max.* Four duplications occurred in common ancestors found the legume clade, which inlcluded all sequences except those found in *P. crispum* and *A. thaliana.* One duplication occurred in the common ancestor to all of the legumes. Another duplication occurred in the common ancestor that has the clade made up of PAL1, PALB, PALC, PALD, the three *P. vulgaris* genes, the two *P. sativum* genes, and the *M. sativa* gene. The clade had a total of ten genes, not including the possible lost genes. One duplication occurred in the common ancestor that has the clade made up of PAL1, PALB, PALC, PALD, the first two *P. vulgaris* genes, the two *P. sativum* genes, and the *M. sativa* gene. The clade had a total of nine genes, not including the possible lost genes. Another duplication occurred in the common ancestor that has the clade made up of PAL1, PALB, PALC, and the first *P. vulgaris* gene. The clade had a total of four genes, not including the possible lost genes.

4.8. SYNONYMOUS AND NONSYNONYMOUS ANALYSIS

PAL2NAL generated a cropped codon aligned nucleotide alignment of the sequences. This allows for comparison of the segment where all the genes align with each other, but it could potentially leave out information that would help determine relatedness of the sequences. However, when looking at the synonymous and nonsynonymous changes it was necessary to look at sites without gaps for accurate calculations. A codon alignment was also required input for SNAP.

In Figure 3.12, more synonymous changes than nonsynonymous changes were seen for each codon in the alignment of the sequences. The rate of changes is also linear for synonymous changes. In Table 3.11, the average ds/dn calculated by SNAP based on all pairwise comparisons of the sequences was 23.2033. Since this number is greater than one, which indicates more synonymous changes than nonsynonymous changes, it supports the idea that all of the sequences are from the same gene family.

When comparing the PAL gene family members in *Glycine max* (Table 3.12), all of the members showed more synonymous changes in pairwise comparisons. This supports that they are actual gene family members. The ds/dn scores were all much greater than one, which also indicates membership in the same gene family. The smallest ds/dn was 12.7986 when comparing PAL1 with PALB. The highest ds/dn was 47.4722 when comparing PALC with PALD.

4.9. NETWORK ANALYSIS

In the Neighbor-net network generated by SplitsTree4 (Figure 3.15), PAL1 seemed to have a more significant relationship with the first *P. vulgaris* gene. It also seemed to have a somewhat significant relationship to PALD. PALB seemed to have a significant relationship with both PALC and the second *P. vulgaris* gene. PALC seemed to have the most significant relationship with the third *A. thaliana* gene. PALD seemed to have a significant relationship with the group of the first *P. vulgaris* gene and PAL1 as well as the second *P. vulgaris* gene.

In the Split Decomposition network generated by SplitsTree4 (Figure 3.16), PAL1 shared its node with the first *P. vulgaris* gene. PALB showed a possible significant relationship with the group of PALC and the third *A. thaliana* gene. PALC shared a node with the third *A. thaliana* gene. PALD did not have a clear significant relationship.

In the Parsimony Splits network generated by SplitsTree4 (Figure 3.17), PALB, PALC, and PALD all seemed to have significant relationships with each other. A relationship also seemed to be indicated between PALC and the third *A. thaliana* gene. It was difficult to significant relationships for PAL1.

In the median network generated by SplitsTree4 (Figure 3.18), PALC seemed to have a significant relationship with the third *A. thaliana* gene. Due to the setup of the network, it was too difficult to tell the relationships for PAL1, PALB, and PALD.

The median network generated by Spectronet (Figure 3.20) was also difficult to interpret. PALC seemed to still share a node with the third *A. thaliana* gene.

The network data shows that networks can be another useful way for viewing relationships between gene family members. They can support previous analyses, such as phylogenetic trees. Networks can also give new information or help clear up conflicting information from trees. For example, the third *A. thaliana* gene had different placements in the phylogenetic trees (Figures 3.4 – 3.6). However, the Split Decomposition network (Figure 3.16) supports PALC having a significant relationship with the third *A. thaliana* gene. Networks can be easy to generate with available programs. However, they can also be difficult to interpret. Each network must be interpreted according to the method used to generate the network. For a new user, distinguishing relationships can be difficult, especially when some areas of the networks can become cluttered with lines. In this case significant interpretations were difficult or impossible. Out of the five generated networks, the split decomposition network (Figure 3.16) was the clearest and potentially easiest to understand. The neighbor-net network (Figure 3.15) and the Parsimony Splits network (Figure 3.17) were the next clearest networks.

4.10. ANALYSIS OF EXPRESSION

The ESTs that were used to form PALB, PALC, and PALD each came from specific libraries and genotypes. For comparison, ESTs that also matched up with PAL1 were included when looking at expression. The occurrences of the genotypes and libraries can be seen in Appendix D. For the ESTs that matched PAL1, the Williams genotype was seen most often at 55.55%. The library seen most often was Gm-c1084 at 34.34%. For the ESTs belonging with PALB, the Williams genotype was seen most often at 57.14%. There were two equal libraries, gmrtDrNs01 and USDA-IFAFS, seen at 28.57% each. For the ESTs belonging with PALC, the Williams genotype was seen most often at 36.58%. The USDA-IFAFS library was seen most often at 17.07%. For the ESTs belonging with PALD, the Williams genotype was seen most often at 46.67%. The library seen most often was Gm-c1084 at 13.33%.

While the Williams genotype was seen most often in all of the PAL groups, the significance is not known because their percentages would need to be compared to the percentage of the Williams genotype among all of the ESTs in the database. The significance of the Gm-c1084, USDA-IFAFS, and gmrtDrNs01 libraries being seen most often in the PAL groups is also not known. Their percentages of occurrence in the PAL groups would need to be compared to their overall percentages of occurrences in the EST database.

When looking at stressed libraries, the number of ESTs as well as the number of stressed “Gm” libraries were determined. The number of ESTs from stressed libraries can be seen in Table 3.13. For the ESTs matching PAL1, 55.55% of the ESTs were from stressed libraries. For PALB, 71.43% of the ESTs were from stressed libraries. For PALC, 51.22% of the ESTs were from stressed libraries. For PALD, 53.33% of the ESTs were from stressed libraries. These percentages indicate that it is common to find PAL genes expressed in soybean plants under stress. The higher percentage for PALB may be due to the smaller sample size. PALB only had 7 ESTs. When looking at the “Gm” libraries that made up the PAL genes (Table 3.14) , 40% of the libraries under PAL1 were stressed. For PALB, 50% of the libraries were stressed. For PALC, 26.67% of the libraries were stressed. For PALD, 50% of the libraries were stressed. When looking at all of the possible “Gm” libraries, only 17.65% of them were stressed. This also supports that PAL genes can be found in stressed soybean plants. The smaller percentage in PALC may be due to a larger sample size (making it more accurate) or the diversity of the libraries from which the ESTs came. The dominant library in PALC was USDA-IFAFS, which is not a “Gm” library.

The type of tissue that the ESTs of the PAL genes came from was also considered. The number of ESTs for each tissue type can be seen in Table 3.15. For PAL1, most of the matching ESTs came from stem tissue. Three ESTs were from stem tissue. Two ESTs were from root tissue, two ESTs were from flower tissue, and one EST was from pod tissue. For PALB, most of the ESTs came from root and stem tissue evenly. Three ESTs were from root tissue and three were from stem tissue. One EST was from embryo tissue. For both PALC and PALD, most of the ESTs came from stem tissue. In PALC, fifteen ESTs were from stem tissue. In PALD, four ESTs were from stem tissue. The stem tissue of soybeans seems to be the tissue were PAL expression is most likely to be found. However, the significance would need to be determined by comparing how often ESTs were found in the tissue types in the PAL genes to how often all ESTs were found in the tissue types in the EST database.

Statistical analyses would have been ideal for analyzing expression. The lack of tools and knowledge for performing those analyses prevented their use. However, the basic analyses that were performed do give some general information about expression. They allowed for some observations about expression to be made. Their exact significance is unknown due to the lack of the application of statistical methods.

5. CONCLUSION

Using PAL1 in Glycine max, similar ESTs in G. max were found from an EST database. These ESTs were assembled into contigs based on similarity. The contigs were assembled into groups representing possible new PAL genes. The contigs in the groups were mapped again PAL1 to view overlap. New PAL gene family members in G. max were determined. These new gene family members were compared using phylogenetic analyses and synonymous and nonsynonymous analysis. The expression of the ESTs that made up the new family members was also studied.

From this method, three new PAL genes in *Glycine max* were identified. They were named PALB, PALC, and PALD. The sequences representing these genes were not full sequences, however. The sequences lined up with exon II of PAL1 in *G. max*. Percent similarities indicated that the three PAL genes were family members with PAL1. Synonymous and nonsynonymous analysis also supported family membership. Looking at the EST details, approximately half of the ESTs came from stressed libraries for each family member.

This method could be used to find PAL gene family members in other plant species, other genes in G. max, and other genes in other plant species. Any automation of the steps would allow the whole process to be completed faster.

Complete sequences for the three new PAL genes would be ideal. The partial sequences could be used for guidance to sequence the actual genes from soybean plants. Successful sequencing of the gene family members would further support this method of finding new gene family members.

APPENDIX A.

ACCESSION NUMBERS

Accession numbers of ESTs from BLAST search:

4290589	9264549	14125989	17153758	22930644	37994408
4291177	9564686	14205587	17400947	23057120	37994428
4395675	9565356	14205596	17401412	23734096	37994452
4396122	9901399	14205605	17518654	23735169	37994913
4396630	10237524	14205606	17519256	26047205	37995071
4397103	10237656	14206408	17519452	26047404	37995193
5057871	10237743	14258962	17964373	26047927	37995515
5509314	10237795	14516272	17998799	26056245	37995770
5605808	10237889	14516273	17998839	26056380	37995839
5606491	10237906	14989996	19346743	26057538	37995872
5677498	10709119	14990644	19935555	26057650	37995991
6482967	10709154	14990959	19935557	26268860	37996037
6667012	10709666	15000839	19938241	27424231	37996067
6667182	10709868	15203390	20075547	31306218	37996181
6846594	10709925	15285981	20812230	31307526	37996200
6848882	10843183	15287543	21256881	31308827	37996285
6848895	10845793	15287581	21479895	31309360	37996397
6914562	11411934	15336939	21600542	31466076	37996801
6951362	12772587	15337807	21601763	31467171	37997230
7029285	13311363	15664149	21602754	31467226	37997435
7234039	13311645	15664594	21637794	31467227	37997569
7234197	13311913	15813572	21638256	31561762	37997633
7640002	13311980	15815750	21676329	33388475	37997720
7686543	13312271	15816014	21676900	33390233	38191098
7692154	13312772	16105142	21678163	33390341	41145961
7692476	13477608	16345016	21887608	37994134	48575449
7796351	13479342	16346064	21888790	37994190	51337607
8282448	13480813	16346726	21993773	37994248	58016604
8283795	13481542	16349046	22541806	37994280	58016886
9264539	13788872	17022034	22927963	37994395	

APPENDIX B.

CONTIG SEQUENCES

Contig 0009

aaaacctccactcattccataaatctctgtttactctcctcgattttaccgcaacatgacacaagaaggaaatggcaacaccaacttc
tgtatgagtgttaacaacaacggctacattagcgctaatgacccgttgaactggggcgcggcggcggaggcgatggccggga
gccacctcgacgaggtcaagcgcatggtggaggagtaccggaggccggtcgtgaagctcggcggcgagaccctgacgatct
cgcaggtggcggcgatcgcggcgcacgaccagggggtgaaggtggagctggcggagtcctccagggccggggttaaggc
cagcagcgactgggtgatggagagcatggacaagggcactgacagctacggcgtcaccactgggttcggtgctacctcccac
cggagaaccaaacaaggcggtgccttgcagaaggagctaattaggtttttgaatgctggaatatttggcaatggtacagagtcca
attgcaccctaccccacacagcaaccagagcagctatgctagtgagaatcaacacactcctccaaggctactcaggaattaggtt
tgaaattttggaggcaatcacaaagcttctgaacaacaacattaccccatgtttgccacttaggggaacaatcacagcatctggtg
accttgttcctttgtcctacattgctggtttgctaactggtagaccaaactccaaggctgttggacctctggtgaattctgaatgcma
agaagcctttgaattggccacattagtgctgagtctttgagttgcaactaaggaaggcttgcctt

Contig 0013

gcaccagggaacaaggcacttcatggtggtaacttccaaggaactcctattggagtctccatggataatacacgtttggctcttgct
tcaattggtaaactcatgtttgctcaattctctgagcttgtcaatgattattacaacaatggtttgccttcaaatctcactgccagcaga
aaccccagcttggattatggattcaagggagctgaaattgccatggcatcttattgttctgaacttcaatatttggcgaatccggtga
caagccacgtgcaaagcgcggagcaacacaaccaagatgtgaactctctggggctgatttcatcaaggaagactcatgaggct
attgagatcctcaagctcatgtcctccactttcctggtcgccctttgccaagccattgacttgaggcatttggaggagaatttgaaga
acacggtcaagaacgttgtgagtcaagttgctaagaggactctcaccacaggtgtcaatggagagcttcacccttcaaggttttgt
gagaaggacttgctcaaggttgttgatagggagtacacatttgcatacattgatgacccctgcagtggaacatacccttttgatgca
aaagctaaggcaagtgcttgtggactatgcattggccaatggagagaacgagaagaacacaaacacatcaatcttccaaaagat
tgcaacatttgaggaagagttgaagacccttttgcctaaggaagtggaaggtgcaagagttgcatatgagaatgaccaatgtgca
attccaaacaagatcaaggaatgcaggtcttaccccttgtacaagtttgtgagagaggagttggggacagcattgctaactggtg
aaagggttatctcaccgggtgaagagtgtgacaaagtgttcactgctttgtgccaagggaagatcattgatccacttttggaatgcc
ttggggagtggaatggggcacctcttccaatatgttagtttttcttattttctgttttcttgaagagtggtttcttttctgtacacgtgtttgt
gttgatattaagcatttggtttgtctatataaggctgtggcaaatcaatccacatacaacaacttcccagttttccttgatgtatgccatg
caaggaacttgtaattcataatgtaatagaattccatttgtttgccgtagctttgcgtgcaaatatcaataaaaaaaaaaaa

Contig 0016

ggtgaaattctgaatgccaaagaagcctttgaattggccaacattagtgctgagttctttgagttgcaacctaaggaaggccttgcc
cttgtgaatggcactgctgttggttctggcttggcttcaattgttctttttgaagccaacatcattgctgtcttgtctgaggttatttcagc

aatttttgctgaagtgatgcaagggaagccagagttcactgaccatttgactcataaactaaagcaccaccctggacagattgaag
ctgctgctatcatggaacacattttggaaggaagctcttacatcaaagctgctaagaagttgcatgagattgatcctttgcaaaagc
ctaaacaagaccgctatgcacttaggacttcaccacaatggcttggtcctcaaattgaagtgattagattctctaccaagtcaattga
gagggagataaactcagtcaatgacaaccctttgattgatgtctcaaggaacaaggcccttcatggtggtaacttccaaggaaca
cctattggagtgtccatggataacacccgtttggctcttgcatcaattggcaagctcatgtttgctcaattctctgagcttgtcaatga
ctattacaacaatgggttgccctcaaatctcactgccagcagaaaccccagcttggattatggattcaagggagctgaaattgcaa
tggcctcttattgctctgaactccaatacttggcgaacccggtgacgagccacgtgcaaagcgccgagcaacacaaccaagatg
tgaactctctcgggctgatttcatcaaggaagacacatgaggctattgagatcctcaagctcatgtcctccactttcctcattgcactt
tgccaagccattgacttgaggcatttggaggagaatttgaagaacacggtgaagaacgttgtgagccaagttgctaagcggact
ctcaccacaggtgtcaatggagagcttcacccttcaaggttttgtgagaaggacttgctcaaggttgttgatagggagtacacattt
gcatacattgatgacccctgcagtggcacataccctttgatgcaaaagctgaggcaagtgcttgtggactatgcattggccaatgg
ggagaacgagaagaacacgaacacatcaatcttccaaaagatcgcaacatttgaggaggagttgaagacccttttgcctaagga
agtggaaggtgcaagagttgcatatgagaatgaccaatgtgctattcccaacaagatcaaggaatgcaggtcttaccccttgtac
aagtttgtgagagaggagttggggacagcattgcttactggtgaaagggttgtctcaccgggtgaagagtgtgacaaagtttttac
tgctatgtgccaagggaagatcattgatccacttttggaatgccttggagagtggaatggtgctycmmytymawttg

Contig26 (Reverse Complement)

agagggtggaacataccctttgatgcaaaagctaagsmmrrkgcttgtggactatgcattggccaatggagagaacgagaag
aacacaagcacatcaatcttccaaaagattgcaacatttgaggaagagttgaagacccttttgcctaaggaagtggaaggtgcaa
gagttgcatatgagaatgaccaatgtgcaattccaaacaagatcaaggaatgcaggtcttaccccttgtacaagt-
ttgtgagagaggagttggggacagcattgctaactggtgaaagggttatctcaccgggtgaagagtgtgacaaagtgttcactgc
tttgtgccaagggaagatcattgatccacttttggaatgccttggggagtggaatggggcacctcttccaatatgttagtttttcttatt
ttctgttttcttgaagagtggtttcttttctgtacacgtgtttgtgttgatattaagcatttggtttgtctatataaggctgtggcaaatcaat
ccacatacaacaacttcccagttttccttgatgtatgccatgcaaggaacttgtaattcataatgtaatagaattccatttgtttgccgta
gctttgcgtgcaaatatcaatacatggccttccatgtgaaggatgttttctcttaaaaaaaaaa

Contig 0041 (Reverse Complement)

ctaagaagttgcatgagattgatccattgcaaaagccaaaacaagatcgatatgcccttagaacttcaccacaatggcttggtcct
ctcattgaagtgattcgtttctcgactaagtcaattgagagagagattaactctgtgaatgacaaccctttgattgatgtctcaaggaa
caaggcattacatggtgtcattctccaaggaaccccaattggagtctctatggacaacacgcgtctggctcttgcatctattggcaa
actcatgtttgctcaattctctgagcttgtcaatgatttttacaacaatgggttgccttcaaatctcactgctagcagaaatcctagcttg

gactatgggttcaagggagctgaaattgccatggcttcttactgctctgaactccaatatcttgcaaatccagtaactagccatgtcc
aaagtgctgagcagcataaccaggatgtgaactctttgggtttaatttcatccagaaagacaaatgaagctatcgagatcnttaag
ctcatgtcttccacattcttgattgcactttgccaagcgattgacttgaggcatttggaggagaatttgaaaaactcggtcaagaaca
ctgtgagccaagtttccaaaaggattcttaccacaggtgtcaatggagaactccatccttcaagattttgtgaaaaggatctgctaa
aagtggttgatagggagtacgtattttcctacattgatgacccctgcagtgctacatacccattgatgcaaaaacttaggcaagtgc
ttgtagatcatgccttggtaaatgcagagaatgagaaggatatgaacacatccatctttcaaaagatagcaaactttgaggaggag
ttgaagaatttcttgccaaaagaggttgaaagtgcaagggttgcttatgagagtggcaaagctgcaattccgaacaagatccaag
aatgcagatcttacccactgtacaagtttgtgagagaggaattagggactgggttgctaactggagagaaggtcaggtcaccag
gtgaagagtttgacaaattattcacagcaatgtgccagggcaaaattattgatcctcttctggagtgccttggggagtggaatgga
gctcctcttccaatctgttgattttactataacttttacaaatattttctttgtacctatgcaagtgcaaccataatcatttggtttgtcaatc
ctttaacaaatgttcctttaatgtcaaataggaccttgtaatttaatattttaatggaatttcagtagtttgccggagctttggttctawtat
ata

Contig 0051

ggcacgagaattggccatatcggtgctgagttctttgagttgcaacctaaggaaggccttgcccttgtgaatggcactgctgttggt
tctggcttggcctcaattgttctatttgaagccaacatcattgctgtcttgtctgaagttatttcagcaatttttgctgaagttatgcaagg
aaagcctgaattcactgaccatttgactcataaactaaagcaccaccctggtcagattgaagctgctgctattatggaacacattttg
gaaggaagctcttacgtgaaagctgctaagaagttgcatgagattgatcctttacaaaagcctaaacaggaccgttatgctcttag
gacttcaccacaatggcttggtcctctaattgaagtgattagattctctaccaagtcaattgagagggagattaactcagncaatga
caacccyttgattgatgtgtcaaggaacaaggcacttcatggtggtaacttccaaggaactcctattggagtctccatggataatac
acgtttggctcttgcttcaattggtaaactcatgtttgctcaattctctgagcttgtcaatgattattacaacaatggtttgccttcaaatct
caccgccagcagaaaccccagcttggattatggattcaagggagctgaaattgccatggcatcttattgttcttaacttcaatatttt
gcgaatccggtgacaagccacgtccaaac

Contig 0052

caataacaatattattctcctcattccttcatttttaaacctagctccatctccctccactcaccataacatggcatcagaagcaaatgc
tgccaacaccaacttctgtgtaaatgttagcaacaatggctacattagtgctaatgaccccttgaactggggtgcggcggcggag
gctatggctgggagccacctcgacgaggtcaagcgcatgctagaggagtaccggaggcccgtcgtcaagctcggtggagag
accctgaccatctcgcaggtcgcggcgatcgcggcccacgaccagggggtgaaggtggagctggcggagtcctccagggc
cggtgttaaggccagcagtgactgggtgatggagagcatgaacaagggcactgacagctacggcgtcaccaccgggttcggt
gctacctcccaccggagaaccaaacagggcggtgccttgcagaaggagctaattaggtttttgaatgctggaatatttggcaatg

gtacagagtccaattgcaccctaccccacacagcaaccagagcagctatgctagtgagaatyaacacactcctccaaggctact
caagaatcaggtttgaaattttggaggcaatcacaaagcttctgaacaacaacattaccccatgtttgccacttaggggaacaatc
acagcatctggtgatcttgttcctttgtcctacattgctgggttgctaactgggaaaacaaactccaaggctgttggaccctccggtg
agattctgaatgccaaa

Contig 0055

gtttggaaggaagctcttacgtgaaagctgctaagaagttgcatgagattgatcctttacaaaagcctaaacaggaccgttatgctc
ttaggacttcaccacaatggcttggtcctctaattgaagtgattagattctctaccaagtcaattgagagggagattaactcagtcaa
tgacaaccctttgattgatgtgtcaaggaacaaggcacttcatggtggtaacttccaaggaactcctattggagtctccatggataa
tacacgtttggctcttgcttcaattggtaaactcatgtttgctcaattctctgagcttgtcaatgattattacaacaatggtttgccttcaa
atctcactgccagcagaaaccccagcttggattatggattcaagggagctgaaattgccatggcatcttattgttctgaacttcaata
tttggcgaatccggtgacnagccacgtgcaaagcgcsgagcaacacaaccaagatgtgaactctctggggctgatttcatcang
gaagactcatgaggctattgagatcctcaagctcatgtccctcactttcctggccgccctttggcaagccattgacttgaggcatttt
gaggagaatttgaagacccggtcaagaacggtttgagtcaagttgctagaggactctccccaaggtgcaatggaagctccaccc
tcaaggtttgaaaaagacttgcttcaggtgtta

Contig 0059

ggcacgaggtccacagattgaaatcatccggtattcgaccaaatcaattgaaagggaaataaactcagtaaatgacaatcccttg
attgatgtcacaangnaataaggcactgaatggtggtaatttccaaggaaccccaattggagtttcaatggataatgcacgtttag
ctgttgcttcaattggcaaactcatctttgcccaatttactgagctagtcaatgatttgtataacaatgggttgccatcaaatctttctgc
tggtagaaacccaagtctggattacggtttcaaggcatctgaagttgccatggctgcttattgttctgaacttcaatatctagcaaatc
cagtaacgagccatgtgcaaagtgctgagcagcacaaccaagatgtgaactctttgggcttaatttctgctttgaaaactgtcgaa
gccgttganatattaaagctcatgtcttcgacttatctggttgcactctgccaagctattgacttgaggcatttggaggaaaatttcaa
gantacggtcaagaatactgtaagcaganttgcacagaaaacattaattacagaaggcaaagaagaaattaacccatttcgacttt
gtgagaaagatttgcttaaagtggtcgatagagagtacgtattttcctacattgatgatc

APPENDIX C.

CONSENSUS SEQUENCES OF NEW PAL GENE FAMILY MEMBERS

PALB

aagaagttgcatgagattgatccattgcaaaagccaaaacaagatcgatatgcccttagaacttcaccacaatggcttggtcctct
cattgaagtgattcgtttctcgactaagtcaattgagagagagattaactctgtgaatgacaacccttgattgatgtctcaaggaac
aaggcattacatggtgtcattctccaaggaaccccaattggagtctctatggacaacacgcgtctggctcttgcatctattggcaaa
ctcatgtttgctcaattctctgagcttgtcaatgatttttacaacaatgggttgccttcaaatctcactgctagcagaaatcctagcttgg
actatgggttcaagggagctgaaattgccatggcttcttactgctctgaactccaatatcttgcaaatccagtaactagccatgtcca
aagtgctgagcagcataaccaggatgtgaactctttgggtttaatttcatccagaaagacaaatgaagctatcgagatcnttaagct
catgtcttccacattcttgattgcactttgccaagcgattgacttgaggcatttggaggagaatttgaaaaactcggtcaagaacact
gtgagccaagtttccaaaaggattcttaccacaggtgtcaatggagaactccatccttcaagattttgtgaaaaggatctgctaaaa
gtggttgatagggagtacgtattttcctacattgatgacccctgcagtgctacatacccattgatgcaaaaacttaggcaagtgcttg
tagatcatgccttggtaaatgcagagaatgagaaggatatgaacacatccatctttcaaaagatagcaaactttgaggaggagttg
aagaatttcttgccaaaagaggttgaaagtgcaagggttgcttatgagagtggcaaagctgcaattccgaacaagatccaagaat
gcagatcttacccactgtacaagtttgtgagagaggaattagggactgggttgctaactggagagaaggtcaggtcaccaggtg
aagagtttgacaaattattcacagcaatgtgccagggcaaaattattgatcctcttctggagtgccttggggagtggaatggagctc
ctcttccaatctgt

PALC

ttgcatgagattgatcctttacaaaagcctaaacaggaccgttatgctcttaggacttcaccacaatggcttggtcctctaattgaagt
gattagattctctaccaagtcaattgagagggagattaactcagncaatgacaacccyttgattgatgyrycarggaacaaggca
cttcatggtggtaacttccaaggaactcctattggagtctccatggataatacacgtttggctcttgcttcaattggtaaactcatgttt
gctcaattctctgagcttgtcaatgattattacaacaatggtttgccttcaaatctcacygccagcagaaaccccagcttggattatg
gattcaagggagctgaaattgccatggcatcttattgttctkaacttcaatatttkgcgaatccggtgacaagccacgtscaaassg
cggagcaacacaaccaagatgtgaactctggggctgatttcatcaaggaagactcatgaggctattgagatcctcaagctcat
gtcctccactttcctggtcgccctttgccaagccattgacttgaggcatttggaggagaatttgaagaacacggtcaagaacgttgt
gagtcaagttgctaagaggactctcaccacaggtgtcaatggagagcttcacccttcaaggttttgtgagaaggacttgctcaag
gttgttgatagggagtacacatttgcatacattgatgacccctgcagtggaacataccctttgatgcaaaagctaaggcaagtgctt
gtggactatgcattggccaatggagagaacgagaagaacacaaacacatcaatcttccaaaagattgcaacatttgaggaagag
ttgaagacccttttgcctaaggaagtggaaggtgcaagagttgcatatgagaatgaccaatgtgcaattccaaacaagatcaagg
aatgcaggtcttaccccttgtacaagtttgtgagagaggagttggggacagcattgctaactggtgaaagggttatctcaccgggt
gaagagtgtgacaaagtgttcactgctttgtgccaagggaagatcattgatccacttttggaatgccttggggagtggaatggggc
acctcttccaatat

PALD

aagaagttgcatgagattgatcctttgcaaaagcctaaacaagaccgctatgcacttaggacttcaccacaatggcttggtcctca

aattgaagtgattagattctctaccaagtcaattgagagggagataaactcagtcaatgacaacccttttgattgatgtctcaaggaa

caaggcccttcatggtggtaacttccaaggaacacctattggagtgtccatggataacacccgtttggctcttgcatcaattggcaa

gctcatgtttgctcaattctctgagcttgtcaatgactattacaacaatgggttgccctcaaatctcactgccagcagaaaccccagc

ttggattatggattcaagggagctgaaattgcaatggcctcttattgctctgaactccaatacttggcgaacccggtgacgagcca

cgtgcaaagcgccgagcaacacaaccaagatgtgaactctctcgggctgatttcatcaaggaagacacatgaggctattgagat

cctcaagctcatgtcctccactttcctcattgcactttgccaagccattgacttgaggcatttggaggagaatttgaagaacacggtg

aagaacgttgtgagccaagttgctaagcggactctcaccacaggtgtcaatggagagcttcacccttcaaggttttgtgagaagg

acttgctcaaggttgttgatagggagtacacatttgcatacattgatgacccctgcagtggcacataccctttgatgcaaaagctga

ggcaagtgcttgtggactatgcattggccaatggggagaacgagaagaacacgaacacatcaatcttccaaaagatcgcaacat

ttgaggaggagttgaagacccttttgcctaaggaagtggaaggtgcaagagttgcatatgagaatgaccaatgtgctattcccaa

caagatcaaggaatgcaggtcttacccttgtacaagtttgtgagagaggagttggggacagcattgcttactggtgaaagggtt

gtctcaccgggtgaagagtgtgacaaagtttttactgctatgtgccaagggaagatcattgatccacttttggaatgccttggagag

tggaatggtgctycmmytymawttg

APPENDIX D.

TISSUE EXPRESSION DATA

EST Information For PAL Genes

Gene	EST Accession #	Library	Genotype	Tissue Description (Tissue Type)
B	37994190	USDA-IFAFS	Harosoy	Phytophthora sojae-infected hypocotyl
	37996181	USDA-IFAFS	Harosoy	Phytophthora sojae-infected hypocotyl
	13788872	Gm-c1075	Jack	differentiating somatic embryos cultured on MSM6AC
	41145961	gmrhRww6	Williams 82	root hairs (cDNA clones generated from soybean root hair tissue treated with Bradyrhizobium japonicum for 6 hours)
	58016604	gmrtDrNS01	Williams 82	Water stressed 48h segment 2 (Droughted Roots)
	58016886	gmrtDrNS01	Williams 82	Water stressed 48h segment 2 (Droughted Roots)
	16105142	Gm-c1084	Williams 82	etiolated hypocotyls, inoculated with Phytophthora sojae race 1
C	26268860	Gm-c1048	Clark	whole seedling, 1 week old, greenhouse grown
	27424231	Gm-c1048	Clark	whole seedling, 1 week old, greenhouse grown
	11411934	Gm-c1051	Corolla	floral meristem
	13312772	Gm-c1051	Corolla	floral meristem
	22541806	Gm-c1054	Harosoy	leaf, 3 week old, greenhouse grown
	37994248	USDA-IFAFS	Harosoy	Phytophthora sojae-infected hypocotyl
	37994280	USDA-IFAFS	Harosoy	Phytophthora sojae-infected hypocotyl
	37994395	USDA-IFAFS	Harosoy	Phytophthora sojae-infected hypocotyl
	37994408	USDA-IFAFS	Harosoy	Phytophthora sojae-infected hypocotyl
	37996200	USDA-IFAFS	Harosoy	Phytophthora sojae-infected hypocotyl
	37996285	USDA-IFAFS	Harosoy	Phytophthora sojae-infected hypocotyl
	37997633	USDA-IFAFS	Harosoy	Phytophthora sojae-infected hypocotyl
	31306218	Soybean induced by Salicylic Acid	Kefeng 1	Seedlings

	31307526	Soybean induced by Salicylic Acid	Kefeng 1	Seedlings
	31308827	Soybean induced by Salicylic Acid	Kefeng 1	Seedlings
	31309360	Soybean induced by Salicylic Acid	Kefeng 1	Seedlings
	31467171	Soybean induced by Salicylic Acid	Kefeng 1	Seedlings
	31467226	Soybean induced by Salicylic Acid	Kefeng 1	Seedlings
	33388475	cDNA Peking library 2, 4 day SCN3	Peking	Roots
	33390233	cDNA Peking library 12hr SCN3	Peking	Roots
	10237524	Gm-c1062	Raiden	stem, 1 month old plants, greenhouse grown
	10237906	Gm-c1062	Raiden	stem, 1 month old plants, greenhouse grown
	10709154	Gm-c1062	Raiden	stem, 1 month old plants, greenhouse grown
	26047927	Gm-c1062	Raiden	stem, 1 month old plants, greenhouse grown
	8283795	Gm-c1028	Supernod	roots inoculated with Bradyrhizobium japonicus root
	4290589	Gm-c1004	Williams	entire roots of 8 day old seedlings
	5057871	Gm-c1009	Williams	entire roots of 2 month old plants
	5606491	Gm-c1013	Williams	whole seedlings, 2-3 week old seedlings, greenhouse grown
	6667182	Gm-c1013	Williams	whole seedlings, 2-3 week old seedlings, greenhouse grown
	7692154	Gm-c1027	Williams	cotyledons of 3- and 7-day-old seedlings
	9564686	Gm-c1044	Williams	hypocotyl, 9-10 day old etiolated seedlings
	9565356	Gm-c1044	Williams	hypocotyl, 9-10 day old etiolated seedlings
	15203390	Gm-c1076	Williams 82	wounded cotyledons, 11 day old seedlings

	19346743	Gm-c1068	Williams 82	Leaf, drought stressed, 1 month old plants, greenhouse grown
	21602754	Gm-c1087	Williams 82	Soybean roots without phosphate 11 days after germination
	21676329	Gm-c1073	Williams 82	seedlings induced for symptoms of SDS (Sudden Death Syndrome) disease
	21676900	Gm-c1087	Williams 82	Soybean roots without phosphate 11 days after germination
	21678163	Gm-c1045	Williams 82	hypocotyl, 9-10 day old etiolated seedlings
	21888790	Gm-c1045	Williams 82	hypocotyl, 9-10 day old etiolated seedlings
	48575449	Glycine max mixed library H. glycines, early library	Williams 82	Root
	22930644	Gm-r1088		
D	17998839	Forrest infected Subtraction Library	Forrest	Root
	20812230	Gm-c1052	Harosoy	whole seedling, 1 week old, greenhouse grown
	14205587*	Gm01_AAFC_ECORC_Glycine_max_cold_stressed_leaves	Maple Arrow	Leaves
	17153758	Gm-c1072	PI567374	seedlings induced for symptoms of SDS (Sudden Death Syndrome) disease
	10237743	Gm-c1062	Raiden	stem, 1 month old plants, greenhouse grown
	8282448	Gm-c1028	Supernod	roots inoculated with Bradyrhizobium japonicus root
	6667012	Gm-c1009	Williams	entire roots of 2 month old plants
	15813572	Gm-c1065	Williams	germinating shoot, cold stressed, 3 day old seedlings
	7640002	Gm-c1016	Williams 82	immature flowers, field grown plants
	16349046	Gm-c1068	Williams 82	leaf, drought stressed, 1 month old plants, greenhouse grown
	17519452	Gm-c1074	Williams 82	seedlings induced for HR (hypersensitive response)

	19935555	Gm-c1084	Williams 82	etiolated hypocotyls, inoculated with Phytophthora sojae race 1
	19935557	Gm-c1084	Williams 82	etiolated hypocotyls, inoculated with Phytophthora sojae race 1
	16345016	Gm-r1083		
	9264539	Soybean hypocotyls Lambda Zap library		long hypocotyls of dark grown seedlings
PAL1	15664149	Gm-c1081	Bragg	roots, 7 day old seedlings, mock-infected 48 hours before harvest
	13311913	Gm-c1051	Corolla	floral meristem
	17998799	Forrest infected Subtraction Library	Forrest	Forrest roots were inoculated with Fusarium solani f. sp. glycinae and samples were collected after 14 days of inoculation
	12772587	Gm-c1071	Williams	immature pods (2 cm), greenhouse grown seed pod
	6951362	Gm-c1015	Williams 82	mature flowers, field grown plants
	16346726	Gm-c1084	Williams 82	etiolated hypocotyls, inoculated with Phytophthora sojae race 1
	19938241	Gm-c1084	Williams 82	etiolated hypocotyls, inoculated with Phytophthora sojae race 1
	15815750	Gm-c1084	Williams 82	etiolated hypocotyls, inoculated with Phytophthora sojae race 1
	51337607	Gm-r1089		
	*14205587 was replaced by 92233570			

Genotype Information for PAL Genes

Gene	Genotype	Genotype %	Number of ESTs	Total ESTs
PAL1	Bragg	11.11%	1	9
	Corolla	11.11%	1	
	Forrest	11.11%	1	
	Williams	55.55%	5	
PALB	Harosoy	28.57%	2	7
	Jack	14.29%	1	
	Williams	57.14%	4	
PALC	Clark	4.88%	2	41
	Corolla	4.88%	2	
	Harosoy	19.51%	8	
	Kefeng 1	14.63%	6	
	Peking	4.88%	2	
	Raiden	9.76%	4	
	Supernod	2.44%	1	
	Williams	36.58%	15	
PALD	Forrest	6.67%	1	15
	Harosoy	6.67%	1	
	Maple Arrow	6.67%	1	
	PI567374	6.67%	1	
	Raiden	6.67%	1	
	Supernod	6.67%	1	
	Williams	46.67%	7	

PAL Library Information for PAL Genes

Gene	Library	Library %	Number of ESTs	Total ESTs
PAL1	Gm-c1015	11.11%	1	9
	Gm-c1051	11.11%	1	
	Gm-c1071	11.11%	1	
	Gm-c1081	11.11%	1	
	Gm-c1084	34.34%	3	
	Gm-r1089	11.11%	1	
	Forrest infected Subtraction Library	11.11%	1	
PALB	Gm-c1075	14.29%	1	7
	Gm-c1084	14.29%	1	
	gmrhRww6	14.29%	1	
	gmrtDrNS01	28.57%	2	
	USDA-IFAFS	28.57%	2	
PALC	cDNA Peking library 12hr SCN3	2.44%	1	41
	cDNA Peking library 2, 4 day SCN3	2.44%	1	
	Glycine max mixed library H. glycines, early library	2.44%	1	
	Gm-c1004	2.44%	1	
	Gm-c1009	2.44%	1	
	Gm-c1013	4.88%	2	
	Gm-c1027	2.44%	1	
	Gm-c1028	2.44%	1	
	Gm-c1044	4.88%	2	
	Gm-c1045	4.88%	2	
	Gm-c1048	4.88%	2	
	Gm-c1051	4.88%	2	
	Gm-c1054	2.44%	1	
	Gm-c1062	9.76%	4	

	Gm-c1068	2.44%	1	
	Gm-c1073	2.44%	1	
	Gm-c1076	2.44%	1	
	Gm-c1087	4.88%	2	
	Gm-r1088	2.44%	1	
	Soybean induced by Salicylic Acid	14.63%	6	
	USDA-IFAFS	17.07%	7	
PALD	Forrest infected Subtraction Library	6.67%	1	15
	Gm-c1009	6.67%	1	
	Gm-c1016	6.67%	1	
	Gm-c1028	6.67%	1	
	Gm-c1052	6.67%	1	
	Gm-c1062	6.67%	1	
	Gm-c1065	6.67%	1	
	Gm-c1068	6.67%	1	
	Gm-c1072	6.67%	1	
	Gm-c1074	6.67%	1	
	Gm-c1084	13.33%	2	
	Gm-r1083	6.67%	1	
	Gm01_AAFC_ECORC_Glycine_max_cold_stressed_leaves	6.67%	1	
	Soybean hypocotyls Lambda Zap library	6.67%	1	

Stress Information for Libraries

Gene	Library	Number ESTs	Stressed	Description	Total ESTs
PAL1	Gm-c1015	1	No	mature flowers, field grown plants	9
	Gm-c1051	1	No	floral meristem	
	Gm-c1071	1	No	immature pods (2 cm), greenhouse grown seed pod	
	Gm-c1081	1	Yes	roots, 7 day old seedlings, mock-infected 48 hours before harvest	
	Gm-c1084	3	Yes	etiolated hypocotyls, inoculated with Phytophthora sojae race 1	
	Gm-r1089	1	--	--	
	Forrest infected Subtraction Library	1	Yes	Forrest roots were inoculated with Fusarium solani f. sp. glycinae and samples were collected after 14 days of inoculation	
PALB	Gm-c1075	1	No	differentiating somatic embryos cultered on MSM6AC	7
	Gm-c1084	1	Yes	etiolated hypocotyls, inoculated with Phytophthora sojae race 1	
	gmrhRww6	1	No	root hairs (cDNA clones generated from soybean root hair tissue treated with Bradyrhizobium japonicum for 6 hours)	
	gmrtDrNS01	2	Yes	Water stressed 48h segment 2 (Droughted Roots)	
	USDA-IFAFS	2	Yes	Phytophthora sojae-infected hypocotyl	
PALC	cDNA Peking library 12hr SCN3	1	Yes	Roots	41
	cDNA Peking library 2, 4 day SCN3	1	Yes	Roots	

	Glycine max mixed library H. glycines, early library	1	Yes	Root	
	Gm-c1004	1	No	entire roots of 8 day old seedlings	
	Gm-c1009	1	No	entire roots of 2 month old plants	
	Gm-c1013	2	No	whole seedlings, 2-3 week old seedlings, greenhouse grown	
	Gm-c1027	1	No	cotyledons of 3- and 7-day-old seedlings	
	Gm-c1028	1	No	roots innoculated with Bradyrhizobium japonicus root	
	Gm-c1044	2	No	hypocotyl, 9-10 day old etiolated seedlings	
	Gm-c1045	2	No	hypocotyl, 9-10 day old etiolated seedlings	
	Gm-c1048	2	No	whole seedling, 1 week old, greenhouse grown	
	Gm-c1051	2	No	floral meristem	
	Gm-c1054	1	No	leaf, 3 week old, greenhouse grown	
	Gm-c1062	4	No	stem, 1 month old plants, greenhouse grown	
	Gm-c1068	1	Yes	leaf, drought stressed, 1 month old plants, greenhouse grown	
	Gm-c1073	1	Yes	seedlings induced for symptoms of SDS (Sudden Death Syndrome) disease	
	Gm-c1076	1	Yes	wounded cotyledons, 11 day old seedlings	
	Gm-c1087	2	Yes	Soybean roots without phosphate 11 days after germination	
	Gm-r1088	1	--	--	
	Soybean induced by Salicylic Acid	6	Yes	Seedlings	

	USDA-IFAFS	7	Yes	Phytophthora sojae-infected hypocotyl	
PALD	Forrest infected Subtraction Library	1	Yes	Root	15
	Gm-c1009	1	No	entire roots of 2 month old plants	
	Gm-c1016	1	No	immature flowers, field grown plants	
	Gm-c1028	1	No	roots innoculated with Bradyrhizobium japonicus root	
	Gm-c1052	1	No	whole seedling, 1 week old, greenhouse grown	
	Gm-c1062	1	No	stem, 1 month old plants, greenhouse grown	
	Gm-c1065	1	Yes	germinating shoot, cold stressed, 3 day old seedlings	
	Gm-c1068	1	Yes	leaf, drought stressed, 1 month old plants, greenhouse grown	
	Gm-c1072	1	Yes	seedlings induced for symptoms of SDS (Sudden Death Syndrome) disease	
	Gm-c1074	1	Yes	seedlings induced for HR (hypersensitive response)	
	Gm-c1084	2	Yes	etiolated hypocotyls, inoculated with Phytophthora sojae race 1	
	Gm-r1083	1	--	--	
	Gm01_AAFC_ECORC_Glycine_max_cold_stressed_leaves	1	Yes	Leaves	
	Soybean hypocotyls Lambda Zap library	1	No	long hypocotyls of dark grown seedlings	

Tissue Type for ESTs from PAL Genes

Gene	EST Accession #	Library	Tissue Type
B	37994190	USDA-IFAFS	Stem
	37996181	USDA-IFAFS	Stem
	13788872	Gm-c1075	Embryo
	41145961	gmrhRww6	Root
	58016604	gmrtDrNS01	Root
	58016886	gmrtDrNS01	Root
	16105142	Gm-c1084	Stem
C	26268860	Gm-c1048	Seedling
	27424231	Gm-c1048	Seedling
	11411934	Gm-c1051	Flower
	13312772	Gm-c1051	Flower
	22541806	Gm-c1054	Leaf
	37994248	USDA-IFAFS	Stem
	37994280	USDA-IFAFS	Stem
	37994395	USDA-IFAFS	Stem
	37994408	USDA-IFAFS	Stem
	37996200	USDA-IFAFS	Stem
	37996285	USDA-IFAFS	Stem
	37997633	USDA-IFAFS	Stem
	31306218	Soybean induced by Salicylic Acid	Seedling
	31307526	Soybean induced by Salicylic Acid	Seedling
	31308827	Soybean induced by Salicylic Acid	Seedling
	31309360	Soybean induced by Salicylic Acid	Seedling
	31467171	Soybean induced by Salicylic Acid	Seedling
	31467226	Soybean induced by Salicylic Acid	Seedling
	33388475	cDNA Peking library 2, 4 day SCN3	Root

	33390233	cDNA Peking library 12hr SCN3	Root
	10237524	Gm-c1062	Stem
	10237906	Gm-c1062	Stem
	10709154	Gm-c1062	Stem
	26047927	Gm-c1062	Stem
	8283795	Gm-c1028	Root
	4290589	Gm-c1004	Root
	5057871	Gm-c1009	Root
	5606491	Gm-c1013	Seedling
	6667182	Gm-c1013	Seedling
	7692154	Gm-c1027	Cotyledons
	9564686	Gm-c1044	Stem
	9565356	Gm-c1044	Stem
	15203390	Gm-c1076	Cotyledons
	19346743	Gm-c1068	Leaf
	21602754	Gm-c1087	Root
	21676329	Gm-c1073	Seedling
	21676900	Gm-c1087	Root
	21678163	Gm-c1045	Stem
	21888790	Gm-c1045	Stem
	48575449	Glycine max mixed library H. glycines, early library	Root
	22930644	Gm-r1088	
D	17998839	Forrest infected Subtraction Library	Root
	20812230	Gm-c1052	Seedling
	14205587*	Gm01_AAFC_ECORC_Glycine_max_cold_stressed_leaves	Leaf
	17153758	Gm-c1072	Seedling
	10237743	Gm-c1062	Stem
	8282448	Gm-c1028	Root
	6667012	Gm-c1009	Root
	15813572	Gm-c1065	Seedling
	7640002	Gm-c1016	Flower

	16349046	Gm-c1068	Leaf
	17519452	Gm-c1074	Seedling
	19935555	Gm-c1084	Stem
	19935557	Gm-c1084	Stem
	16345016	Gm-r1083	
	9264539	Soybean hypocotyls Lambda Zap library	Stem
PAL1	15664149	Gm-c1081	Root
	13311913	Gm-c1051	Flower
	17998799	Forrest infected Subtraction Library	Root
	12772587	Gm-c1071	Pod
	6951362	Gm-c1015	Flower
	16346726	Gm-c1084	Stem
	19938241	Gm-c1084	Stem
	15815750	Gm-c1084	Stem
	51337607	Gm-r1089	
	*14205587 was replaced by 92233570		

BIBLIOGRAPHY

[1] Canadian Food Inspection Agency. "The Biology of Glycine max (L.) Merr. (Soybean)," The Canadian Food Inspection Agency, Health Canada, October 1996. http://www.inspection.gc.ca/english/plaveg/bio/dir/t11096e.shtml, April 2008.

[2] PLANTS Profile for Glycine max (soybean), http://plants.usda.gov/java/profile?symbol=GLMA4, April 2008.

[3] Hymowitz, T. "On the domestication of the soybean," Economic Botany, 1970, 24: 408-421.

[4] Hymowitz, T. and J. R. Harlan. "Introduction of soybeans to North America by Samuel Bowen in 1765," Economic Botany, 1983, 37: 371-379.

[5] "Why Sequence the Soybean?" - JGI, http://www.jgi.doe.gov/sequencing/why/soybean.html, April 2008.

[6] National Statistics for Soybeans, http://www.nass.usda.gov/Statistics_by_Subject/index.asp, April 2008.

[7] Singh, Bijay K. Plant Amino Acids: Biochemistry and Biotechnology. New York: Marcel Dekker, Inc, 1999.

[8] Bohm, Bruce A. Introduction to Flavonoids – Chemistry and Biochemistry of Organic Natural Products, Vol 2. Australia: Hardwood Academic Publishers, 1998.

[9] Frank, RL and LO Vodkin. "Sequence and structure of a phenylalanine ammonia-lyase gene from Glycine max," DNA Seq, 1991, 1(5):335-46.

[10] Nucleotide Home, http://www.ncbi.nlm.nih.gov/sites/entrez?db=nuccore, April 2008.

[11] Hurles, M. "Gene Duplication: The Genomic Trade in Spare Parts," PLoS Biol, 2004, 2(7): e206 doi:10.1371/journal.pbio.0020206

[12] "What are Gene Families?" - Genetics Home Reference, http://ghr.nlm.nih.gov/handbook/howgeneswork/genefamilies, April 2008.

[13] "Programs and Activities" - NCBI, http://www.ncbi.nlm.nih.gov/About/glance/programs.html, April 2008.

[14] "ESTs Factsheet" - A Science Primer at NCBI, http://www.ncbi.nlm.nih.gov/About/primer/est.html, April 2008.

[15] Staden, R. "A new computer method for the storage and manipulation of DNA gel reading data," Nucleic Acids Res, 1980, 8, 3673-3694.

[16] Krane, Dan E. and Michael L Raymer. Fundamental Concepts of Bioinformatics. San Francisco: Benjamin Cummings, 2003.

[17] "The BLAST Sequence Analysis Tool" - NCBI handbook , http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.chapter.ch16, April 2008.

[18] BLAST: Basic Local Alignment Search Tool, http://www.ncbi.nlm.nih.gov/blast/Blast.cgi, April 2008.

[19] Hall BG. Phylogenetic trees made easy: A how-to manual for molecular biologists. Sunderland, Massachusetts, Sinauer Associates, Inc, 2001.

[20] "Phylogenetics Factsheet" - A Science Primer at NCBI, http://www.ncbi.nlm.nih.gov/About/primer/phylo.html, April 2008.

[21] D. Durand, B. V. Halldorsson, B. Vernot. A Hybrid Micro-Macroevolutionary Approach to Gene Tree Reconstruction. Journal of Computational Biology , 2006, 13(2):320-335.

[22] B. Vernot, M. Stolzer, A. Goldman, D. Durand. Reconciliation with Non-Binary Species Trees. In Computational Systems Bioinformatics: CSB2007 Conference Proceedings, Imperial College Press, 2007: 441-452.

[23] D. H. Huson and D. Bryant, Application of Phylogenetic Networks in Evolutionary Studies, Mol. Biol. Evol., 2006, 23(2):254-267.

[24] K. T. Huber, M. Langton, D. Penny, V. Moulton and M. Hendy . "Spectronet: A package for computing spectra and median networks," Applied Bioinformatics 1(3), 2002, 159-161.

[25] Korber B. HIV Signature and Sequence Variation Analysis. Computational Analysis of HIV Molecular Sequences, Chapter 4, pages 55-72. Allen G. Rodrigo and Gerald H. Learn, eds. Dordrecht, Netherlands: Kluwer Academic Publishers, 2000.

[26] SNAP web version, http://hcv.lanl.gov/content/sequence/SNAP/SNAP.html, April 2008.

[27] Sequencher – Gene Codes Corporation [http://www.genecodes.com/]

[28] AssemblyLIGN - Oxford Molecular

[29] MacVector – MacVector, Inc. [http://www.macvector.com/index.html]

[30] Swofford, D. L. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts, 2003.

[31] Page, R. D. M. 1996. TREEVIEW: An application to display phylogenetic trees on personal computers. Computer Applications in the Biosciences 12: 357-358.

[32] Wojciechowski M. F., Lavin M., Sanderson M. J. "A phylogeny of legumes (Legumenosae) based on analyses of the plastid matK gene resolves many well-supported subclades within the family," Am. J. Bot, 2004, 91:1846–1862.

[33] MrBayes, http://mrbayes.csit.fsu.edu/index.php, April 2008.

[34] Mikita Suyama, David Torrents, and Peer Bork. "PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments," Nucleic Acids Res, 2006, 34, W609-W612.

[35] "Index of Soybean cDNA (EST) Libraries" - Soybean Genomics Initiative , http://soybean.ccgb.umn.edu/documents/soy_libraries/index_EST.html, April 2008.

[36] Granger, C., V. Coryell, A. Khanna, P. Keim, L. Vodkin, and R.C. Shoemaker. "Identification, structure, and differential expression of members of a BURP domain containing protein family in soybean," Genome, August 2002, Volume 45, Number 4, pp. 693-701(9).

[37] Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D.G. Higgins. "The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools," Nucleic Acids Research, 1997, 24:4876-4882.

[38] Sequence Manipulation Suite: Reverse Translate, http://www.bioinformatics.org/sms2/rev_trans.html, April 2008.

VITA

Erin Kathleen Pringle was born in Columbia, Missouri, USA on March 29, 1982. In December of 2005, she received her B.S., magna cum laude, in Biological Sciences from the University of Missouri-Rolla, Rolla, Missouri, USA. In May 2008, she received her M.S. degree in Applied and Environmental Biology from the Missouri University of Science and Technology.