In Silico Analysis of Candidate Genes Involved in Lignin Biosynthesis from Woody Species and Phylogenetic Relationship

Wong Lai Kuan

(20256)

A Thesis submitted in partial fulfillment of
the requirements for the degree of Bachelor of Science with Honours
(Resource Biotechnology)

Faculty of Resource Science and Technology
UNIVERSITI MALAYSIA SARAWAK
2010

ACKNOWLEDGEMENT

I sincerely thank my supervisor, Dr. Ho Wei Seng, for his valuable guidance and advices in this project. I am grateful to his patient in guiding me. Besides that, I am also thankful to the Forest Genomics Laboratory members for concerning and giving advices for me. My thanks also go to my family and friends for supporting and providing helps when I get into problems.

Table of Content

List of Abbreviation

4CL	4-coumarate coenzymeA ligase
C3H	para-coumarate-3- hydroloase
C4H	cinnamate-4-hyhroxylase
CAD	cinnamyl alcohol dehydrogenase
CCoAOMT	caffeoyl coenzymeA methyltransferase
CCR	cinnamoyl coenzymeA reductase
COMT	caffeic/5-hydroxyferulic acid O-methyltransferase
F5H	ferulate 5-hydrolase
HCT	hydroxycinamoyl transferase
PAL	phenylalanine ammonia-lyase
(G) units	guaiacyl units
(S) units	syringyl units
(H) units	para-hydroxyhenyl units
NJ	Neighbor Joining
MP	Maximum Parsimony
ML	Maximum Likelihood
XML	eXtensible Markup Language

List of Tables

List of Figures

In Silico Analysis of Candidate Genes Involved in Lignin Biosynthesis from Woody Species and Phylogenetic Relationship

Wong Lai Kuan

Resource Biotechnology Program
Faculty of Resource Science and Technology
University Malaysia Sarawak

Abstract

The lignin biosynthesis genes are highly conserved throughout vascular plants and thus it can provide useful information to better understand the evolution of land plants. Three Angiosperm trees namely *Arabidopsis thaliana, Populus trichocarpa, Eucalyptus globulus,* and a gymnosperm tree, *Pinus taeda* were selected in this study. Motif and domain for each gene were identified by scanning it with PROSITE and Pfam databases. In this study, only six of the genes: *4CL, CAD, PAL, C3H, C4H* and *F5H* have motif identified and represented in regular expression. The motif in regular expression can be used to estimate the protein function. Domain analysis showed that two lignin proteins; COMT, CAD are multidomains proteins while C3H, C4H and F5H shared the same domain, p450 which involve in oxidative metabolism of natural compounds. The phylogenetic analyses were carried out using two different approaches implemented in MEGA software: distance based (NJ) and character based (MP). The phylogenetic analysis of lignin genes showed the divergence between angiosperm and gymnosperm species.

Keywords: lignin biosynthesis pathway genes, motif and domain identification, phylogenetic analysis

Abstrak

Gen-gen biosintesis lignin adalah terpelihara dalam tumbuhan-tumbuhan vaskular dan ini membekalkan maklumat yang berguna dalam memahami evolusi tumbuhan-tumbuhan darat. Tiga pokok-pokok Angiosperm iaitu Arabidopsis thaliana, Populus trichocarpa, Eucalyptus globulus, dan satu pokok Gymnosperm Pine taeda terlibat dalam pengkajian ini. Analisasi gen-gen lignin dalam kajian ini menunjukkan hanya enam gen-gen lignin mempunyai motif dalam bentuk regular expression iaitu 4CL, CAD, PAL, C3H, C4H, dan F5H. Motif dalam bentuk regular expression membolehkan sesuatu fungsi protein dijangkakan dengan cepat dan mudah. Analisis dalam gen-gen lignin menunjukkan terdapat dua gen-gen lignin iaitu CAD dan COMT mempunyai multidomain dan domain p450 dikongsi oleh C3H, C4H, dan F5H. Analisis filogenetik telah dijalankan dengan cara NJ dan MP dalam perisian MEGA . Analisis filogenetik menunjukkan kebezaaan angiosperm dan gymnosperm dalam tumbuhan.

Kata Kunci: gen-gen pusat biosintesis lignin, identificasi motif dan domain, filogenetik analisis

CHAPTER 1

INTRODUCTION

A remarkable event in plant evolution is the occurrence of vascular plants in dry land which it took part during the Silurian period around 430 million years ago (Zhong and Ye, 2009). The successful adaptation of vascular plants had facilitated the emersion of large size plants in terrestrial habitat (Boudet, 2000). However, all these will still be illusory without lignin as lignifications of vascular elements provide mechanical support and the hydrophobicity of tracheary elements for water conduction (Zhong and Ye, 2009).

Wood is an important natural and endlessly renewable source of energy. Wood has robust characteristic due to the formation of complex chemicals during the secondary growth of trees such as cellulose, hemicellulose, and lignin synthesis (Plomion *et al*., 2001). Lignin is the second most abundant biopolymer of plant after cellulose which it comprises of about 30% of the dry mass of wood in some tree species (Boerjan *et al*., 2003). Typically, wood is formed by the successive addition of secondary xylem, which differentiates from the vascular cambium (Ko *et al*., 2004). The cells in the vascular cambium also go through a series of processes include cell division, cell expansion, secondary cell wall formation and apoptosis of cells to form wood (Sterky *et al*., 1998). The contributions of wood are mainly as sawn timbers, energy sources and for the production of pulp and paper. (Plomion *et al*., 2001).

Although lignin presence in the trees contributes robust characteristic in wood, research on lignin biosynthesis pathway of woody species has been actively carried out over past two decades mainly to reduce the amount of lignin in wood (Boerjan *et al*., 2003). The presence of lignin in wood is not favour in some industries such as paper production industry, animal forage, biofuel production industry and etc (Xu *et al*., 2009). For example,

lignin have to be separated from pulps for the production of high quality paper and the chemical processes are usually costly, energy consuming and not environmentally friendly (Boerjan *et al*., 2003). In addition, high lignin content in plants also affects the forage digestibility in animals (Xu *et al*., 2009).

The biochemical pathways of lignin biosynthesis are highly conserved throughout vascular plants, and this can provide crucial insight into the understanding of evolution of land plants (Xu *et al*., 2009). Besides that, deep understanding of the functions and evolution of lignin biosynthesis pathway genes will provide researchers to develop strategies for modifying and improving plant species especially those have economic important. The objectives of this study are to determine the functional properties of the genes involved in the lignin biosynthesis pathway through protein motifs and domains identification and to infer the phylogenetic relationship of lignin biosynthesis genes from four different tree species, namely arabidopsis, poplar, pine and eucalyptus.

CHAPTER 2

LITERATURE REVIEW

2.1 Species involved in the study

There are four woody species which normally known as arabidopsis, poplar, pine and eucalyptus were involved in this study. Among these trees, only pine is gymnosperm while the rest of the four tree species are angiosperm.

2.1.1 *Populus trichocarpa*

Populus trichocarpa is from the family of Salicaceae, genus of *Populus* also known as black cottonwood, balsam cottonwood, western balsam poplar and California poplar, is the largest hardwood tree in western North America (Tuskan *et al*., 2006). *P. trichocarpa* is a large and fast growing tree which can grow to 30-50m of height and the diameter of the tree trunk can be over 2 m. The bark of *P. trichocarpa* is grey in colour and is covered with lenticels which allow gas exchange between the atmosphere and the internal tissues. The stem of the tree is grey in the older parts and is light brown in the younger parts. The leaves are 7-20 cm long with a glossy dark green at the upper side and light grey-green at the underside. The leaves are alternate, elliptic with a crenate margin and an acute tip, and reticulate venation. The petiole of *P. trichocarpa* is reddish and the buds are conical, long, narrow and sticky, with a strong scent when they open. It also has a deep widespread root system.

Populus trichocarpa has huge economics importance where it can be used for timber, veneer and fibre products. The living trees also provide windbreak due to its tallness. *Populus trichocarpa* can reach reproductive maturity in as few as 4 to 6 years

under appropriate conditions, permitting selective breeding for large-scale sustainable plantation forestry (Tuskan *et al*., 2006).

P. trichocarpa is a good model species for trees and has been extensively studied. The tree was selected as the first model of forest species for genome sequencing because of its modest genome size, rapid growth, easy to be manipulated in experiment, and its range of available genetic tools (Tuskan *et al*., 2006). The genome of the tree was published in 2006. The genome of *P. trichocarpa* has been subjected to sequencing for eight times and its size is approximately 403 Mb large with an arrangement into 19 chromosomes (Tuskan *et al*., 2006). According to Tuskan *et al*., the poplar genome is about four times larger than the small flowering plant *A. thaliana* but is 50 times smaller compare to pine genome. Topics of studies involved *P. trichocarpa* include lignin biosynthesis, wood formation, engineer the amount and structure of lignin in plants and etc.

2.1.2 *Pinus taeda*

Loblolly pine is the common name of *Pinus taeda* which is originated from southeast parts of United States (Vidakovic, 1991). According to Vidakovic, this gymnosperm tree can grow up to 46m high and 1,6m in diameter. The barks of young trees are gray or yellowish in colour which later grow thick into blackish-gray to reddish-brown. The tree's needles are in fascicles of three, thin, stiff, 12-23 cm long, light green, and able to persisting on the tree for three years. The trees are normally flowering between February to April. The cones produced by the trees are sessile, 2-5 together, ovate, symmetrical, 6-12cm long, and reddish colour. *P. taeda* starts to produced seed at the age of 3 to 4 but it's sexually maturity is reached in its 10^{th} year. It is estimate that there are approximately 30-40 seeds are located in a cone where the seed is brown-reddish colour. *P. taeda* has 24 chromosomes.

2.1.3 *Arabidopsis thaliana*

Arabidopsis thaliana is a small flowering plant which is a member in the Brassicaceae family. This dicotyledonous plant is native to Europe, Asia, and northwest Africa. *A. thaliana* also commonly known as thale cress, and mouse- eared cress. According to Nottingham *Arabidopsis* Stock Centre (n.d.), *A.thaliana* can be found in open free draining ground such as poor sandy.

A. thaliana does not have economic value like other woody species but it is a good plant model for genetic, biochemical and physiological studies. This typical flowering plant is popular as it has simple genome, rapid life cycle (about six weeks from germination to mature seed), prolific seed production, the availability of numerous mutations and great adaptability (Goodman *et al*., 1995). Due to its favourable characteristics for researching, *A. thaliana* was chosen for sequencing and it becomes the first genome sequenced plant in 2000. *A. thaliana* has a total of five chromosomes and its genome contains 25,498 genes encoding proteins from 11,000 families (The Arabidopsis Initiative, 2000). However, there are only 34 genes were indentified in arabidopsis genome which are involved in the lignin biosynthesis pathway (Raes *et al*. 2003)

2.1.4 *Eucalyptus globulus*

Eucalyptus globulus belongs to Myrtaceae family, *Eucalyptus* genus is an evergreen tree that originated from Australia. The common names of *E. globulus* include Tasmanian bluegum. *E. globulus* is one of the best known eucalypt trees and has competitive advantage compare to other species as the young foliage of *E. globulus* is seldomly browsed by cattle or sheep (Food and Agriculture Organization of the United Nations, 1979).

Most of the eucalypt plantation in the world are managed as short rotation coppice crop around 5 to 15 years to provide roundwood at low cost (Eldridge *et al.,* 1994). These fast growing trees with normal height from 30-55m are harvested before the formation of much heartwood which affects pulping process (Eldridge *et al.,* 1994). According to Eldridge *et al.* (1994) wood from young eucalypt plantations can be used as firewood, charcoal, poles and posts, mine timbers, pulpwood and reconstituted wood. Only minority of the eucalyptus have the chance to growth as large diameter of the tree bark is needed for sawn timber.

Besides it commercial value as energy source, eucalypt leaves can be used to produce essential oil. Eucalyptus oils are clear liquids with aroma characteristic have medical uses and are well known to the aborigines in Australia thousands years ago (Coppen, 2002). However, only *E. globulus* along with other five eucalypt species have been exploited commercially for eucalyptus oil production (Coppen, 2002).

2.2 Lignin Biosynthesis Pathway

Lignin is a three dimensional polymer of phenylpropanoid alcohols that can be found in all vascular plants (Boerjan *et al*., 2003) especially in woody plant vascular tissue (Xu *et al*., 2009). Lignin formation in plant cells involves a pathway known as lignin biosynthesis pathway. The pathway is consists of ten genes namely cinnamate-4-hyhroxylase (*C4H*), para-coumarate-3- hydroloase (*C3H*), caffeic/5-hydroxyferulic acid O-methyltransferase (*COMT*), ferulate 5-hydrolase (*F5H*), caffeoyl-CoA-methyltransferase (*CCoAOMT*), cinnamyl alcohol dehydrogenase (*CAD*), 4-coumarate CoA ligase (*4CL)*, hydroxycinamoyl transferase (*HCT*), phenylalanine ammonia-lyase (*PAL*) and cinnamoyl CoA reductase (*CCR*) which are responsible for the lignin monolignols formation.

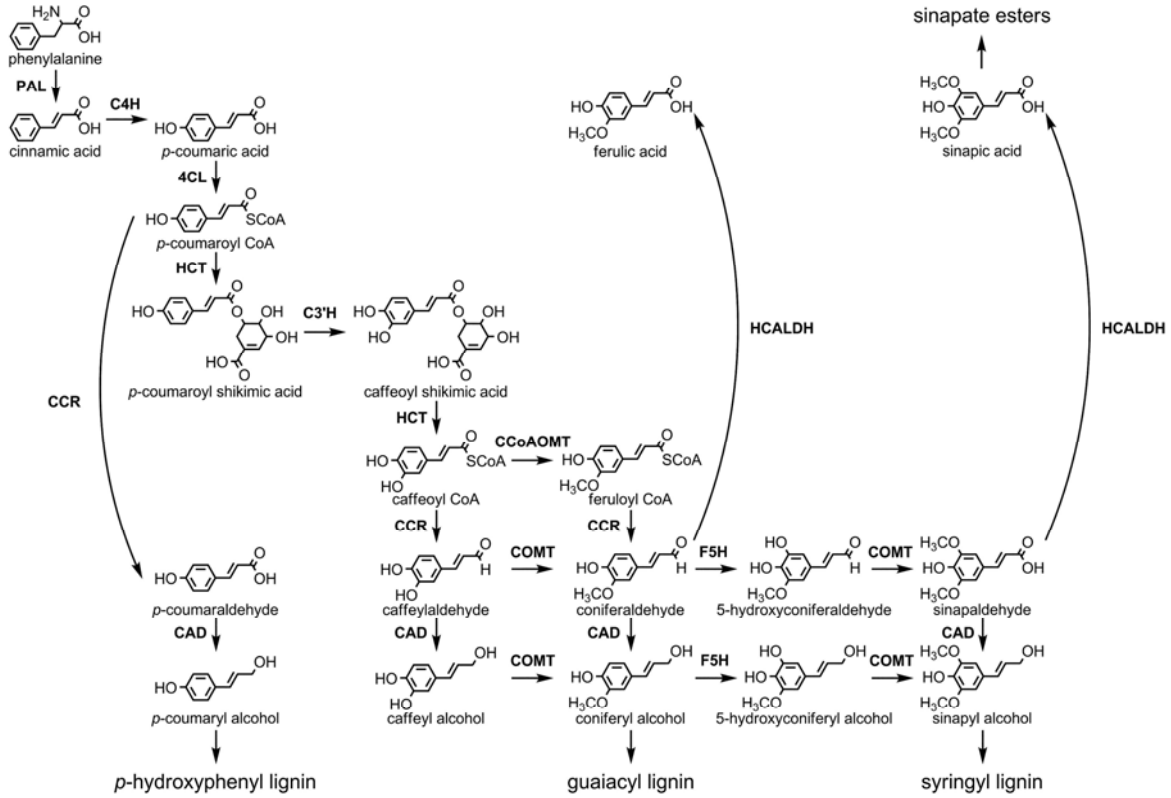

Figure 2.1: Lignin biosynthesis pathway. (Adapted from Jing *et al.*, 2008).

The monolignol biosynthesis pathway involves many intermediates and enzymes. In addition, there are several multifunctional enzymes are involved in the biosynthesis pathway and these enzyme also correspond to diverse gene families (Xu *et al*., 2009). Example of multifunctional enzyme is F5H enzyme which can acts on several substrates such as ferulic acid, coniferaldehyde and coniferyl alcohol (Xu *et al*., 2009).

According to Boerjan *et al*. 2003, lignin is produce from oxidative polymerization of mainly three hydroxycinnamyl alcohol monomers, para-coumaryl (H), coniferyl (G) and sinapyl (S) alcohol. The three lignin monomers differ from each other by the degree of methoxylation of the aromatic ring where there is an increase of methoxylation from H units to G units follow by S units (Boerjan *et al*., 2003). The presence of three different kinds of monolignols leads to lignin content variability in plants in the terms of monolignols ratio. For example, the lignin composition between angiosperms and gymnosperms are slightly different as angiosperms mainly contain guaiacyl (G) units and syringyl (S) units while gymnosperms contain G and para-hydroxyhenyl (H) units. The increase in the degree of methoxylation also corresponds to the decrease of reactive site on the aromatic ring which reduces the number of potential cross-links between monomers during polymerization (Whettena and Sederoff, 1995). Due to this reason angiosperm becomes a better choice for paper production compare to gymnosperm.

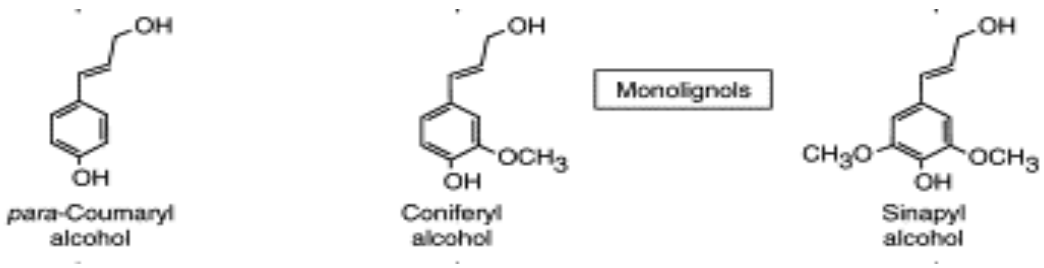

Figure 2.2: Monolignols of lignin (Whettena and Sederoff ,1995)

2.3 Motif and Domain

Proteins are comprised of recognizable smaller sequence domains which recur in other proteins in various combinations (Pearl *et al*., 2007). Proteins can contain more than one domain which can be linked by single or multiple connections (Pearl *et al*., 2007). Domain usually consists of secondary and supersecondary structures and have and average size of 150 ± 50 residues (Pearl *et al*., 2007). Protein domain also shows a certain level of evolutionary conservation (Koonin and Galperin, 2003).

Sequence motifs are usually short and fixed length sequence pattern which it can be either DNA, RNA or protein molecules (Bailey, 2008). Protein motif is known as supersecondary structure is a combination of several secondary structural elements which are produced by the folding of adjacent section of the polypeptide chain into a specific three dimensional configuration (Mount, 2005). Protein motifs can represent enzymes active site, and involved in the structure and stability determination of protein (Bailey, 2008).

There are two general approaches use to representing the consensus information of motif and domain, consensus pattern approach and statistical based approach (Xiong, 2005). The consensus pattern is also known as regular expression which it is a brief and simple way for representing sequences by a string of letters (Xiong, 2005). The pattern also represents the common characteristic of a protein family or multiple sequence alignment (Baxevanis and Ouellette, 2005). Example of a consensus pattern can be written as [LIVMFY] - {E} - G - x - [KR] which is consists of five amino acid residues. The first residue started with either L, I,V, M, F,Y followed by any residues but E, followed by G, an unspecific amino acid residue (X) and finally either K or R residue. Residues that are placed within the bracket indicate the position has multiple alternative conserved residues

while residues that are placed in the curly bracket shown the exclusion of the particular residues (Xiong, 2005).

Statistical based approach used profile to include possibility information derived from multiple sequence alignment. Profile is a numerical representation of multiple sequence alignment with intrinsic information about the common characteristic of particular sequence (Baxevanis and Ouellette, 2005). Example of profile is position weight matrix (PWM). PWM defines the probability of each letter at a specific position and assumes that each position in the motif is independently of the other (Bailey, 2008).

2.4 Tree Building Method

Phylogenetic tree construction is used to study the evolutionary changes of organism genetic materials and make implicit assumptions about the sequence data. A scoring function is required to discriminate between trees by quantifying how well a phylogenetic tree describes the sequences (Whelan, 2008). Up to date, there are many and different scoring functions have been proposed and it can be briefly summarizes as distance based and parsimony based (Baxevanis and Ouellette, 2005).

2.4.1 Distance-based Methods

Distance-based methods use the amount of dissimilarity or the distance between the aligned sequences to derive trees (Baxevanis and Ouellette, 2005). Examples of distance-based method are unweighted pair group method with arithmetic mean (UPGMA), neighbor joining (NJ), and minimum evolution (ME).

Unweighted pair group method with arithmetic mean (UPGMA) is a popular and simple method in building phylogenetic trees. There are two important assumptions in UPGMA method for trees building where it assumes that evolution occurs at the same rate

on all tree branches, and distance between two leaves is the total length of edges that connecting them (Westhead *et al*., 2002). However, these assumptions can lead to create of incorrect trees as not all organisms evolve at the same rate.

Neighbor joining is the most commonly applied distance-based method and it uses evolutionary distance data to reconstruct phylogenetic trees. Different from UPGMA, neighbor joining disagree with the assumption of constant evolutionary rate and the principle of neighbor joining is to find operational taxonomic units (OTU) or neighbors that minimize the total branch length. The final tree with branch length is made by star decomposition where the most similar terminals are joined and a branch will be inserted between them (Baxevanis and Ouellette, 2005).

2.4.2 Character-based method

Character- based method uses character data in all the steps in the analysis which allows assessing the reliability of each base position in an alignment (Baxevanis and Ouellette, 2005). Examples of character-based method are maximum parsimony (MP) and maximum likelihood (ML).

Maximum parsimony infers a phylogenetic tree by reducing the total number of evolutionary steps required to explain a given set of data (Baxevanis and Ouellette, 2005). On other word, trees are constructed base on a basis of minimum number of mutation for a given data set (Westhead *et al*., 2002). Maximum parsimony method identifies and analyses of the minimum number of substitutions required from converting one amino acid to other in an aligned sequence. The final tree is generated by grouping sequences that can be interconnected with the smallest of overall changes (Westhead *et al*., 2002).

Maximum likelihood analysis the changes of each position of the sequence and hence produce the most reliable trees. This is due to the reason that maximum likelihood

allows users to incorporate as expected model of sequence changes, which weights the probability of mutation (Westhead *et al.,* 2002). Thus, maximum likelihood is time consuming compare to other methods.

2.5 XML

XML or eXtensible Markup Language is introduced in 1996 by the World Wide Web Consortium (W3C) with the aim to for structuring documents (Achard *et al*., 2001). XML is derived from the Standard Generalized Markup Language (SGML) which is the international standard for defining structure and content description of different types of electronic documents (Achard *et al*., 2001). Different from SGML, XML is built on the strength of SGML with less complexity compare to SGML (Cerami, 2005).

XML is a tool that focuses on document semantics where users can identify a specific part in a document easily without concern about the specific presentation and layout of the document (Cerami, 2005). Thus XML can be used to create highly structured document for data storage (Holzner, 2009). Besides that, XML also allow its users to create elements. XML documents consist of elements which are the textual data structured by tags and each element has a start or end pair. The data between the two tags is XML element (Achard *et al*., 2001).Thus, XML is often used as data exchange language to facilitate the exchange of scientific information due to its simplicity (Achard et al., 2001).

XML compulsory users to strictly fulfil requirements in order to create a well formed xml document. It is important to follow the criteria to avoid xml parsing error. The requirements include every start tag must have corresponding end tag except empty element tag syntax, elements must be nested, all attribute values must have quotes, XML document must has one root element and reserved characters are defined as markup (Cerami, 2005).

CHAPTER 3

MATERIALS AND METHODS

3.1 Data Mining

Lignin biosynthesis pathway genes sequences in the form of nucleotide and protein with other sequences details such as accession number, molecular weight were retrieved from online public database like NCBI databases (URL: http://www.ncbi.nlm.nih.gov/), and TIGR Arabidopsis thaliana Database (URL:http://www.tigr.org/tdb/e2k1/ath1).

3.2 XML Database

The downloaded data from gene bank was presented in an organized form using XML (eXtensible Markup Language) document. Each lignin biosynthesis pathway gene's XML document contains information about organism name, sequence accession number, both protein and nucleotide sequences, and the sequence length of protein and nucleotide.

3.3 Multiple Sequence Alignment

Three to four protein sequences were randomly chosen from each tree species gene for performing multiple sequence alignment. The protein sequences were aligned using ClustalW programme in Molecular Evolutionary Genetics Analysis (MEGA) software. The aligned protein sequences were subsequent edited manually. The ClustalW parameters in alignment of protein were set to use default parameters in pairwise alignment but gap opening penalty of 3, gap extension penalty of 1.8 were used in multiple alignments. The protein weight matrix used in alignment is Gonnet.

3.4 Motif and Domain Identification

Domain for each lignin genes were scanned with Pfam database using the *Pfam Domain Search* tool which is available in CLC Protein Workbench 5.3 software. The gene domains were identified by matching query sequences with Pfam database and the result were listed out with the availability of E value and Score for reliability evaluation. Motif identification of lignin genes were performed by matching lignin biosynthesis pathway genes in the PROSITE database (URL: http://expasy.org/prosite/)

3.5 Phylogenetic analysis

Phylogenetic analysis of each lignin genes were performed on aligned protein sequences using two different approaches: Neighbor Joining (NJ) and Maximum Parsimony (MP) which implemented in Molecular Evolutionary Genetics Analysis (MEGA) software. Bootstrap tests were conducted using 1000 replicates. Tables of comparison between pairwise sequences were also created to show the differences between two sequences being compared.

CHAPTER 4

RESULTS & DISCUSSION

Many genes are distributed in the genome as multiple copies rather than existing as individual copies producing a single protein (Page and Holmes, 1998). Normally, these genes are derived from evolution events such as mutation, gene duplication, translocation and etc (Russell, 2006). Polyploidy which is one of the events in gene duplication and it is recognised as the common evolution in plants especially in angiosperm which it is estimated that 70% of angiosperm are polyploidy (Gottlieb, 2003). The occurrence polyploidy is more readily in plant species compare to other organism as most of the time plants are reproduce through vegetative propagation (Page and Holmes, 1998).

The divergence of the duplicated ancestral genes leads to new functional and structural specialisation to the new genes (Tavares *et al*., 2000). In the case of lignin biosynthesis gene, most of the genes are derived from multigenes family and an example is the *CAD* multigenes family in *Arabidopsis thaliana* which is consisted of nine members (Kim *et al*., 2004). Among the nine CAD genes only six show catalytically competent for NADPH-dependent reduction of p-coumaryl, caffeyl, coniferyl, 5-hydroxyconiferyl, and sinapyl aldehydes, whereas the other three displayed very low activity even at very high substrate concentrations (Kim *et al*., 2004). Hence, proteins belonging to the same family have the same biochemical function, but not necessarily have the same biological role (Tavares *et al*., 2000).

4.1 Data Mining

Data mining in the NCBI and TIGR public databases yield large amount of nucleotide and protein sequences especially for *Arabidopsis thaliana* and *Populus trichocarpa*. Thus, three to four sequences of *A. thaliana* and *P. trichocarpa* were randomly selected to be included in the study. Table 4.1 below showed the summary of selected sequences from bioinformatics analysis.

Table 4.1: Summary of sequences selected for bioinformatics analysis.

gene	Arabidopsis				Poplar				Eucalyptus				Pine			
	accession no.	cds	aa	c/p	accession no.	cds	aa	c/p	accesion no.	cds	aa	c/p	accesion no.	cds	aa	c/p
4CL	AT4G19010.1	1701	566	c	ACC63867.1	1623	540	c	BAF93472.2	1635	544	c	AAA92669	1614	537	c
~60000	AT5G63380.1	1689	562	c	ACC63868.1	1632	543	c	BAI47543.1	1635	544	c	AAA92668	1614	537	c
	AT5G38120.1	1641	546	c	ACC63869.1	1677	558	c								
CCR	AT1G15950.2	1014	337	c	CAA12276.1	1017	338	c	AAM34502.1	1011	336	c	AAL47684	975	324	c
~36000	AT2G33590.1	966	321	c	ACC63879.1	1017	338	c	AAT74876.1	1011	336	c				
	AT5G58490.1	975	324	c	XP_002298395.1	978	325	c	AAT74877.1	1011	336	c				
PAL	AT2G37040.1	2178	725	c	ACC63887.1	2148	715	c					AAA84889	2265	754	c
~7700	AT3G53260.1	2154	717	c	ACC63889.1	2136	711	c								
	AT3G10340.1	2124	707	c	XP_002322884.1	2148	715	c								
F5H	AT5G04330.1	1539	512	c	ACC63881.1	1542	513	c	BAI47544.1	1590	529	c				
~5800	AT4G36220.1	1563	520	c	CAB65335.1	1542	513	c	BAI47545.1	1590	529	c				
					ACC63880.1	1545	514	c								
CCoAMT	AT4G26220.1	699	232	c	ACC63876.1	744	247	c	AAD50443.1	744	247	c	AAD02050.1	780	259	c
~26000	AT1G67990.1	702	233	c	ACC63877.1	708	235	c	AAC26191.1	741	246	c				
	AT4G34050.2	447	148	c	ACC63878.1	744	247	c								
COMT	AT1G21100.1	1122	373	c	ACC63884.1	1095	364	c	AAD50439.1	938	313	p	AAC49708	1146	381	c
~41000	AT5G54160.1	1092	363	c	ACC63885.1	1056	351	c	AAD50440.1	938	312	p				
	AT1G77530.1	1146	381	c	ACJ76442.1	1095	364	c								
C4H	AT2G30490.1	1518	505	c	ACC63871.1	1518	505	c					AAD23378.1	1521	506	c
~57800					ACC63872.1	1614	537	c								
					ACC63873.1	1518	505	c								
					XP_002325638.1	1518	505	c								
C3H	AT2G40890.1	1527	508	c	ACC63870.1	1527	508	c					AAL47685	1539	512	c
~57900	AT1G74540.1	1494	497	c	XP_002323218.1	1530	509	c								
	AT1G74550.1	1464	487	c	XP_002336101.1	1530	509	c								
					XP_002336354.1	1530	509	c								
HCT	AT5G48930.1	1302	433	c	ACC63882.1	1302	433	c								
~4800					XP_002303858.1	1374	457	c								
					XP_002324534.1	1293	430	c								
					XP_002334783.1	1338	445	c								
CAD	AT1G72680.1	1068	355	c	XP_002322761.1	1089	362	c	AAC07987.1	1071	356	c	CAA86072	1074	357	c
~39000	AT2G21730.1	1131	376	c	XP_002313293.1	1083	360	c					CAA86073	1074	357	c
	AT4G34230.1	1074	357	c	XP_002300211.1	1089	362	c								

Note: The sequence length details of cds and amino acid (aa) were included, and the column (c/p) which indicate the corresponding sequence is either complete (c) or partially complete (p).