African Journal of Biotechnology Vol. 12(1), pp. 8-13, 2 January, 2013
Available online at http://www.academicjournals.org/AJB
DOI: 10.5897/AJB11.3692
ISSN 1684–5315 ©2013 Academic Journals

Full Length Research Paper

Computational identification of putative cytochrome P450 genes in soybean (*Glycine max*) using expressed sequence tags (ESTs)

Muhammad Azam Chattha, Jianyu Liu, Fang Huang, Lingyong Li and Deyue Yu*

National Center for Soybean Improvement, National Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing 210095, China.

Accepted 10 May, 2012

Cytochrome P450 is a group of monooxygenase that exists as a gene superfamily and plays an important role in metabolizing physiologically important compounds in plants. However, to date only a limited number of P450s have been identified and characterized in soybean (*Glycine max*.). In this work, a computational study of expressed sequence tags (ESTs) of soybean was performed by data mining methods and bio-informatics tools and as a result 78 putative P450 genes were identified, including 57 new ones. These genes were classified into five clans and 20 families by sequence similarities and among those 57 new families, 18 new subfamilies were found which have not been observed previously in soybean. This work may provide a basis for further functional dissection of P450 genes in soybean and other legumes.

Key words: Expressed sequence tags (ESTs), *in silico*, soybean (*Glycine max*.), P450.

INTRODUCTION

Cytochrome P450 monooxygenase (P450s) is a complex gene superfamily of proteins which metabolize physiologically important compounds in organisms ranging from protists to plants and to humans (Nelson et al., 1996). P450s represents the largest family of plant protein identified to date (Morant et al., 2003). P450s are found in all major plant biosynthetic pathways, including those UV protectants (flavonoids, coumarins, sinapoly-esters), pigments (anthocyamins), defense compounds (isoflavonoids, phytoalexins, hydroxamic acids), fatty acid hormones (gibberellins, brassinosteroids), signaling molecules (salicylic acid, jasmonic acid and etc.), accessory pigments (carotenoids) and structural polymers (lignins).

Plant P450s are also responsible for catabolizing some endogenous signaling molecules as well as enhancing exogenous compounds in the environment (Chaudry et al., 2002; Harvey et al., 2002). Due to the usefulness of oxygen in building complex molecules, cytochrome P450 enzymes are abundant in these pathways and comprise approximately 1% of the genes in plant genomes such as *Arabidopsis*, rice and poplar (Nelson, 2006). Complexities existing in the biochemical reactions and genomes of the more diverse phyla, which includes 230,000 named species in angiosperms (Margulis et al., 1998), that there is still much to be discovered, particularly in the prolife-ration of P450-mediated reactions already characterized in plants (Kahn and Durst, 2000; Werck-Reichhart et al., 2002).

P450s are widespread in the plant kingdom and con-stitute a gene superfamily. There are two main classes of P450s containing 10 clans and 62 families in plants, however, only clans 71, 85, 86, 74 and 97, containing a total of 22 families, have been identified in *Fabales* (Nelson et al., 2004). The Fabacea (legumes) are the third largest plant family of flowering plants, comprising more than 650 genera and 18,000 species. Economically, legumes represent the second most important family of crop plants after *Poacea* (grass family), accounting for

*Corresponding author. E-mail: dyyu@njau.edu.cn. Tel/Fax: 86 25 84396410.

around 27% of world crop production (Graham and Vance, 2003).

Soybean (*Glycine max*) is one of the most economically important species in legumes (VandenBosch and Stacey, 2003). The large scale of soybean expressed sequence tags (ESTs) database (NCBI) contains over 1,386,618 sequence entries from 73 non-normalized complementary deoxyribonucleic acid (cDNA) libraries, representing various vegetative and reproduction organs (20, February 2008). These resources are very helpful to get in acquiring a greater knowledge of genomes and their roles. Now these databases are offering an opportunity to identify previously uncharacterized genes, and to assess the frequency and tissue specificity of their expression *in silico*.

The genome of the plant kingdom is entering from genome sequencing into the post-genomic area. There are still large quantities of genes to be annotated. This study used ESTs which is a good approach for computational work. *In silico* resources and bio-informatics tools were applied to detect, identify and annotate putatively functional P450 encoding sequences in soybean. The deduced amino acid sequences which are based on phylogenetic analysis have allowed identification of paralogous genes and clusters of orthologous groups, allowing further characterization of P450 genes with both known and unknown functions.

MATERIALS AND METHODS

Collection of putative P450 sequences from *G. max*

The National Center for Bio-technology Information (NCBI) database (http://www.ncbi.nlm.nih.gov/) was used to retrieve soybean nucleotide and 70 nucleotides were discovered. The strategy used to discover the soybean P450 family at NCBI is as following: Each nucleotide was retrieved at NCBI with the tblastn option. Unregistered accession numbers of ESTs were BLASTed using blastn with nt/rt option. The ESTs with 'no significant similarity found' were BLASTed using balstn with ESTs option. These ESTs were BLASTed at (http://www.phytozome.net) using soybean as target genome for organism genome. Genomic regions were selected with E-value less than 0.5, 5 to 10 Kb of genome sequence were obtained for each accession number by using download sequence file in reports and analysis with soybean in data source. These genomic sequences were researched at (http://www.softberry.com) by online software, using Gene finding in eukaryotes, FGENESH HMM based gene structure predication. The sequences were pasted and *Medicago* (legume plant) selected to obtain ORF and corresponding protein sequences. The protein sequences were then BLASTed at NCBI using pblast to detect P450 conserved domains. The resulting 78 new P450 protein sequences were discovered. A total of 133 protein sequences of soybean were collected with 55 already known and termed as P450.

Phylogenetic analysis of putative P450s

Predicted P450 protein sequences from soybean and representative members of known P450 families from other plants were used for alignment and phylogenetic analysis. The alignment of multiple sequence of P450 proteins was performed using the CLUSTALX program (version 1.81) (Thompson et al., 1997).The phylogenetic analysis was carried out by the neighbor-joining (N-J) method (Saitou and Nei, 1987) and a neighbor-joining tree was constructed using protein in CLUSTALX. The significance level of the neighbor-joining analysis was examined by bootstrap testing with 1000 repeats. The tree was represented using N-J plot in the CLUSTALX program.

RESULTS

Identification of putative P450 genes in soybean

We identified 78 putative P450 genes in soybean. All these putative P450s genes presented here were based on sequence similarity searches. Query proteins were from representative members of each plant P450 family, but functional identification in each family member was preferred. The sequence similarity based search was refined by multiple alignments and searched for conserved domains as mentioned in methods part. All the sequences analyzed possessed the structures typical of P450 family. Thus, we annotated 57 P450 genes according to the standardized system of P450 nomenclature (Nelson et al., 1996). All these P450 sequences were distributed among 20 P450 families and five clans (clans 71, 72, 74, 85 and clan 86) which suggested P450 genes existed as a superfamily in soybean.

Approximately 73% (57 out of 78) of the putative P450s identified were annotated as P450, P450-like, or related to P450 without any indication of similarity to certain families and were given names related to P450s (Table 1). Therefore, they could not have been related to a certain families solely based on the description given by the database. The annotation of these "anonymous" P450s was improved in this study by identifying their similarities as characterized P450s that were not available at the time of annotation of sequences. ESTs were performed by similarity comparisons to previously identified genes and were improved in this study.

Phylogenetic analysis of predicated P450 families

A N-J phylogenetic tree for identified and predicated sequences from soybean and representative members of P450 families was constructed using MEGA4.1 (Kumar et al., 2008) (Figure 1). It can be seen from Figure 1, that plant P450s genes were classified into two branches, A-type and non-A-type. All the soybean putative P450 sequences were clustered into their corresponding branches. The proteins encoded by the A-type genes from a single clan containing 75% soybean sequences (43 out of 57); represent many of plant-specific enzymes functioning in the synthesis of secondary products (phenlypropanoids, glucosinolates, isoprenoids and

Table 1. Predicated and annotated P450s.

Clan	Family	Name*	EST	Previous annotation
71	71	CYP71A33	EV271564	CYP71A
		CYP71AU9	CF922481	CYP71A
		CYP71D105	CX701446	CYP71D
		CYP71D99	BE822612	CYP71D
		CYP71D101	BU547920	CYP71D
		CYP71D103	BE346101	CYP71D
		CYP71D100	BE609824	CYP71D
		CYP71D104	BE822684	CYP71D
		CYP71D102	BI969861	CYP71D
	75	CYP75B40	CO981271	P450
	76	CYP76E4	AW308831	CYP76E
		CYP76E5	CO983935	CYP76E
		CYP76F17	CO985671	CYP76F
	78	CYP78A43	BG881893	CYP78A
		CYP78A44	BQ612617	CYP78A
	81	CYP81EV4	BQ079433	CYP81E
		CYP81E18	BQ080894	CYP81E
		CYP81E19	BI321275	CYP81E
		CYP81E24	BI423886	CYP81E
		CYP81E17	BU766794	CYP81E
		CYP81E11	CF806405	CYP81E
		CYP81E22	CA937281	CYP81E
	82	CYP82A3	AW424163	CYP82A
		CYP82D25	AW279047	CYP82D
		CYP82D26	AW704257	CYP82D
		CYP82D27	CF808171	CYP82D
		CYP82D28	AW734404	CYP82D
		CYP82A18	BM527722	CYP82A
	83	CYP83E13	BE612106	CYP83E
		CYP83E14	EH224922	CYP83E
		CYP83E15	BQ628050	CYP83E
		CYP83G3	BE823818	CYP83G
	89	CYP89A43	DW247110	CYP89A
	93	CYP93A19	BU080998	CYP93A
	706	CYP706A10	EH259119	P450
		CYP706K1	BI971415	P450
	736	CYP736A28	BM178003	P450
		CYP736A29	BU577301	P450
		CYP736A30	EV264217	P450
		CYP736A31	CF806276	P450
		CYP736A32	CX704908	P450
		CYP736A33	EV278728	P450
		CYP736A34	BQ452803	P450

alkaloids, etc.). The proteins encoded by the non-A-type genes represent multi-kingdom enzymes functioning in the synthesis of more general compounds (sterol, oxygenated fatty acids, etc.) and in the synthesis of hormones

Table 1. Continued.

Clan	Family	Name*	EST	Previous annotation
72	72	CYP72A120	EH223296	CYP72A
		CYP72A121	EH223622	CYP72A
	714	CYP714A9	BI973557	P450
	734	CYP734A17	CO983699	P450
74	74	CYP74C16X	CA800304	CYP74C
85	90	CYP90A20	BU081866	CYP90A
	707	CYP707A16	EV277045	CYP707A
		CYP707A45	AW733509	CYP707A
	716	CYP716G1	BI785161	#P450
	722	CYP722A1	BM523065	P450
86	94	CYP94C18	BE803693	CYP94C
		CYP94C19	BM094927	CYP94C
		CYP94C20	CK769142	CYP94C
		CYP94D24V2	AI794818	CYP94D

and other molecules.

DISCUSSION

P450 genes exist as a superfamily in soybean

In this study, 11 families of clan 71 were discovered in soybean in which eight new subfamilies and three new families of clan 71, not reported previously (family 75, 706, 736) were designated as CYP75B40, CYP706A10, CYP706K1, CYP736A28, CYP736A29, CYP736A30, CYP736A31, CYP736A32, CYP736A32, CYP736A33 and CYP736A34. Constantly in the angiosperms, clan 71 is by far the biggest, with one third of the plant P450s in any genome. CYP71 clan sequences are noted in other plant genomes and seem to have begun early in plant lands (Durst and Nelson, 1995). Furthermore, four clans were identified in this study, namely clans 72, 74, 85 and 86. In clan 72, three families had not been explored previously, one was found to have new subfamily and two new families in soybean. Two new families annotated as CYP714A9 and CYP734A17. Clan CYP72 inactive brassinolide particularly found in CYP734A subfamily (Turk et al., 2005). During the identification, only one new subfamily was found in clan 74 for soybean, which was not reported previously and was named CYP74C16X which may be an incomplete sequence or pseudo gene. It has been noted that the CYP74 family is important in modifying unsaturated fatty acid hydroperoxides derived from linolenic acid or α-linolenic acids and includes oxide synthases, hydroperoxide lyases and divinely synthesis (Nelson, 2006).

In clan 85, four families were discovered in which two are new subfamilies and the other two new families in soybean which were not previously reported. The annotations of the two new families are CYP716G1 and CYP722A1. The CYP85 clan has several functions, including synthesis of kaurenoic acid oxidase (Winkler and Helentjaris, 1995; Helliwell et al., 2000). During the investigation of clan 86, four new subfamilies of family CYP94 (named CYP94C18, CYP94C19, CYP94C20 and CYP94D24V2) were found which were not previously explored in soybean. The CYP86 clan had three families CYP86, CYP94 and CYP704. The functions of these families appear to have been established very early in plant evolution with land plants needing to protect themselves against water loss (Nelson, 2006). Further studies using multiple approaches like cloning, single nucleotide polymorphisms (SNPs) and functional association will be required to resolve function and divergence in this gene superfamily and there are still big knowledge gaps surrounding plant's P450s, in-spite of the model plants *Arabidopsis* and rice, as these can hardly be representative of their 170,000 dicot and 65,000 monocot relatives.

Phylogenetic analysis revealed new putative P450 families in soybean

This study found seven new families and 18 new subfamilies not previously explored in soybean (Table 1 and Figure 1). The five clans (clans 71, 72, 74, 85 and 86)

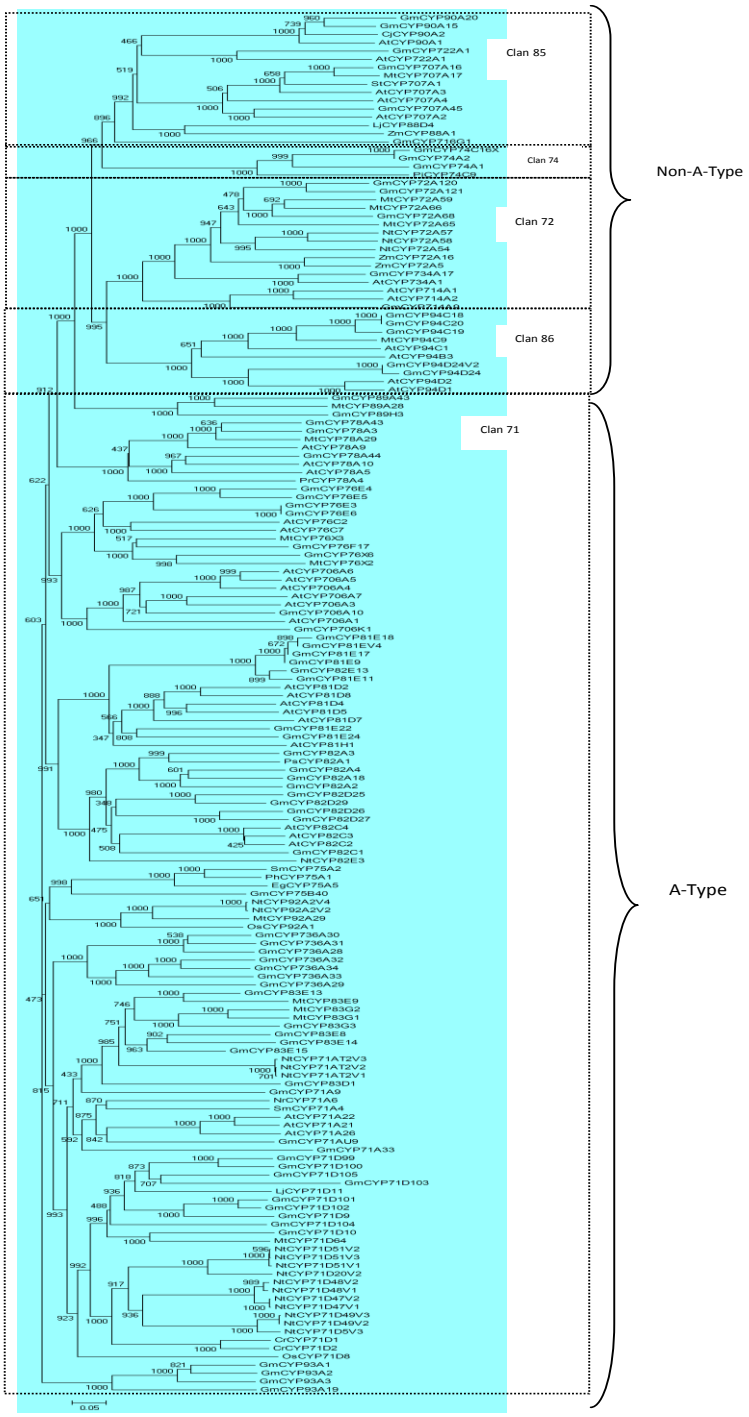

Figure 1. Phylogenetic analysis of predicated and annotated P450s. Phylogenetic tree of the collected *G. max* P450s and the representative members of P450 families. The unrooted phylogentic tree of P450s was depicted by the CLUSTAL X (version 1.81) program and the neighbor-joining (N-J) method. A N-J tree was constructed using MEGA 4.1; the significance level of the N-J analysis was examined by bootstrap testing with 1,000 repeats. The numbers beside the branches represent bootstrap values based on 1,000 replications. At, *Arabidopsis thaliana*; Cj, *Camellia japonica*; Cr, *Catharanthus roseus*; Eg, *Eustoma grandiflorm*; Lj, *Lotus japonica*; Mt,*Medicago truncatula*; Nt, *Nicotiana tabacum*; Os, *Oryza sativa*; Ph, *Petunia hybrid*; Pi, *Petunia inflata*; Pr, *Pinus radiata*; Ps, *Pisum sativa*; Sm,*Solanum melongena*; St, *Solanum tubersome*; Zm, *Zea mays*.

can be found for the dicot plant clans with corresponding members in soybean. With the exception of the CYP71 clan, the other four clans seems involved in conserved functions that relate to sterol and isoprenoid biosynthesis (clan 85), fatty acid metabolism (clan 86), biosynthesis of oxylipids (some subfamilies of clan 74) and plant hormone homeostasis (clan 72) (Werck-Reichhart et al., 2002). Many of the families of clan 71 and some particular subfamilies appear to be species specific and represent the success in recruiting P450s for evolutionary novel (Li et al., 2007). With the completion of the soybean genome project, there is a speculation that more P450s will be discovered. This study may provide a basis for further functional dissection of P450 in soybean and other legumes.

ACKNOWLEDGEMENT

This work was supported in part by National Natural Science Foundation of China (31000718, 31171573)

REFERENCES

Chaudry Q, Scroder P, Werck-Reichhart D, Grajek W, Marecik R (2002). Prospects and limitations of phytoremediation for the removal of persistant pesticides in the environment. Environ Sci. Pollut. Res. Int., 9: 4-17.

Durst F, Nelson DR (1995). Diversity and evolution of plant P450 and P450-reductases. Drug Metab. Drug Interact., 12: 189-206.

Graham PH, Vance CP (2003). Legumes: importance and constraints to greater use. Plant Physiol., 131:872-877.

Harvey PJ, Campanella BF, Castro PM, Harms H, Lichtofous E (2002). Phytoremediation of polyaromatic hydrocarbons, anilines and phenols. Environ. Sci. Pollut. Res. Int., 9: 29-47.

Helliwell CA, Chandler PM, Poole A, Dennis ES, Peacock WJ (2001). The CYP88A cytochrome P450, ent-kaurenoic acid oxidase, catalyzes three steps of the gibberellin biosynthesis pathway. Proc. Natl. Acad. Sci. USA., 98: 2065-2070.

Kahn R, Durst F (2000). Function and evolution of plant cytochrome P450. Recent Adv. Phytochem., 34: 151-189.

Kumar S, Dudley J, Nei M, Tumara K (2008). A biologist-centric software for evolutionary analysis of DNA and protein sequences. Briefing Bioinforma. 9:299-306.

Li L, Cheng H, Gai J, Yu D (2007). Genome wide identification and characterization of putative cytochrome P450 genes in the model legume *Medicago truncatula*. Planta 226:109-123.

Margulis L, Schwartz KV (1998). Five Kingdoms. An illustrated Guide to the Phyla of Life on Earth Ed 3. W.H. Freeman, New York.

Morant M, Bak S, Moller BL, Werck-Reichhart D (2003). Plant cytochromes P450: tools for pharmacology, plant protection and phytoremediation. Curr. Opin. Biotechnol. 14:151-162.

Nelson DR (2006). Plant cytochrome P450s from moss to poplar. Phytochem. Rev. 5:193-204

Nelson DR, Koymans L, Stegeman JJ, Feyereisen R, Waxman DJ , Waterman MR, Gotoh O, Coon MJ , Esatabrook RW, Gunsalus IC, Nebert DW (1996). P450 superfamily: up-date on new sequences, gene mapping, accession numbers and nomenclature. Pharmacogenetics 6:1-41.

Nelson DR, Schuler MA, Paquette SM, Werck-Reichhart D, Bak S (2004). Comparative genomics of rice and *Arabidopsis.* Analysis of 727 cytochrome P450 genes and pseudogenes from a monocot and a dicot. Plant Physiol. 135:756-772.

Saitou N, Nei M (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406-425.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997). The CLUSTAL-X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25:4876-4882.

Turk EM, Fujioka S, Seto H, Shimada Y, Takatsuto S, Yoshida S, Wang H, Torres QI, Ward JM, Murthy G, Zhang J, Walker JC, Neff MM (2005). BASI and SOB7 act redundantly to modulate *Arabidopsis* photomorphogensis via unique brassinosteroid inactivation mechanisms. Plant J. 42:23-34.

VandenBosch KA, Stacey G (2003). Summaries of legume genomics projects from around the globe. Community resources for crops and models. Plant Physiol. 131:840-865.

Winkler RG, Helentjaris T (1995). The maize *Dwarf3* gene encodes a cytochrome P450-mediated early step in gibberellins biosynthesis. Plant Cell 7:1307-1317.

Werck-Reichhart D, Bak S, Paquette S (2002). Cytochrome P450. In: Somerville CR, Meyerowitz EM (eds) .The *Arabidopsis* book, American society of Plant Biologist, Rockville, MD. doi:/10.1199/tab.0028, http://www.aspb.org/publications/arabidopsis.