

## Identification of cytochrome P450 heme motif in plants proteome

Akansha Saxena<sup>1</sup>, Priyanka Singh<sup>2</sup>, Dharmendra K. Yadav<sup>3</sup>, Pooja Sharma<sup>3</sup>, Sarfaraz Alam<sup>3</sup>, Feroz Khan<sup>3\*</sup>, Sanjog T. Thul<sup>4</sup>, Rakesh K. Shukla<sup>4</sup>, Vikrant Gupta<sup>4</sup>, Neelam S. Sangwan<sup>3</sup>

<sup>1</sup>Biomedical Informatics Center, Indian Council of Medical Research, New Delhi, India

<sup>2</sup>Bioinformatics Infrastructure facility, Department of Biochemistry, University of Lucknow, Lucknow, India

<sup>3</sup>Metabolic and Structural Biology Department, <sup>4</sup>Biotechnology Division, CSIR-Central Institute of Medicinal and Aromatic Plants, Lucknow, India

\*Corresponding author: f.khan@cimap.res.in

### Abstract

Plant cytochrome (CYP) P450 monooxygenases are heme-containing enzymes that take part in the production of a wide variety of secondary products, a number of which are inhibitory to the survival of insects, pathogens and vertebrate herbivores. Main class of characterized natural plant compounds terpenoids are mostly substrates for plant CYPs. Metabolic engineering of plants using genetic modification of the P450 enzymes has remarkable implications for molecular farming of natural plant chemicals used as pharmaceuticals, disease and insect deterrents. However, molecular information regarding the metabolism of herbicides by plant CYPs is limited and most detoxifying CYPs are expressed at low levels in plants. Therefore, identification of plant CYP enzymes may help in characterization and increase in expression levels followed by design of herbicide-resistant plants. Therefore, in this study we have developed a method to explore the available plant's proteome for the identification of CYPs on the basis of conserved heme motif. For this we have used position-specific scoring matrix method and correlated the availability of CYPs on the basis of heme motif distribution. We have identified the proteome-wide CYPs on the basis of occurrence and positional arrangements of heme motif in the publicly available nine plant's proteome viz., *Arabidopsis thaliana*, *Vitis vinifera*, *Zea mays*, *Triticum aestivum*, *Sorghum bicolor*, *Glycine max*, *Brassica napus*, *Brassica oleracea*, and *Solanum lycopersicum*. Results showed that *V. vinifera* and *A. thaliana* have high conservation of heme motif in CYP enzymes as well as hypothetical sequences i.e., 88.46%, therefore characterizations of CYP enzymes from these prioritized plants may help to increase the CYPs expression levels followed by design of resistant plants. Among the heme motif sequence, three most variable amino acids i.e., alanine, isoleucine and proline can be targeted to enhance the possibility of functional role in the biosynthesis of secondary metabolites and consequently may increase the production of phytochemicals as a precursor for synthesis of active molecules for drug discovery and phytoremediation.

**Keywords:** Weight matrix, cytochrome P450, Heme motif, plants CYPs, proteome, herbicide, xenobiotics, metabolism.

**Abbreviations :** Position specific scoring matrix (PSSM); CYP, cytochrome P450; Alanine, Ala, A; Isoleucine, Ile, I; Leucine, Leu, L; Valine, Val, V; Phenylalanine, Phe, F; Tryptophan, Trp, W; Tyrosine, Tyr, Y; Asparagine, Asn, N; Cysteine, Cys, C; Glutamine, Gln, Q; Methionine, Met, M; Serine, Ser, S; Threonine, Thr, T; Aspartic acid, Asp, D; Glutamic acid, Glu, E; Arginine, Arg, R; Histidine, His, H; Lysine, Lys, K; Glycine, Gly, G; Proline, Pro, P.

### Introduction

Advancement in sequencing techniques and development of biological databases has opened the door for computational research. Large numbers of genome sequencing projects have resulted in exponential growth of biological sequences. This data inspire researchers to use the computational tools and techniques to identify conserved functional enzyme motifs present in cytochrome (CYP) sequences (Nelson DR, 1999). CYPs play a very prominent role in the biosynthesis of secondary metabolites in plants (Ayabe and Akashi, 2006). Plant CYP P450 monooxygenases are heme-containing enzymes that take part in the production of a wide variety of secondary products, a number of which are inhibitory to the survival of insects, pathogens and vertebrate herbivores (Kaspera and Croteau, 2006). In plants, these CYP enzymes also play an important role in the biosynthesis of several compounds such as hormones, defensive compounds and fatty acids. These enzymes are involved in electron transfer process forming CYP system which plays an important role in different metabolic biochemical reactions such as hydroxylation, oxidation, detoxification and synthesis of

various compounds in the cell and thus play a crucial role for the existence of life. Main class of characterized natural plant compounds terpenoids are mostly substrates for plant CYPs. Recently, metabolic engineering of plants using genetic modification of the P450 enzymes has remarkable implications for molecular farming of natural plant chemicals used as pharmaceuticals, disease and insect deterrents (Fischer et al., 2007; Kumar et al., 2012b). Moreover, plant CYP enzymes have been used to remove the herbicides and industrial contaminants apart from other methods (Morant et al., 2003). Plant CYPs catalyze herbicide metabolism and contribute to detoxification or activation of other agrochemicals in crop plants. However, molecular information regarding the metabolism of herbicides and other chemicals by plant CYPs is limited, and most herbicide detoxifying CYPs are expressed at low levels in plants. Therefore, identification of plant CYP enzymes may assist in characterization and increase expression levels followed by design of resistant plants. At present the first two plant CYPs that were identified are CYP76B1 from *Helianthus tuberosus*

and CYP71A10 from soybean. These CYP enzymes actively metabolize herbicides, namely phenylureas. Further, CYP81B2 and CYP71A11 were isolated from tobacco, which were shown to metabolize chlortoluron (Kumar et al., 2012b). Although these CYP enzymes play some role in herbicide metabolism, these plants do not contain a complete and efficient catabolic pathway to detoxify herbicides and other toxic compounds. Therefore, bacterial and mammalian CYP enzymes have been introduced in various plants for effective phytoremediation of these chemical. Presently, around 5100 sequences of plant CYPs have been identified and annotated which include 3651 CYPs from 11 completely sequenced plant genomes viz. *Arabidopsis*, *Brachypodium distachyon*, *Chlamydomonas reinhardtii*, grapevine, papaya, *Physcomitrella patens*, rice, soybean, *Selaginella moellendorffii*, tomato and *Volvox carteri* (Nelson and Werck-Reichhart, 2011). The analysis of the different CYPs indicate that all these metalloenzymes exhibits the same features so called 'P450 signature', a motif of 10 amino acids long, e.g., Phe-XX-Gly-Xb-XX-Cys-X-Gly. This motif facilitates the binding of heme iron to the CYP enzyme (Gribskov et al., 1987). Other than non-conserved residues, the conserved motif signature includes the hydrophobic aromatic residue phenylalanine (Phe), invariant hydrophobic cysteine (Cys) and simple small amino acid glycine (Gly). The amino acid refer by 'X' indicates any residue, while 'Xb' refer a basic amino acid that plays a key role in interactions with the reductase partner (Reichhart and Feyereisen, 2001; Nebert and Russell, 2002). In CYP, the basic amino acid residue interacts with two cysteine residues. Arginine has also an important role in the interactions between the oxygenase and the iron-sulfur enzyme that acts as electron transferase (Meunier et al., 2004). These motifs or patterns have biological significance and play an important role in functional analysis of associated enzymes (Degtyarenko KN, 1995; Degtyarenko and Kulikova, 2001). Keeping in mind this importance, we identified CYP enzymes in the available nine plants proteome by using heme motif box as identifying marker signature through a statistical weight matrix approach.

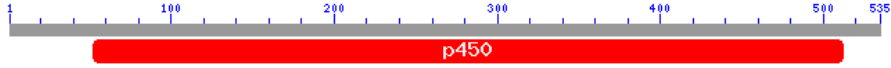
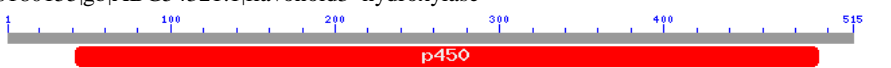

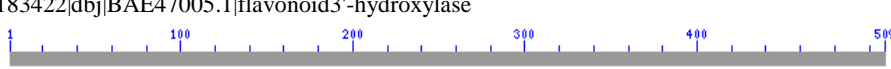


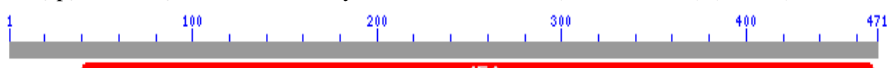
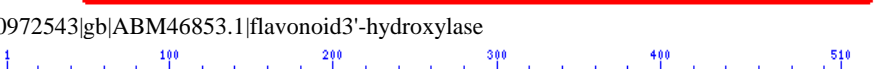

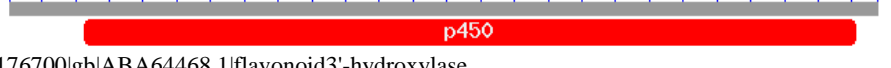
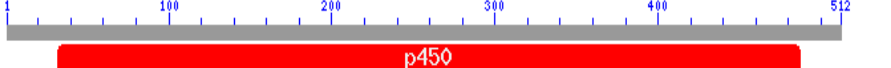
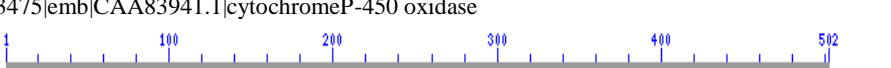

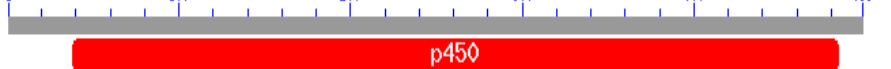

Proteome wide identification of CYP enzymes product with conserved heme motifs and their functional analysis is not widely studied in plants. Besides, there is no simple and reliable computational method to predict the CYP enzymes with the help of only motif signature in plant proteomes. This is the first study to explore the CYP enzymes with conserved heme motif in proteome of selected nine plants viz., *Arabidopsis thaliana*, *Vitis vinifera*, *Zea mays*, *Triticum aestivum*, *Sorghum bicolor*, *Glycine max*, *Brassica napus*, *Brassica oleracea*, and *Solanum lycopersicum* with the help of heme-box weight matrix approach. Selection of plants for searching CYP enzymes was done on the basis of availability of proteome data from public databases. Before carrying out proteome-wide search, we first constructed the weight matrix by using the training data set of experimentally characterized CYP enzyme sequences with heme motif of eighteen plants viz., *Oryza sativa* (*japonica* cultivar), *S. bicolor*, *V. vinifera*, *Catharanthus roseus*, *A. thaliana*, *Persea americana*, *Ageratina adenophora*, *Lobelia erinus*, *Gerbera* hybrid cultivar, *Mentha piperita*, *Asparagus officinalis*, *Beta vulgaris*, *Berberis stolonifera*, *Capsicum annuum*, *Coptis japonica*, *Gentiana triflora*, *Perilla frutescens* var. *crispa*, and *Ruta graveolens*. Conserved CYP enzyme domain was predicted through conserved domain database (CDD) at NCBI server. Known patterns of heme binding sites were identified through InterProScan and PPsearch tool at EBI server. Weight matrix was developed by MEME motif

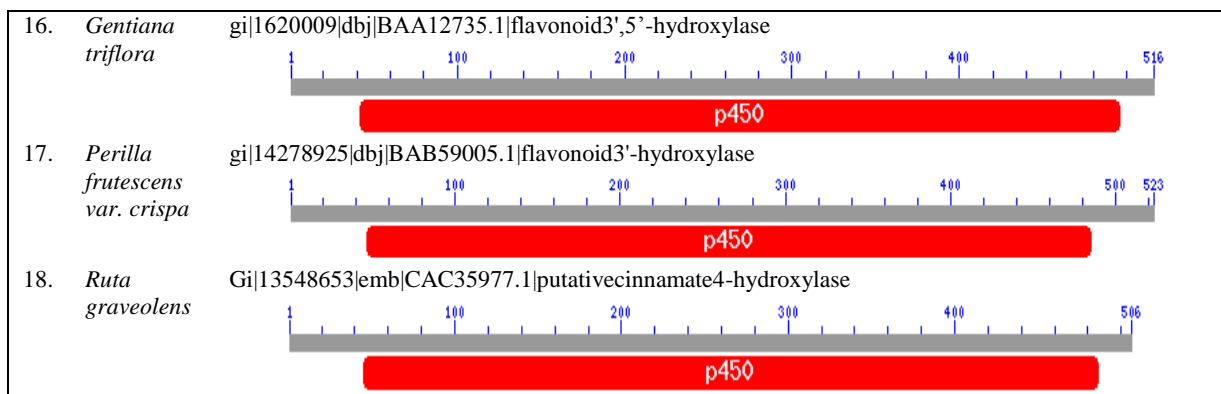
discovery and proteome wide search was performed by PoSSuMsearch in Linux RHv.4 operating system. Results indicate more conserved nature of heme motif in *V. vinifera* and *A. thaliana* (88.46%) and less conserved nature in *B. oleracea* (<50%). Among the heme motif consensus sequences (FGAGRRICPG) [Phenylalanine- Glycine- Alanine- Glycine- Arginine- Arginine- Isoleucine- Cysteine- Proline- Glycine], three most variable amino acids i.e., Alanine (A), Isoleucine (I) and Proline (P) can be targeted to enhance the possibility of functional role in the biosynthesis of secondary metabolites and consequently may increase the production of phytochemicals as a precursor for synthesis of active molecules for drug discovery research.

## Results and Discussion

### *Plants Cytochrome P450 and conserved heme motif signature*

Cytochrome P450 is membrane bound large and diverse superfamily of heme containing monooxygenase enzymes which is present in all divisions of life i.e. bacteria, plants, animals. These enzymes participate in oxidative metabolism of the large range of exogenous and endogenous compounds. CYP is colored cellular enzyme with heme iron pigment which absorbs light of wavelength 450 nm and reduced hence named as CYP. Due to large diversity, on the basis of sequence similarity CYP is divided into 70 families. In Plants, CYP participate in various biosynthetic pathways which result in biosynthesis of various hormones, fatty acid, steroid, drug and terpenoids in plants. These enzymes are involved in electron transfer and play an important role in many biological processes and the synthesis of various compounds in the cell (Gribskov et al., 1987; Fischer et al., 2007). They are also reported to be involved in signaling in regulation of plant growth and stress responses (Chaban et al., 2003; Cheng et al., 2010). CYP has large binding site so that it can bind to exogenous and endogenous compound easily. CYP play an important role in drug metabolism and drug interactions. CYP is highly diverged enzyme family but when it is studied at the sequence level, some conserved motifs are identified in all CYP enzymes based on their tertiary structure and the function which it performs. One of the signature motif which is regarded as the heme binding site is conserved. In this motif only cysteine residue is highly conserved in all P450 enzymes. This heme-iron motif provide binding sites for oxygen and various compounds which may act as drugs, but it cannot bind to all substrates. This motif is necessary for binding of heme. The heme iron is involved in the electron transport chain of cell via various cyclic oxidation and reduction reaction step (Nelson, 1999). The heme iron binds with the six ligands in octahedral manner and complex with four nitrogen atoms of protoporphyrin IX, and, two axial ligands lie normal to porphyrin plane. The activity of CYP shows high variation because it catalyzes large number of biological reactions. CYP enzymes show polymorphism as it is present in multiple forms (Fischer et al., 2007). Various types of biologically active motifs are present in CYP; heme binding site is one of them. The proteomic level analysis of the different CYP indicates that all these metalloenzymes exhibit the same pattern i.e., "P450 signature" motif of 10 amino acids, including the invariant cysteine residue that ligates the heme iron to the enzyme, Phe-XX-Gly-Xb-XX-Cys-X-Gly (Degtyarenko, 1995). The consensus sequence for the heme binding site is FGAGRRICPG (Suppl. Table 1) (Fig 1-2). This study aimed to predict the heme binding site-like patterns in selected plant

S. No	Plant name	Conserved CYP domain
1.	<i>Oryza sativa</i>	gi 21672032 gb AAM74394.1 AC119149_9 cytochromeP450 (pfam00067) 
2.	<i>Sorghum bicolor</i>	gi 110180155 gb ABG54321.1 flavonoid3'-hydroxylase 
3.	<i>Vitis vinifera</i>	gi 78183422 dbj BAE47005.1 flavonoid3'-hydroxylase 
4.	<i>Catharanthus roseus</i>	gi 78183422 dbj BAE47005.1 flavonoid3'-hydroxylase 
5.	<i>Arabidopsis thaliana</i>	gi 15225510 ref NP_182079.1 CYP76C4 (cytochromeP450, family76, subfamily-C, polypeptide4); oxygen binding 
6.	<i>Persea americana</i>	gi 117188 sp P24465.1 C71A1_PERAECytochromeP45071A1 (CYPLXXIA1) (ARP-2) 
7.	<i>Ageratina adenophora</i>	gi 120972543 gb ABM46853.1 flavonoid3'-hydroxylase 
8.	<i>Lobelia erinus</i>	gi 133874242 dbj BAF49324.1 flavonoid3'-hydroxylase 
9.	<i>Gerbera hybrid cultivar</i>	gi 77176700 gb ABA64468.1 flavonoid3'-hydroxylase 
10.	<i>Mentha piperita</i>	gi 493475 emb CAA83941.1 cytochromeP-450 oxidase 
11.	<i>Asparagus officinalis</i>	gi 40645046 dbj BAD06417.1 cytochromeP450 
12.	<i>Beta vulgaris</i>	gi 6979556 gb AAF34537.1 AF195816_1isoflavonesynthase1 
13.	<i>Berberis stolonifera</i>	Gi 642386 gb AAC48987.1 cytochromeP-450CYP80] 
14.	<i>Capsicum annuum</i>	gi 6739506 gb AAF27282.1 AF122821_1cytochromeP450 
15.	<i>Coptis japonica</i>	gi 9971208 dbj BAB12433.1 (S)-N-methylcoclaurine-3'-hydroxylase 



**Fig 1.** Identification of conserved functional CYP domain in the experimental data set of selected plant's CYP sequences used for weight matrix development.

proteome through weight matrix machine learning approach (Suppl. Table 2). A weight matrix can simply predict the motif of interest present in sequences of the genome or proteome. Experimental identification of conserved motifs in proteome is very tedious and time consuming, thus, we used weight matrix method for proteome-wide search. In genome or proteome-wide studies, weight matrix and other machine learning techniques are so far found to be reliable (Suppl. Table 3).

#### **Identification of Heme motif in selected plant's proteome**

To identify the heme-binding site like patterns, scanning through derived heme weight matrix was performed on proteome of nine plants *i.e.*, *Arabidopsis thaliana*, *Vitis vinifera*, *Zea mays*, *Triticum aestivum*, *Sorghum bicolor*, *Glycine max*, *Brassica napus*, *Brassica oleracea* and *Solanum lycopersicum*. Results showed that a large number of heme-binding site like patterns are localized in CYP enzyme and hypothetical sequences of these plants, which are not yet identified experimentally. The comparison of results revealed that weight matrix score for heme motif like patterns in CYP enzymes is higher for *V. vinifera* and *A. thaliana* (*i.e.*, 88.46% sequence similarity refer by MSSP) in compare to low score putative uncharacterized, hypothetical and other functional category protein sequences. This suggests presence of high level of CYP enzymes in these plants. On the basis of distribution of predicted heme motif like patterns, selected plants were prioritized for presence of CYP enzymes. Results suggest that lowest heme motif distribution and weight matrix score found in case of *B. oleracea* (*i.e.*, <50%), thus indicates low level of CYP enzymes (Fig 4-5). The conserved domain study also showed similarities in the P450 domain amongst all the eighteen enzymes of training data set which was used for weight matrix construction. Multiple sequence alignment of all eighteen homologs showed highly conserved heme binding site motif. Weight matrix scores of a predicted heme binding site-like patterns was in the range of 76 to 156. Derived weight matrix successfully predicted the heme motif-like patterns at minimum and maximum cut-off scores of 78 and 138, respectively (Suppl. Table 3). At score 138, weight matrix predicted the 100% similar heme motif-like patterns. In *B. oleracea*, significant patterns were predicted in other than CYP enzymes at lower score in compare to other plants. In case of *A. thaliana*, *V. vinifera* and *S. bicolor*, lower cut-off used were 129, 135 and 135, respectively.

#### **Proteome-wide identification of heme binding site motif in *A. thaliana***

In *A. thaliana*, proteome-wide identification of conserved heme motif like pattern through derived weight matrix was performed on cut-off score  $\geq 129$  (*i.e.*, 82.69% sequence similarity refer by MSSP) to avoid false positive predictions. Results showed significant hits and indicate a maximum of 88.46% sequence similarity of predicted motifs with experimental patterns. These highly conserved motifs were found in mostly enzymes belong to cytochrome family after proteome wide search. Twenty one cytochrome P450 family enzymes and an uncharacterized enzyme (Q304B3\_ARATH) showed the presence of two types of heme motif-like patterns FGAGRRICPA and FGAGRRICPG with 88.46% motif sequence similarity. Similarly, a flavonoid 3'-monooxygenase (Flavonoid 3'-hydroxylase) (AtF3'H) (Cytochrome P450 75B1) enzyme showed the presence of a conserved heme motif-like pattern FGAGRRICAG with 86.53% sequence similarity. Thirty four enzymes mainly cytochrome P450 and four uncharacterized enzymes (A4VCM5, Q8LPP9, Q5E922 and Q3EB00) showed the presence of two types of conserved heme motif-like patterns FGSRRICPA and FGSRRICPG with 83.33% sequence similarity. Results indicate a comparatively less conservation of heme motif in four cytochrome P450 family enzymes, which showed the presence of two types of conserved patterns FGGRRICPA and FGGRRICPG with 82.69% motif sequence similarity (Suppl. Table 4). Molecular modeling results showed a significant hits and sequence similarity of predicted binding site residues *i.e.*, hydrophobic and non-polar PHE-393, hydrophobic polar GLY-394, GLY-396 and GLY-402, basic ARG-398, and polar residue CYS-400. Predicted heme motif (FGAGRRICPG) with 88.46% MSSP showed highly conserved residues PHE-393, GLY-394, GLY-396, ARG-398, CYS-400, GLY-402 interacting with hemoenzyme domain in *A. thaliana* (Fig-6a).

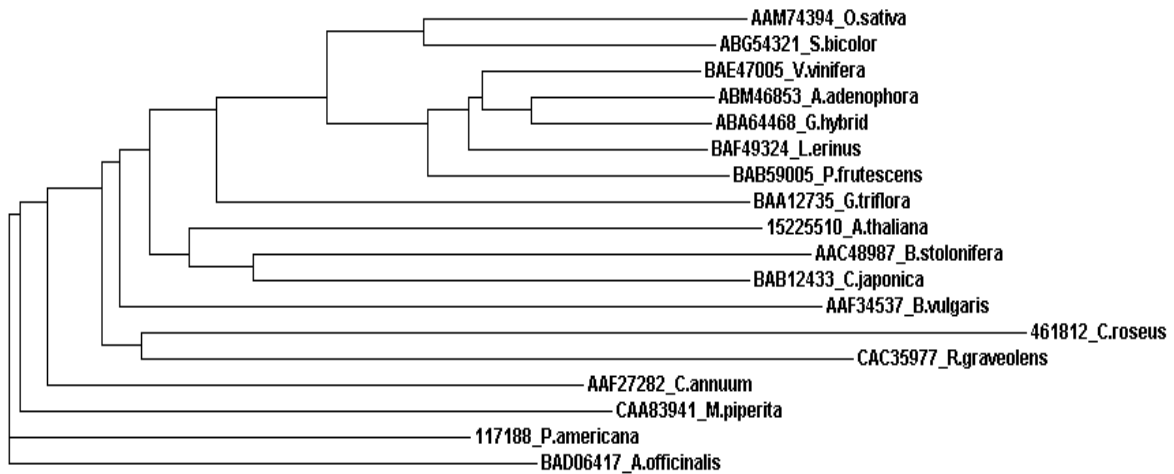
#### **Proteome-wide identification of heme binding site motif in *V. vinifera***

In *V. vinifera*, proteome-wide identification of conserved heme motif-like pattern through derived weight matrix was performed on a cut-off score of  $\geq 135$  (*i.e.*, 86.53% sequence similarity refer by MSSP) and significant hits were found in putative uncharacterized proteins rather than CYPs.

1. AAM74394_ <i>O. sativa</i>	VDVKGND FGLI PFGAGRRICAGLSWGLRMVTMTAATLVHAFDWQLPA--- 496
2. ABG54321_ <i>S. bicolor</i>	VDVKGSD FFEI PFGAGRRICAGLSWGLRMVTIMTATLVHAFDWDLAD--- 476
3. ABM46853_ <i>A. adenophora</i>	VDVVRGND FEVI PFGAGRRICVGMTLGLRMVQLLVATLVQT FDWELAK--- 471
4. ABA64468_ <i>G. hybrid</i>	TDIKGNDFEVI PFGAGRRICVGMISLGLRMVQLLTATLVHAFDWELAD--- 471
5. BAE47005_ <i>V. vinifera</i>	ADVVRGND FEVI PFGAGRRICAGMSLGLRMVHLLTATLVHAFNWELPE--- 471
6. BAF49324_ <i>L. erinus</i>	VDVKGND FEVI PFGAGRRVCAGLSLGLRMVQLVLTATLVHSDWELAD--- 477
7. BAB59005_ <i>P. frutescens</i>	VDVVRGND FEI PFGSGRRICAGMNLGIRMVQLLIATMVHAFDFELAN--- 483
8. BAA12735_ <i>G. triflora</i>	IDPRGNHFELI PFGAGRRICAGTRMGILLVEYI LGTLVHSDWKLGL--- 479
9. AAC48987_ <i>B. stolonifera</i>	IEYNGKQFQFI PFGSGRRICPGRPLAVRIIPLVLASLVHAFGWELPD--- 456
10. BAB12433_ <i>C. japonica</i>	VDYKGNDFELI PFGGGRRICPGLPLASQFSNLIIVATLVQNFWSLPLQ--- 454
11. 15225510_ <i>A. thaliana</i>	IDVKGNDYELT PFGGGRRICPGLPLAVKTVSIMLASLLYSFDWKLPN--- 477
12. AAF34537_ <i>B. vulgaris</i>	LDLRGSHFQLL PFGSGRRMCPGVNLTSGTATLLASLIQC FDLQVLGPQG 464
13. 117188_ <i>P. americana</i>	VDFKQDFQLI PFGAGRRGCPGI AFGISSVEISLANLLYWFNWELP--- 469
14. BAD06417_ <i>A. officinalis</i>	IDFRGQCFEVP PFGAGRRICPGMHFAAANLELALANIMYRFDWELPD--- 463
15. CAA83941_ <i>M. piperita</i>	VDFKGLDFELI PFGAGRRGCPGTTFPMATLEFTLANIMQKFDWELP--- 473
16. AAF27282_ <i>C. annuum</i>	VDFLGSHHQFI PFGAGRRICPGMLFGLANVGQPLAQLLYHFDKRLPN--- 470
17. CAC35977_ <i>R. graveolens</i>	VEANGNDFRYI PFGVGRRCSPGI ILALPILGITIGRMVQNFELLP PPG-- 475
18. 461812_ <i>C. roseus</i>	ANATKNNVTYL PFSWGRVCLGONFALLQAKLGLAMILQRKFQDVA--- 497



**Fig 2.** Identification of conserved residues of heme motif in the selected orthologous set of CYP enzyme sequences. Logo represent the conserved nature of heme motif amino acid residues at position 1, 2, 4, 5, 6, 8, and 10.



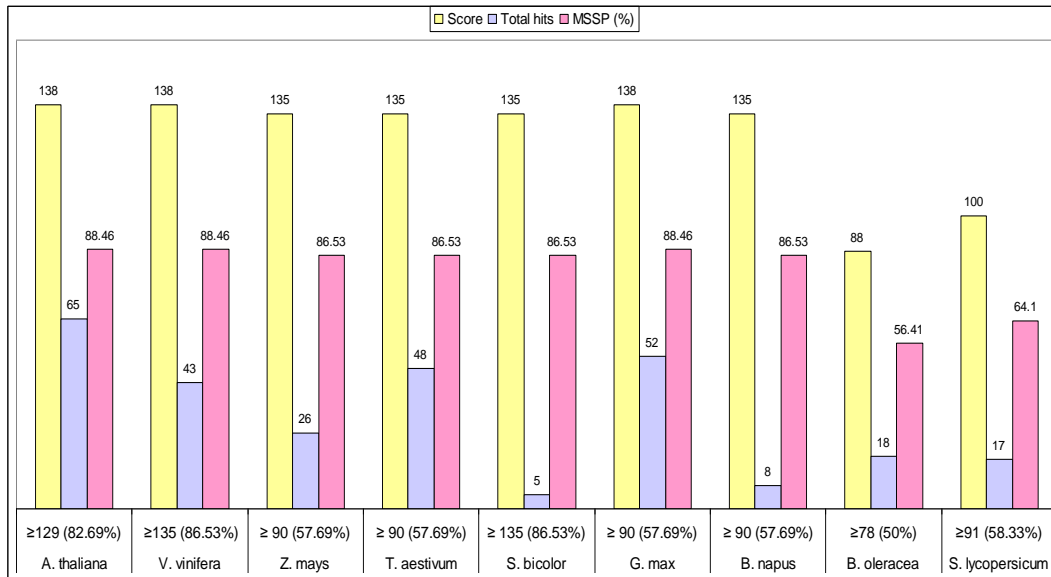
**Fig 3.** Phylogenetic tree showing unrooted phylogram representing the evolutionary relationship among orthologous set of selected plant's CYP 450 enzyme sequences. Evolutionary close relationship showed by *O. sativa*, *S. bicolor*, *V. vinifera*, *A. adenophora*, *G. hybrid*, *L. erinus*, and *P. frutescens*. Evolutionary distant relationship showed by *G. triflora*, *A. thaliana*, *B. stolonifera*, *C. japonica*, *B. vulgaris*, *C. roseus*, *R. graveolens*, *C. annuum*, *M. piperita*, *P. americana*, and *A. officinalis*.

This suggests that these uncharacterized proteins might belong to CYP protein family and may play an important role in plant metabolism. These high score predicted heme motifs belong to mostly uncharacterized proteins which indicate their functional similarity with CYP family proteins. Results showed that 35 enzymes revealed localization of three types of conserved heme motif-like patterns FGAGRRICPE, FGAGRRICPA, and FGAGRRICPG with 88.46% motif sequence similarity. Out of these 35, nineteen were putative uncharacterized enzymes and sixteen were chromosome chr6 scaffold\_3 unknown enzyme sequences. Similarly, a total of 9 enzymes showed the presence of a conserved heme motif-like pattern FGAGRRICAG with 86.53% sequence similarity. Out of 9, eight were putative uncharacterized enzymes and a flavonoid 3' 5' hydroxylase (A6XHG1) enzyme (Suppl. Table 5). Molecular modeling results showed significant hits and sequence similarity of predicted binding

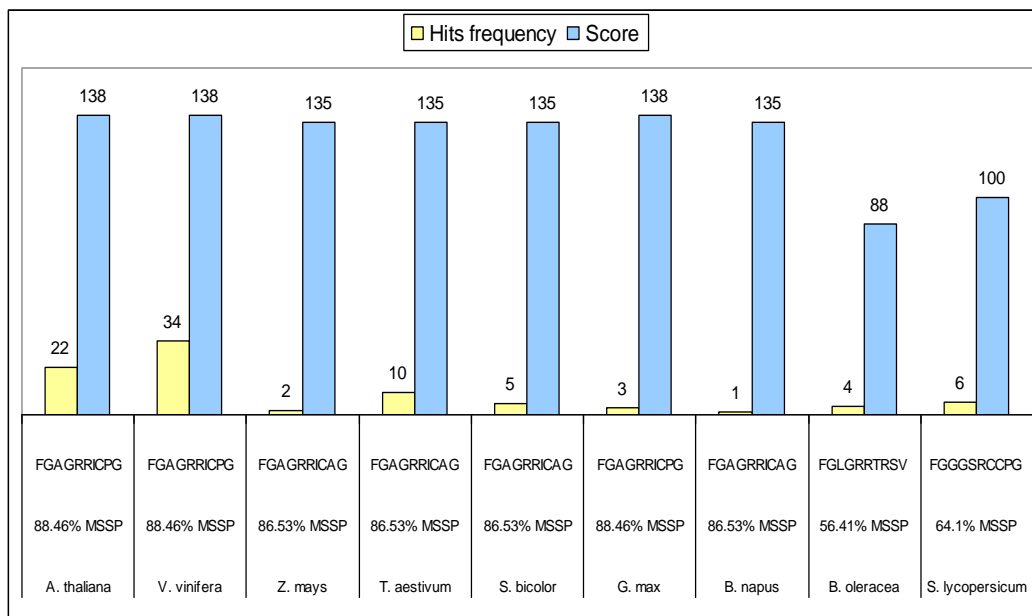
site residues i.e., hydrophobic, non-polar PHE-393, hydrophobic polar GLY-394, GLY-396 and GLY-402, basic ARG-398, and polar CYS-400. Predicted heme motif (FGAGRRICPG) with 88.46% MSSP showed conserved residues PHE-393, GLY-394, GLY-396, ARG-398, CYS-400, GLY-402 interacting with hemoenzyme domain in *V. vinifera* (Fig-6b).

**Proteome-wide identification of heme binding site motif in *Z. mays***

In *Z. mays*, proteomic identification of conserved heme motif-like pattern through derived heme-box weight matrix was performed on a cut-off score of  $\geq 90$  (i.e., 57.69% sequence similarity refer by MSSP) and significant hits were obtained in mostly CYP proteins family, Cell wall invertase, Myb protein, Proteasome subunit beta type, and putative uncharacterized proteins. The uncharacterized proteins



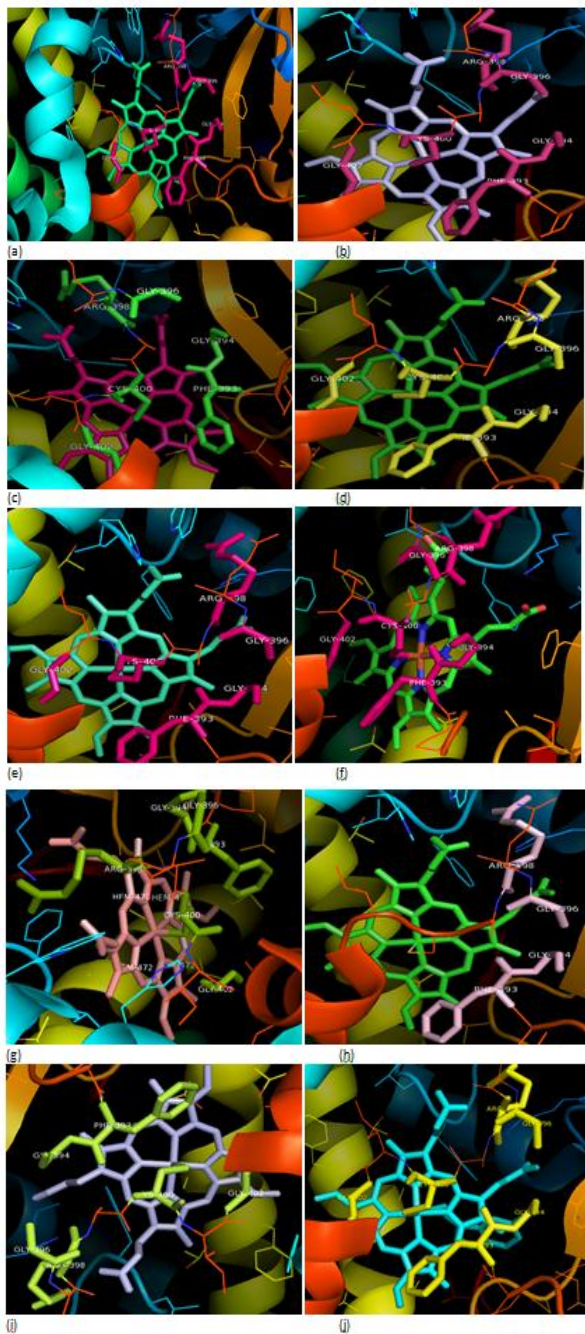
**Fig 4.** Distribution of predicted heme motif in each plant's proteome and comparison of score, total hit frequency and motif sequence similarity percentage (MSSP) at given cut-off parameter. Maximum score achieved was 138 in *A. thaliana*, *V. vinifera*, and *G. max* with total predicted motif hit frequency of 65, 43, and 52 respectively. Distribution of low match motif showed by *B. oleracea* and *S. lycopersicum* with total hit frequency of 18 and 17 at lowest score of 88 (56.41% MSSP) and 100 (64.1% MSSP), respectively.



**Fig 5.** Frequency of predicted high score heme motif in each plant's proteome and comparison of pattern hit frequency with score value. Maximum hit frequency showed by *V. vinifera* and *A. thaliana* at maximum score value of 138 (88.46% MSSP). Low match motif showed by *B. oleracea* and *S. lycopersicum* at minimum score of 88 (56.41% MSSP) and 100 (64.1% MSSP), respectively

predicted to be CYP family proteins. Results showed that a highly conserved pattern FGAGRRICAG was found in two cytochrome P450 family enzymes (O04980 and Q43250) with 86.53% sequence similarity to heme motif. A more conserved heme motif-like pattern FGSGRRICPG was found in three cytochrome P450 family enzymes (Q8VYA8, Q43256 and Q43255) with 83.33% sequence similarity. Similarly, a pattern FGAGRRVCPG in cytochrome P450 enzyme was detected with 82.05% similarity. Two cytochrome P450 family enzymes showed a moderately conserved pattern FGSGRRMCPG with 76.28% similarity. Surprisingly, two putative uncharacterized enzymes showed a conserved pattern LGAGRRFTPS with 69.23% similarity. At the same time, another four cytochrome P450 enzymes

showed conserved patterns FGWGPRICIG and FGGGPRICIG with heme motif sequence similarity of 66.02% and 66.66%, respectively. Results also indicate a conserved pattern FGAGGRTCIL with heme motif similarity of 64.10% in two cell wall invertase enzyme sequences. A P-type R2R3 MYB enzyme with conserved pattern PGPGRPGCPS also showed heme motif similarity of 63.46%. Similarly, a cytochrome P450 88A1 (Dwarf3 enzyme) with conserved pattern FGLGARLCPG also showed sequence similarity of 62.82% with heme motif. A cytochrome P450-like enzyme (Q94KE5) showed the presence of a conserved pattern FQAGPRICLG with 62.17% sequence similarity. A glutathione S-transferase GST 30 (Q9FQA9) showed the presence of a less conserved heme



**Fig 6.** Representation of predicted heme motif residues with hemoenzyme domain in CYP (PDB ID: 2HPD). (a) Predicted heme motif (FGAGRRICPG; 88.46% MSSP) residues PHE-393, GLY-394, GLY-396, ARG-398, CYS-400, GLY-402 (pink color) with hemoenzyme domain (green color) in *A. thaliana*, (b) Predicted heme motif (FGAGRRICPG; 88.46% MSSP) residues PHE-393, GLY-394, GLY-396, ARG-398, CYS-400, GLY-402 (pink color) with hemoenzyme domain (purple color) in *V. vinifera*, (c) Predicted heme motif (FGAGRRICAG; 86.53% MSSP) residues PHE-393, GLY-394, GLY-396, ARG-398, CYS-400, GLY-402 (green color) with hemoenzyme domain (pink color) in *Z. mays*, (d) Predicted heme motif (FGAGRRICAG; 86.53% MSSP) residues PHE-393, GLY-394, GLY-396, ARG-398, CYS-400, GLY-402 (yellow color) with hemoenzyme domain (green color) in *T. aestivum*, (e) Predicted heme motif

(FGAGRRICPG; 86.53% MSSP) residues PHE-393, GLY-394, GLY-396, ARG-398, CYS-400, GLY-402 (pink color) with hemoenzyme domain (green color) in *S. bicolor*, (f) Predicted heme motif (FGAGRRICPG; 88.46% MSSP) residues PHE-393, GLY-394, GLY-396, ARG-398, CYS-400, GLY-402 (pink color) with hemoenzyme domain (green color) in *G. max*, (g) Predicted heme motif (FGAGRRICAG; 86.53% MSSP) residues PHE-393, GLY-394, GLY-396, ARG-398, CYS-400, GLY-402 (green color) with hemoenzyme domain (pink color) in *B. napus*, (h) Predicted heme motif (FGLGRRTRSV; 56.41% MSSP) residues PHE-393, GLY-394, GLY-396, ARG-398 (pink color) with hemoenzyme domain (green color) in *B. oleracea*, (i) Predicted heme motif (FGGGTRQCPG; 64.10% MSSP) residues PHE-393, GLY-394, GLY-396, ARG-398, CYS-400, GLY-402 (green color) with hemoenzyme domain (purple color) in *S. lycopersicum*, and (j) Predicted heme motif (FGAGRRICAG; 86.54% MSSP) residues PHE-393, GLY-394, GLY-396, ARG-398, CYS-400, GLY-402 (yellow color) with hemoenzyme domain (blue color) in *O. sativa*. motif-like pattern FGAVGRIIGS with 60.25% sequence similarity. A putative gag-pol enzyme (Q7XBD9) showed the presence of pattern FGAPRLIFES with 59.61% sequence similarity. Two putative uncharacterized enzymes, a Pyruvate Pi dikinase regulatory enzyme and cytochrome P450 monooxygenase CYP72A28 showed the presence of a less conserved heme motif-like patterns AAAGRRFTPS, SPAGRRLPPS, FGWGPRTCIG with 58.33% sequence similarity (Suppl. Table 6). Predicted heme motif (FGAGRRICAG) with 86.53% MSSP indicates highly conserved residues PHE-393, GLY-394, GLY-396, ARG-398, CYS-400, GLY-402 interacting with hemoenzyme domain in *Z. mays* (Fig-6c).

#### **Proteome-wide identification of heme binding site motif in *T. aestivum***

In *T. aestivum*, proteome-wide search was performed on a cut-off score of  $\geq 90$  (57.69% sequence similarity refer by MSSP). Results indicate a significant conservation among the identified CYP enzymes i.e., within predicted motif matching score range 135 (86.53% MSSP) to 103 (66.02% MSSP). Below score value 94 (62.25% MSSP) predicted low score motif patterns found in the proteins of Alternative splicing regulator, Cytochrome P450 51 (Obtusifoliiol 14-alpha demethylase), Cell wall invertase, Fructan 1-exohydrolase precursor, Fructan exohydrolase, H-protein, Resistance-related receptor-like kinase, Pre-mRNA processing factor, Proteasome subunit beta type, and WRKY protein.

Results showed that a total of thirteen cytochrome P450 family enzymes have conserved heme motif-like patterns FGSGRRICPG, FGAGRRVCPG and FGTGRRICPG with sequence similarity of 83.33%, 82.05% and 81.41%, respectively. Similarly, nine cytochrome P450 family enzymes and a flavonoid 3'-hydroxylase which is a P450-dependent monooxygenase (CYP) has F3'H activity and showed a highly conserved heme motif like pattern FGAGRRICAG with 86.53% sequence similarity (Suppl. Table 7). Predicted heme motif (FGAGRRICAG) with 86.53% MSSP indicates highly conserved residues PHE-393, GLY-394, GLY-396, ARG-398, CYS-400, GLY-402 interacting with hemoenzyme domain in *T. aestivum* (Fig-6d).

### ***Proteome-wide identification of heme binding site motif in S. bicolor***

In *S. bicolor*, proteomic search was done at a cut-off score of  $\geq 135$  (86.53%). Significant conservation was detected in mostly in CYP family enzymes. The results indicate that four flavonoid 3'-hydroxylase, a P450-dependent monooxygenase (CYP) and other cytochrome P450 enzyme showed a highly conserved heme motif-like pattern FGAGRRICAG with 86.53% sequence similarity (Suppl. Table 8). Predicted heme motif (FGAGRRICPG) with 86.53% MSSP indicates highly conserved residues PHE-393, GLY-394, GLY-396, ARG-398, CYS-400, GLY-402 interacting with hemoenzyme domain in *S. bicolor* (Fig-6e).

### ***Proteome-wide identification of heme binding site motif in G. max***

In *G. max*, proteome-wide search was performed on a cut-off score of  $\geq 90$  (57.69%). Significant conservation was identified in predicting motif patterns in *G. max* CYP enzymes. Except hypothetical DNA binding protein (Q2HWE6) at score 90 (57.69% MSSP) all the predicted high score (138 to 91) motifs were belong to CYP proteins family (88.46% to 58.33% MSSP). The results indicated that seven flavonoid 3'-hydroxylase enzymes revealed a highly conserved heme motif-like pattern FGAGRRICAG with 86.53% similarity. Similarly, three cytochrome P450 monooxygenase enzymes indicated a highly conserved pattern FGAGRRICPG with 88.46% sequence similarity (Suppl. Table 9). Predicted heme motif (FGAGRRICPG) with 88.46% MSSP indicates conserved residues PHE-393, GLY-394, GLY-396, ARG-398, CYS-400, GLY-402 interacting with hemoenzyme domain in *G. max* (Fig-6f).

### ***Proteome-wide identification of heme binding site motif in B. napus***

In *B. napus*, proteome-wide search was made at the cut-off score of  $\geq 90$  (57.69%). Significant conservation was found in predicting motif of CYP protein family members. High score (135 to 101) predicted motifs belong to CYP protein family (within 86.53% to 64.74% MSSP range). The results indicate that a flavonoid 3'-hydroxylase enzyme sequence contain a highly conserved heme motif-like pattern FGAGRRICAG with 86.53% similarity. Also, two cytochrome P450-dependent monooxygenase and a ferulate-5-hydroxylase (Q0PW93) enzyme indicate a moderately conserved pattern FGSGRRSCPG with 76.28% similarity. Similarly, cinnamate 4-hydroxylase isoform 1 and 2 enzymes showed a less conserved pattern FGVGRRSCPG with 75% sequence similarity (Suppl. Table 10). Predicted heme motif (FGAGRRICAG) with 86.53% MSSP showed conserved residues PHE-393, GLY-394, GLY-396, ARG-398, CYS-400, GLY-402 interacting with hemoenzyme domain in *B. napus* (Fig-6g).

### ***Proteome-wide identification of heme binding site motif in B. oleracea***

In *B. oleracea*, no significant hits were found for conserved heme motif through derived heme-box weight matrix in the proteome-wide search at a cut-off score of  $\geq 78$  (50%). A putative uncharacterized enzyme (Q25BL3) indicated the presence of a less conserved pattern FGLGRRTRSV, while an S-locus enzyme 11 (Q84KT7) indicated the presence of a

pattern KGTFRRRCGS, both at score value 88 (56.41% MSSP). Two S-locus linked 3 (SLL3) enzymes (Q2A9I4 and Q2A9F7) indicated the presence of pattern FGLGRRTRSV with a maximum of 56.41% sequence similarity to heme motif. Low score motif found in putative uncharacterized proteins (Suppl. Table 11). Predicted heme motif (FGLGRRTRSV) with 56.41% MSSP showed conserved residues PHE-393, GLY-394, GLY-396, ARG-398 interacting with hemoenzyme domain in *B. oleracea* (Fig-6h).

### ***Proteome-wide identification of heme binding site motif in S. lycopersicum***

In *S. lycopersicum*, no significant hits were found for conserved heme motif through derived weight matrix in proteome-wide search at a cut-off score of  $\geq 91$  (58.33%). However, three cytochrome P450 family enzymes and three 6-deoxocastasterone oxidases showed moderately conserved heme motif-like patterns FGGGSRCCPG, FGGGTRLCPG and FGGGTRQCPG with a maximum of 64.10% sequence similarity (MSSP) at score value 100. High score motifs were found in CYP proteins at score range 100 to 92 (i.e., 64.10% to 58.97% MSSP) (Suppl. Table 12). Predicted heme motif (FGGGTRQCPG) with 64.10% MSSP showed conserved residues PHE-393, GLY-394, GLY-396, ARG-398, CYS-400, GLY-402 interacting with hemoenzyme domain in *S. lycopersicum*, (Fig-6i).

### ***Comparative study for distribution of predicted heme motifs in the selected plants***

To identify and prioritized the plants based on CYP enzymes presence, the derived heme weight matrix was successfully used to predict the heme binding site like patterns in the selected plants proteome in different ratios. Distribution of predicted heme motif in each plant's proteome and comparison of score, total hit frequency and MSSP at different cut-off parameters revealed that maximum hit score achieved was 138 in *A. thalina*, *V. vinefera*, and *G. max* with total predicted motif hit frequency of 65, 43, and 52 respectively. Distribution of low match motif showed by *B. oleracea* and *S. lycopersicum* with total hit frequency of 18 and 17 at lowest score of 88 (56.41% MSSP) and 100 (64.1% MSSP) respectively (Fig 4). Similarly, frequency of predicted high score heme motif in each plant's proteome and comparison of pattern hit frequency with score value revealed that higher hit frequency showed by *V. vinefera* and *A. thalina* at maximum score value of 138 (88.46% MSSP). On the other hand, low match motif showed by *B. oleracea* and *S. lycopersicum* at minimum score of 88 (56.41% MSSP) and 100 (64.1% MSSP) respectively (Fig 5). This study suggests that low motif score plants should be avoided while selecting the plants for CYP activity.

### ***Functional annotation of uncharacterized proteins on the basis of heme motif***

Putative function of the uncharacterized (unknown) and hypothetical proteins was predicted as CYP like enzymes on the basis of localization of highly conserved heme motif like patterns in their protein sequences. Characterization of these unexplored proteins is a subject of further research work. This way, we have identified the new CYP like enzymes in the studied plants which are not yet identified experimentally. Large numbers of conserved patterns were predicted in



selected plant's proteome, which were identical or similar to experimental heme motifs.

### **Comparison of plant's proteome wide predicted heme motifs**

Proteome-wide prediction in *A. thaliana* showed that significant patterns were present in a large number of CYP enzymes. All the sequences of *A. thaliana* were retrieved through PlantGDB database. A total of 50164 enzyme sequences were scanned for heme binding site prediction through the weight matrix at cut-off value  $\geq 129$ . A similar approach was used for proteomes of *V. vinifera*, *Z. mays*, *T. aestivum*, *S. bicolor*, *G. max*, *B. napus*, *B. oleracea*, and *S. lycopersicum*. In total, 54367, enzymes were searched for highly similar motifs in *V. vinifera* at the cut-off 135. In *Z. mays*, a total of 3978 enzymes showed highly similar motifs *i.e.*, 86.53% true positive at cut-off  $\geq 90$ . In *T. aestivum*, a total of 2857 enzymes showed highly similar motifs *i.e.*, 86.53% true positive at cut-off  $\geq 90$ . In *S. bicolor*, a total of 771 enzymes showed highly similar motifs *i.e.*, 86.53% true positive at cut-off  $\geq 135$ . *S. bicolor* enzyme sequences were also present in the training data set. In the training data set, heme motifs were predicted at a cut-off value of 135. In *G. max*, a total of 2118 enzymes showed highly similar motifs *i.e.*, 86.53% true positive at cut-off  $\geq 90$ . Similarly, a total of 1065 enzyme sequences from *B. napus* were scanned for the heme binding site. In *B. oleracea*, a total of 786 enzymes were searched for highly similar motifs *i.e.*, 86.53% true positive at cut-off  $\geq 90$  which could not identify any motif, therefore, search were done at lower score cut-off which identified insignificant patterns. Lastly, in *S. lycopersicum*, a total of 1887 enzymes were searched for highly similar motifs *i.e.*, 86.53% true positive at cut-off  $\geq 90$  (Suppl. Table 4-12).

## **Materials and methods**

### **Training data set selection**

Weight matrix was derived through multiple sequence alignment of the training data set sequences of experimentally identified CYP enzymes belong to eighteen plants *viz.*, *O. sativa* (*japonica*), *S. bicolor*, *V. vinifera*, *C. roseus*, *A. thaliana*, *P. Americana*, *A. adenophora*, *L. erinus*, *G. hybrid cultivar*, *M. piperita*, *A. officinalis*, *B. vulgaris*, *B. stolonifera*, *C. annuum*, *C. japonica*, *G. triflora*, *P. frutescens* *var. crispa*, and *R. graveolens* (Suppl. Table 1).

These CYP enzyme sequences were selected on the basis of CYP enzyme sequence and functional protein domain similarity by using BLASTp program at NCBI web server ([www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)). Enzyme sequences of CYP-P450 from different plant families were retrieved from GenBank database at NCBI web server.

### **Test data set of selected plants proteome**

Based on availability of plants proteome data on the Plant Genome Database (PGDB) ([www.plantgdb.org](http://www.plantgdb.org)), proteome of only nine plants *viz.*, *A. thaliana* (PGDB proteome reference ID: 3702), *V. vinifera* (PGDB proteome reference ID: 29760), *Z. mays* (PGDB proteome reference ID: 4577), *T. aestivum* (PGDB proteome reference ID: 4565), *S. bicolor* (PGDB proteome reference ID: 4558), *G. max* (PGDB proteome reference ID: 3847), *B. napus* (PGDB proteome reference ID: 3708), *B. oleracea* (PGDB proteome reference ID: 3712), and *S. lycopersicum* (PGDB proteome reference ID: 4081) were retrieved from PGDB database.

### **CYP P450 conserved domain identification**

Evidence of CYP was verified on the basis of presence of CYP P450 conserved domain localized in all the selected eighteen CYP enzyme sequences used in training data set. Conserved Domain Database (CDD) at NCBI, USA web server (<http://www.ncbi.nlm.nih.gov/cdd/>) was used to predict the conserved CYP P450 protein domain in the training data set. Heme motif was localized within CYP P450 domain.

### **Evolutionary relationship study by sequence alignment and phylogenetics**

Prior to heme motif weight matrix development, evidence of CYP family relationship was confirmed through multiple sequence alignment and phylogenetic studies for the selected training set CYP enzymes by using Clustal-W2 program at EBI web server, UK (<http://www.ebi.ac.uk/Tools/msa/clustalw2/>). All showed close evolutionary relationship among orthologous sequences.

### **Identification of Heme motif in the training set of CYP enzymes**

Experimentally known motif of heme binding sites was identified through InterProScan (<http://www.ebi.ac.uk/Tools/pfa/iprscan/>) and PPsearch (<http://www.ebi.ac.uk/ppsearch/>) software's at EBI web server, UK. PPsearch tool searches the Prosite database (<http://prosite.expasy.org/>) for the identification of conserved CYP enzyme motifs. MEME tool (<http://meme.sdsc.edu/meme/intro.html>) was used to detect heme motif patterns in training data sets first and later predicted patterns were used for the development of the alignment matrix followed by weight matrix. The expectation maximization method MEME appears to be a more robust motif search algorithm, as MEME resulted in a significant higher rate of identification of the motifs. In a few cases, there was a clear enrichment of motifs at particular positions.

### **Logo representation of conserved heme motif**

Distribution of amino acid residues at specific position detected by multiple sequence alignment of experimentally identified heme motif patterns of CYP enzymes of training data set. Conservation was represented by consensus sequence represented in the form of logo predicted by WebLogo tool (<http://weblogo.berkeley.edu/logo.cgi>), where the bit scale value of each residue is related to its conservation or frequency of amino acid residues at that position.

### **Structural molecular modeling of CYP**

The 3D crystal structure of CYP enzyme was retrieved through Protein Data Bank (RCSB PDB) (<http://www.rcsb.org/pdb/>), so that to explore the crystal structure of CYP and showed the experimental as well as predicted heme motif sequence interacting with hemoprotein. Structural molecular modeling was performed on CYP hemoenzyme domain of CYP (PDB ID: 2HPD). Confirmation studies related to presence of CYP enzyme secondary structure and conserved domain or motif was performed through the PDB, SCOP (<http://scop.mrc-lmb.cam.ac.uk/scop/>) and Pfam (<http://pfam.sanger.ac.uk/>) databases. Known patterns of heme binding sites were identified through EMBOSS Stretcher Alignment

([www.ebi.ac.uk/Tools/psa/](http://www.ebi.ac.uk/Tools/psa/)) tool at the EBI server. EMBOSS Stretcher tool searches an optimal global alignment of two protein sequences using a modification of the classical dynamic programming algorithm which uses linear space.

### Development of weight matrix

Position specific scoring matrix (PSSM) or weight matrix is frequently used for scanning of large data sets of biological sequences like genome or proteome. A high PSSM score in some region of a sequence often indicates a possible biological relationship of sequences to the family or motif characterized by the PSSM (Beckstette et al., 2006; Wu et al., 2000). There are several databases utilizing PSSM for function assignment and annotation, e.g., PROSITE, PRINTS and BLOCKS (Henikoff et al., 2000). A weight matrix is a representation of multiple sequence alignment that has no gaps. A weight matrix is a table of score values that gives a weighted match to any given substring of fixed length. It has one row for each symbol of the alphabet and one column for each position in the pattern (Sen et al., 2009). The score assigned by the matrix to a substring:

$$S = (S_j)_{j=1}^N \text{ is defined as } \sum_{j=1}^N m_{s_j,j}$$

Score (matrix) represented as:

$$S = m_{\alpha, j}$$

Where, j represents the position of the substring,  $s_j$  is the symbol at position j in the substring, and  $m_{\alpha, j}$  is the score in a row  $\alpha$ , column j of the matrix.

The matrix may be made from a multiple sequence alignment or by searching for patterns of the same length in a set of sequences using pattern-finding or statistical methods, e.g., expectation maximization, Gibbs sampling, ASSET, and by aligning the conserved patterns (Beckstette et al., 2006; Wu et al., 2000). The consecutive columns of the matrix represent columns of the aligned patterns and the rows represent the distribution of amino acids in each column of the alignment. The PSSM columns include log odds scores for evaluating matches with a target sequence. The matrix is then used to search a sequence for comparing patterns by sliding the matrix along the sequence and at each position in the sequence, evaluating the match at each column position using the matrix values for that column. The log odds scores for each column are added to obtain a log odds score for the alignment to that sequence position. High log odds scores represent a significant match. Using a scoring matrix instead of a single query sequence can enhance a database search because the matrix represents the greater amount of sequence variation found in a multiple sequence alignment. The similar approach was successfully used in genome-wide identification of transcription factors binding sites e.g., SinR (Khan et al., 2009), LexA (Khan et al., 2008), PhoB (Khan et al., 2006), NodD (Khan et al., 2005a,b), GltC (Khan et al., 2003), and recently DREB protein characterization in wheat (*Triticum aestivum*) (Sazegari and Niazi, 2012). Amino acid representation in each column of the alignment is also reflected in the matrix scores for that column; the more common an amino acid, the higher the score for a match to that amino acid. Weight matrix scores were later transformed into the percentage form by using standard approach (Khan et al., 2009).

The motif sequence similarity percentage (MSSP) was then calculated by using the following formula:

$$\text{MSSP} = \text{Os/Es} * 100$$

Where, Os = observed score value for each predicted motif, and

Es = maximum score value obtained for 100% matching (i.e., 156).

Several approaches and techniques are used for the prediction of conserved motifs and repeats in the DNA and protein sequences. Here, the weight matrix approach was used to predict conserved motif in the selected plant proteome to explore the yet unidentified CYP enzymes/genes. Experimentally reported heme binding sites was used for weight matrix model development. PoSSuMsearch (Beckstette et al., 2006), a Linux based standalone software was used to run the weight matrix file to scan the selected nine plant proteome. Successful application of PoSSuMsearch software reported in the studies of transcription factor binding site e.g., SinR (Khan et al., 2009) and LexA (Khan et al., 2008). After setting the standard cut-off parameters for true positive predictions, the newly designed heme-box matrix was first run on training data sets as part of validation process and subsequently on the selected nine plant proteome.

### Training set validation and Identification of heme motif in plant proteome

Before proteome wide predictions, the heme motif-like patterns were first predicted in the training set by the derived weight matrix. After internal validation, predictions were made on the selected plants proteome data one by one using same weight matrix. The weight matrix detects the heme motifs at minimum and maximum cut-off of 78 and 138, respectively. To filter the spurious results, we mostly selected the hits at cut-off  $\geq 90$  in each plants proteome. After validation of derived weight matrix, selected plant's proteomes were scanned for conserved patterns similar to heme binding sites.

### Search algorithm and parameters used in PoSSuMsearch

One of esa (enhanced suffix array algorithm), lahead (Look ahead search algorithm) or simple (Simple search algorithm) provided with PoSSuMsearch software (Beckstette et al., 2006). Output restriction in the form of E-Val: an E-value to determine the threshold from sequence set, P-Val: a p-value to determine the threshold from sequence set, MSSTH: a matrix similarity score (MSS) to determine the threshold from sequence set or motif, RAWTH: a raw threshold value, and Best: search for 'k' best matches for each PSSM. Output format was in the form of one of human, cismil or tabs. MSSP refer the motif sequence similarity percentage. E-value refer an expected by chance error probability to determine the threshold or cutoff for significant motif hit. P-value refer a probability value to determine the threshold or cutoff for significant predicted motif and MSS refer a matrix similarity score to determine the threshold or cutoff for predicted motif.

### Conclusion

This study provides a proteome-wide identification of CYP enzymes in the selected nine plants proteome based on conserved heme binding sites (or motif) by using statistical weight matrix approach and prioritized the plants accordingly. The weight matrix approach is a most widely

used method for genome-wide or proteome-wide prediction of conserved motifs. Understanding the architecture of the heme binding site is central to establishing the mechanism by which the basic metabolic machinery assembles and facilitates for the formation of secondary metabolites. Our results revealed that there are numerous other enzymes in plants, which have conserved patterns similar to heme binding sites other than CYP. We assumed that these predicted proteins might have some important CYP like role in the plant cellular metabolism and therefore predicted these uncharacterized or hypothetical proteins as CYP like enzymes, which are not yet characterized experimentally. The identified heme like motifs in the selected plant proteome can act as a novel target for engineering the secondary metabolism. Besides, through these targets, one can also identify the enzymes of secondary metabolic pathways which are not yet functionally annotated and currently submitted in the databases as hypothetical or unknown enzymes. This would enable to reconstruct the secondary metabolic pathways in plants. The derived weight matrix for the heme binding site can be a useful tool in the detection of conserved patterns similar to experimental heme binding sites in other plants. This approach was found to be very efficient and of immense utility towards the identification of CYP enzyme sequences which are yet to be functionally annotated in plants.

#### Acknowledgements

The authors are grateful to the Director, CSIR-Central Institute of Medicinal and Aromatic Plants (CIMAP), Council of Scientific and Industrial Research (CSIR), Lucknow for providing necessary facilities and encouragement. We acknowledge the 'Council of Science & Technology, U.P.' (CST, UP), Lucknow for financial support to Mr. Sarfaraz Alam, Project Assistant-II under GAP247 project at CSIR-CIMAP, Lucknow. We also acknowledge the financial support of the 'Department of Science & Technology' (DST), New Delhi to research fellow Ms. Pooja Sharma and the 'Indian Council of Medical Research' (ICMR), New Delhi to research fellow Mr. Dharmendra K. Yadav at CSIR-CIMAP, Lucknow, India.

#### References

Ayabe SI, Akashi T (2006) Cytochrome P450 in flavonoid metabolism. *Phytochem Rev.* 5:271-282

Beckstette M, Homann R, Giegerich R, Kurtz S (2006) Fast index based algorithms and software for matching position-specific scoring matrices. *BMC bioinformatics.* 7:389

Chaban C, Waller F, Furuya M, Nick P (2003) Auxin responsiveness of a novel cytochrome P450 in rice coleoptiles. *Plant Physiol.* 133:2000-2009

Cheng DW, Lin H, Takahashi Y, Walker MA, Civerolo EL, Stenger DC (2010) Transcriptional regulation of the grape cytochrome P450 monooxygenase gene CYP736B expression in response to *Xylella fastidiosa* infection. *BMC Plant Biol.* 10:135-145

Degtyarenko KN (1995) Structural domains of P450-containing monooxygenase systems. *Protein Eng.* 8:737-747

Degtyarenko KN, Kulikova TA (2001) Evolution of bioinorganic motifs in P450-containing systems. *Biochem Soc T.* 29:139-147

Fischer M., Knoll M, Sirim D, Wagner F, Funke S, Pleiss J (2007) The Cytochrome P450 Engineering Database: a navigation and prediction tool for the cytochrome P450 enzyme family. *Bioinformatics.* 23:2015-2017

Gribskov M, McLachlan M, Eisenberg D (1987) Enzyme Analysis: Detection of Distantly Related Enzymes. *Proc Natl Acad Sci USA.* 84:4355-4358

Henikoff J, Greene E, Pietrokovski S, Henikoff S (2000) Increased Coverage of Enzyme Families with the Blocks Database Servers. *Nucleic Acids Res.* 28:228-230

Kaspera R, Croteau R (2006) Cytochrome P450 oxygenases of Taxol biosynthesis. *Phytochem Rev.* 5:433-444

Khan F, Agrawal S, Mishra BN (2003) Identification of GltC transcription factor binding DNA motifs and its novel co-regulated genes in nitrogen fixing bacteria. *Online J Bioinformatics.* 4:106-114

Khan F, Agrawal S, Mishra BN (2005) Genomic-wide identification of DNA binding motifs of NodD-factor in *Sinorhizobium meliloti* and *Mesorhizobium loti*. *J of Bioinfo Comp Biol.* 3:773-801

Khan F, Agrawal S, Mishra BN (2005) *In silico* Identification of cis-regulatory elements in *Mesorhizobium loti*. *Online J Bioinformatics.* 6:129-141

Khan F, Sharma R, Shukla R, Meena A, Shasany AK, Sharma A (2009) Genomic identification of SinR transcription factor binding sites in nitrogen fixing bacterium *Bradyrhizobium japonicum*. *Open Bioinformatics J.* 10:8-17

Khan F, Singh SP, Mishra BN (2006) Conservation of the regulatory pho-box in acetoacetyl-CoA reductase promoters across bacterial PHB biosynthetic metabolic pathway. *Online J Bioinformatics.* 7:57-68

Khan F, Singh SP, Mishra BN (2008) Conservation of the LexA repressor binding site in *Deinococcus radiodurans*. *J Integr Bioinformatics.* 5:86-98

Kumar S, Jin M, Weemhoff JL (2012) Cytochrome P450-Mediated Phytoremediation using Transgenic Plants: A Need for Engineered Cytochrome P450 Enzymes. *J Pet Environ Biotechnol.* 3:127

Meunier B, de Visser SP, Shaik S (2004) Mechanism of Oxidation Reactions Catalyzed by Cytochrome P450 Enzymes. *Chem Rev.* 104:3947-3980

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

Nebert DW, Russell DW (2002) Clinical importance of the cytochromes P450. *Lancet.* 360:1155-1162

Nelson D, Werck-Reichhart D (2011) A P450-centric view of plant evolution. *Plant J.* 66:194-211

Nelson DR (1999) Cytochrome P450 and the Individuality of Species. *Arch Biochem Biophys.* 369:1-10

Reichhart DW, Feyereisen R (2000) Cytochromes P450: a success story. *Genome Biol.* 6:3003.1-3003.9

Sazegari S, Niazi A (2012) Isolation and molecular characterization of wheat (*Triticum aestivum*) Dehydration Responsive Element Binding Factor (DREB) isoforms. *Aust J Crop Sci.* 6(6):1037-1044

Sen N, Mishra M, Meena A, Khan F, Sharma A (2009) D-MATRIX: A web tool for constructing weight matrix of conserved DNA motifs. *Bioinformation*. 3:415-418

Wu T, Nevill-Manning C, Brutlag D (2000) Fast Probabilistic Analysis of Sequence Function using Scoring Matrices. *Bioinformatics*. 16:233-244