Indian Journal of Biochemistry & Biophysics
Vol. 50, December 2013, pp 537-547

Characterization and expression of codon optimized soybean *phytase* gene in *E. coli*

Pritee Singh, Mansi Punjabi, Monica Jolly, R D Rai and Archana Sachdev*

Division of Biochemistry, Indian Agricultural Research Institute, Pusa Campus, New Delhi 110 012, India

Received 06 March 2013; revised 28 October 2013

Phytic acid, the major storage form of phosphorus in plant seeds is degraded by the phytases to yield inositol and free phosphate, contributing thereby to the improved bioavailability of phytate phosphorus and essential minerals in plant foods and simultaneous reduction in phosphorus pollution of the terrestrial and aquatic ecosystems. As a possible strategy for altering seed phytate levels, the approach involving reduction of phytate content by ectopically expressing endogenous phytase gene during seed development of soybean (*Glycine max* L. *cv.* Pusa-20) was attempted in the present study. Semi-quantitative RT-PCR revealed the maximum expression of *phytase* gene transcripts in germinating cotyledons (~10 days after germinations), compared to other vegetative tissues. A full-length phytase cDNA was amplified from the germinating seedlings by splicing by overlap extension (SOE)-PCR and its sequence analysis revealed an open-reading-frame of 1644 bp, including an N terminal signal peptide of 28 amino acids. Predicted amino acid sequence (547-aa) of molecular mass 62 kDa on alignment with related purple acid phosphatases in other plants shared five conserved domains and seven invariant amino acids involved in coordination of the metals in the binuclear center of purple acid phosphatases. Owing to a large number of *E. coli* low-usage codons in soybean *phytase* gene, the modified gene was cloned into a prokaryotic expression vector pET-28a (+) and its expression in *E. coli* was confirmed by SDS-PAGE and Western blot analysis. Bioassay of the crude expression product in E. coli revealed a functional *phytase* gene, showing a great potential for developing low phytate transgenic soybean through its seed-specific overexpression in the early stages of seed development.

Keywords: Glycine max, *Phytase* gene, SOE-PCR, Prokaryotic expression, *E. coli*

Soybean (*Glycine max* L.) is a global crop well-known for its high protein and oil content. It provides an excellent source of vegetable protein and essential fatty acids for human consumption and is also commonly used in animal feed worldwide. Around 60-80% of the phosphorous in soybean seeds is stored as phytic acid (*myo*-Inositol-1,2,3,4,5,6-hexa*kis*phosphate), a form nutritionally unavailable to monogastric animals due to the lack of hydrolytic phytases in their digestive tract[1]. In addition, phytic acid is a strong chelating agent that binds metal ions, affecting the bioavailability of many essential minerals, such as iron, zinc, calcium and magnesium[2], leading to their potential deficiencies in human beings and animals. Therefore, high-phytate grain-based diets are often feared to exacerbate iron and zinc malnutrition in both the developing countries, where well-balanced diets are not available and people rely mainly on staple foods, such as rice or legumes as well as in developed countries, where livestock is primarily fed with grain-based feed and the excreted phytate contributes to environmental phosphorous pollution by being washed into surface waters[3].

A possible solution to the above-stated problems could be the diet supplementation of phosphorus in the form of monocalcium or dicalcium phosphate or an alternative use of commercial phytase supplements for *in vitro* hydrolysis of phytates, such as microbial phytases[4,5], both of which increase the diet costs. Over the years, several genetic methods have also been exploited, including the production of transgenic plants by expressing a recombinant bacterial[6] or fungal *phytase* gene in their seeds[7,8] or by overexpressing the endogeneous *phytase* gene, such as soybean *phytase* gene (*GmPhy*) in developing cotyledons[9].

Recently, development of transgenic *Brassica napus* with improved soil organic phosphorus utilization and seed phytase activity by overexpression

*Author for correspondence:
Tel: 011-25842038
Mobile: +91-9810683014
E-mail: arcs_bio@yahoo.com

Abbreviations: DAG, days after germination; IPTG, isopropyl β-D-1-thiogalactopyranoside; PAPs, purple acid phosphatases. PBST, phosphate buffered saline tween; RT-PCR, reverse transcriptase polymerase chain reaction; SOE-PCR, splicing by overlap extension polymerase chain reaction.

of *phyA* and *appA* genes[10] has been achieved. Reduction in soybean seed phytic acid phosphorous has also been achieved through silencing (partial) the *myo*-inositol-1-phosphate synthase (MIPS) gene using *RNAi* technology[11], but is found to affect seed development and emergence. Soybean with reduced phytate has also been produced using mutagenesis[12]. However, breeding of high yielding, low phytic acid (*lpa*) soybean varieties has been hindered by the inherent defects of the *lpa* mutations. Therefore, amongst all the possible strategies for altering phytate levels, the approach involving the reduction of phytate by expressing a *phytase* transgene during seed development seems to be the most suitable option.

Use of the endogenous soybean *phytase* gene has distinct advantages over *phytase* genes of foreign origin, such as bacterial and fungal. The soybean *phytase* gene contains the necessary signals for appropriate localization of the enzyme to the protein bodies and moreover optimally suited to digestion of soybean phytate. Thus, in the present study, a full-length SOE-PCR amplified *phytase* cDNA from soybean has been cloned and expressed in *E. coli*. The induced recombinant *E. coli* lysate is found to display higher phytase activity than the non-recombinant control cells, thus making this recombinant *phytase* gene available for use in genetic engineering to improve phosphorous and mineral bioavailability and the overall plant nutrient content.

Materials and Methods

Plant material

Seeds of *Glycine max* L *cv.* Pusa-20 obtained from Division of Genetics, Indian Agricultural Research Institute, New Delhi were used in the study. Following surface sterilization by 0.15% $HgCl_2$ (10 min) and thorough washings with double-distilled water, the seeds were placed on damp paper towels for germination at 25°C in a B.O.D incubator. For detection of expression profile of *phytase* gene in germinated seedlings, samples of various tissues, including cotyledons, leaves, stems and roots were collected at 5, 10, 15 and 20 days after the germination (DAG). The samples obtained at various time points were immediately frozen in liquid nitrogen and stored at -80°C for further use.

RNA Isolation and RT-PCR expression analysis of *GmPhy*

Total RNA from the above-sampled tissues was extracted using TRIzol reagent (Invitrogen, CA, USA). The first strand cDNA synthesis was carried out using 5 µg of total RNA with the help of Revert Aid first strand cDNA synthesis kit (Fermentas, St. Leon-Roth, Germany) according to the manufacturer's protocol. The resulting soybean cDNA was used for semi-quantitative RT-PCR amplification of *phytase* gene to measure its expression pattern in various seedling tissue samples. Polymerase chain reactions (PCR) were performed using 100 ng of cDNA and FP-B and RP-B primers, designed from *phytase* gene (Table 1). PCR program was initiated with a 94°C denaturation for 4 min, followed by 30 cycles of 94°C for 30 s, 57°C for 30 s, 72° C for 40 s.

Table 1—List of primers

Primer	Sequence
For genomic DNA amplification	
FP-A	ATGGCGTCAATTACTTTTTCTCTTC
RP-A	CTGTAACCCAAGATATCCAAACGG
For cDNA amplification	
FP-B	CTCTCTACCTCCCATGACTC
RP-B	TCTCCAACCACTTGTATTGTTC
FP-C	GAACAATACAAGTGGTTGGAGA
RP-C	TTATATCGAAGCAATGCAATCTAT
For amplification of core *phytase* without signal peptide	
FP-D	TGGATCCTGCCATATTCCGTCAACCCT
RP-D	TGAATTCTTATATCGAAGCAATGCAAT
For codon mutagenesis	
FP-M1	TGGATCCTGCCATATTCCGTCAACCCT CGA AGGTCCGTTTGATCCGGTC
RP-M1	AATCTGTTCGGGCTCGAAACCCCGGAC GC GGCGGCGAACACGAGGATC
FP-M2	ACCGTTCCGTTCGACCCGGCCTTGCGC GGC GTCGCCGTCGACTTGCCG
RP-M2	TGAATTCTTATATCGAAGCAATGCAAT C TATGTTCACCCTTTGATG

Isolation of *GmPhy*

Owing to high secondary structures in the isolated RNA and very low expression levels of *phytase* gene, it was difficult to obtain its full-length cDNA. According to the nucleotide sequence (GenBank Accession Nos. AF272346 and EU715238) of *phytase* gene obtained from *G. max*, three pairs of primers were designed (Table 1). Splicing by overlap extension (SOE) PCR was performed with these primers using Phusion DNA polymerase (Finnzymes) with proof-reading activity. At first, primers FP/RP-A, FP/RP-B and FP/RP-C were used to amplify three sets of overlapping fragments — A (277 bp) from genomic DNA, B (810 bp) and C (610 bp)

from cDNA of germinating seedlings (~10 days after germination) through 30 cycles of denaturation (94°C for 30 s), annealing (57°C for 30 s) and extension (72°C for 30 s) with an overlap of 42 bp between A and B and 22 bp between B and C.

The amplified fragments B and C were further used as template for the successive PCR amplification with external primers FP-B and RP-C using the same conditions to produce SOE product BC. Finally, the full-length phytase cDNA-ABC (1644 bp) was amplified using overlapping fragments A and BC as template with external primers FP-A and RP-C, following the above-mentioned amplification conditions. The putative full-length cDNA was cloned via blunt end ligation into the linearized cloning vector pBluescript SK^+ (Stratagene)[13] and the recombinant was analyzed via restriction digestion with EcoRI and HindIII and DNA sequencing.

Characterization of *GmPhy*

The DNA and protein sequence alignments were carried out using the BLASTN and BLASTP programs, respectively (http://www.ncbi.nlm.nih.gov/BLAST/). Identity matrix between different phytase proteins of plants, bacteria and fungi were obtained by BioEdit version 7.0.9.1. The presence of a signal peptide in the deduced amino acid sequence was predicted using SignalP (http://www.cbs.dtu.dk/services/SignalP/) with the sub-cellular location prediction performed by ProtComp version 6.1 of Softberry web server (http://linux1.softberry.com/berry.phtm?topic=protcomppl&group=programs&subgroup=proloc). Conserved domain architecture analysis was conducted using ALL-IN-ONE-SEQ-ANALYZER version 1.36. The secondary structure prediction was done using PSS finder package of Softberry web server (http://linux1.softberry.com/berry.phtml?topic=pps&group=programs&subgroup=propt). Multiple alignments of protein sequences were performed using the ClustalW program (http://www.ebi.ac.uk/clustalW/). A phylogenetic tree was constructed using the Neighbor–Joining (NJ) algorithm in MEGA version 4[14], with its reliability assessed by 500 bootstrap repetitions. Estimation of various physical and chemical parameters of phytase protein was carried out using the ProtParam package of ExPasy web server.

Expression of *GmPhy* in *E. coli* and *phytase* gene activity assay of induced protein codon optimization and PCR amplification of core region of *GmPhy*

The core region of *GmPhy* gene was PCR amplified using degenerate primers FP/RP-D (Table 1) containing BamHI and EcoRI restriction sites, respectively. Point mutations were induced to nullify the effect of rare codons at the 5'-end of core *GmPhy* by optimizing the sequence in accordance with *E. coli* codon usage. Mutations were induced (Table 2) for P(37), P(40), R(62) with primers FP/RP-M1 and P(47) with primers FP/RP-M2 using overlap extension PCR to generate M1 and M2 fragments, respectively[15]. The thermal cycling conditions were optimized to 30 cycles of denaturation (94°C for 30 s), annealing (57°C for 30 s) and extension (72°C for 30 s) for M1 and denaturation (94°C for 30 s), annealing (57°C for 30 s) and extension (72°C for 1min 40 s) for M2. Finally, the full-length core *GmPhy*

Table 2—Codon substitutions in the 5' region of soybean *phytase* cDNA according to *E.coli* codon preference

Amino acid	Codon	Frequency and No. of codons	Amino acid	Codon	Frequency and No. of codons
Phe	UUU	28.9 (109)	Ser	UCU	8.5 (32)
	UUC	18.8 (71)		UCC	8.0 (30)
Leu	UUA	17.5 (66)		UCA	6.1 (23)
	UUG	18.6 (70)		UCG	11.4 (43)
	CUU	12.7 (48)	Pro*	CCU	5.8 (22)
	CUC	14.1 (53)		CCC	2.4 (9)
	CUA	3.4 (13)		CCA	7.4 (28)
	CUG	54.9 (207)		CCG	24.9 (94)
Ile	AUU	33.9 (128)	Thr	ACU	7.7 (29)
	AUC	31.0 (117)		ACC	25.2 (95)
	AUA	5.0 (19)		ACA	6.1 (23)
Met	AUG	37.4 (141)		ACG	14.6 (55)
Val	GUU	19.6 (74)	Ala	GCU	13.8 (52)
	GUC	14.3 (54)		GCC	25.5 (96)
	GUA	10.6 (40)		GCA	19.6 (74)
	GUG	33.9 (128)		GCG	32.6 (126)
Tyr	UAU	18.6 (70)	Cys	UGU	4.2 (16)
	UAC	8.5 (32)		UGC	5.8 (22)
	UAA	1.9 (7)		UGA	0.8 (3)
	UAG	0.3 (1)	Trp	UGG	12.7 (48)
His	CAU	9.3 (35)	Arg*	CGU	16.4 (62)
	CAC	7.2 (27)		CGC	18.8 (71)
Gln	CAA	13.5 (51)		CGA	2.4 (9)
	CAG	24.7 (93)		CGG	5.0 (19)
Asn	AAU	21.2 (80)	Ser	AGU	9.0 (34)
	AAC	15.9 (60)		AGC	14.3 (54)
Lys	AAA	29.2 (110)	Arg	AGA	2.4 (9)
	AAG	8.8 (33)		AGG	2.1 (8)
Asp	GAU	30.0 (113)	Gly	GGU	24.4 (92)
	GAC	15.1 (57)		GGC	33.1 (125)
Glu	GAA	29.4 (111)		GGA	8.2 (31)
	GAG	18.0 (68)		GGG	14.3 (54)

No. of codons are indicated in brackets
*Optimized codons through mutagenesis, Pro (37, 40, 47) & Arg (62).

cDNA fragment with mutated codons was amplified using FP-M1 and RP-M2 primer pair under the following amplification conditions: 30 cycles of denaturation (94°C for 30 s), annealing (57°C for 30 s) and extension (72°C for 1 min 40 s).

Molecular cloning of mutated *GmPhy* to construct a constitutive expression vector

The *GmPhy* fusion protein was inserted into the bacterial expression vector pET-28a to generate pET-GmPhy construct and transformed into the host *E. coli* BL21 (DE3) for expression. The transformants were screened on LA plates supplemented with 50 µg/ml kanamycin and cultured overnight at 37°C taking transformants with plasmid pET-28a as control.

Protein induction and preparation of crude recombinant phytase

The positive transformants harboring recombinant pET-GmPhy expression vector were inoculated in LB broth supplemented with 50 µg/ml kanamycin and cultured by shaking at 37°C, 200 rpm until OD_{600nm} reached 0.7. *GmPhy* was induced by shaking the culture with 1 mM final concentration of isopropyl β-D-1-thiogalactopyranoside (IPTG) for 5 h at 37°C and 200 rpm. The induced cultures were harvested, washed by centrifugation at 8000 x *g*, 4°C for 5 min, suspended in 5 ml of 0.2 M sodium acetate (pH 5) and subjected to six 10 s cycle repeats of ultrasonication using an ultrasonic homogenizer (400 W model) at maximum output in ice. After disruption by ultrasonication and removal of cell debris by centrifugation at 10,000 x *g*, 4°C for 20 min, the supernatant obtained was used as a crude recombinant phytase preparation. The lysate from the transformants with plasmid pET-28a was used as the control.

SDS-PAGE and Western blot analysis

The crude recombinant phytase preparation obtained above was confirmed for the presence of recombinant phytase on a denaturing SDS-PAGE system and analyzed by staining with Coomassie brilliant blue R-250 using SDS molecular weight marker (molecular mass 29-205 kDa; Sigma). To confirm the recombinant *phytase* gene expression as a His-tagged fusion protein, Western blot analysis was carried out using monoclonal mouse-anti-His-Tag antibody (Santa Cruz Biotechnology) as the primary antibody and goat anti-mouse IgG antibody conjugated with horseradish peroxidase (HRP; Santa Cruz Biotechnology) as the secondary antibody. After SDS-PAGE, the proteins were transferred on to a nitrocellulose membrane for 2 h at 25 V. The membrane was blocked with 3% BSA in Tris-buffered saline for 1 h at 37°C and then incubated with the mouse-anti-His-Tag antibody at a dilution of 1:500 for 3 h at 37°C with gentle shaking and washed three-times for 10 min in 0.05% phosphate buffered saline tween (PBST, pH 7.4). Subsequently, the membrane was incubated with goat anti-mouse IgG antibody conjugated with HRP (Santa Cruz Biotechnology) at a dilution of 1:1000 for 40 min at 37°C. After washing in PBST two-times, the membrane was developed using a chromogenic substrate (Sigma Fast Dab Urea tablet sets).

Results

Expression patterns of *GmPhy* in various tissues of seedling

Semi-quantitative RT-PCR analysis used to elucidate the expression levels of *GmPhy* in cotyledons, leaves, stems and roots of the seedlings showed a single band of 800 nucleotides in all the tissues at varying intensities (Fig. 1). Low levels of *GmPhy* transcripts could be detected in cotyledons 5 DAG. At the time point of 10 DAG, the expression level of *GmPhy* in cotyledons was dramatically enhanced[16] (with highest *phytase* RNA levels detected at 8 DAG[17]) and then decreased steadily until a constant low level was reached after 20 days. In comparison, the expression level of *phytase* gene was low in other seedling tissues.

Sequence homology of *GmPhy*

Using SOE-PCR involving both RT-PCR and genomic DNA amplification full-length *phytase* cDNA (1644 bp) encoding 547 amino acids (GenBank Accession No. HM113371) was isolated from *G. max*. The nucleotide sequence analysis using BLAST software at NCBI server revealed 100% homology to previously reported soybean *phytase* gene (GenBank Accession No. AF272346). Further analysis revealed that it shared high degree of homology with *phytase* genes of *Lupinus albus, Medicago truncatula, Vigna radiata* and purple acid phosphatases of *Arabidopsis lyrata, Nicotiana*

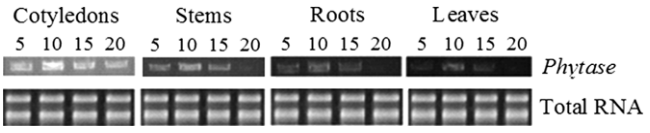


Fig. 1—Expression patterns of *phytase* obtained by RT-PCR in cotyledons, leaves, stems and roots at different time points i.e. 5, 10, 15 and 20 days after germination, respectively

tabacum, Oryza sativa, Triticum aestivum, Zea mays, Hordeum vulgare and *Sorghum bicolor*. Protein sequence comparison with other *phytases* like plant, fungi and bacteria revealed 81.9% identity with *L. albus*, 71.4% with *M. truncatula* and 68% with *V. radiata,* whilst a much lower degree of amino acid identity of 5.3% with *Z. mays*, 7.4% with *Aspergillus ficum*, 8.9% with *A. fumigatus*, 7.6% with *Penicillium oxalicum*, 5.6% with *Bacillus* spp. and 6.7% with *Citrobacter braakii* (Table 3).

Biophysical and biochemical characteristics of *GmPhy*

Sub-cellular analysis indicated the presence of a 28-aa signal peptide at the N-terminal of *GmPhy* with a putative cleavage site located between amino acids 28 and 29 (GHC-HI), indicating that it could be targeted at the cytoplasmic membrane or secreted into the rhizosphere after post-translation modifications (Fig. 2). The predicted molecular mass of *GmPhy* was found to be 62.25 kDa with an isoelectric point of 5.21 and extinction coefficient of 121130 m^{-1} cm^{-1}. Instability index depicted a value of 39.93, suggesting the stable nature of this protein. The hydropathy plot of *GmPhy* protein sequence revealed the presence of about 270 hydrophilic amino acid residues (from amino acid 250 to 520) close to the COOH-terminus, suggesting the hydrophilic nature of the protein, which was also confirmed by the hydropathicity value of -0.264.

```
     MASITFSLLQFHRAPILLLI LLAGFGHCHIPSTLEGPFDPVTVPFDPALRGVAVDLPETD
61   PRVRRRVRGFEPEQISVSLSTSHDSVWISWVTGEFQIGLDIKPLDPKTVS SVVQYGTSRF
121  ELVHEARGQSLIYNQLYPFEGLQNYTSGII HHVQLKGLEPSTLYYYQCGDPSLQAMSDIY
181  YFRTMPISGSKSYPGKVAVVGDLGLTYNTTTTIGHLTSNE PDLLLLIGDVTYANLYLTNG
241  TGSDCYSCSFPLTPIHETYQPRWDYWGRFMQNLVSNVPIMVVEGNHEIEKQAENRTFVAY
301  SSRFAFPSQESGSSSTFYYS FNAGGIHFIMLGAYINYDKTAEQYKWLERDLENVDRSITP
361  WLVVTWHPPWYSSYEAHYREAECMRVEMEDLLYAYGVDII FNGHVHAYERSNRVYNYNLD
421  PCGPVYITVGDGGNREKMAI KFADEPGHCPDPLSTPDPYM GGFCATNFTF GTKVSKFCWD
481  RQPDYSAFRESSFGYGILEV KNETWALWSWYRNQDSYKEVGDQIYIVRQPDICPIHQRVN
541  IDCIASI
```

Fig. 2—Predicted signal peptide by SignalP (http://www.cbs.dtu.dk/services/SignalP/)

***In silico* modeling of the protein**

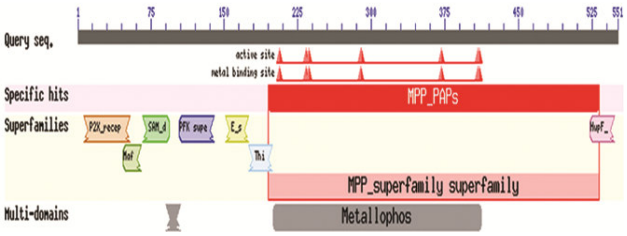
The post-translational modification showed the presence of 6 Asn-glycosylation sites, 11 microbodies binding sites, 4 protein kinase phosphorylation sites and 2 prenyl group binding sites. Conserved domain architecture analysis revealed the presence of purple acid phosphatase and metallophosphoesterase domains at C-terminal (Fig. 3). The secondary structure of phytase protein predicted by using PSS Finder Softberry web server showed a repeating beta and alpha helical structure as β-α-β-α-β loop (Fig. 4).

Phylogenetic analysis of *GmPhy*

The top scoring sequences from BLAST search were used to generate a phylogenetic tree (Fig. 5) by neighbor joining method in the MEGA program version 4. The phylogenetic study showed that *G. max* phytase and *L. albus* phytase protein, Vitis vinifera hypothetical protein and *Arabidopsis thaliana* PAP15 clustured on a single branch of the tree, which was clearly distinct from purple acid phosphatases of other plants such as *O. sativa*, *Z. mays*, *T. aestivum*, *H. vulgare*, *N. tobacum* and *Ricinus comunnis*.

Table 3—Sequence identity matrix showing the sequence identity in 0-1 scale among protein sequences of protein sequences of phytases from plant, fungi and bacteria

	G. max(N)	*G. max*	*Z. mays*	*L. albus*	*V. radi-ata*	*M. truncatula*	*A. ficuum*	*A. fumigatus*	*P. oxalic-um*	*D. caste-llii*	*Y. kristen-senii*	*Bacillus sp.*	*C. amalonat icus*	*Y. pestis*	*C. braakii*
G. max (N)	ID														
G. max	1	ID													
Z. mays	0.05	.03	ID												
L. albus	0.81	0.8	0.06	ID											
V. radiata	0.68	0.6	0.05	0.66	ID										
M.truncatul	0.71	0.71	0.05	0.70	0.63	ID									
A. ficuum	074	07	0.06	0.07	0.08	0.079	ID								
A. fumigats	0.08	08	0.06	0.08	0.09	0.077	0.641	ID							
P. oxalicum	076	07	0.07	0.07	0.07	0.08	0.579	0.585	ID						
D. castellii	0.08	08	0.06	0.07	0.06	0.088	0.152	0.147	0.164	ID					
Y. kristenseni	0.06	06	0.06	0.07	0.06	0.06	0.095	0.104	0.093	0.09	ID				
Bacillus sp.	0.056	06	0.07	0.06	0.06	0.062	0.065	0.073	0.064	0.04	0.05	ID			
C. amalonas	0.053	05	0.06	0.06	0.05	0.049	0.101	0.101	0.100	0.09	0.417	0.066	ID		
Y. pestis	0.070	07	0.05	0.07	0.06	0.066	0.087	0.097	0.089	0.09	0.807	0.055	0.404	ID	
C. braakii	0.067	06	0.07	0.06	0.05	0.053	0.097	0.101	0.089	0.09	0.441	0.066	0.684	0.434	ID



List of domain	Description
P2X_receptor super family	-
Maf_Ham1 super family	Maf, a nucleotide binding protein
SAM_decarbox super family	This is a family of S adenosylmethionine decarboxylase (SAMDC) proenzymes
RhaD_aldol-ADH	-
PFK super family	Phosphofructokinase, catalyze the phosphorylation of fructose-6-phosphate
E_set super family	E or "early" set of sugar utilizing enzymes
Thioredoxin_like super family	Function as protein disulfide oxidoreductases (PDOs), altering the redox state of target proteins via the reversible oxidation of their active site dithiol
MPP_PAPs	Purple acid phosphatases (PAPs) belong to a diverse family of binuclear metallohydrolases
Superfamily Metallophosphatases (MPPs)	Also known as metallophosphoesterases, phosphodiesterases (PDEs), binuclear metallophosphoesterases
Metallophos, pfam00149	Family includes a diverse range of phosphoesterases
HupF_HypC super family	-

http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

Fig. 3—Predicted conserved domains of phytase protein and their function

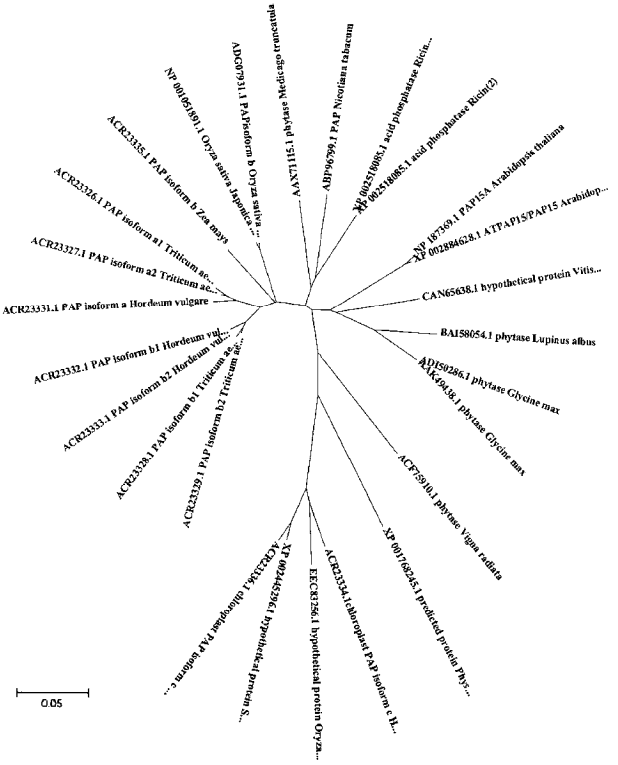


Fig. 5—Phylogenetic tree showing relationship among *phytase* proteins and PAPs

Fig. 4—Predicted secondary structure of phytase protein [The coloured alphabets showing b-a-b-a-b motif important for catalytic activity of enzyme) determined using PSS Finder Softberry web server]

ClustalW analysis of the *phytase* gene sequence and other homologous sequence revealed several regions of high homology shared between them. The presence of five highly conserved sequence motifs (G**D**LG/G**D**VT**Y**/GN**H**E/VTW**H**/G**H**V**H**; bold letters represent metal ligating residues) showed its similarity to purple acid phosphatase (see Supplementary Figure). It is important to mention here that most *G. max* PAP genes are probably found involved in phosphorous acquisition and recycling[18].

Expression of *GmPhy* in *E. coli* and *phytase* gene activity assay of the induced protein

The codon optimized core *GmPhy* region without signal peptide was cloned into the expression vector pET-28a and the recombinant plasmid pET-GmPhy (Figs 6,7) containing *GmPhy* gene was transformed into *E. coli* BL21 (DE3). The positive BL21 (DE3)/pET-GmPhy transformant obtained was induced by IPTG. After disruption by ultrasonication and removal of cell debris of the induced cells, the

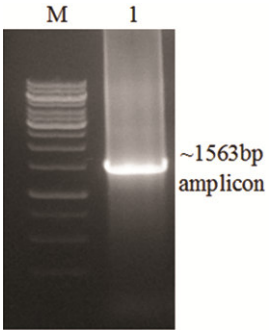


Fig. 6—Electrophoresis of PCR amplified core *phytase* on 1% agarose gel [Lane M, 1 kb DNA Ladder (Fermentas); and lane 1, ~1563 bp core *phytase* amplicon]

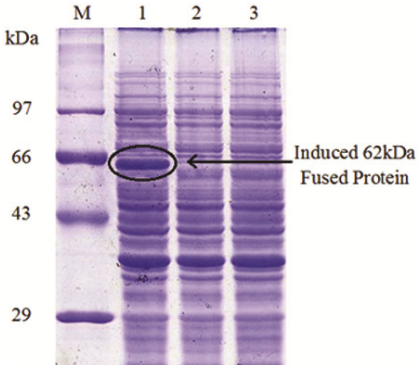


Fig. 8—SDS-PAGE analysis of *GmPhy* expressed in *E. coli* BL21 [Lane M, Protein Mr marker (Fermentas); lane 1, IPTG-induced recombinant pET-GmPhy fusion protein indicated by a strong band; lane 2, crude extract from uninduced recombinant pET-GmPhy vector showing no band of expression; and lane 3, crude extract from empty pET-28a transformed cells taken as control]

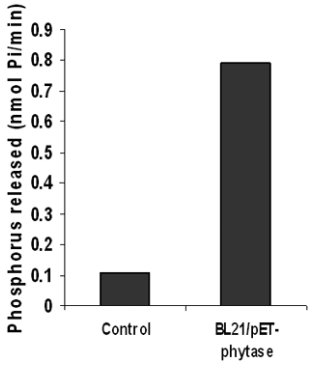


Fig. 10—Graph showing about 7.2-folds higher Pi release from phytate in induced cells of recombinant BL21/ pET GmPhy as compared to control

crude recombinant phytase activity in the supernatant obtained was determined and the cell-free extract from the transformants with only plasmid pET-28a was used as the control.

The results of SDS-PAGE analysis (Fig. 8) indicated that cell-free extracts from the induced cells

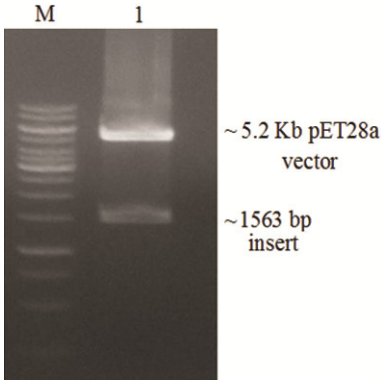


Fig. 7—Electrophoresis of recombinant pET-GmPhy plasmid following BamHI and HindIII restriction digestion on a 1% agarose gel [Lane M, 1 kb DNA ladder; lane 1, ~1563 bp *GmPhy* insert]

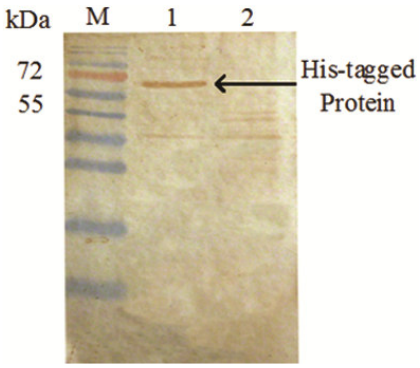


Fig. 9—Western blot analysis [Lane M, PageRuler prestained protein ladder; lane1, Western of induced *GmPhy* with anti-His antibody; and lane 2, control host]

harboring recombinant pET-GmPhy exhibited one specific band with a molecular mass of about 62 kDa, corresponding to the calculated value with no specific banding in the controls. Western blot analysis using anti-His antibody confirmed the recombinant expression of *phytase* gene as a His-tagged fusion protein (Fig. 9). The uninduced transformant and the control showed no bands of induction. The phytase activity assay in the induced cells of BL21 (DE3)/pET-GmPhy was found to be 7.2-fold higher than in the control (Fig. 10).

Discussion

Amongst the diverse technologies and approaches available to alter seed phytate and hence total phosphorous content, the most promising approach is to engineer crops to express high levels of *phytase* gene in seeds[19]. In a previous transgenic approach, fungal *phytase* gene was utilized for plant tranformation[11], which due to its distant evolutionary

relationship was preferred over by *phytase* gene from plant origin[20], therefore making the *phytase* gene from *G. max* in our current study a good candidate for molecular improvement of phosphorous acquisition in plants.

In the first step of our study, *phytase* gene was isolated and its mRNA levels were examined in different tissues (cotyledons, leaves, stems and roots) at different time intervals by RT-PCR using *phytase* gene-specific primers. The phytase expression in cotyledons was induced as early as 5 DAG, reaching maximum levels at 10 DAG until a constant low was observed after 20 DAG. This suggested its role in hydrolysis of phytic acid and its derivatives stored in seeds during the seed germination and the early growth stage of seedlings.

For expression cloning of *GmPhy*, attempts were made to amplify the full-length *phytase* cDNA from 10 days old cotyledons in a single step by using gene-specific primers FP-A and RP-C designed from the terminal ends of *phytase* cDNA, but no amplification could be obtained, possibly due to the low levels of expression of the endogenous *phytase* and apparent secondary structure near the 5' end of the mRNA, as evident from RT-PCR analysis. Thus, in order to generate full-length *phytase* cDNA, the SOE-PCR approach was used[20]. Compared with traditional methods for isolating new genes, including genomic library construction and screening, direct protein separation etc, the SOE-PCR method is a much easier, direct and a more sensitive method for target region amplification.

Previously, *phytase* cDNA was isolated from *G. max* by rapid amplification of cDNA ends (RACE) technique using peptide sequence data generated from the purified enzyme[20] and from *Z. mays* by screening a maize seedling cDNA expression library[21]. In the present study, full-length *phytase* cDNA (1644 bps) was isolated using three rounds of PCR. To circumvent apparent secondary structures near the 5'-end of the cDNA, a small 5' portion of the coding sequence was amplified from genomic DNA. Structural analysis of the sequence revealed 100% identity to previously reported soybean *phytase* gene (Accession number: AF272346), sharing similar features, such as length of signal sequence at the amino terminus[20]. *G. max* phytase protein showed high degree of homology with PAPs of different plant sources, *Arabidopsis lyrata, N. tabacum, O. sativa, T. aestivum, Z. mays, H. vulgare* and *S. bicolor,* which was also confirmed by phylogenetic tree analysis (figure not shown). It shared 81.92% similarity with *L. albus phytase*, 71.4% with *Medicago truncatula phytase*, and 68% with *Vigna radiata phytase*. On the other hand, it showed least similarity (5-8%) with *Z. mays* and other fungal and bacterial *phytase*s, which were classified as histidine acid phosphatases and contained the amino acid motif RHGXRXP[21,22]. ClustalW alignment of highly homologous sequences revealed that soybean phytase protein contained five conserved amino acid motifs and seven invariant amino acids (G**D**LG/G**D**VT**Y**/G**N**HE/VTW**H**/G**H**V**H**; bold letters representing metal-ligating residues) involved in coordination of the metals in the binuclear center, a distinct characteristic of PAPs[23].

In order to express *phytase* gene in *E. coli*, the ORF without 28-aa signal peptide sequence at the N terminal of *GmPhy* was cloned downstream of IPTG-inducible T7 promoter in the expression vector pET-28a and transformed to the host *E. coli* BL21 (DE3) having T7 RNA polymerase derived from the bacteriophage integrated into the bacterial chromosome. Initial experiments of expression in *E. coli* host were a failure due to codon usage bias between *GmPhy* coding sequence and *E. coli* host[24]. Construction of a system capable of overexpressing soybean *phytase* gene in *E. coli* required coding sequence optimization. As most low usage codons are found within the first 25 codons in *E. coli*[25], therefore, initial 35 codons were optimized according to *E. coli* codon usage.

In most genes, an exchange of low usage codons with more optimal codons increases the yield of expression[26]. So, mutation of the rare codons CCCpro (37), CCCpro (40), CCCpro (47) and CGAarg (62) in the 5′ end of *phytase* gene to their common counterparts CCGpro and CGTarg was done, which promoted the expression of soybean *phytase* gene in *E. coli* BL21(DE3). The IPTG-induced recombinant pET-GmPhy was confirmed and characterized by SDS-PAGE and Western analysis. The molecular weight of expressed phytase protein was less than that of previously reported natural phytase from *G. max*[20]. This difference in molecular mass might be due to the difference in glycosylation. In prokaryotes, glycosylation is rare[27] which might have led to low molecular weight of phytase expressed in *E. coli* host. A total of six putative N-glycosylation sites were present in the primary structure of soybean phytase. In general, additional glycosylation of protein has

been known to facilitate folding and increase stability[28]. Phytic acid bioassay of crude expression product expressed in *E. coli* revealed 7.2-fold increase in phytase activity in the induced cells than that in the uninduced control. The functional *phytase* gene product thus isolated was capable of catalyzing the hydrolysis of phytic acid, making it potentially useful for low phytate transgenic development. Isolation and manipulation of the soybean *phytase* gene might further provide the opportunity to alter seed phytic acid metabolism.

References

1 Wodzinski R J & Ullah A H (1996) *Adv Appl Microbiol* 42, 263-302

2 Brown K H & Solomons N W (1991) *Infect Dis Clin North Am* 5, 297-317

3 Lott J N A, Ockenden I, Raboy V & Batten G D (2000) *Seed Sci Res* 10, 11-33

4 Jongbloed A W, Kemme P A, Morz Z, & Van Diepen H T M (2000) *Proceedings of the Sixteenth Alltech Anuual Symposium*, Nottingham University Press, Nottingham, U.K. pp. 111-129

5 Almeida F N & Stein H H (2010) *J Anim Sci* 88, 2968-2977

6 Bilyeu K D, Zeng P, Coello P, Zhang Z J, Krishnan H B, Bailey A, Beuselinck P R & Polacco J C (2008) *Plant Physiol* 146, 2468-2477

7 Chen R, Xue G, Chen P, Yao B, Yang W, Ma Q, Fan Y, Zhao Z, Tarczynski M C & Shi J (2006) *Transgenic Res* 17, 633-643

8 Denbow D M, Grabau E A, Lacy G H, Kornegay E T, Russell D R & Umbeck P F (1998) *Poult Sci* 77, 878-881

9 Chiera J M, Finer J J & Grabau E A (2005) *Plant Mol Biol* 56, 895-904

10 Wang Y, Ye X, Ding G & Xu F (2013) *PLoS ONE* 8(4), e60801. doi:10.1371/journal.pone.0060801

11 Nunes A C S, Vianna G R, Cuneo F, Amaya-Farfán J, de Capdeville G, Rech E L & Aragão F J L (2006) *Planta* 224, 125-132

12 Hitz W D, Carlson T J, Kerr P S & Sebastian S A (2002) *Plant Physiol* 128, 650-660

13 Sambrook J & Russel D W (2001) *Molecular Cloning: A Laboratory Manual,* 3rd edn, Vol. 3, Cold Spring Harbor Laboratory Press, NY, USA

14 Tamura K, Dudley J, Nei M & Kumar S (2007) *Mol Biol Evol* 24, 1596-1599

15 Ge L & Rudolph P (1997) *Biotechniques* 22, 28-30

16 Gibson D M & Ullah A H J (1988) *Arch Biochem Biophys* 260, 503–513

17 Chaudhary V N & Sachdev A (2010) *Proceedings International Conference on Chemical Engineering and Applications*, Singapore indexed by Thomson ISI, British Library and Nelson

18 Li1 Ch, Gui Sh, Yang T, Walk Th, Wang Xi & Liao H (2012) *Annals Bot* 109, 275–285

19 Raboy V (2009) *Plant Sci* 177, 281-296

20 Hegeman C E & Grabau E A (2001) *Plant Physiol* 126, 1598-1608

21 Maugenest S, Martinez I & Lescure A M (1997) *Biochem J* 322,511-517

22 Gu W, Huang H, Meng K, Yang P, Fu D, Luo H, Wang Y, Yao B & Zhan Z (2009) *Appl Microbiol Biotechnol* 157, 113–123

23 Li D, Zhu H, Liu K, Liu X, Leggewie G, Udvardi M & Wang D (2002) *J Biol Chem* 277, 27772-27781

24 Makrides S C (1996) *Microbiol Mol Biol Rev* 60, 512-538

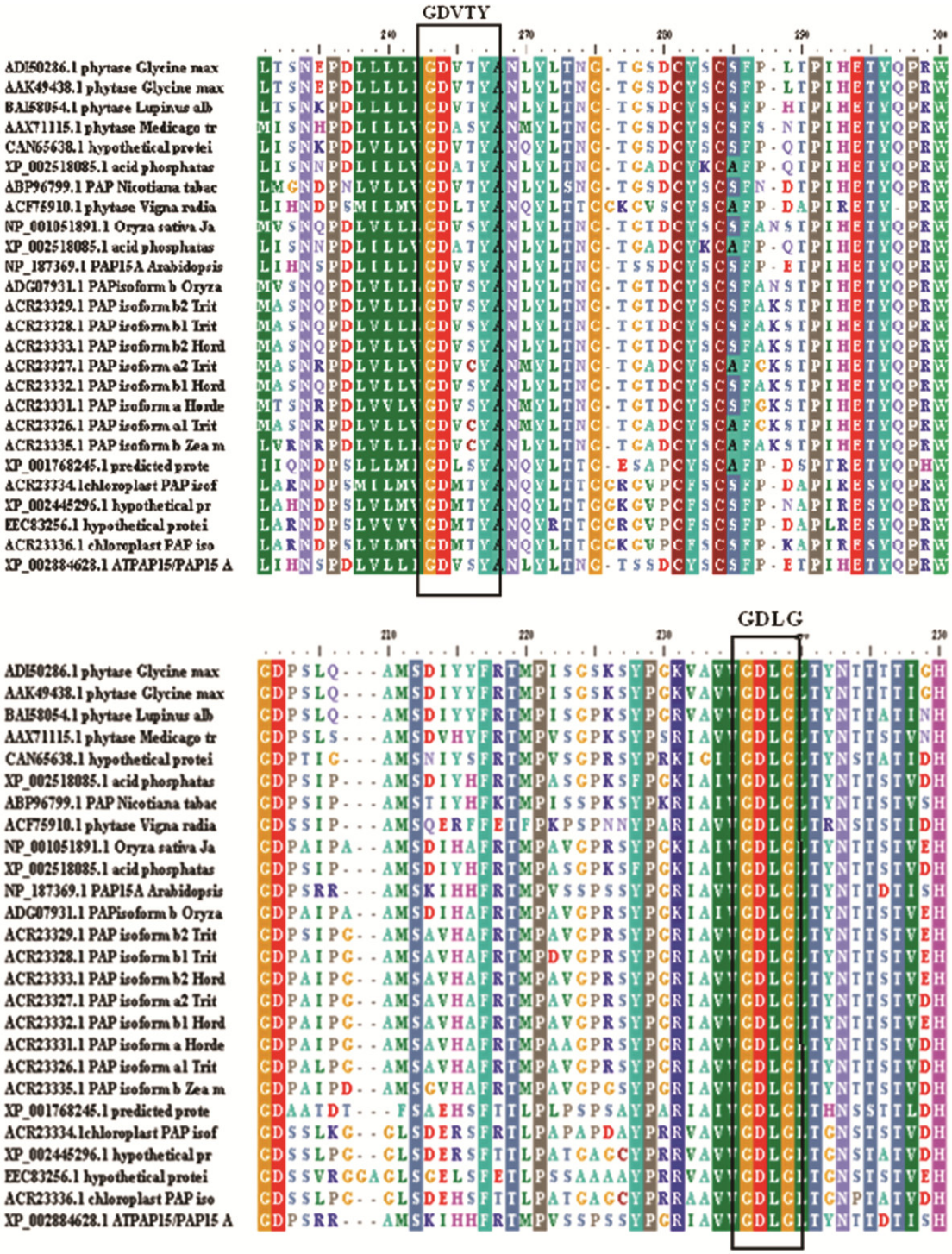
25 Chen G T & Inouye M (1994) *Genes Dev* 8, 2641-2652

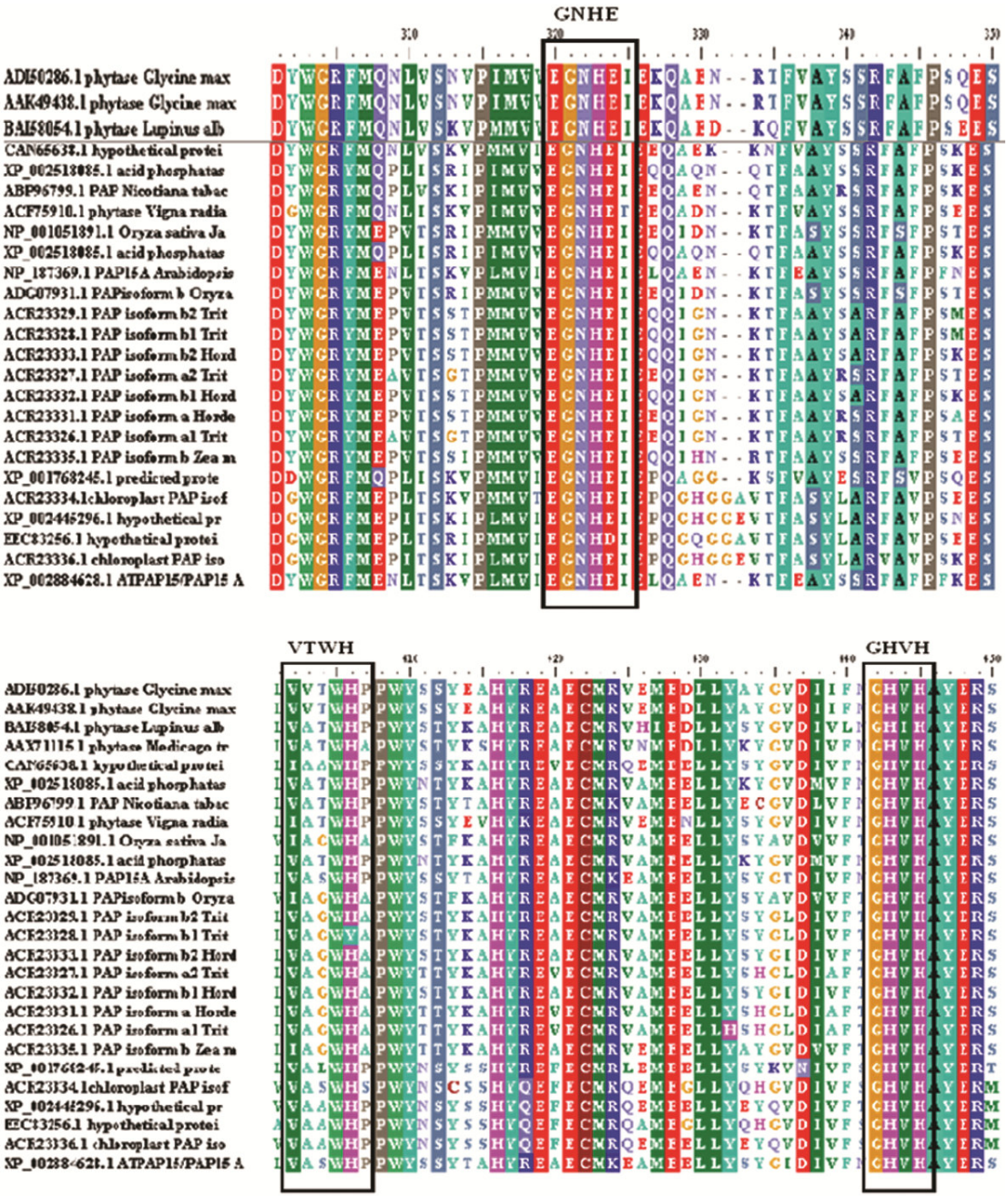
26 Rangwala S H, Finn R F, Smith C E, Berberich S A, Salsgiver W J, Stallings W C, Glover G L & Olins P O (2003) *J Neurosci Methods* 124, 75-81

27 Sherlock O, Dobrindt U, Jensen J B, Munk V R & Klemm P (2006) *J Bacteriol* 188, 1798-1807

28 Haraguchi M, Yamashiro S, Furukawa K, Takamiya K, Shiku H & Furukawa K (1995) *Biochem J* 312, 273-280

Supplementary Figure—ClustalW analysis of deduced amino acid sequences of various *phytase*s and purple acid phosphatases like sequences showing five conserved motifs (G**D**LG/G**D**VT**Y**/G**N**HE/VTW**H**/G**H**V**H**; bold letters represent metal ligating residues)




GNHE
VTWH
GHVH
ADI50286.1 phytase Glycine max
AAK49438.1 phytase Glycine max
BAI58054.1 phytase Lupinus alb
CAN65638.1 hypothetical protei
XP_002518085.1 acid phosphatas
ABP96799.1 PAP Nicotiana tabac
ACF75910.1 phytase Vigna radia
NP_001051891.1 Oryza sativa Ja
XP_002518085.1 acid phosphatas
NP_187369.1 PAP15A Arabidopsis
ADG07931.1 PAPisoform b Oryza
ACR23329.1 PAP isoform b2 Trit
ACR23328.1 PAP isoform b1 Trit
ACR23333.1 PAP isoform b2 Hord
ACR23327.1 PAP isoform a2 Trit
ACR23332.1 PAP isoform b1 Hord
ACR23331.1 PAP isoform a Horde
ACR23326.1 PAP isoform a1 Trit
ACR23335.1 PAP isoform b Zea m
XP_001768245.1 predicted prote
ACR23334.1chloroplast PAP isof
XP_002446296.1 hypothetical pr
EEC83256.1 hypothetical protei
ACR23336.1 chloroplast PAP iso
XP_002884628.1 ATPAP15/PAP15 A
AAX71115.1 phytase Medicago tr