Molecular characterization, modeling, and docking analysis of late phytic acid biosynthesis pathway gene, *inositol polyphosphate 6-/3-/5-kinase*, a potential candidate for developing low phytate crops

Mansi Punjabi^{1,2} · Navneeta Bharadvaja¹ · Archana Sachdev² · Veda Krishnan²

Received: 8 February 2018 / Accepted: 6 July 2018 / Published online: 28 July 2018
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract

The coding sequence of *inositol polyphosphate 6-/3-/5-kinase* (*GmIPK2*) gene was identified and cloned from popular Indian soybean cultivar Pusa-16. The clone was predicted to encode 279 amino acids long, 30.97 kDa protein. Multiple sequence alignment revealed an inositol phosphate-binding motif, PxxxDxKxG throughout the IPK2 sequences along with other motifs unique to inositol phosphate kinase superfamily. Eight α -helices and eight β -strands in antiparallel β -sheets arrangement were predicted in the secondary structure of *GmIPK2*. The temporal analysis of *GmIPK2* revealed maximum expression in the seed tissues during later stages of development while spatially the transcript levels were lowest in leaf and stem tissues. Endosperm-specific cis-regulatory motifs (GCN4 and Skn_1) which support high levels of expression, as observed in the developing seeds, were detected in its promoter region. The protein structure of *GmIPK2* was modeled based on the crystal structure of inositol polyphosphate multikinase from *Arabidopsis thaliana* (PDB:4FRF) and subsequently docked with inositol phosphate ligands (PDB: 5GUG-I3P and PDB: 4A69-I0P). Molecular dynamics (MD) simulation established the structural stability of both, modeled enzyme and ligand-bound complexes. Docking in combination with trajectory analysis for 50 ns MD run confirmed the participation of Lys105, Lys126 and Arg153 residues in the formation of a network of hydrogen bonds to stabilize the ligand-receptor interaction. Results of the present study thus provide valuable information on structural and functional aspects of *GmIPK2* which shall assist in strategizing our long-term goal of achieving phytic acid reduction in soybean by genetic modification of its biosynthetic pathway to develop a nutritionally enhanced crop in the future.

Keywords Glycine max · Phytic acid · Inositol polyphosphate 6-/3-/5-kinase (IPK2) · Low phytate crops, spatiotemporal expression · Homology modeling · Molecular docking · Molecular dynamics simulation

Introduction

Soybean is a phenomenal crop with a unique nutrient profile. It is the only plant source which provides a complete protein with quality equivalent to animal protein yet its human

consumption is found to be very limited, even in the countries where its production is very high. Besides, majority of its uses in the food industry owe primarily to its functional properties rather than nutritional gain. The above stated can be attributed to the presence of a variety of bioactive components including phytic acid (PA), saponins, protease inhibitors, isoflavones, lectins, etc. Amongst these, PA is probably the most well-known antinutrient that is found in abundance in cereal grains and legumes. It is a saturated cyclic acid whose metabolism has been substantiated to regulate phosphorous and myo-inositol homeostasis. Due to its highly negative chemical structure, it forms phytate–mineral–protein complexes which result in reduced bioavailability of phosphorus, other associated minerals and proteins and thus contribute to the debatable health effects associated with its consumption. Therefore, in a need to reduce the level of this compound several approaches were undertaken which blocked its constitutive

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s13205-018-1343-7>) contains supplementary material, which is available to authorized users.

✉ Archana Sachdev
arcs_bio@yahoo.com

¹ Department of Biotechnology, Delhi Technological University (Formerly Delhi College of Engineering), New Delhi 110042, India

² Division of Biochemistry, Indian Agricultural Research Institute, New Delhi 110012, India

synthesis, but often resulted in crop with deplorable agronomic performance (Bilyeu et al. 2008; Cichy and Raboy 2009; Feng and Yoshida 2004; Kuwano et al. 2009; Nunes et al. 2006). Recently trending reverse genetic approaches that target tissue-specific knockdown provide a promising alternative to eliminate this negative impact (Ali et al. 2013a, b). Research is thus underway to characterize PA biosynthetic pathway enzymes (Josefsen et al. 2007; Krishnan et al. 2015; Stiles et al. 2008; Sun et al. 2007; Sweetman et al. 2006, 2007) and study their expression patterns (Bhati et al. 2014; Fileppi et al. 2010; Suzuki et al. 2007) to generate adequate information for developing *lpa* soybean through metabolic engineering.

PA biosynthesis follows a succession of phosphorylation steps (Brearley and Hanke 1996a, b). One of its late pathway enzyme inositol polyphosphate 6-/3-/5-kinase (IPK2, EC 2.7.1.151) is a multiple specificity enzyme which phosphorylates the D-6, D-3, and/or D-5 positions on a variety of inositol polyphosphate substrates (InsP3/InsP4/InsP5) (Frederick et al. 2005) that makes it a key enzyme in regulating PA turnover and is thus a crucial target for perturbing PA dynamics (Stevenson-Paulik et al. 2005). The enzyme has previously been studied in *Arabidopsis* (Stevenson-Paulik et al. 2002), yeast (Saiardi et al. 2000; York et al. 1999), mouse (Frederick et al. 2005) and human (Majerus 1992, 1996). A study has also been reported in soybean (Stiles 2007), however, it discusses only the gene's expression and enzyme kinetics. Therefore, with the intention to extend the current knowledge on its transcript expression (both spatial and temporal) as well as to understand its structure, biochemical features and evolutionary history, we investigated this gene in soybean and compared it with those of another 19 homologs in plants through various bio-computational tools. Furthermore, to elucidate complete molecular and biochemical mechanisms regulating PA synthesis as well as inositol metabolism and signalling in plants, an insight into IPK2 protein's structural features is required. Very less information regarding the three-dimensional structure of plant IPK2 is available. To facilitate the same, we determined the three-dimensional (3D) model of soybean *GmIPK2* protein through homology modeling and performed its docking simulation with the substrate [1D-*myo*-inositol 1,4,5-trisphosphate (PDB: 5GUG-I3P) and 1D-*myo*-inositol-1,4,5,6-tetrakisphosphate (PDB: 4A69-I0P)] molecules.

Through this work, we made an effort to study *IPK2* gene to lay a groundwork for achieving the main goal of our work which is, to develop low phytic acid soybean, a nutritionally and agriculturally important crop.

Materials and methods

Plant material

Field grown soybeans (*Glycine max* [L.] Merr. cv. Pusa-16) were procured from Division of Genetics, IARI, New Delhi for use in this study. For spatial and temporal expression profiling, root, stem, leaf, and flower tissues were collected from 30-day-old plants while developing seeds were collected regularly after flowering until maturation and sorted based on their sizes. After collection, the tissue samples were immediately frozen in liquid nitrogen and stored at -80°C until used.

PCR amplification, cloning and sequencing of partial *GmIPK2* sequence

Total RNA was isolated from soybean seeds 8 mm in size using TRIzol reagent method (Invitrogen, USA). Approximately, 1 μg of isolated total RNA was used for single stranded cDNA synthesis by reverse transcription with RevertAid™ H Minus First Strand cDNA synthesis kit (ThermoFisher Scientific, USA). Prior to cDNA synthesis, the RNA was treated with RNase-free DNase I (Thermo Scientific, USA) for removal of any residual genomic DNA. Polymerase chain reaction (PCR) amplification was performed using 0.2 μg of the initial RT reaction optimized at 94°C for 4 min; 35 cycles of 94°C for 30 s, 62.5°C for 30 s, 72°C for 30 s; 72°C for 10 min, using the *GmIPK2* specific primers: *IPK2F* 5'-ATGCTCAAGATCCGGAG-3' and *IPK2R* 5'-CAGTTAGTCTGCGACACTAATTCAAGC-3'. On completion, 2 μl of amplified product was analyzed by electrophoresis on 1% agarose gel and purified using gel/PCR DNA Fragments Extraction Kit (DF100) by Geneaid. The purified amplicon was subsequently ligated in pGEM®-T Easy vector (Promega, USA) by TA cloning following protocol described in the manual and used directly for transformation of competent *Escherichia coli* DH5 α cells. The recombinant plasmids were identified by blue-white screening and confirmed by restriction digestion with *EcoRI* to release the amplicon fragment. The nucleotide sequences of *GmIPK2* thus isolated was determined by DNA sequence analysis of overlapping plasmid clones using universal primers (SP6 and T7) on an automated sequencer (ABI 3730xl DNA Analyzer, USA). The nucleotide sequence data was submitted to INSDC database GenBank.

Gene expression analysis by semi-quantitative reverse transcription PCR and quantitative real-time PCR

To analyze *GmIPK2* gene expression in different soybean tissues, we first performed semi-quantitative reverse transcription PCR (RT-PCR) to obtain an expression pattern

and then further estimated the transcript levels by quantitative real-time PCR (qRT-PCR). For RT-PCR, first strand cDNA was synthesized from total RNA using RevertAid™ H Minus First strand cDNA Synthesis Kit (Thermo Scientific, USA) and PCR was subsequently performed using the same pair of *GmIPK2* specific primers and thermal cycling conditions which were described for its cloning. We further monitored the quantitative amplification of *GmIPK2* by performing qRT-PCR analysis on a PikoReal 96 Real-Time PCR platform (ThermoFisher Scientific, USA). The reactions were set up using DyNamo Flash SYBR Green qPCR Kit (Thermo Scientific, USA) with cDNA first strands as the template DNA. The expression of *GmIPK2* was normalized to an endogenous control, the housekeeping gene phosphoenolpyruvate carboxylase (*PEPCo*) (Sugimoto et al. 1992; Tuteja et al. 2004). The primers for the gene (*qIPK2F* 5'-CGCGGATCCGCGTTGCAGAAGCTCAAG-3' and *qIPK2R* 5'-TCCCCGCGGGGAGCGACACTAATCAAG-3') and the internal control (*qPEPCoF* 5'-CATGCCAAAGGGTGTGTTT-3' and *qPEPCoR* 5'-TTTTCGCGCAGCTATCTCTC-3') were designed using PrimerQuest tool by IDT, USA. The reactions were setup following the standard protocol provided in DyNamo ColorFlash SYBR Green qPCR Kit (Thermo Scientific). The thermal profile used for PCR amplification was: 95 °C for 4 min; 40 cycles of 95 °C for 15 s, 60 °C for 30 s and fluorescence data collection. To minimize variation in the output, three technical replicates were carried out for each of the three biological replicates. The baseline data was collected for first 15 cycles to generate a baseline-subtracted plot of the logarithmic increase in fluorescence signal (ΔR_n) versus cycle number. A standard fluorescence threshold (R_n) was set to 0.5 on the log fluorescence scale to determine the fractional cycle number (C_t value). The relative abundance of *GmIPK2* was calculated using the $2^{-\Delta\Delta C_t}$ method (Livak and Schmittgen 2001). Dissociation curve analysis from 60 to 95 °C was also performed at the end of the assay to check for any non-specific amplification and/or contamination.

Sequence analysis and phylogenetic tree construction

Homologous *IPK2* sequences from other plants were identified with National Center for Biotechnology Information (NCBI) Basic Local Alignment Search Tool (Protein BLAST, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>) using the above deduced *GmIPK2* sequence as a query. The primary sequence composition of *IPK2* sequences were computed using PEPSTATS (Rice et al. 2000). The physio-chemical parameters of proteins were predicted using the ProtParam tool of ExPASy web server (Gasteiger et al. 2005). TargetP 1.1 Server using the cutoffs of 95% specificity was implemented for subcellular location prediction of *GmIPK2*

protein (Emanuelsson et al. 2000) and the outcome was compared to predictions obtained from MemType-2L (Chou and Shen 2007), WoLF PSORT (Horton et al. 2007), SubLoc v1.0 (Hua and Sun 2001) and CELLO v2.5 (Yu et al. 2006). Its transmembrane topology was predicted with PSIPred (Buchan et al. 2013), TMPred (Hofmann and Stoffel 1993) and NPS@ web programs (Rost et al. 1996). A final consensus was drawn manually and the topology was generated and visualized using Protter version 1.0 (Omasits et al. 2014). Presence of potential secretory signal peptides or mitochondrial targeting peptides was analyzed with SignalP 4.1 web server (Petersen et al. 2011). M-Coffee multiple sequence alignment (MSA) of the selected amino acid sequences was carried out to produce quality alignments which served as the basis for phylogenetic analysis (Wallace et al. 2006) to detect its evolutionary placement and phylogenetic similarity with other similar genes. The evolutionary tree was constructed using neighbor joining (NJ) clustering method to compute distances and poisson model for amino acid substitution in MEGA Version 6.0 (Molecular Evolutionary Genetic Analysis, Tamura et al. 2013). Bootstrap replications were set at 1000 to assess the degree of confidence for each clade of the observed tree. The final image was rendered with the Interactive Tree of Life server (iTOL) (Letunic and Bork 2016).

Promoter isolation and prediction of regulatory motifs

A motif search was carried out to define putative cis-elements in the promoter sequences involved in the regulation of *IPK2* expression using PlantCARE (Lescot et al. 2002). To identify these cis-regulatory elements, around 2 kb upstream sequence of the *IPK2* homologs were retrieved by NCBI's nucleotide BLAST program and fed to PlantCARE web tool. The tool's database identified regulatory elements in the isolated upstream sequences and each of the elements observed were analyzed with previously reported properties of the particular element. Unlike CDS, the regulatory sequences indirectly influence their immediate phenotype.

Secondary structure analysis and domain prediction

M-Coffee multiple sequence alignment (MSA) of the selected amino acid sequences was carried out and the alignment file was imported to Jalview (Waterhouse et al. 2009) to identify and shade the conserved amino acid sequences. Ungapped motifs were also detected using MEME web tool available on MEME suite 4.11.1 (Bailey et al. 2009). The motifs present were further verified using My Hits motif scan tool (Pagni et al. 2007). The domain composition was analyzed using CDD tool (<http://www.ncbi.nlm.nih.gov/cdd>) on NCBI server. A secondary structure consensus from

amino acid sequences was built based on the joint prediction with SOPMA (nearest-neighbor method) and PHD (neural networks method) correctly predicting 82.2% of the residues for 74% of co-predicted amino acids (https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_seccons.html) (Geourjon and Deléage 1995; Rost and Sander 1993). Cysteine species and disulfide connectivity of protein sequences were determined using web tool DiANNA (Ferrè and Clote 2005). A secondary structure topology map of the 3D model was generated using ProMotif (<http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/Generate.html>) and rendered using TopDraw (Bond 2003).

Homology modeling and quality assessment of predicted model

The 3D model of GmIPK2 was constructed by template-based homology modeling using automated comparative protein modeling servers, SWISS-MODEL (Biasini et al. 2014) and PHYRE2 (Kelley et al. 2015) as well as a standalone comparative modeling program MODELLER 9.16 (Webb and Sali 2016). Comparative modeling consists of five main steps: search for related protein structures, selection of one or more appropriate templates, target-template alignment, model building and model evaluation. The final step of comparative modeling was additionally performed with RAMPAGE (Lovell et al. 2003), VERIFY 3D (Eisenberg et al. 1997) and ProSA servers (Wiederstein and Sippl 2007) to evaluate the stereochemical and energetic properties of the obtained models. In addition to this, the quality of models can also be assessed by structural comparison to the template using the MatchMaker tool in UCSF Chimera (Pettersen et al. 2004), a molecular visualization software package and calculating the C α root mean square deviation scores (RMSDs) for each of the comparative models. The secondary structures of the final and template proteins were also compared by pairwise 3D alignment using MATRAS 1.2 (Kawabata 2003). Agreement on the best model was made on the basis of the majority in best scores from these different quality analyses.

Refinement of the predicted homology model

Molecular dynamics (MD) simulation was performed to optimize the obtained GmIPK2 model. Simulation of the model was conducted in explicit solvent using the GROMACS (Groningen Machine for Chemical Simulations) 4.5.5 package (Pronk et al. 2013). The model was solvated with simple point charge (SPC216) water in a cubic box with edges that were 0.7 nm from the molecular boundary. Initially, energy minimization (maximum number of steps: 1000) was performed to remove steric conflicts between the protein and water molecules, using the steepest descent

integrator. The system was then equilibrated by optimizing the solvent molecules surrounding the energy-minimized model with NVT (constant Number of particles, Volume, and Temperature) and NPT (constant Number of particles, Pressure, and Temperature) ensembles using Berendsen thermostat and Parrinello-Rahman barostat, respectively. Finally, the system was simulated for 50 ns thrice maintaining the same temperature (300 K) and pressure (1 bar) using the Particle Mesh Ewald (PME) electrostatics method. After completion of production run, we collected the data using a post-processing tool, trjconv (strips out coordinates and correct for periodicity) and used this corrected trajectory as the input for studying the conformational stability of our simulated protein via *g_rms* and *g_rmsf* tools. We subsequently generated and analyzed the output plots using a simple plotting program, xmgrace (Turner 2005).

Active site prediction, molecular docking and MD simulation of the docked complexes

The refined protein model thus generated was docked to characterize the 3D structure of the complex, to gain an insight into the crucial amino acid residues also referred to as active residues that are involved in the complex formation. Based on the available literature for IPK2, 1D-*myo*-inositol-1,4,5-trisphosphate (PDB: 5GUG-I3P) and 1D-*myo*-inositol-1,4,5,6-tetrakisphosphate (PDB: 4A69-I0P) ligands were selected for optimized GmIPK2 protein receptor. Docking of these ligands was performed using AutoDock Vina (version 1.1.2) (Trott and Olson 2010) following the semi-flexible approach of docking. AutoDock Vina reads all the molecules in a simplified PDB file representation, termed PDBQT and thus the coordinates of both GmIPK2 protein and its ligands (downloaded from the RCSB protein data bank) were prepared using MGL tools (version 1.5.4) AutoDock tools (Morris et al. 2009) prior to docking. All the water and solvent atoms of the protein were removed and the polar hydrogen atoms were added. The GmIPK2 molecule was kept rigid while the ligands were allowed to rotate and explore more flexible binding pockets. Vina used a customized rectangular 3D cartesian grid for specifying the binding site of the protein and for efficient geometric scoring. The dimensions of the grid were customized to make sure that the size of the search space is large enough for the ligand to rotate in. Once set, the docking run produced ligand poses each with a definite binding energy (kcal/mol) calculated based on the scoring function used in Vina. The conformations with the lowest binding affinity were chosen and the interaction diagrams were generated using Discovery Studio Visualizer 4.1 (Accelrys Software Inc., USA 2013). The amino acid residues present at a distance of 2 Å were considered as the binding partners of the ligands. The active amino acid residues were also predicted by combining results of

three different interface prediction web servers, CASTp (Dundas et al. 2006), FTSite (Kozakov et al. 2015) and FunFOLD2 (Roche et al. 2013) into a consensus. The final complexes were equilibrated by MD simulation by following the procedure described for protein model in explicit water in the previous section but by applying restraints on the ligands to prevent them from moving away from the binding site. Once equilibrated, we simulated the complexes for 50 ns and analyzed the resulting trajectories for their stability using *g_rms*, *g_rmsf*, *g_hbond* and *g_mmpbsa* tools.

Results and discussion

Cloning and sequencing of *GmIPK2*

We first performed a BLASTN analysis in the plant comparative genomics portal Phytozome v9.1 (Goodstein et al. 2012) with the soybean *IPK2* gene sequence available in NCBI (GenBank: NM_001250522) to retrieve a 1241 bp transcript sequence. Based on the sequence data thus derived, we designed primers specific to amplify complete *GmIPK2* coding sequence (CDS) as well as a fragment of its 3' untranslated region (UTR) with the aim to design silencing construct in the future specific to this highly conserved region because of its substantial role in gene regulation. To accomplish this, we converted total RNA isolated from 8 to 10 mm developing seed stage to cDNA and performed PCR amplification from the synthesized cDNA template following the protocol described under sect. “PCR amplification, cloning, and sequencing of partial *GmIPK2* sequence”. The amplicon was then cloned into pGEM-T easy vector system and introduced into the bacterial host *E. coli* (DH5 α). The putative cDNA clones were verified by restriction analysis with *EcoRI* and sequence characterized to ~910 bp residues in length (Fig. 1). It contains a single open reading frame ~840 bp long which potentially encode a single polypeptide of 279 amino acid residues and a ~70 bp 3' UTR fragment. We submitted the obtained CDS data to NCBI (GenBank: KF297702) and used the information as query to conduct protein homology search using PSI-BLAST algorithm. Amongst the sequences producing significant alignment, we shortlisted 30 plant *IPK2* sequences based on percentage sequence identity to carry out further in silico analysis.

Spatial and temporal expression profiling

As discussed above, tissue-specific modulation of *IPK2* gene to generate a *lpa* mutant is essential to evade any possible pleiotropic effects. Before this can be achieved it is vital to investigate its spatial expression profile in different tissues as well as its temporal expression profile in developing

Fig. 1 Agarose gel showing PCR amplification product of *GmIPK2* gene isolated from developing seeds (8–10 mm) of *Glycine max* cv. Pusa-16. Lane M: 500 bp DNA ladder, lanes 1–3: ~910 bp *GmIPK2* fragment

seeds, to provide an initial point for strategic achievement of a desired level of silencing.

To study the expression of *GmIPK2* gene in different tissues of a soybean plant and during seed development, semi-quantitative as well as real-time PCR expression analysis was performed with total RNA isolated from root, stem, leaf and flower tissues of 30-day-old *G. max* plants and developing seeds ranging from 0 to 16 mm in size distributed in eight different progressive stages. Semi-quantitative PCR analysis revealed a differential pattern of *GmIPK2* transcript expression across the set of experimental tissues analyzed, with the highest level of transcripts observed in seeds (Fig. 2a). The same was confirmed by steady-state qRT-PCR analysis which also detected highest level of transcripts in seeds (Fig. 2b). This suggests that *GmIPK2* play a key role in PA biosynthesis in this tissue for use as a primary source of energy during germination. In both the analyses, same level of amplification was observed for *PEPCo* house-keeping gene. Since PA is also required for several other vital functions throughout the plant system, a basal level of it is observed in all the tissues as well. Thus, amongst the other tissues analyzed, expression of *GmIPK2* was also recorded in roots and flowers, but at a level lower than that observed in the cotyledons. The strong presence of *GmIPK2* transcripts in these tissues may be attributed to its role in regulating cytosolic calcium gradient which correlates with pollen germination, pollen tube growth (Franklin-Tong et al. 1996; Malho 1998; Pierson et al. 1994; Xu et al. 2005), root growth, and root hair development (Bibikova et al. 1997; Felle and Hepler 1997; Wymer et al. 1997). A low level of expression was also observed in stems and leaves as it is studied to be involved in regulating a vital function of axillary shoot branching by participating in auxin signaling (Zhang et al. 2007).

Fig. 2 **a** RT-PCR expression analysis of *GmIPK2* gene in different plant tissues of Pusa-16 cultivar using soybean housekeeping gene *PEPCo* as an internal control. **b** Relative quantification of *GmIPK2* transcript levels in the samples analyzed above by qRT-PCR, normalized to soybean housekeeping gene *PEPCo*. The leaf tissue was taken as calibrator. The data are mean of technical triplicates of each of the three biological replicates with error bars indicating standard deviation (SD)

We then analyzed its temporal expression pattern during eight progressive seed development stages. In both semi-quantitative as well as real-time PCR expression analysis, we observed that its expression increased as the development progressed and reached a peak value at the later stages of seed development (Fig. 3). This pattern of expression coincides evenly with the pattern of accumulation of PA which is linear throughout most of seed development (Raboy and Dickinson 1987). Similar results were obtained in a microarray transcriptome study conducted in the past in our laboratory (GEO: GSE69821). This observation can be explained by the increase in production of phosphorous compounds required to support growth and development during initial stages of seed development when the synthesis of phosphorous reserve, PA is minimal (Raboy and Dickinson 1987).

The spatiotemporal analysis in summary identifies seed as the major tissue for its expression with maximum relative expression occurring during the later stages of its development. Thus, from the present study we can hypothesize that targeting the *GmIPK2* gene expression during late seed development stages may provide a potential strategy for generating *lpa* soybean with enhanced nutritional value.

Characterization of regulatory motifs in *IPK2* gene promoter

To understand the mechanism regulating the spatiotemporal expression pattern of *GmIPK2* gene thus observed, we analyzed the promoter region of all its homologs and identified different cis-regulatory elements located therein.

Fig. 3 **a** RT-PCR expression analysis of *GmIPK2* gene in developing seeds of Pusa-16 cultivar using soybean housekeeping gene *PEPCo* as an internal control. **b** Relative quantification of *GmIPK2* transcript levels in the samples analyzed above by qRT-PCR, normalized to soybean housekeeping gene *PEPCo*. The 0–2 mm seed stage was taken as calibrator. The data are mean of technical triplicates of each of the three biological replicates with error bars indicating SD

A 2 kb sequence upstream to the open reading frame was identified and subjected to PlantCARE analysis. Database search revealed the presence of many motifs related to seed-specific promoters, hormone-responsive cis-elements (HRE) and cis-elements responsive to stresses (DSRE) that together contribute to the differential regulation of our gene (Table 1). Light responsive elements (Arguello-Astorga and Herrera-Estrella 1996; Feldbrugge et al. 1996; Lois et al. 1989; Lopez-Ochoa et al. 2007; Sessa et al. 1995) were observed most frequently which suggest a probable diurnal regulation of *IPK2* expression. The GCN4 and Skn_1 motifs highly conserved in the promoters of cereal seed storage protein genes were located within the *IPK2* promoter. These cis-acting elements play a central role in controlling endosperm-specific gene expression (Onodera et al. 2001; Takaiwa et al. 1996; Washida et al. 1999; Wu et al. 1998). Multiple HREs particularly those known to be involved in abscisic acid (ABA) and gibberellic acid (GA3) sensing were also identified. A ABA-responsive element (ABRE) with the core sequence PyACGTG/TC and three GA-responsive elements (GARE) AAACAGA, TCTGTTG and TAT CCAC/T (Niu et al. 2016; Mongkolsiriwatana et al. 2009; Yamaguchi-Shinozaki et al. 1989) were identified which previously were reported to correlate with PA accumulation during grain filling (Abe et al. 2003; Matsuno and Fujimura 2014). Aggarwal et al. 2015 reported that *IPK2* is an ABA-induced gene which is antagonistically suppressed by GA3 underlining the crucial role played by these hormones in regulating PA pathway genes. Putative elements responsive

Table 1 Potential cis-acting elements identified in the 5' regulatory sequences of plant *IPK2s*

Classification	Name	Sequence	Source organism	References
Endosperm	GCN_4 motif	GAAGCCA, TGAGTCA	<i>Oryza sativa</i>	Takaiwa et al. (1996), Onodera et al. (2001)
	Skn-1 motif	CAAGCCA, TGTGTCA, GTCAT	<i>Oryza sativa</i>	Washida et al. (1999)
ABA	ABRE	CACGTG, TACGTG	<i>Arabidopsis thaliana</i>	Yamaguchi-Shinozaki et al. (1989)
Gibberellin	GARE	AAACAGA, TCTGTTG, TATCCAC/T	<i>Brassica oleracea</i> <i>Oryza sativa</i>	Mongkolsirawatana et al. (2009), Jun Niu et al. (2016)
Light	Box 4	ATTAAT	<i>Petroselinum crispum</i>	Lois et al. (1989)
	CATT-motif	GCATTC	<i>Zea mays</i>	Arguello-Astorga and Herrera-Estrella (1996)
	G-Box	CACGTG	<i>Pisum sativum</i> , <i>Arabidopsis thaliana</i> , <i>Zea mays</i>	Sessa et al. (1995), López-Ochoa et al. (2007)
	ATCC/T-motif	AATCTAATCC/T	<i>Pisum sativum</i> , <i>Arabidopsis thaliana</i>	Arguello-Astorga and Herrera-Estrella (1996)
	TCT-motif	TCTTAC	<i>Arabidopsis thaliana</i>	Arguello-Astorga and Herrera-Estrella (1996)
	Box-I	TTTCAAA	<i>Pisum sativum</i>	Arguello-Astorga and Herrera-Estrella (1996)
	ACE	AAAACGTTTA	<i>Petroselinum crispum</i>	Feldbrugge et al. (1996)
	GA-motif	AAGGAAGA	<i>Glycine max</i>	Arguello-Astorga and Herrera-Estrella 1996
	CGTCA-motif	CGTCA	<i>Hordeum vulgare</i>	Kim et al. (1993), Rouster et al. (1997)
	TGACG-motif	TGACG	<i>Hordeum vulgare</i>	
Drought	MBS	T/CAACTG	<i>Arabidopsis thaliana</i>	Shinozaki and Yamaguchi-Shinozaki (2000), Abe et al. (2003)
Anaerobic	ARE	TGGTTT	<i>Zea mays</i>	Nguyen et al. (2003)

to methyl-jasmonate, drought inducibility and anaerobic induction were also observed which help combat assorted types of abiotic factors that plants are exposed to under natural environment (Abe et al. 2003; Kim et al. 1993; Nguyen et al. 2003; Rouster et al. 1997; Shinozaki and Yamaguchi-Shinozaki 2000).

Computation of physicochemical parameters and subcellular localization prediction

GmIPK2 sequence similarity search using PSI-BLAST revealed homology to IPK2 protein sequences from different plant sources showing maximum similarity with *Glycine soja* (98%) and *Vigna radiata* (74%) (Table 2). Primary sequence analysis of these homologs indicate that leucine is the most abundant amino acid which makes upto approximately 9–11 mol percent of its backbone residues whilst the percentage of tryptophan and methionine were found to be the least at approx. 1% across species. The isoelectric points (pI) of all IPK2 proteins were computed to be under 7 suggesting that they are most likely to precipitate in acidic buffers except *Glycine soja*, *Medicago truncatula*, *Zea mays* and *Sorghum bicolor* which has a pI of 7.1, 8.57, 10.48 and 8.26, respectively indicating their solubility in basic buffers. The calculated pI will be useful for empirical protein purification by isoelectric focusing and ion exchange chromatography. The extinction coefficient (EC) (Gill and

Von Hippel 1989) of IPK2 proteins measured at 280 nm in water was found ranging from 25,440 to 34,380 M⁻¹ cm⁻¹ with respect to their concentration of aromatic amino acids (11–15%) and cystine (disulfide bonds). These EC values can be used to calculate protein concentration in a solution which in turn help in the quantitative study of biochemical interactions (protein–protein and protein–ligand). The instability indices (Ii) computed for selected IPK2 proteins are used to determine their in vivo half-lives (Guruprasad et al. 1990). Rogers et al., 1986 reported that proteins having Ii values greater than 40 have an in vivo half-life of less than 5 h while those proteins having Ii values less than 40 have a longer in vivo half-life of 16 h. Our study showed that Ii values of all the homologs are less than 40 and hence are thermally stable with a long half-life except for *Lotus japonicas* and *Zea mays* kinases that have a Ii above 40 which indicate their possible thermal instability. Thermostability of proteins result from a combination of several factors acting synergistically. Here we assess thermal stability of our proteins in direct proportionality with their aliphatic index (Ai) which is a measure of the relative volume occupied by aliphatic side chains (Ikai 1980). The Ai values determined for IPK2 kinases ranged from 68.79 to 96.70 with those from brassicaceae family showing lowest thermal stability. This in turn is indicative of their greater flexibility at a wide range of temperatures when compared to proteins of other families. Grand average of hydropathicity (GRAVY) number reflect

Table 2 Physiochemical parameters of shortlisted plant IPK2 sequences computed using the ProtParam tool

Organism	Accession no.	Sequence length	MW	pI	EC	Ii	Ai	GRAVY	−R	+R
<i>Glycine max</i>	AGW99177.1	279	30,979.2	6.29	29,910	29.72	89.07	−0.18	35	30
<i>Glycine soja</i>	KHN19419.1	178	19,898.87	7.10	23,950	29.73	95.73	−0.059	21	21
<i>Vigna radiata</i>	XP_022634636.1	286	31,482.10	6.59	28,420	38.17	93.67	0.045	29	27
<i>Cajanus cajan</i>	XP_020232596.1	267	29,463.71	6.86	28,420	33.70	91.31	−0.132	31	30
<i>Phaseolus vulgaris</i>	XP_007133857.1	264	28,951.0	6.59	26,930	38.90	90.38	−0.017	27	25
<i>Lotus japonicus</i>	AFK39224.1	283	31,348.76	5.66	32,430	42.36	87.74	−0.139	39	30
<i>Cicer arietinum</i>	XP_004510840.1	296	33,061.70	6.12	26,930	31.59	91.11	−0.250	38	33
<i>Medicago truncatula</i>	XP_003627882.1	426	47,961.8	8.57	29,910	31.89	85.02	−0.344	49	53
<i>Arachis hypogaea</i>	ALT56981.1	297	32,684.01	6.45	28,420	36.41	84.01	−0.227	34	32
<i>Arachis duranensis</i>	XP_015937330.1	297	32,883.25	6.45	28,420	37.35	84.01	−0.248	35	33
<i>Corchorus capsularis</i>	OMO94774.1	292	32,498.97	5.84	28,420	40.75	92.43	−0.159	38	33
<i>Theobroma cacao</i>	EOY22693.1	304	34,213.77	6.32	31,400	37.61	84.61	−0.292	39	36
<i>Herrania umbratica</i>	XP_021285810.1	305	34,392.01	6.56	36,900	40.38	83.38	−0.307	38	36
<i>Durio zibethinus</i>	XP_022738659.1	295	32,954.54	6.71	28,420	38.72	87.83	−0.237	36	35
<i>Solanum lycopersicum</i>	XP_004235863.1	294	32,476.9	5.93	28,420	34.32	83.23	−0.261	35	30
<i>Solanum tuberosum</i>	NP_001335929.1	408	44,295.0	5.96	28,420	20.21	93.19	−0.254	50	43
<i>Brassica rapa</i>	XP_009112000.1	274	30,640.7	6.25	34,380	30.25	80.66	−0.283	36	33
<i>Capsicum baccatum</i>	PHT32205.1	374	40,995.36	5.97	32,890	33.47	90.19	−0.297	50	44
<i>Nicotiana attenuata</i>	XP_019263239.1	368	40,865.55	5.51	31,400	30.30	93.51	−0.179	48	40
<i>Capsicum annuum</i>	PHT66064.1	374	41,015.41	5.91	32,890	32.27	92.01	−0.277	50	43
<i>Capsicum chinense</i>	PHU00941.1	374	40,986.37	5.81	32,890	32.27	92.27	−0.271	51	43
<i>Brassica napus</i>	XP_013749343.1	285	31,961.0	5.87	35,870	30.23	75.86	−0.387	40	34
<i>Arabidopsis lyrata</i>	XP_020870073.1	300	33,671.08	5.82	31,400	29.76	83.47	−0.324	42	35
<i>Prunus avium</i>	XP_021815221.1	281	31,100.28	6.20	31,400	26.98	86.41	−0.209	34	30
<i>Trifolium pratense</i>	PNY08557.1	273	30,275.83	5.50	23,950	29.99	96.70	−0.054	37	26
<i>Arabidopsis thaliana</i>	NP_200984.1	300	33,486.7	5.72	31,400	25.96	80.23	−0.329	40	31
<i>Prunus persica</i>	XP_007209445.1	282	31,303.57	6.50	31,400	28.39	85.39	−0.252	35	33
<i>Lepidium latifolium</i>	ACK86969.2	297	33,212.5	6.59	34,380	29.25	81.95	−0.285	31	29
<i>Zea mays</i>	XP_008649440.2	240	26,665.45	10.48	29,450	65.74	68.79	−0.517	22	37
<i>Aegilops tauschii</i>	XP_020147020.1	287	30,607.86	6.13	26,930	38.85	88.43	−0.052	33	29
<i>Sorghum bicolor</i>	XP_002452184.1	322	34,550.64	8.26	28,420	39.34	90.93	−0.013	32	34

MW molecular weight (g/mol), pI isoelectric point, EC extinction coefficient (M−1cm−1), Ii instability index, Ai aliphatic index, GRAVY grand average hydropathy, (−R) number of negative residues, (+R) number of positive residues

the average hydropathy of a protein, the positively rated being hydrophobic and negatively rated being hydrophilic in nature Kyte and Doolittle (1982). GRAVY index for IPK2 kinases was found ranging from −0.517 to −0.013 indicating that these proteins will interact favourably with water except for *Vigna radiata* with an index of 0.045 and hence is a potential hydrophobic protein.

Protein localization and target peptide predictions are significant studies as they aid in in silico protein function characterization as well as genome annotation. The acidic amino acid composition of IPK2 homologs determined by the physiochemical analysis conducted above suggest that they are cytoplasmic in nature as opposed to the basic amino acid composition of membrane proteins for their stability

(Schwartz et al. 2001). Further sequence analysis based on TargetP scores (cTP: 0.163, mTP: 0.066, SP: 0.087, other: 0.906) also suggest that IPK2 kinases may be located anywhere in the cell besides chloroplast and mitochondria. The sequences were not predicted to have any signaling pre-sequence which was indicated by their low signal peptide (SP) score reinforced that they are soluble in nature. Besides, a consensus of predictions obtained from WoLF PSORT, CELLO v2.5, SubLoc v1.0 and MemType-2L servers (Table S1) also established the cytoplasmic character of IPK2 protein. However, TMPred, PSIPRED, NPS@ and DAS servers also identified a single consensus C-terminal transmembrane region positioned at 259–274. This could have been mistaken due to the presence of large regions

of hydrophobic residues in the C-terminal soluble region, as hydrophobicity alone is used as the criterion to predict membrane-spanning regions. The absence of any transmembrane helix is also well documented in the hydrophobicity plot generated using waveTM server (Fig S1) (Pashou et al. 2004) as majority of amino acids show negative for hydrophobicity.

Motif analysis and secondary structure characterization

Multiple sequence alignment (MSA) of the selected plant IPK2 homologs was performed using M-Coffee web server which is a meta-method for assembling MSA by combining the output of several individual methods into one to generate the best possible alignment. The consensus alignment thus generated revealed several significantly conserved motifs and sites unique to inositol phosphate kinases (Fig. 4). A signature inositol phosphate-binding motif, PxxxDxKxG was identified in all the aligned sequences (Odom et al. 2000; Saiardi et al. 1999) which confirms that they belong to inositol phosphate kinase (IPK) superfamily of IP kinases. Holmes and Jogl, 2006 state that the members of this superfamily share several strictly conserved signature motifs with each other and are predicted to assume the same overall fold, despite the low sequence conservation. The core catalytic tyrosine kinase motif, RxxxExxxY was also discovered in all the sequences which suggest that they are tyrosine-specific protein kinases (Cooper et al. 1984). IPK2 sequences from Solanaceae and Rosaceae families were found to contain a Glycine-rich consensus ATP-binding GxGxxG motif characteristic of protein kinase C (PKC) catalytic domain (Steinberg 2008). The classical PKC and plant CDPKs recognized phosphorylation S/TxK/R motif (Nishikawa et al. 1997; Neumann et al. 1996; Roberts and Harmon 1992) was also identified in some of the sequences speculating their role in lipid-dependent PA biosynthetic pathway. Such promiscuous kinase activity suggests that both lipid-dependent and independent pathways regulate PA biosynthesis as well as basic nuclear and cellular processes in plants (Josefsen et al. 2007; Stevenson-Paulik et al. 2002). A protein recognition LxxLL motif common to all of the aligned sequences indicate their participation in protein–protein interactions, regulating cell signalling, cell adhesion, and transcription (Plevin et al. 2005). Further analysis by MEME suite web server identified a total of 11 conserved ungapped motifs, with motif PxxxDxKxG being the most conserved amongst all IPK2 homologs as indicated by its lowest *E* value of 2.9e-487 (Fig S2). The obtained motifs were subjected to BLASTP analysis for conformation of their annotations which established that they all belong to IPK superfamily domain (CDD Acc: c112283) and thus substantiate our previous results. We further explored GmIPK2 protein sequence

by Motif Scan which recognized diverse protein kinase phosphorylation sites (Table S2). Since phosphorylation acts as a molecular switch in modulating protein function, structural rearrangement and cellular localization it can be suggested that GmIPK2 play a critical role in many biological regulatory events of signalling, proliferation, differentiation, and apoptosis. Its role in biological processes as diverse as mRNA export (York et al. 1999), DNA repair (Hanakahi et al. 2000), regulation of chromatin structure (Shen et al. 2003; Steger et al. 2003), maintenance of basal resistance to plant pathogens (Murphy et al. 2008) and apoptosis (Agarwal et al. 2009) have been studied in the past.

The secondary structure of a protein is more conserved than its nucleotide sequence and is, therefore, a prized source of information in understanding its classification, function, molecular evolution, and interaction with macromolecules (Reehana et al. 2013). In addition, secondary structure provides the first framework for homology based prediction of a protein 3D-model. Thus, in the current study, we inferred the secondary structure composition of IPK2 kinases from a three-state prediction done using NPS@ web server (Table 3). A high coil content was observed in most of the sequences including GmIPK2 while some showed them in nearly equal proportion with α -helix. This structural state can be justified based on the rich content of highly flexible glycine and kink inducing proline amino acid residues. The percentage of extended strands (% Ee) in all the kinases were found ranging from 13 to 29% except for the Poaceae family kinases which showed a low % Ee conformation (below 10%). PDBSum tool PROMOTIF analysis of GmIPK2 polypeptide identified total eight α -helices and eight β -strands arranged to form three antiparallel β -sheets, interspersed throughout by regions of coil or turn conformations (Fig. 5). We also recognized a varying number of bonded half-cysteine pairs in all the IPK2 protein sequences using DIANNA server. It revealed the presence of 8 Cys residues in GmIPK2 and the most probable half-cysteine pairs predicted by CYSPREC were 94–121, 123–276 and 158–186. These potential long-term interactions participate in stabilizing the native conformations of our proteins and may as well contribute to differences in their tertiary structures.

Evolutionary analysis

Conserved motif analysis of IPK2 protein sequences point at a distinct evolutionary association between these kinases. Phylogenetic analysis would provide a further basis to determine their relatedness as well as to understand their collective evolution from a common ancestor. Previous research has reported that IPK superfamily of kinases to which IPK2 belongs evolved from a common ancestor (Irvine and Schell 2001; Shears 2004). In our study, based on the alignment obtained in “Motif analysis and secondary structure

a **b**

Fig. 4 M-Coffee multiple sequence alignment diagram of selected plant IPK2 protein sequences rendered with Jalview. The sequence motifs shared amongst all the representatives are coloured according

to percentage identity. The consensus row at the bottom shows the most frequent residue at each column or a '+' if two or more residues are equally abundant

c

Fig. 4 (continued)

characterization” of GmIPK2 protein and its homologs, we constructed a neighbor-joining phylogenetic tree using MEGA 6.0 software (Fig. 6). The tree topology derived was supported by high bootstrap values. The IPK2s were clustered into six well delineated groups. The clusters consist of members of the Poaceae, Brassicaceae, Malvaceae, Rosaceae, Solanaceae and Fabaceae families. The Poaceae family of monocots (*Zea may*, *Sorghum bicolor* and *Aegilops tauschii*) was found to be most distantly related to *G. max* in comparison to the fabaceae family of eudicots (*Glycine soja*, *Phaseolus vulgaris*, *Vigna radiata*, *Cajanus cajan*, *Trifolium pratense*, *Cicer arietinum*, *Medicago truncatula*, *Lotus japonicas*, *Arachis duranensis*, and *Arachis hypogaea*) which are most closely related.

Three-dimensional model construction

Model building

To derive structural information about GmIPK2 protein we built its theoretical model by following homology modeling approach since no X-ray crystal or NMR structure of it is available. The homology modeling technique takes advantage of structural conservation found in similar proteins that have evolved from a common ancestor. Yeast IPK2 protein was the first member of inositol multikinase family whose

crystal structure was determined (Holmes and Jogl 2006), but it was found to show a very low sequence similarity with GmIPK2 protein. As we know, aligning two sequences can be a difficult process if the sequence similarity is low, we therefore, used the above derived GmIPK2 protein sequence as a query in PSI-BLAST to find more similar sequences amongst the PDB database proteins with resolved 3D structures to use as a potential template. The only closest homologous sequence available in PDB was that of chain A of *Arabidopsis thaliana* inositol phosphate multikinase (PDB:4FRF) which showed 55% sequence identity with an e value of 9e-99. The initial comparative models, i.e., GmIPK2-S and GmIPK2-P were built using fully automated SWISS-MODEL and PHYRE2 servers, respectively, which also identified 4FRF_A as the most reliable template using sensitive hidden markov model searches and used the same as the structural input. A global quality estimation score (GMQE) of 0.62 was provided by SWISS-MODEL which indicates a reasonably reliable structure. Homology model was additionally built using MODELLER 9.16 program from the X-ray crystal structure coordinates of the previously identified template structure (4FRF_A). The software generated five different models by optimizing the objective function of spatial restrains in a cartesian space. Three different energy scores viz. molpdf, DOPE and GA341 were computed for each of these models and compared to one

Table 3 Three-state description of secondary structure content and disulfide pattern prediction of IPK2 sequences

Organism	α -Helix (%Hh)	Extended strands (%Ee)	Random coil (%Cc)	Disulphide bridge prediction
<i>Glycine max</i>	26.88	13.26	31.54	94–121, 123–276, 158–186
<i>Glycine soja</i>	34.27	29.78	35.96	20–117, 22–85, 161–175
<i>Vigna radiata</i>	29.72	17.13	24.48	127–190, 263–281
<i>Cajanus cajan</i>	25.84	13.11	29.21	119–177
<i>Phaseolus vulgaris</i>	24.62	21.21	28.79	116–176, 208–249
<i>Lotus japonicus</i>	25.44	14.49	26.86	94–215, 149–187, 261–283
<i>Cicer arietinum</i>	21.96	14.53	28.72	122–185, 217–279, 260–274
<i>Medicago truncatula</i>	18.54	15.02	31.69	98–262, 325–357
<i>Arachis hypogaea</i>	24.58	14.48	33.67	120–188, 223–288, 269–292
<i>Arachis duranensis</i>	24.58	14.48	33.33	223–288, 269–292
<i>Corchorus capsularis</i>	19.52	20.21	25.34	120–187, 222–267
<i>Theobroma cacao</i>	19.74	16.78	39.14	50–120, 187–222, 267–298, 288–296
<i>Herrania umbratica</i>	18.69	14.75	42.95	120–296, 222–298, 267–288
<i>Durio zibethinus</i>	18.98	22.37	26.10	93–120, 149–222, 267–288
<i>Solanum lycopersicum</i>	18.03	18.37	30.27	120–187, 222–267
<i>Solanum tuberosum</i>	17.40	17.65	26.23	120–187, 222–283
<i>Brassica rapa</i>	13.50	21.17	22.63	144–186, 221–264
<i>Capsicum baccatum</i>	17.65	17.38	30.75	120–187, 222–283
<i>Nicotiana attenuata</i>	21.47	15.76	25.82	156–223, 258–303
<i>Capsicum annum</i>	19.52	16.58	30.48	120–187, 222–267
<i>Capsicum chinense</i>	20.32	16.31	30.21	120–187, 222–267
<i>Brassica napus</i>	18.25	18.25	24.56	120–220, 144–185, 264–281
<i>Arabidopsis lyrata</i>	22.00	17.00	23.33	120–187, 222–272
<i>Prunus avium</i>	17.44	17.44	27.05	120–222, 156–187
<i>Trifolium pratense</i>	21.61	15.75	31.87	173–205, 203–250
<i>Arabidopsis thaliana</i>	22.00	16.67	24.67	120–187, 197–272
<i>Prunus persica</i>	17.02	17.02	29.79	120–222, 156–187
<i>Lepidium latifolium</i>	17.85	17.17	28.62	78–269, 120–187, 222–285
<i>Zea mays</i>	16.25	10.42	44.17	101–203, 192–200
<i>Aegilops tauschii</i>	28.57	7.67	31.36	110–188, 124–267
<i>Sorghum bicolor</i>	30.12	12.42	26.40	13–300, 137–160

The secondary structure data were generated by joint prediction with SOPMA and PHD while disulfide bonding pattern was determined using DiANNA (DiAminoacid Neural Network Application) 1.1 server

another to select the best 3D structure (Table S3). Model 4 (GmIPK2-M4), with the lowest molpdf and DOPE scores of 1734.48291 and –27846.58203 respectively, was chosen as the principal conformational structure.

Structure validation

We then assessed the accuracy and reliability of all the three predicted models viz. GmIPK2-S, GmIPK2-P and GmIPK2-M4 using various online diagnostic tools (Table 4). RAM-PAGE server which evaluates the 3D-structures based on Ramachandran plot calculations showed variable distribution of torsion angles in all the models. Fig S3a shows Ramachandran plot for GmIPK2-S model, showing 92.5% residues in favourable region and 4.9% in allowed region

with just 2.7% residues in outlier region of the plot, which reflects its superior backbone geometry. We then utilized ProSA-web computational engine to analyze overall quality of the models based on their z scores and local quality based on their residue energies. GmIPK2-S model showed a better z score of –6.56 which is displayed in the energy distribution plot derived from a group of experimentally determined protein structures of similar size. The Z score thus observed was very much within the range of scores typically found for native conformations of this group which indicates a good overall quality of the modelled protein (Fig S3b). Moreover, its residue energy was computed to be largely negative which further reflects that the local regions in the protein are modelled well (Fig S3c). The accuracy of 3D-models was further analyzed from energy profiles obtained by

Fig. 5 Topology map of GmIPK2 generated using ProMotif. There are a total of eight α -helices (1–8) and three β -sheets (β -sheet 1: strands A1 and A2; β -sheet 2: strands B1, B2, B3 and B4; β -sheet 3: strands C1 and C2)

Verify3D program. Figure S3(d) shows 3D-1D profile for GmIPK2-S model with 97.82% of the model residues showing an average 3D-1D profile score ≥ 0.2 and hence validate that majority of its amino acid sequence reconcile to its

environment in the 3D structure. Additionally, the pairwise 3D structural alignment of GmIPK2-S model with template protein, 4FRF_A using MATRAS 2.1 program revealed that both the structures shared 91.3% secondary structure identity (Fig S4a) and the average distance between the C α backbone atoms of their 3D structures, i.e., root mean square deviation (RMSD) measured through superimposition was 0.44 Å (between 196 atom pairs) (Fig S4b). Based on the majority of winning scores, GmIPK2-S (Fig. 7) was chosen as the best comparative model for energy minimization and further analyses.

Molecular dynamics simulation

We subsequently performed MD simulation on our predicted GmIPK2-S model using GROMOS96 53A6 force field to compute its stability and dynamics. Initial potential energy minimization of solvated model showed that the maximum force dropped below the defined value of $1000 \text{ kJ mol}^{-1} \text{ nm}^{-1}$ in nearly 500 steps. The protein structure was then subjected to 50 ns of equilibration run at a constant temperature and pressure to obtain its molecular trajectory. The trajectory thus obtained was used to determine RMSD of C α backbone atoms of the model using its starting structure as the reference, to determine its convergence towards an equilibrium state. Figure 8a shows RMSD as a function of simulation time. We observed that the protein stabilized around 10 ns of production run and converged to $\sim 0.55 \text{ nm}$ at 50 ns. The initial increase of RMSD could be attributed to the restraints in the system applied in the equilibration phase and their release later at the beginning of production phase. Besides RMSD, we also calculated root mean square fluctuations (RMSFs) to study mobility of the protein structure to draw an idea of its flexibility regions. From the RMSF plot (Fig. 8b), we identified Leu (75, 135, 272), Asp (76, 79, 127), Ala (77), Ser (78, 130–132), Gly (80, 136), His (84),

Fig. 6 Phylogenetic tree showing evolutionary relationship of 31 plant IPK2 sequences divided into different clades, colour coded to indicate the plant family to which they belong. The posterior probability values are indicated corresponding to every node

Table 4 Validation parameters computed for the energy-minimized 3D models of GmIPK2 protein built using different programs

Model	RAMPAGE Percentage of residues in favoured region	ProSA Z score	Verify3D Percentage of residues with 3D-1D score ≥ 0.2
GmIPK2-S	92.5	-6.56	97.82
GmIPK2-P	82.2	-6.28	84.95
GmIPK2-M4	87.7	-5.84	79.93

Fig. 7 Homology model of GmIPK2_S protein rendered using PyMOL

Lys (126, 154), Glu (129, 230), Arg (128, 153) residues to be more flexible. This implies that the mentioned residues show greater movement from their native position, i.e., are dynamic in nature and thus are functionally more relevant. Thus, overall, the simulation results highlight the stable and reliable nature of our protein model and find it fit to be used for further active site predictions.

Molecular docking with inositol phosphates

Active site predictions

Once the protein model was refined, a comparative study was performed to detect possible binding pocket residues using CASTp, FTSite and FunFOLD2 binding site prediction servers (Fig S5). Lys105, Thr110, Lys122, Lys126, Ser130, Lys138, Ile139, Pro140, Arg153, Lys154, Gln157, and Ser219 were determined as possible active site residues

Fig. 8 **a** Root mean square deviations and **b** Root mean square fluctuations of the C α backbone atoms in GmIPK2_S model over 50 ns MD simulation

in GmIPK2-S model. Similar predictions were also made in template protein (4FRF_A) which identified Arg104, Thr105, Pro108, Phe137, Lys149, Arg152, His216, Asn218, Ser219, Gln242, and Val246 as the probable binding site residues. From these studies, we could decipher that residues Lys, Arg, Pro, Thr, Gln and Ser are highly conserved in active sites of functionally identical model and template proteins.

Docking and residue interaction analysis

Molecular recognitions are vital to many biological processes. However, experimental determination of structures of molecular interactions is cost intensive, demand time, and expertise. We therefore, chose computational molecular docking to model our protein–ligand binding and characterize the interactions between its binding pocket residues and

known active ligands. Multiple substrate specificities have been described previously for IPK2 gene product, catalyzing primarily 5GUG-I3P and 4A69-I0P (Stevenson-Paulik et al. 2002). We thus docked these centroid ligands into the binding cavity of GmIPK2 protein using molecular docking program Autodock Vina based on a semi-flexible docking approach with the scaling factor defined within 0.1 nm to predict their bound geometry. VINA uses its iterated local search global optimizer algorithm to produce 9 different poses of which pose 1 corresponding to each ligand was identified as the best binding mode based on their lowest binding affinity score of -6.2 kcal/mol for 5GUG-I3P and -5.8 kcal/mol for 4A69-I0P computed by VINA's default statistical scoring function. We further inspected the molecular interactions between these protein substrate poses using Discovery Studio to predict functionally important amino acid residues and found that both the ligands were stabilized in their active site area by strong hydrogen-bonding interactions (Fig. 9a, b). Lys105 and Lys126 were identified as H-donors to the phosphate group oxygen of both 5GUG-I3P

and 4A69-I0P while Arg153 a H-acceptor for hydrogen bonds formation suggesting that these binding pocket residues may play a pivotal role in enzymes function and protein structure stability. Based on the previous work conducted by Holmes and Jogl, we hypothesize that side chains of these amino acid residues may assist in inducing conformational changes on inositol phosphate binding, enabling the enzyme to interact with differently phosphorylated inositol polyphosphates in different orientations, thus endorsing its substrate versatility. Besides Ser217 was also found to form stabilizing hydrogen bond with I0P ligand. Gln157 and Pro140 were observed to form an unconventional carbon–oxygen hydrogen bond with 5GUG-I3P and 4A69-I0P respectively indicating their possible contribution to ligand binding affinity and ligand recognition (Klaholz and Moras 2002). Lengths and angles of hydrogen bonds stabilizing the GmIPK2-I3P and GmIPK2-I0P complexes are enlisted in Table 5a, b respectively. Moreover, our analysis indicates that Thr110, Lys138 and Lys154 show non-bonding interactions with the ligands. 2D protein GmIPK2-I3P/I0P ligand

Fig. 9 Molecular docking of substrates to the GmIPK2_S homology model. Hydrogen-bonding interactions of **a** 5GUG-I3P and **b** 4A69-I0P substrates with residues in the active site of GmIPK2_S protein. The substrate is depicted in stick representation with carbon atoms coloured turquoise, oxygen atoms red, hydrogen atoms yellow and

phosphorus atoms blue. The interacting protein side chains are represented as maroon sticks in **a** and green sticks in **b**. **c**, **d** 2D-schematic representation of the interactions shown in **a**, **b** respectively drawn using Discovery Studio Visualizer

interaction diagrams (Fig. 9c, d) were also generated using Discovery Studio.

Molecular dynamics simulation of GmIPK2-I3P and GmIPK2-I0P complexes

We subsequently subjected our docked protein–ligand complexes to MD simulation using GROMOS96 43A1 force field to understand their stability and dynamics. The trajectories obtained were utilized to construct their respective RMSDs, RMSFs and H-bond interactions. The GmIPK2-I3P and GmIPK2-I0P complexes exhibited a deviation between ~0.45–0.65 nm (Fig S6) and ~0.45–0.58 nm (Fig S7a), respectively, that converged to ~0.6 and ~0.55 nm, respectively, at 50 ns. This suggests that the structures were stabilized following simulation. From their RMSF plots, we observed fluctuations up to ~0.38 nm in the GmIPK2-I3P complex (Fig S6b) and ~0.29 nm in the GmIPK2-I0P complex (Fig S7(b)) which reveal the characteristic regional flexibilities of functional significance in each complex. We also analyzed the hydrogen bonds which participate in the maintenance of these complexes. Three main hydrogen bonds help to stabilize the ligands, 5GUG-I3P and 4A69-I0P, within the enzyme's active site, two of them acting as acceptors and only one as a donor. In particular, the hydrogen bonds formed with Lys105 and Lys126 and the oxygen atom of each ligand's phosphate group, were constantly held. Figure 10 shows the time evolution of these main H-bonds. Other less significant hydrogen bonds were formed

sporadically between the ligands and the protein molecule. Lastly, we calculated the binding free energies ($\Delta G_{\text{binding}}$) of both the complexes using *g_mmpbsa* for their respective trajectories. $\Delta G_{\text{binding}}$ values of –12.4 and –11.9 kcal/mol were estimated for GmIPK2-I3P and GmIPK2-I0P complexes respectively. This indicates that the binding of ligand molecules is thermodynamically favourable and thus validates the reliability of our simulation.

Conclusion

The present work describes expression analysis and molecular characterization of *IPK2* multifunctional kinase involved in the late phase of PA biosynthesis. To initiate the study, we identified and cloned the partial gene sequence of *IPK2* from *G. max* cv. Pusa-16. *IPK2* transcripts, when assessed during seed development showed a predominant expression in the later stages of its development which consequently suggests that perturbing *IPK2* gene at this stage can be a viable strategy for manipulating PA levels in soybean seeds to achieve a *lpa* trait. Computational analysis of the gene further highlighted its molecular features including cis-acting promoter elements potentially regulating its observed expression; primary and secondary structural features shedding light on its physico-chemical features, conserved functional motifs and major structural elements; and its evolutionary relationship, which can be used to design several experimental studies.

Table 5 Hydrogen bonds between the active site residues of GmIPK2 and its substrates (a) 5GUG-I3P and (b) 4A69-I0P along with their distances and angles measured using Accelrys Discovery Studio Visualizer 4.1

(a)						
GmIPK2			5GUG-I3P		Distance (Å°)	Angle (Degree°)
Residue	Atom	Chemistry	Atom	Chemistry		
Lys105	HZ3	H-Donor	O7	H-Acceptor	2.074	49.9029
Lys126	HZ3	H-Donor	O14	H-Acceptor	1.852	44.2482
Arg153	O	H-Acceptor	H39	H-Donor	2.168	45.6080
Gln157	CA	H-Donor	O18	H-Acceptor	3.348	52.1869
(b)						
GmIPK2			4A69-I0P		Distance (Å°)	Angle (Degree°)
Residue	Atom	Chemistry	Atom	Chemistry		
Lys105	HZ2	H-Donor	O18	H-Acceptor	2.693	47.8602
Lys105	HZ3	H-Donor	O24	H-Acceptor	2.078	28.0171
Lys126	HZ3	H-Donor	O28	H-Acceptor	2.179	35.7367
Lys126	O	H-Acceptor	H40	H-Donor	2.121	39.3959
Pro140	CA	H-Donor	O4	H-Acceptor	3.732	60.3277
Arg153	O	H-Acceptor	H36	H-Donor	2.187	16.0359
Ser217	HG	H-Donor	O27	H-Acceptor	2.899	22.5987

Fig. 10 Time evolution of main hydrogen bonds formed between the GmIPK2_S protein and the 5GUG-I3P and 4A69-I0P ligands over 50 ns MD simulation

The 3D model of IPK2 protein developed can assist in the experimental determination of its 3D model as well as can be considered as a working model for generating hypothesis to make more accurate predictions of protein function and catalytic mechanism in the future. The docking studies performed subsequently can be used for structure-based designing of potent IPK2 inhibitors, useful in studying IPK2 active site and substrate selectivity and for studying PA biosynthesis pathway in detail. In conclusion, the obtained results provide very important preliminary data needed to manipulate PA content in soybean seeds as well as other crops for improving their nutritional quality by biotechnological intervention in the future.

Acknowledgements Financial support for the work was provided by funding from the National Funds for Basic, Strategic and Frontier Application Research in Agriculture [Grant No. NFBSFARA/RNAi-2011/2011-12], ICAR, Government of India. The authors would also like to thankfully acknowledge the Supercomputing Facility for Bioinformatics and Computational Biology at IIT Delhi for the use of its facilities.

Compliance with ethical standards

Conflict of interest The authors declare that there is no conflict of interest regarding the publication of this article.

References

- Abe H, Urao T, Ito T, Seki M, Shinozaki K, Yamaguchi-Shinozaki K (2003) Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signalling. *Plant Cell* 15:63–78
- Accelrys Software Inc (2013) Discovery studio modeling environment, release 4.0. Accelrys Software Inc., San Diego
- Agarwal R, Mumtaz H, Ali N (2009) Role of inositol polyphosphates in programmed cell death. *Mol Cell Biochem* 328:155–165
- Aggarwal S, Shukla V, Bhati KK, Kaur M, Sharma S, Singh A, Mantri S, Pandey AK (2015) Hormonal regulation and expression profiles of wheat genes involved during phytic acid biosynthesis pathway. *Plants (Basel)* 4:298–319
- Ali N, Paul S, Gayen D, Sarkar SN, Datta SK, Datta K (2013a) Development of low phytate rice by RNAi mediated seed-specific

- silencing of inositol 1,3,4,5,6-pentakisphosphate 2-kinase gene (IPK1). *PLoS One* 8:e68161
- Ali N, Paul S, Gayen D, Sarkar SN, Datta SK, Datta K (2013b) RNAi mediated down regulation of myo-inositol-3-phosphate synthase to generate low phytate rice. *PLoS One* 6:12
- Arguello-Astorga GR, Herrera-Estrella LR (1996) Ancestral multipartite units in light-responsive plant promoters have structural features correlating with specific phototransduction pathways. *Plant Physiol* 112:1151–1166
- Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 37:W202–W208
- Bhati KK, Aggarwal S, Sharma S, Mantri S, Singh S, Bhalla S, Kaur J, Tiwari S, Roy J, Tuli R, Pandey AK (2014) Differential expression of structural genes for the late phase of phytic acid biosynthesis in developing seeds of wheat (*Triticum aestivum* L.). *Plant Sci* 224:74–85
- Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L, Schwede T (2014) SWISS-MODEL: modeling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res* 42:W252–W258
- Bibikova TN, Zhigiler A, Gilroy S (1997) Root hair growth in *Arabidopsis thaliana* is directed by calcium and an endogenous polarity. *Planta* 203:495–505
- Bilyeu Kristin D, Zeng P, Coello P, Zhang Zhanyuan J, Krishnan Hari B, Bailey A, Beuselinck Paul R, Polacco Joe C (2008) Quantitative conversion of phytate to inorganic phosphorus in soybean seeds expressing a bacterial phytase. *Plant Physiol* 146:468–477
- Bond CS (2003) TopDraw: a sketchpad for protein structure topology cartoons. *Bioinformatics* 19:311–312
- Brearley CA, Hanke DE (1996a) Inositol phosphates in barley (*Hordeum vulgare* L.) aleurone tissue are stereochemical similar to the products of breakdown of Ins P6 in vitro by wheat bran phytase. *Biochem J* 318:279–286
- Brearley CA, Hanke DE (1996b) Metabolic evidence for the order of addition of individual phosphate esters to the myo-inositol moiety of inositol hexakisphosphate in the duckweed *Spirodela polyrhiza* L. *Biochem J* 314:227–233
- Buchan DWA, Minneci F, Nugent TCO, Bryson K, Jones DT (2013) Scalable web services for the PSIPRED protein analysis workbench. *Nucleic Acids Res* 41:W340–W348
- Chou KC, Shen HB (2007) MemType-2L: a Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem Biophys Res Commun* 360:339–345
- Cichy K, Raboy V (2009) Evaluation and development of low-phytate crops. In: Hari Krishnan B (eds) Crop Science Society of America, Soil Science Society of America, American Society of Agronomy, Madison, pp 177–200
- Cooper JA, Esch FS, Taylor SS, Hunter T (1984) Phosphorylation sites in enolase and lactate dehydrogenase utilized by tyrosine protein kinases in vivo and in vitro. *Biol Chem* 259:7835–7841
- Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J (2006) CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res* 34:W116–W118
- Eisenberg D, Lüthy R, Bowie JU (1997) VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymol* 277:396–404
- Emanuelsson O, Nielsen H, Brunak S, Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Mol Biol* 300:1005–1016
- Feldbrugge M, Hahlbrock K, Weisshaar B (1996) The transcriptional regulator CPRF1: expression analysis and gene structure. *Mol Gen Genet* 251:619–627
- Felle HH, Hepler PK (1997) The cytosolic Ca^{2+} concentration gradient of *Sinapis alba* root hairs as revealed by Ca^{2+} selective microelectrode tests and fura-dextran ratio imaging. *Plant Physiol* 114:39–45
- Feng X, Yoshida KT (2004) Molecular approaches for producing low-phytic-acid grains in rice. *Plant Biotechnol* 21:183–189
- Ferrè F, Clote P (2005) DiANNA: a web server for disulfide connectivity prediction. *Nucleic Acids Res* 33:W230–W232
- Fileppi M, Galasso I, Tagliabue G, Daminati M, Campion B, Doria E, Sparvoli F (2010) Characterisation of structural genes involved in phytic acid biosynthesis in common bean (*Phaseolus vulgaris* L.). *Mol Breed* 25:453–470
- Franklin-Tong VE, Drobak BK, Allan AC, Trewavas AJ (1996) Growth of pollen tubes of *Papaver rhoeas* is regulated by a slow moving calcium wave propagated by inositol 1,4,5-trisphosphate. *Plant Cell* 8:1305–1321
- Frederick JP, Mattiske D, Wofford JA, Megosh LC, Drake LY, Chiou ST, Hogan BLM, York JD (2005) An essential role for an inositol polyphosphate multikinase, Ipk2, in mouse embryogenesis and second messenger production. *Proc Natl Acad Sci* 102:8454–8459
- Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A (2005) Protein identification and analysis tools on the ExPASy server. In: Walker JM (ed) The proteomics protocols handbook. Humana Press, pp 571–607
- Geourjon C, Deléage G (1995) SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput Appl Biosci* 11:681–684
- Gill SC, Von Hippel PH (1989) Calculation of protein extinction coefficients from amino acid sequence data. *Anal Biochem* 182:319–326
- Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* 40:D1178–D1186
- Guruprasad K, Reddy BVB, Pandit MW (1990) Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng* 4:155–161
- Hanakahi LA, Bartlett-Jones M, Chappell C, Pappin D, West SC (2000) Binding of inositol phosphate to DNA-PK and stimulation of double-strand break repair. *Cell* 102:721–729
- Hofmann K, Stoffel W (1993) TMbase—a database of membrane spanning proteins segments. *Biol Chem* 374:166
- Holmes W, Jogi G (2006) Crystal structure of inositol phosphate multikinase 2 and implications for substrate specificity. *Biol Chem* 281:38109–38116
- Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K (2007) WoLF PSORT: protein localization predictor. *Nucl Acids Res* 35:W585–W587
- Hua S, Sun Z (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* 17:721–728
- Ikai AJ (1980) Thermostability and aliphatic index of globular proteins. *J Biochem* 88:1895–1898
- Irvine RF, Schell MJ (2001) Back in the water: the return of the inositol phosphates. *Nat Rev Mol Cell Biol* 2:327–338
- Josefsen L, Bohn L, Sorensen M, Rasmussen S (2007) Characterization of a multifunctional inositol phosphate kinase from rice and barley belonging to the ATP-grasp superfamily. *Gene* 397:114–125
- Kawabata T (2003) MATRAS: a program for protein 3D structure comparison. *Nucleic Acids Res* 31:3367–3369
- Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* 10:845–858

- Kim SR, Kim Y, An G (1993) Identification of methyl jasmonate and salicylic acid response elements from the nopaline synthase (nos) promoter. *Plant Physiol* 103:97–103
- Klaholz B, Moras D (2002) C–H ... O hydrogen bonds in the nuclear receptor RAR γ —a potential tool for drug selectivity. *Structure* 10:1197–1204
- Kozakov D, Grove LE, Hall DR, Bohnuud T, Mottarella SE, Luo L, Xia B, Beglov D, Vajda S (2015) The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat Protoc* 10:733–755
- Krishnan V, Jain P, Tripathi V, Hada A, Manickavasagam M, Ganapathi A, Rai RD, Sachdev A (2015) Molecular modeling and in-silico characterization of Glycine max inositol (1,3,4) tris 5/6 kinase-1 (Gmitpk1)—a potential candidate gene for developing low phytate transgenics. *Plant Omics* 8:381–391
- Kuwano M, Mimura T, Takaiwa F, Yoshida KT (2009) Generation of stable ‘low phytic acid’ transgenic rice through antisense repression of the 1D-myo-inositol 3-phosphate synthase gene using the 18-kDa oleosin promoter. *Plant Biotechnol* 7:96–105
- Kyte J, Doolittle RF (1982) A simple method for displaying the hydrophobic character of a protein. *J Mol Biol* 157:105–132
- Lescot M, Dehais P, Thijs G, Marchal K, Moreau Y, Van de Peer Y, Rouze P, Rombauts S (2002) PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res* 30:325–327
- Letunic I, Bork P (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucl Acids Res* 44:W242–W245
- Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using realtime quantitative PCR and the $2^{-\Delta\Delta C(T)}$ method. *Methods* 25(4):402–408
- Lois R, Dietrich A, Hahlbrock K, Schulz W (1989) A phenylalanine ammonia-lyase gene from parsley: structure, regulation and identification of elicitor and light responsive cis-acting elements. *EMBO J* 8:1641–1648
- Lopez-Ochoa L, Acevedo-Hernández G, Martínez-Hernández A, Argüello-Astorga G, Herrera-Estrella L (2007) Structural relationships between diverse cis-acting elements are critical for the functional properties of a rbcS minimal light regulatory unit. *J Exp Bot* 58:4397–4406
- Lovell SC, Davis IW, Arendall WB III, Bakker PIW, Word JM, Prisant MG, Richardson JS, Richardson DC (2003) Structure validation by C_α geometry: Φ , Ψ and C_β deviation. *Proteins* 50:437–450
- Majerus PW (1992) Inositol phosphate biochemistry. *Annu Rev Biochem* 61:225–250
- Majerus PW (1996) Inositols do it all. *Genes Dev* 10:1051–1053
- Malho R (1998) Role of 1,4,5-inositol trisphosphate-induced Ca^{2+} release in pollen tube orientation. *Sex Plant Reprod* 11:231–235
- Matsuno K, Fujimura T (2014) Induction of phytic acid synthesis by abscisic acid in suspension-cultured cells of rice. *Plant Sci* 217–218:152–157
- Mongkolsiriwatana C, Pongtongkam P, Peyachoknagul S (2009) In silico promoter analysis of photoperiod-responsive genes identified by DNA microarray in rice (*Oryza sativa* L.). *Kasetsart J (Nat Sci)* 43:164–177
- Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ (2009) Autodock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J Comput Chem* 16:2785–2791
- Murphy AM, Otto B, Brearley CA, Carr JP, Hanke DE (2008) A role for inositol hexakisphosphate in the maintenance of basal resistance to plant pathogens. *Plant J* 56:638–652
- Neumann GM, Thomas I, Polya GM (1996) Identification of the site on potato carboxypeptidase inhibitor that is phosphorylated by plant calcium-dependent protein kinase. *Plant Sci* 114:45–51
- Nguyen T, Sherratt PJ, Cecil B, Pickett CB (2003) Regulatory mechanisms controlling gene expression mediated by the antioxidant response element. *Annu Rev Pharmacol Toxicol* 43:233–260
- Nishikawa K, Tokar A, Johannes FJ, Songyang Z, Cantley LC (1997) Determination of the specific substrate sequence motifs of protein kinase C isozymes. *Biol Chem* 272:952–960
- Niu J, Wang J, Hu H, Chen Y, An J, Cai J, Sun R, Sheng A, Liu X, Lin S (2016) Cross-talk between freezing response and signalling for regulatory transcriptions of MIR475b and its targets by miR475b promoter in *Populus suaveolens*. *Sci Rep* 6:20648
- Nunes ACS, Vianna GR, Cuneo F, Amaya-Farfán J, de Capdeville G, Rech EL, Aragao FJL (2006) RNAi mediated silencing of the myo-inositol-1-phosphate synthase gene (GmMIPS1) in transgenic soybean inhibited seed development and reduced phytate content. *Planta* 224:125–132
- Odom AR, Stahlberg A, Went SR, York JD (2000) A role for nuclear inositol 1,4,5-trisphosphate kinase in transcriptional control. *Science* 287:2026–2029
- Omasits U, Ahrens CH, Müller S, Wollscheid B (2014) Protter: interactive protein feature visualization and integration with experimental proteomic data. *Bioinformatics* 30:884–886
- Onodera Y, Suzuki A, Wu CY, Washida H, Takaiwa F (2001) A rice functional transcriptional activator, RISBZ1, responsible for endosperm-specific expression of storage protein genes through GCN4 motif. *J Biol Chem* 276:14139–14152
- Pagni M, Ioannidis V, Cerutti L, Zahn-Zabal M, Jongeneel CV, Hau J, Martin O, Kuznetsov D, Falquet L (2007) MyHits: improvements to an interactive resource for analyzing protein sequences. *Nucleic Acids Res* 35:W433–W437
- Pashou EE, Litou ZI, Liakopoulos TD, Hamodrakas SJ (2004) WaveTM: wavelet-based transmembrane segment prediction. *In Silico Biol* 4:127–131
- Petersen TN, Brunak S, Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* 8:785–786
- Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF chimera—a visualization system for exploratory research and analysis. *J Comput Chem* 25(13):1605–1612
- Pierson ES, Miller DD, Callahan DA, Shipley MA, Rivers BA, Cresti M, Hepler PK (1994) Pollen tube growth is coupled to the extracellular calcium ion flux and the intracellular calcium gradient: effect of BAPTA-type buffers and hypertonic media. *Plant Cell* 6:1815–1828
- Plevin MJ, Mills MM, Ikura M (2005) The LxxLL motif: a multifunctional binding sequence in transcriptional regulation. *Trends Biochem Sci* 30:66–69
- Pronk S, Páll S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, Shirts MR, Smith JC, Kasson PM, Spoel D, Hess B, Lindahl E (2013) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* 29:845–854
- Raboy V, Dickinson DB (1987) The timing and rate of phytic acid accumulation in developing soybean seeds. *Plant Physiol* 85:841–844
- Reehana N, Ahamed AP, Ali DM, Suresh A, Kumar RA, Thajuddin N (2013) Structure based computational analysis and molecular phylogeny of C-Phycocyanin gene from the selected cyanobacteria. *Int J Biol Vet Agric Food Eng* 7:47–51
- Rice P, Longden I, Bleasby A (2000) Emboss: the European molecular biology open software suite. *Trends Genet* 16:276–277
- Roberts DM, Harmon AC (1992) Calcium modulated proteins: targets of intracellular calcium signals in higher plants. *Annu Rev Plant Physiol Plant Mol Biol* 43:375–414
- Roche DB, Buenavista MT, McGuffin LJ (2013) The FunFOLD2 server for the prediction of protein-ligand interactions. *Nucleic Acids Res* 41:W303–W307

- Rogers S, Wells R, Rechsteiner M (1986) Amino acid sequences common to rapidly degraded proteins: the PEST hypothesis. *Science* 234:364–368
- Rost B, Sander C (1993) Prediction of protein secondary structure at better than 70% accuracy. *Mol Biol* 232:584–599
- Rost B, Casadio R, Fariselli P (1996) Refining neural network predictions for helical transmembrane proteins by dynamic programming. *Proc Int Conf Intell Syst Mol Biol* 4:192–200
- Rouster J, Leah R, Mundy J, Cameron-Mills V (1997) Identification of a methyl jasmonate-responsive region in the promoter of a lipoxygenase 1 gene expressed in barley grain. *Plant J* 11:513–523
- Saiardi A, Erdjument-Bromage H, Snowman A, Tempst P, Snyder SH (1999) Synthesis of diphosphoinositol pentakisphosphate by a newly identified family of higher inositol polyphosphate kinases. *Curr Biol* 9:1323–1326
- Saiardi A, Caffrey JJ, Snyder SH, Shears SB (2000) Inositol polyphosphate multikinase (ArgR11) determines nuclear mRNA export in *Saccharomyces cerevisiae*. *Biol Chem J* 275:24686–24692
- Schwartz R, Ting CS, King J (2001) Whole proteome pI values correlate with subcellular localizations of proteins for organisms within the three domains of life. *Genome Res* 11(5):703–709
- Sessa G, Meller Y, Fluhr R (1995) A GCC element and a G-box motif participate in ethylene-induced expression of the PRB-1b gene. *Plant Mol Biol* 28:145–153
- Shears SB (2004) How versatile are inositol phosphate kinases? *Biochem J* 377:265–280
- Shen X, Xiao H, Ranallo R, Wu WH, Wu C (2003) Modulation of ATP-dependent chromatin-remodeling complexes by inositol polyphosphates. *Science* 299:112–114
- Shinozaki K, Yamaguchi-Shinozaki K (2000) Molecular responses to dehydration and low temperature: differences and cross-talk between two stress signaling pathways. *Curr Opin Plant Biol* 3:217–223
- Steger DJ, Haswell ES, Miller AL, Wente SR, O'Shea EK (2003) Regulation of chromatin remodeling by inositol polyphosphates. *Science* 299:114–116
- Steinberg SF (2008) Structural basis of protein kinase C isoform function. *Physiol Rev* 88:1341–1378
- Stevenson-Paulik J, Odom AR, York JD (2002) Molecular and biochemical characterization of two plant inositol polyphosphate 6-/3-/5-kinases. *Biol Chem J* 277:42711–42718
- Stevenson-Paulik J, Bastidas GJ, Chiou ST, Frye RA, York JD (2005) Generation of phytate-free seeds in *Arabidopsis* through disruption of inositol polyphosphate kinases. *Proc Natl Acad Sci* 102:12612–12617
- Stiles AR (2007) Identification and characterization of late pathway enzymes in phytic acid biosynthesis in *Glycine max*. Dissertation, Virginia Polytechnic Institute and State University
- Stiles A, Qian X, Shears S, Grabau E (2008) Metabolic and signaling properties of an ITPK gene family in *Glycine max*. *FEBS Lett* 582:1853–1858
- Sugimoto T, Kawasaki T, Kato T, Whittier RF, Shibata D, Kawamura Y (1992) cDNA sequence and expression of a phosphoenolpyruvate carboxylase gene from soybean. *Plant Mol Biol* 20:743–747
- Sun Y, Thompson M, Lin G, Butler H, Gao Z, Thornburgh S, Yau K, Smith D, Shukla V (2007) Inositol 1,3,4,5,6-pentakisphosphate 2-kinase from maize: Molecular and biochemical characterization. *Plant Physiol* 144:1278–1291
- Suzuki M, Tanaka K, Kuwano M, Yoshida K (2007) Expression pattern of inositol phosphate-related enzymes in rice (*Oryza sativa* L.): Implications for the phytic acid biosynthetic pathway. *Gene* 405:55–64
- Sweetman D, Johnson S, Caddick S, Hanke D, Brearley C (2006) Characterization of an *Arabidopsis* inositol 1,3,4,5,6-pentakisphosphate 2-kinase (AtIPK1). *Biochem J* 394:95–103
- Sweetman D, Stavridou I, Johnson S, Green P, Caddick S, Brearley C (2007) *Arabidopsis thaliana* inositol 1,3,4-trisphosphate 5/6-kinase 4 (AtITPK4) is an outlier to a family of ATP-grasp fold proteins from *Arabidopsis*. *FEBS Lett* 581:4165–4171
- Takaiwa F, Yamanouchi U, Yoshihara T, Washida H, Tanabe F, Kato A, Yamada K (1996) Characterization of common cis-regulatory elements responsible for the endosperm-specific expression of members of the rice glutelin multigene family. *Plant Mol Biol* 30:1207–1221
- Tamura K, Stecher G, Peterson D, Filipowski A, Kumar S (2013) MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* 30:2725–2729
- Trott O, Olson AJ (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *Comput Chem* 31:455–461
- Turner PJ (2005) XMGRACE, version 5.1.19. center for coastal and land-margin research. Oregon Graduate Institute of Science and Technology, Beaverton
- Tuteja JH, Clough SJ, Chan WC, Vodkin LO (2004) Tissue-specific gene silencing mediated by a naturally occurring chalcone synthase gene cluster in *glycine max*. *Plant Cell* 16:819–835
- Wallace IM, Sullivan O, Higgins DG, Notredame C (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res* 34:1692–1699
- Washida H, Wu CY, Suzuki A, Yamanouchi U, Akihama T, Harada K, Takaiwa F (1999) Identification of cis-regulatory elements required for endosperm expression of the rice storage protein glutelin gene GluB-1. *Plant Mol Biol* 40:1–12
- Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25(9):1189–1191
- Webb B, Sali A (2016) Comparative protein structure modeling using modeller. *Curr Protoc Bioinform* 54:5.6.1–5.6.37
- Wiederstein M, Sippl MJ (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res* 35:W407–W410
- Wu CY, Adach T, Hatano T, Washida H, Suzuki A, Takaiwa F (1998) Promoters of rice seed storage protein genes direct endosperm-specific gene expression in transgenic rice. *Plant Cell Physiol* 39:885–889
- Wymer CL, Bibikova TN, Gilroy S (1997) Cytoplasmic free calcium distribution during the development of root hairs of *Arabidopsis thaliana*. *Plant J* 12:427–439
- Xu J, Brearley CA, Lin WH, Wang Y, Ye R, Mueller-Roeber B, Xu ZH, Xue HW (2005) A role of *Arabidopsis* inositol polyphosphate kinase, AtIPK2 α , in pollen germination and root growth. *Plant Physiol* 137:94–103
- Yamaguchi-Shinozaki K, Mundy J, Chua NH (1989) Four tightly linked rab genes are differentially expressed in rice. *Plant Mol Biol* 14:29–39
- York JD, Odom AR, Murphy R, Ives EB, Wente SR (1999) A phospholipase C-dependent inositol polyphosphate kinase pathway required for efficient messenger RNA export. *Science* 285:96–100
- Yu CS, Chen YC, Lu CH, Hwang JK (2006) Prediction of protein subcellular localization. *Proteins* 64:643–651
- Zhang ZB, Yang G, Arana F, Chen Z, Li Y, Xia HJ (2007) *Arabidopsis* inositol polyphosphate 6-/3-Kinase (AtIpK2 β) is involved in axillary shoot branching via auxin signaling. *Plant Physiol* 144:942–951