Plant Omics Journal POJ

POJ 8(5):381-391 (2015) ISSN:1836-3644

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

Veda Krishnan[1,3], Priyanka Jain[2], Vinutha T[1], Alkesh Hada[1,3], Manickavasagam M[3], Ganapathi A[3], Raj D Rai[1]and Archana Sachdev[1,*]

[1]Division of Biochemistry, Indian Agricultural Research Institute (IARI), New Delhi, India
[2]National Research Centre on Plant Biotechnology (NRCPB), New Delhi, India
[3]Department of Biotechnology and Genetic Engineering, Bharathidasan University, Tiruchirappalli, India

***Corresponding author: arcs_bio@yahoo.com**

Abstract

Inositol (1, 3, 4) tris 5/6 kinase (*Itpk*) is a key player in lipid-independent pathway of phytate biosynthesis. In this study, the full length coding sequence of *Gmitpk*1 was cloned and blasted to retrieve the available inositol phosphate kinases (ipks) from the public domain. Sequence analysis of the selected 18 plant kinases revealed a consensus 'ATP-grasp' domain. Secondary structure predictions showed high alpha helix content (80.02%) which justified the structural flexibility as well as the versatility of these kinases. Homology modelling of the *Gm*itpk1 performed using the template crystal structure of inositol tetrakisphosphate-1-kinase (2q7d.1.A) from *Homo sapiens* revealed the presence of N and C-terminal domains with a mixed α/β topology having an active site located in the deep cleft between the domains. The model was further refined using intrinsic dynamic tools like ProSA, Verify3D, WEBnm and ElNemo. This study has enabled the elucidation of the 3Dstructure of *Gm*itpk1and the data has been submitted to protein model data base (PMD) - PM0079572 (first report). Ligand binding residues and energy computations have revealed Mg^{2+}, ATP and ADP as the most likely ligands for *Gm*itpk1. The study throws light on some novel insights into the structural features of *Gm*itpk1, a potential candidate for developing low phytate transgenic soybean.

Keywords: Inositol (1, 3, 4) tris 5/6 kinase – 1; soybean; phytate; cloning; molecular modeling; intrinsic dynamics.
Abbreviations: ITPK_Inositol (1, 3, 4) tris 5/6 kinase; ORF_Open reading frame; EST_Expressed sequence tags; NJ_Neighbor-joining; MW_Molecular weight; pI_Isoelectric pH; EC_Extinction coefficient; Ai_Aliphatic index; Ii_Instability index; GRAVY_Grand average hydropathy; PMD_Protein model database; NMA_Normal mode analysis.

Introduction

Phytic acid ($InsP_6$) is the primary storage compound of phosphorus in soybean seeds accounting for up to 80% of the total seed phosphorus and contributing as much as 1.5% to the seed dry weight. Soybean is widely used as a primary source of protein for human consumption as well as, as an animal feed; but the inability of humans and other monogastrics to digest $InsP_6$ has anti-nutrient effects due to its property to chelate mineral cations such as iron, zinc, calcium and potassium (Raboy et al., 2001). The undigested $InsP_6$ excreta, which are used as manures, lead to phosphorous pollution and eutrophication (Abelson et al., 1999). Therefore there is a need to generate low phytate soybean using molecular tools. Researchers are therefore currently underway to develop low phytate transgenics using metabolic pathway engineering of several inositol intermediate kinases like inositol multi kinases, the reversible Ins $(1,3,4)P_3$ and Ins $(3,4,5,6)P_4$ kinases and diphosphoinositol polyphosphate kinases, while the catalytic versatility and the structural complexities of these intermediate ipks is a major limitation. Among this, Inositol (1,3,4) tris 5/6 kinase (Itpk)-1 [EC: 2.7.1.159] plays a key role in the lipid-independent pathway of phytate biosynthesis and has been characterized previously in *Arabidopsis* (Sweetman et al., 2007), rice (Josefsen et al., 2007) and soybean (Amanda et al., 2008). The structural characteristics of this enzyme are however not yet known and the amino acid residues that contribute to its catalytic site have not been fully deciphered yet. Besides, the enzyme has no major sequence similarities at the nucleotide level with other Ipks of phytate pathway across the domain. In soybean, the *itpk* gene family comprises of four isoforms (Amanda et al., 2008) and are best known for their Ins $(1,3,4)P_3$ 5/6 kinase and Ins $(3,4,5,6)P_4$-1 kinase activities (Yang et al., 2000). Amanda (2008) studied the K_m values of itpk's for their substrates and verified that they preferentially phosphorylated the position-1 of Ins $(3,4,5,6)P_4$. The K_m values for either Ins $(3,4,5,6)P_4$ or Ins $(1, 3, 4)P_3$ as substrate (0.1-0.5µM) were found to be identical with the K_i values determined, when each substrate was found to inhibit the phosphorylation of the other. It is therefore reasonable to assume that the enzyme uses a single catalytic site to perform all its ipk activities making it further relevant to study the catalytic residues and the combinatorial mode of ligand recognition which are not yet explored. It has how ever been suggested that the catalytic promiscuity towards different inositol phosphates is not an evolutionary compromise, it is rather exploited to facilitate the tight regulation of the metabolic pathway (Stephen et al., 2004) making it further intriguing as to how a biologically active inositol phosphate interacts with the target enzyme (itpk1) in a highly specific manner and what is the structural basis of

this specificity? The present study was undertaken using an *in silico* approach to gain some insights into the structural features of *Gm*itpk1.

Results and Discussion

Cloning of Glycine max Itpk1

Itpk1 cDNA was amplified from *Glycine max*. Pusa 16 cultivar, by designing primers from the retrieved sequence (NM_001250900.1) available in NCBI. The 1053 bp fragment was cloned in pGEM-T vector system, sequenced and submitted in NCBI as KF913641 (Fig. 1). The full length cDNA sequence showed an Open Reading Frame (ORF) of 1053 nucleotides that could potentially code for a single polypeptide.

Nature of amino acid composition, disulfide bond topology and physiochemical parameter analysis

Table 1 show sequences of the selected 18 *ipk*s retrieved from NCBI on blasting the *Gm*itpk1 as a query sequence. The primary structural analysis of the selected kinases was done by computing different parameters using ExPASy's ProtParam and PEPSTATS tools and tabulated as Table 2. The analysis revealed Leu as the most abundant amino acid, accounting for 10% of the enzyme's primary structure in all the kinases under study. The least common amino acids were Trp and Cys. The predicted average MW calculated was 40658.4 g mol^{-1} (Table 2). The computed average pI of the ipks was 6.3 indicating that the enzyme is likely to precipitate in acidic buffers (Table 2) and this information will be very much useful for developing buffer systems for the purification of recombinant proteins by isoelectric focussing. Stability of ipks was studied by analysing the values for Ii, Ai and GRAVY. The Ii of these plant enzymes were found to be stable in *Medicago truncatula* (37.77), *Solanum lycopersicum (34.86)* and *S. tuberosum* (34.96) as they were below 40. According to the ProtParam server, a protein whose Ii is larger than 40 may be unstable (Gasteiger et al., 2005). GRAVY scores indicate the relative value for the hydrophobic residues of the protein. Although the positional or interaction effects of the adjacent residues are not taken into consideration, it provides some indication about the physical state of the protein (Vriend et al., 1990). GRAVY index of these plant kinases was shown to be in a range from -0.0 to -0.5 (Table 2). All ipks had negative GRAVY scores attesting to their solubility in hydrophilic solvents. The Ai which evaluates the relative volume occupied by the side chains of hydrophobic amino acids (Ala, Val, Ile, Leu) ranged from 84.18 – 96.86, which might be acting as a positive factor for increasing the thermal stability of proteins. Among the selected ipks, the EC value was found to be the highest in *Solanum* sps reflecting high concentration of Cys, Trp and Tyr whereas it was least observed in *Cicer arietinum* (Table 2). Although ExPASy's ProtParam tool computes the EC for a range of (276,278,279,280,282nm), 280 nm is favoured because proteins absorb strongly at this wavelength and other substances commonly in protein solutions do not interfere. EC of ipks ranged from 20525 to 36120 $M^{-1}cm^{-1}$ with respect to the concentration of Cys, Trp and Tyr. The subcellular location of a protein being closely correlated to its biological function, we predicted using Cello 4 and found that these kinases were cytoplasmic. Disulfide predictions using DiANNA server revealed, cysteine residues were prevalent in all kinases except in *Hordeum vulgare* which reflected upon the absence of disulphide bridges in it as compared to other ATP dependent enzymes (Table 3).

Conserved domain identification and phylogenetic analysis

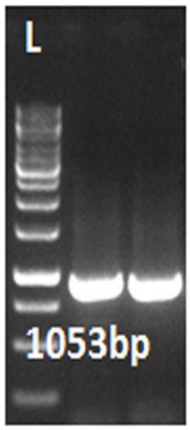
Clustal W2 server was used to multiple align the selected amino acid sequences of inositol phosphate kinases (Supplementary Fig. 1). Most significantly conserved sequences among all the selected plant species included RLHNRISML, ADGSAKSHKM. The tri-peptide sequences FGIP, IAKP, VLQE and VNHG were also found to be conserved and indicated the beta stranded nature of these kinases. The predictive modelling studies and BLAST analysis suggested that itpk's have a structure similarity to the members of "ATP-grasp" family, in which ATP is normally held between two anti-parallel beta sheets (Cheek et al., 2002). The comparison suggested that there might be a relatively close evolutionary link between inositol kinases. Phylogenetic analysis provides a useful framework to understand the relationship of different forms and how they have evolved from a common ancestor. Hence a phylogenetic tree was constructed from the aligned protein sequences (Fig. 2) using NJ method and it indicated that the enzymes belonged to two major branches or clades. First clade consisted of two sub clades; with Fabales, Rosales, Myrtales, Cucurbitales, Malphigiales, Vitales in the first and Solanales completely occupied the second sub clade. The itpk1 homologue from *Hordeum vulgare* formed a distinct group (second clade) which is reflected by its amino acid composition as well as by the absence of disulphide bridges in it. The phylogram thus illustrated a high degree of divergence among the legumes, cereals, and the horticultural crops.

Secondary structures with α & β topology identification

The predicted secondary structure composition of itpk1 from *Glycine max* was determined using CFSSP server which generates a consensus report from twelve secondary structure prediction methods (Peter et al., 1974). More detailed analysis of the secondary structural elements (Table 3) and the schematic diagram showing secondary motif map (Fig. 3 A) was performed using PDBsum tool. The secondary structure prediction server revealed that 80.02 % of amino acids resided in the alpha helix conformation, whereas 54.88 % of residues were in beta sheets and 13.34 % in coils. This high percentage of helices in the structure provides an exceptional flexibility to inositol kinases. Alexandr (2008) also emphasized on the role of α helix scaffold for the assembly of active protein kinases. Low content of coils as structural components is due to lack of amino acids such as Gly and Pro and the latter has a unique property of creating kinks in the polypeptide chains disrupting the ordered secondary structure. The rest of the amino acids were found in other conformations such as beta hair pins and beta turns. The topology of itpk1 is illustrated in Fig. 3 (B). The protein's main chain was found to consist of a C-terminal and N-terminal domains that were both folded into a mixed α/β topology. The predicted secondary structures generated by the PDBsum tool were generally in agreement with the three dimensional structures. However, while the enzyme's schematic diagram illustrated the existence of α-helices and β-sheets in the N-terminal between residues 7 and 40, no such structures were found in the corresponding region of the 3D model. This may be due to the fact that ipks, across species, align poorly in this particular area of the N-terminal domain. The Motif scan tool identified the phosphorylation sites (Table 4). Predominantly phosphorylation sites of inositol 1, 3, 4-trisphosphate 5/6-kinases were predicted from

Table 1. Inositol phosphate kinase sequences retrieved from NCBI database (htp:www.ncbi.nlm.nih.gov/)

Organism	Accession number	Nucleotide specificity
Glycine max	AHJ25994.1	ATP
Phaseolus vulgaris	XP_007156414.1 (ID: 18636338)	ATP
Cicer arietinum	XP_004509737.1 (ID: 101510649)	ATP
Medicago truncatula	XP_003628787.1 (ID: 11428245)	ATP
Morus notabilis	EXB66541.1	ATP
Prunus mume	XP_008238214.1 (ID: 103336869)	ATP
Malus domestica	XP_008373591.1 (ID: 103436916)	ATP
Prunus persica	XP_007210517.1 (ID: 18777458)	ATP
Fragaria vesca	XP_004300416.1 (ID: 101312158)	ATP
Theobroma cacao	XP_007040636.1 (ID: 18606781)	ATP
Eucalyptus grandis	KCW77969.1	ATP
Vitis vinifera	XP_002279736.1 (ID: 100242009)	ATP
Populus trichocarpa	XP_006368380.1(ID: 18094146)	ATP
Cucumis melo	XP_008448667.1 (ID: 103490768)	ATP
Cucumis sativus	XP_004148764.1 (ID: 101211366)	ATP
Solanum lycopersicum	XP_004244396.1 (ID: 101262540)	ATP
Solanum tuberosum	XP_006367451.1 (ID: 102593136)	ATP
Hordeum vulgare	CAL49035.1	ATP



>gi|585087515|gb|KF913641.1| Glycine max cultivar Pusa 16 inositol-1,4,3-tris phosphate 5/6 kinase mRNA, complete cds

ATGTCAGAGCGGTTACGAAGCCCAGTTATTGTATCGACTCACTATAGGGCGATTGCGACGTTGCATGCTT
CCCGTCGCATGCGCGCGGGACTCGATATGCGGAGAAGAGATCGGCGTGATAGGCGAATCGCGAGGGATGA
GCTCGTGCGCGTGGACTCCGACAAGCCCCTCGCGGATCAGGGCCCCTTCGACTGCGTCCTCCACAAGCTC
TACGGCGACGACTGGAAGCGCCAGCTCCAGGAATTCCACACCCTCTACCCAAACGCCGTCATCCTCGACG
CCCCCGAGGCGATCGAGCGCCTCCACAACCGAATCTCGATGCTGCAGGTCGTGTCGGAGCTCCGAATTGA
GGACCGACCCGAGACCTTTGGGATCCCCAAGCAGATTGTGATATACGACAAGGCCACCCTCCTGGACCCG
CAGGCCTGGGAGAGCCTCAAGTTTCCCGTGATAGCGAAGCCTCCCGTTGCCGACGGCAGCGCCAAGTCGC
ACAAGATGGCGCTCGTGTTCACGCGCGACGCACTGAACAAGCTCAAGCCGCCGATCGTGCTGCAGGAGCT
CGTGAACCACGGTGGCGTGATCTTCAAGGTGTACGTGGTGGGCGAGCACGTGCGCTGCGTGAAGCGCAAG
TCCCTCCCCGACGTGTCCGACGAGGAGAAGGCGCTCGGCGGCGTTTCCGAGGACCTGATGTCGTTCTCGC
AGGTCTCCAACCTCGCCACCGTGAACGACTGCGACGGCTATTACCGTTTGATGCATCTCGACGACGACAC
CGAAATGCCCCCCGACGCCTTCGTCGTCGACATCGCTGGGGGCCTCAGGCGCGCCCTCAAGCTCAACCTC
TTCAATTTCGATGTCATAAGAGACGCTCGATACGGGAACCGTTATCTTATCATTGATATCAACTATTTCC
CTGGGTACGCCAAAATGCCCGGCTACGAGGCCGTTTTGACCCAGTTTTTCTGCGAAGTAATGCTCAAGAA
GAAACAGCGAGAGGAACAACAGCAAGAAGAGGGTAACGCTCCTAAGGAGAAAGAGGAATCTCTTCAAGCT
TGA

(A) (B)

Fig 1. Cloning of *Gmitpk* 1 cDNA in pGEMT easy vector system. (A) PCR amplification of 1053bp fragment isolated from *Glycine max cv*. Pusa 16 using gene specific primers and separated on 1% agarose gel along with 1 kb molecular weight marker. (B) *Gmitpk* 1 sequence submitted in NCBI (KF913641).

amino acid residues spanning from 20-322. An extension like protein repeat was also predicted from 149-158 with very low Raw and N scores. Cheek (2002) classified itpk as group 1 kinases along with S/T-Y protein kinase-like/SAICAR synthase/atypical protein kinases having similar phosphorylation domains assisted by binding of ATP. It is also noteworthy that these kinases showed more similarity in their N terminal domain compared to the C-terminal domain, except the region containing the three-stranded anti-parallel β-sheet and associated α-helices, which included the αββ unit which is essential for nucleotide binding (Grishin et al., 1999).

Protein disorder analysis

The total intrinsic protein disorder was detected by PrDOS showed that total disordered amino acid residues were 34 (10.3%). However, they were spread over the protein in 3 regions (Fig. 4). The longest disordered region was spread from Ser216 to Leu250. The N-terminal residues responsible for ATP binding motifs predicted by 'My Hits', were not disordered. Disordered regions differ considerably from the well-structured regions of a protein (Wright and Dyson, 1999) and these disordered regions may play significant roles in protein-ligand interactions (Patil and Nakamura, 2006). Our results too indicate that these disordered regions may interact with various ligands providing versatility in the action of kinases which may be facilitating the tight regulation of inositol phosphate signalling pathway.

Homology modeling & stereo chemical evaluation of the predicted model

Homology models of proteins are of great interest for planning and analysing biological experiments when no experimental 3D structures are available. But now 3D structures of proteins can be predicted from amino acid sequences by various web based homology modeling servers at different levels of complexity. Evolutionarily, the structures have been found to be more stable and their changes are much slower than the changes in their associated sequences, so that similar sequences adopt practically identical structures and distantly related sequences still fold into similar structures (Chothia and Lesk, 1986). The homology model of itpk1 was generated with SWISS-MODEL software using 2q7d.1.A. (Inositol tetrakisphosph-

Table 2. Parameters of inositol phosphate kinases (plants) calculated using the ProtParam program: molecular weight (MW) (g/mol); isoelectric point (pI); extinction coefficient (EC) (M^{-1} cm^{-1}); instability index (Ii); aliphatic index (Ai); grand average hydropathy (GRAVY); number of negative residues (-R); number of positive residues (+R).

Organism	Seq length	MW	pI	EC	Ii	Ai	GRAVY	-R	+R
Glycine max	350	40136.9	6.77	29130	59.49	90.51	-0.426	53	52
Phaeolusvulgaris	341	38845.6	6.21	27515	41.13	91.70	-0.303	45	48
Cicer arietinum	345	38795.1	5.32	20525	41.72	91.19	-0.318	43	53
Medicagotruncatula	385	43271.9	7.10	26025	37.77	96.86	-0.271	56	56
Morus notabilis	368	42064.1	5.60	30620	51.18	88.18	-0.418	53	60
Prunus mume	359	40829.4	5.14	27640	46.13	89.83	-0.397	58	45
Malus domestica	359	40726.2	5.19	29130	46.88	90.11	-0.350	45	56
Prunus persica	359	40878.4	5.10	27640	48.14	89.55	-0.411	60	46
Fragaria vesca	359	40837.6	5.30	29130	46.92	91.45	-0.363	47	58
Theobroma cacao	369	42304.3	6.49	27765	47.43	87.59	-0.434	56	54
Eucalyptus grandis	331	37616.2	8.11	27765	49.32	87.70	-0.331	44	46
Vitis vinifera	368	42042.9	6.36	36120	46.37	84.18	-0.499	54	52
Populustrichocarpa	359	40981.9	5.57	33390	45.62	93.04	-0.360	55	48
Cucumis melo	363	41584.8	6.42	26150	44.67	88.84	-0.432	54	56
Cucumis sativus	363	41471.1	7.05	26150	42.36	88.02	-0.426	53	53
Solanumlycopersicum	349	39674.1	7.59	33600	34.86	96.56	-0.192	48	49
Solanum tuberosum	363	41488.1	8.53	36580	35.96	96.34	-0.286	50	54
Hordeum vulgare	347	38302.7	5.78	30370	49.94	95.01	-0.110	38	46

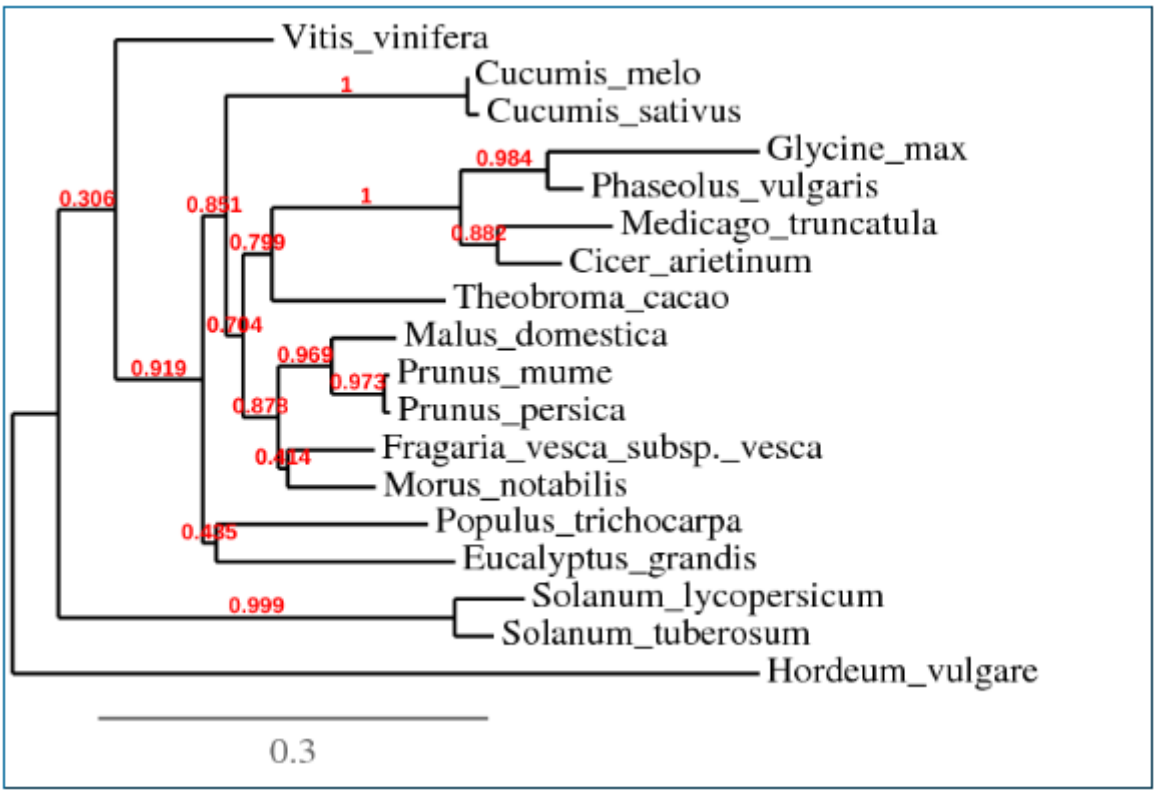


Fig 2. Phylogram of selected inositol phosphate kinases based on the amino acid sequences in Table 1 using Treedyn showing similarity with other relatives. Scale bar indicates nucleotide substitutions per site.

ate-1- kinase from *Homo sapiens*) as the PDB template (Fig. 5). The overall fold of itpk1was found to be similar to that of template as expected with a sequence identity of 27.36 % and ranging from 25-328 with 0.85 query coverage. This final averaged and optimized model passed all the tests implemented in the stereochemistry evaluating with VERIFY3D program, which analyses the compatibility of an atomic model (3D) with its own amino acid sequences (1D). It also uses the contact potentials to assess whether the modelled amino acid residues occur in the environment as typical for globular proteins with hydrophobic core and solvent exposed surface (Eisenberg et al., 1997). Moreover reasonable energies are rarely observed for misfolded structures and thus, the scores reported for our model by VERIFY 3D (average score 0.256) suggests that the 3D model generated is reasonable and geometrically viable. The predicted model was submitted to PMD (accession code - PM0079572). The stereo chemical quality and accuracy of the predicted protein model was verified for further refine-

Table 3. Predicted consensus secondary structure content and predicted disulfide patterns of inositol phosphate kinases. The data was generated using CFSSP (Chou and Fasman Secondary Structure Prediction server) and DiANNA (DiAminoacid Neural Network Application) 1.1 srver.

Organism	Alpha heix (%)	Beeta sheets (%)	Coils (%)	Disulfide bond (Signature Sequences)
Glycine max	82.0	58.0	12.6	ATVND**C**DGYYR - LTQFF**C**EVMLK
Phaseolus vulgaris	85.6	46.9	12.6	GEHVR**C**VKRKS - LTQFF**C**DVMLK
Cicer arietinum	87.2	40.3	15.1	GEHVR**C**VKRKS - LTDFF**C**DLMNNK
Medicago truncatula	84.9	60.8	16.6	GNHVR**C**VKRKS - FVDLM**C**KKELG
Morus notabilis	80.2	32.9	13.9	ADSRG**C**DKEMR - VSNAY**C**SDGDD
Prunus mume	77.7	35.9	13.1	QGPFD**C**VMHKL - GEHVK**C**VKRKS
Malus domestica	77.9	38.9	12.3	YEVLS**C**DDEVR - RKIVS**C**DGDDG
Prunus persica	77.7	35.4	13.1	QGPFD**C**VMHKL - GEHVK**C**VKRKS
Fragaria vesca	81.1	63.0	13.1	LTDFF**C**DIVQK - YEALS**C**DKEVR
Theobroma cacao	81.6	62.3	15.2	LTDFF**C**DIVNR - GLAVE**C**SQEKV
Eucalyptus grandis	80.7	40.5	12.7	KNLAK**C**SGKED - SGKED**C**SEKSK
Vitis vinifera	75.5	60.1	13.6	IPMRS**C**DKDSR - YGSDD**C**CSDGE
Populus trichocarpa	78.3	65.5	13.9	SLLSL**C**KSKGV - QGPFD**C**VLHKL
Cucumis melo	73.8	41.0	15.4	EGRRF**C**IGYAL - QGPFD**C**ILHKF
Cucumis sativus	70.8	43.8	15.4	VFNHD**C**LNKLK - GQYVK**C**VKRKSL
Solanum lycopersicum	87.4	65.3	12.3	QGPFD**C**VLHKL - ELEIK**C**ETTSF
Solanum tuberosum	85.4	67.2	12.1	QGPFD**C**VLHKL - ELEIK**C**ETTSF
Hordeum vulgare	82.7	57.1	11.5	Nil

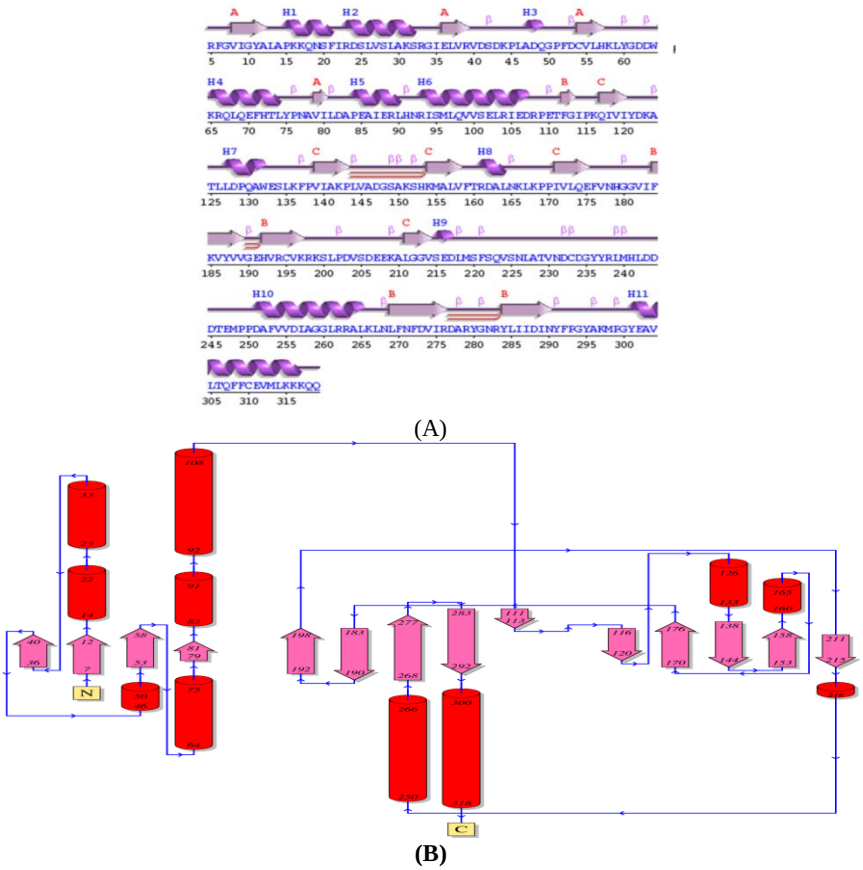


Fig 3. (A) Schematic diagrams showing the secondary structural elements in *Gm*itpk 1. Alpha helices are labeled with letter "H" and beta strands are lettered in upper case "B" and "C" for coils. The secondary motif map was calculated using the PDBsum tool. (B) Topology diagram *Gm*itpk 1 represents Helices as cylinders and beta strands as arrows. "N" and "C" terminal domains are also represented. The topology diagram was calculated using PDBsum tool.

Table 4. Motifs identified using Motif scan.

Motif information	No. of sites	Amino acid residues	Raw score	N score
Inositol 1, 3, 4-trisphosphate	2	20-322,	446.6,	139.40,
5/6-kinase		47-322	476.2	150.984
Extensin-like protein repeat	1	149-158	1.9	7.122

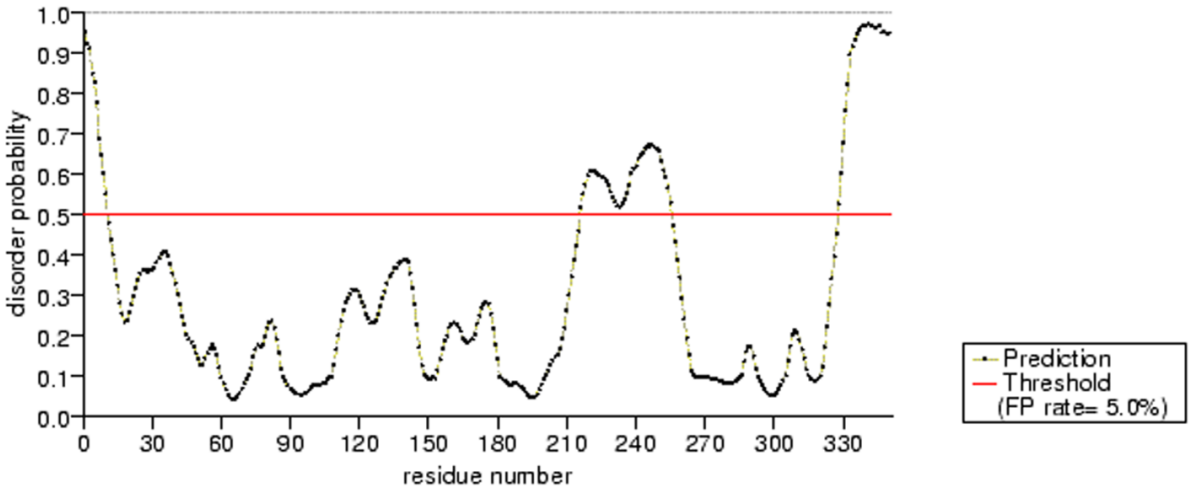


Fig 4. Schematic diagram of disordered regions of *Gm*itpk1 sequence using PrDOS program. and total disordered amino acid residues were 34 (10.3%). The longest disordered region was spread from Ser216 to Leu250 and the N-terminal residues.

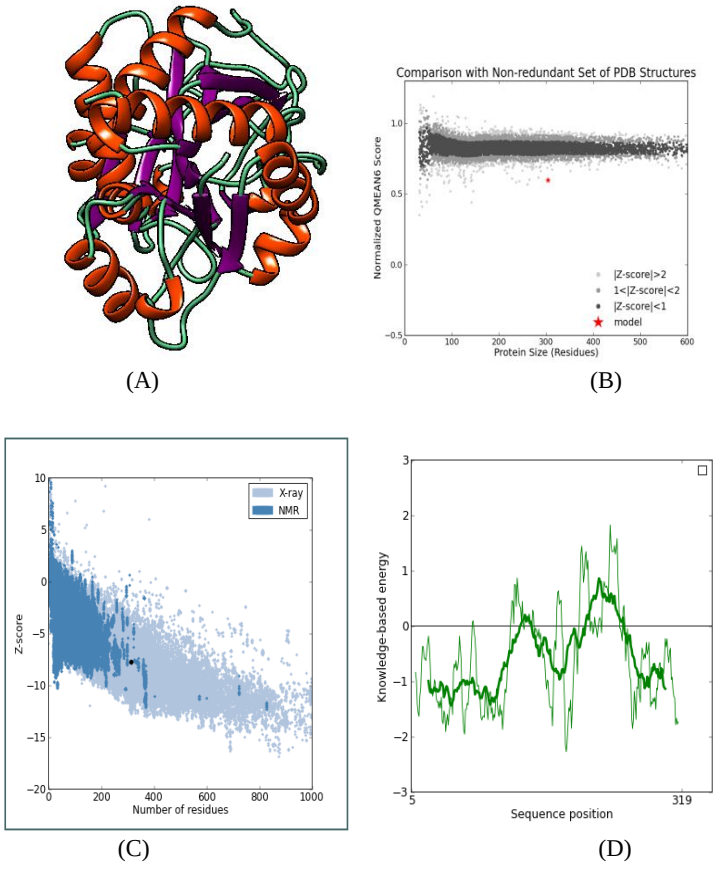


Fig 5. (A) Predicted 3D structure of inositol-1, 3, 4 tris phosphate 5/6 kinase -1. The models were generated with SWISS-MODEL using PDB template 2q7d.1.A and the ribbon shape model was visualized in chimera (B) Comparison of the predicted model with non-redundant set of PDB structures available. (C) Z-score plot from ProSA web server showing the quality of the predicted model in NMR region (dark blue). (D) Energy plot showing all residues of predicted model at very stable position (dark green).

ment process using RAMPAGE. The reliability of the target proteins was examined by torsion angles φ and ψ and; a percentage quality measurement of the protein structure was used, in which four sorts of occupancies were observed - "core, most favoured, generally allowed and outlier regions". According to the Ramachandran plot generated with the RAMPAGE server, 90.7 % of residues were found in the most favoured region, while 7.3% of residues were found to reside in the generously allowed region and 1.9% of residues were found in the outlier regions (Fig. 6). The modelled structure was validated by ProSA for reducing potential errors. This program displays 2 quality measures of the input structure; Z score and a plot of its residue energies. Z score indicates overall model quality and measures the deviation of the total energy of the structure with respect to an energy distribution derived from random conformations. As shown in Figure 4, the Z score of *Gm*itpk1were well within the range of scores (-7.75) typically found for proteins of similar size, indicating a highly reliable structure. The energy plot shows the local model quality by plotting energies as a function of amino acid sequence position (Benkert et al., 2011).

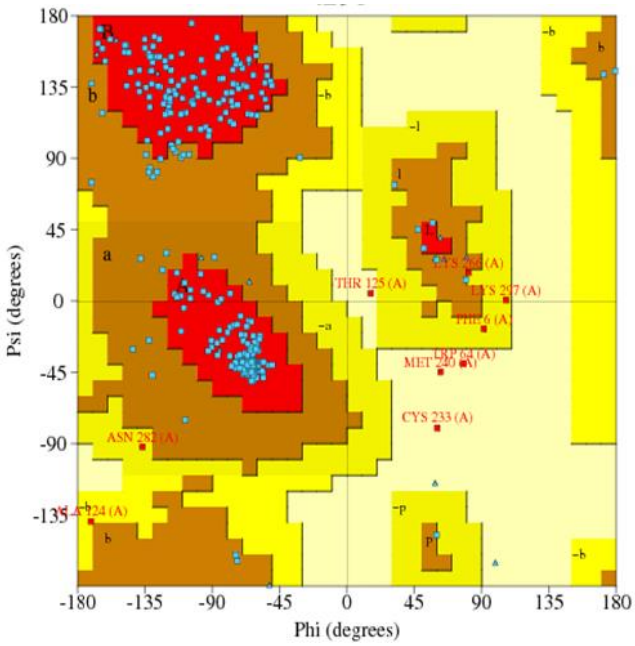


Fig 6. (a) General amino acid representation of Ramachandran plot for the modeling of inositol-1, 3, 4 tris phosphate 5/6 kinase-1. The plot calculations were done by PROCHECK program and the plot reveals 90.7 % of all residues of *Gm*itpk1 in the most favoured regions. 7.3 % of residues are in the most allowed regions and 1.9% of residues are found in the outlier regions. The plot was generated using RAMPAGE program.

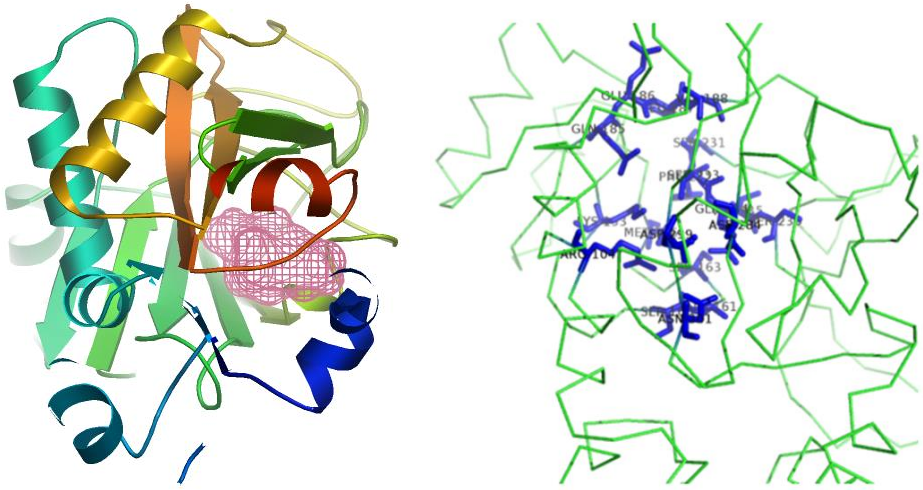
Cleft analysis, ligand predictions and intrinsic dynamics evaluation of ITPK1

The protein is organized into a C-terminal and an N-terminal domain with mixed α/β content. The active site is located in a deep cleft between the two domains. The identification of catalytic residues is a key to understand the functioning of the enzymes. With the information from other functionally similar sequences with known crystallographic structures, we can identify the ligand binding residues. But one of the great unmet promises of molecular biology is the calculation of protein-ligand binding affinities. Although thousands of protein structures have been determined to atomic resolution, interpreting these structures to understand their function, and modulate it by designed ligands, remains elusive and hence *in-silico* approach adds momentum to it. The energy computations of itpk1 were done by GROMOS 96 of SWISS PDB viewer with bond (2777.53KJ/mol), angles (3399.68KJ/mol), torsion (1539.34KJ/mol) and non-bonded as 11523.21 KJ/mol energies (Supplementary Fig. 2). By calculating the interaction energy of the protein moiety by *in-silico* strategies revealed the ability of itpk1's to bind various ligand moieties. Cleft analysis followed by PyMOL generated image of ligand binding residues ofitpk1using FunFold server predicted Arg104, Lys153, Ser160, Ala161, Ser163, Met166, Gln185, Glu186, Leu187, Val188, Ser231, Phe232, Ser233, Gln234, Val235, Ser236, Asp284, Asp299, Asn301,as ligand binding amino acid residues (Fig. 7) and the most likely ligand type predicted were Mg^{2+}, ATP and ADP (Roche et al., 2011). It is believed that a considerable free energy change can cause severe electrostatic and steric constraints within these molecules which result in catalytic activities of kinases (Stephens et al., 1993). However, the results obtained to date are only preliminary and further structural and mutational studies need to be performed to throw more light on the catalytic versatility of the itpk's. Interest in such versatility of these multifaceted ipks is growing as its structural, functional, evolutionary and therapeutic implications are becoming more widely appreciated. Intrinsic dynamics of Gmitpk1 was analysed using normal mode analysis (NMA). The first six modes matching with global rotation and translation of the system in NMA are generally ignored (Hollup et al., 2005) and hence we focussed on the lowest frequency mode of concern is the seventh one. Deformation energies calculated according to Hinsen (1998) revealed that itpk1 showed lower deformation energy at modes 7 (854.98), 8 (1048.29), 9 (1404.71), 10 (1893.35) compared to higher modes like 11 (3306.67) and 12 (3677.18) (Fig. 8) which reflects the lower rigidity in those amino acid regions. The normalized atomic displacement plot indicates the vibrational and thermal properties of the protein at atomic level. B factors calculated from ElNemo analysis were based on the first 100 normal modes and were scaled to match the overall B factors. The B factors calculated from ElNemo analysis signifies that itpk1 contain less rigid regions and are flexible (result not shown). Figure 9, shows the solvent accessibility plot of itpk1. ASA View analysis of the solvent accessibility for the modelled protein pointed out that the accessible residues were present in the outermost ring of the spiral. The majority of negatively charged residues and polar uncharged residues were present on the outermost surface and hydrophobic residues were confined to the inner rings of the spiral. However, few positively charged and hydrophobic amino acid residues were also present.

Materials and Methods

Cloning of itpk1cDNA from Glycine max Pusa 16 cultivar

Soybean variety (Pusa 16) was obtained from Division of Genetics (IARI). RNA was isolated from the developing soybean seeds using RNeasy Plant Mini Kit (Qiagen). Using ClustalW2 (http://www.ebi.ac.uk/Tools/clustalw2/index.html), conserved nucleotide regions of *itpk* isoforms were determined and specific primers for *itpk*1 [5'-ATGGCGGAGAAGAGATTCGGCG-3'] & [5'



(A) **(B)**

Fig 7. Cleft analysis of *Gm*itpk1 by FTsite server (A) and predicted ligand binding residues using FunFold server (B). Ligand binding residues are shown as blue sticks on PyMol generated model of inositol-1, 3, 4 tris phosphate 5/6 kinase-1. Mg^{2+}, ATP and ADP are the most likely predicted ligands.

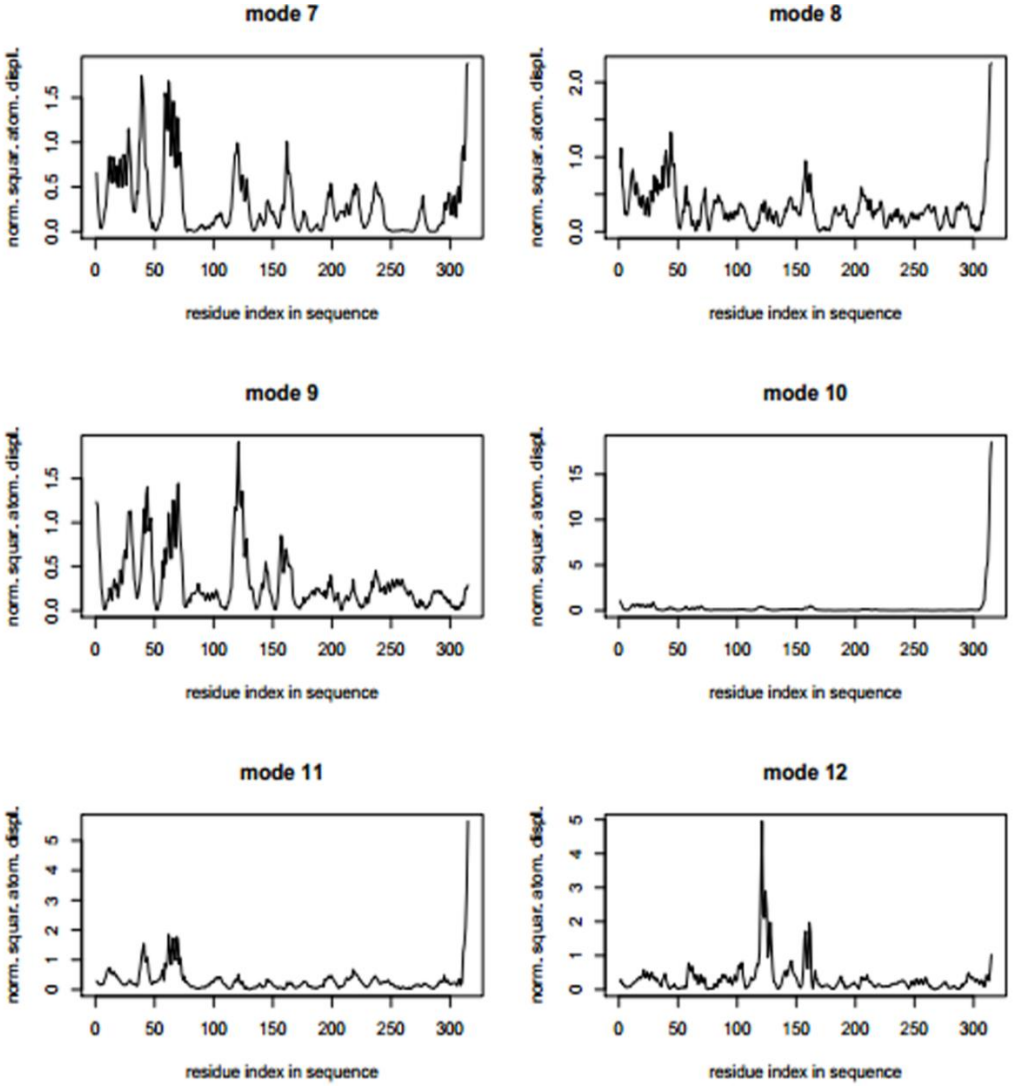


Fig 8. Normalized atomic displacement plot calculated for modes 7 to 12 in *Gm*itpk1. The figure shows the plot of the normalized square atomic displacements which represents the square of the displacement of each C-α atom. The highest values corresponded to the most displaced regions and residues with maximum displacements associated with functional sites. X and Y axis denote residue index in sequence and normal mode of square atomic displacement, respectively.

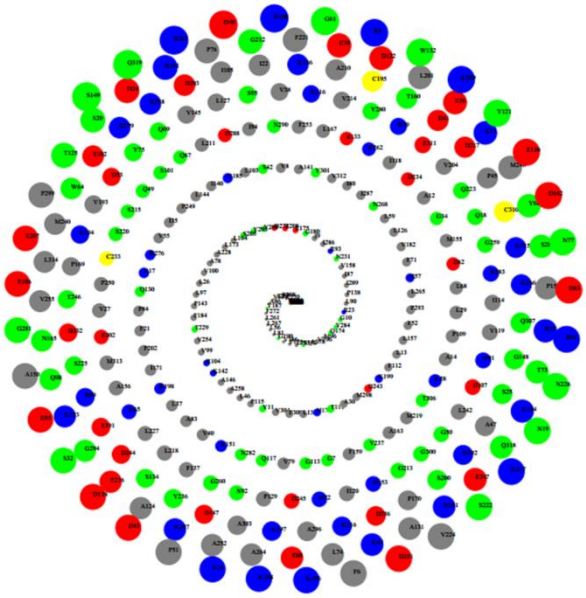


Fig 9. Solvent accessibility plot of *Gm*itpk1.using ASA-View. Blue-Positive charged residues (R,K,H), Red-Negative charged residues (D,E) Green-Polar uncharged residues (G,N, Y,Q,S,T,W), Yellow- C & Gray-Other hydrophobic residues.

CTTGAAGAGATTCCTCTTTCTCCTTAGGAGCG-3']
were designed based on EST (NM_001250900.1) sequence from NCBI (http://www.ncbi.nlm.nih.gov/). With total RNA as template, a one-step RT-PCR (Qiagen) reaction was carried out with gene specific primers to amplify the *itpk1* gene sequence. The amplicon was cloned into pGEM-T vector (Promega) and transformed into *E.coli* (DH5α). The recombinant plasmids were confirmed by restriction analysis and sequenced. The nucleotide sequence data was submitted to the Gen Bank (Accession number KF913641).

Inositol kinases sequences retrieval

For *in silico* characterization of *itpk1*, the sequence of the cloned *itpk1*fragment was subjected to homology search using BLASTN on NCBI server (http://www.ncbi.nlm.nih.gov/) and the translated protein sequence for the complete ORF was searched for homology using BLASTP on NCBI server (http://blast.ncbi.nlm.nih.gov/Blast.cgi). The search of protein database yielded 18 ipks, which were selected after excluding the redundant sequences. Plant sequences obtained from NCBI protein database are represented in Table 1. The sequences were converted to FASTA format using Read SeQ sequence conversion server (Gilbert et al., 2003).

Sequence Alignments

Multiple sequence alignments were performed with Clustal W2 (Larkin et al., 2007) and the alignment file was imported to BioEdit programme (Thomas et al., 1999) to identify and shade the conserved amino acid sequences. Phylogenetic analyses of the protein sequences were generated by Treedyn (Dreeper et al., 2010) using Neighbor Joining (NJ) method with the poisson correction implemented in the server. Boot strapping was used to evaluate the degree of support for particular grouping pattern in the phylogenetic tree. For the tree construction, itpk1 *Glycine max* protein sequence was used as query to perform BLASTP search in the GenBank to retrieve amino acid sequences of inositol phosphate kinases.

Physiochemical parameter analysis

For understanding the structural attributes, the amino acid composition of the ipk sequences were computed using PEPSTATS (http://emboss.bioinformatics.nl/cgi-bin/emboss/pepstats) analysis tool (Rice et al., 2000). The physiochemical parameters such as the molecular weight (MW), isoelectric point (pI), extinction coefficient (EC), aliphatic index (Ai), instability index (Ii), amino acid property and grand average hydropathy (GRAVY) were calculated using ProtParam (http://web.expasy.org/protparam/) tool (Gasteiger et al., 2005) of the ExPASy proteomics server and tabulated (Table 2). Sub-cellular localization was predicted by Cello-4 (http://cello.life.nctu.edu.tw/) (Yu et al., 2004).

Identification of domains and motifs

Secondary structure elements predictions were performed using CFSSP (Chou and Fasman Secondary Structure Prediction) (http://www.biogem.org/tool/chou-fasman/) (Ashok et al., 2013). The consensus secondary structure contents and the predicted disulphide patterns of ipks are tabulated (Table 3). The presence of disulphide bridges were analysed using DiANNA server (http://clavius.bc.edu/~clotelab/DiANNA/) (Ferre et al., 2005) and compared with CYS-REC tool, which predicts the most probable bonding patterns between the available cysteine residues (http://linux1.softberry.com/berry.phtml). The secondary motif map and topology diagram were created using PDBSum tool (Laskowski et al., 2005). Motif scan was done using My Hits (http://myhits.isb-sib.ch/cgi-bin/motif_scan) analysis tool (Sigrist et al., 2010).

Homology modeling and refinement

The 3D models ipks were constructed using the protein structure homology building program SWISS-MODEL with energy minimization parameters (Arnold et al., 2006). SWISS PDB viewer (Guex et al., 1997) was used to visualize and refine the models using a suitable PDB template 2q7d.1.A. The 3D structures were evaluated and validated using RAMPAGE program (http://mordred.bioc.cam.ac.uk/~rapper/rampage.php) (Vriend et al., 1990). The final model obtained was further assessed and verified by Verify 3D (http://nihserver.mbi.ucla.edu/Verify_3D/), a program that analyses the compatibility of an atomic model (3D) with its own amino acids (1D)and ProSA server (https://prosa.services.came.sbg.ac.at/prosa.php), a program used for refinement and validation of protein model (Wiederstein et al., 2007). This model was further refined by molecular dynamics method and the final stable structure of itpk1was obtained. Ligand binding residue PyMOL model of itpk1 was generated using FunFold server (http://www.reading.ac.uk/bioinf/servlets/nFOLD/). The energy computations of itpk1 were done by GROMOS 96 of SWISS PDB viewer. The structural dynamics task accomplished using WEBnm (http://apps.cbu.uib.no/webnma/home) (Hollup et al., 2005) was used to calculate the slowest modes and related deformation energies. ElNemo (Suhre and Sanejouand, 2004) was utilized to calculate to the corresponding protein movements. Solvent accessibility of the amino acid residues in the modeled protein was

determined by ASA-view (http://www.abren.net/asaview/) (Ahmad et al., 2004).

Conclusions

A potentially useful gene (*Gmitpk*1) for the development of low phytate containing transgenic soybean was cloned and computationally characterized. Primary structure analyses of the 18 selected ipks revealed that majority of the proteins were hydrophobic in nature containing disulphide linkages and had a cytoplasmic localization. Physiochemical characterization on parameters such as pI, EC, Ai, GRAVY and stability provided essential and vital information on the structural and functional attributes of these proteins. The predicted phylogenetic tree revealed a relative close evolutionary linkage between the inositol kinases selected under study. Secondary structure analysis suggested an exceptional flexibility to the inositol kinases on account of high intensity of alpha helices. A geometrically possible 3D model of itpk1 after energy minimization was also constructed and further refined using Verify3D and ProSA. Ligand binding residues clearly predicted Mg^{2+}, ATP and ADP to be the most likely ligands for itpk1. NMA of itpk1 demonstrated low deformation energies and less rigidity which underlines the catalytic flexibility possessed inositol phosphate kinases. The reliable and stable 3 D model predicted in the present study for the very first time threw some novel insights into the structural features of *Gm*itpk1 and will be able to build testable hypotheses for unravelling the molecular functions of inositol phosphate kinases. RNAi silencing constructs of *Gm*itpk1 can be transformed into soybean in future to reduce the phytate content and thus would tackle the alarming issues of phosphorous pollution, eutrophication and mineral bioavailability.

Acknowledgement

We gratefully acknowledge Indian Agricultural Research Institute (IARI) and Indian Council of Agricultural Research (ICAR) for the financial assistance.

References

Abelson PH (1999) A potential phosphate crisis. Science. 283(5410):2015.

Ahmad SM, Gromiha M, Fawareh H, Sarai A (2004) ASA View: Solvent accessibility graphics for proteins. BMC Bioinformatics. 1(5):51.

Alexandr PK, Susan ST, Lynn FT (2008) A helix scaffold for the assembly of active protein kinases. Proc Natl Acad Sci. 105(38):14377-14382.

Amanda RS, Xun Q, Stephen BS, Elizabeth AG (2008) Metabolic and signalling properties of an *ITPK* gene family in *Glycine max.* FEBS Lett. 582(13):1853-1858.

Arnold KL, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL work space: A web-based environment for protein structure homology modelling. Bioinformatics. 22(2):195-201.

Peter YC, Gerald DF (1974) Prediction of protein conformation. Biochemistry. 13(2): 222–245.

Benkert P, Biasini M, Schwede T (2011) Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics. 27(3):343-350.

Cheek S, Zhang H, Grishin NV (2002) Sequence and structure classification of kinases. J Mol Biol. 320(4):855-881.

Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J. 5(4):823-826.

Dereeper A, Audic S, Claverie JM, Blanc G (2010) BLAST-EXPLORER helps you building datasets for phylogenetic analysis. BMC Evol Biol. 10:8.

Eisenberg D, Luthy R, Bowie JU (1997) VERIFY 3D: Assessment of protein models with three-dimensional profiles. Methods Enzymol. 277**:**396-404.

Ferre F, Clote P (2005) (Web Server issue): W230-2.DiANNA: a web server for disulfide connectivity prediction. Nucleic Acids Res.1**:**33.

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 J (ed) The Proteomics Protocols Handbook, Humana Press, New York.

Gilbert D (2003) Sequence files format conversion with commandline read seq. In: Baxevanis AD, Petsko GA, Stein LD, Stormo GD (ed) Current protocols in Bioinformatics, Wiley, New York.

Grishin NV (1999) Phosphatidylinositol phosphate kinase: a link between protein kinase and glutathione synthase folds. J Mol Biol. 291(2)**:**239–247.

Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss pdb Viewer: an environment for comparative protein modeling. Electrophoresis. 18(5)**:**2714-2723.

Hinsen K (1998) Analysis of domain motions by approximate normal mode calculations. Proteins. 33(3):417-429.

Hollup SM, Sælensminde G, Reuter N (2005) WEBnm@: a web application for normal mode analysis of proteins. BMC Bioinformatics. 6:52.

Josefsen L, Bohn L, Sorensen S, Green P, Caddick SE, Brearley CA (2007) *Arabidopsis thaliana* inositol phosphate kinase from rice and barley belonging to the ATP-grasp super family. Gene. 397**:**114-125.

Larkin MA, Blackshields G, Brown NP, Chenna R and Gettigan M (2007) Clustal W and clustal X version 2.0. Bioinformatics. 23(21)**:**2947-2948.

Laskowski RA, Chistyakov VV, Thornton JM (2005) PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids, Nucleic Acids Res. 33**:**266-268.

Patil A, Nakamura H (2006) Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks. FEBS Lett. 580(8)**:**2041-2045.

Raboy V (2001) Seeds for a better future: 'low phytate' grains help to overcome malnutrition and reduce pollution. Trends Plant Sci. 6(10)**:**458-462.

Rice P, Longden I, Bleasby A (2000) Emboss: The European molecular biology open software suite. Trends Genet. 16(6)**:** 276-277.

Roche DB, Tetchner SJ, McGuffin LJ (2011) FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins. BMC Bioinformatics. 12:160.

Sigrist CJ, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N (2010) PROSITE a protein domain database for functional characterization and annotation. Nucleic Acids Res. 38:161-166.

Stephen BS (2004) How versatile are inositol phosphate kinases? Biochem J. 377(Pt2): 265-280.

Stephens, LR, Radenberg T, Thiel U, Vogel G, Khoo KH, Dell A, Jackson TR, Hawkins PT, Mayr GW (1993) The detection, purification, structural characterization and metabolism of diphospho inositol pentakisphosphate(s) and bisdiphospho inositol tetrakisphosphate(s). J Biol Chem. 268(6)**:**4009-4015.

Suhre K, Sanejouand YH (2004) ElNemo: a normal mode web-server for protein movement analysis and the generation of templates for molecular replacement. Nucleic Acids Res. 32:610-614.

Sweetman D, Stavridou I, Johnson S, Green P, Caddick SE, Brearley CA (2007) *Arabidopsis thaliana* inositol 1,3,4-trisphosphate5/6 kinase is an outlier to a family ATP-grasp fold proteins from *Arabidopsis*. FEBS Lett. 581(22): 4165-4171.

Thomas AH (1999) BioEdit: A user friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. 41:95-98.

Vriend (1990) WHAT IF: A molecular modeling and drug design program. J Mol Graphics. 8(1): 52-56.

Wiederstein M, Sippl MJ (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res. 35:407-410.

Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. J Mol Biol. 293(2):321-331.

Yang X, Shears SB (2000) Multitasking in signal transduction by a promiscuous Human Ins (3,4,5,6)P_4 1-Kinase/Ins(1,3,4)P_3 5/6 kinase. Biochem J. 351**:**551-555.

Yu CS, Lin CJ, Hwang JK (2004) Predicting subcellular localization of proteins for gram-negative bacteria by support vector machines based on n-peptide compositions. Prot Sci. 13(5):1402-1406.