INNSPUB

International Journal of Biosciences | IJB |
ISSN: 2220-6655 (Print), 2222-5234 (Online)
http://www.innspub.net
Vol. 6, No. 2, p. 93-99, 2015

RESEARCH PAPER OPEN ACCESS

Phylogenetic relationship and *in silico* expression profile of *PELPK1* of *Arabidopsis thaliana* (L.) Heynh*

Abdur Rashid[1], Michael Deyholos[2]**

[*]*Department of Biological Sciences, University of Alberta, Edmonton, Alberta, TG6 2E9, Canada*
[1]*Current address: 401C Plant Science Bldg., University of Kentucky, Lexington, KY 40546, USA*
[2]*Current address: IK Barber School of Arts & Sciences, The University of British Columbia, Okanagan Campus, SCI 316, 1177 Research Road, Kelowna, Canada*

Key words: *Arabidopsis thaliana* (L.) (Heynh.), *PELPK,* phylogeny, bioinformatics, expression, inducibility, *in silico*.

http://dx.doi.org/10.12692/ijb/6.2.93-99 Article published on January 18, 2015

Abstract

Based on bioinformatics and computational analyses, the *Arabidopsis thaliana* (L.) (Heynh.) gene, *At5g09530* was earlier annotated as *PELPK1* and its putative paralog, *At5g09520* as *PELPK2*. In the present report, further *in silico* analyses were carried out to determine the phylogenetic relationship of PELPK1 with the proteomes of *A. thaliana* (L.) (Heynh.) and other species of angiosperms; and the expression pattern and inducibility of *PELPK1* in Arabidopsis plants. The data suggest that PELPK1 and PELPK2 are closely related and form a subgroup of hydroxyproline rich glycoprotein (HRGP)-family of proteins; PELPK and similar motifs are conserved to large evolutionary distances involving both monocot and dicot; *PELPK1* is highly responsive to certain biotic factors and elicitors, moderately responsive to abiotic factors, unresponsive to common growth hormones, and is negatively responsive to natural phytosterols; and that it is predominantly expressed during early and late stages of Arabidopsis growth. The above *in silico* data correspond partially with our experimental observations reported earlier.

Corresponding Author: Abdur Rashid ✉ abdur.rashid@uky.edu

Introduction

When the Arabidopsis genome sequence was first published (AGI 2000), the majority of its genes (~70%) were automatically assigned to functional categories based on their sequence homology to other genes of known function. Less than 10% of the genes were characterized experimentally, and about 30% remained unclassified as they had no closely related sequences of known function (TAIR, http://www.arabidopsis.org/).

Over the past more than a decade, although many annotations have been verified manually, the number of genes with no annotation appears to be unchanged. For example, in 2004, TAIR had 35% of the genes with unknown molecular function, and the figure has not been changed significantly during last several years. Gutiérrez *et al.* (2004) estimated that more than half of about 4,000 plant specific Arabidopsis proteins are with unknown function.

TAIR, that provides annotations to Gene Ontology (GO, http://www.geneontology.org/) relies mostly on information from public sequence databases such as GeneBank (Wortman *et al.* 2003). The advantage of the GeneBank is that anyone can deposit sequences and annotations. However, the drawback is that errors such as redundant gene names, misplacement of gene families, and clustering of genes (Schlueter *et al.* 2005) can be introduced and propagated. Thus, a thorough database analysis of the gene of interest (GOI) is necessary in order to acquire accurate information to design experimentations for further characterization.

For instance, depending on protein alignments, the *A. thaliana* (L.) (Heynh.) gene, *At5g09530* has been annotated differently such as an extensin-like protein, an HRGP (hydroxyproline-rich glycoprotein) family protein (TAIR database), an HRGP family member containing Pro-rich extensin domains (NCBI REFSEQ: NP_196515), a periaxin-like protein (NCBI accession: AAK96839), and a member of the PRP (proline-rich protein) family representing PRP10 (Showalter *et al.* 2010). However, by extensive bioinformatics and computational analyses, this gene has been annotated accurately as *PELPK1* and its presumptive paralog as *PELPK2* (*AT5G09520*); for a review, refer to Rashid and Deyholos (2011).

In recent years, various bioinformatics tools have been developed and extensively used for database analysis to predict the structure and function of many genes (Attwood 2000; Hvidsten 2001; Marcotte *et al.* 1999; Pavlidis *et al.* 2001; Syed and Yona, 2003). For example, bioinformatics tools have been successfully utilized in data mining of several microorganisms including *Saccharomyces cerevisiae*, *Escherichia coli* and *Mycobacterium tuberculosis* (Clare and King, 2003; King *et al.*, 2000).

Also many bioinformatics resources available around the world that allow researchers to access and analyze large amounts of genetic, genomic, proteomic, and biological data through the internet. In the current report, further database analyses were carried out to determine the evolutionary relationship of PELPK1 with the other proteins of *A. thaliana* (L.) (Heynh.) and the proteomes of higher plants, compile *in silico* expression and inducibility data of *PELPK1* in Arabidopsis plants, and to compare these *in silico* data with our experimental results reported earlier (Rashid and Deyholos 2011; Rashid *et al.* 2013ab; Rashid 2014) based on expression, mutational, and proteomic analyses of *PELPK1*.

Materials and Methods

Proteome search was conducted in *A. thaliana* (L.) (Heynh.) for proteins similar to PELPK1 by using BLAST$_P$. The neighbor joining tree was constructed from the 20 Arabidopsis genes with the highest protein similarity to PELPK1 (BLAST$_P$ e-values <1e^{-4}) using MEGA 4.
(http://www.megasoftware.net/mega.html).

The gene cluster that includes PELPK1 and PELPK2 was generated by the Phytozome database (gene family 22878593, www.phytozome.org). *In silico* expression profiling of PELPK1 was conducted using Genevestigator V3 (ATH1: 22k full genome Affymetrix

GeneChip, https://www.genevestigator.com/gv/user/serveAppleT.jsp (Zimmermann *et al.* 2004); Bio Analytic Resource (eFP Arabidopsis, http://bar.utoronto.ca (Winter *et al.* 2007); The Arabidopsis Information Resource (http://www.arabidopsis.org/), and the seed-specific database (http://seedgenenetwork.net/).

Results and discussion

Relationship of PELPK1 with the proteome of *A. thaliana (L.) Heynh.*

The constructed neighbor-joining tree has been labeled with current annotations for each of the proteins (Fig. 1). According to this inferred phylogeny, PELPK1 and PELPK2 are most closely related compared to other proteins annotated as HRGP and PRP (proline rich protein) family members, including At5g09480, which were identified based on the presence of six PELPK-like motifs (Rashid and Deyholos 2011). While inferring relationships of highly repetitive proteins is complicated, and not well-suited to standard tools of phylogenetics, these results show that PELPK1 and PELPK2 do form a distinct group within the Arabidopsis genome (Fig. 1). The above conclusion is supported by the previous report indicating that insertional inactivation of PELPK1 alone failed to exhibit phenotype, whereas knock-down of both PELPK1 and PELPK2 by RNAi exhibited phenotype (Rashid and Deyholos 2011). The other related proteins of PELPK1 in *A. thaliana* (L.) (Heynh.) proteome include: hydroxyproline-rich glycoproteins (HRGPs), proline rich proteins (PRPs), protease inhibitor proteins (PIPs), seed storage proteins (SSPs), late embryogenesis abundant proteins (LEAs), leucine rich repeat proteins (LRRs), C-protein immunoglobulins(CPIs), F-box family proteins FFRs), and pollen Ole e 1 allergens (POAs), Fig. 1.

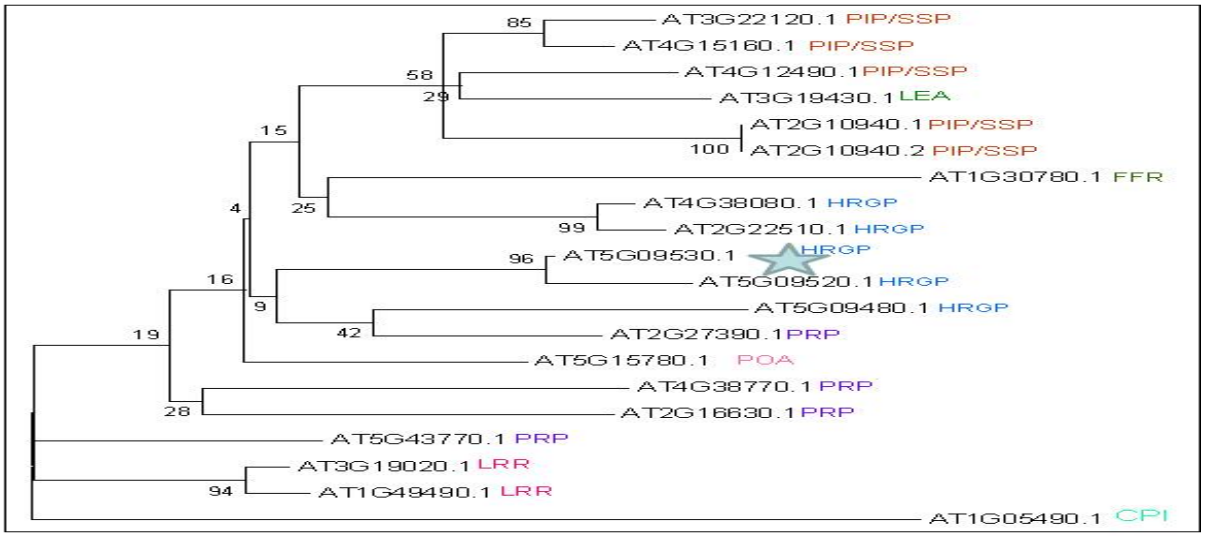

Fig. 1. A bootstrap neighbor joining tree constructed using MEGA4 is showing the relationship of PELPK1 with twenty other closely related *A. thaliana* proteins. The tree was drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method (Tamura *et al.* 2007) and are in the units of the number of amino acid substitutions per site. All positions containing gaps and missing data were eliminated from the dataset (complete deletion option). Bootstrap values (in %) on the branches are calculated as the number of times that a particular grouping of sequences appears during the bootstrap analysis. A 96% bootstrap value for the grouping of PELPK1 (AT5G09530) with PELPK2 (AT5G09520) indicates that in the 1000 bootstrap replicates selected, that grouping was found 960 times. HRGP, Hydroxyproline rich glycoproteins; PRP, Proline rich proteins; PIP, Protease inhibitor proteins; SSP, Seed storage proteins; LEA, Late embryogenesis abundant; LRR, Leucine rich repeats; CPI, C-protein immunoglobulin; FFR, F-box family protein; POA, pollen Ole e 1 allergen.

Relationship of PELPK1 with the proteomes of other angiosperms

The gene cluster generated by including PELPK1 and PELPK2 (Fig. 2) shows that beyond the Arabidopsis genome, there is evidence for conservation of the PELPK motif and PELPK1 protein. The proteins that are most similar to PELPK1 are from its close relative *Arabis lyrata* and, interestingly from the much more distantly related *Glycine max* (L.) Merr. (Fig. 2). Indeed, the PELPK motif was found repeated in several species, including both monocot and dicot, such as *Arabis lyrata* (L.), *Glycine max* (L.) Merr., *Linum usitatissimum* (L.), *Ricinus communis* (L.), *Populus trichocarpa* (Torr. & A.Gray), *Vitis vinifera* (L.), *Carica papya* (L.), *Cucumis sativus* (L.), *Sorghum bicolor* (L.) Moench.), *Brachypodium distachyon* (L.) (P.Beauv.), *Oryza sativa* (L.), and *Zea mays* (L.). Although the evolutionary history of this repeat-rich protein is still unclear, the conservation of PELPK1-like sequences over large evolutionary distances suggests that there might be a specific function associated with these repeated motifs.

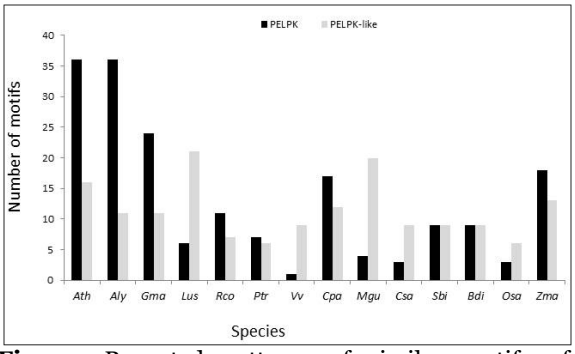

Fig. 2. Repeated patterns of similar motifs of PELPK1 in higher plants. Ath, *Arabidopsis thaliana* (L.) Heynh.); Aly, *Arabis lyrata* (L.); Gma, *Glycine max* (L.) Merr.); Lus, *Linum usitatissimum* (L.); Rco, *Ricinus communis* (L.); Ptr, *Populus trichocarpa* (Torr. & A.Gray),; Vvi,= *Vitis vinifera* (L.); Cpa, *Carica papya* (L.); Csa, *Cucumis sativus* (L.); Sbi, *Sorghum bicolor* (L.); Bdi, *Brachypodium distachyon* (L.); Osa, *Oryza sativa* (L.); Zma, *Zea mays* (L.).

In silico expression profiles of *PELPK1*

Database analysis for transcript expression of *PELPK1* (probe set: 250500_at/TAIR Accession: AASequence: 1009132415) was found to vary depending on external and developmental factors. The results are summarized below.

Response of PELPK1 to external factors

Survey of different databases stated above revealed that *PELPK1* was highly up-regulated by certain biotic factors and elicitors (Fig. 3A and B), moderately up-regulated by common abiotic factors (Fig. 4A), unresponsive to common growth hormones (Fig. 4B), and down-regulated by certain steroid hormones (Fig. 4B). Among the external factors surveyed, those that significantly up-regulated the *PELPK1*, as determined by the ratios (in parenthesis) of infected/uninfected or treated/untreated *in silico* data are graded as follows: *Pseudomonas syringae* (a biotic factor; 69.9)> cabbage leaf curl virus, CalCuV (a biotic factor; 23.5)> syringolin A, SRG (an elicitor; 15.6)≥ lipopolysaccharide, LPS (an elicitor/endotoxin; 15.3)> flagellin, FLG-22 (an elicitor; 12.0)> *Phytophthora infestans, Botrytis cinerea, Blumeria graminis,* and *Erysiphe orontii* (biotic factors; ± 10.5)> elevated CO_2 (an abiotic factor; 8.5)> hypersensitive bacterial toxin, HrpZ (an elicitor; 7.0)> cold, drought, and salt stresses (common abiotic factors; ± 6.0)> osmotic stress, and N_2 and Fe deficiencies (abiotic factors; ≥ 5.0)> wounding (an abiotic factor; 4.2)> oxidative stress (an abiotic factor; 3.8)≥ ABA and TIBA (abiotic/hormonal factors; 3.5-3.8)> ethylene, PCIB, and 2,4,6-T (abiotic/hormonal factors; ~2.0). On the other hand, the factors that significantly down-regulated the *PELPK1* include, cycloheximide (CHX), an inhibitor of protein biosynthesis in eukaryotes (0.15), and natural phytosterols such as campestanol (CAS) and castasterone (CS) (≥ 0.08). The common growth hormones e.g., IAA and GA_3 did not show any influence on *in silico* transcript expression profile of *PELPK1* (Fig. 4B).

The results of *in silico* analysis of *PELPK1* presented above correspond to some extent with our previously reported experimental results. For instance, the results reported earlier showed that the transgenic plants harboring a *PELPK1*-promoter::*β-glucoronidase* (GUS)-reporter fusion were highly

responsive to biotic factors, particularly to the pathogen, *Pseudomonas syringae* (Rashid *et al.* 2013). However, some of the above *in silico* observations do not correspond with the previously reported experimental data. For example, the present *in silico* data showed that the levels of *PELPK1* expression in response to defense hormones such as methyl jasmonate (MeJa) and/or salicylic acid (SA) were lower (Fig. 3B) than certain abiotic factors (Fig.4A). While it was previously shown that none of the abiotic factors tested including salt and osmotic stresses had any influence on the expression of the *PELPK1;* however, MeJa and SA had significantly up-regulated the *PELPK1* expression (Rashid 2010). The reason for the discrepancy between *in silico* versus our experimental observations is unclear at this time; however, it is generally accepted that transcript expression may not always represent actual translation of the genes.

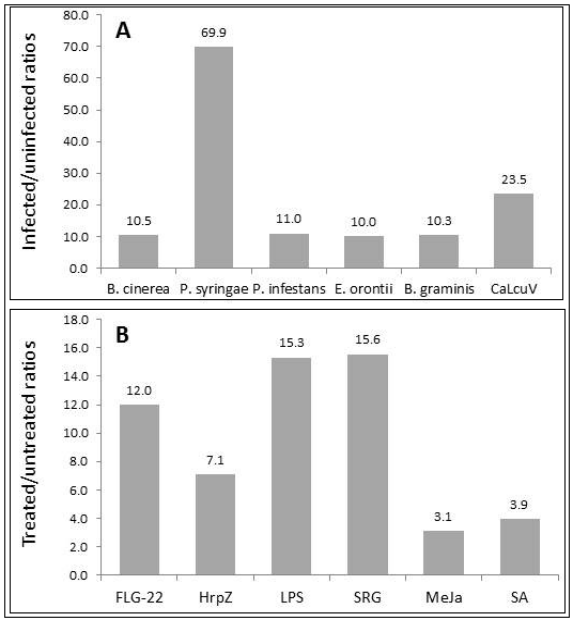

Fig. 3. Database analyses for transcript expression profiles of *PELPK1* in *Arabidopsis thaliana* (probe set: 250500_at/TAIR Accession: AASequence: 1009132415). A: biotic factors, B: elicitors. Infected/uninfected or treated/untreated ratios were calculated using signal intensities. *B, Botrytis; P, Pseudomonas; P, Phytophthora*; *E, Erysiphe; B, Blumeria;* CaLcuV, Cabbage leaf curl virus; MeJa, methyl jasmonate; SA, salicylic acid; HrpZ, viral coat protein; FLG-22, flagellin; SRG, syringolin; LPS, lipopolysaccharide.

Response of PELPK1 to developmental cues

These *in silico* data are summarized under two categories: (i) tissue-specific expression profile (Fig.5A), and (ii) growth stage-specific expression profile (Fig. 5B). The tissue specific expression profile of *PELPK1* exhibited the following ranking of transcript abundance (arbitrary units in parenthesis): radicle (15.0)> root (13.0)> hypocotyl (6.5)> seed (5.5)> seedling (4.0)> silique (3.0)> inflorescence/leaf/root-hair (~1.5) (Fig. 5A). On the other hand, the growth-stage specific expression profile of *PELPK1* showed highest expression during the stage of green seed maturation (15.0), followed by silique formation (~11.0), seed germination and early rosette formation (≥ 8.0), seedling growth (3.0), late rosette stage (1.4), and flower initiation and bolting stages (± 0.2) (Fig. 5B).

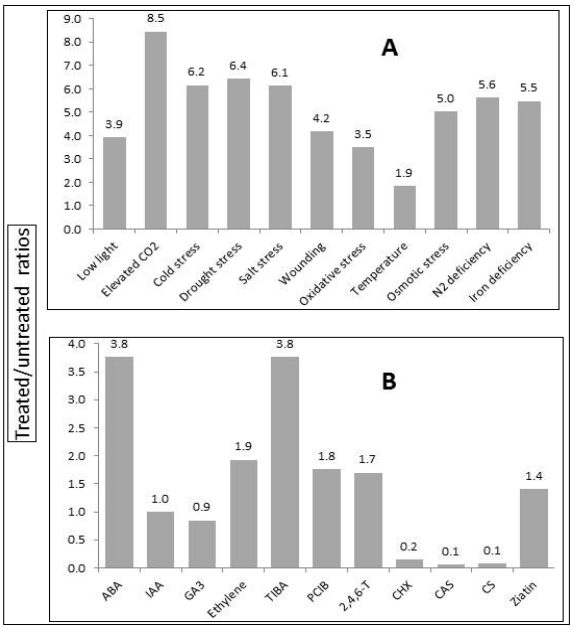

Fig. 4. Database analyses for transcript expression profiles of *PELPK1* in *Arabidopsis thaliana*. A: abiotic factors, B: hormonal factors. Treated/untreated ratios were calculated as mentioned above. ABA, abscisic acid; IAA, indole acetic acid; GA_3, gibberellic acid; TIBA, triiodobenzoic acid; PCIB, p-chlorophenoxyisobutyric acid; 2,4,6-T, trichlorophenoxyacetic acid; CHX, cycloheximide; CAS, compestanol; CS, cathasterone.

While the above database analyses do not show any specific pattern of transcript abundance of *PELPK1*

during the developmental stages of *Arabidopsis* plants, relatively higher levels of transcript expression were observed in the early and late stages of *Arabidopsis* growth. For instance, the tissue-specific expression profile showed that *PELPK1* transcript level was higher in radicle and root tissues (Fig. 5A); whereas the growth-stage specific expression profile showed that it was higher during the silique formation and seed maturation stages (Fig. 5B). These *in silico* data appear to suggest that the expression of *PELPK1* was transient, changing with the growth and developmental stages of *Arabidopsis* plants and perhaps also with the local growth conditions. This assumption is correlated with the previously made prediction that PELPK1 might belong to intrinsically disordered proteins (IDPs) with a wide range of structural flexibility (Rashid *et al.* 2013; Rashid 2010). The above *in silico* data also correspond to some extent with the previously reported experimental results that showed (i) that *PELPK1*-promoter activity was higher during the stage seed germination (Rashid *et al.* 2013), and (ii) that *PELPK1* encoded protein was predominantly deposited to the seed coat during seed germination, and to the cell walls of mature tissues (Rashid 2014).

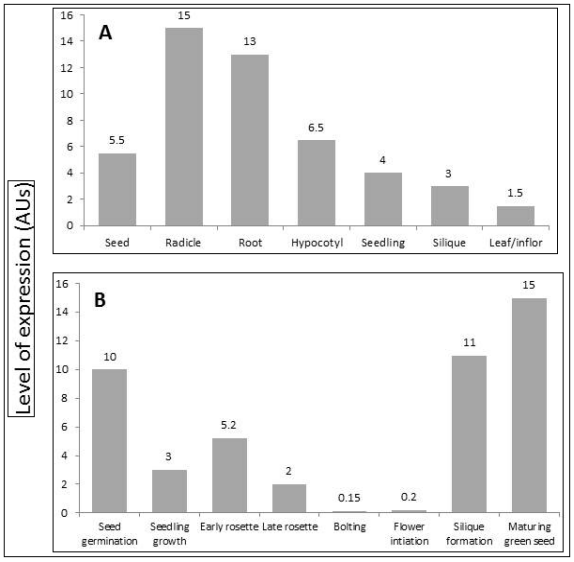

Fig. 5. Database analyses for tissue- and growth stage-specific transcript expression profiles of *PELPK1* in *Arabidopsis thaliana*. A: tissue-specific expression profile; B: growth-stage specific expression profile. The level of expression is presented in arbitrary units (AUs).

In conclusion, the information generated from the survey of public databases appear to suggest that (i) PELPK1 and PELPK2 form a sub-group of HRGPs, (ii) PELPK-like motifs are conserved to a large evolutionary distance involving both monocots and dicots, (iii) *PELPK1* is highly responsive to biotic factors and elicitors but negatively responsive natural phytosterols, and (iv) that it is predominantly expressed during the early and late stages of *Arabidopsis* growth. These *in silico* data are partially consistent with some of our previously reported experimental observations.

References

AGI. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408,** 796–815.
http://dx.doi:10.1038/35048692

Attwood TK. 2000. The quest to deduce protein function from sequence: the role of pattern databases. The International Journal of Biochemistry and Cell Biology **32,** 139–155.
http://dx.doi:10.1016/S1357-2725(99)00106-5

Clare A, King RD. 2003. Predicting gene function in it *Saccharomyces cerevisiae*. Bioinformatics **19,** ii42–ii49.
http://dx.doi:10.1.1.97.2507

Gutie´rrez RA, Green PJ, Keegstra K, Ohlrogg JB. 2004. Phylogenetic profiling of the *Arabidopsis thaliana* proteome: what proteins distinguish plants from other organisms? Genome Biology **5,** R53.
http://dx.doi:10.1186/gb-2004-5-8-r53

Hvidsten TR. Komorowski J, Sandvik AK, Laegreid A. 2001. Predicting gene function from gene expressions and ontologies. Pacific Symposium on Biocomputing **6,** 299-310.
http://dx.doi:10.1142/9789814447362_0030

King RD, Karwath A, Clare A, Dehaspe L. 2000. Accurate prediction of protein functional class from sequence in the *Mycobacterium tuberculosis*

and *Escherichia coli* genomes using data mining. Yeast **1,** 283-293.
http://dx:doi.org/10.1002/1097-0061

Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. 1999. A combined algorithm for genome-wide prediction of protein function. Nature, **402,** 83–86.
http://dx.doi:10.1038/47048

Pavlidis P, Weston J, Cai J, Grundy W. 2001. Gene functional classification from heterogenous data. In Proceedings of RECOMB.
http://dx.doi:10.1.1.152.8534

Rashid A. 2012. Developing transgenic plants through application of reverse genetics. Book, pp 232; ISBN-13: 978-3-8465-4625-3; LAP LAMBERT Academic Publishing GmbH & Co. KG; Heinrich-Böcking-Str. 6-8, 66121, Saarbrücken, Germany

Rashid A. 2014. Sub-cellular localization of PELPK1 in *Arabidopsis thaliana* as determined by translational fusion with green fluorescent protein reporter. Molecular Biology **48,** 258 – 262.
http://dx.doi:10.1134/S0026893314020162

Rashid A, Deyholos MK. 2011. PELPK1 contains a unique pentapeptide repeat and is a positive regulator of germination in *Arabidopsis thaliana*. Plant Cell Reporter **30,** 1735 -1745.
http://dx.doi:10.1007/s00299-011-1081-3

Rashid A, Hobson N, Deyholos MK. 2013a. A genomic region upstream of *Arabidopsis thaliana* PELPK1 promotes transcription in aleurone tissues and in response to *Pseudomonas syringae* and *Pythium irregulare*. Plant Molecular Biology Reporter **31,** 1025 –1030.
http://dx.doi:10.1007/s11105-012-0553-0

Rashid A, Bhadan A, Deyholos M, Kav K. 2013b. Proteomic profiling of the aleurone layer of mature *Arabidopsis thaliana* seed. Plant Molecular biology Reporter **31,** 464-469.
http://dx.doi:10.1007/s11105-012-0498-3

Schlueter SD, Wilkerson MD, Huala E, Rhee SY, Brendel V. 2005. Community-based gene structure annotation. Trends in Plant Science **10,** 9-14.
http://dx.doi.org/10.1016/j.tplants.2004.11.002

Showalter AM, Keppler B, Lichtenberg J, Gu D, Welch LR. 2010. A bioinformatics approach to the identification, classification, and analysis of hydroxyproline-rich glycoproteins. Plant Physiology **153,** 485-513.
http://dx.doi:10.1104/pp.110.156554

Syed U, Yona G. 2003. Using a mixture of probabilistic decision trees for direct prediction of protein function. In Proceedings of RECOMB.ACM, Berlin, Germany.
http://dx.doi:10.1145/640075.640114

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. Molecular Biology Evolurion **24,** 1596-1599.
http://dx.doi:10.1093/molbev/msr121

Winter D, Vinegar B, Nahal H, Ammar R, Wilson GV, Provart NJ. 2007. An electronic fluorescent pictograph browser for exploring and analyzing large-scale biological data sets. *PLoS ONE* 2 e718.
http://dx.doi:10.1371/journal.pone.0000718

Wortman JR, Haas BJ, Hannick LI, Smith RK, Maiti M, Ronning CM, Chan AP, Yu C, Ayele M, Whitelaw CA, White OR, Town CD. 2003. Annotation of the Arabidopsis Genome. Plant Physiology **132,** 461-468.
http://dx.doi.org/10.1104/pp.103.022251

Zimmermann P, Hirch-Hoffman M, Henning L, Gruissem W. 2004. Genevestigator, *Arabidopsis* microarray database and analysis toolbox. Plant Physiology **136,** 2621 – 2632.
http://dx.doi.org/10.1104/pp.104