Indian J. Genet., 74(2): 229-237 (2014)
DOI: 10.5958/0975-6906.2014.00160.6

In silico identification of late blight susceptibility genes in *Solanum tuberosum*

Tanmaya Kumar Sahu, A. R. Rao*, Sasmita Dora, Satakshi Gupta and Anil Rai

Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, New Delhi 110 012

(Received: January 2014; Revised: April 2014; Accepted: May 2014)

Abstract

Late blight of potato, caused by *Phytophthora infestans,* is responsible for high economic loss world-wide. The expression of late blight susceptibility genes (S-genes) in potato cultivars during the infection often favours the pathogenicity. Thus, the identification of the host S-genes, required to enhance susceptibility towards the pathogens under compatible interaction, is highly essential to control the infection. However, to our limited knowledge, fewer late blight S-genes have been identified in potato till date. Therefore, an attempt was made here to identify such genes through a two step classification approach using the primary and derived sequence information of potato proteins. The results revealed that WRKY transcription factor 6, Catalase protein, Shaggy-Like protein kinase NtK-1 and OTU-Like Cysteine protease were found closer to the candidate susceptibility proteins (S-proteins). These proteins were also classified into susceptible category when validated through the computer intensive techniques like Support Vector Machine, Random Forest and Artificial Neural Network. The EST database search for the above proteins has confirmed their expression under the compatible interaction. Besides, the chromosomal locations of the genes encoding these proteins were also identified, so that, the information can be utilized to develop the resistant cultivars. Thus, the predicted S-genes can be used as potential effector targets for late blight resistance in potato.

Key words: *Phytophthora infestans*, cluster analysis, S-genes, compatible interaction, computer intensive techniques

Introduction

Late blight is one of the devastating diseases occurs in potato caused by the water mold pathogen *Phytophthora infestans*. It is a major constraint in organic potato production and results in severe loss of yield and agronomic inputs. In presence of the susceptibility genes in the host and under favourable environmental conditions, the pathogen can kill off a field of potatoes just in a few days [1].

The host-pathogen interactions are generally of two types: i) compatible ii) incompatible. In compatible interaction, the host lacks the ability to defend the pathogen. The specific genes in the host, required for the infection during the compatible interaction are the S-genes [2] that increase the risk of susceptibility. Whereas, in the incompatible interaction the pathogen lacks the ability to infect the host and the genes expressed here, convey disease resistance against pathogens by producing resistance proteins (R-proteins), being referred as resistance genes (R-genes). Sometimes, the R-genes behave like S-genes, being a victim of the proteolytic activity of the pathogen virulent proteins under compatible interaction [3].

Though wet lab studies have been made related to late blight in potato, not many S-genes are reported so far to our limited knowledge. Hence, there is a need to perform extensive bio-computational analysis to identify the putative late blight S-genes in potato. Keeping the above in view, an effort was made to identify the putative late blight S-genes in potato by using computer intensive classification techniques on the primary protein structure (sequence) and the physico-chemical properties of the potato proteins.

Materials and methods

Late blight susceptibility proteins

The S-proteins related to late blight infection in different crops, including potato, were identified from literature.

*Corresponding author's e-mail: arrao@iasri.res.in; rao.cshl.work@gmail.com

Published by the Indian Society of Genetics & Plant Breeding, F2, First Floor, NASC Complex, PB#11312, IARI, New Delhi 110 012
Online management by indianjournals.com

These genes include the genes expressed in a susceptible potato cultivar during late blight infection and having high similarities with the genes of *Arabidopsis thaliana, Pisum sativum* and *Solanum tuberosum* [4]. The protein sequences of above identified genes were collected from the protein database of National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/protein/). Besides, P69B subtilase of tomato, usually a resistance gene inhibited by Extra Cellular Protease Inhibitor (EPI1) of *P. infestans* in a compatible interaction as a S gene [5], was collected from NCBI and subjected to NCBI BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) against proteome of *S. tuberosum*. From BLAST results, a subtilisin-like serine protease of potato was found with a high similarity and thus considered as an S-gene. Simko and Jones [6] have submitted an unknown gene sequence of potato, reported as an S-gene for late blight, was also collected from NCBI. The S-genes along with their protein products are given in Table 1.

Proteins expressed in late blight infected potato

In addition to the S-genes reported in Table 1, the NCBI protein databases were searched to find protein sequences submitted from late blight infected potato cultivars and a total of 78 protein sequences were filtered out for the study based on their isolation from the infected plant parts (leaves or tubers).

Physico-chemical properties of the proteins

It is evident from literature that the physico-chemical properties of proteins are related with their expression level [7, 8, 9] and function [10]. Hence, the sequences of all the proteins under study were subjected to ProtParam tool of ExPASy Proteomics Server (http://web.expasy.org/protparam/) to collect the values of the protein parameters like molecular weight, theoretical isoelectric point (pI), amino acid composition, extinction coefficient, instability index, aliphatic index and grand average of hydropathicity (GRAVY) etc. The observed length dependent parameters (molecular weight, total number of negatively charged residues, total number of positively charged residues, extinction coefficients with cysteine, extinction coefficient without cysteine, each amino acid composition) were made length independent due to the fact that the similar proteins often vary in length in spite of having same functional domains or motifs. Further, all the parametric values were standardized for classification analysis.

Prediction of late blight susceptibility genes in potato

To predict the late blight S-genes in potato, a two step approach was followed *viz*., sequence alignment based cluster analysis and protein physico-chemical properties based cluster analysis as explained below:

Step 1: Sequence alignment based cluster analysis

Since, domains are the structural components of the proteins constituting the active site residues and functional similarities in proteins correspond to the existence of similar domains, all the protein sequences under study were aligned using the Alignment tool of ClustalX2 (http://www.clustal.org) in Multiple Alignment Mode. A phylogenetic tree was constructed from the resulted MSA profile based on the Neighbour Joining method with 1000 boot strap replications in ClustalX2 and it was analysed in the interface of MEGA 5.10

Table 1. List of S-genes with accession number, protein product and the source organism

S.No.	Accession gene/EST	Protein id	Protein name	References	Organism
1	AF234984	AAF43210	Pseudouridine synthase	Evers *et al.* [4]	*Arabidopsis thaliana*
2	Z86094	CAB06698	Plastid protein	Evers *et al.* [4]	
3	X00806	CAA25390	Ribulose bisphosphate carboxylase	Evers *et al.* [4]	*Pisum sativum*
4	X52387	CAA36613* CAA36614* CAA36615* CAA36616*	Copia-like transposable element	Evers *et al.* [4]	*Solanum tuberosum*
5	DQ066722	AAY63882	Subtilisin-like serine protease	Tian *et al.* [5]	
6	AY059429	AAL30115	Unknown (reported as susceptible to late blight) Gene Bank direct submission	Simko *et al.* [6]	

*These four proteins belong to a family of retrotransposons and contain different domains being translated from the same gene (X00806) (http://www.megasoftware. net/mega_beta.php) to find out the grouping of candidate S-proteins with other proteins under study. The number of distinct groups obtained in this step was used as *a priori* information for the execution of Step 2.

Step 2: Physico-chemical properties based cluster analysis

Clustering methods and distances

There are several clustering methods and distances available for clustering objects into homogenous groups [11]. The most commonly used non-hierarchical method, *i.e.,* K-Means Clustering and the hierarchical methods *viz.* Between Group Linkage (BGL) and Ward's Minimum Variance (WMV) with the distance measures as Euclidean, Squared Euclidean and Minkowski distances [12] were used to cluster the proteins based on their standardized physico-chemical properties. Both the K-Means and hierarchical clustering analyses were performed using Statistical Package for Social Sciences (SPSS) [13] on the physico-chemical parametric values. The results obtained from K-Means clustering and six hierarchical method-distance combinations were analysed to identify the proteins close to the candidate S-proteins found under different clusters.

Criterion for identification of susceptibility proteins

The proteins, which are functionally as well as physico-chemically closer to the candidate S-proteins, were considered based on Euclidean distance:

$$\mathbf{d}(X, Y) = \sqrt{(X\text{-}Y)'(X\text{-}Y)}$$

where $\boldsymbol{X}$ is the vector of physico-chemical parametric values of the known susceptibility protein and $\mathbf{Y}$ is the vector of the physico-chemical parametric values of other proteins under study.

Validation of susceptibility genes using Computer Intensive Techniques (CIT)

Most widely used CITs like non-linear Support Vector Machine (SVM)[14], Random Forest (RF)[15] and Artificial Neural Network (ANN)[16] were used to computationally validate the new S-proteins that are predicted from the previously explained two-step approach. The SVM model was constructed with *C* classification and Gaussian *Radial* Karnel function and the RF model was tuned for minimum classification error with the parameters, *i.e.*, *mtry*=5 and *ntree*= 5000. Similarly, the ANN model was used with the *Backpropagation* learning function. The SVM, RF and ANN models described using e1071[17], randomForest [18] and RSNNS [19] packages of R software respectively were trained with the training dataset [Susceptible (S): s1-s9, Resistance (R): r1-r9] The predictions were then made for the new S-genes (t1-t4) identified based on sequence information and physico-chemical properties.

Sequence based expression analysis

The putative S-proteins identified from the analysis based on two step approach as well as computationally validated from CITs were subjected to *tblastn* program of NCBI against the Expressed Sequence Tags (EST) of *S. tuberosum*, to check whether these genes are expressed indeed.

Identification of chromosomal location

The putative S-proteins were again subjected to *tblastn* program of Solanaceae Genomics Resource (http://solanaceae.plantbiology.msu.edu/pgsc_download.shtml) to map them on to the genome of *S. tuberosum* group Phureja [20] so as to find the corresponding chromosomal locations.

Results and discussion

Various species of the genus *Phytophthora* severely affect the agriculturally important plants like potato, tomato, tobacco etc. in their natural habitat leading to high yield losses in agriculture. In most of the cases, pathogen effectors prevent recognition or suppress

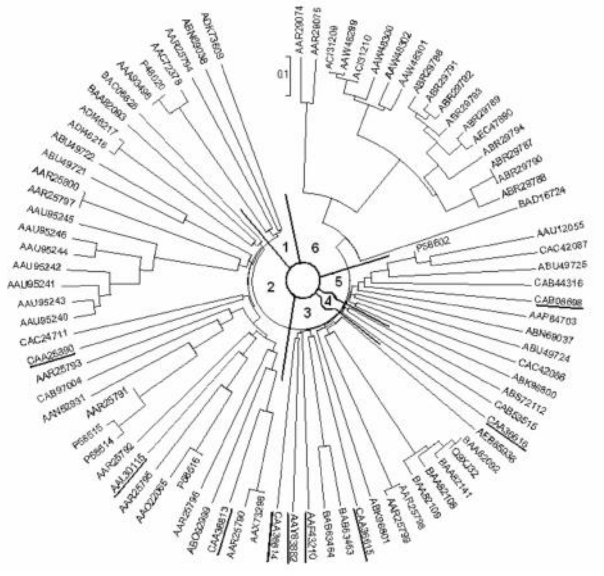


Fig. 1. MSA based dendrogram showing six different clusters with S-proteins being underlined

Table 2. A comparative analysis indicating the proteins closer to the candidate S-proteins based on physico-chemical parameters and sequence analysis. The proteins shown in bold face are consistently grouped with the candidate S-proteins in all classification methods

Reported late blight susceptibility Protein	Other late blight related proteins of potato clustering with the reported susceptibility proteins							
	Between group linkage method				Ward's method			
	Euclidean distance	Squad Euclidean distance	Minkoswiki distance	Euclidean distance	Squared Euclidean distance	Minkoswiki distance	K-means	Sequence analysis
AAY63882	CAB53515 AAR25800 AEB65936	CAB53515 AAR25800	CAB53515 AAR25800 AEB65936	CAB53515 AAR25800 AEB65936 AAR25799	CAB53515 AAR25800	CAB53515 AAR25800 AEB65936 AAR25799	CAB53515 ABR29794	BAC06825 ABU49722 ABU49721
AAF43210	AAR25794	AAR25794 AAO22065	AAR25794	AAR25794	AAR25794 CAC42087	AAR25794	CAC42087 AAC72378 ABU49724	AAN52931 BAD16724
CAA36614	CAC42087 AAU12055	CAC42087	CAC42087 AAU12055	CAC42087 AAU12055 ABU49721	CAC42087 AAU12055 AAR25799	CAC42087 AAU12055	BAA82108 AAU12055	BAC06825 BAA82141 ABU49722
CAA36615	BAC06825	BAC06825	BAC06825	BAC06825 AAR29075 AAR29074	AAR29075 BAC06825	BAC06825 AAR29075 AAR29074	ADI46216 ABS72112	BAB63463 BAB63464
CAA36616	BAC06825	BAC06825	BAC06825	BAC06825 AAR29075 AAR29074	BAC06825 AAR29075	BAC06825 AAR29075 AAR29074	AAR25797 BAC06825	ABS72112 CAB53515
CAA25390	AAN52931	AAN52931 ABU49724	AAN52931	AAN52931 AAU12055 CAC42087	AAN52931 AAU12055	AAN52931 AAU12055 CAC42087	AAO22465 AAR25795 AAP84703 P58515	CAC2471
CAB06698	**ABU49725** ABU49724	**ABU49725** AAR25795 ABU49724	**ABU49725** ABU49724 ABU49725	AAR25795 **ABU49725** ABU49724	ABU49725 **AAR25795** ABU49724	AAR25795 **ABU49725** ABU49724	ABU49722 ABN69037 ABK96800 ABN69038 ABU49721 **ABU49725**	CAB44316 **ABU49725** CAC42087 AAU12055 P58602
CAA36613	CAB97004 AAR25793 **AAR25790** **AAR25796**	CAB97004 AAR25793 **AAR25790** **AAR25796**	CAB97004 AAR25793 **AAR25790** **AAR25796**	CAB97004 AAR25793 **AAR25796** **AAR25790**	CAB97004 AAR25793 **AAR25796** **AAR25790**	CAB97004 AAR25793 **AAR25796** **AAR25790**	AAC72378 AAR25793 **AAR25790** **AAR25796** CAC42086	**AAR25790** AAX73296 ABO92999 **AAR25796**
AAL30115	AAR25790 AAR25796 **AAR25793** CAB97004	AAR25790 AAR25796 **AAR25793** CAB97004	AAR25790 AAR25796 **AAR25793** CAB97004	**AAR25793** CAB97004 AAR25796 AAR25790	**AAR25793** CAB97004 AAR25796 AAR25790	**AAR25793** CAB97004 AAR25796 AAR25790	**AAR25793** AAR25790 AAR25796 AAC72378 CAC42086	AAR25795 AAO22065 P58516 P58514 P58515 AAR25791 CAB97004 AAR25793

host defence mechanism. However, successful suppression of host defence is not always sufficient for pathogenesis, which requires further host-components like proteins, metabolites etc. that meet the demands of pathogen development and nutrition [21]. However, the disease susceptibility can be avoided by inhibition of these negative regulators of defence.

The S-proteins considered in this study are based on Evers *et al.* [4], Tian *et al.* [5] and Simko and Jones [6]. These proteins belong to mainly five different classes that are plastid protein, ribulose bisphosphate carboxylase, copia-like transposable element, subtilisin-like serine protease, pseudouridine synthase and evidenced to have connection with disease susceptibility [4, 22-24, 25]. To our limited knowledge, pseudouridine synthase with late blight susceptibility is not reported so far in potato. However, a protein dyskerin, having pseudouridine synthase domain, was reported to be associated with the Cajal bodies and nucleolus that are required for systemic viral infections in plants [26].

The phylogenetic tree constructed from MSA profile of all the proteins is shown in Fig. 1 six major clusters were observed. The candidate S-proteins were found distributed in different clusters based on their functional similarity. The second and the sixth clusters were found as the larger clusters. The candidate S-proteins found in second cluster are CAA25390, CAA36613 and AAL30115 whereas in third cluster are CAA36614, AAF43210, AAY63882 and CAA36615. CAB06698 is the only candidate S-protein found in the fifth cluster. The fourth cluster contains only one candidate S-protein, *i.e.*, CAA36615. Besides, third and fourth clusters were found closer to each other than other clusters. However, first and sixth clusters were not observed with any of the candidate S-proteins.

The cluster membership of each protein was obtained after application of the K-Means clustering procedure (K=6) on the standardized parametric values. Out of six clusters, cluster 2 was observed with the S-proteins, *viz*., CAB06698, AAF43210, CAA36613, AAL30115 whereas all other S-proteins were observed in cluster 4. The clusters 1, 3, 5 and 6 were found to contain none of the S-proteins. The dendrograms obtained from hierarchical clustering for six method distance combinations are presented in Supplementary Figs. 1 and 2 (available online at http://www.isgpb.co.in) where the known S-proteins are shown in bold face and underlined format. The results revealed that the proteins CAB53515 and AAR25800 are clustered with the S-protein AAY63882. Also, the proteins AAR25794 and AAN52931 were found closer with S-proteins AAF43210 and CAA25390 respectively. Whereas, the proteins CAC42087 and AAU12055 were found clustered with the S-protein CAA36614 in all methods.

From all the method-distance combinations, the protein BAC06825 was found with the two S-proteins CAA36615 and CAA36616 in one cluster and the proteins ABU49725 and ABU49724 remained with the S-protein CAB06698 in another cluster. Besides, the proteins CAB97004, AAR25793, AAR25790 and AAR25796 were found together with two S-proteins CAA36613 and AAL30115 in the same cluster under BGL and WMV methods. The cluster numbers used here are specific to a particular method. For example, the cluster numbered as 1 in one method need not necessarily same as that in other methods. In

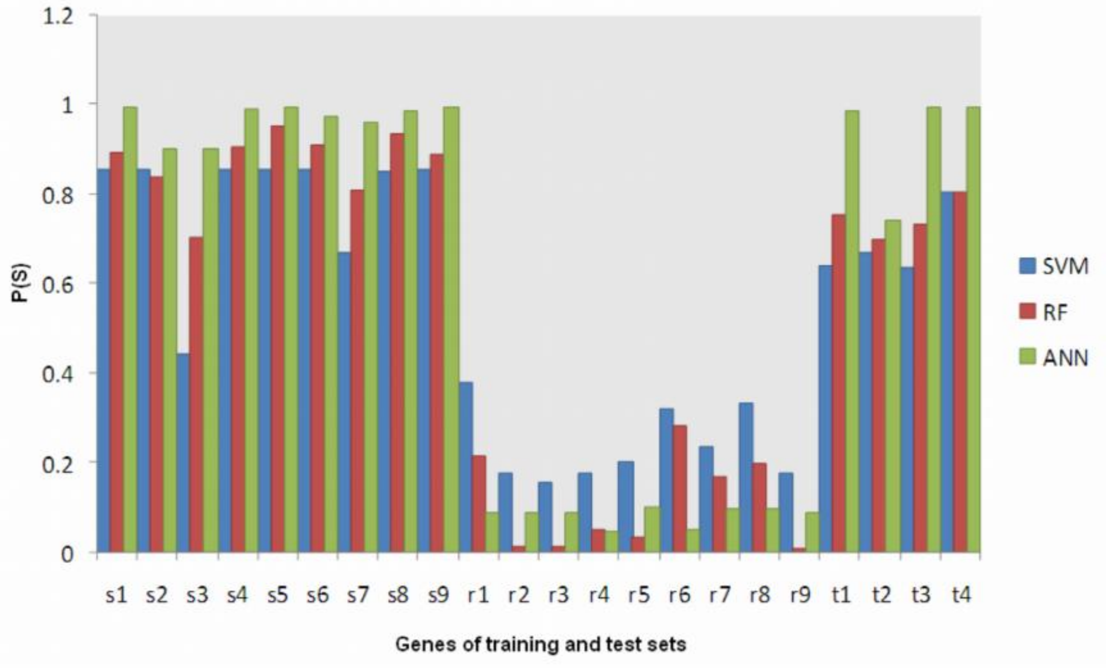


Fig. 2. Comparison of three CITs for classification of genes into S category. P(S) – probability of being classified as S-category

summary, the results obtained from the two-step approach revealed that the proteins, *viz*., ABU49725, AAR25790, AAR25793 and AAR25796 were found consistently grouped with the candidate S-proteins in all classification methods (Table 2). Hence, these proteins are referred as putative S-proteins from now onwards.

The results obtained from the classification analysis based on sequence alignment reveal a total of six clusters consisting of both the candidate S-proteins as well as other proteins expressed in late blight infected potato cultivars. This information was further used in statistical analysis for grouping these proteins into clusters, with the expectation that the proteins functionally closer to each other are likely to possess similar physico-chemical properties. Hence, the property based clustering was done with K=6 for K-means clustering method and grouped into 6 clusters under hierarchical methods as well. The clustering results showed that the candidate S-proteins CAB06698 and AAL30115 were found closer with ABU49725 and AAR25793 respectively whereas another reported S-protein CAA36613 was found clustered with two proteins, *viz.,* AAR25790, AAR25796. From Table 3, it is depicted that the newly predicted Late blight S-proteins of potato are generally catalytic enzymes and transcription factors, which are likely to be involved in host-pathogen interaction during infection. The protein with unknown function, *i.e.,* AAR25796 is also expected to be an S-protein as it was clustered with the candidate S-protein, *i.e.,* CAA36613.

The CAB06698, a plastid protein of *A. thaliana* (having no identified domain) was found to be closer to the ABU49725 (WRKY Transcription Factor 6). Shang *et al.* [27] studied the association of WRKY Transcription Factors with resistance genes as well as the interaction between the WRKY proteins and the chloroplast/plastid-localized ABA receptor. Probably, due to such interaction, the WRKY Transcription Factor 6 of *S. tuberosum* was found together with the plastid protein of *A. thaliana* in the same cluster. Besides, Dellagi *et al.* [28] reported that WRKY1 of *S. tuberosum* is strongly up-regulated in the compatible interaction whereas weakly in incompatible interaction, which confirms the association of WRKY domain with the late blight pathogenesis.

The protein AAL30115, an S-protein of potato with unknown function was found clustered with the Shaggy Like Protein Kinase NtK 1 of *S. tuberosum*, which belongs to PKc like superfamily. Avrova *et al.* [29] found a cloned cDNA sequence from a susceptible

Table 3. Physico-chemical properties and functional domains, related ESTs and the chromosomal locations of the identified putative Late blight S-proteins of potato. PN: Protein Name, Chr: chromosome, CDS: coding sequence (Exons), coord: coordinates, ID: identity

Accessions		ABU49725	AAR25790	AAR25793	AAR25796
PN		WRKY transcription factor 6	Catalase	Shaggy-like protein kinase NtK-1	unknown protein
Domain		WRKY super family	Catalase-like heme-binding proteins and protein domains	Protein Kinases, catalytic domain	No domains detected
Function		DNA binding transcription factor	Catalyses the conversion of hydrogen peroxide to water and molecular oxygen	Catalyses the transfer of the gamma-phosphoryl group from ATP to hydroxyl groups in specific substrates such as serine, threonine, or tyrosine residues of proteins	Similar to Arabidopsis thaliana unknown protein deposited in GenBank Accession number AY085012
Chr_ location	Chr_no	9	12	1	7
	Strand	+ve	+ve	-ve	+ve
	Chr_coord	18097665-18099258	55392140-55392535	71491585-71489934	39795667-39797295
	CDS	18097665-18097985, 18098556-18098663, 18098776-18099258	55392140-55392406, 55392491-55392535	71491585-71491526, 71491426-71491364, 71490207-71490085, 71489996-71489934	39795667-39795771, 39796369-39796446, 39797239-39797295
	%ID	100%	100%	@ 100%	100%

potato cultivar similar to Shaggy Like Protein Kinase (NtK-1) of *Nicotiana tabacum* under a compatible interaction that confirms its relation with the susceptibility of potato cultivars towards *P. infestans* infection.

CAA36613 (unnamed protein product), a known susceptibility gene from Kennbee potato cultivar, was found closer to two proteins, *viz.,* AAR25790 (Catalase Protein of potato) and AAR25796 (unknown protein). Chumakov and Zakharova [30] stated that, in general, catalases split hydrogen peroxide (H_2O_2) which is an antimicrobial endogenous agent protecting plants against pathogens. Therefore, catalases are often considered as components of pathogen aggression [31]. Hence, the catalase protein, *i.e.,* AAR25790 of potato is expected to behave as a late blight S-protein on it's over expression.

The function of unknown protein AAR25796 was identified using Conserved Domain Database (CDD) search (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) and BLASTP of NCBI. The results showed that it possess 55-78% of identity with most of the hypothetical, predicted, unknown or uncharacterised protein. But, interestingly, it showed a maximum of 75% identity with two known proteins NP_566704 and NP_974352 of *A. thaliana* that belongs to OTU (Ovarian Tumor) like cystein protease superfamily. Besides, it was also found similar with some other known proteins of rice and zea mays (Supplementary Table 2 available online at http://www.isgpb.co.in) having cysteine protease activity. The BLAST result showed that AAR25796 (1-80) was found aligned with the 38-115 residues of the closest known sequence, OTU-like cysteine protease of *A. thaliana, i.e.,* NP_566704 whereas the OTU Like Cystein protease superfamily was found at residue positions 107-219 on the subject sequence (Supplementary Figure 4 and 5 available online at http://www.isgpb.co.in). It implies that AAR25796 contains initial 9 residues of the OTU Like Cystein Protease superfamily and expected to be a partial sequence of OTU Like Cystein protease of potato. The viral OTU proteases inhibit Uband ISG15-dependent antiviral pathways in host [32]. Therefore, it is expected that the pathogen could utilize the host OTU-like cysteine protease to inhibit the antiviral pathway in potato.

The computational validations of the putative S-proteins by CITs have shown that the trained CIT models have classified all the putative S-proteins (4 test proteins: t1-t4) under the S category with high probabilities. The graphical representation for the probabilities of training and test observations that were classified under the S or R categories by different CITs are given in Fig. 2 created based on Supplementary Table 1(available online at http://www.isgpb.co.in).

The sequence search against EST database for all the putative S-proteins have indicated that the accession ABU49725 has 99% of identity with top three hits (CK265690, BI434362, BG592222) out of which the last two are from *P. infestans* challenged potato leaf under compatible (susceptible) interaction. The accession AAR25790 showed 100% identity with 68 sequences (41 with 100% query coverage) out of which 8 sequences are from *P. infestans* challenged potato leaf. Though AAR25793 has shown 80-99% of identity with many ESTs, two of them (BG590895, BG592192) were also found from *P. infestans* challenged potato leaf with 92% and 86% identity respectively. The accession AAR25796 has confirmed 100% identity with complete query coverage against three EST sequences (CN213326, AM908303, AM908432) and it was also found 57% identical with another EST, *i.e.,* BI435076, sequenced from a *P. infestans* challenged potato leaf. Further, all the putative S-genes were found to map with the chromosomes of *S. tuberosum* group Phureja with around 100% identity and hence their chromosomal locations were identified. The chromosomal locations of these proteins (genes) on the genome of *S. tuberosum* group Phureja are also given in Table 3. The physical mapping of the proteins on to the genome of *S. tuberosum* group Phureja is shown in Supplementary Fig. 3 (available online at http://www.isgpb.co.in).

The CITs were trained with the known susceptible (s1-s9; Table 1) and resistance genes (r1-r9; reported in the NCBI Sequence profile) to predict the test proteins, *viz.,* ABU49725, AAR25790, AAR25793 and AAR25796 (t1-t4). Among the three CITs, ANN followed by RF and SVM have shown high discrimination between susceptible(S) and resistant(R) categories (Figure 2). Besides, the putative S-proteins (t1-t4) have shown high identity with the ESTs. Further, these proteins have also shown identities with ESTs isolated from the *P. infestans* challenged potato leaves. This indicates that the genes encoding these putative S-proteins are expressed indeed and are probably expressed during the compatible interaction. The genes were found to be well-mapped with around 100% identity on to the genome of *S. tuberosum* group Phureja confirming that the predicted S-genes are indeed real and exist. The information on the chromosomal locations of the putative S-proteins provided here can thus be used in the field of genetic engineering for the development of late blight resistant potato cultivars.

In conclusion, the genes encoding WRKY transcription factor 6, Catalase ptotein, Shaggy-like protein kinase NtK-1 and OTU Like cysteine protease of potato are the putative late blight S-genes as they showed similarity with the candidate susceptibility proteins in terms of both amino acid sequences and physico-chemical properties. Some of these S-genes can be the potential effector targets and hence can be used in breeding for resistance.

Acknowledgements

This study was supported by World Bank Funded National Agricultural Innovation Project (NAIP), ICAR Grants 30(68)/2009/Bio Informatics/NAIP/ O&M.

References

1. **Volk T. J.** 2001. Phytophthora infestans, cause of late blight of potato and the Irish potato Famine. University of Wisconsin-La Crosse. (http://botit.botany.wisc.edu/toms_fungi/m2001alt.html)
2. **Eckardt N. A.** 2002. Plant disease susceptibility genes? Plant Cell, **14**: 1983-1986, doi: 10.1105/tpc.140910.
3. **Bouwmeester K., Klamer S., Gouget A., Haget N., Canut H. and Govers F.** 2008. Lectin receptor kinase 79, a putative target of the Phytophthora infestans effector IPI-O.*In*: Biology of Plant–Microbe Interactions, vol. VI (edS. M. Lorito, SL. Woo and F. Scala). International Society for Molecular Plant-Microbe Interactions, St. Paul, Minnesota, USA.
4. **Evers D., Ghislain M., Hausman J. F. and Dommes J.** 2003. Differential gene expression in two potato lines differing in their resistance to Phytophthora infestans. J. Plant Physiol., **160**: 709-712.
5. **Tian M., Benedetti B. and Kamoun S.** 2005. A second kazal-like protease inhibitor from Phytophthora infestans inhibits and interacts with the apoplastic pathogenesis-related protease P69B of tomato. Plant Physiol., **138**: 1785-1793.
6. **Simko I. and Jones R.W.** 2001. Direct submission of unknown gene of potato AY059429, conceptually translated to a protein AAL30115.1 in GenBank, GI: 16933579.
7. **Idicula-Thomas S. and Balaji P. V.** 2005. Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in Escherichia coli. Protein Sci., **14**: 582-592.
8. **Madhavan V., Bhatt F. and Jeffery C. J.** 2010. Recombinant expression screening of P. aeruginosa bacterial inner membrane proteins. BMC Biotechnol., **10**: 83, doi: 10.1186/1472-6750-10-83.
9. **Price W. N., Handelman S. K., Everett J. K., Tong S. N., Bracic A., Luff J. D.** *et al*. 2011. Large-scale experimental studies show unexpected amino acid effects on protein expression and solubility *in vivo* in *E. Coli*. Microb. Inform. Exp., **1**: 6, doi: 10.1186/2042-5783-1-6.
10. **Shenoy S. R. and Jayaram B.** 2010. Proteins: sequence to structure and function-current status. Curr. Protein Pept. Sci., **11**: 498-514.
11. **Johnson R. A. and Wichern D. W.** 2007. Applied multivariate statistical analysis. 6th Edition. Prentice Hall, New Jersey.
12. **Wahi S. D., Dash S. and Rao A. R.** 2009. An empirical inve stigation on classical clustering methods. ICFAI Univ. J. Genet. Evol., **2**: 74-79.
13. **Argyrous G.** 2005. Statistics for research: with a guide to SPSS, 2nd Edition. Sage, London.
14. **Cortes C. and Vapnik V.** 1995. Support-Vector Networks. Mach. Learn., **20**: 273-297.
15. **Breiman L.** 2001. Random Forests. Mach. Learn., **45**: 5-32.
16. **Haykin S.** 1998. Neural Networks: a comprehensive foundation. Prentice Hall, Upper Saddle River, New Jersey.
17. **Meyer D., Dimitriadou E., Hornik K., Weingessel A. and Leisch F.** 2012. e1071: Misc functions of the Department of Statistics (e1071), TU Wien, R package version 1.6-1. (http://CRAN.R-project.org/package =e1071)
18. **Liaw A. and Wiener M.** 2002. Classification and Regression by random. Forest. R. News, **2**: 18-22.
19. **Bergmeir C. and Benýtez J. M.** 2012. Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS. J. Stat. Soft., **46**: 1-26.
20. **Potato Genome Sequencing Consortium, Xu X., Pan S., Cheng S., Zhang B., Mu D.** ***et al*.** 2011. Potato Genome Sequencing Consortium: Genome sequence and analysis of the tuber crop potato. Nature,**475**: 189-195.
21. **Hückelhoven R., Eichmann R., Weis C., Hoefle C. and Proels R. K.** 2013. Genetic loss of susceptibility: a costly route to disease resistance. Plant Pathol., doi: 10.1111/ppa.12103.
22. **Conrath U., Linke C., Jeblick W., Geigenberger P., Quick W. P. and Neuhaus H. E.** 2003. Enhanced resistance to Phytophthora infestans and Alternaria solani in leaves and tubers, respectively, of potato plants with decreased activity of the plastidic ATP/ADP transporter. Planta, **217**: 75-83.

23. **Tian M., Huitema E., Da Cunha L., Torto-Alalibo T. and Kamoun S.** 2004. A Kazal-like extracellular serine protease inhibitor from Phytophthora infestans targets the tomato pathogenesis-related protease P69B. J. Biol. Chem., **279**: 26370-26377.

24. **Kombrink E. and Hahlbrock K.** 1990. Rapid, systemic repression of the synthesis of ribulose 1, 5-bisphosphate carboxylase small-subunit mRNA in fungus-infected or elicitor-treated potato leaves. Planta, **181**: 216-219.

25. **Slotkin R. K. and Martienssen R.** 2007. Transposable elements and the epigenetic regulation of the genome. Nat. Rev. Genet., **8**: 272-285, doi:10.1038/nrg2072.

26. **Kim S. H., Ryabov E. V., Kalinina N. O., Rakitina D. V., Gillespie T., MacFarlane S. *et al*.** 2007. Cajal bodies and the nucleolus are required for a plant virus systemic infection. EMBO J., **26**: 2169-2179, doi:10.1038/sj.emboj.7601674.

27. **Shang Y., Yan L., Liu Z. Q., Cao Z., Mei C., Xin Q. *et al*.** 2010. The Mg-chelatase H subunit of Arabidopsis antagonizes a group of WRKY transcription repressors to relieve ABA responsive genes of inhibition. Plant Cell, **22**: 1909-1935.

28. **Dellagi A., Helibronn J., Avrova A. O., Montesano M., Palva E. T., Stewart H. E. *et al*.** 2000. A potato gene encoding a WRKY-like transcription factor is induced in interactions with Erwinia carotovora subsp. atroseptica and Phytophthora infestans and is coregulated with class I endochitinase expression. Mol. Plant Microbe Interact., **13**: 1092-1101.

29. **Avrova A. O., Taleb N., Rokka V. M., Heilbronn J., Campbell E., Hein I.** *et al***.** 2004. Potato oxysterol binding protein and cathepsin B are rapidly up-regulated in independent defence pathways that distinguish R gene-mediated and field resistances to Phytophthora infestans. Mol. Plant Pathol., **5**: 45-56.

30. **Chumakov A. E. and Zakharova T. I.** 1990. Vredonosnost' boleznei sel'skokhozyaistvennykh kul'tur (Harmfulness of Crop Diseases). Agropromizdat, Moscow.

31. **Bolwell G. P. and Daudi A.** 2009. Reactive oxygen species in plant-pathogen interactions. *In*: Reactive Oxygen Species in Plant Signaling. Signaling and Communication in Plants. (ed. L.A. Rio and A. Puppo), Springer Berlin Heidelberg, Berlin: pp 113-133.

32. **Frias-Staheli N., Giannakopoulos N. V., Kikkert M., Taylor S.L., Bridgen A., Paragas J. *et al*.** 2007. Ovarian Tumor (OTU)-domain containing viral proteases evade Ubiquitin- and ISG15-dependent innate immune responses. Cell Host Microbe, **2**: 404-416.

Supplementary Materials

Supplementary Table 1. Probabilities of the S-genes, R-genes and test genes, being classified under Susceptible(S) and Resistant(R) groups obtained from the CITs (SVM, RF and ANN). Here, s1 to s9 are candidate S-genes, r1 to r9 are the candidate R-genes and t1 to t4 are the test genes. The test genes were predicted as putative S-genes from sequence and physico-chemical property based analysis and also these genes were validated through CITs.

Accession	Category	SVM		RF		ANN	
		S	R	S	R	S	R
AAF43210	s1	0.854168	0.145832	0.891	0.109	0.994017	0.006039
AAL30115	s2	0.854249	0.145751	0.8378	0.1622	0.902947	0.094169
AAY63882	s3	0.440428	0.559572	0.7034	0.2966	0.902981	0.098502
CAA25390	s4	0.854133	0.145867	0.9036	0.0964	0.989617	0.010366
CAA36613	s5	0.85424	0.14576	0.9522	0.0478	0.995132	0.004953
CAA36614	s6	0.854168	0.145832	0.9106	0.0894	0.973172	0.02668
CAA36615	s7	0.670964	0.329036	0.807	0.193	0.958184	0.041366
CAA36616	s8	0.850871	0.149129	0.9354	0.0646	0.983238	0.016604
CAB06698	s9	0.85413	0.14587	0.8864	0.1136	0.9941	0.005991
AAU95246	r1	0.380591	0.619409	0.215	0.785	0.089862	0.908953
AAW48300	r2	0.175783	0.824217	0.0134	0.9866	0.08657	0.912776
AAW48302	r3	0.156392	0.843608	0.0114	0.9886	0.08753	0.911951
ABO92999	r4	0.175734	0.824266	0.0498	0.9502	0.048117	0.95106
ABR29791	r5	0.202138	0.797862	0.0326	0.9674	0.098991	0.900047
ACI31209	r6	0.318411	0.681589	0.2826	0.7174	0.051634	0.947392
AAP84703	r7	0.23418	0.76582	0.1672	0.8328	0.098304	0.901097
AAU95242	r8	0.331076	0.668924	0.1986	0.8014	0.096328	0.902909
AAW48299	r9	0.175622	0.824378	0.0082	0.9918	0.087741	0.911525
AAR25790	t1	0.641527	0.358473	0.7552	0.2448	0.984392	0.016029
AAR25793	t2	0.671474	0.328527	0.698	0.302	0.741503	0.25608
AAR25796	t3	0.637633	0.362367	0.734	0.266	0.993512	0.006557
ABU49725	t4	0.802905	0.197095	0.8056	0.1944	0.993111	0.007007

(*ii*)

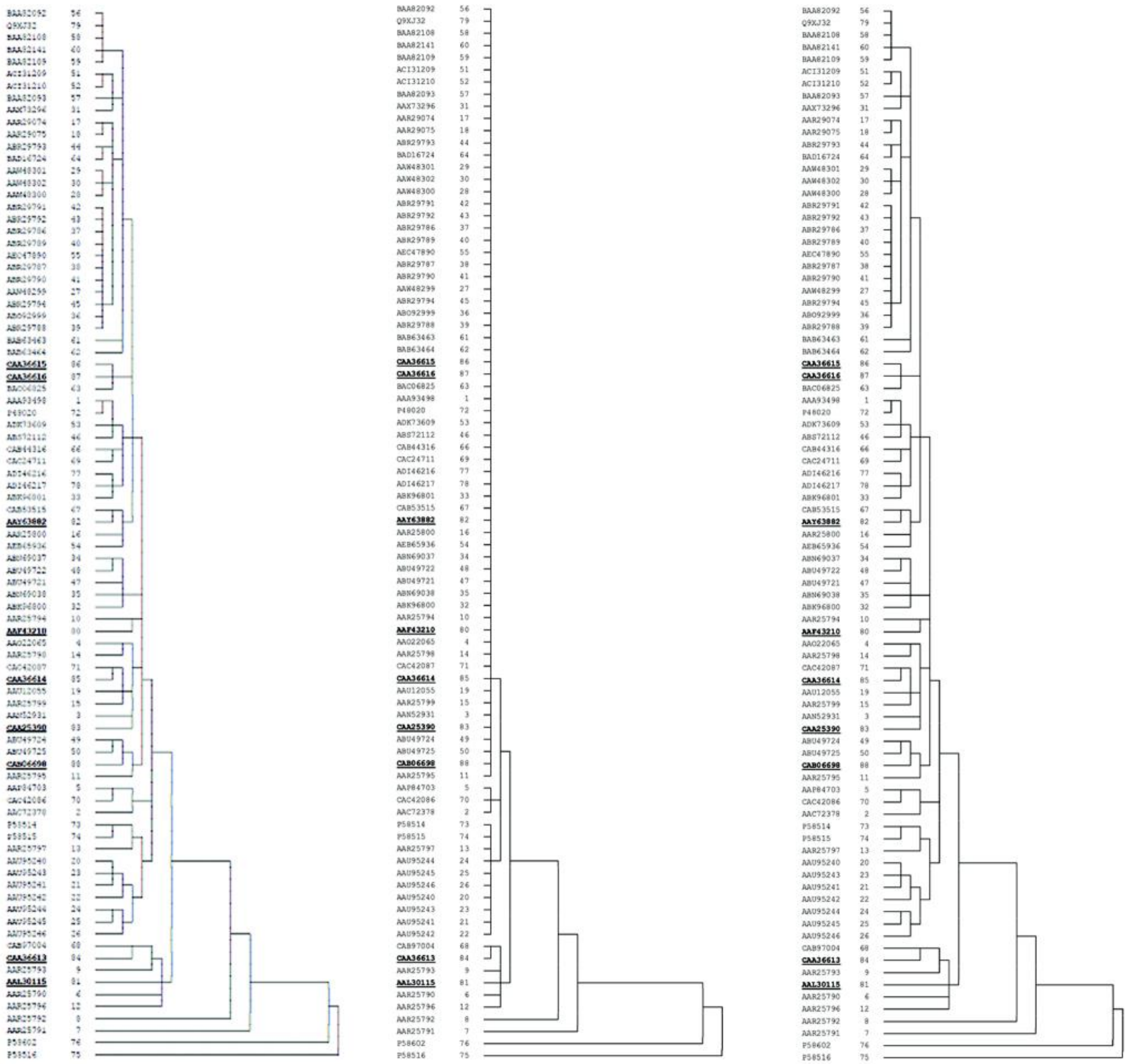
Supplementary Table 2. BLAST result of AAR25796 showing the Score, Query Coverage (QC), Expectation Value (E Value) and percentage of identity (ID) with the known proteins

Accession	Description	Score	QC	E value	% identity
NP_566704.1	OTU-like cysteine protease family protein [*Arabidopsis thaliana*] >gb|AEE76613.1| OTU-like cysteine protease family protein [*Arabidopsis thaliana*]	116	100%	1.00E-31	75%
NP_974352.1	OTU-like cysteine protease family protein [*Arabidopsis thaliana*] >ref|NP_001189948.1| OTU-like cysteine protease family protein [*Arabidopsis thaliana*] > gb|ABH04458.1| At3g22260 [*Arabidopsis thaliana*] gb|AEE76614.1| OTU-like cysteine protease family protein [*Arabidopsis thaliana*] >gb|AEE76615.1| OTU-like cysteine protease family protein [*Arabidopsis thaliana*]	117	100%	1.00E-31 >	75%
NP_001151701.1	LOC100285337 [*Zea mays*] >gb|ACG44045.1| cysteine-type peptidase [*Zea mays]*	99.4	88%	6.00E-25	67%
NP_186856.1	cysteine proteinase-like protein [*Arabidopsis thaliana*] >gb|AAF14829.1|AC011664_11 unknown protein [*Arabidopsis thaliana*] >gb|AAS76700.1| At3g02070 [*Arabidopsis thaliana*] >gb|AAS92324.1| At3g02070 [*Arabidopsis thaliana*] >dbj|BAH20151.1| AT3G02070 [*Arabidopsis thaliana*] >gb|AEE73759.1| cysteine proteinase-like protein [*Arabidopsis thaliana*]	91.3	100%	7.00E-22	55%
CAD40788.2	OSJNBb0012E08.12 [*Oryza sativa Japonica* Group] >emb|CAD40683.2| OSJNBb0118P14.1 [*Oryza sativa Japonica* Group]	89	100%	2.00E-21	52%
ACG28098.1	cysteine-type peptidase [*Zea mays*]	88.6	98%	9.00E-21	53%
NP_001147603.1	cysteine-type peptidase [*Zea mays*] >gb|ACG28056.1| cysteine-type peptidase [*Zea mays*] >gb|ACR38119.1| [*Zea mays*]	85.5	100%	1.00E-19	51% unknown
NP_001148776.1	cysteine-type peptidase [*Zea mays*] >gb|ACG32866.1| cysteine-type peptidase [*Zea mays*]	74.3	88%	8.00E-15	46%
NP_568136.1	OTU-like cysteine protease family protein [*Arabidopsis thaliana*] >ref|NP_001119168.1| OTU-like cysteine protease family protein [*Arabidopsis thaliana*] >gb|AAM64524.1| unknown [*Arabidopsis thaliana*] >dbj|BAD95211.1| hypothetical protein [*Arabidopsis thaliana*] >gb|ABD85149.1| At5g04250 [*Arabidopsis thaliana*] >gb|AED90718.1| OTU-like cysteine protease family protein [*Arabidopsis thaliana*] >gb|AED90719.1| OTU-like cysteine protease family protein [*Arabidopsis thaliana*]	71.2	98%	1.00E-13	44%
NP_001048535.1	Os02g0819500 [*Oryza sativa Japonica* Group] >dbj|BAD22969.1| OTU-like cysteine protease-like [*Oryza sativa Japonica* Group] >dbj|BAD23098.1|	69.3	72%	1.00E-13	50%

(*iii*)

	OTU-like cysteine protease-like [*Oryza sativa Japonica* Group] >dbj|BAF10449.1| Os02g0819500 [*Oryza sativa Japonica* Group] >dbj|BAG97492.1| unnamed protein product [*Oryza sativa Japonica* Group]				
NP_001051970.1	Os03g0859800 [*Oryza sativa Japonica* Group] >gb|ABG00013.1| OTU-like cysteine protease family protein, putative, expressed [*Oryza sativa Japonica* Group] >dbj|BAF13884.1| Os03g0859800 [*Oryza sativa Japonica* Group] >dbj|BAG92774.1| unnamed protein product [*Oryza sativa Japonica* Group] >gb|EEE60341.1| hypothetical protein OsJ_13452 [*Oryza sativa Japonica* Group]	70.5	98%	2.00E-13	43%
ABG00014.1	OTU-like cysteine protease family protein, putative, expressed [*Oryza sativa Japonica* Group] >gb|ABG00015.1| OTU-like cysteine protease family protein, putative, expressed [*Oryza sativa Japonica* Group]	68.9	98%	4.00E-13	43%
BAD22968.1	OTU-like cysteine protease-like [*Oryza sativa Japonica* Group] >dbj|BAD23097.1| OTU-like cysteine protease-like [*Oryza sativa Japonica* Group] >dbj|BAG91428.1| unnamed protein product [*Oryza sativa Japonica* Group] >dbj|BAG99627.1| unnamed protein product [*Oryza sativa Japonica* Group]	68.9	72%	6.00E-13	50%
EEE58057.1	hypothetical protein OsJ_08894 [*Oryza sativa Japonica* Group]	68.9	72%	6.00E-13	50%
EEC74259.1	hypothetical protein OsI_09471 [*Oryza sativa Indica* Group]	68.9	72%	6.00E-13	50%
XP_003590457.1	Cysteine-type peptidase [*Medicago truncatula*] >gb|AES60708.1| Cysteine-type peptidase [*Medicago truncatula*]	67.8	71%	2.00E-12	50%
NP_001050580.1	Os03g0589300 [*Oryza sativa Japonica* Group] >gb|AAV35815.1| OTU-like cysteine protease domain containing protein [*Oryza sativa Japonica* Group] >gb|ABF97379.1| OTU-like cysteine protease family protein, expressed [*Oryza sativa Japonica* Group] >dbj|BAF12494.1| Os03g0589300 [*Oryza sativa Japonica* Group] >dbj|BAG93237.1| unnamed protein product [*Oryza sativa Japonica* Group]	65.5	73%	2.00E-11	47%

(*iv*)

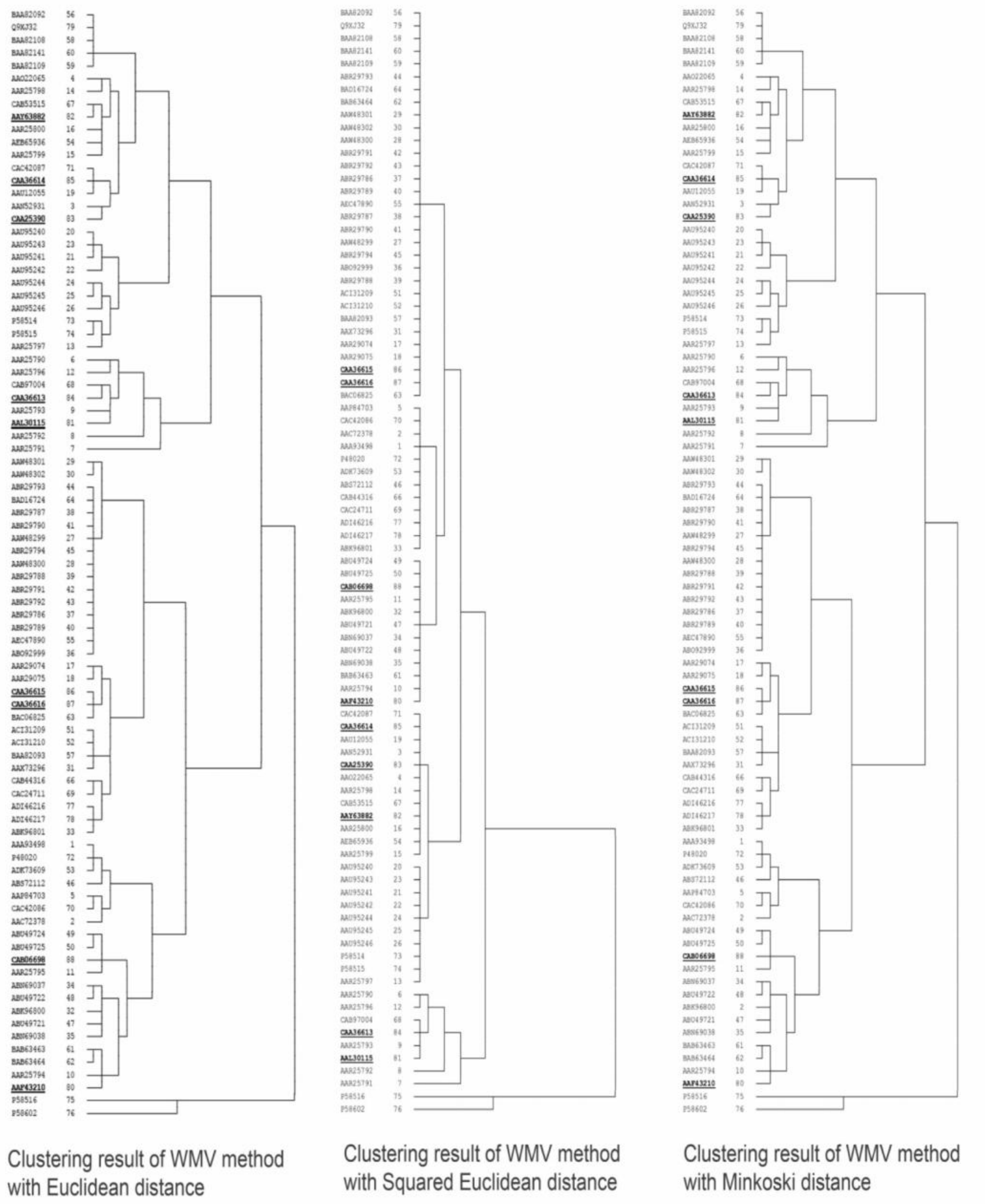


Clustering result of BGL method with Euclidean distance

Clustering result of BGL method with Squared Euclidean distance

Clustering result of BGL method with Minkoski distance

Supplementary Fig. 1. Result of hierarchical clustering analysis (BGL) showing the candidate S-proteins in bold face and underlined format



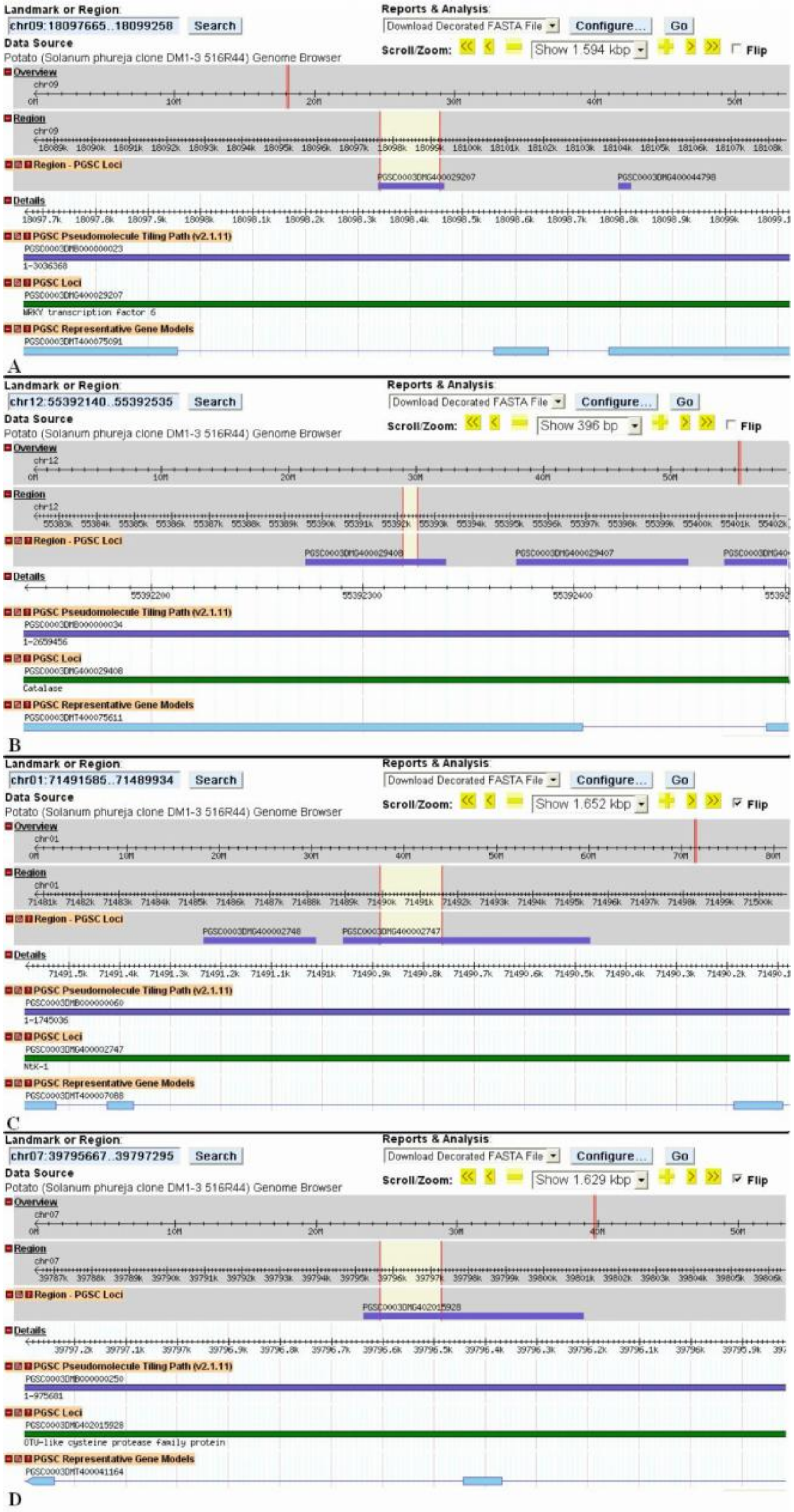
Clustering result of WMV method with Euclidean distance

Clustering result of WMV method with Squared Euclidean distance

Clustering result of WMV method with Minkoski distance

Supplementary Fig. 2. Result of hierarchical clustering analysis (WMV) showing the candidate S-proteins in bold face and underlined format

(*vi*)



Landmark or Region: chr09:18097665..18099258 Search
Data Source
Potato (Solanum phureja clone DM1-3 516R44) Genome Browser
Reports & Analysis: Download Decorated FASTA File Configure... Go
Scroll/Zoom: Show 1.594 kbp □ Flip

Overview
chr09
0M 10M 20M 30M 40M 50M
Region
chr09
18089k 18090k 18091k 18092k 18093k 18094k 18095k 18096k 18097k 18098k 18099k 18100k 18101k 18102k 18103k 18104k 18105k 18106k 18107k 18108k
Region - PGSC Loci
PGSC0003DMG400029207
PGSC0003DMG400044798
Details
18097.7k 18097.8k 18097.9k 18098k 18098.1k 18098.2k 18098.3k 18098.4k 18098.5k 18098.6k 18098.7k 18098.8k 18098.9k 18099k 18099.1
PGSC Pseudomolecule Tiling Path (v2.1.11)
PGSC0003DMB000000023
1-3036368
PGSC Loci
PGSC0003DMG400029207
WRKY transcription factor 6
PGSC Representative Gene Models
PGSC0003DMT400075091
A

Landmark or Region: chr12:55392140..55392535 Search
Data Source
Potato (Solanum phureja clone DM1-3 516R44) Genome Browser
Reports & Analysis: Download Decorated FASTA File Configure... Go
Scroll/Zoom: Show 396 bp □ Flip

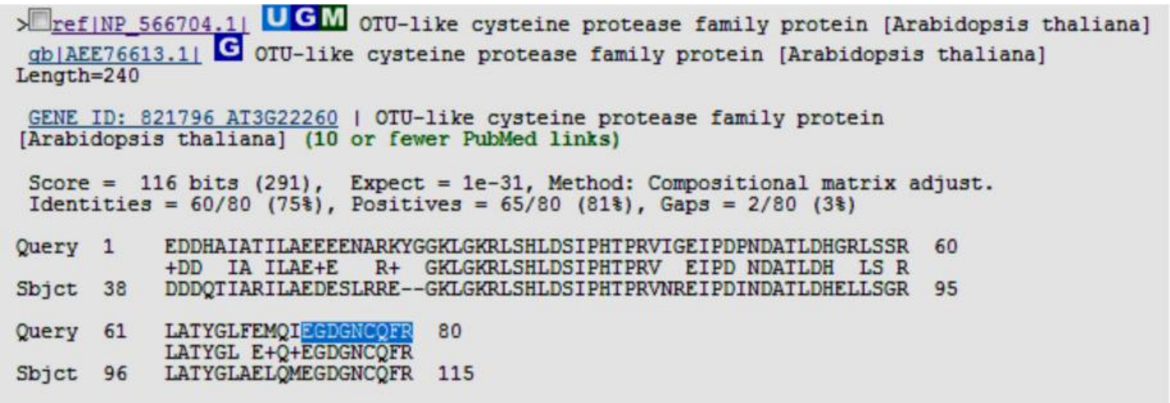
Landmark or Region: chr01:71491585..71489934 Search
Data Source
Potato (Solanum phureja clone DM1-3 516R44) Genome Browser
Reports & Analysis: Download Decorated FASTA File Configure... Go
Scroll/Zoom: Show 1.652 kbp ☑ Flip

Overview
chr01
0M 10M 20M 30M 40M 50M 60M 70M 80M
Region
chr01
71481k 71482k 71483k 71484k 71485k 71486k 71487k 71488k 71489k 71490k 71491k 71492k 71493k 71494k 71495k 71496k 71497k 71498k 71499k 71500k
Region - PGSC Loci
PGSC0003DMG400002748
PGSC0003DMG400002747
Details
71491.5k 71491.4k 71491.3k 71491.2k 71491.1k 71491k 71490.9k 71490.8k 71490.7k 71490.6k 71490.5k 71490.4k 71490.3k 71490.2k 71490.1
PGSC Pseudomolecule Tiling Path (v2.1.11)
PGSC0003DMB000000060
1-1745036
PGSC Loci
PGSC0003DMG400002747
NtK-1
PGSC Representative Gene Models
PGSC0003DMT400007088
C

Landmark or Region: chr07:39795667..39797295 Search
Data Source
Potato (Solanum phureja clone DM1-3 516R44) Genome Browser
Reports & Analysis: Download Decorated FASTA File Configure... Go
Scroll/Zoom: Show 1.629 kbp ☑ Flip

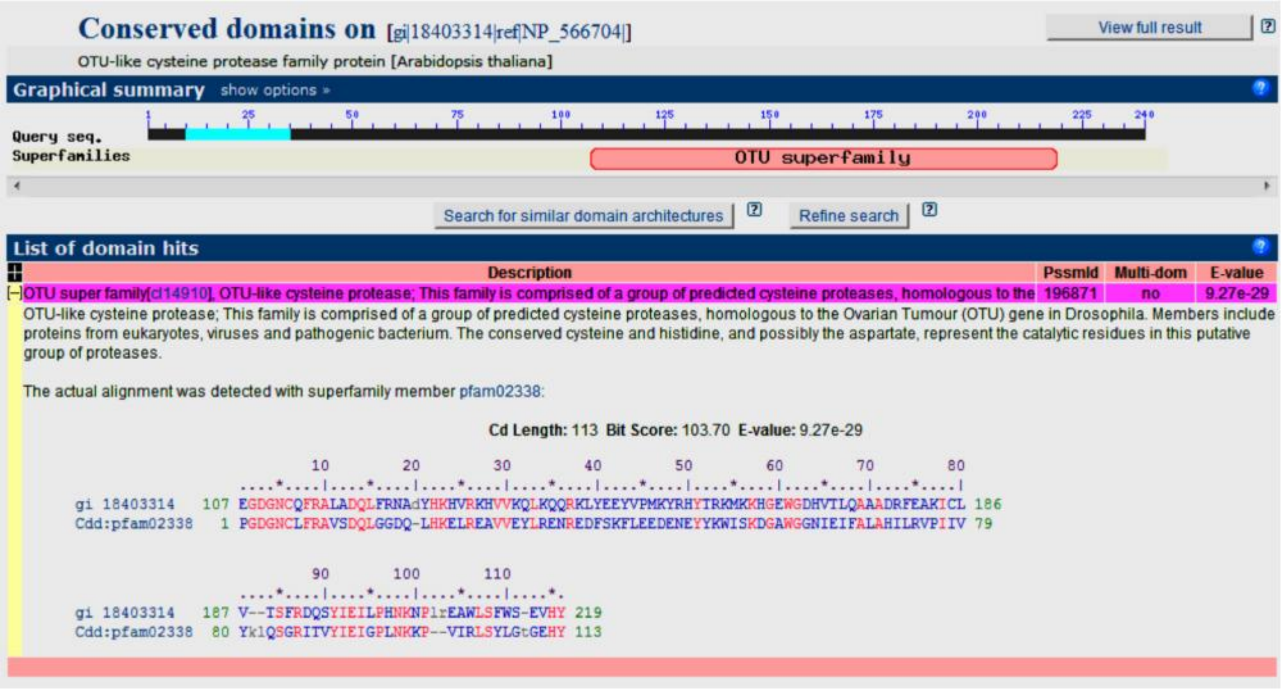
Overview
chr07
0M 10M 20M 30M 40M 50M
Region
chr07
39787k 39788k 39789k 39790k 39791k 39792k 39793k 39794k 39795k 39796k 39797k 39798k 39799k 39800k 39801k 39802k 39803k 39804k 39805k 39806k
Region - PGSC Loci
PGSC0003DMG402015928
Details
39797.2k 39797.1k 39797k 39796.9k 39796.8k 39796.7k 39796.6k 39796.5k 39796.4k 39796.3k 39796.2k 39796.1k 39796k 39795.9k 397
PGSC Pseudomolecule Tiling Path (v2.1.11)
PGSC0003DMB000000250
1-975681
PGSC Loci
PGSC0003DMG402015928
OTU-like cysteine protease family protein
PGSC Representative Gene Models
PGSC0003DMT400041164
D

Supplementary Fig. 3. The identified S-proteins A:ABU49725, B: AAR25790, C: AAR25793 and D:AAR25796 mapped on to the genome of *S. tuberosum* group Phureja. The images were generated from the genome browser of Solanaceae Genomics Resource (http://solanaceae.plantbiology.msu.edu/cgi-bin/gbrowse/potato)



```
>ref|NP_566704.1| UGM OTU-like cysteine protease family protein [Arabidopsis thaliana]
 gb|AEE76613.1| G OTU-like cysteine protease family protein [Arabidopsis thaliana]
Length=240

 GENE ID: 821796 AT3G22260 | OTU-like cysteine protease family protein
[Arabidopsis thaliana] (10 or fewer PubMed links)

 Score =  116 bits (291),  Expect = 1e-31, Method: Compositional matrix adjust.
 Identities = 60/80 (75%), Positives = 65/80 (81%), Gaps = 2/80 (3%)

Query  1    EDDHAIATILAEEEENARKYGGKLGKRLSHLDSIPHTPRVIGEIPDPNDATLDHGRLSSR  60
            +DD  IA ILAE+E   R+  GKLGKRLSHLDSIPHTPRV  EIPD NDATLDH  LS R
Sbjct  38   DDDQTIARILAEDESLRRE--GKLGKRLSHLDSIPHTPRVNREIPDINDATLDHELLSGR  95

Query  61   LATYGLFEMQIEGDGNCQFR  80
            LATYGL E+Q+EGDGNCQFR
Sbjct  96   LATYGLAELQMEGDGNCQFR  115
```

Supplementary Fig. 4. Pair wise alignment of query protein AAR25796 with subject protein NP566704.The last nine residues of AAR25796 have 100% identity with 107-115 residues of NP566704 that indicates the highlighted fragment as a part of OTU like cysteine protease domain



Conserved domains on [gi|18403314|ref|NP_566704|]

OTU-like cysteine protease family protein [Arabidopsis thaliana]

Description	PssmId	Multi-dom	E-value
OTU super family[cl14910], OTU-like cysteine protease; This family is comprised of a group of predicted cysteine proteases, homologous to the	196871	no	9.27e-29

OTU-like cysteine protease; This family is comprised of a group of predicted cysteine proteases, homologous to the Ovarian Tumour (OTU) gene in Drosophila. Members include proteins from eukaryotes, viruses and pathogenic bacterium. The conserved cysteine and histidine, and possibly the aspartate, represent the catalytic residues in this putative group of proteases.

The actual alignment was detected with superfamily member pfam02338:

Cd Length: 113 **Bit Score:** 103.70 **E-value:** 9.27e-29

```
                           10        20        30        40        50        60        70        80
                   ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
gi 18403314    107 EGDGNCQFRALADQLFRNAdYHKHVRKHVVKQLKQQRKLYEEYVPMKYRHYTRKMKKHGEWGDHVTLQAAADRFEAKICL 186
Cdd:pfam02338    1 PGDGNCLFRAVSDQLGGDQ-LHKELREAVVEYLRENREDFSKFLEEDENEYYKWISKDGAWGGNIEIFALAHILRVPIIV 79

                           90       100       110
                   ....*....|....*....|....*....|....*.
gi 18403314    187 V--TSFRDQSYIEILPHNKNPlrEAWLSFWS-EVHY 219
Cdd:pfam02338   80 YklQSGRITVYIEIGPLNKKP--VIRLSYLGtGEHY 113
```

Supplementary Fig. 5. CDD result of NP566704 showing OTU superfamily on the sequence from 107-219 residues