Plant Omics Journal POJ

POJ 9(4):248-260 (2016) ISSN: 1836-3644
DOI: 10.21475/poj.16.09.04.pne103

A computational study on genetic diversity of *shatterproof1* (*shp1*) and *shatterproof2* (*shp2*) genes in some members of *Oleraceae* and its molecular implications

Shachi Gahoi[1‡], Neetesh Pandey[2‡], B.V. Suresh[1], Monendra Grover[2]*, S.S. Marla[1] and Anil Rai[2]

[1]National Bureau of Plant Genetic Resources, New Delhi-110012, India
[2]Centre for Agricultural Bioinformatics, ICAR-IASRI, Library Avenue, Pusa, New Delhi -110012, India

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

‡These authors contributed equally to the paper as first authors.

Abstract

Dispersal and maturation of seed is a complex event in flowering plants. The genes *shatterproof1 (shp1)* and *shatterproof2 (shp2)* are essential for fruit dehiscence in Arabidopsis. In this study, we have analyzed the diversity in these two genes and their molecular implications in some members of *Oleraceae*. We have studied the gene organization of these two genes and various biochemical and biophysical parameters of the proteins encoded by these two genes. Though there are some similarities, there also exist some notable differences. These differences could be exploited for creating a library of synthetic alleles (neutral or advantageous) to be used for genetic engineering, thus ensuring a wide genetic base**.** This diversity analysis may be significant to create diversity in the transgenic plants for shattering resistance using genetic engineered methods. This analysis explores the possible correlation of results of this study with the phenotypic data to derive functional significance of the diversity in SHP genes.

Keywords: Shatterproof1 (SHP1), Shatterproof2 (SHP2), *Arabidopsis thaliana, Arabidopsis lyrata*, *Brassica napus*.
Abbreviations: *A. lyrata: Arabidopsis lyrata, A.thaliana: Arabidopsis thaliana,* MULTALIN: multiple sequence alignment, SHP1: Shatterproof1, SHP2: Shatterproof2.

Introduction

The variety of genes within a species is described as genetic diversity. The variety of different alleles must be conserved to conserve the genetic diversity. This is indispensable for adaptability to environmental changes and is therefore vital to species survival. More variation ensures that there are better chances of at least some individuals having some allelic variant that is suited to extreme weather conditions. These individuals may produce offspring with the variant that will be carried throughout successive generations.

Loss of genetic diversity in agriculture is dangerously leading to extinction. We must preserve the complex interrelationships that hold the natural world together. Reducing the diversity of life, narrows our options for the future and raises question on our own survival. Though we aim to be grain-rich, we should not be gene-poor. The genetic engineering of crops should also aim to create gene rich base. Thus it is important to have a library of alleles (neutral or advantageous) which can be used for genetic transformation and molecular breeding. It is important to characterize these genes and proteins coded by them in this context. Some of the properties from the library of genes could be used to construct novel gene constructs (Butler and Reichhardt 1999). Crops can be enhanced in a surprising number of ways, and alteration of genes to improve yield is just one. Twenty to fifty percent of a rapeseed crop harvested and crushed to yield canola oil, can be lost because the pods exposed and release the seeds before the farmer can harvest them. But when two nifty genes called *shatterproof1 (shp1)* and *shatterproof2 (shp2)* are mutated, the seed pods fail to shatter, or spurt. The Yanofsky team reports the discovery of two weakened *shatterproof* genes in *Arabidopsis thaliana*, a minute flowering weed geneticists study to isolate genes imperative in plant development. Genes that control pod shattering in *Arabidopsis* are likely to be present in close relatives like cauliflower, brussels sprouts, broccoli, peas, soya beans and other important food crops. Discovery of weak versions of *shatterproof* and inserting them into rapeseed strength increase the production of crop per hectare, making the land more productive and reducing the amount of water, pesticides and fertilizer a farmer needs. It might be one of the more immediate applications of the team's discovery (Liljegren et al. 2000).

In this study, we have analyzed the diversity in *SHP1* and *SHP2* genes and their molecular implications. This could be used to create wider secondary or tertiary genetic pools.

A coherent theory of neutral evolution was proposed by Motoo Kimura in 1968 (Kimura 1968) and by King and Jukes independently in 1969. Researcher suggested that the vast majority of molecular differences are selectively "neutral" when the genomes of existing species are compared (King and Jukes 1969). The neutral theory of molecular evolution further holds that at the molecular level most evolutionary changes and most of the variation within and between species is caused by genetic drift of mutant alleles that are neutral and not by natural selection. A neutral mutation is defined as a mutation that does not affect an organism's survival or reproduction abilities. The neutral theory allows for the possibility that most mutations are

deleterious but they do not make significant contributions to variation and are rapidly purged by natural selection. Other mutations (not deleterious) are assumed to be mostly neutral. In addition to assuming the primacy of neutral mutations, the theory also assumes that the fate of neutral mutations is determined by the sampling processes described by specific models of random genetic drift (Kimura 1983). This theory is proposed to be applicable at molecular level and not at phenotypic level.

This concept of neutral mutations is primarily based on the degenerate genetic code, in which the third position of the codon may differ and yet encode the same amino acid. As a result many potential single-nucleotide changes are in effect synonymous in other words "silent" or "unexpressed". Essentially such changes are presumed to have little or no biological effect. A second hypothesis of the neutral theory is that the genetic drift acting on neutral alleles is the cause of most evolutionary changes is the result of genetic drift acting on neutral alleles. After appearing by mutation, a neutral allele may become more prevalent within the population by genetic drift. In rare cases it may become fixed, in the population. It is notable however that neutral theory does not deny the occurrence of natural selection.

In this study we have tried to characterize some of the diversity in the gene pool of *Oleraceae* in context of SHP1 and SHP2 genes. The alleles used in this study may be neutral or advantageous. This important study lays groundwork for correlation of molecular diversity with the phenotypic diversity and better understanding of gene pool available with respect to SHP1 and SHP2 genes in *Oleraceae*. The results used in this study could further be used for rational genetic manipulation of important crop plants.

Results and Discussion

The vast genetic diversity in plants is an important resource. The diversity at molecular level is particularly important and can be used to rationally engineer crop plants. In this paper we have characterized the diversity in the SHP1 and SHP2 at gene and protein level.

Gene Organization

The organization of the gene has important functional implications**.** The exon/intron organization of genes from the plants which were selected for protein analysis was determined using FGENESH (Solovyev et al. 2006). (http://www.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind). The results are shown in Table 4 and 5. There is only a single exon in all the SHP1 and SHP2 genes studied except the *Brassica napus* which has 7 exons.

Protein analysis

For protein analysis NCBI protein blast was performed with SHP1and SHP2 proteins from *Arabidopsis thaliana*. The tabulated proteins were used in this study from obtained hits.

Domain analysis

The search for domains in the listed proteins in the NCBI database revealed that MADS box and K boxes are present in the proteins studied. The distribution of MADS boxes and K boxes in the SHP1 and SHP2 proteins from different plants is shown in Table 6 and Table 7. MADS box is present in all the analyzed SHP1 proteins. Similar distribution of MADS box domain was observed in the SHP2 proteins. The positions of MADS box domain in SHP1 and SHP2 proteins were broadly conserved in the analyzed proteins which may have important functional implications. MADS box is a DNA-binding domain found in many eukaryotic regulatory proteins: such as MCM1, the regulator of cell type-specific genes in fission yeast; DSRF, a trachea development factor found in Drosophila ; the MEF2 family of myocyte-specific enhancer factors; and the Agamous and Deficiens families of homeotic proteins found in plants. Proteins belonging to the MADS family function as dimers. The primary DNA-binding element of these proteins is an anti-parallel coiled coil of two amphipathic alpha-helices. One alpha helix each is contributed by each subunit and the basic N-termini of the helices fit into the DNA major groove. The chain extending from the helix N-termini penetrates into the minor groove. The MADS-box domain is commonly found associated with K-box region (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR002100#PUB00000821).

The K box domain was found to be present in SHP1 and SHP2 genes. The K box domain was found in SHP1 protein in all the plants in all the studied SHP1 proteins. The K-box was also found to be present in all the analyzed SHP2 proteins. The position of K-box in SHP1 and SHP2 proteins was broadly conserved in all the proteins studied. The above mentioned fact may be functionally relevant. The K-box region is frequently found associated with SRF-type transcription factors. The K-box is a possible coiled-coil structure and has a possible role in multimer formation (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR002487).

Multiple sequence alignment

The SHP1 and SHP2 proteins were subjected to multiple sequence alignment (Figure 1 and 2) using Mulatalin. The positional variations in these genes from different plants have been listed in Tables 8 and 9 respectively.

Secondary structure analysis

The secondary structure analysis in the listed SHP1 and SHP2 proteins was performed using GOR IV tool. These results are shown in Figure 3, 4, 5 and 6. The positional variations in secondary structure from SHP1 and SHP2 proteins have been tabulated in Tables 10 and 11. The conspicuous differences have been highlighted.

Analysis of physico-chemical properties

The analysis of physico-chemical properties of SHP1 and SHP2 proteins was done using Protparam. The Protparam analysis (Table 12) shows that the SHP1 and SHP2 proteins studied in the plants in this study are unstable. The 'stability–function hypothesis' has been proposed by (Meiering et al. 1992) and further developed by (Schreiber et al. 1994) and (Shoichet et al. 1995). This theory states that protein stability is not maximized; in contrast, there is a balance between stability and function. Residues which contribute to ligand binding or catalytic activity may be suboptimal for stability. Thus it is probable that during the evolution of SHP1 and SHP2 proteins the stability has been traded off for functional residues.

The aliphatic index of a protein is defined as the relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine). It may be regarded as a positive

Table 1. NCBI BLAST was performed with SHP1 protein from *Arabidopsis thaliana*. From obtained hits SHP1 and SHP1 like proteins from plants were selected for further studies.

S.N.	Protein Accession No	Species	E-value
1	NP_191437.1	*Arabidopsis thaliana*	9e-143
2	AAU82055.1	*Arabidopsis lyrata*	5e-129
3	AAK00646.1	*Brassica napus*	2.00E-131
4	ACD76827.1	*Capsella bursa-pastoris*	1.00E-134

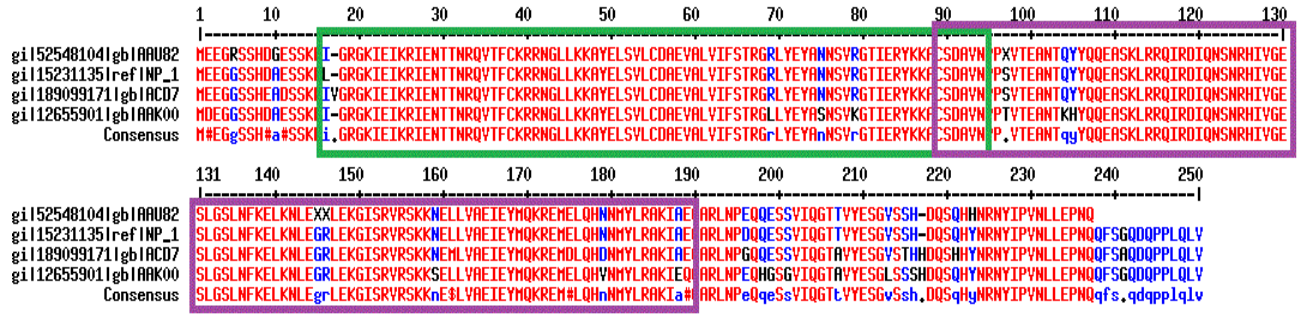

Fig 1. MultAlin alignment of SHP1 protein from Arabidopsis *lyrata*, *Arabidopsis thaliana*, *Brassica napus* and *Capsella bursa-pastoris*. MADS-Box domain is highlighted green rectangle and K-Box domain is highlighted by purple rectangle.

Table 2. NCBI BLAST was performed with SHP2 protein from Arabidopsis thaliana. From obtained hits SHP2 and SHP2 like proteins from plants were selected for further studies.

S.N.	Protein Accession No	Species	E-value
1	NP_565986.1	*Arabidopsis thaliana*	2.00E-140
2	AAU82080.1	*Arabidopsis lyrata*	2.00E-127
3	ACA42768.1	*Brassica napus*	2.00E-127
4	ACD76825.1	*Capsella bursa-pastoris*	1.00E-114

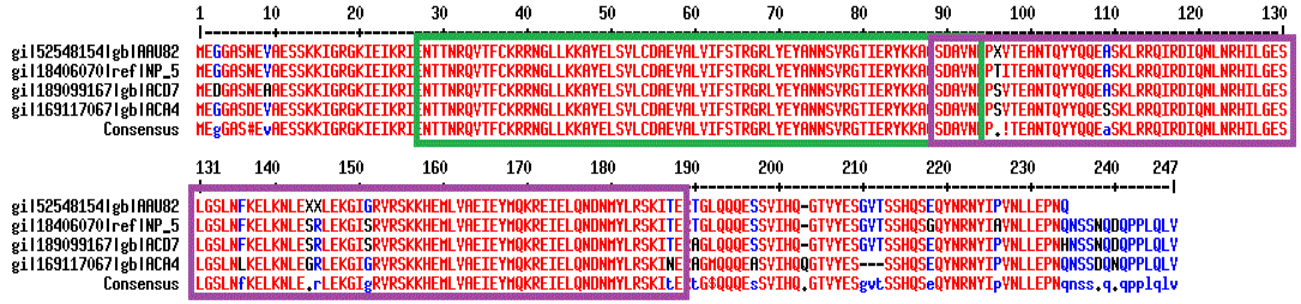

Fig 2. MultAlin alignment of SHP2 protein from *Arabidopsis lyrata*, *Arabidopsis thaliana*, *Brassica napus and Capsella bursa-pastoris*. MADS-Box domain is highlighted by enclosing in green rectangle and K-Box domain is highlighted by purple rectangle.

Table 3. Tools used in this study.

S.No.	Analysis	Tool (URL address)
1	Determination of Exon/Intron structure	FGENESH. (http://www.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind)
2	Selection of Proteins for analysis	BlastP (https://blast.ncbi.nlm.nih.gov/Blast.cgi)
3	Domain Analysis	NCBI (http://www.ncbi.nlm.nih.gov/)
4	Multiple sequence alignment	Multalin (http://multalin.toulouse.inra.fr/multalin)
5	Secondary Structure prediction	GOR (https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor4.html)
6	Analysis of Physico-chemical properties	ProtParam (http://expasy.org/tools/protparam.html)
7	De novo detection of repeats	RADAR (http://www.ebi.ac.uk/Tools/Radar/index.html)
8	Detection of PEST motifs	ePEST (http://emboss.bioinformatics.nl/cgi-bin/emboss/epestfind)
9	Phosphorylation sites prediction	NetPhos (www.cbs.dtu.dk/services/NetPhos)
10	Prediction of Kinase specific phosphorylation sites	NetPhosK (www.cbs.dtu.dk/services/NetPhosK)
11	Protein-protein interactions	STRING database (http://string-db.org/)

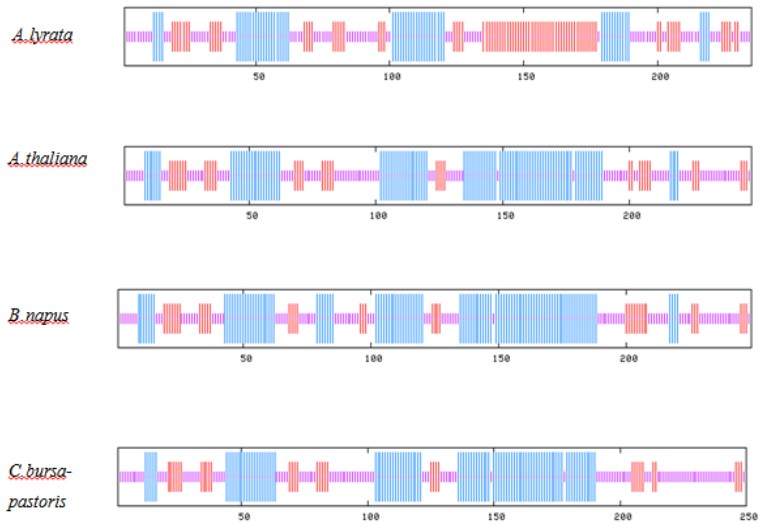

Fig 3. Diagrammatic representation of Secondary structures of SHP1 protein from different species.

Table 4. Predictedexonic regions for *shp1* gene by FGENESH.

Gene	Length (bp)	No. of Exons	Exonic Region
Capsella bursa-pastoris	1024	1	78-830 (753bp)
Arabidopsis lyrata	1182	1	327-1073 (747bp)
Arabidopsis thaliana	1209	1	322-1068 (747bp)
	3549	7	247-273 (227bp)
			1625-1706 (82bp)
			1980-2041 (62bp)
			2395-2494 (100bp)
			2546-2617 (72bp)
			2690-2731 (42bp)
			2808-3002 (195bp)

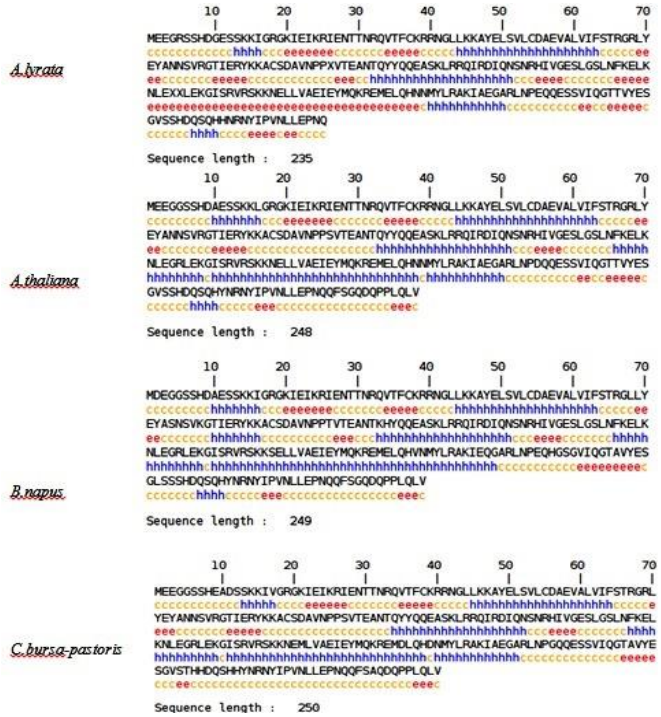

Fig 4. Secondary structure details of SHP1 proteins. The helixes (Hh) are shows by blue, extended strand (Ee) are shows by red and random coil (Cc) is shows by yellow color.

Table 5. Predicted exonic regions for *shp2* gene by FGENESH.

Gene	Length (bp)	No. of exons	Exonic Region
Capsella bursa-pastoris	995	1	97-837(741bp)
Arabidopsis thaliana	1135	1	188-928(741bp)
Brassica napus	840	1	12-746(735bp)

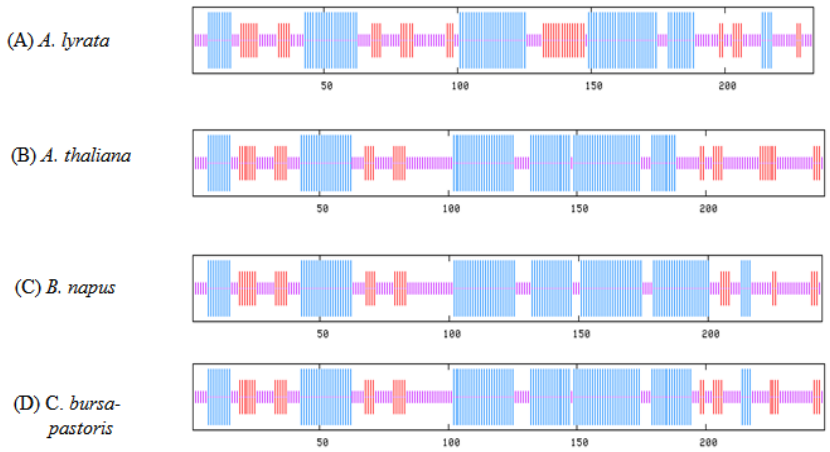

Fig 5. Diagrammatic representation of Secondary structures of SHP2 protein from different species.

Table 6. MADS domain region (from NCBI).

MADS-Box	*A. lyrata*	*A. thaliana*	*B. napus*	*C. bursa-pastoris*
SHP1	28 – 93	17 - 93	28 – 93	18 – 94
SHP2	28 – 93	28 - 93	28 – 93	28 – 93

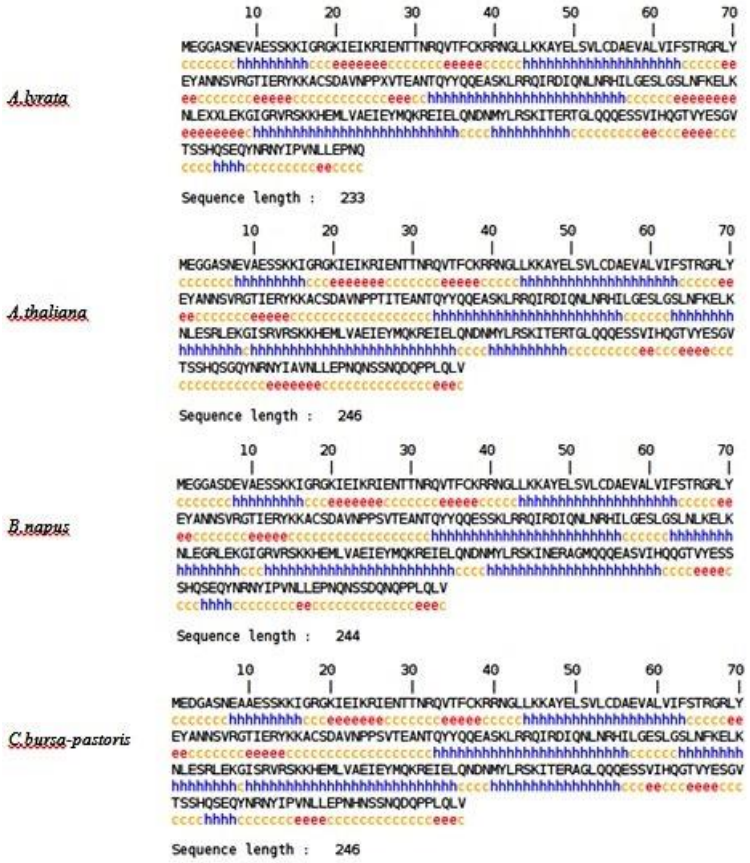

Fig 6. Secondary structure details of SHP2 proteins. The helixes (Hh) are shows by blue, extended strand (Ee) are shows by red and random coil (Cc) is shows by yellow color.

Table 7. K-Box domain region (from NCBI).

K-Box	*A. lyrata*	*A. thaliana*	*B. napus*	*C. bursa-pastoris*
SHP1	89 – 188	89 - 188	97 – 187	90 – 189
SHP2	89 – 188	97 - 188	89 – 188	89 – 188

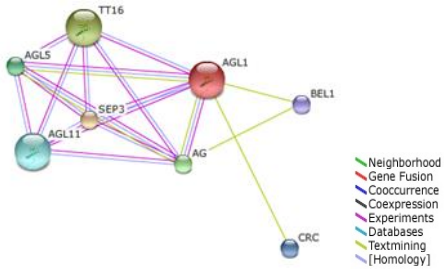

Fig 7. Interacting partners of SHP1 proteins from STRING database. Different line colors represent the types of evidence for the association. SHP1 is represented by AGL1.

factor for the increase of thermostability of proteins (Atsushi 1980)**.** The aliphatic index of the studied proteins is broadly similar as demonstrated by Protparam analysis.

De novo detection of repeats

De novo detection of repeats in SHP1 and SHP2 protein sequences was done using RADAR. Of all the proteins studied, *Arabidopsis thaliana*, *Arabidopsis lyrata* and *Brassica napus* show 1 set of repeats (Table 13). The other proteins studied have no repeats. Many large proteins have evolved by internal duplication and many internal sequence repeats may correspond to functional and structural units (Heger and Holm 2000).

Detection of PEST motifs

The detection of PEST motifs in SHP1 and SHP2 was done using ePESTFind. PEST motifs reduce the half-lives of proteins dramatically and hence, that they target proteins for proteolytic degradation The Pest motifs share high local concentrations of amino acids proline (P), glutamic acid (E), serine (S), threonine (T) and to a lesser extent aspartic acid (D) (Rechsteiner et al. 1987; Rechsteiner and Rogers 1996). The ePEST analysis (Table 14) shows that all the proteins studied have PEST motifs and most of the proteins studied have two motifs. However the length of motifs is highly variable.

Prediction of O-GlcNAC sites

The Prediction of O-GlcNAc sites in SHP1 and SHP2 was done using (YinOYang 1.2 www server) (Gupta 2001) Table 15 displays predictions for O-ß-GlcNAc attachment sites in the studied proteins. The addition of a carbohydrate moeity to the side chain of a residue in a protein chain influences the physicochemical properties of the protein. Glycosylation is known to alter proteolytic resistance, protein solubility, stability, and local structure. In plant cells, protein *N*-glycosylation surprises in the ER by the co- or post-translational transmission of an oligosaccharide precursor, $Glc_3Man_9GlcNAc_2$, from a dolichol lipid carrier onto exact Asn residues constitutive of the *N*-glycosylation consensus sequence Asn–X–Ser/Thr (X is any amino acid excluding Pro). The number of modified residues is variable in the proteins studied (Gomord et al. 2010; Hounsell et al. 1996; Lis and Sharon 1993).

Prediction of phosphorylation sites

Protein phosphorylation is a post-translational modification of proteins in which a serine, a threonine or a tyrosine residue is phosphorylated by a protein kinase by the addition of a covalently bound phosphate group. Protein phosphorylation at serine, threonine or tyrosine residues affects a multitude of cellular signaling processes. Post-translational modifications modulate the activity (Blom et al. 1999)**.** The analysis using Net Phos (Table 16) shows that all the studied proteins are potentially phosphorylated. However the sites of phosphorylation vary among the studied proteins.

Prediction of kinase specific phosphorylation sites

The prediction of kinase specific phosphorylation sites was done using NetPhosK. Table 17 shows kinase specific phosphorylation sites in studied proteins. It is shown that all the prtoeins studied are potentially phosphorylated by PKCs. Protein kinase Cs also known as PKCs are a family of enzymes that are involved in regulating the function of other proteins through the phosphorylation of serine and threonine amino acid residues on these proteins. PKC enzymes in turn are activated by signals such as increases in the concentration of diacylglycerol or Ca^{2+} (MELLOR and PARKER 1998). All the proteins are phosphorylated at same residue by Protein kinase C highlighting the possible importance of Protein Kinase signaling in the pathways involving SHP proteins.

Prediction of sumoylation sites

The Prediction of Sumoylation sites in SHP1 and SHP2

Table 8. Positional variation in multiple sequence alignment (MULTALIN) in SHP1 protein from different species.

S.N.	*Position*	*Arabidopsis lyrata*	*Arabidopsis thaliana*	*Capsella bursa-partoris*	*Brassica napus*
1	*10th*	*G*	*A*	*A*	*A*
2	*16th*	*I*	*L*	*I*	*I*
3	*17th*	-	-	*V*	-
4	*69th*	*R*	*R*	*R*	*L*
5	*79th*	*R*	*R*	*R*	*K*
6	*96th*	*X*	*S*	*S*	*T*
7	*104th*	*Q*	*Q*	*Q*	*K*
8	*105th*	*Y*	*Y*	*Y*	*H*
9	*145th*	*X*	*R*	*R*	*R*
10	*150th*	*X*	*R*	*R*	*R*
11	*159th*	*N*	*N*	*N*	*S*
12	*179th*	*N*	*N*	*D*	*V*
13	*188th*	*A*	*A*	*A*	*E*
14	*196th*	*E*	*D*	*G*	*E*
15	*198th*	*Q*	*Q*	*Q*	*H*
16	*199th*	*E*	*E*	*E*	*G*
17	*207th*	*T*	*T*	*A*	*A*
18	*213th*	*V*	*V*	*V*	*L*
19	*215th*	*S*	*S*	*T*	*S*
20	*216th*	*H*	*H*	*H*	*S*
21	*217th*	-	-	*H*	*H*
22	*221th*	*Q*	*Q*	*Q*	*H*
23	*223th*	*H*	*Y*	*Y*	*Y*
24	*241th*	-	*G*	*A*	*G*

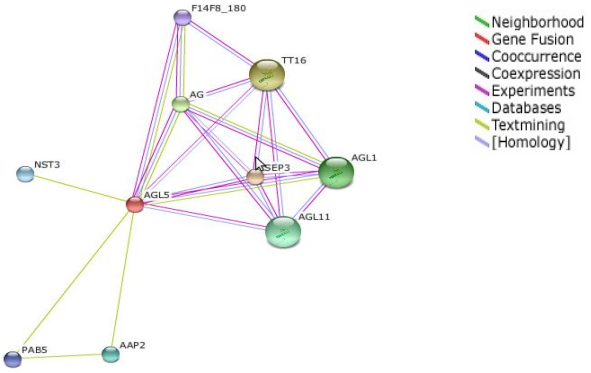

Fig 8. Interacting partners of SHP2 proteins from STRING database. SHP2 is represented by AGL5.

Table 9. Positional variation in multiple sequence alignment (MULTALIN) in SHP2 protein from different species.

S.N.	Posiotion	*Arabidopsis lyrata*	*Arabidopsis thaliana*	*Capsella bursa-partoris*	*Brassica napus*
1	3rd	G	G	D	G
2	9th	V	V	A	V
3	96th	X	T	S	S
4	109th	A	A	A	S
5	136th	F	F	F	L
6	144th	X	S	S	G
7	151th	G	S	S	G
8	187th	T	T	T	N
9	190th	T	T	A	A
10	197th	S	S	S	A
11	203th	-	-	-	Q
12	210th	G	G	G	-
13	211th	V	V	V	-
14	212th	T	T	T	-
15	218th	E	G	E	E
16	226th	P	A	P	P
17	234th	Q	Q	H	Q
18	238th	-	N	N	D
19	240th	-	D	D	N

Table 10. Positional variations in Secondary structure of SHP1 from different species.

A. lyrata	*A. thaliana*	*B. napus*	*C. bursa-partoris*
1-12 (Coil)	1-9 (Coil)	1-9 (Coil)	1-12 (Coil)
13-16 (Helix)	10-16 (Helix)	10-16 (Helix)	13-17 (Helix)
17-19 (Coil)	17-19 (Coil)	17-19 (Coil)	18-21 (Coil)
20-26 (Sheet)	20-26 (Sheet)	20-26 (Sheet)	22-27 (Sheet)
27-33 (Coil)	27-33 (Coil)	27-33 (Coil)	28-34 (Coil)
34-38 (Sheet)	34-38 (Sheet)	34-38 (Sheet)	35-39 (Sheet)
39-43 (Coil)	39-43 (Coil)	39-43 (Coil)	40-44 (Coil)
44-63 (Helix)	44-63 (Helix)	44-63 (Helix)	45-64 (Helix)
64-68 (Coil)	64-68 (Coil)	64-68 (Coil)	65-69 (Coil)
69-72 (Sheet)	69-72 (Sheet)	69-72 (Sheet)	70-73 (Sheet)
73-79 (Coil)	73-79 (Coil)	73-79 (Coil)	74-80 (Coil)
80-84 (Sheet)	80-84 (Sheet)	80-86 (Helix)	81-85 (Sheet)
85-96 (Coil)	85-102 (Coil)	87-96 (Coil)	86-103 (Coil)
97-99 (Sheet)	-	97-99 (Sheet)	-
100-101 (Coil)	-	100-102 (Coil)	-
102-121 (Helix)	103-121 (Helix)	103-121 (Helix)	104-122 (Helix)
122-124 (Coil)	122-124 (Coil)	122-124 (Coil)	123-125 (Coil)
125-128 (Sheet)	125-128 (Sheet)	125-128 (Sheet)	126-129 (Sheet)
129-135 (Coil)	129-135 (Coil)	129-135 (Coil)	130-136 (Coil)
136-178 (Sheet)	136-148 (Helix)	136-148 (Helix)	137-149 (Helix)
	149-149 (Coil)	149-149 (Coil)	150-150 (Coil)
	150-178 (Helix)	150-189 (Helix)	151-178 (Helix)
179-179 (Coil)	179-179 (Coil)	-	179-179 (Coil)
180-190 (Helix)	180-190 (Helix)	-	180-191 (Helix)
191-200 (Coil)	191-200 (Coil)	190-200 (Coil)	192-205 (Coil)
201-202 (Sheet)	201-202 (Sheet)	201-209 (Sheet)	206-210 (Sheet)
203-204 (Coil)	203-204 (Coil)	-	-
205-209 (Sheet)	205-209 (Sheet)	-	-
210-216 (Coil)	210-216 (Coil)	210-217 (Coil)	211-213 (Coil)
217-220 (Helix)	217-220 (Helix)	218-221 (Helix)	214-215 (Sheet)
221-224 (Coil)	221-225 (Coil)	222-226 (Coil)	216-246 (Coil)
225-228 (Sheet)	226-228 (Sheet)	227-229 (Sheet)	247-249 (Sheet)
229-229 (Coil)	229-244 (Coil)	230-245 (Coil)	250-250 (Coil)
230-231(Sheet)	245-247 (Sheet)	246-248 (Sheet)	-
232-235 (Coil)	248-248 (Coil)	249-249 (Coil)	-

Table 11. Positional variations in Secondary structure of SHP2 from different species.

A.lyrata	*A.thaliana*	*B.napus*	*C.bursa-partoris*
1-7 (Coil)	1-7 (Coil)	1-7 (Coil)	1-7 (Coil)
8-16 (Helix)	8-16 (Helix)	8-16 (Helix)	8-16 (Helix)
17-19 (Coil)	17-19 (Coil)	17-19 (Coil)	17-19 (Coil)
20-26 (Sheet)	20-26 (Sheet)	20-26 (Sheet)	20-26 (Sheet)
27-33 (Coil)	27-33 (Coil)	27-33 (Coil)	27-33 (Coil)

34-38 (Sheet)	34-38 (Sheet)	34-38 (Sheet)	34-38 (Sheet)
39-43 (Coil)	39-43 (Coil)	39-43 (Coil)	39-43 (Coil)
44-63 (Helix)	44-63 (Helix)	44-63 (Helix)	44-63 (Helix)
64-68 (Coil)	64-68 (Coil)	64-68 (Coil)	64-68 (Coil)
69-72 (Sheet)	69-72 (Sheet)	69-72 (Sheet)	69-72 (Sheet)
73-79 (Coil)	73-79 (Coil)	73-79 (Coil)	73-79 (Coil)
80-84 (Sheet)	80-84 (Sheet)	80-84 (Sheet)	80-84 (Sheet)
85-96 (Coil)	85-102 (Coil)	85-102 (Coil)	85-102 (Coil)
97-99 (Sheet)	-	-	-
100-101 (Coil)	-	-	-
102-126 (Helix)	103-126 (Helix)	103-126 (Helix)	103-126 (Helix)
127-132 (Coil)	127-132 (Coil)	127-132 (Coil)	127-132 (Coil)
133-148 (Sheet)	133-148 (Helix)	133-148 (Helix)	133-148 (Helix)
149-149 (Coil)	149-149 (Coil)	149-151 (Coil)	149-149 (Coil)
150-175 (Helix)	150-175 (Helix)	152-175 (Helix)	150-175 (Helix)
176-179 (Coil)	176-179 (Coil)	176-179 (Coil)	176-179 (Coil)
180-189 (Helix)	180-189 (Helix)	180-201 (Helix)	180-195 (Helix)
190-198 (Coil)	190-198 (Coil)	-	196-198 (Coil)
199-200 (Sheet)	199-200 (Sheet)	-	199-200 (Sheet)
201-203 (Coil)	201-203 (Coil)	202-205 (Coil)	201-203 (Coil)
204-207 (Sheet)	204-207 (Sheet)	206-209 (Sheet)	204-207(Sheet)
A.lyrata	A.thaliana	B.napus	C.bursa-partoris
208-214 (Coil)	208-221 (Coil)	210-213 (Coil)	208-214 (Coil)
215-218 (Helix)	-	214-217 (Helix)	215-218 (Helix)
219-227 (Coil)	222-228 (Sheet)	218-225 (Coil)	219-225(Coil)
228-229 (Sheet)	229-242 (Coil)	226-227 (Sheet)	226-229 (Sheet)
230-233 (Coil)	243-245 (Sheet)	228-240 (Coil)	230-242 (Coil)
	246-246 (Coil)	241-243 (Sheet)	243-245 (Sheet)
		244-244 (Coil)	246-246 (Coil)

Table 12. Analysis of physico-chemical properties of SHP1 and SHP2 proteins.

Description	SHP1 *Arabidopsis thaliana*	SHP1 *Arabidopsis lyrata*	SHP1 *Brassica napus*	SHP1 *Capsella bursa-pastoris*	SHP2 *Arabidopsis thaliana*	SHP2 *Arabidopsis lyrata*	SHP2 *Brassica napus*	SHP2 *Capsella bursa-pastoris*
Instability Index:	54.53 (unstable)	54.53 (unstable)	56.28 (unstable)	51.79 (unstable)	58.68 (unstable)	56.85 (unstable)	60.64 (unstable)	62.80 (unstable)
Aliphatic Index:	77.15	77.15	80.64	78.76	82.14	80.73	80.7	80.45

Table 13. De novo detection of repeats in SHP1 and SHP2 protein sequences.

Protein	No of repeats	Total score	Length	From	To	Sequence
SHP1 *Arabidopsis thaliana*	1 set	80.89	24	2	27	EEGRSSHDgeSSKKIGRGKIEIKRIE
				209	232	ESGVSSHD..QSQHHNRNYIPVNLLE
SHP1 *Arabidopsis lyrata*	1 set	80.89	24	2	27	EEGRSSHDgeSSKKIGRGKIEIKRIE
				209	232	ESGVSSHD..QSQHHNRNYIPVNLLE
SHP1	1 set	77.57	24	6	33	SSHDaeSSKKIGRGKIEIKRIEntTNRQ

Brassica napus						
				214	237	SSHD..QSQHYNRNYIPVNLLE..PNQQ
SHP1 *Capsella bursa-pastoris*	No repeats	-----	----	-----	-----	----------
SHP2 *Arabidopsis thaliana*	No repeats	-----	-----	-----	-----	----------
SHP2 *Arabidopsis lyrata*	No Repeats	-----	----	-----	----	----------
SHP2 *Brassica napus*	No repeats	-----	----	-----	-----	----------
SHP2 *Capsella bursa-pastoris*	No repeats	------	-----	------	----	----------

Table 14. Detection of PEST motifs in SHP1 and SHP2 proteins.

Protein	No of motifs	Motif Number	Score	Position	Length	Sequence
SHP1 *Arabidopsis thaliana*	2	1	-5.63	86-111	24	KACSDAVNPPXVTEANTQYYQQEASK
		2	-1.46	191-215	23	RLNPEQQESSVIQGTTVYESGVSSH
SHP1 *Arabidopsis lyrata*	2	1	-5.63	86-111	24	KACSDAVNPPXVTEANTQYYQQEASK
		2	-1.46	191-225	24	RLNPEQQESSVIQGTTVYESGVSSH
SHP1 *Brassica napus*	2	1	-12.43	224-249	25	RNYIPVNLLEPNQQFSGQDQPPLQLV
		2	-2.47	86-103	16	KACSDAVNPPTVTEANTK
SHP1 *Capsella bursa-pastoris*	2	1	-12.74	225-250	25	RNYIPVNLLEPNQQFSAQDQPPLQLV
		2	-3.23	87-112	24	KACSDAVNPPSVTEANTQYYQQEASK
		3	-4.55	192-216	23	RLNPGQQESSVIQGTTVYESGVSTH
SHP2 *Arabidopsis thaliana*	2	1	-11.54	223-248	25	RNYIAVNLLEPNQNSSNQDQPPLQLV
		2	-3.33	86-111	24	KACSDAVNPPTITEANTQYYQQEASK
SHP2 *Arabidopsis lyrata*	1	1	-5.63	86-111	24	KACSDAVNPPXVTEANTQYYQQEASK
SHP2 *Brassica napus*	2	1	-9.24	219-244	25	RNYIPVNLLEPNQNSSDQNQPPLQLV
		2	-1.08	86-111	24	KACSDAVNPPSVTEANTQYYQQESSK
SHP2 *Capsella bursa-pastoris*	2	1	-3.23	86-111	24	KACSDAVNPPSVTEANTQYYQQEASK
		2	-10.27	233-246	13	HNSSNQDQPPLQLV

Table 15. Prediction of O-GlcNAc sites in SHP1 and SHP2 proteins.

	SHP1 *Arabidopsis thaliana*	SHP1 *Arabidopsis lyrata*	SHP1 *Brassica napus*	SHP1 *Capsella bursa-pastoris*	SHP2 *Arabidopsis thaliana*	SHP2 *Arabidopsis lyrata*	SHP2 *Brassica napus*	SHP2 *Capsella bursa-pastoris*
No. of positions:	7	7	5	6	7	6	9	8
Positions:	6, 30, 98, 206, 210, 213, 214	6, 30, 98, 206, 210, 213, 214	6, 30, 98, 210, 215	6, 31, 90, 99, 207, 214	30, 98, 199, 210, 213, 214, 237	30, 98, 197, 208, 211, 212	30, 89, 98, 102, 109, 110, 205, 210, 233	30, 89, 98, 197, 208, 211, 212, 235
Residues:	S T T T S S S	STTTSSSS	S T T S S	S T S T T S	T T S S T S S	TTSSTS	T S T T S S T S S	T S T S S T S S

Table 16. Prediction of phosphorylation sites in SHP1 and SHP2 proteins.

	SHP1 *Arabidopsis thaliana*	SHP1 *Arabidopsis lyrata*	SHP1 *Brassica napus*	SHP1 *Capsella bursa-pastoris*	SHP2 *Arabidopsis thaliana*	SHP2 *Arabidopsis lyrata*	SHP2 *Brassica napus*	SHP2 *Capsella bursa-pastoris*
Predicted Positions:	Ser: 14 Thr: 5 Tyr: 2	Ser: 13 Thr: 4 Tyr: 2	Ser: 13 Thr: 2 Tyr: 1	Ser: 13 Thr: 5 Tyr: 2	Ser: 10 Thr: 6 Tyr: 5	Ser: 7 Thr: 4 Tyr: 4	Ser: 10 Thr: 4 Tyr: 5	Ser: 12 Thr: 6 Tyr: 3

Table 17. The prediction of kinase specific phosphorylation sites in SHP1 and SHP2 proteins.

	SHP1 *Arabidopsis thaliana*	SHP1 *Arabidopsis lyrata*	SHP1 *Brassica napus*	SHP1 *Capsella bursa-pastoris*	SHP2 *Arabidopsis thaliana*	SHP2 *Arabidopsis lyrata*	SHP2 *Brassica napus*	SHP2 *Capsella bursa-pastoris*
Site:	S-12	S-12	S-12	S-12	S-12	S-12	S-12	S-12
Kinase:	PKC	PKC	PKC	PKC	PKC	PKC	PKC	PKC

Table 18. Prediction of Sumoylation sites in SHP1 and SHP2 proteins.

Protein	No. of Positions	Positions	Peptides	Type
SHP1 *Arabidopsis thaliana*	2	157 170	RSKKNEL VMQKREM	Non-Consensus type
SHP1 *Arabidopsis lyrata*	2	157 170	RSKKNEL YMQKREM	Type II- Non consensus
SHP1 *Brassica napus*	3	157 170 185	RSKKSEL YMQKREM LRAKIEQ	DO DO Type-I
SHP1 *Capsella bursa-pastoris*	4	158 171 408 421	RSKKNEM YMQKREM RSKKNEM YMQKREM	DO Do Do Do
SHP2 *Arabidopsis thaliana*	1	157	RSKKHEM	DO
SHP2 *Arabidopsis lyrata*	2	157 170	RSKKHEM YMQKREI	Type II- Non consensus
SHP2 *Brassica napus*	2	157 170	RSKKHEM YMQKREI	Non-consensus DO
SHP2 *Capsella bursa-pastoris*	2	157 170	RSKKHEM YMQKREI	DO DO

Table 19. Protein interacting partners for SHP1 protein from different species. Interacting partners are same for all species (*A.thaliana, A.lyrata, B.napus, C.bursa-pastoris*).

S. N.	Protein name	Description	Protein Length	STRING score
1	SEP_3	Developmental protein SEPALLATA 3. Probable transcription factor active in inflorescence development and floral organogenesis.	251	0.998
2	TT16	Protein TRANSPARENT TESTA 16. Transcription factor involved in the developmental regulation of the endothelium and in the accumulation of proanthocyanidins (PAs)	252	0.986
3	AG	Floral homeotic protein AGAMOUS; Probable transcription factor involved in the control of organ identity during the early development of flowers.	252	0.882
4	AGL5	Agamous-like MADS-box protein AGL5; Probable transcription factor	248	0.880
5	AGL11	Agamous-like MADS-box protein AGL5; Probable transcription factor	256	0.88
6	CRC	Transcription factor CRC; Transcription factor required for the initiation of nectary development.	181	0.45
7	BEL1	Homeobox protein BEL1 homolog; Plays a major role in ovule patterning and in determination of integument identity via its interaction with MADS-box factors.	611	0.43

Table 20. Protein interacting partners for SHP2 protein from different species. Interacting partners are same for all species (*A.thaliana, A.lyrata, B.napus, C.bursa-pastoris*).

S.N.	Protein name	Description	Protein Length	STRING score
1	SEP_3	Developmental protein SEPALLATA 3. Probable transcription factor active in inflorescence development and floral organogenesis.	251	0.998
2	TT16	Protein TRANSPARENT TESTA 16. Transcription factor involved in the developmental regulation of the endothelium and in the accumulation of proanthocyanidins (PAs)	252	0.986
3	AG	Floral homeotic protein AGAMOUS; Probable transcription factor involved in the control of organ identity during the early development of flowers.	252	0.882
4	AGL1	Agamous-like MADS box protein AGL1/shatterproof 1; Probable transcription factor	248	0.880
5	AGL11	Agamous-like MADS-box protein AGL5; Probable transcription factor	256	0.88
7	NST3	NAC domain-containing protein 12; Transcription activator of genes involved in biosynthesis of secondary walls.	358	0.502
8	PAB5	Polyadenylate-binding protein 5; Binds the poly(A) tail of mRNA	668	0.488
9	F14F8_180	Developmental protein SEPALLATA 1; Probable transcription factor.	251	0.400

proteins was done using SUMO (Sampson et al. 2001) in some plants. **S**mall **U**biquitin-like **Mo**difier or **SUMO** (http://sumosp.biocuckoo.org/) proteins are a family of small proteins that are attached to and detached from other proteins to modify their function. SUMOylation could alter the sub-cellular localization, activity or stability etc. of proteins (Fernandez-Lloris et al. 2006; Mahajan et al. 1997; Matunis et al. 1996). Besides protein sumoylation sites can play an important role in a variety of biological processes, such as transcriptional regulation, signaling transduction, cell cycle progression and differentiation (Deyrieux et al. 2007; Montpetit et al. 2006; Seeler and Dejean 2003), etc. The number of sumoylated residues vary from 1-4. Positions 157 and 170 are sumoylated in majority of proteins (Table 18).

Protein-Protein interactions

Protein-protein interactions in SHP1 and SHP2 proteins were analyzed using string database. The results are shown in Tables 19 and 20. The interacting partners were similar for all the plants with respect to SHP1 and SHP2 genes. These interactions are depicted pictorially in Figure 7 and 8.

Materials and Methods

Selection of proteins for analysis

The Arabidopsis SHP1 and SHP2 proteins sequences were retrieved from NCBI and were subjected to blast analysis using BLASTP against the nr-Protein Database. Some of the hits were selected for further analysis (Table 1 and 2). The genes corresponding to these proteins were used for FGENESH analysis. Table 3 shows various tools used for analysis in this study. The plants species viz., *Arabidopsis thaliana*, *Arabidopsis lyrata, Brassica napus, Capsella bursa-pastoris* have been taken for this study.

Conclusion

In conclusion we have analyzed diversity in SHP genes and proteins in this study. This diversity may be important to create diversity in the transgenic plants engineered for shattering resistance. This analysis also needs to be complemented with phenotypic data to derive functional significance of the diversity in SHP genes. This will aid in rational manipulation of crop plants for shattering resistance.

Acknowledgement

We are thankful to the Indian Council of Agricultural Research (ICAR) for financial assistance under the network project of Centre for Agricultural bioinformatics Scheme (CABin project code 1004936). We are also thankful to our colleagues from ICAR-Indian Agricultural Statistics Research Institute, New Delhi who provided insight and expertise that greatly assisted the research,

References

Atsushi I (1980) Thermostability and aliphatic index of globular proteins. J Biochem. 88(6):1895-1898.

Blom N, Gammeltoft S and Brunak S (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol. 294(5):1351-1362.

Butler D and Reichhardt T (1999) Assessing the threat to biodiversity on the farm. Nature, London (United Kingdom).

Deyrieux AF, Rosas-Acosta G, Ozbun MA and Wilson VG (2007) Sumoylation dynamics during keratinocyte differentiation. J Cell Sci. 120(1):125-136.

Fernandez-Lloris R, Osses N, Jaffray E, Shen L, Vaughan O, Girwood D, Bartrons R, Rosa J, Hay R and Ventura F (2006) Repression of SOX6 transcriptional activity by SUMO modification. FEBS Lett. 580(5):1215-1221.

Gomord V, Fitchette AC, Menu-Bouaouiche L, Saint-Jore-Dupas C, Plasson C, Michaud D and Faye L (2010) Plant-specific glycosylation patterns in the context of therapeutic protein production. Plant Biotechnol J. 8(5):564-587.

Gupta R (2001) Prediction of glycosylation sites in proteomes: from post-translational modifications to protein function (Ph.D. thesis).

Heger A and Holm L (2000) Rapid automatic detection and alignment of repeats in protein sequences. Proteins. 41(2):224-237.

Hounsell EF, Davies MJ and Renouf DV (1996) O-linked protein glycosylation structure and function. Glycoconjugate J. 13(1):19-26.

Kimura M (1968) Evolutionary rate at the molecular level. Nature. 217(5129):624-626.

Kimura M (1983) Rare variant alleles in the light of the neutral theory. Mol Biol Evol. 1(1):84-93.

King JL and Jukes TH (1969) Non-darwinian evolution. Science. 164(3881):788-798.

Liljegren SJ, Ditta GS, Eshed Y, Savidge B, Bowman JL and Yanofsky MF (2000) SHATTERPROOF MADS-box genes control seed dispersal in Arabidopsis. Nature. 404 (6779):766-770.

Lis H and Sharon N (1993) Protein glycosylation. Eur J Biochem. 218(1):1-27.

Mahajan R, Delphin C, Guan T, Gerace L and Melchior F (1997) A small ubiquitin-related polypeptide involved in targeting RanGAP1 to nuclear pore complex protein RanBP2. Cell. 88(1):97-107.

Matunis MJ, Coutavas E and Blobel G (1996) A novel ubiquitin-like modification modulates the partitioning of the Ran-GTPase-activating protein RanGAP1 between the cytosol and the nuclear pore complex. J Cell Biol. 135(6):1457-1470.

Meiering EM, Serrano L and Fersht AR (1992) Effect of active site residues in barnase on activity and stability. J Mol Biol. 225(3):585-589.

Mellor H and Parker PJ (1998) The extended protein kinase C superfamily. Biochem J. 332(2):281-292.

Montpetit B, Hazbun TR, Fields S and Hieter P (2006) Sumoylation of the budding yeast kinetochore protein Ndc10 is required for Ndc10 spindle localization and regulation of anaphase spindle elongation. J Cell Biol. 174(5):653-663.

Rechsteiner M, Rogers S and Rote K (1987) Protein-structure and intracellular stability. Trends Biochem Sci. 12(10):390-394.

Rechsteiner M and Rogers SW (1996) PEST sequences and regulation by proteolysis. Trends Biochem Sci. 21(7):267-271.

Sampson DA, Wang M and Matunis MJ (2001) The small ubiquitin-like modifier-1 (SUMO-1) consensus sequence mediates Ubc9 binding and is essential for SUMO-1 modification. J Biol Chem. 276(24):21664-21669.

Schreiber G, Buckle AM and Fersht AR (1994) Stability and function: two constraints in the evolution of barstar and other proteins. Structure. 2(10):945-951.

Seeler J-S and Dejean A (2003) Nuclear and unclear functions of SUMO. Nat Rev Mol Cell Bio. 4(9):690-699.

Shoichet BK, Baase WA, Kuroki R and Matthews BW (1995) A relationship between protein stability and protein function. P Natl Acad Sci. 92(2):452-456.

Solovyev V, Kosarev P, Seledsov I and Vorobyev D (2006) Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biol. 7(1):1.