Int J Pharm Bio Sci 2014 April ; 5 (2) : (B) 718 - 735

Research Article **Bioinformatics**

International Journal of Pharma and Bio Sciences **ISSN 0975-6299**

COMPUTATIONAL ANALYSIS OF ABIOTIC STRESS INDUCIBLE GENES AND PROTEINS FROM RICE (*ORYZA SATIVA* L.)

SUPRATIM BASU[1] AND ARYADEEP ROYCHOUDHURY[2]*

[1]University of Arkansas, 115, Plant Sciences Building, Fayetteville, AR 72701, USA
[2]Post Graduate Department of Biotechnology, St. Xavier's College (Autonomous), 30, Mother Teresa Sarani, Park Street, Kolkata-700016, West Bengal, India

ABSTRACT

In this communication, various *in silico* based prediction tools were employed to localize various stress-inducible genes [dehydrin (viz., RAB16A), polyamine-biosynthesis enzyme (viz., SAMDC), transcription factors (OSBZ8, NAC1, DREBP2, CRT/DRE-BP, WRKY24/51/71), ion channels and membrane transporters (HKT1, NHX1, SOS3) and protein kinases (SAPK5/7)] on rice chromosomes and establish their exon-intron organization, characterize the upstream sequences of these genes or analyze the regulatory domains in the upstream regions. The selected genes were found in seven of the 12 rice chromosomes. A comparative search for the conserved elements in the 5'-upstream region of these genes revealed fourteen common and most frequent relevant potential regulatory motifs. Their significance was evaluated by searching for their presence in transcription factor binding site databases. In addition, bioinformatic tools were used to determine the amino acid composition, signal sequences and secondary structures of the corresponding proteins; their subcellular localization, phosphorylation sites and relative solvent accessibility, i.e., percentage of solvent exposed residues. The positional information of these motifs would provide a basis for designing *in vivo* experiments for more accurate promoter function annotation and in validating the expression of these genes/proteins during abiotic stress in rice.

KEYWORDS: Abiotic stress, *in silico* analysis, *Oryza sativa*, Promoters, Secondary protein structure

ARYADEEP ROYCHOUDHURY
Post Graduate Department of Biotechnology, St. Xavier's College (Autonomous), 30, Mother Teresa Sarani, Park Street, Kolkata-700016, West Bengal, India, E-mail: aryadeep.rc@gmail.com

INTRODUCTION

Abiotic stresses resulting from high salinity or drought are the major factors constituting serious threats to agriculture and leading to worldwide crop loss [1, 2]. Thus, elucidation of mechanism governing plant response to such stress is of considerable scientific and economic importance. Several genes are involved that play a significant role in controlling plant adaptation to stress tolerance. The genes having similar expression patterns contain common motifs in their promoter regions [3]. Certain common promoter motifs are the key signatures for a family of co-regulated genes and are usually present in the regions where complex protein interactions occur [4]. However, in some cases, single motifs can bind various transcription factors (TFs), thereby bringing the genes under multiple regulatory controls [5]. Extensive studies on the upstream regions of yeast promoters suggest that regulatory elements are commonly present in those regions [6]. In eukaryotes, the computational detection of regulatory sites is difficult, as the sequences where TFs bind are generally much shorter than in prokaryotes [7]. In addition, they are generally active in both orientations and can be dispersed over a large distance. Sometimes, they can be present in introns and also in distal parts of the promoter [6]. When a plant promoter with a unique expression pattern is identified, it is desirable to characterize the promoter as well as the corresponding gene to gain a detailed insight into its function. Although the exact information to delineate a promoter and its regulatory elements requires experimental approaches like promoter deletions, substitutions and linker scanning, prior computational analysis of the sequence can serve as a guide to establish a platform for further promoter analysis[31] A number of public databases and software tools are available for analyzing the putative *cis*-acting motifs and regulatory elements in any promoter region [8-11]. Similarly, comparative sequence alignment, analysis of functional domains of a gene and homology based molecular modeling of the protein provide a useful foundation for systematic determination of gene function. A large number of computational programs are available to predict the architecture of a plant gene and conformation of the protein structure [12, 13]. We have earlier studied the physiological and molecular responses of rice plants to abiotic stresses like salinity and dehydration [14, 15]. Using different available web-based software tools, the present study attempted for *in silico* characterization of the upstream sequences and detection of various regulatory motifs of several abiotic stress-inducible genes of rice like *Responsive to Abscisic acid* (*Rab16A*) belonging to dehydrin or group 2 *late embryogenesis abundant* (*LEA*) gene, *S-adenosylmethionine decarboxylase* (*SamDC*) encoding the enzyme SAMDC for spermidine biosynthesis, along with the genes for ion transporters (*NHX1*, *SOS3* and *HKT1*), TFs (*OSBZ8*, *NAC1*, *DREBP2*, *CRT/DRE-BP*, *WRKY*s) and protein kinases (*SAPKs*). Our study has identified some of the relevant *cis*-regulatory elements in the promoters of the selected genes. Based on the *in silico* analysis, we also report here the amino acid composition of the corresponding proteins, hydrophobicity, probable phosphorylation sites, subcellular localization and prediction of their secondary structures, thereby suggesting their relative solvent accessibilities, functional motifs and polymorphic sites.

MATERIALS AND METHODS

Cis-acting regulatory elements and TF-binding sites

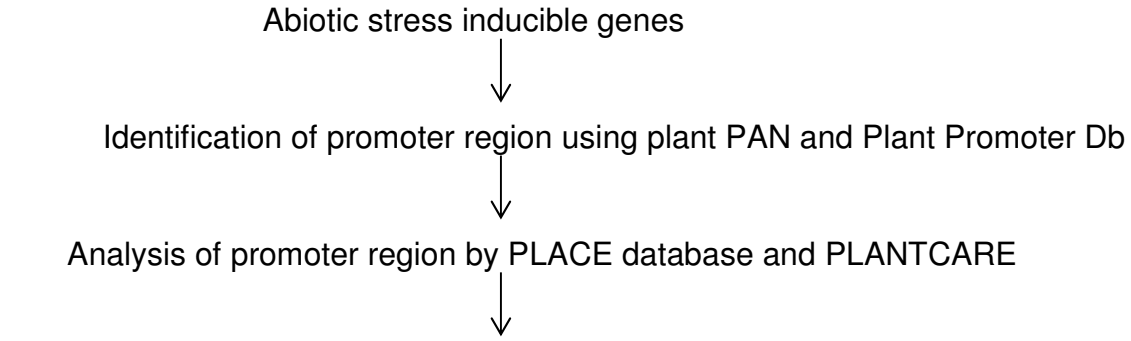
The promoter regions of the selected genes comprising of 910 bp upstream and 90 bp downstream sequences obtained from Plant Promoter Db (http://ppdb.gene.nagoya-u.ac.jp/cgi-bin/index.cgi) and Plant PAN (http://plantpan.mbc.nctu.edu.tw) was used as an input sequence for identifying *cis*-acting regulatory elements (CAREs) and TF-binding sites (TFBs). The software programs used

were PLACE (http://www.dna.affrc.go.jp/PLACE) [16], PlantCARE(http://bioinformatics.psb.ugent.be/webtools/plantcare/html) [17], and NSITE-PL (Softberry; http://www.softberry.com/berry.phtml). The potential TFs, binding to some important *cis*-acting regulatory motifs, identified from the above studies were further analyzed by searching the factor table of TRANSFAC® 7.0 database (http://www.gene-regulation.com/pub/databases.html#transfac) [18]. The common regulatory motifs in the upstream sequences of these genes in *Oryza sativa* and other plant species were identified by NSITE-PL (http://linux1.softberry.com/cgi-bin/programs/promoter/nsitep.pl).

In silico analysis of the gene and gene product structures

To determine the chromosomal location of all the above-mentioned genes, we determined the position of each gene on rice chromosome pseudomolecules available from Rice Genome Annotation Project (http://rice.plantbiology.msu.edu) and IRGSP (http://rgp.dna.affrc.go.jp/IRGSP). The genomic organization, i.e., the orientations of exons and introns of each gene were also determined utilizing TIGR database. The amino acid composition was determined using Expasy proteomics server (http://www.iut-arles.up.univ-mrs.fr/w3bb/d_abim/compo-p.html). To judge the post translational modification in these stress-inducible proteins, *in silico* analysis was performed with their protein sequence using NetPhos 2.0 (http://www.cbs.dtu.dk/services/NetPhos/) [19], an online phosphorylation-site prediction tool and confirmed by using another online software tool Scansite 2.0 [20]. The probable organelle targeting signal sequence was detected using TargetP 1.1 (http://www.cbs.dtu.dk/services/TargetP) [21], Mitoprot v1.0a4 (http://ihg.gsf.de/ihg/mitoprot.html) [22], Predotar v1.03 (http://urgi.infobiogen.fr/predotar) [23] and ProtComp v6.0 (SoftBerry Inc) (http://www.softberry.com). The signal sequences were identified by the program TargetP [21] and SignalP 3.0 server (http://www.cbs.dtu.dk/services/SignalP) [24]. Other analyses like secondary structure predictions or relative solvent accessibility (RAS) were assessed using the tool PROFsec and PredictProtein (www.predictprotein.org) [25] and GOR4 (http://npsapbil.ibcp.fr/cgibin/npsaautomat.pl?page=/NPSA/npsa gor4.html) [26]. The globularity of the proteins was tested by the software tool GLOBE of Predict Protein. The domain structures of these genes were analyzed using SMART database (http://smart.embl-heidelberg.de) [27], while secondary structures and polymorphic sites were predicted using SABLE (http://sable.cchmc.org) [28].

A flowchart for the overall methodology is depicted as follows:



RESULTS AND DISCUSSION

In silico analysis of the stress-inducible promoters for regulatory motifs and TF-binding sites

The abiotic-stress inducible genes that we studied were all present on seven of the twelve rice chromosomes (Fig. 1). Their genomic organization, including exon-intron structure is shown (Fig. 2), revealing that only one of these genes was intronless (*CRT/DRE-BP*). Among those having introns, the number varied from one to thirteen (Table 1). Their exact position in rice chromosome pseudomolecules is given in Table 1. In order to understand the evolution of these genes in rice, both segmental and tandem duplication was studied in detail (data not shown). The c*is*-acting regulatory elements or CAREs are the important DNA sequences, with short consensus sequences serving as binding sites for TFs to initiate transcription and modulate the complex gene regulation process by combinatorial interaction. Identification of these short regulatory sequences is therefore crucial to understand the nature of promoter function. Computer-aided search for *cis*-regulatory elements primarily serve as a platform for experimental identification of actual regulatory elements in the promoter. Locating such *cis*-elements in a plant promoter requires careful selection of specialized databases, which compare and analyze the sequence information in order to identify biologically relevant elements present in the promoter region. We carried out a detailed *in silico* search for identification of potential *cis*-acting elements or TF-binding sites in the upstream of these genes (1000 bp)[31] Out of the 120 motifs detected, we selected fourteen, that were the most frequent (they appeared in >46% of promoters, a minimum six or more promoters) (Fig. 3). The maximum occurrence in these promoters was around 100%. Three motifs were found with either 62% occurrence (nine out of fourteen) or more. The total information content (IC) [29] of each motif group is given in Figure 4. The minimum value for the total IC for each motif was >66% of the maximum possible values, while the maximum was 87% of those values. This indicated that the motifs in individual groups are very homogeneous (i.e., very similar to each other). The positional information of the potential regulatory elements related to the stress tolerance, deduced from PLACE and PlantCARE tool, will provide an important guidance for the *in vivo* analysis of the stress-inducible promoters in future experiments. All these motifs are supposedly involved in thtranscriptional enhancement. A comparative summary of the motifs obtained by different *cis*-element search programs is provided in Table 2. The regulatory motifs were compared amongst other plant species (Table 3), where we have seen motifs like abscisic acid responsive elements (ABREs), dehydration responsive elements (DREs) or Myb-binding sites, occurring in the upstream of these genes in most plant species. The TFs binding to these *cis*-acting elements are known and well characterized. However, there are some motifs like CTCACCAACCC, CCCCTCCTC and AAAATTTTTATA, for which the interacting TFs are not yet known and further works need to be done to characterize these TFs. Other important elements detected by PLACE and PlantCARE were related to gibberellic acid (GA) response, wounding and floral meristem development, which cannot be explained by the promoter activity until further experiments validate their functional presence. All these genes contained a TATA box; except for DREBP2 and WRKY24, where TATA box is absent and seemed to be a TATA-less promoter.

Structural prediction of the proteins and their probable localization in the cell

From amino acid composition analysis, it was revealed that HKT1, NHX1 and RAB16A are positively charged while OSBZ8 and WRKY51 are negatively charged. The detailed results of amino acid composition analysis are shown in Table 4. The prediction services for identifying putative subcellular localization through TargetP v1.1, Predotar v1.03, Mitoprot v1.0a4,

ProtComp v6.0 (SoftBerry Inc.) and PlantRBP (Plant RNA Binding Protein Database), all showed their probable localization in different parts of the cell with high scores (Table 5). SignalP 3.0 server showed the specific signal sequence for the proteins (Fig. 5). The secondary structure of these proteins was predicted using GOR4 prediction server which revealed that HKT1, NHX1, NAC1 and SAMDC contain a maximum percentage of random coils, while SOS3, which has a calcium binding domain, contains a maximum percentage of α helix; the details are shown in Table 6. This analysis tells us that none of these proteins has a predominant α helix but a mixed secondary structure. Using the NetPhos 2.0 and Scansite 2.0 tools, the proteins like OSBZ8, WRKY24, WRKY51 and HKT1 showed >15 serine residues with higher confidence value (Table 7), while all others like RAB16, and SAMDC showed threonine and tyrosine residues with lower confidence value. Using the tool PROFsec, the solvent accessibility of these proteins was found to be > 45% of all amino acid residues exposed by more than 16% of their surface. The software tool GLOBE of Predict Protein showed that most of these proteins, like the proteins belonging to the WRKY family, OSBZ8 and DREBP2 were not globular, while only SAMDC was found to be globular and compact as a domain (Table 8). The SOS3, SAPK5 and SAPK7 may be globular but will not be compact as a domain. All these rice abiotic stress-inducible proteins were analyzed for domain structures employing SMART database and compared with each other. The domains identified in these proteins are depicted in Figure 6. After analyzing all these stress-inducible proteins from rice, we noted that membrane transporters like HKT1 have a potassium transporter domain while NHX1 and SOS3 have a sodium/proton exchanger domain and a calcium binding motif respectively. On the contrary, RAB16 and SAMDC have a dehydrin and S-adenosyl methionine decarboxylase domain respectively. The TFs like WRKY24, WRKY51 and WRKY71 have similar sequence and highly conserved catalytic domains (W-box), thus likely exhibiting a common function. NAC1 has a NAM domain while OSBZ8 has a basic leucine zipper (bZIP) domain. Finally DREBP2 and CRT/DRE-BP share a common AP2 domain, suggesting that they have a similar function. We have also tried to find out the secondary structure, transmembrane domain and polymorphic sites of each protein using the SABLE and the results are depicted in Figure 7.

Table 1
Position of the genes on rice chromosome as obtained from TIGR and IRGSP

DNA	Position on chromosome	Gene ID/Locus	No. of introns	Predicted protein length	Predicted pI	Predicted protein molecular weight (kDa)
HKT-1	29,537,936 to 29,540,221	LOC_Os06g48810.1	2	531	9.82	59.29
SOS3	23,886,903 to 23,893,655	LOC_Os03g42840.1	7	226	4.5	25.8
NHX1	28,163,419 to 28,168,342	LOC_Os07g47100.1	13	536	8.65	59.07
SAPK5	35,159,194 to 35,161,130	LOC_Os04g59450.1	3	371	6.38	42.16
SAPK7	21,242,549 to 21,248,009	LOC_Os04g35240.1	8	360	6.08	41.32
SamDC	24,731,102 to 24,732,716	LOC_Os04g42095.1	1	451	4.72	48.79
NAC1	34,158,969 to 34,160,416	LOC_Os03g60080.1	1	317	7.15	35.17

Rab16A	14,856,502 to 14,857,840	LOC_Os11g26750.1	8	152	9.81	15.55
OSBZ8	26,799,016 to 26,804,464	LOC_Os01g46970.1	11	361	9.66	38.55
DREBP-2	3,355,383 to 3,360,204	LOC_Os01g07120.1	2	282	6.07	31.59
CRT/DRE-BP	6,812,714 to 6,814,546	LOC_Os01g12440.1	0	381	10.74	40.50
WRKY24	35,346,935 to 35,349,602	LOC_Os01g61080.1	4	556	7.10	59.30
WRKY51	12,416,278 to 12,418,171	LOC_Os04g21950.1	2	327	10.33	34.30
WRKY71	4,542,759 to 4,544,980	LOC_Os02g08440.1	3	349	7.91	37.22

Table 2
Comparative result of regulatory motif search in the Promoter regions of the stress inducible genes

Genes	Programs used	Number of TATA (T) and CAAT (C) motifs	Light responsive elements	Stress- inducible elements	Other important elements
NHX1	PlantCARE	T-9, C-8	G-box, GT1, I-box, L box, MNF1, SP1	ABRE, MBS	WUN, GCN4
	PLACE	T-3, C-4	GATA box, GT1, SORLIP, I-box	ABRE, DRE, W-box, MYB	SKN1, GCN4, NOD, OSE
HKT1	PlantCARE	T-12, C12	I-box, SP1	MBS, W-box	GCN4, HDzip1, SKN1
	PLACE	T-3, C-9	GATA box, GT1, SORLIP, I-box	W-box, MYB	GCN4, NOD, OSE
SOS3	PlantCARE	T-15, C-7	L-box , SP1	HSE, LTR	OCT
	PLACE	T-2, C-5	GATA box, GT1, SORLIP, I-box	ABRE, W-box, MYB	SKN1, GCN4, NOD
SAPK5	PlantCARE	T-12, C-3	G-box, GT1, I-box, GATA box	HSE, MBS	SKN1
	PLACE	T-2, C-1	GATA box, GT1, I-box	ABRE, W-box, MYB	SKN1, NOD, OSE
SAPK7	PlantCARE	T-15, C-8	G-box, GT1	ABRE, HSE, LTR	GCN4, SKN1
	PLACE	T-1, C-8	GATA box, GT1, SORLIP, I-box	ABRE, W-box, MYB	GCN4, NOD, OSE
NAC1	PlantCARE	T-12, C-5	G-box, I-box, MNF1, SP1	ABRE, LTR, MBS, MRE	GCN4, SKN1
	PLACE	T-1, C-4	GATA box, GT1, SORLIP, I-box	ABRE, W-box, MYB	GCN4, NOD, OSE
SamDC	PlantCARE	T-16, C-7	G-box,, L-box, SP1	HSE, LTR, MBS, W-box	GCN4, SKN1
	PLACE	T-1, C-3	GATA box, GT1, I-box	ABRE, W-box, MYB	GCN4, NOD, OSE
Rab16A	PlantCARE	T-7, C-12	G-box, GT1, SP1	ABRE, LTR, MBS	GCN4, SKN1
	PLACE	T-2, C-13	GATA box, GT1, SORLIP, I-box	ABRE, W-box, MYB	SKN1, GCN4, NOD
DREBP2	PlantCARE	T-N/A, C-N/A	MNF1, SP1	N/A	N/A
	PLACE	T-N/A, C-N/A	SORLIP, GATA- box	ABRE, W-box, MYB	SKN1
CRT/DRE-BP	PlantCARE	T-7, C-9	G-box, GT1, I-box, L-box, MNF1, SP1	ABRE, LTR, MBS, W-box	GCN4, SKN1
	PLACE	T-1, C-5	GATA box, GT1, SORLIP	ABRE, DRE, MYB, W-Box	NOD, OSE
OSBZ8	PlantCARE	T-5, C-7	GT1, I-box, L-box, MNF1, SP1	ABRE, LTR, MBS, W-box	GCN4, SKN1
	PLACE	T-2, C-5	GATA box, GT1, SORLIP, I-box	ABRE, MYB, W-box.	NOD, OSE

WRKY24	PlantCARE PLACE	T-12, C-1 T-N/A, C-9	G-box, GT1, SP1 GATA box, GT1, SORLIP, I-box	ABRE, MBS ABRE, W-box, MYB	SKN1 SKN1, GCN4
WRKY51	PlantCARE PLACE	T-6, C-5 T-3, C-2	G-box, SP1 GATA box, GT1, SORLIP, I-box	DRE, HSE, MBS, W-box ABRE, W-box, MYB	N/A GCN4, NOD, OSE
WRKY71	PlantCARE PLACE	T-20, C-3 T-5, C-2	I-box, MRE, SP1 GATA box, GT1, SORLIP, I-box	CRT/DRE, HSE, MBS ABRE, W-box, MYB	SKN1 SKN1, GCN4, NOD, OSE

Table 3
Common motifs in the upstream of the stress inducible genes in Oryza sativa and other plant species

Genes	Sequence in *Oryza sativa*	Motif found in other plant species	Binding factor
NHX1	GCCACTTGTC AAACCGAgAAAA AGAGAGAGA TaTtTAATTTTTT CTCACCAACCC cCTCACCAACCC	*Brassica napus* *Catharanthus roseus* *Arabidopsis thaliana* *Pinus sylvestris* *Petroselinum crispum* *Daucus carota*	ABI3 GT-1 BPC1 Seed nuclear proteins Unknown DcMYB1
HKT1	CGTGatGTCCATGCgT CAATAATTG AAAATTCAAT	*Zea mays* *Craterostigma plantagineum* *Arabidopsis thaliana*	ABRE binding factor CpbZIP1 CBF-C
SOS3	ATAATAAAA AAaATTTTTATa GcGGGTAGGTGa	*Pisum sativum* *Lycopersicon esculentum* *Arabidopsis thaliana*	PCF1 Unknown AtMYB84
SAPK5	AAaAATAcTAATtATAAAaA ACCGAGtCGTG TCCAtCCATCCA	*Glycine max* *Zea mays* *Daucus carota*	Unknown DRE binding factor DcMYB1
SAPK7	TACAaTTTTGG CCACGTCA AgaCGCCTCCTC	*Arabidopsis thaliana* *Triticum aestivum* *Zea mays*	OBP1 HBP11 ABF
NAC1	CTCGTgCTTATCTC ttTTGACTGATA CACACGTGCC CTCTATATAT	*Lycopersicon esculentum* *Arabidopsis thaliana* *Zea mays* *Phaseolus vulgaris*	Unknown WRKY EmBP1 (+VP1) Unknown
SamDC	CCCCTCCTC	*Lycopersicon esculentum*	Unknown
Rab16A	AAcATTTTTATc AGATGCCGACGCAa TGACTCAATG TCTCCCGCC	*Lycopersicon esculentum* *Hordeum vulgare* *Arabidopsis thaliana* *Zea mays*	Unknown HvCBF2 WRKY E2F/DP
DREBP2	CCACGTCA	*Triticum aestivum*	HBP11

CRT/DRE-BP	TGAATTTGTGc CCcACGTGGCGG AgAGAATCAAA AAGGTCCCT CCcCCAAcCACA	*Lycopersicon esculentum* *Hordeum vulgare* *Pisum sativum* *Vicia faba* *Pisum sativum*	GT-1 ABRE binding factor Unknown Unknown Unknown
OSBZ8	AcCCAATCCC CCACgTCATC TCCAtCCcTCCA ATaTTTTTGaAAATTg	*Lemna gibba* *Triticum aestivum* *Daucus carota* *Zea mays*	Unknown HBP-11 DcMYB1 HMG proteins
WRKY24	GCTTAATTA	*Lycopersicon esculentum*	Unknown
WRKY51	CATCCAACG TTTTcCTTTTcaTTTC CATCCAACG	*Triticum aestivum* *Lycopersicon esculentum* *Triticum aestivum*	Unknown Unknown WZF-1
WRKY71	GACTaGACCATCcTC GATGTGGTTTTT CTCTATATAT TTaTGAAATGA	*Hordeum vulgare* *Zea mays* *Phaseolus vulgaris* *Pisum sativum*	Seed proteins Unknown Unknown Unknown

Table 4
Comparative analysis of amino acid composition of the gene products

Proteins	Non-polar amino acids (%)	Polar amino acids with no charge (%)	Polar amino acids with positive charge (%)	Polar amino acids with negative charge (%)
NHX1	58.9	23.7	9.8	6.4
HKT1	58.2	25.1	10.8	5.9
SOS3	48.5	21.8	11.6	11.9
RAB16A	40.4	28.5	19.2	11.9
SAMDC	40.08	29.25	17.63	13.04
DRE-BP2	39.04	29.92	16.06	14.98
CRT/DRE-BP	41.12	27.68	8.7	22.5
SAPK5	48.3	22.3	16.4	13.0
SAPK7	43.5	25.8	16.2	14.5
NAC1	53.1	17.9	15.3	13.7
OSBZ8	49.72	23.61	7.78	18.95
WRKY24	46.83	32.25	11.17	9.75
WRKY51	44.56	23.92	8.75	22.96
WRKY71	49.98	22.98	14.66	12.38

Table 5
Summary of prediction results for subcellular localization of proteins

Proteins	Tool	Plastid	Mitochondria	Nucleus	Signal peptide (SP)	MB	Prediction
NHX1	TargetP Mitoprot Protcom	0.004	0.153 0.0096		0.530	0.499 1.71	SP Mitochondria Membrane Bound
HKT1	TargetP Mitoprot Protcom	0.014	0.245 0.8103		0.080	0.728 1.84	None Mitochondria Membrane Bound
SOS3	TargetP Mitoprot Protcom	0.070	0.107 0.0298		0.627	0.309 2.10	SP Mitochondria Membrane Bound
SAPK5	TargetP Mitoprot Protcom	0.050	0.1240 0.0106		0.031	0.903 0.99	None Mitochondria Cytoplasm
SAPK7	TargetP Mitoprot Protcom	0.033	0.1820 0.0366		0.027	0.894 0.59	None Mitochondria Cytoplasm
NAC1	TargetP Mitoprot Protcom	0.059	0.280 0.0416	 6.7	0.090	0.867	None Mitochondria Nucleus
SAMDC	TargetP Mitoprot Protcom	0.256	0.118 0.0105		0.078	0.733 1.8	None Mitochondria Membrane Bound
RAB16A	TargetP Mitoprot Protcom	0.120	0.213 0.0042		0.035	0.839 6.3	None Mitochondria Cytoplasm
DREBP2	TargetP Mitoprot Protcom	0.081	0.180 0.0548	 9.7	0.028	0.801	None Mitochondria Nucleus
CRT/DRE-BP	TargetP Mitoprot Protcom	0.915	0.032 0.7831	 5.0	0.024	0.304	Chloroplast Mitochondria Nucleus
OSBZ8	TargetP Mitoprot Protcom	0.097	0.067 0.1312	 8.6	0.027	0.925	None Mitochondria Nucleus
WRKY24	TargetP Mitoprot Protcom	0.715	0.043 0.0674	 9.1	0.016	0.353	Chloroplast Mitochondria Nucleus
WRKY51	TargetP Mitoprot Protcom	0.019	0.181 0.0595	 9.3	0.031	0.931	None Mitochondria Nucleus
WRKY71	TargetP Mitoprot Protcom	0.056	0.304 0.0164	 7.8	0.051	0.638	None Mitochondria Nucleus

Table 6
Prediction of Secondary Structure using GOR4

Proteins	α helix (%)	Extended strand (%)	Random coil (%)
NHX1	34.21	23.18	42.62
HKT1	37.55	16.98	45.47
SOS3	52.00	5.78	42.22
SAPK5	32.16	19.46	48.38
SAPK7	39.28	16.71	44.01
NAC1	24.84	15.92	59.24
SAMDC	18.59	23.62	57.79
RAB16A	19.21	20.53	60.26
DREBP2	25.18	17.52	57.30
CRT/DRE-BP	30.83	20.95	48.22
OSBZ8	30.56	12.50	56.94
WRKY24	18.74	12.61	68.65
WRKY51	16.97	15.76	67.27
WRKY71	25.29	13.51	61.21

Table 7
Prediction of phosphorylation sites of proteins by NetPhos 2.0

Proteins	Serine	Threonine	Tyrosine
NHX1	17	3	2
HKT1	11	3	2
SOS3	7	3	1
SAPK5	10	3	5
SAPK7	10	5	6
NAC1	9	4	4
SAMDC	9	3	7
RAB16A	10	3	0
DREBP2	7	2	1
CRT/DRE-BP	8	7	0
OSBZ8	25	10	0
WRKY24	28	9	3
WRKY51	15	6	3
WRKY71	17	3	2

Table 8
Relative Solvent Accessibility of proteins and prediction of their globularity

Proteins	All other residues	Residues exposed with more than 16% of their surface	Globularity
HKT1	69.06	30.94	May be globular
SOS3	36.44	63.56	May be globular
NHX1	72.52	27.48	No
SAPK 5	42.70	57.30	May be globular
SAPK 7	45.66	54.34	May be globular
SAMDC	48.49	51.51	Compact as a domain
NAC1	31.85	68.15	No
RAB16A	6.62	93.38	No
OSBZ8	33.33	66.97	No
DREBP2	26.64	73.36	No
CRT/DRE-BP	35.97	64.03	No
WRKY24	34.77	65.23	No
WRKY51	26.67	73.33	No
WRKY71	27.30	72.70	No

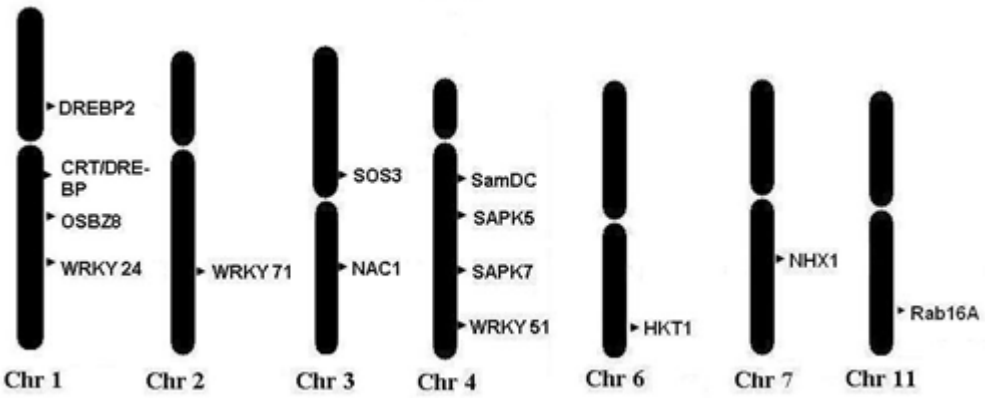


Figure 1
Chromosomal localization of abiotic stress-responsive genes of rice. The position of centromeres on the chromosomes is indicated.

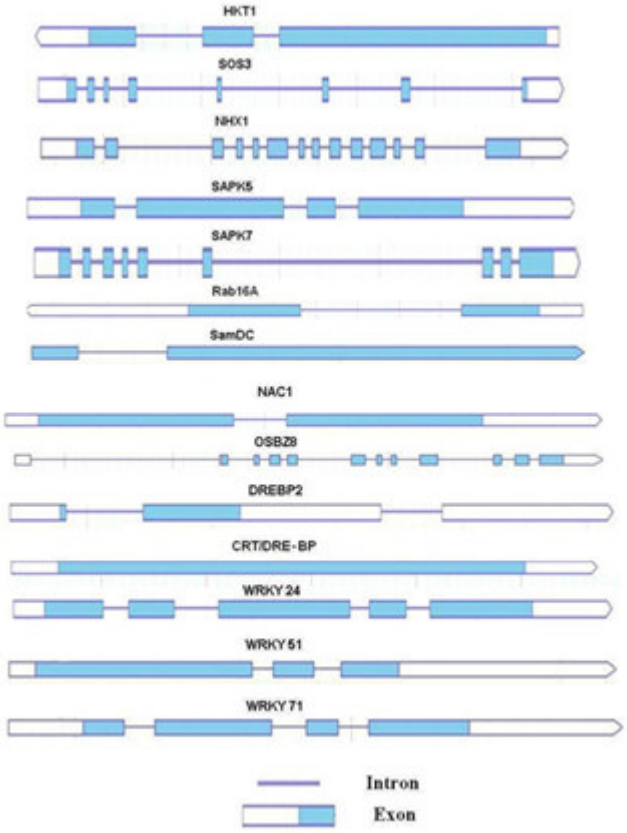


Figure 2

Genomic structure of abiotic stress inducible genes showing the exon-intron regions

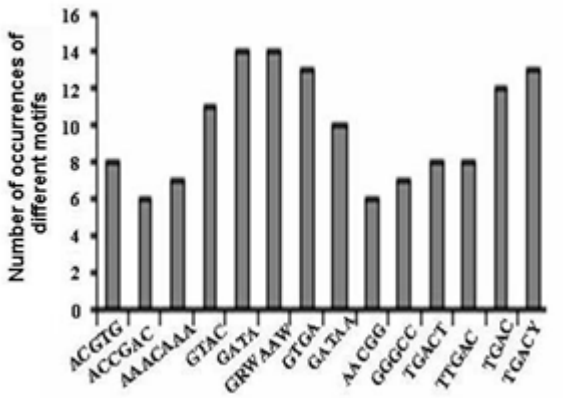


Figure 3

Fourteen different motif groups detected in the promoter sequences of fourteen abiotic stress-inducible genes involved in stress response. Each motif group is presented by the consensus sequence for the group. The number of promoters that contain the motif for specific groups is given.

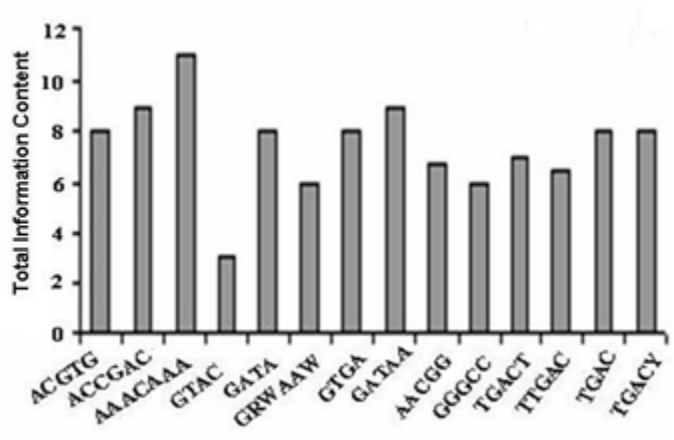


Figure 4

The total information content (IC) for each motif group is given. These indicate how homogeneous the motifs are in an individual motif group. The smaller the difference between the maximum IC and the total IC for the group, the more homogeneous are the motifs in the group, i.e. they are more similar to each other. All motif groups have a total IC >66% of the maximum possible IC for motifs of that length, indicating that motifs in the detected motif groups are very similar to each other within the group.

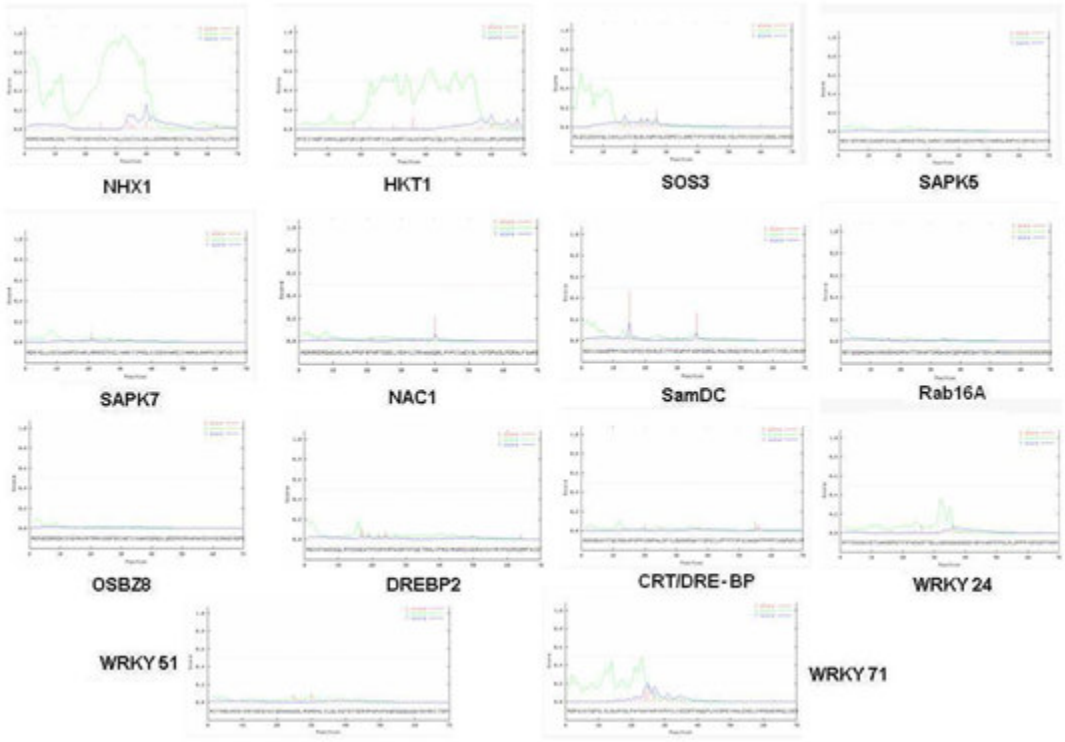


Figure 5

SignalP prediction of targeting signal in the various stress-inducible proteins.

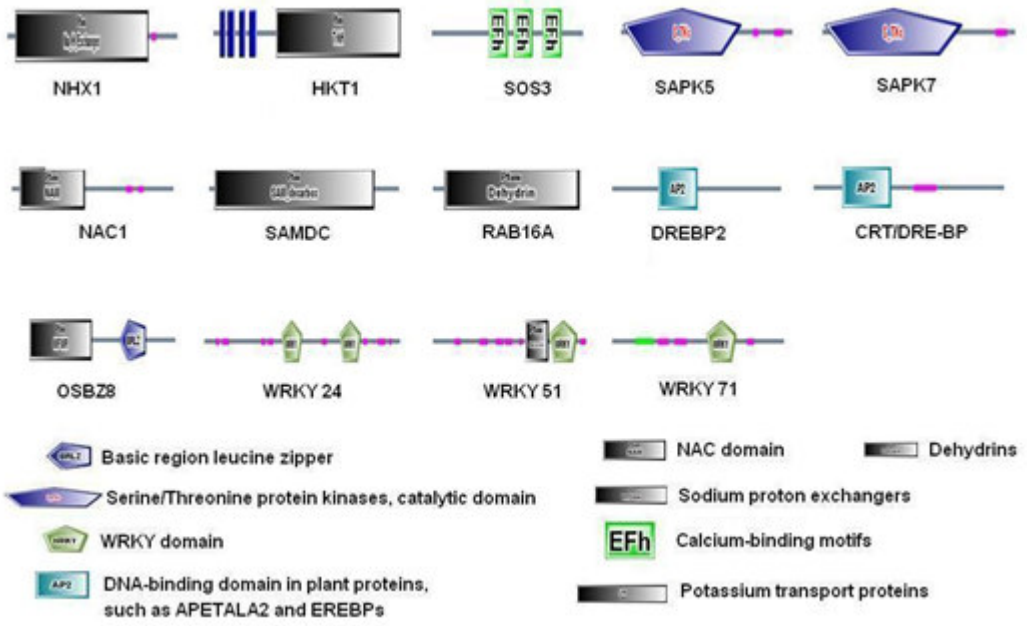


Figure 6

Domains present in different proteins encoded by rice stress-responsive genes. The presence of catalytic and other domains, as identified by SMART, was used to deduce the domain structure. Different colors are used for different domains.

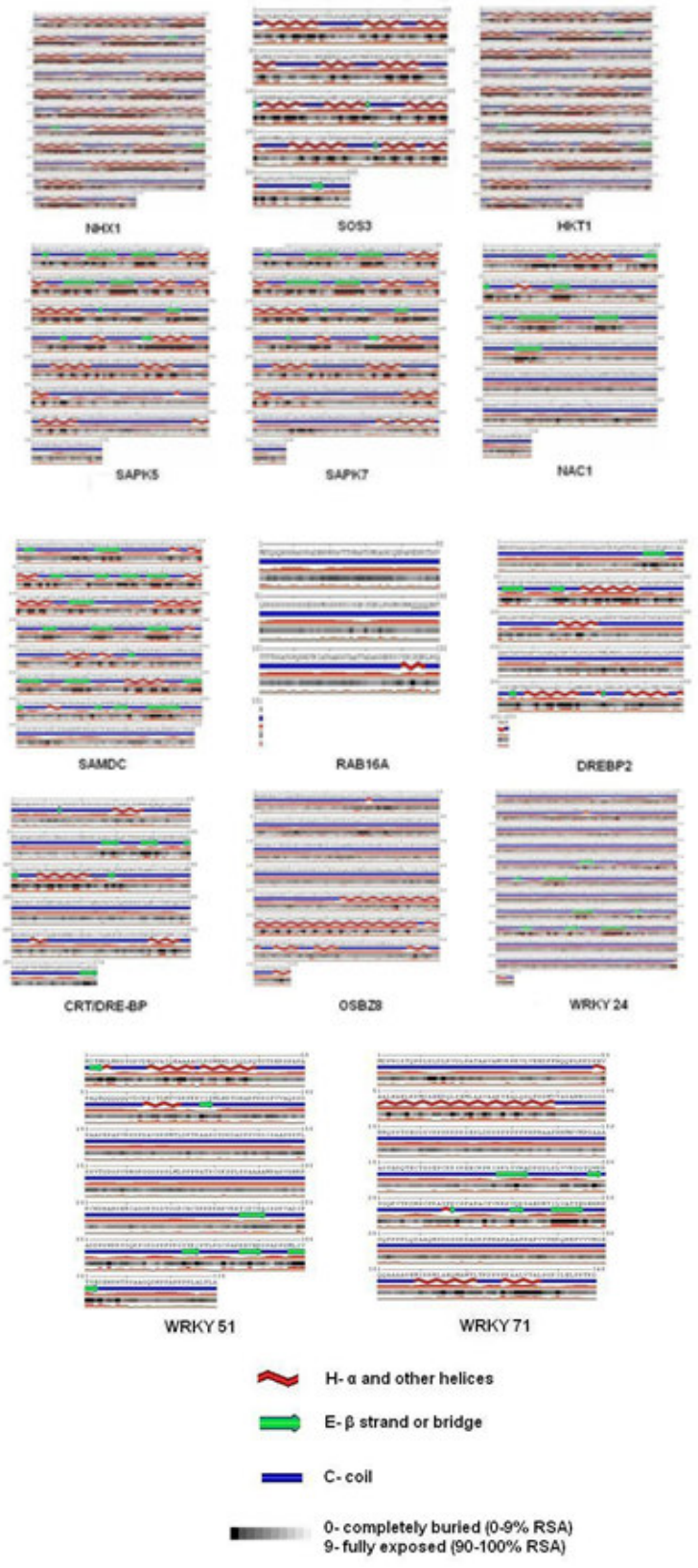


Figure 7

Prediction of secondary structure and relative solvent accessibility of different abiotic stress-inducible proteins using SABLE prediction server.

CONCLUSION

Computational programs serve as an invaluable tool for *in silico* analysis of DNA and protein sequences and are applied for identification of various *cis*-acting elements, [31].The TF-binding sites, structural features of genes and promoters as well as structural analysis of proteins are very important in deciphering the complex genetic regulatory networks prior to experimental validation. Employing the *cis*-regulatory motif prediction tools, we have identified few important *cis*-acting regulatory elements and the TFs binding to the promoter of stress-inducible genes. We also presume their involvement in stress tolerance of rice. The information obtained about the presence of the cis elements in the upstream region of the stress inducible genes was tested with the genes under study after salinity, drought, ABA, cold stress and light treatment by Reverse transcriptase-polymerase chain reaction (RT-PCR) in susceptible and tolerant rice cultivars (data unpublished). Furthermore gel mobility shift assay was done using different cis-elements as probe to find the regulators of *SamDC* gene in rice [30]. Bioinformatic analysis of such proteins thus provides an insight into their probable localization in the cell, about their secondary structural composition and also about their conserved domains. The information obtained from the present analysis will serve as an important resource in validating the expression of these genes/proteins in *Oryza sativa* and the regulation they exert in plant stress tolerance.

ACKNOWLEDGEMENT

Financial grant from the Science and Engineering Research Board, Department of Science and Technology, Government of India (SR/FT/LS-65/2010) to Dr. Aryadeep Roychoudhury is gratefully acknowledged.

REFERENCES

1. Grover, A., Pareek, A., Singla, S.L., Minhas, D., Katiyar, S. & Ghawana, S. (1998). Engineering crops for tolerance against abiotic stresses through gene manipulation. *Curr. Sci.,* 75: 689–696.
2. Khush G.S. & Baenziger P.S. (1998). *Crop improvement: emerging trends in rice and wheat* (Eds. Chopra, V.L., Singh, R.B. and Verma, A.) Crop productivity and sustainability - shaping the future, New Delhi: Oxford and BH publishing, pp.113–125.
3. Vilo, J., Brazma, A., Jonassen, I., Robinson, A. & Ukkonen, E. (2000). Mining for putative regulatory elements in the yeast genome using gene expression data. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology,* 8: 384–394.
4. Wang, Z., Dalkilic, M. & Kim, S. (2004). Guiding motif discovery by iterative pattern refinement. ACM Symposium on Applied Computing, Nicosia, Cyprus pp.162–166.
5. Jin, H. & Martin, C. (1999). Multifunctionality and diversity within the plant MYB-gene family. *Plant Mol. Biol.,* 41: 577–585.
6. Caselle, M., Di Cunto, F. & Provero, P. (2002). Correlating overrepresented upstream motifs to gene expression: a computational approach to regulatory element discovery in eukaryotes. *BMC Bioinformatics,* 3: 7.
7. van Helden, J., Andre, B. & Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.,* 281: 827-842.
8. Wingender, E., Dietze, P., Karas, H. & Knüppel, R. (1996). TRANSFAC: a database on transcription factors and their

DNA binding sites. *Nucleic Acids Res.,* 24: 238-241.

9. Fickett, J. W. & Hatzigeorgiou, A. G. (1997). Eukaryotic promoter recognition. *Genome Res.,* 7: 861-878.
10. Rombauts, S., Florquin, K., Lescot, M., Marchal, K., Rouzé, P. & van de Peer, Y. (2003). Computational approaches to identify promoters and cis-regulatory elements in plant genomes. *Plant Physiol.,* 132: 1162-1176.
11. 11 Molina, C. & Grotewold, E. (2005). Genome wide analysis of *Arabidopsis* core promoters. *BMC Genomics,* 6: 25.
12. Mathe, C., Sagot, M. F., Schiex, T. & Rouze, P. (2002). Current methods of gene prediction, their strength and weaknesses. *Nucleic Acids Res.,* 30: 4103-4117.
13. Davuluri, R. V. & Zhang, M. Q. (2003). Computer software to find genes in plant genomic DNA. *Methods Mol. Biol.,* 236: 87-108.
14. Roychoudhury, A., Basu, S., Sarkar, S.N. & Sengupta, D.N. (2008). Comparative physiological and molecular responses of a common aromatic indica rice cultivar to high salinity with non-aromatic indica rice cultivars. *Plant Cell Rep.,* 27: 1395–1410.
15. Basu, S., Roychoudhury, A., Saha, P.P. & Sengupta, D.N. (2010). Differential antioxidative responses of indica rice cultivars to drought stress. *Plant Growth Regul.,* 60: 51–59.
16. Higo, K., Ugawa, Y., Iwamoto, M. & Korenaga, T. (1999). Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res.,* 27: 297-300.
17. Lescot, M., Déhais, P., Thijs, G., Marchal, K., Moreau, Y., van de Peer, Y., et al. (2002). PlantCARE, a database of plant *cis*-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences. *Nucleic Acids Res.,* 30: 325-327.
18. Matys, V., Fricke, E., Geffers, R., Gößling, E., Haubrock, M., Hehl, R., et al. (2003). TRANSFAC®: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.,* 31: 374-378.
19. Blom, N., Gammeltoft, S. & Brunak, S. (1999). Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.,* 294: 1351-1362.
20. Obenauer, J.C., Cantley, L.C. & Yaffe, M.B. (2003). Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.,* 31: 3635-3641.
21. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.,* 300: 1005-1016.
22. Claros, M.G. & Vincens, P. (1996). Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.,* 241: 779-786.
23. Small, I. D., Peeters, N., Legeai, F. & Lurin, C. (2004). Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics,* 4: 1581-1590.
24. Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.,* 340: 783-795.
25. Rost, B., Yachdav, G. & Liu, J. (2004). The PredictProtein Server. *Nucleic Acids Res.,* 32: W321-W326.
26. Garnier, J., Gibrat, J.F. & Robson, B. (1996). GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.,* 266: 540–543.
27. Letunic, I., Doerks, T. & Bork, P. (2009). SMART 6: recent updates and new developments. *Nucleic Acids Res.,* 37: Database issue D229-D232.
28. Adamczak, R., Porollo, A. & Meller, J. (2005). Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins: Structure, Function and Bioinformatics*, 59: 467-475.

29. Stormo, G.D. (2000). DNA binding sites: representation and discovery. *Bioinformatics,* 16: 16–23.
30. Basu, S., Roychoudhury, A. & Sengupta, D.N. (2014) Identification of trans-acting factors regulating *SamDC* expression in *Oryza sativa*. BBRC (accepted in press). http://dx.doi.org/10.1016/j.bbrc.2014.02.004.
31. http://www.bioinfo.de/isb/2006/07/0002/main.html