Plant Mol Biol (2009) 69:167–178
DOI 10.1007/s11103-008-9414-5

An *in silico* strategy identified the target gene candidates regulated by dehydration responsive element binding proteins (DREBs) in *Arabidopsis* genome

Shichen Wang · Shuo Yang · Yuejia Yin · Xiaosen Guo · Shan Wang · Dongyun Hao

Received: 20 March 2008 / Accepted: 1 October 2008 / Published online: 18 October 2008

Abstract Identification of downstream target genes of stress-relating transcription factors (TFs) is desirable in understanding cellular responses to various environmental stimuli. However, this has long been a difficult work for both experimental and computational practices. In this research, we presented a novel computational strategy which combined the analysis of the transcription factor binding site (TFBS) contexts and machine learning approach. Using this strategy, we conducted a genome-wide investigation into novel direct target genes of dehydration responsive element binding proteins (DREBs), the members of AP2-EREBPs transcription factor super family which is reported to be responsive to various abiotic stresses in *Arabidopsis*. The genome-wide searching yielded in total 474 target gene candidates. With reference to the microarray data for abiotic stresses-inducible gene expression profile, 268 target gene candidates out of the total 474 genes predicted, were induced during the 24-h exposure to abiotic stresses. This takes about 57% of total predicted targets. Furthermore, GO annotations revealed that these target genes are likely involved in protein amino acid phosphorylation, protein binding and Endomembrane sorting system. The results suggested that the predicted target gene candidates were adequate to meet the essential biological principle of stress-resistance in plants.

Keywords Target genes prediction · Transcription factor binding site · DREBs · SVM · *Arabidopsis*

Abbreviations

TFs	Transcription factors
TFBS	Transcription factor binding site
DBD	DNA binding domain
DREBs	Dehydration responsive element binding proteins
DFSs	DRE frame sequences
nDFSs	Non-DRE frame sequences
MGs	Master genes
SVM	Support vector machine

Introduction

The identification of downstream target genes of specific transcription factors (TFs) is necessary in understanding cellular responses to environmental stimuli. Most existing structures of gene regulatory network are highly complicated as it involves cooperative interactions and feedback regulations. The discovery of the direct targets of TFs is a fundamental step to elucidate the construction of regulatory networks. Genes regulated by a given transcription factor is partially determined by the DNA binding domain (DBD) of a protein (Pabo and Sauer 1992). The DBD in TFs binds to

Electronic supplementary material The online version of this article (doi:10.1007/s11103-008-9414-5) contains supplementary material, which is available to authorized users.

S. Wang
College of Animal Science and Veterinary Medicine,
Jilin University, Changchun 130062, People's Republic of China

S. Yang · X. Guo · S. Wang · D. Hao (✉)
Key Lab for Molecular Enzymology and Engineering
of the Ministry of Education, Jilin University, Changchun
130023, People's Republic of China
e-mail: dyhao@cjaas.com

Y. Yin · D. Hao
Biotechnology Research Centre, Jilin Academy of Agricultural
Sciences (JAAS), Changchun 130033,
People's Republic of China

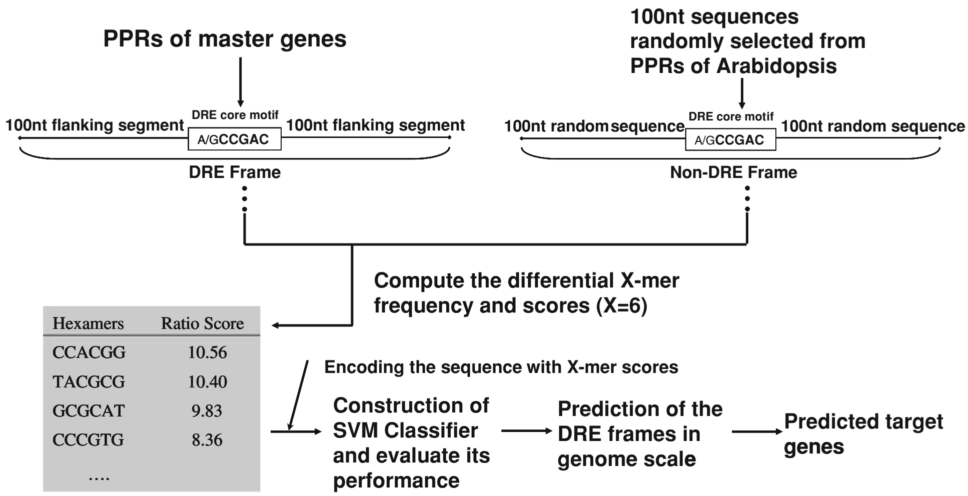
a specific DNA motif at the regulatory region of the target genes.

Availability of genome sequences made it possible to discover the target genes of a specific transcription factor by looking for the locations of the specific recognition motifs in genome. In practice, however, the task is still difficult due to the complication of plant genomes (Holstege and Clevers 2006). The development of microarray technology in measurement of mRNA expression profile promoted the identification of transcription factor target genes in genome (Young 2000). This approach allows the discovery of those genes that exhibit significant changes in mRNA levels upon inactivation or activation of a transcription factor (Qian et al. 2003, 2005; Maruyama et al. 2004). Despite this advancement, microarray profiling data might not be always sufficient in accurately identifying the direct target genes of TFs. For those organisms whose genomes are completed, an advanced computational strategy would be desirable to explore in genome-wide the binding motifs of the target genes for a given transcription factor.

During the last decade many computational methods have been developed to identify the target genes of TFs successfully (GuhaThakurta and Stormo 2001; Dieterich et al. 2003; Qian et al. 2003; Bigelow et al. 2004; Chan and Kibler 2005; Jolly et al. 2005; Zhang et al. 2005; Horsman et al. 2006; Redestig et al. 2007). Among the methods, the positional weight matrix (PWM) was the technique most widely used in describing the transcription factors binding sites (TFBS) and scanning the TFBS in the genome scale. However, owing to the looseness of the TFBS's conservation (Vavouri and Elgar 2005), these strategies were not capable of effectively identifying TFBS in genome scale. For this reason, the approach, including the PWM and the analysis of TFBS contexts, were developed to overcome the shortage (Kel et al. 2001; Rebeiz et al. 2002; Qian et al. 2005). The fundamental nature of the aforementioned approaches was in fact to develop appropriate algorithms that will describe the properties of the TFBSs and their contexts.

Dehydration responsive element binding proteins (DREBs) are important TFs that induce the expression of a series of abiotic stress-related genes and impart stress endurance to plants (Yamaguchi-Shinozaki and Shinozaki 1994; Agarwal et al. 2006). They belong to the ethylene responsive element binding factors (AP2-EREBPs) super family of 124 members (so-called ERF proteins), and among which 57 proteins are in the DREB subfamily (Riechmann et al. 2000). The ERF proteins share a conserved DNA binding domain (ERF domain) of 58–60 amino acids that, reportedly, binds to two typical *cis*-acting elements, that is, the GCC-box, and the C-repeat CRT/dehydration responsive element (DRE) motif and involves in the expression of cold and dehydration responsive genes (Ohme-Takagi and Shinshi 1995; Hao et al. 1998; Fujimoto et al. 2000; Brown et al. 2003; Song et al. 2005). It is important to identify the target genes of DREBs in *Arabidopsis* since the DREBs play a vital role in various types of biotic and abiotic stress responses. Maruyama et al. identified the downstream genes of the DREB1A/CBF3 using two microarray systems (Maruyama et al. 2004). Fowler and Thomashow, Taji et al. also reported the downstream genes of DREBs proteins (Fowler and Thomashow 2002; Taji et al. 2002). Nevertheless, the overall target genes of DREBs are yet to be discovered.

In this paper, we reported a novel computational strategy to determine the DREB transcription factor binding sites in *Arabidopsis* genome by combination of the context analysis for the TFBS and machine learning approach. As illustrated in the flowchart (Fig. 1), the differences between the DRE frame sequences (DFSs, see section "Materials and methods") and non-DRE frame sequences (nDFSs, see section "Materials and methods") were given focus. A machine learning approach, specifically the support vector machine (SVM) based classifier, was developed to


PPRs of master genes
100nt sequences randomly selected from PPRs of Arabidopsis
100nt flanking segment
DRE core motif
A/GCCGAC
100nt flanking segment
DRE Frame
100nt random sequence
DRE core motif
A/GCCGAC
100nt random sequence
Non-DRE Frame
Compute the differential X-mer frequency and scores (X=6)
Encoding the sequence with X-mer scores
Construction of SVM Classifier and evaluate its performance
Prediction of the DRE frames in genome scale
Predicted target genes

Hexamers	Ratio Score
CCACGG	10.56
TACGCG	10.40
GCGCAT	9.83
CCCGTG	8.36
....	

Fig. 1 A flowchart showing the strategy for genome-wide identification of DREBs-targeted genes

categorize DRE-containing sequences into DFSs and nDFSs. Our results suggested that this algorithm was effective in the discovery of the DREB binding sites in the promoter region of the target genes, so as to infer the target genes of DREBs in *Arabidopsis*. Furthermore, we predicted 474 candidate genes as the direct targets of DREBs. With Reference to the AtGenExpress microarray data, we achieved the 268 direct targets of DREBs that was inducible by abiotic stress stimuli such as cold, salinity and drought during a 24 h observation.

Materials and methods

The flowchart of working method reported in this study was illustrated in Fig. 1. The data sets used were listed in Table 1. Three definitions operationally used in this study were made as:

> Master genes (MGs)—The DREB-targeted genes that were experimentally identified contained the detectable DRE motifs in their putative promoter regions (PPRs).
> DRE frame sequences (DFSs)—DNA fragments of 206 bp, which were retrieved from the PPRs of MGs, contained a DRE motif (A/GCCGAC) at their center region.
> Non-DRE frame sequences (nDFSs)—DNA fragments of 206 bp, which were collected randomly from the PPRs of *Arabidopsis* genome, with a DRE motif inserted artificially at their center region.

Collection of *Arabidopsis* putative promoter regions

Arabidopsis putative promoter regions (PPRs), the sequences of 1000 bp upstream of the start codon, were retrieved based on the TAIR (www.arabidopsis.org) genome annotations. All PPRs were scanned to explore the DRE motifs containing the A/GCCGAC consensus.

Collection of MGs

We searched for target genes of DREBs in public databases, including PubMed searching. A total of 32 genes (Seki et al. 2001; Fowler and Thomashow 2002; Taji et al. 2002; Sakuma et al. 2006) were extracted which were further confirmed to be the target genes of DREBs by experimental approaches such as microarray analyses and real time PCR. The detailed information for these genes was listed in Table 2.

Analysis of the differential hexamers in the context of DRE sites

We collected 48 DFSs from the PPRs of MGs as the positive dataset and 1000 nDFSs as the negative dataset. HexDiff algorithm (Chan and Kibler 2005) was implemented to compute the hexamers frequency and ratio scores. Accordingly, the frequency of hexamers for positive $F_p(h)$ and negative $F_n(h)$ datasets on both strands were counted and used to calculate the ratio, R(h), as:

$$R(h) = \frac{F_p(h)}{F_n(h)}$$

The hexamers of larger ratio scores than the threshold were placed into H_d. In total we obtained 83 hexamers whose ratio scores R_h were greater than the threshold of 3.0. As a result of this selection process, H_d contained hexamers presented over in DRE frame rather than in non-DRE frame sequences. In this case, once the hexamers for H_d were chosen, the number of each hexamer N_h in a frame was counted and used to calculate the frame score S_f as:

$$S_f = \frac{\sum_{h \in H_d} N_h R(h)}{\text{Length of frame}}$$

The hexamer score S_h in a frame was calculated as:

$$S_h = N_h R(h), h \in H_d$$

Encoding the DRE-containing sequences

The hexamers in H_d were selected as the features used to encode the 206 bp DRE-containing sequences. Each sequence was encoded into an 84-dimension vector with 83 hexamer scores S_h and one label of either +1 or −1 to distinguish the DFSs (+1) or nDFSs (−1).

Table 1 The datasets used in this research

Dataset	Number of sequences	Description
DRE frame	48	Retrieved from the PPRs of master genes
Non-DRE frame	1000	Retrieved randomly from the PPRs
Random dataset	1000	Random sequence with equal nucleotide frequency p(A) = p(T) = p(G) = p(C) = 0.25 (used to evaluate the basic false positive rate, basic-FPR)

Table 2 List of the reported target genes of DREBs

AGI	DRE motif	Motif position	Functional description	Publications
AT5G52310	CATGGACCGACTACTA	−267	Hydrophilic protein	1, 3
	ATCATACCGACATCAG	−217		
	ATACTACCGACATGAG	−160		
	ATCAAGCCGACACAGA	−123		
AT2G17840	AAGCGACCGACCGACA	−204	Chloroplast related	2, 3, 5, 6
AT2G23120	CTACAGCCGACATCAG	−121	Unknown protein	1, 2, 5
AT5G62350	TCTGGACCGACGTTTA	−554	Endomembrane system	5, 6
	TGCTTGCCGACCTCTA	−186		
	TTTGTACCGACTATAA	−34		
AT5G15960	TAGCTACCGACATAAG	−79	Late embryogenesis-abundant protein	4, 5
AT5G15970	AAGCTACCGACATAAG	−125	Late embryogenesis-abundant protein	2, 5
AT1G01470	ATTCCACCGACGTGCA	−371	Late embryogenesis-abundant protein	1, 5, 6
	CATCGACCGACTTCAT	−23		
AT1G20440	ACATGACCGACATCTA	−989	Dehydrin	1, 3, 4, 5, 6
	TCAAAGCCGACCATTC	−960		
	CATCTACCGACTTCAA	−155		
AT1G20450	TTTCTGCCGACGTGGC	−603	Dehydrin	3, 4, 5
	ACATGACCGACATCCA	−567		
AT2G42530	TGATGGCCGACCTCTT	−188	Late embryogenesis-abundant protein	1, 2, 5
AT2G42540	TGTTGGCCGACATACA	−83	Hydrophilic protein	5, 6
AT3G50970	ACTACACCGACGTCTT	−207	Dehydrin	5, 6
AT4G15910	ATCTCACCGACCTCTT	−574	Late embryogenesis-abundant protein	5, 6
AT2G21660	CGTAAACCGACTCTAA	−81	Glycine-rich RNA-binding protein	4, 5, 6
AT1G22730	CATTGACCGACAATTC	−479	Zinc finger DNA binding protein	5, 6
AT5G58670	GGACAGCCGACAGATC	−154	Late embryogenesis-abundant protein	5, 6
AT3G24190	GTGGCGCCGACGTAGC	−302	Chloroplast	1, 5, 6
	TCGTTGCCGACGTAAT	−257		
AT4G24960	CTCTCACCGACCGACG	−329	HVA22 family	2, 3, 6
AT5G17460	GCCACGCCGACATAGT	−318	Mitochondrion	1, 5, 6
	AACAGGCCGACATAAT	−132		
AT5G04340	AAGTAGCCGACTTAAT	−406	Zinc finger DNA	1, 3, 5, 6
	TCTTAGCCGACTTCCA	−244	Binding protein	
AT2G17840	AAGCGACCGACCGACA	−204	Unknown protein	2, 3, 4, 5, 6
AT4G33070	TTTAGACCGACATAAA	−181	Chloroplast	3, 4, 5, 6
AT4G35300	ATTATGCCGACATTAA	−398	Sugar transport protein	5
AT4G14000	CACAGACCGACTTTAA	−976	Unknown protein	5
	TGGAAGCCGACTAAAA	−640		
	CGGAAGCCGACCAAAG	−553		
	CTCGTACCGACCGGTT	−44		
AT4G15910	ATCTCACCGACCTCTT	−574	Unknown protein	2, 5, 6
AT1G52690	CTCTCGCCGACCAAGA	−41	Late embryogenesis-abundant protein	5, 6
AT1G69870	ATATGACCGACAACAC	−334	Proton-dependent oligopeptide transport family protein	5, 6
AT1G22985	TCAAAACCGACTTGAT	−117	AP2 domain-containing DNA binding protein	4, 5, 6
AT5G52300	CGTGGACCGACTAAAA	−166	Hydrophilic protein	2, 5, 6
AT3G09390	ATTGAACCGACGTACG	−145	Metallothionein protein	5
AT1G01470	ATTCCACCGACGTGCA	−371	Late embryogenesis-abundant protein	5, 6
	CATCGACCGACTTCAT	−23		
AT2G23120	CTACAGCCGACATCAG	−121	Unknown protein	5, 6

Publications refer to the literatures that reported the genes as the regulatory targets of DREBs. 1. Fowler and Thomashow (2002); 2. Seki et al. (2002); 3. Taji et al. (2002); 4. Maruyama et al. (2004); 5. Sakuma et al. (2006); 6. Yamaguchi et al. (2006)

Support vector machines (SVMs)

The SVMs were standard supervised machine learning algorithms based on recent developments in statistical learning theory (Burges 1998; Vapnik 1998; Schölkopf et al. 1999; Schölkopf and Smola 2002). SVMs were binary classification tools that provided nonlinear function approximations by nonlinearly mapping the input vectors into feature spaces and using linear methods for regression or classification in feature space. Thus SVMs, and more generally kernel methods, combined the advantages of linear and nonlinear methods. For more details of SVMs, please refer Burges (1998), Vapnik (1998), Schölkopf and Smola (2002). Furthermore, SVMs have been used in many fields of bioinformatics such as splice site prediction, protein structural relationship prediction and microarray analysis and so on (Yu et al. 2003; Phan et al. 2005; Baten et al. 2006; Ogul and Mumcuoglu 2006; Towsey et al. 2006; Wee et al. 2006; Huang et al. 2007). In this study, we used SVM to discriminate DFSs from nDFSs. The Perl package of LIBSVM [http://www.csie.ntu.edu.tw/∼cjlin/libsvm], an implementation of SVM by Chih-Chung Chang et al., was used to construct our classifier.

Prediction of the target genes of DREBs

Once the construction of SVM classifier was completed, the SVM classifier was then used to scan the PPRs in order to get the DFSs in *Arabidopsis* genome. Using Regular Expression technique, we first scanned the PPR of genes in genome-scale to discover the DREB consensus (A/GCCGAC). After that, to determine whether those consensuses with the 200 bp flanking sequences on both sides were DRE frames or not, they were classified through the optimized SVM classifier. In this case, the gene containing at least one DFS in its PPR was considered to be the target gene candidate of DREBs. Thereafter, the gene ontology (GO) enrichments of the candidate genes were then investigated using GO annotation system (Ashburner et al. 2000). The GO term file for *Arabidopsis* was downloaded from TAIR (filename: gene_association.tair.gz) and GO-TermFinder (version 0.81) (Boyle et al. 2004) Perl module was used to get the term distribution of genes and the significance of the each groups.

Collection of gene expression data

The AtGenExpress gene expression datasets, GSE5620 (for control), GSE5621 (cold stress), GSE5623 (salinity stress) and GSE5624 (drought stress), were downloaded from the gene expression omnibus (GEO) of the National Center for Biotechnology Information (NCBI). Those genes were considered as inducible genes if their expression level fluctuated over 2-fold (either up or down) after cold, drought or salinity stress treatments.

Results

Being a group of important TFs responsible for abiotic stresses in plants (Chen et al. 2003; Kasukabe et al. 2004; Li et al. 2005; Agarwal et al. 2006, 2007; Wang et al. 2006; Xiong and Fei 2006), some DREBs were studied in the last decade and only 32 direct target genes of DREBs were reported (Fowler and Thomashow 2002; Seki et al. 2002; Taji et al. 2002; Maruyama et al. 2004; Sakuma et al. 2006; Yamaguchi et al. 2006). The DFSs in the PPRs of master genes were collected using home-made Perl scripts. The data was presented in Table 2. It was apparent that the collected DRE motifs of the MGs located within the 1000 bp upstream of the translation start codon (TSC). Thus, sequences of 1000 bp upstream of TSC were selected as the PPRs for this study. Given that not much information about promoters of non-DRE binding proteins target genes was available, as a control, the nDFSs were randomly extracted from the PPRs with DRE consensus (A/GCCGAC) added artificially (Fig. 1). In addition, a random sequences dataset was generated and was used to evaluate the false positive rate (Table 1).

Discovery of over-represented hexamers in the DFSs

To identify the over-represented hexamers distributed among DFSs, the hexamers frequency ratios R(h) between DFSs and nDFSs were calculated in each dataset. Because there were only 48 DFSs found in positive dataset, which was indeed far less than the number of the nDFSs in the negative dataset, the sampling with replacement was then carried out to increase the positive dataset from 48 DRE frames to 500 DRE frames, which was about half the scale of the negative dataset. Appling the threshold varying from 2.0 to 5.0 with an interval of 0.5, we were able to get seven hexamer collections. Furthermore, we accessed the performance of SVM classifiers trained preliminarily on datasets encoding with each hexamer collections. Such was done to find out the satisfactory hexamer collections that were used as sequence feature. The results showed that the threshold at 3.0 gave rise to an optimal SVM classifier training in terms of accuracy and sensitivity (See Electronic supplementary material). Consequently, we collected 83 hexamers (See Electronic supplementary material) with ratio scores >3.0 that were considered to be the distinctive motifs for further analysis.

Optimization of SVM parameters and performance evaluation

In general, the SVM involves two classes of parameters: the capacity parameter *C* and kernel type. *C* is a regularization parameter that controls the tradeoff between maximizing the margin and minimizing the training error. In the radial basis function (RBF) kernel used in our study, γ is an important parameter to dominate the generalization ability of SVM by regulating the amplitude of the kernel function. Accordingly, two parameters, *C* and γ, should be optimized. Thus, the parameter optimization was performed using a grid search approach within a limited range.

Using the optimized value of *C* and γ, the prediction model was thereafter constructed based on the training set with the RBF kernel. Moreover, to verify whether the hexamer features used in our study were able to effectively discriminate the DRE frames from the non-DRE frames and to avoid the dependence of the nDFSs dataset, five datasets were then constructed. Each dataset contained the 48 DFSs and 100 nDFSs that were randomly selected from the non-DRE frame dataset. The performances of SVM classifiers training each dataset were accessed by 5-fold cross-validation, and the data obtained from such procedure were presented in Table 3. It was shown that our classifier achieved the average accuracy of 85.6% with the average basic false positive rate of 9.12%. The parameters were likewise taken (C, $\gamma = 64$, 16) to scan the genome, as they were in the best prediction accuracy (87.2%). We also constructed the prediction models using three other kernel functions, i.e. polynomial, sigmoid, and linear functions, on the same data sets. However, the prediction accuracy of the aforementioned kernel functions achieved was not as satisfactory as that of the RBF kernel (data not shown).

Hunting the target gene candidates of CRT/DRE binding factors in genome scale

The regular expression protocol was used at first to explore the putative DFSs in the PPRs of genes in genome scale. After each sequence was encoded as vectors with 83 hexamers scores, the SVM classifier was then used to categorize whether the sequence was a DFS or not. Apparently, each gene occupied at least one predicted DFS among its PPRs was considered to be the target gene candidate of DREBs. From such procedure, a total of 474 target gene candidates, including 26 MGs, were obtained. The average number of targets per DREB transcription factor was about 8. The locus identifiers of target genes and DRE frame scores were listed in (See Electronic supplementary material). Mapping the positions of each predicted target gene to chromosomes indicated that the target genes candidates were well-distributed among the all five chromosomes in *Arabidopsis*.

Target genes responsive to abiotic stress

In this study, the expression profiles of those target gene candidates were investigated after 30-min exposure to the stresses such as cold, drought and salinity. Under such condition, a number of candidate genes induced were obtained and were presented in the Venn diagram (Fig. 2). A total of 160 genes (See Electronic supplementary material) were considered to be the direct targets of DREB TFs responsive to the abiotic stresses. Among them, 27, 91 and 106 genes were respectively responsive to drought, salinity and cold stress, respectively. In addition, there were ten genes were identified to be responsive to all the three types of abiotic stresses, indicating that they act as the cross-talkers in all likelihood among the drought, salinity and cold responsive pathways.

The positional distributions of the DRE motif occurred in each promoter of stress responsive genes were investigated. It was shown that the 160 target gene candidates

Table 3 Performances of SVMs trained on different datasets

Dataset	(C, γ)	Accuracy (%)	TP	FN	TN	FP	Sn (%)	Sp (%)	Basic-FPR[a] (%)
1	8, 128	85.2	32	16	94	6	66.7	94.0	8.44
2	64, 16	87.2	33	15	96	4	68.8	96.0	13.74
3	2, 512	85.9	33	15	94	6	68.8	94.0	14.32
4	8192, 0.0625	85.2	32	16	94	6	66.7	94.0	5.18
5	256, 4	84.5	30	18	95	5	62.5	95.0	9.76
Average		85.6 ± 1.02					66.7 ± 2.6	94.6 ± 0.9	9.12 ± 3.08

Accuracy = (TP + TN)/(TP + FN + TN + FP)

Sn = TP/(TP + FN), Sp = TN/(TN + FP)

[a] Basic-FPR was evaluated by the random dataset

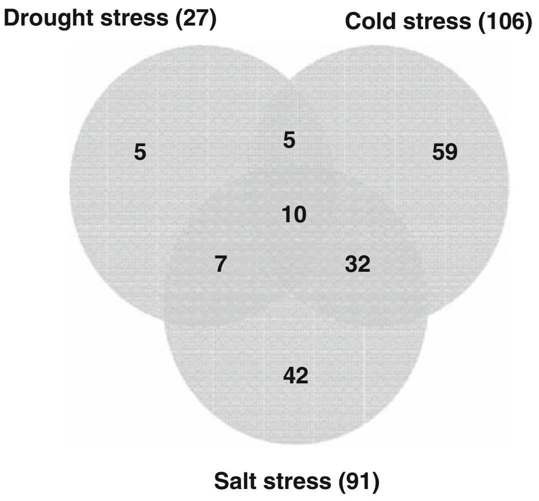


Fig. 2 Venn diagram showing the target genes induced after 30 min exposure to cold, drought or salinity

harbored the most DRE motifs within the 300 bp upstream of the start codon (Table 4). By analyzing of the gene expression profiles for 24 h, other 108 candidates were detected to be responsive to abiotic stresses (See Electronic supplementary material). In total, 268 target gene candidates (about 57% of predicted candidate genes) exhibited to be responsive to abiotic stresses.

Gene Ontology enrichment analysis of candidate targets of DREBs

Through the gene ontology enrichment analysis, it was found out that DREB target gene candidates were diverse, in terms of their functionality, molecular process and cell component categories. Approximately 5% of the candidates (38) were categorized as responsive to stress, and other 4.7% candidates (33) were functionally assigned as responsive to abiotic or biotic stimulus. Figure 3 illustrated that candidate target genes were among other categories significantly represented in “GO: 0012505 (endomembrane system)”, “GO: 0006355 (regulation of transcription)”, “GO: 0006468 (protein amino acid phosphorylation) and “GO: 0005515 (protein binding)”. In addition, the target gene candidates induced by abiotic stresses showed the similar functional distributions pattern as the total predicted candidates (see Fig. 3).

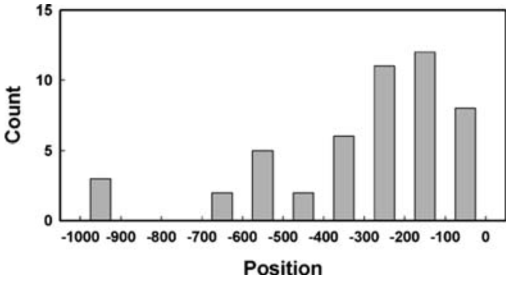
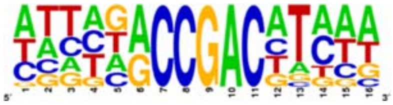
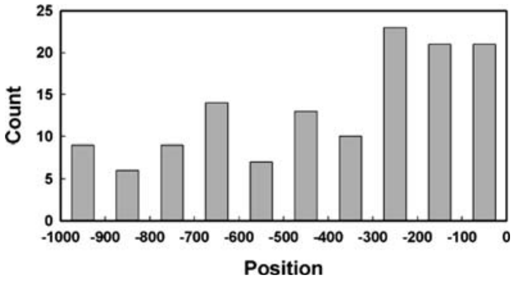
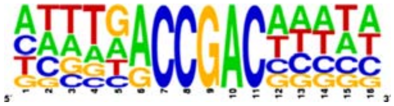
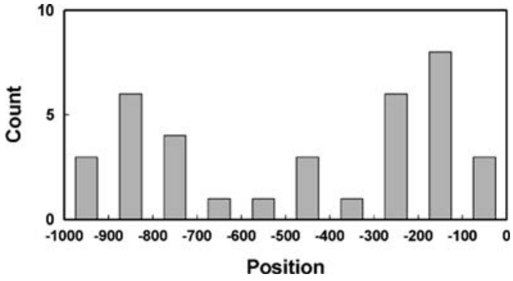
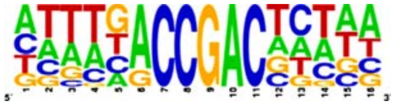
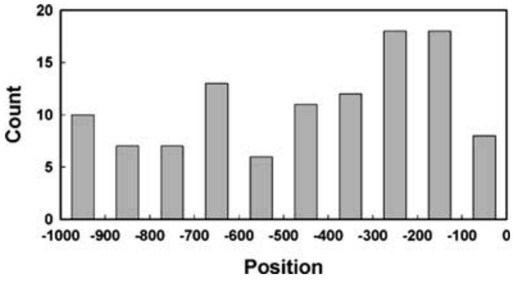
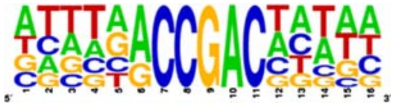
Discussion

In this study, attempts were made to discover computationally the DREBs-specific target genes through exploration of the DREBs binding sites in genome scale. Owing to the looseness of the DRE motif, it is insufficient to infer the targets of DREBs by simply scanning the DRE consensus in the PPRs of genes. A number of studies have already revealed that *cis*-acting elements distributed among the context of conserved TFBS also influenced the recognition by TFs (Wasserman and Fickett 1998; Kel et al. 2001; Krivan and Wasserman 2001; Zhang et al. 2005). It can be presumed that, in the PPRs of DREBs target genes, there were not only the DRE motifs but also other *cis*-acting elements were flanking to the DRE motifs. Those *cis*-acting elements were thus used as supplementary information to promote the inference of the target genes of DREBs. The HexDiff algorithm, which was previously reported to predict novel *cis*-regulatory modules (CRMs) (Chan and Kibler 2005), was employed to get the differential *cis*-acting elements between the DFSs and the nDFSs. Yamamoto et al. and Gertz et al. identified the promoter constitution and regulatory regions using similar algorithms (Gertz et al. 2006; Yamamoto et al. 2007). The requirements of HexDiff algorithm suggested that it worked effectively in a well-defined system where the TFBS frames and non-TFBS frames were already clarified. As with many bioinformatic analyses, there is restriction based on the initial dataset on which predictions are based. In this study, we used 48 DFSs as positive data, which were relatively less than the negative data. It is expected that the accuracy of our algorithm would be improved when more DFSs are discovered and used for training in the future.

In machine learning, the imbalance problem (Japkowicz 2000) has to be considered when there existed great disparity between the size of the positive and negative training sets. In general, there are two kinds of method which were commonly introduced to overcome such imbalance problem, the methods in which the class represented by a small dataset gets over-sampled so as to match the size of the other class and the methods in which the class represented by the large dataset can be down-sized so as to match the size of the other class. In this study, the imbalance problem was encountered since there were more negative samples (nDFSs, 1000) than positive samples (DFSs, 48). In the process of discovering of over-represented hexamers, we augmented the size of positive dataset by the sampling with replacement to match the size of the negative dataset. While in the process of SVM training, we used 100 negative samples by downsizing the negative dataset that will match the positive dataset (48 samples).

A further attempt to understand the meaning of the hexamers was done by searching for hexamers in the databases of JASPAR (Sandelin et al. 2004) and TRANSFAC (Matys et al. 2003). Based on the gathered information, some hexamers were shown to be the conserved sequences of known *cis*-acting elements, such as the abscisic acid responsive element which was reported to be the coupling *cis*-acting element with the DRE motif in the stress

Table 4 The position distribution and sequence logo of DRE motifs in the master genes (MGs) and stress responsive genes

	Position distribution of DRE motifs	DRE motif frequency logo
Master genes (38)		
Cold responsive genes (106)		
Drought responsive genes (27)		
Salt responsive genes (91)		

The DRE motif frequency logos were generated by the WEBLOGO (Crooks et al. 2004) online server (http://weblogo.berkeley.edu/). The overall height of each stack indicated the sequence conservation at that position, whereas the height of symbols within the stack presented the relative frequency of the corresponding nucleic acid at that position

response (Zhang et al. 2005) and the Myb responsive element which was reported to be effective to increase the tolerance of the cold, drought and salinity stress (Dai et al. 2007). This information was valuable in analyzing the cooperative gene regulation of TFs.

The frame scores of DFSs and nDFSs showed that 43% of nDFSs scores were under 0.05, while only 5% accounted for DFSs scores (Fig. 4). It was difficult to give out a cut-off value to discriminate between DFSs and nDFSs as that was used in the reference (Chan and Kibler 2005). For this

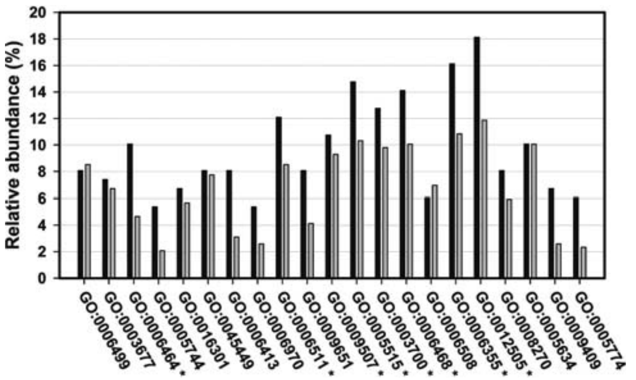


Fig. 3 Gene Ontology enrichments of target gene candidates. Black bars represented the relative frequency of GO classes in the set of total 474 target genes identified; grey bars indicated the relative frequency of GO classes for the set of stress responsive target genes. GO classes covering <2% of the genes were omitted for clarity and unknown classes were not shown. Significant overrepresented classes for the total 474 target genes (black bars) were indicated (*, $P < 0.01$). (GO:0006499, *N*-terminal protein myristoylation; GO:0003677, DNA binding; GO:0006464, protein modification process; GO:0005744, mitochondrial inner membrane presequence translocase complex; GO:0016301, kinase activity; GO:0045449, regulation of transcription; GO:0006413, translational initiation; GO:0006970, response to osmotic stress; GO:0006511, ubiquitin-dependent protein catabolic process; GO:0009651, response to salt stress; GO:0005515, protein binding; GO:0003700, transcription factor activity; GO:0006468, protein amino acid phosphorylation; GO:0006508, proteolysis; GO:0006355, regulation of transcription, DNA-dependent; GO:0012505, endomembrane system; GO:0008270, zinc ion binding; GO:0005634, nucleus; GO:0009409, response to cold; GO:0005774, vacuolar membrane)

reason, we introduced the machine learning technique. To describe the characteristics of the TFBS context, the applications of linear classification model (Kel et al. 2001) and logistic regression model (Krivan and Wasserman 2001) were previously reported. The SVM was for the first time introduced in our study to discriminate the characters of the TFBS context. The results suggested that our computational approach was satisfactory and the resulting accuracy (more than 85%) was superior to that reported by Krivan and Wasserman (2001).

In this study, we mined 474 genes as the direct target genes of DREBs from the pool of genes with the DRE motifs in its PPRs. Among the 474 target genes, 56 genes (12%) were annotated as the stress responsive genes by GO annotations. Additionally, the functional distribution of the predicted 474 target genes was diverse. The top four categories involved were the endomembrane system, the regulation of transcription, protein binding and protein amino acid phosphorylation. Endomembrane system was reported to be a proteinase-sorting system involving in assistance of the plant cells under various stress conditions (Matsushima et al. 2002; Chitteti and Peng 2007; Wang et al. 2007; Prak et al. 2008). Protein phosphorylation was another mechanism which was largely reported to be responsible for the stress-relative proteins activation or deactivation (Kiegerl et al. 2000; D'Angelo et al. 2006; Boudsocq et al. 2007; Chae et al. 2007; Schweighofer and Meskiene 2008). Moreover, the phosphoproteome analyses of rice and *Arabidopsis* revealed the important roles of the phosphorated proteins responsive to different stresses (Khan et al. 2005; Chitteti and Peng 2007). These results suggested that the target gene candidates of DREBs predicted through the strategy proposed in this study were adequate in meeting the essential biological principle of stress-resistance in plants.

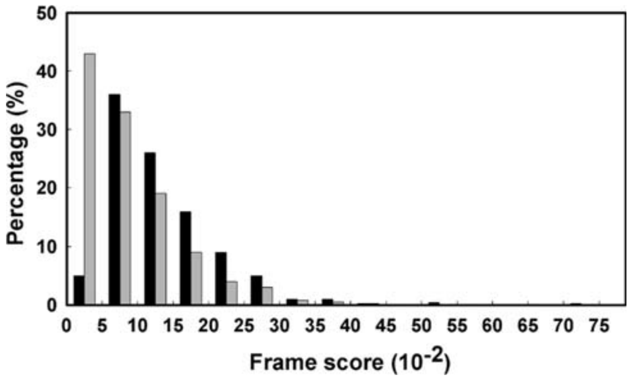


Fig. 4 Histogram showing the distribution of the frame scores in the DFSs and nDFSs datasets. Black bars indicated the frame score distribution of the DFSs dataset; grey bars indicated the frame score distribution of the nDFSs dataset

Exposing to the stresses for 30 minutes, 27, 91 and 106 genes out of the 474 candidate genes were induced respectively by drought, salinity and cold. Extending the stress exposure time to 24 h, further 108 genes out of the remaining genes in *Arabidopsis* were induced. The remaining genes may be the stress-inducible genes that appeared after the 24-h exposure to abiotic stresses, though the possibility of the false positives occurring in this case was not ruled out in this study. Nevertheless, our strategy was able to produce an immediate prediction with the existing TFBS knowledge and the microarray experiments that would be useful as a first step in the discovery of TF target genes.

Conclusions

In this paper, we reported a novel computational strategy. Using the method proposed in this study, the analyses of TFBS contexts and SVM classification led to the identification of 474 target gene candidates of DREBs, a group of TFs playing a vital role in abiotic stresses responsive regulatory network in plants. With reference to the relevant DNA microarray data, the results showed that 160 candidate genes were induced in 30 min and other 108 candidate genes were induced during 24-h exposure to abiotic stresses. The results obtained in this study provided the primary information that warranted further experimental investigation regarding the anti-stress regulatory network of DREBs in plants.

Acknowledgments This research was financially supported by the National Natural Science Foundation of China (Grant No. 30470159/C01020304) and the National High-Technology Research and Development Program ("863"Program) of China (Grant No. 2007AA10Z110).

References

Agarwal PK, Agarwal P, Reddy MK, Sopory SK (2006) Role of DREB transcription factors in abiotic and biotic stress tolerance in plants. Plant Cell Rep 25:1263–1274. doi:10.1007/s00299-006-0204-8

Agarwal P, Agarwal PK, Nair S, Sopory SK, Reddy MK (2007) Stress-inducible DREB2A transcription factor from *Pennisetum glaucum* is a phosphoprotein and its phosphorylation negatively regulates its DNA-binding activity. Mol Genet Genomics 277:189–198. doi:10.1007/s00438-006-0183-z

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 25:25–29. doi:10.1038/75556

Baten A, Chang B, Halgamuge S, Li J (2006) Splice site identification using probabilistic parameters and SVM classification. BMC Bioinformatics 7(Suppl 5):S15. doi:10.1186/1471-2105-7-S5-S15

Bigelow HR, Wenick AS, Wong A, Hobert O (2004) CisOrtho: a program pipeline for genome-wide identification of transcription factor target genes using phylogenetic footprinting. BMC Bioinformatics 5:27. doi:10.1186/1471-2105-5-27

Boudsocq M, Droillard MJ, Barbier-Brygoo H, Lauriere C (2007) Different phosphorylation mechanisms are involved in the activation of sucrose non-fermenting 1 related protein kinases 2 by osmotic stresses and abscisic acid. Plant Mol Biol 63:491–503. doi:10.1007/s11103-006-9103-1

Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G (2004) GO: TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 20:3710–3715. doi:10.1093/bioinformatics/bth456

Brown RL, Kazan K, McGrath KC, Maclean DJ, Manners JM (2003) A role for the GCC-box in jasmonate-mediated activation of the PDF1.2 gene of Arabidopsis. Plant Physiol 132:1020–1032. doi:10.1104/pp.102.017814

Burges CJC (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 2:121–167. doi:10.1023/A:1009715923555

Chae MJ, Lee JS, Nam MH, Cho K, Hong JY, Yi SA, Suh SC, Yoon IS (2007) A rice dehydration-inducible SNF1-related protein kinase 2 phosphorylates an abscisic acid responsive element-binding factor and associates with ABA signaling. Plant Mol Biol 63:151–169. doi:10.1007/s11103-006-9079-x

Chan BY, Kibler D (2005) Using hexamers to predict cis-regulatory motifs in Drosophila. BMC Bioinformatics 6:262. doi:10.1186/1471-2105-6-262

Chen JQ, Dong Y, Wang YJ, Liu Q, Zhang JS, Chen SY (2003) An AP2/EREBP-type transcription-factor gene from rice is cold-inducible and encodes a nuclear-localized protein. Theor Appl Genet 107:972–979. doi:10.1007/s00122-003-1346-5

Chitteti BR, Peng Z (2007) Proteome and phosphoproteome differential expression under salinity stress in rice (*Oryza sativa*) roots. J Proteome Res 6:1718–1727

Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14:1188–1190. doi:10.1101/gr.849004

D'Angelo C, Weinl S, Batistic O, Pandey GK, Cheong YH, Schultke S, Albrecht V, Ehlert B, Schulz B, Harter K, Luan S, Bock R, Kudla J (2006) Alternative complex formation of the Ca-regulated protein kinase CIPK1 controls abscisic acid-dependent and independent stress responses in Arabidopsis. Plant J 48:857–872. doi:10.1111/j.1365-313X.2006.02921.x

Dai X, Xu Y, Ma Q, Xu W, Wang T, Xue Y, Chong K (2007) Overexpression of an R1R2R3 MYB gene, OsMYB3R-2, increases tolerance to freezing, drought, and salt stress in transgenic Arabidopsis. Plant Physiol 143:1739–1751. doi:10.1104/pp.106.094532

Dieterich C, Herwig R, Vingron M (2003) Exploring potential target genes of signaling pathways by predicting conserved transcription factor binding sites. Bioinformatics 19(Suppl 2):II50–II56. doi:10.1093/bioinformatics/btg1059

Fowler S, Thomashow MF (2002) Arabidopsis transcriptome profiling indicates that multiple regulatory pathways are activated during cold acclimation in addition to the CBF cold response pathway. Plant Cell 14:1675–1690. doi:10.1105/tpc.003483

Fujimoto SY, Ohta M, Usui A, Shinshi H, Ohme-Takagi M (2000) Arabidopsis ethylene-responsive element binding factors act as transcriptional activators or repressors of GCC box-mediated gene expression. Plant Cell 12:393–404

Gertz J, Fay JC, Cohen BA (2006) Phylogeny based discovery of regulatory elements. BMC Bioinformatics 7:266. doi:10.1186/1471-2105-7-266

GuhaThakurta D, Stormo GD (2001) Identifying target sites for cooperatively binding factors. Bioinformatics 17:608–621. doi:10.1093/bioinformatics/17.7.608

Hao D, Ohme-Takagi M, Sarai A (1998) Unique mode of GCC box recognition by the DNA-binding domain of ethylene-responsive element-binding factor (ERF domain) in plant. J Biol Chem 273:26857–26861. doi:10.1074/jbc.273.41.26857

Holstege FC, Clevers H (2006) Transcription factor target practice. Cell 124:21–23. doi:10.1016/j.cell.2005.12.026

Horsman S, Moorhouse MJ, de Jager VC, van der Spek P, Grosveld F, Strouboulis J, Katsantoni EZ (2006) TF Target Mapper: a BLAST search tool for the identification of transcription factor target genes. BMC Bioinformatics 7:120. doi:10.1186/1471-2105-7-120

Huang WL, Tung CW, Huang HL, Hwang SF, Ho SY (2007) ProLoc: prediction of protein subnuclear localization using SVM with automatic selection from physicochemical composition features. Biosystems 90:573–581

Japkowicz N (2000) The class imbalance problem: significance and strategies. Proceedings of the 2000 International Conference on Artificial Intelligence, pp 111–117

Jolly ER, Chin CS, Herskowitz I, Li H (2005) Genome-wide identification of the regulatory targets of a transcription factor using biochemical characterization and computational genomic analysis. BMC Bioinformatics 6:275. doi:10.1186/1471-2105-6-275

Kasukabe Y, He L, Nada K, Misawa S, Ihara I, Tachibana S (2004) Overexpression of spermidine synthase enhances tolerance to multiple environmental stresses and up-regulates the expression of various stress-regulated genes in transgenic *Arabidopsis* *thaliana*. Plant Cell Physiol 45:712–722. doi:10.1093/pcp/pch083

Kel AE, Kel-Margoulis OV, Farnham PJ, Bartley SM, Wingender E, Zhang MQ (2001) Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors. J Mol Biol 309:99–120. doi:10.1006/jmbi.2001.4650

Khan M, Takasaki H, Komatsu S (2005) Comprehensive phosphoproteome analysis in rice and identification of phosphoproteins responsive to different hormones/stresses. J Proteome Res 4:1592–1599. doi:10.1021/pr0501160

Kiegerl S, Cardinale F, Siligan C, Gross A, Baudouin E, Liwosz A, Eklof S, Till S, Bogre L, Hirt H, Meskiene I (2000) SIMKK, a mitogen-activated protein kinase (MAPK) kinase, is a specific activator of the salt stress-induced MAPK, SIMK. Plant Cell 12:2247–2258

Krivan W, Wasserman WW (2001) A predictive model for regulatory sequences directing liver-specific transcription. Genome Res 11:1559–1566. doi:10.1101/gr.180601

Li XP, Tian AG, Luo GZ, Gong ZZ, Zhang JS, Chen SY (2005) Soybean DRE-binding transcription factors that are responsive to abiotic stresses. Theor Appl Genet 110:1355–1362. doi:10.1007/s00122-004-1867-6

Maruyama K, Sakuma Y, Kasuga M, Ito Y, Seki M, Goda H, Shimada Y, Yoshida S, Shinozaki K, Yamaguchi-Shinozaki K (2004) Identification of cold-inducible downstream genes of the *Arabidopsis* DREB1A/CBF3 transcriptional factor using two microarray systems. Plant J 38:982–993. doi:10.1111/j.1365-313X.2004.02100.x

Matsushima R, Hayashi Y, Kondo M, Shimada T, Nishimura M, Hara-Nishimura I (2002) An endoplasmic reticulum-derived structure that is induced under stress conditions in Arabidopsis. Plant Physiol 130:1807–1814. doi:10.1104/pp.009464

Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31:374–378. doi:10.1093/nar/gkg108

Ogul H, Mumcuoglu EU (2006) SVM-based detection of distant protein structural relationships using pairwise probabilistic suffix trees. Comput Biol Chem 30:292–299. doi:10.1016/j.compbiolchem.2006.05.001

Ohme-Takagi M, Shinshi H (1995) Ethylene-inducible DNA binding proteins that interact with an ethylene-responsive element. Plant Cell 7:173–182

Pabo CO, Sauer RT (1992) Transcription factors: structural families and principles of DNA recognition. Annu Rev Biochem 61:1053–1095. doi:10.1146/annurev.bi.61.070192.005201

Phan J, Moffitt R, Dale J, Petros J, Young A, Wang M (2005) Improvement of SVM algorithm for microarray analysis using intelligent parameter selection. Conf Proc IEEE Eng Med Biol Soc 5:4838–4841

Prak S, Hem S, Boudet J, Viennois G, Sommerer N, Rossignol M, Maurel C, Santoni V (2008) Multiple phosphorylations in the C-terminal tail of plant plasma membrane aquaporins. Role in subcellular trafficking of AtPIP2;1 in response to salt stress. Mol Cell Proteomics 7:1019–1030. doi:10.1074/mcp.M700566-MCP200

Qian J, Lin J, Luscombe NM, Yu H, Gerstein M (2003) Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. Bioinformatics 19:1917–1926. doi:10.1093/bioinformatics/btg347

Qian J, Esumi N, Chen Y, Wang Q, Chowers I, Zack DJ (2005) Identification of regulatory targets of tissue-specific transcription factors: application to retina-specific gene regulation. Nucleic Acids Res 33:3479–3491. doi:10.1093/nar/gki658

Rebeiz M, Reeves NL, Posakony JW (2002) SCORE: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data Site clustering over random expectation. Proc Natl Acad Sci USA 99:9888–9893. doi:10.1073/pnas.152320899

Redestig H, Weicht D, Selbig J, Hannah MA (2007) Transcription factor target prediction using multiple short expression time series from *Arabidopsis thaliana*. BMC Bioinformatics 8:454. doi:10.1186/1471-2105-8-454

Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR, Creelman R, Pilgrim M, Broun P, Zhang JZ, Ghandehari D, Sherman BK, Yu G (2000) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. Science 290:2105–2110. doi:10.1126/science.290.5499.2105

Sakuma Y, Maruyama K, Osakabe Y, Qin F, Seki M, Shinozaki K, Yamaguchi-Shinozaki K (2006) Functional analysis of an Arabidopsis transcription factor, DREB2A, involved in drought-responsive gene expression. Plant Cell 18:1292–1309. doi:10.1105/tpc.105.035881

Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res 32:D91–D94. doi:10.1093/nar/gkh012

Schölkopf B, Smola AJ (2002) Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT Press, Cambridge, MA

Schölkopf B, Burges CJC, Smola AJ (1999) Advances in kernel methods: support vector learning. MIT Press, Cambridge, MA

Schweighofer A, Meskiene I (2008) Regulation of stress hormones jasmonates and ethylene by MAPK pathways in plants. Mol Biosyst 4:799–803. doi:10.1039/b718578m

Seki M, Narusaka M, Abe H, Kasuga M, Yamaguchi-Shinozaki K, Carninci P, Hayashizaki Y, Shinozaki K (2001) Monitoring the expression pattern of 1300 Arabidopsis genes under drought and cold stresses by using a full-length cDNA microarray. Plant Cell 13:61–72

Seki M, Narusaka M, Ishida J, Nanjo T, Fujita M, Oono Y, Kamiya A, Nakajima M, Enju A, Sakurai T, Satou M, Akiyama K, Taji T, Yamaguchi-Shinozaki K, Carninci P, Kawai J, Hayashizaki Y, Shinozaki K (2002) Monitoring the expression profiles of 7000 Arabidopsis genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. Plant J 31:279–292. doi:10.1046/j.1365-313X.2002.01359.x

Song CP, Agarwal M, Ohta M, Guo Y, Halfter U, Wang P, Zhu JK (2005) Role of an Arabidopsis AP2/EREBP-type transcriptional repressor in abscisic acid and drought stress responses. Plant Cell 17:2384–2396. doi:10.1105/tpc.105.033043

Taji T, Ohsumi C, Iuchi S, Seki M, Kasuga M, Kobayashi M, Yamaguchi-Shinozaki K, Shinozaki K (2002) Important roles of drought- and cold-inducible genes for galactinol synthase in stress tolerance in *Arabidopsis thaliana*. Plant J 29:417–426. doi:10.1046/j.0960-7412.2001.01227.x

Towsey MW, Gordon JJ, Hogan JM (2006) The prediction of bacterial transcription start sites using SVMs. Int J Neural Syst 16:363–370. doi:10.1142/S0129065706000767

Vapnik V (1998) Statistical learning theory. Wiley, New York

Vavouri T, Elgar G (2005) Prediction of cis-regulatory elements using binding site matrices—the successes, the failures and the reasons for both. Curr Opin Genet Dev 15:395–402. doi:10.1016/j.gde.2005.05.002

Wang JW, Yang FP, Chen XQ, Liang RQ, Zhang LQ, Geng DM, Zhang XD, Song YZ, Zhang GS (2006) Induced expression of DREB transcriptional factor and study on its physiological effects of drought tolerance in transgenic wheat. Yi Chuan Xue Bao 33:468–476

Wang M, Zhang Y, Wang J, Wu X, Guo X (2007) A novel MAP kinase gene in cotton (*Gossypium hirsutum* L.), GhMAPK, is involved in response to diverse environmental stresses. J Biochem Mol Biol 40:325–332

Wasserman WW, Fickett JW (1998) Identification of regulatory regions which confer muscle-specific gene expression. J Mol Biol 278:167–181. doi:10.1006/jmbi.1998.1700

Wee LJ, Tan TW, Ranganathan S (2006) SVM-based prediction of caspase substrate cleavage sites. BMC Bioinformatics 7(Suppl 5):S14. doi:10.1186/1471-2105-7-S5-S14

Xiong Y, Fei SZ (2006) Functional and phylogenetic analysis of a DREB/CBF-like gene in perennial ryegrass (*Lolium perenne* L.). Planta 224:878–888. doi:10.1007/s00425-006-0273-5

Yamaguchi-Shinozaki K, Shinozaki K (1994) A novel cis-acting element in an Arabidopsis gene is involved in responsiveness to drought, low-temperature, or high-salt stress. Plant Cell 6:251–264

Yamaguchi K, Lee SH, Kim JS, Wimalasena J, Kitajima S, Baek SJ (2006) Activating transcription factor 3 and early growth response 1 are the novel targets of LY294002 in a phosphatidylinositol 3-kinase-independent pathway. Cancer Res 66:2376–2384. doi:10.1158/0008-5472.CAN-05-1987

Yamamoto YY, Ichida H, Matsui M, Obokata J, Sakurai T, Satou M, Seki M, Shinozaki K, Abe T (2007) Identification of plant promoter constituents by analysis of local distribution of short sequences. BMC Genomics 8:67. doi:10.1186/1471-2164-8-67

Young RA (2000) Biomedical discovery with DNA arrays. Cell 102:9–15. doi:10.1016/S0092-8674(00)00005-2

Yu GX, Ostrouchov G, Geist A, Samatova NF (2003) An SVM-based algorithm for identification of photosynthesis-specific genome features. Proc IEEE Comput Soc Bioinformatics Conf 2:235–243

Zhang W, Ruan J, Ho TH, You Y, Yu T, Quatrano RS (2005) Cis-regulatory element based targeted gene finding: genome-wide identification of abscisic acid- and abiotic stress-responsive genes in *Arabidopsis thaliana*. Bioinformatics 21:3074–3081. doi:10.1093/bioinformatics/bti490