An enumeration-based putative dyad predicting algorithm for promoter analysis in plants

Doctoral (Ph.D.) thesis

Mátyás Cserháti

Doctoral School of Biology

Biological Research Center (Hungarian Academy of Sciences), Institute of Plant Biology

SZTE TTIK

Supervisors: Dr. Sándor Pongor and Dr. János Györgyey

Szeged

2011

"…replenish the earth, and subdue it: and have dominion over the fish of the sea, and over the fowl of the air, and over every living thing that moveth upon the earth. And God said, Behold, I have given you every herb bearing seed, which is upon the face of all the earth, and every tree, in the which is the fruit of a tree yielding seed; to you it shall be for meat." (Genesis 1:28-29)

Table of contents

1. List of abbreviations

ABA: abscisic acid
ABF: ABA binding factor
ABRE: ABA Responsive Element
AGRIS: Arabidopsis Gene Regulatory Information Service
AKR: aldo-keto reductase
AP2: Apetala2
ATAF: Arabidopsis thaliana transcription factor
AREB: ABA Responsive Element-Binding protein
ARF: Auxin Responsive Factor
AS-1: Activation Sequence 1
AUC: area under curve
AuxRR: Auxin-Responsive Region
BLAST: Basic Local Alignment Search Tool
bZIP: basic Leucine Zipper Domain
CAAT: CAAT box element
CBF: C-repeat Binding Factor
cdr: cumulative difference ratio
COR: Cold Responsive
CRT: C-repeat
CUC: Copper Chaperonin
DRE: Dehydration Responsive Element
DREB: Dehydration Responsive Element Binding Transcription Factor
EEC: Enhancer Element Consensus
EIRE: Elicitor Responsive Element
ERD: Early Responsive to Dehydration
ERE: Elicitor Responsive Element
EREBP: Ethylene Responsive Element Binding Protein
FN: false negative
FP: false positive
FPR: false positive rate
GEO: Gene Expression Omnibus
GO: Gene Ontology
GT-1: Trihelix transcription Factor
HMG: High Mobility Group box
HMMER: Hidden Markov Model program
HSE: Heat Shock Element
ICE: Inducer of CBF Expression
IP: Inositole Phosphate
IUPAC: International Union of Pure and Applied Chemistry
JERE: Jasmonate and Elicitor Responsive Element
KIN: Antigenic Determinant of recA Protein Homolog
MADS: MCM1, AGAMOUS, DEFICIENS, SRF protein
MYB: myeloblastosis viral oncogene
MYC: myelocytomatosis viral oncogene

MYBRS: MYB Recognition Sequence
MYCRS: MYC Recognition Sequence
NAC: NAM, ATAF, CUC binding genes
NAM: No Apical Meristem gene
NCBI: National Center for Biotechnology Information
NCED: 9-cis-epoxycarotenoid dioxygenase
NLS: nuclear localization signal
PC: performance coefficient
PLACE: Plant cis-acting regulatory DNA elements database
PPV: positive predictive value
PSWM: position specific weight matrix
PR: Pathogen Response
RD: responsive to dehydration
REP: regulatory element pair
ROC: receiver operating characteristic, relative operating characteristic
ROS: reactive oxygen species
RSR: root specific region
RYRE: RY-repeat
TAIR: The Arabidopsis Information Resource
TATA: TATA box element
TF: transcription factor
TFBS: transcription factor binding site
tfpl: TRANSFAC/PLACE
TIGR: The Institute for Genome Research
TN: true negative
TP: true positive
TPR: true positive rate
TSS: transcription start site
UTR: untranslated region
WAR: Wounding Activating Region
WUN: Wound-Responsive Element
YMF: Yeast Motif Finder

2. Introduction

In understanding how plants adapt to suboptimal environmental conditions (cold, salt, drought and osmotic stress), it is of great importance to uncover those genetic elements which regulate the expression of different genes which take part in stress response. This could be useful in the case of a number of agriculturally important crop species. In such a way certain genetic modifications could then be made in them making them more resistant to stress.

As a possible tool for achieving this, we have developed an enumeration-based algorithm in order to find putative stress response elements. Since these genetic regulatory pathways are co-regulated and are interconnected to each other, in many cases multiple regulatory elements act in concert with each other to bring about stress response. Therefore our algorithm entails finding putative dyad elements, or pairs of motifs belonging to a specific regulatory network. Afterwards using the set of defined abiotic stress responsive dyad elements we can find other abiotic stress-responsive genes through a promoterome search to find promoters which also contained such elements.

The algorithm was first defined and tested in *Arabidopsis thaliana*, and then applied to the promoterome of rice (*Oryza sativa*). The algorithm was also applied in two special cases to the promoter analysis of a set of wheat aldo-keto reductase genes studied by our colleagues Zoltán Turóczy and Gábor V. Horváth as well as to the Oryza sativa promoter orthologues of a set of glucanase, chitinase and pathogen-responsive genes studied by our colleagues Vera Pós and Noémi Lukács. The algorithm can be also used to discover putative regulatory elements and new genes in other sets of co-regulated genes, and has its own online website which can be used to find dyad elements in a set of input promoter sequences according to a number of parameters.

3. Literature overview

3.1. Abiotic stress in plants

Besides biotic stress (viruses, bacteria, fungi, insects and herbivores), different kinds of abiotic stress also play a major role in decreasing crop yield (McGloughlin, 2010). The most common abiotic stress factors are cold, salinity, drought, osmotic, oxidative, anaerobic, heat, radiation, chemical, wind, flooding stress, and nutrient deprivation. Only the first four types of abiotic stress are discussed in the present thesis.

Abiotic stress in plants in general is characterized by an altered physiological state which induces a molecular, genetic, and biochemical response in order to either adapt to the new environmental conditions, or to return the plant's to its normal physiological state. What these types of abiotic stresses have in common with each other is that they all have something to do with the reduction of water potential within the cell, either by the formation of ice crystals during cold stress, or by ionic imbalance during salinity and osmotic stress, or by the withdrawal of water altogether during drought stress. During abiotic stress, injury of the plasma membrane is common, also marked by changes in its fluidity.

During the onset of abiotic stress (the process of which can be seen in Figure 1), the stress factor is transmitted by receptors, ion channels, or different kinds of kinases through the plasma membrane into the cell (Tuteja, 2007). Afterwards a number of secondary messengers and regulatory molecules such as Ca^{2+}, ABA, reactive oxygen species (ROS), and inositol phosphates (IP's) activate different kinds of signal transduction pathways, which in turn activate different kinds of transcription factors, which have a combinatorial effect on either activating or deactivating a number of different genes, which then in turn respond to abiotic stress. Sugars also can act as hormones, thereby regulating plant growth and metabolism. Therefore changes in their levels, detected by the protein hexokinase (HXK), play an important role in signaling the onset of abiotic stress (mainly cold, drought, and salt stress). One of the main kinds of sugar metabolites are sucrose and hexoses, which are produced in large quantities in

plants. Plant hormones such as ABA and ethylene are also induced by the onset of abiotic stress, and are also involved in sensing the level of sugars (Rosa, 2009).

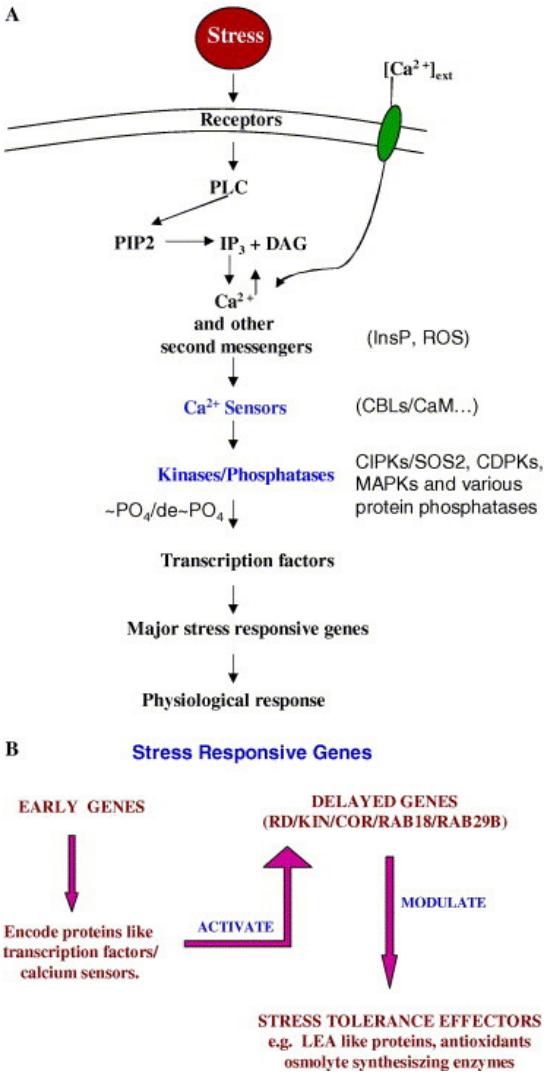


Figure 1. General scheme of signal transduction in abiotic stress in plants (figure taken from Mahajan, 2005)
Abiotic stress signals are transduced by receptors through the plasma membrane, which then activate secondary messengers such as Ca^{2+}, ROS, and other molecules. This in turn activates a phosphorylation cascade through the phosphorylation and dephosphorylation of kinase and phosphatase enzymes, which activate transcription factors, which are responsible for activating genes involved in stress response, which can be categorized as early activated genes which are expressed transiently, or lately activated genes which exhibit a sustained expression level.

Many of the genes as well as transcription factor binding sites take part in response to various forms of abiotic stress; therefore there is a considerable overlap, or

crosstalk between their regulatory pathways. Since abiotic stress has a fundamental effect on the homeostasis of the plant, it follows that a high number of genes are involved in stress response in order to repair the damage done by it.

The reason that the regulation of abiotic stress in plants at the molecular genetic level is so complex is because plants themselves are routinely subjected to a wide variety of environmental stresses. Because of this there is a large overlap between regulatory networks. Plants respond to stress both at the cellular and organism level. The complexity of response to abiotic stress is made even more so because the different structural parts of the plant have to communicate with each other in order to react to stress (for example drought may act as a stimulus on the leaf of the plant, which may induce increased water uptake in the roots).

Abiotic stress response takes one of two forms. First, stress response may be induced early on, within a few minutes by a set of master genes, which regulate the expression of other genes. Permanent stress response is induced by other sets of genes, the goal of which is to protect the plant against stress in the long run, for example enzymes needed for synthesis, proteins needed for the stabilization of the plasma membrane, as well as proteins which synthesize osmolytes and antioxidants, chaperones, proteases, and detoxification enzymes (Zhu, 2002; Dudits, 2006). Examples of such genes are the late embryogenesis abundant genes, or LEA (Late Embryonic Abundant) genes, which play a very important role in abiotic stress response. In the following, we will discuss different forms of abiotic stress in plants.

3.1.1. Cold stress

In plants, cold stress can have severe effects on the physiology of the plant, such as retarding development or forming ice crystals in the cell plasma, which put a strain on the cell wall leading to rupture. Cold stress is also associated with changes in the fluidity of the plasma membrane, leading to wilting and chlorosis. Cold-sensitive plants are characterized by a higher number of saturated fatty acids compared to unsaturated fatty

acids, which have a lower transition temperature, thereby stabilizing the plasma membrane.

Calcium is an important messenger in low temperature response, which regulates a number of specific cold-regulated genes such as COR and KIN genes. Another way that plants respond to cold is by the synthesis of a number of solutes, mainly different sugars, such as sucrose, fructose, trehalose, or fructans, as well as poly-alcohols. The role of these solutes is to stabilize the plasma membrane by maintaining the proper osmotic potential as well as the hydrophilic interactions with the membrane lipids and proteins.

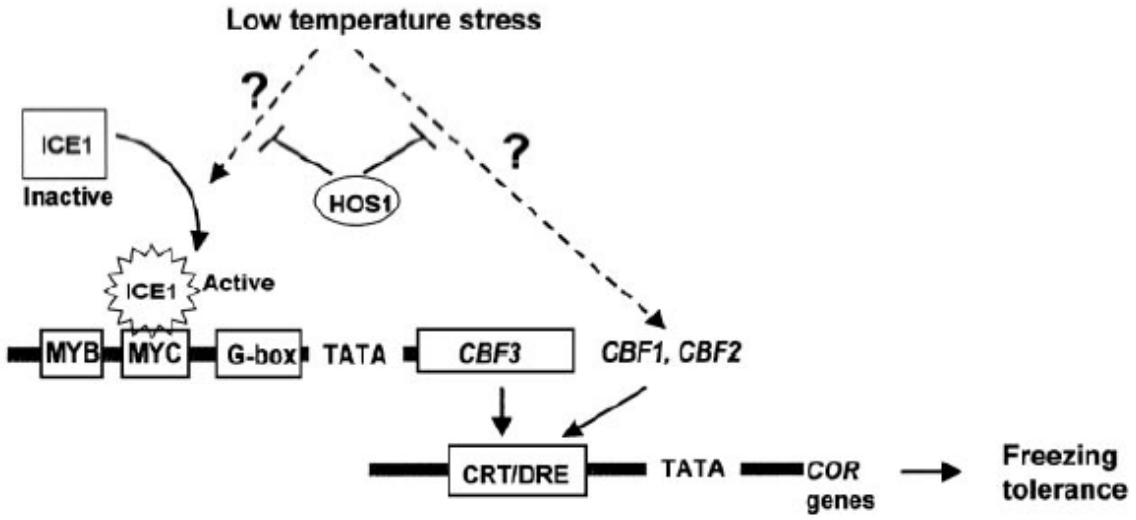


Figure 2. Molecular genetic pathways involved in cold-response in plants (figure taken from Chinnusamy, 2004).
Low temperatures activate the TF ICE1 (Inducer of CBF Expression 1), which is usually inactive under normal conditions. ICE1 then binds to MYB and MYC TFBS's in CBF genes, which then activate COR genes which take part in cold response.

A plethora of genes involved in cold stress response are genes such as COR (cold responsive), KIN, ERD (early responsive to dehydration), LTI (low-temperature-induced) (Wang, 1995), CBF (C-repeat binding factor), ICE (inducer of CBF expression) (Zarka, 2003), DREB (DRE-binding factor), NCED (9-cis-epoxycarotenoid dioxygenase), and RD (responsive to dehydration) genes (Thomashow, 1999). Some of the transcription factor binding sites (TFBS's) involved in cold response are the ABRE motif and the

DRE/CRT element, which all show the significant overlap with dehydration response (Yamaguchi-Shinozaki, 2005).

A general overview of the cold-response regulatory machinery in plants can be seen in Figure 2.

3.1.2. Salt and osmotic stress

In salt stress, the natural balance of the ions K^+, Na^+, Ca^{2+}, and H^+ is disrupted, as these ions play a major role in the homeostasis of the cell. High levels of Na^+ cause osmotic imbalance, and has a deleterious effect on enzymes as well as photosynthesis, whereas K^+ as an enzyme cofactor plays an important role in stomatal opening and closure. As usual, Ca^{2+} plays an important role in what is called the SOS (salt overly sensitive) pathway by binding to the EF motifs of a number of calcium binding proteins (such as calmodulin).

Several genes were discovered by Liu et al. (1998) which take part in this pathway, such as the NHX (Na^+/H^+ exchanger), HKT (hydroxykynurenine transaminase), and Na^+/H^+ antiporter genes, the latter of which function as ionic pumps to pump Na^+ either out of the cell or into the vacuole.

An osmolyte called glycine-betaine produced during salt stress in many plants confers resistance to abiotic stress (Mahajan, 2005; Dudits, 2003). Besides this, certain genes are synthesized which take part in protein and DNA stabilization like certain DNA helicases, such as the PDH proteins (pyruvate dehydrogenase). Important regulatory cascade proteins such as the MAP kinases: MAPK (Mitogen-acivated Protein Kinase), MAPKK (Mitogen-acivated Protein Kinase Kinase), and MAPKKK (Mitogen-acivated Protein Kinase Kinase Kinase) have also been shown to respond to salt and osmotic stress (Yoo, 2008).

The different kinds of TF's and TFBS's which take part in salt and osmotic stress response can be seen in Figure 3.

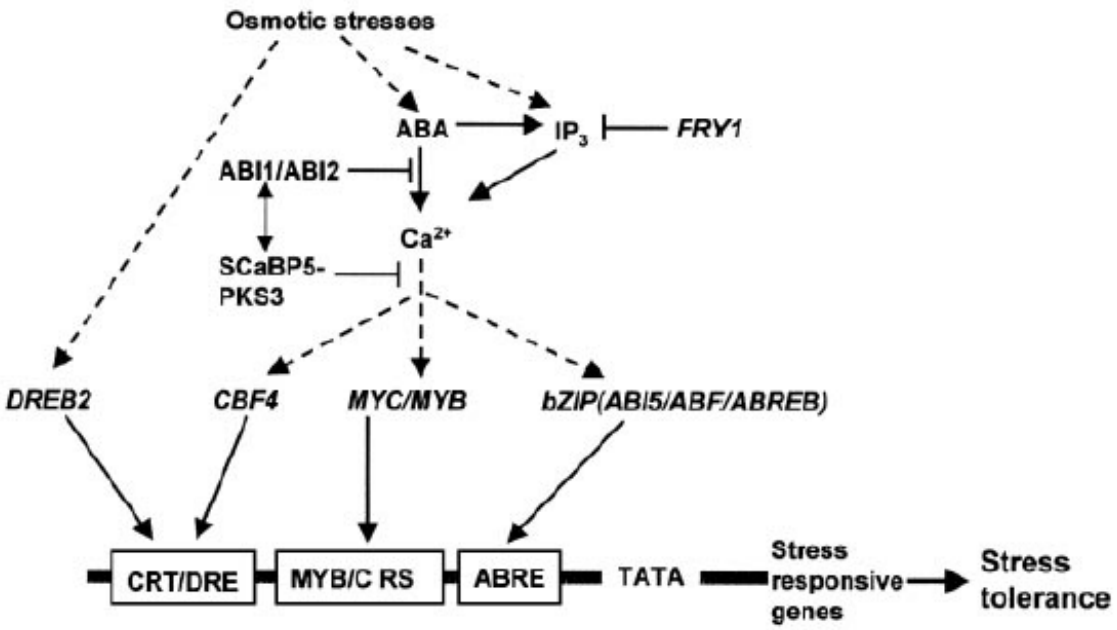


Figure 3 Molecular genetic regulation of the response to salt and osmotic stress in plants (figure taken from Chinnusamy, 2004)
Osmotic stress induces DREB factors in an ABA-independent way. These in turn bind to CRT/DRE elements in different stress responsive genes. Ca^{2+} and IP_3 regulate abiotic stress response in an ABA-dependent manner, by activating the TF's CBF, MYC/MYB, and a number of bZIP TF's. These in turn bind to CRT/DRE, MYC/MYB binding sites and ABRE elements.

3.1.3. Drought stress

The main source of stress during drought is due to the rise in concentration of cellular osmolytes. This leads to injury within the cell, for example in the photosynthetic apparatus, which leads to the generation of different kinds of ROS's. Other effects of drought lead to slower cell division and therefore reduction in vegetative growth, as well as stomatal closure. This fact points to extensive crosstalk between salt stress and drought stress.

Plants respond to drought stress in a number of different ways. For example, a number of hormones such as ABA, ethylene, and cytokinins are produced which induce plant growth and stomatal closure at the right concentration due to change in the pH level within the cell. Certain enzymes are also induced which respond to ROS's, and others which synthesize osmolytes such as sugars or different kinds of amino acids (such as

proline) which serve to increase the water potential within the cell in order for the cell to regain its normal shape.

Genes induced by drought stress in plants include AREB, ABF, MYB, MYC, and DREB, RD (responsive to drought) (Fujita, 2004), dehydrin genes, and the NCED gene, which is important for ABA production. The regulation of drought responsive genes has a significant overlap with cold-induced genes. It is interesting to note that since some of those genes involved in cold and drought stress all take part in the repair of the plasma membrane (such as lipid transfer genes), they also have an effect on enhancing resistance to bacterial pathogens (Hong, 2006).

These genes are mostly regulated through the ABA-dependent pathway. Common TFBS's which play a role in this pathway include DRE, ABRE, MYCRS and MYBRS (Shinozaki, 2003). The promoters of such genes therefore naturally contain copies of the ABRE and DRE elements (Narusaka, 2003). It is interesting to note that the ABRE is not activated by ABA in all cases. For example, the Oryza sativa *pwsi18* gene is induced by water deficit only (Joshee, 1998).

3.2. Regulatory pathways and elements involved in abiotic stress response

Response to abiotic stress is regulated by two different pathways. These include an ABA-dependant and an ABA-independant pathway (Shinozaki, 2003; Yamaguchi-Shinozaki, 2005), both of which, however show somewhat of an overlap with each other, and which play a major role in response to cold, drought and salinity stress (Yamaguchi-Shinozaki, 2006). One of the main mediators between the two pathways is Ca^{2+}. Some genes, such as RD29 can be activated by both pathways, and some metabolites such as proline also accumulate in both ways. Figure 4 depicts the basic regulatory elements which make up the ABA-depenedent and ABA-independant pathways involved in abiotic stress regulation.

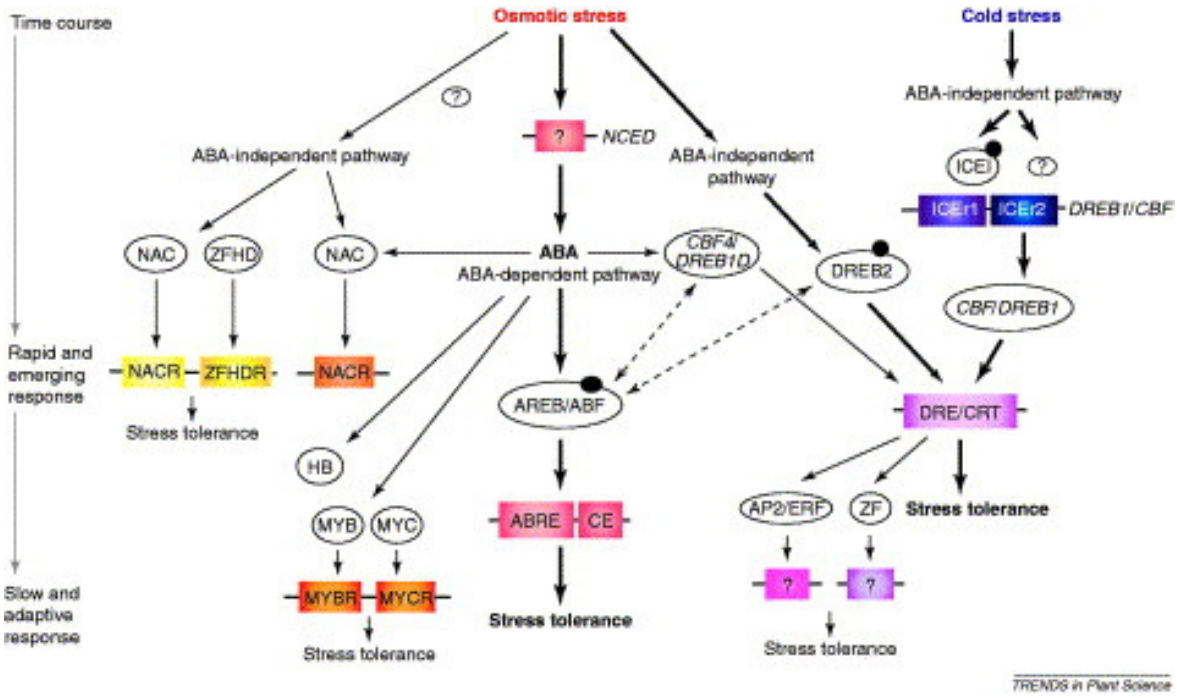


Figure 4. ABA-dependent and -independent regulatory pathways involved in response to cold and osmotic stress (figure taken from Yamaguchi-Shinozaki, 2005)
TF's regulating stress-responsive genes are depicted as ellipses, while genes involved in abiotic stress response are depicted in colored boxes. Small black circles depict modifications in TF's due to stress signals. Early expressed genes are depicted above lately expressed genes.

It is interesting to note that there is also an overlap with genes involved in seed development, and that certain stress conditions may even induce somatic embryogenesis (Fehér, 2005-6; Cooper, 2003). The reason for this may probably be that the plant hormone ABA regulates the plant's water status as well as germination. Seed development itself is basically a dessication process (Tuteja, 2007).

3.2.1. The ABA-dependant regulatory pathway

Both osmotic and drought stress are regulated in an ABA-dependent manner. ABA is such an important phytohormone that it is used in many plant experiments to mimic abiotic stress. Needless to say that the ABRE element is very common in the promoters of genes regulated by this molecule, which usually binds the AREB leucine zipper TF. Genes regulated by the ABA-dependent pathway include RD29A, RD22, COR15A, COR47, and P5CS (1-pyrroline-5-carboxylate synthetase) (Yamaguchi-Shinozaki, 2005; Silva-Ortega, 2008).

ABA itself is produced from β–carotene within the cell in reaction to abiotic stress cues, inducing some of the enzymes which take part in its production, such as ZEP (zeaxanthin epoxidase), NCED (9-cis-epoxycarotenoid dioxygenase), AAO (ABA-aldehyde oxidase), and MCSU (molybdenium cofactor sulfurase) (Tuteja, 2007).

3.2.2. The ABA-independant regulatory pathway

To a large degree cold exerts its effect in an ABA-independent manner as well as drought to a smaller degree, although ABA also effects genes induced by cold (Zhu, 2002). The latter was proven by how the ERD1 gene (which is homologous to a regulatory subunit of the Clp protease) was activated in under 1 hour after the onset of dehydration in Arabidopsis by ZH-HD and NAC prior to the accumulation of ABA in plants (Nakashima, 1997; Chinnusamy, 2004). ABA-independent stress genes are regulated many times by DREB elements such as DREB2A and DREB2B, which bind to the DRE element in the genes' promoter.

3.2.3. Regulatory elements in abiotic stress response

There are many different kinds of TFBS's which take part in abiotic stress response, namely the ABA responsive element (ABRE), characterized by the sequence ACGTGKC, and which is similar to the G-box found in other ABA-induced promoters. The drought responsive element (DRE) is characterized by the sequence RCCGAC, the C-repeat elements (CRT), the low temperature responsive element (LTRE) characterized by the sequence GGCCGACGT (Jiang, 1996), and the MYC and MYB recognition sites (MYBRS and MYCRS), characterized by the sequences TGGTTAG and CACATG (Abe, 1997; Abe, 2003). Many of these TFBS's are conserved among orthologous, paralogous, and co-regulated genes (Tran, 2010). Many of these motifs often act in concert with other TFBS's in order to have an effect on abiotic stress response, such as the DRE/CRT element. Some TFBS's occur within a specific distance from one another, thereby forming dyad elements. This is the case in the ABRE element in Arabidopsis and Oryza sativa as studied by Gómez-Porras et al (Gómez-Porras, 2007). ABRE elements often occur in multiple copies, thereby having a quantitative effect on stress response. ABRE

and DRE elements are also known to occur within different kinds of promoters induced by cold, drought, and salt stress.

Many abiotic stress motifs exert their effect on the basal transcription machinery, while others lie farther upstream within the promoter. Regulatory elements in general form diverse regulatory networks each having an effect on one another (Wray, 2003). As a part of this, certain TF's are induced by abiotic stress themselves. Some of these plant stress TF's include the DREB, WRKY, MYB, bHLH (basic helix-loop-helix), bZIP, and NAC TF families. Protein-protein interactions between different TF's also take part in abiotic stress response, such as bZIP factors during ABRE-mediated cold-regulated gene expression (Jakoby, 2002). Another example is the SCOF-1 protein (soybean cold-inducible factor-1) which interacts with SGBF-1 (soybean G-box binding bZIP transcription factor), which is a bZIP protein in soybean in response to cold (Kim, 2001). A list of well-known plant TFBS's involved in abiotic stress and their specific TF can be seen in Table 1.

***cis*-element**	**sequence**	**transcription factor**	**stress condition**
ABRE	YACGTGKC	bZIP	water deficit, ABA
CE1	TGCCACCGG	ERF/AP2	ABA
CE3	ACGCGTGCCTC	-	ABA
CRT	GGCCGACAT	ERF/AP2	Cold
DRE	TACCGACAT	ERF/AP2	water deficit, cold
ICEr1	GGACACATGTCAGA	-	Cold
ICEr2	ACTCCG	-	Cold
LTRE	GGCCGACGT	ERF/AP2	Cold
MYBR	TGGTTAG	MYB	water deficit, ABA
MYCR	CACATG	bHLH	water deficit, ABA
NACR	ACACGCATGT	NAC	water deficit
ZFHDR	-	ZFHD	water deficit

Table 1. Well known cis-regulatory elements and the transcription factors that bind to them (taken from Yamaguchi-Shinozaki, 2005)

3.3. Promoter and regulatory element databases and regulatory motif discovery programs

What sets regulatory motif discovery programs and algorithms apart from each other is the way the basic model of the binding between the surface of the DNA and TF protein is implemented. In these algorithms an objective functional measure is devised which seperates functional motifs from background motifs with no function. Obviously a good algorithm will be able to distinctly seperate and highly rank true binding motifs from false ones (Li, 2006). The following is a list of features of an algorithm which may influence the way it is tested and used (Li, 2006):

- The size of the dataset
- The length of motifs in the sequences
- The median length of a sequence in the dataset
- The number and density of binding sites in the dataset
- The different number of kinds of binding sites in the dataset
- The fraction of sequences not containing a binding site
- The uniformity of binding sites in the dataset

Program	URL	Type	Reference
Bioprospector	http://bioprospector.stanford.edu/	Probabalistic	Liu, 2001
ConSite	http://www.phylofoot.org	Phylogenetic	Lenhard, 2003
CompMoby	http://genome.ucsf.edu/compmoby/	Probabalistic	Chaivorapol, 2008
Dyadscan	http://bhd.szbk.u-szeged.hu/dyadscan/	Enumeration	Cserháti, 2011
FootPrinter	http://bio.cs.washington.edu/software.html	Phylogenetic	Blanchette, 2003
Gibbs motif sampling	http://bayesweb.wadsworth.org/gibbs/gibbs.html	Probabalistic	Lawrence, 1993
HMMER	http://hmmer.janelia.org/	Probabalistic	Eddy, 2008
MEME	http://meme.sdsc.edu/meme4_3_0/intro.html	Probabalistic	Bailey, 1995
MotifSampler	http://www.esat.kuleuven.ac.be/~thijs/Work/MotifSampler.html	Probabalistic	Thijs, 2001
PatSearch	http://www.pesolelab.it//index.php	Probabalistic	Pesole, 2000
PhyloNet	http://bioinfo.cs.rice.edu/phylonet/	Phylogenetic	Than, 2008
PhyloScan	http://bayesweb.wadsworth.org/cgi-bin/phylo_web.pl	Phylogenetic	Carmack, 2007
W-AlignAce	http://www1.spms.ntu.edu.sg/~chenxin/W-AlignACE/	Probabalistic	Chen, 2008
YMF	http://bio.cs.washington.edu/software.html	Enumeration	Sinha, 2003

Table 2. List of several motif discovery programs

In order to evaluate motif discovery programs, a number of parameters are calculated to measure their effectiveness. Such basic parameters are the true positive rate (TP), which is the ratio of correctly predicted motifs, while the false positive rate (FP) is the ratio of non-motifs which were falsely predicted as true motifs. The true negative rate (TN) is the ratio of non-motifs which were correctly predicted as such, and not detected by the program, while the false negative rate (FN) is the ratio of true motifs which were falsely predicted as non-motifs.

Another common parameter which is calculated is the sensitivity of the given algorithm, which is defined as such: Sens = TP/(TP+FN), which quantifies the proportion of true motifs found compared to non-motifs. The specifity of an algorithm is defined as the ratio of true negatives to all non-motifs, that is Spec = TN/(TN+FP). The positive predictive value is a parameter which quantifies the proportion of true motif sites amongst all predicted ones: PPV = TP/(TP+FP). Finally, the performance coefficient PC = TP/(TP+FP+FN), which summarizes the sensitivity and the positive predictive value (Defrance, 2009).

A study performed by the designers of 13 well-known motif discovery algorithms led by Martin Tompa however found that according to absolute measures of correctness, site sensitivity of these algorithms was at most 0.22, meaning that motif discovery algorithms are far from being perfect and have room for improvement (Tompa, 2005). Indeed this might be because of the many false positives found by such algorithms as not all found sites (according to one calculation, one in every thousand) have an *in vivo* function (Wasserman, 2004). This also means that when a researcher wishes to find regulatory elements within promoter sequences, it would be advisable to use more than one motif discovery programs; ones of different types, which can complement each other's results in order to get the best results.

3.3.1. Promoter structure in plants

Promoter structure in plants can be defined in a number of different ways for *in silico* analysis because of the complex relationship between its different parts and

components binding to it. Whereas the promoter can be defined as that part of the gene which contains all regulatory elements which are needed for its regulation, in practice the promoter sequence is usually defined as the regulatory sequence 500 bp to 5 Kbp upstream from the ATG start. This includes the UTR, which however in some cases is difficult to define, because of multiple ATG start signals found in it. In theory, it is possible to take as large an upstream section as possible, but doing so risks increasing the number of non-functional sequences which introduces noise into the analysis (Das, 2007). The complexity of eukaryotic regulatory elements and networks compared to prokaryotic ones makes the task of extracting promoters even more difficult (Bulyk, 2003). Promoter sequences must be shortened if necessary if they reach into another gene farther upstream. In general, plant promoters tend to have a GC-compositional strand bias (Rombauts, 2003; Fujimori, 2005), as well as having a larger A/T ratio than animals (Szafranski, 2005).

The promoter can also be subdivided into the core, proximal, and distal promoter. The core promoter is where the pre-initiation transcriptional complex forms and actually starts transscription, and is usually the stretch of DNA around 100 bp long before the transcription start site (TSS). The proximal promoter is about 1000 bp long, and is where more specific regulation takes place, and is the part of the promoter where TFBS's accumulate. The distal promoter is that part of the promoter which is usually around 3000 bp long, where enhancer and silencer elements exert their effect (Lichtenberg, 2009). In many cases, intron sequences, especially the sequence of the first intron are also taken into account since such sequences also contain regulatory elements, which have albeit different roles than TFBS's.

Walther et al. have shown in a study of *Arabidopsis* that promoters of upstream genes taking part in physiological processes tend to have larger promoters and a larger density of TFBS's within them, since such genes integrate regulatory signals coming from outside the cell. Such genes themselves encode TF proteins. Therefore because of the increased density of TFBS's in such promoters, TFBS interactions also tend to increase as part of a complex regulatory network. In contrast, downstream genes tend to have shorter promoters and less regulatory elements as their main role in a gene cascade

or biochemical pathway is to produce a protein or enzyme with a specific non-regulatory function (Gómez-Porras, 2007). The TATA-box was found by these researchers in the promoters of many genes involved in abiotic stress response (Walther, 2007). *Therefore we can assume that the number of interactions between TFBS's in abiotic stress promoters is quite common.* Indeed, Yu et al. and Vardhanabhuti et al. discovered separately in yeast and vertebrate promoters that many TFBS's with similar functions occur at a given distance from one another (Yu, 2006; Vardhabanuti, 2007).

3.3.2. On-line promoter databases

A number of on-line integrated, cross-referenced, and annotated databases exist which contain different kinds of upstream sequences which can be used in plant promoter analysis as well as databases which contain information on individual TFBS's. Such databases also include resources for promoter visualization. These databases have become more complete as more genomes are sequenced and more genes get annotated. Up to now, promoter databases have been designed for *Arabidopsis thaliana* and *Oryza sativa*, the first two plant organisms whose genome sequence have been determined. Several smaller but less complete databases also exist for other plants integrated into one single database.

One such database is the Eukaryotic Promoter Database, which contains a non-redundant collection of experimentally defined Pol II promoters (Schmid, 2006). PlantProm is a similar database which contains similar information for plants, as well as position specific weight matrixes (PSWM's) for a number of TFBS's (Shahmarudov, 2003). A PWSM is a matrix containing weight values for the individual positions along a regulatory motif which it represents used to score individual occurrences of the given motif. It has 4 rows and as many columns for each position in the motif. The individual weights represent the weight of a given base appearing at a given position. One quite useful database which holds promoter sequence data for chordates and different plant species is the DoOP database (Database of Orthologous Promoters) (Barta, 2005), which has also been supplemented with a search tool which is capable of finding putative common conserved regulatory motifs (Sebestyén, 2009). Promoter databases for

Arabidopsis and Oryza sativa include ppdb (Yoshiharu, 2007), Athena (O'Connor, 2005), AthaMap (Steffens, 2004), Osiris (Morris, 2008), and the TIGR Rice Genome Annotation Resource (Ouyang, 2006). Athena is a very useful Arabidopsis promoter database, which contains 30,067 predicted Arabidopsis promoter sequences and 105 TFBS's, which can all be visualized in selected promoter sequences.

3.3.3. Transcription factors and cis-regulatory elements in plants

Domain type	Structural characteristics
AP2/EREBP	A 68-amino acid region with a conserved domain that constitutes a putative amphiphatic α-helix
ARF	A 350 amino acid region similar to B3 in sequence
AT-hook motif	A consensus core sequence R(G/P)RGRP with the RGR region contacting the minor groove of A/T-rich DNA
B3	A 120 amino acid conserved sequence at the C-termini of VP1 and ABI3
bZIP	A basic region and a leucine-rich zipper-like motif
HMG-box	L-shaped domain consisting of three α-helices with an angle of about 80° between the arms
Homeodomain	Approximately 60 amino acid residues producing either three or four α-helices and an N-terminal arm
MADS	Approximately 57 amino acid residues that comprise a long α-helix and two β-strands
Myb-related	A basic region with one to three imperfect repeats each forming a helix-helix-turn-helix
Myc b/HLH	A cluster of basic amino acid residues adjacent to a helix-loop-helix motif
Trihelix	Basic, acidic, and proline/glutamine-rich motif which forms a trihelix DNA-binding motif
Zinc finger	Finger motif(s) each maintained by cysteine and/or histidine residues organized around a zinc ion

Table 3. Structures of well-known transcription factor families (adapted largely from Liu, 1999)

Besides general TF's, specific TF's contain several domains which are necessary for their function. These are the DNA-binding region, oligomerization site(s), a transcription regulatory domain, and a nuclear localization signal. Many of these domains are highly conserved because of their function, and also constitute different TF protein families, all characterized by a specific amino acid sequence signature (Liu, 1999).

The secondary structure of the DNA-binding domain is responsible for their selectivity and affinity. Oligomerization domains also convey specificity to the binding of the DNA motif, are quite conserved, with characteristic secondary structures, such as coiled coils, β-sheets, and helix-loop-helix structures. In comparison with the previous

two types of TF domains, the transcriptional regulation domains are highly divergant because of their specific regulatory functions which happen through conformational changes.

Transcriptional regulation takes place in two basic forms, repression or activation. Repression takes place mainly either through competetive binding of factors to a given TFBS, or by binding to another TF, thereby making it incapable of binding DNA. Nuclear localization signals (NLS's) determine the nucleolar localization of a given TF within the cell, and differ in sequence, number, and organization. Plant NLS's are many times enriched in arginine and lysine. A list of the most common TF families in plants can be seen in Table 3.

TFBS's in plants are short stretches of sequence usually around 5-10 bp long (Solovyev, 2010), and can be described by a consensus sequence defined by IUPAC symbols (depicted in Table 4). Such motif sequences are ambiguous many times, because the surface of a TF may not come in contact with the surface of the DNA at that position. Some sequences have spacer regions between well-defined parts of the sequence, thereby forming dyad sequences, as in the case of the Gal4 binding site ($CGGN_{11}CCG$). Yu et al. for example found that in yeast, 75% of all interacting TFBS's which co-occur with each other lay within 166 bp from each other (Yu, 2006). Regulatory complexes can also form at the 3' end of the gene as well as in the 5' UTR and first intron.

3.3.4. Topographic characteristics of the binding relationship between different families of plant TF's and TFBS's

An important factor in promoter regulation is the way TF's bind to their corresponding TFBS's. About 2000 genes encode transcription factors in Arabidopsis, representing around 5-10% of the genome (5-7% for plants in general) (Grotewold, 2008; Mitsuda, 2008). This is more compared to other eukaryotes such as Drosophila melanogaster. TF's are categorized into roughly 50-60 families. TF's are further capable of combinatorically regulating the genome through for example, alternative splicing, post-translational modification, phosphorylation, or miRNA interaction.

TF's usually contain a DNA binding domain (DBD). Those which don't have such a domain can integrate with transcription complexes through protein-protein interactions. DBD's are usually highly conserved, since they come into contact with the DNA, at sites with a specific sequence. These are usually conserved stretches 6-8 bp long. The surface between the TF and the DNA forms a specific „lock and key" structure which is responsible for specific regulation of the gene whose promoter the TFBS is found in.

TF's also contain a regulation domain, which can act as an activator or repressor. Activator domains are many times rich in proline and glutamine. Repressors can act either passively or actively. The former possess neither an activation domain (AD) nor a repression domain (RD), while active repressors have an RD (Mitsuda, 2008). TF's many times also contain interaction domains with which other TF's can come into contact with them. TF's also very commonly formk form dimers or tetrameric complexes through dimerization domains. Such TF's include leucine zippers, zinc fingers, or MADS proteins. Because of their combinatorical regulation capabilities, dimerization domains are not very conserved as compared to DBD's. Dimerization can occur in two ways: either first in solution, and then binding to the DNA, or afterwards (Amoutzias, 2008).

3.3.4.1. The Dof protein family

For example, the Dof (DNA binding with One Finger) family is a common group of Cys2/Cys2 zinc fingers divided into six groups which play a large number of different roles in the cell, including stress response, and mediating DNA-protein and protein-protein interactions. The Dof domain is a conserved stretch of 50-52 amino acids at the N-terminus of the protein, which binds to a common core sequence of AAAG. Serine stretches link it to the transcriptional regulation domain at the C-terminus (Kushwaha, 2010).

3.3.4.2. The bHLH protein family Myc-like bHLH TF's form one of the largest families of plant TF's. The 190 amino acid long N-terminus is comprised of two domains, an interaction domain, which is rich in acidic amino acids, and an activation

domain (see figure 5). The interaction domain binds Myb-like TF's. The bHLH proteins are capable of binding to the E-box, whose consensus sequence is CANNTG. The bHLH region of the protein is capable of homo- and heterodimerization. The activation domain forms a protein-binding surface which interacts with the RNA polymerase II machinery, which initiates transcription. It is through the interaction of other TF's with the interaction domain which can modify the way bHLH proteins induce transcription as can be seen in Figure 5 (Pattanaik, 2008).

3.3.4.4. The bZIP protein family

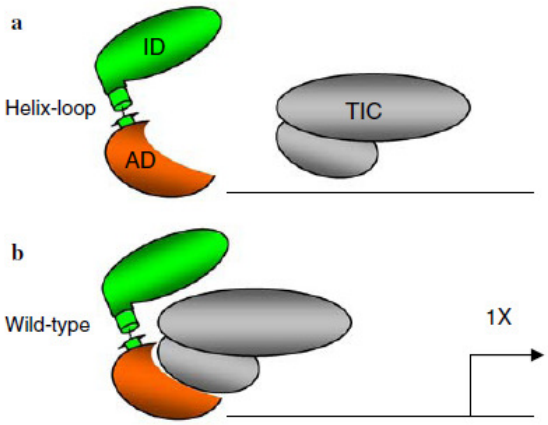


Figure 5. Interaction of the activation domain (AD) with the transcription initiation complex (TIC) in bHLH TF's. ID: interaction domain (figure taken from Pattanaik, 2008).

Basic region/leucine zipper (bZIP) proteins are involved in pathogen defense and stress signalling, with 75 identified putative members in Arabidopsis. The bZIP domain is made up of two parts. First, following a 16 amino acid NLS is an invariant N-x7-R/K motif, which comes in contact with the DNA. Afterwards comes a heptad repeat of leucines at the C-terminus. bZIP proteins are capable of forming homo- and heterodimers, where two subunits bind to each other forming a coiled-coil structure with their hydrophobic surfaces contacting each other. Plant bZIPs preferentially bind to the A-, C-, and G-boxes (TACGTA, GACGTC, and CACGTG) (Jakoby, 2002). A pair of bZIP subunits binding the DNA can be seen in Figure 6.

3.3.4.5. The ARF protein family

Auxin response factors (ARF) and auxin/indole acetic acid (Aux/IAA) proteins take part in the regulation of auxin response genes, which are capable of dimerizing with each other, although additional proteins might also interact with them. 23 and 25 ARFs have been found in Arabidopsis and rice, respectively. 29 and 31 Aux/IAA proteins exist in Arabidopsis and rice. ARF proteins are made up of a DBD, a non-conserved middle region (MR), and a C-terminal dimerization domain (CTD) (see Figure 7). The DBD binds to the auxin

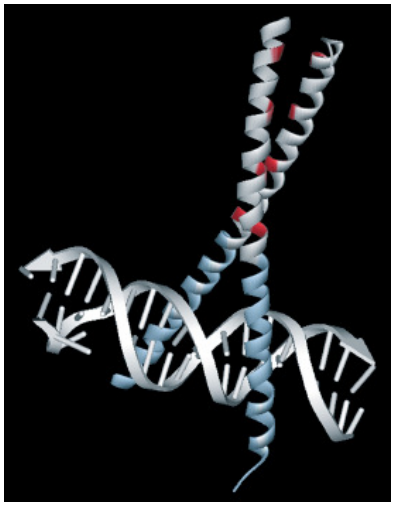


Figure 6. Three-dimensional model of two leucine zipper subunits binding to DNA forming a coiled-coil structure (figure taken from Jakoby, 2002)

response element (AuxRE) which has a sequence of TGTCTC. The CTD is capable of forming homo- and heterodimers. Based on the composition of the MR, ARF proteins are capable of acting either as acivators or repressors. Activators are enriched in glutamine, serine, and leucine, whereas repressors are enriched in serine, leucine, proline, and glycine (Shen, 2010).

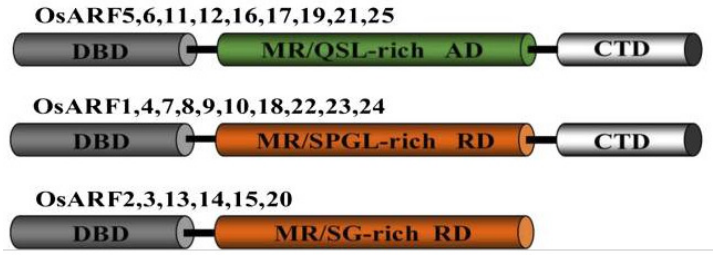


Figure 7. Domain representation of three ARF subfamilies in rice

3.3.4.6. The MADS protein family

MADS proteins are well-known for their involvement in the regulation of floral development, and can be divided into three main classes (A, B, and C). MADS proteins are made up of a highly conserved MADS domain, an I-region, and a K-box (see Figure 8a). The K-region folds into 3 amphipatic alpha helixes (K1, K2, and K3), which are responsible for dimerization (see Figure 8b). The I-region is responsible for partner selection in dimerization.. Dimerization itself is needed for other TF's or accesory factors such as chaperones to interact with MADS proteins, such as bHLH and bZIP proteins. Dimerization is very strongly conserved in these proteins As seen in Figure 8c, MADS proteins bend the DNA 180 degrees, and thus form a quartet structure. Two interacting MADS dimers bind to the CArG motif on the DNA within close proximity to each other. The function of the quartet structure is not well-known (Immink, 2010).

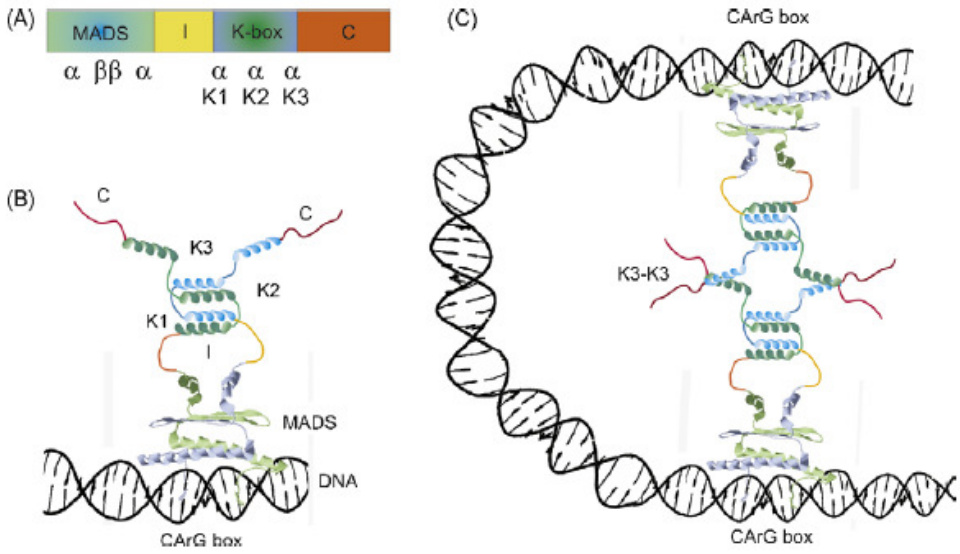


Figure 8. Structure of a MADS protein and physical model of DNA binding in MADS proteins (figure taken from Immink, 2010).

3.3.5. On-line TF and TFBS databases

In plants, two regulatory motif sequence databases are widely known; the PLACE database (Higo, 1999), and the PlantCARE database (Lescot, 2002). The PLACE database contains a number of cis-regulatory elements which can be searched by

matching sequences or by keywords. Also, individual TFBS's can be found within sequences supplied as input in the query interface at this database. The PlantCARE database also contains a number of plant TFBS's which can also be matched to these sequence.

Databases which include data for transcription factors and position weight matrixes include the TRANSFAC database (Matys, 2006), the PlnTFDB (plant transcription factor database) (Riano-Pachon, 2007; Pérez-Rodríguez, 2010), and JASPAR (Bryne, 2008). Amongst these the TRANSFAC database is a widely known database which contains experimentally defined information on TF's and their corresponding TFBS's in eukaryotes, as well as their PSWM's. Some well-known motif finding programs such as MatInd, MatInspector (Quandt, 1995), MATRIX SEARCH (Chen, 1995), SIGNAL SCAN (Prestridge, 1991), and rVISTA (Loots, 2002) have been developed to allow the user to match a number of well-known PSWM's to instances of a given motif within a user-provided input sequence. The TRANSCompel database contains information on composite regulatory elements as well as experimental results describing the cooperative action between them (Kel-Margoulis, 2002). The Transcription Regulatory Regions Database (TRRD) also contains much information on TFBS's, regulatory regions, and expression pattersn (Kolchanov, 2002).

letter	definition
M	A or C
R	A or G
W	A or T
S	C or G
Y	C or T
K	G or T
V	A, C, or G
H	A, C, or T
D	A, G, or T
B	C, G, or T
N	any base

Table 4. List of IUPAC letters and their definition

The JASPAR database contains a number of TFBS matrix models for a number of organisms. The AGRIS database (Palaniswamy, 2006) is an Arabidopsis-specific database which holds information on Arabidopsis-specific cis-regulatory elements and TF's. PlantPAN (Plant Promoter Analysis Navigator) (Chang, 2008) is an interesting plant promoter database, which implements the identification of cis-regulatory elements within a distance constraint, complete with interface visualization. The DRTF database (Database of Rice Transcription Factors) contains different kinds of information for 2,025

putative TF's in Oryza sativa divided into 63 different families. Such information includes sequence features, functional domains, as well as experimental data (Gao, 2006).

3.3.6. Enumeration-based methods

A number of motif discovery algorithms are based on enumerating all possible motif sequences within the input sequences, and then applying some sort of filter to distinguish biologically relevant motifs (or functional motifs) from irrelevant (functionless) ones. The advantage that enumeration-based algorithms have over other algorithms is that they perform an exhaustive search of all possible wordspace, that is, of all possible regulatory motifs. Furthermore, they also outperform heuristic methods in many cases (Tompa, 2005). The dyad finding algorithm presented in this thesis falls into this category.

In general, the longer the motif is, the more specific it is sequentially, and the easier it is to determine whether it is biologically relevant or not. However, the longer a motif is the more degenerate it can become, as well as being a lot scarcer, and may therefore not produce significant or robust statistics.

A filter can be applied in two basic ways, first, by counting all occurrences of the motif in real (functional or positive) promoter data sets, and then comparing their occurrence in either a number of randomized, randomly selected, or non-functional (negative) sequences. It is naturally expected that if the motif is biologically significant, then it will occur significantly more times in the positive set than in the negative set (Sinha, 2002).

This can be measured by a number of statistical measures. Secondly, one can apply a certain type of model based on which one can calculate the number of occurrences of a given motif, and then measure the difference between the number of observed occurrences and expected occurrences. Based on this difference one can then calculate a z-score based on which one can segregate relevant motifs from irrelevant ones (Rombauts, 2003). The z-score is calculated in the following way:

$$(1)\ z(m) = \frac{obs(m) - E(m)}{\sigma(m)}$$

Where obs(m) is equal to the number of observed occurrences of motif m, while E(m) is the number of expected occurrences of the motif based on the distribution model used, usually a binomial or Poisson distribution. σ(m) is the standard deviation for motif m used in the model employed. This measure takes into account the conservedness of a given motif, since it is present in a smaller number of forms, the fewer number of times it is expected to occur (Pavesi, 2004). An enumeration algorithm used by some researchers called the universal algorithm employs a similar z-score: $sign(w) = S \cdot \ln\left(\frac{S}{E_S}\right)$, where S is the number of sequences the word w occurs in, and E_S is the number of sequences the word is expected to occur in (Geisler, 2006).

Very many times the background genome base distribution is calculated for a given organism (in genes, promoters, or on the whole-genome level in general) in order to be able to discriminate relevant sequences from irrelevant ones. For example a Markov model of varying orders can be applied to calculate the prior probability of finding an oligomer motif based on the background base distribution (Ellrott, 2002). Based on this as well as the size of the input sequence, the number of expected oligonucleotide sequences can be then calculated, and then compared to the observed amount (van Helden, 2004). This method is implemented in the oligoscan and dyadscan programs of the RSAT tools (Regulatory Seqence Analysis Tools) designed by van Helden (van Helden, 2003; Thomas-Collier, 2008), which have been used to discover regulatory signals and cis-elements in different kinds of regulatory regions (Janky, 2007; Sand, 2007; Turatsinze, 2008; Defrance, 2008).

TFBS's found by enumeration methods can also be clustered together to form consensus sequences which may be used to characterize the TFBS. Based on the sequence of experimentally defined TFBS's one can also define PSWM's, which can then be used to scan the genomic sequence of a given organism in order to find newer instances of the TFBS.

3.3.7. Phylogenetic methods

Although enumeration-based motif discovery methods prove useful in sequence analysis, it does have its weaknesses. One such weakness is the computation time and memory needed to find motifs in an input sequence. This gets larger and larger the longer the motifs are. Another weakness (especially in the case of PSWM's) is that the underlying assumption is that each base position is independent from neighbouring bases, and does not take into account possible molecular interactions spanning over many bases (Gunewardena, 2008).

The basic assumption behind phylogenetic methods is that a functional motif will be conserved in the promoter sequences of orthologous genes in different kinds of species over time (Wassermann, 2004). This means that the surrounding sequences will be free to mutate, because of the lack of functional constraints on them, thereby leaving behind islands of functional motifs sometimes called "phylogenetic footprints" (Hannenhalli, 2008). However, this is not always so, because small differences in sequence, such as differences in the length of spacer regions between regulatory elements may cause functionally compensatory changes in the regulatory machinery (Bulyk, 2003).

Furthermore, whereas enumeration methods are used mainly to analyze single genomes, phylogenetic footprinting is useful in comparative genomic analysis whereby multiple species are analyzed (Kellis, 2003; Kechris, 2004). In this way, phylogenetic methods are capable of reducing the noise-to-signal ratio, and also increase the selectivity of TFBS's. A study by Lenhard et al. when developing the ConSite software showed that compared to analyses of single sequences, selectivity increased by 85% (Lenhard, 2003). However, one of the drawbacks of phylogenetic footprinting is that since it deals with alignments of conserved motifs, it does not allow for supposed binding site turnover.

Phylogenetic methods usually construct a global multiple alignment of the orthologous promoters used in the study, and then identify conserved regions within the alignment (Das, 2007). The conserved regions are assumed to contain conserved TFBS's. A common tool used in making the alignment is CLUSTALW.

Motifs can be best found by these methods by increasing the number of species involved in the analysis as well as using species which are related to each other at an appropriate distance. If the species are not related, then information may be eroded away altogether. Motifs in the promoters of species which are phylogenetically too closely related to each other will not be able to be seen clearly enough, therefore selecting an outlier species is always important. However, a technique called phylogenetic shadowing makes it possible to discover conservation patterns in more closely related species (Boffelli, 2003). This method takes into account the phylogenetic relationship of the studied species, and thereby localizes regions of conserved and variable sequences. Furthermore, the longer the sequence is, the better these kinds of methods will be able in distinguishing them from random background sequences.

One such widely used phylogenetic program is the FootPrinter program of Blanchette and Tompa (Blanchette, 2003). As input it takes in a number of homologous promoter sequences from different species, and then runs the footprinting program according to a number of parameters set by the user (e.g. motif length, ambiguity, promoter region motifs are to be found in). As output, a graphic interface visualizes the spacing of the motifs along the promoter sequences. Other well-known phylogenetic algorithms include PHYLONET (Than, 2008) and PhyloScan (Carmack, 2007).

3.3.8. Co-occurence based methods

Co-occurrence based methods rely on the fact that TFBS's do not act alone but in networks, along with many other elements within the promoter. Many times they are embedded in tracts of sequences called regulatory modules or enhancers. Indeed, eukaryotes have been known to employ combinatorial strategies to generate a wide variety of expression patterns from a relatively small set of regulatory motifs (Nguyen, 2006). The underlying assumption here (which is also used in the algorithm presented in this thesis) is that since proteins which all take part in the same biochemical process or pathway are basically to a large degree localized in the same cell compartment, and that since many of these genes are active at just about the same time, then this means that all

or most of the genes should be regulated by common regulatory factors, and therefore most or all contain similar regulatory cis-elements (Walhout, 2006).

Here the statistically significant co-occurrence of motifs with each other is what is used to define the regulation of a set of genes, which usually make up a genetic regulatory network. In this way, even those motifs can be detected which have a low occurrence, if they co-occur relatively many times along with another motif. The present thesis makes use of this fact in the definition of the regulatory elements implemented in the algorithm designed by the author.

These kinds of algorithms also make use of co-regulated sets of genes defined by chip and microarray experiments. For example, the program RiCES (Rice Cis-Element Searcher) uses likelihood statistics in order to discover significant relationships between pairs of motifs (Doi, 2008). A number of works have studied synergistic effects of pairs of motifs acting in concert with each other in order to increase the level of expression of genes that they co-occur in, compared to gene expression where only one of the motifs are present. These include studies of the ABRE and DRE elements (Zhang, 2005), and different kinds of yeast motifs, such as the RRPE (ribosomal RNA processing element) and PAC motifs (Puromycin N-acetyl-tranferase) (Pilpel, 2001; Sudarsanam, 2002). In such studies motif networks are visualized through a graph-like map where the number of edges between vertices representing individual motifs shows the strength of the motif interaction with each other.

3.3.9. Probabalistic methods

Probabalistic methods initially begin with an alignment matrix defined by a multiple alignment made up of different instances of a known regulatory motif. In this matrix each column represents the number of times each of the four bases occur at a given position in the motif being represented. The weight matrix can be calculated from the alignment matrix in a number of ways or determined experimentally (Tatusov, 1994; Hertz, 1999).

In the case of alignment and weight matrixes, the log-likelihood ratio is a statistic which measures the difference between the base distribution measured in the alignment and the assumed *a priori* base distribution. In this way, statistically significant motifs represented by such matrixes can be distinguished from random, background motifs. The log-likelihood ratio of a given motif m is given in Equation 2:

$$(2)\ \mathrm{LLR}(m) = \sum_{j=1}^{N}\sum_{i=1}^{4} n_{i,j} \ln \frac{p_i}{f_{i,j}}$$

Where $n_{i,j}$ denotes the number of occurrences of the i^{th} base at position j in the motif, p_i is the a priori probability of the i^{th} base, and $f_{i,j} = n_{i,j} / N$, where N is the number of bases in the motif (Hertz, 1999). The higher this value is, the more significant the motif is statistically.

Another interesting measure used in motif discovery is the information content of a given motif, as can be seen in Equation 3. Here $f_{i,j}$ and p_i are the same as in Equation 2:

$$(3)\ I(m) = \sum_{j=1}^{N}\sum_{i=1}^{4} f_{i,j} \ln \frac{f_{i,j}}{p_i}$$

The information content of a motif is in a sense, related to its thermodynamic content of the protein-DNA interaction between the regulatory motif and the TF. We can grasp intuitively that the greater $f_{i,j}$ differs from p_i (in theory $f_{i,j}$ is equal to the background base distribution meaning that $f_{i,j} = p_i$ meaning 0 information content), the greater its information content and therefore its statistical significance will be.

Also related to the log-likelihood ratio and information content is the p-value of any kind of statistic used in measuring the biological relevance of a given regulatory motif. The p-value is basically the probability of finding a sequence which has larger information content than a given alignment. The lower the p-value is, the more significant a given motif is. In another sense, the information content of a motif sequence profile also describes the overall specifity of a profile, and is measured in bits. Information content also depends upon the length of the motif involved (Lenhard, 2003).

The weight matrix is then optimized through a number of iterations aimed at optimizing the score of a motif at a given position in the genome. The model parameters are also estimated in many cases by maximum likelihood or Bayesian inference (Das, 2007). More advanced algorithms compensate for false motif discovery by comparing possible common motifs with the background base pair distribution (Rombauts, 2003). The motifs found by such algorithms can also be represented by a sequence logo, which signifies the size of a given base at a given position by its size. One of the first implementations of a weight matrix model to find TFBS's was a greedy algorithm called Consensus (Hertz, 1990) which was able to find instances of a motif in all input sequences, while it maximized the information content. It is important to note here that if the matrix is not determined specifically enough, this will generate many false positive TFBS hits.

The expectation-maximalization method (Lawrence, 1990) is a deterministic model which calculates the likelihood that a given motif of a given length belongs to a given motif model, and after each step re-estimates the motif model. However, this method is deterministic, and always assumes that there is one instance of the motif in all promoter sequences, although this may very well not be the case. Another of the first and most widely used probabilistic algorithms for motif discovery is MEME (Multiple Expectation maximization for Motif Elicitation), which does not require a background model, automatically estimates motif length, and allows absence of motifs in some input sequences (Bailey, 2006).

The Gibbs sampling method (first designed for motif detection in proteins but then modified a number of times for detection in DNA sequences) is a stochastic kind of probabilistic method which allows detected motif instances to be replaced with higher scoring ones in order to escape local optima. Because of its stochastic nature, such algorithms must be run many times in order to get reliable results. However, the more pronounced an optimal solution is, the higher the likelihood that Gibbs sampling methods will find them (Rombauts, 2003).

As a subfamily of probabilistic methods, a number of hybrid enumeration-probablistic methods exist, which entail a dictionary-like classification of all possible words (which correspond to regulatory motifs), even of different lengths within a set of input promoter sequences. These algorithms are capable of describing the whole regulatory landscape of a given set of promoters. One such algorithm is the MobyDick algorithm which starts out with the base distribution of a given set of sequences, and then builds up a dictionary of longer words formed out of shorter words, serving as suffixes. These words are also ranked according to a statistical significance score. This algorithm is so adept at finding regulatory sequences that it was also capable of finding several hundred English words in the first 10 chapters of the novel "Moby Dick" (Chaivorapol, 2008).

Overall, probabilistic methods and other non-enumeration based methods which use a heuristic principle to avoid having to analyse the entire motif space have the advantage that they can perform their analysis in less time. However, the trade-off is that they cannot be sure that they have found an optimal TFBS (Pavesi, 2004).

4. Objectives

The objective of this Ph.D. thesis was to develop an enumeration-based dyad prediction algorithm in order to find putative regulatory elements in the promoters of co-regulated genes. As mentioned previously we exploit the fact that many times common TF's regulate the expression of genes which take part in the same physiological process or biochemical pathway. In our case we studied regulatory motifs which take part in abiotic stress response in Arabidopsis, Oryza sativa, and Triticum aestivum.

The algorithm was validated in Arabidopsis and then used in Oryza sativa and Triticum aestivum, as well as in two other test cases in Oryza sativa aldo-keto reductase genes, in chitinase, glucanase, and pathogenesis-related gene promoters from Oryza sativa and Triticum aestivum. One of the main results of the algorithm is a set of putative dyad regulatory elements which can be then used in further studies. With this list of regulatory elements in hand, one can also then perform a promoterome search of a given organism in order to predict the involvement of other genes in abiotic/biotic stress whose promoter sequences contain similar regulatory elements. This may then facilitate the annotation of genes of unknown function.

5. Materials and methods

5.1. Dyad definition

In our algorithm we defined a dyad as a pair of oligonucleotide motifs with a spacer region between them with a well-defined characteristic length, as depicted in Figure 9. The dyad can also be represented by the formula $M_1N_nM_2$ where the first motif (M_1) is called the head motif, and the second one (M_2) the tail motif. For example, the dyad in Figure 9 can be depicted as such: ACGTC{N_x}TCAGG. Both motifs are of the same length, and were set to 5bp. This means $4^5 \times 4^5$=1,048,576 different kinds of dyads, all different in their head and tail motif sequences (since this is the total possible number of ways the four bases can be selected in two pentamer motifs). The spacer length was set to maximum 52 bp due to computational constraints (because the original version of the algorithm was tested on hexamer pairs which were represented in binary on a 64-bit computer system). This length covers five full turns in the DNA double helix, and allows an additional 2 bp of wobbling.

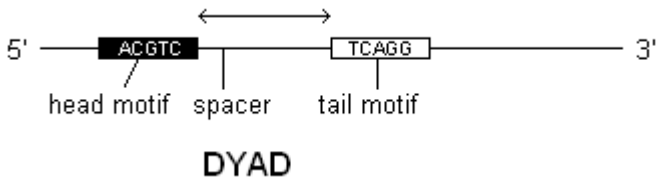


Figure 9. Schematic drawing of a dyad element analyzed in the algorithm. The dyad is made up of a head and tail motif and a spacer region of a defined length

Many motifs occur at different positions quite flexibly at long distances from each other. At longer distances between motifs a lot less free energy is needed to form a DNA loop between the motifs. Therefore at larger distances, the distance itself ceases to be an influencing factor upon the dynamics of the cooperation between the transcription factors binding to their individual DNA motifs (therefore motifs can occur at any distance relative to each other). Therefore our algorithm is specially tuned to identify motif pairs which are found closer together and therefore form a much stable transcription unit along with their respective transcription factors.

Our basic assumption was that motifs occur at a well-defined distance from each other within the proximal promoter because of functional constraints, and that such a distribution of motifs is non-random, thereby possibly forming a smaller part of a regulatory network. Conversely, if the two motifs occur at totally random positions from each other, then this is a sign that they do not associate with each other, and are therefore non-functional (Cserháti, 2006).

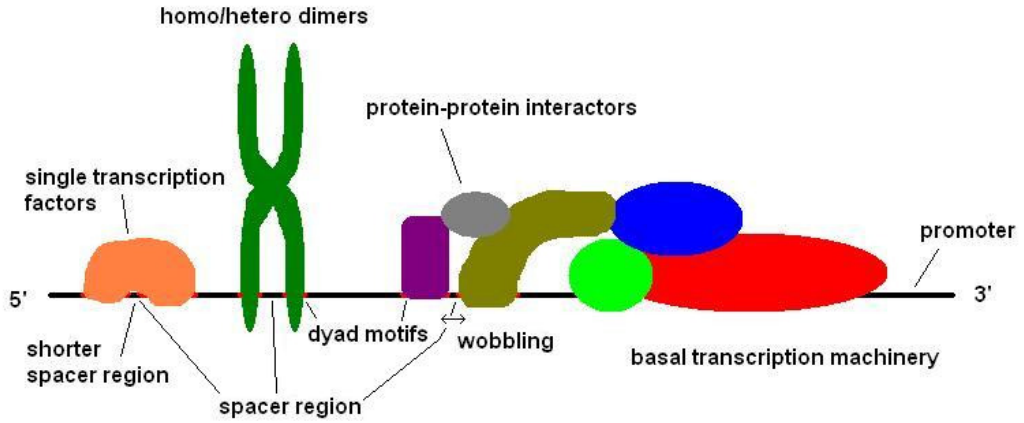


Figure 10. Basic kinds of interactions between dyad sequences (red) and transcription factors (different colored shapes) in the proximal promoter.

For example, some TF's may interact with the basal transcription machinery, either activating or repressing it. As seen in Figure 10, the individual head and tail motifs of some dyads may be the binding site for different kinds of single TF's. Other dyads may bind dimerized proteins. A whole class of transcription factors, the BZIP proteins, binds to sites on the DNA molecule which are separated from one another by a stretch of DNA of unspecified sequence. One such bZIP protein, OsAREB1 binds to the well known ABRE element (consensus sequence: (C/T)ACGTGGC) (Jin, 2010). Others include leucine zippers, which take part in abiotic stress response. It may be the case that a single TF may bind to both the head and tail motif of a dyad leaving a short undefined spacer region in between. Slight wobbling (a few bp) is allowed between the TF's binding the head and tail motif of the dyad.

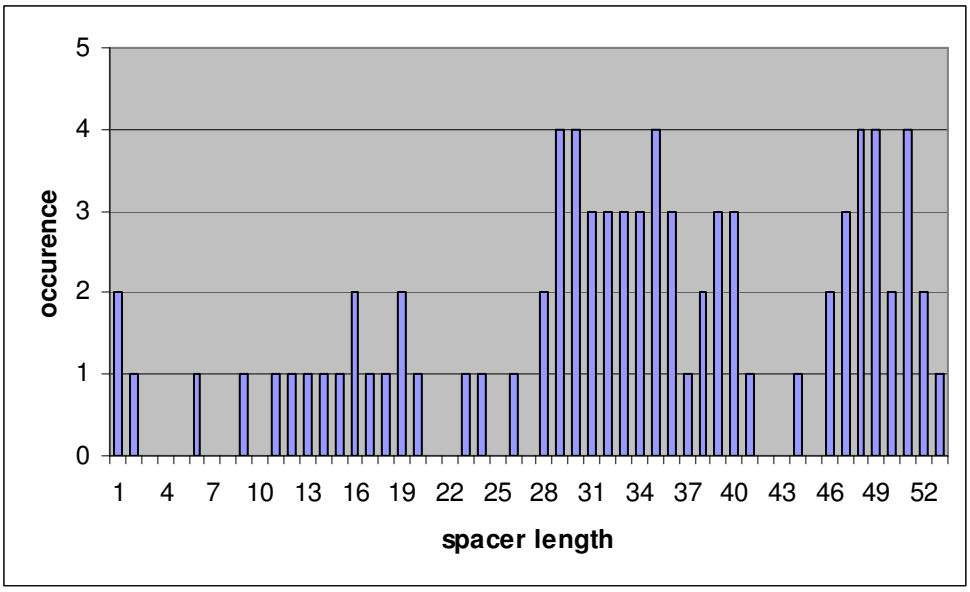
5.2. Calculation of dyad score

Mathematically, a given dyad's score (*cdr*, cumulative difference ratio) can be calculated by the following formula:

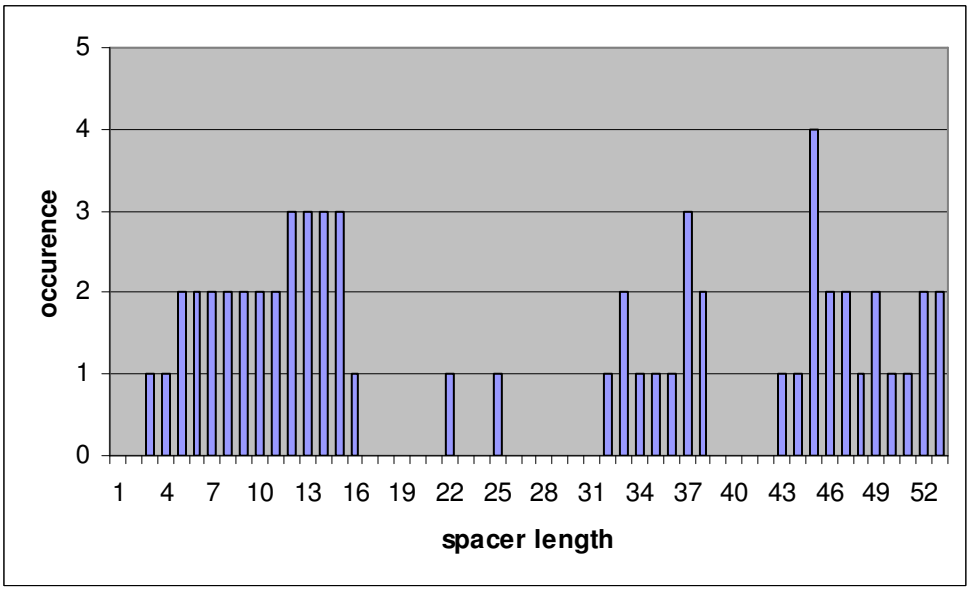
$$(4)\ cdr = \frac{N_{positive} - N_{negative}}{N_{positive}}$$

Here $N_{positive}$ is the number of promoters in the positive learning promoter set that the dyad occurs in, and $N_{negative}$ is the number of promoters in the negative learning promoter set that the dyad occurs in. Here the positive learning promoter set is the set of promoters which the algorithm is trained on. That is, it contains instances of the biologically relevant motifs we wish to „teach" the algorithm to discover. In contrast the negative learning promoter set contains relatively little or none of these motifs.

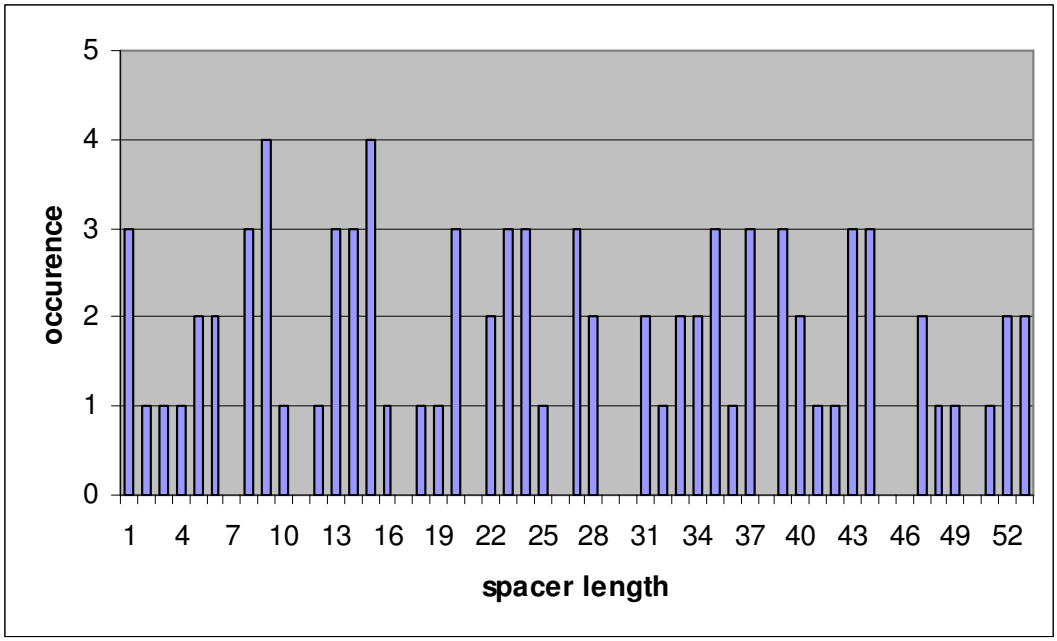
The *cdr* value ranges from $-\infty$ to 1, with a score of 1 meaning that the dyad was found only within the positive promoter set ($N_{negative} = 0$). Therefore, the higher the score, the more significant a dyad is. Only dyads with an occurence in the positive learning set were taken into consideration. In our approach we searched for all occurences of all possible dyad where each dyad was given a characteristic spacer length beforehand, meaning $53 \times 4^5 \times 4^5$=55,574,528 possible dyads in total. This approach was used in order to avoid picking up repetitive sequences. The larger the size of the learning promoter set with longer sequences, the larger the probability that all possible dyads will appear in it.



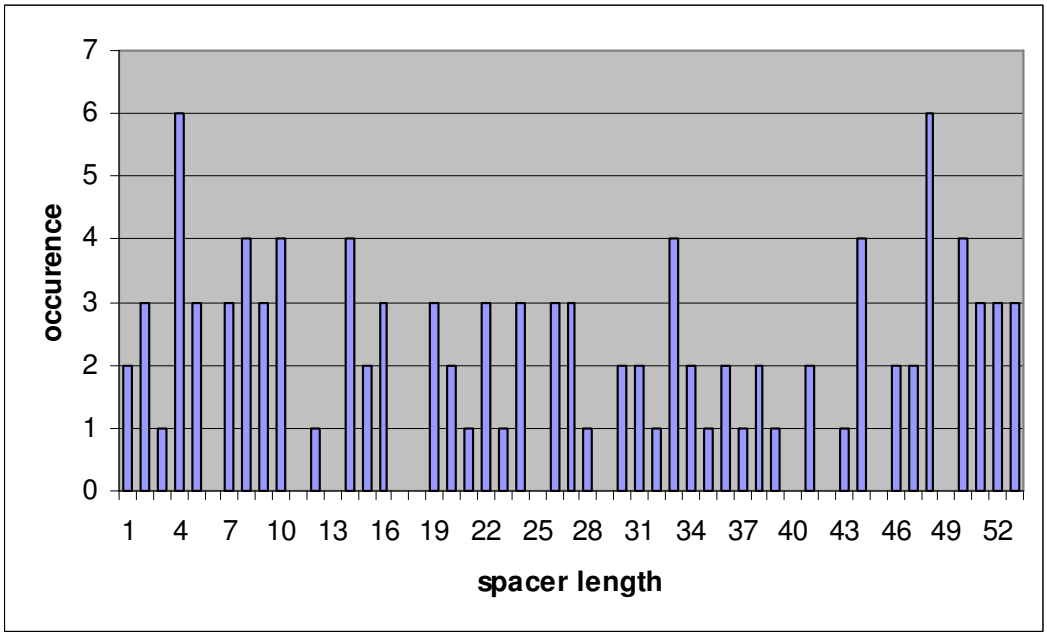
A. Distribution of the dyad ACGTG TTTTT in the stress learning promoter set



B. Distribution of the dyad ACGTG TTTTT in the nonstress learning promoter set



C. Distribution of the dyad ATGAT TTTAT in the stress learning promoter set



D. Distribution of the dyad ATGAT TTTAT in the nonstress learning promoter set

Figure 11. Example distributions of a biologically relevant and irrelevant dyad in the positive and negative learning promoter sets in Arabidopsis

For example in Figure 11, we can see an example of a biologically relevant dyad (ACGTG{N_n}TTTTT), and a biologically irrelevant one (ATGAT{N_n}TTTAT). The head motif of the biologically relevant dyad is ACGTG, which corresponds to the ABRE element, which is a well-known to take part in abiotic stress (Lee, 2010). In this case the positive set corresponded to abiotic stress promoters while the negative set corresponded to non-stress promoters. The first dyad occurs in 11 positive promoters, and in 0 negative promoters. Therefore, its *cdr* score is (11-0)/11 = 1.0. The second dyad occurs in 8 positive promoters, but in 11 negative promoters. Therefore its *cdr* score is (8-11)/8 = -0.375, which is very low, even below zero.

5.3. Calculation of promoter score

The *cdr* score thus calculated is used in the testing phase for calculating the score of a given promoter, given by the following equation:

$$(5)\ S_{promoter} = \sum_{i}^{N} n_i \cdot cdr_i$$

where N is the number of optimal dyads found after the test phase of the analysis, and n_i is the occurence of the i^{th} dyad, and cdr_i denotes its *cdr* score.

Similarly, the score for an individual promoter is

$$(6)\ S_{promoter} = \sum_{i}^{N} n_i \cdot cdr_i + \sum_{i}^{37} n_i \cdot cdr_i$$

where the sum of the *cdr* scores of the 37 TRANSFAC and PLACE stress motifs (described in section 5.5.1.) is also added to the promoter's score (if present) in the case where they were also included in the analysis.

5.4. Overview of algorithm

Figure 12 gives us an overview of the steps implemented in the algorithm.

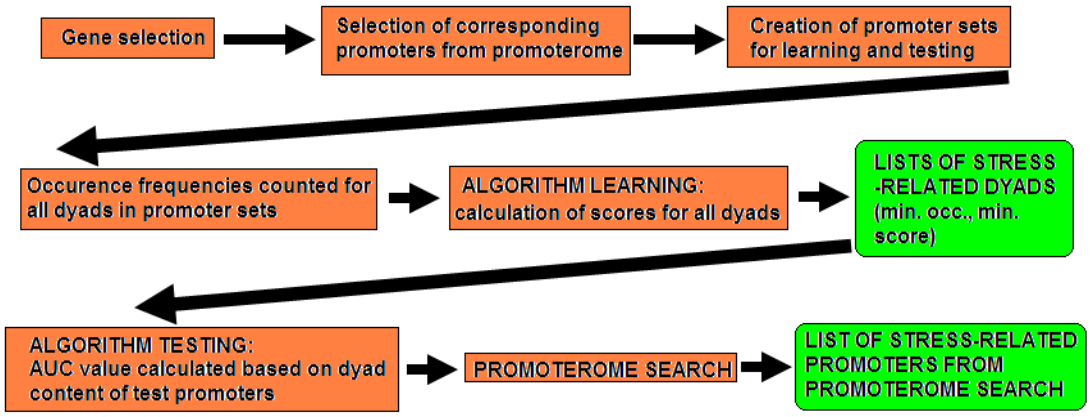


Figure 12. Flowchart of steps taken during the execution of the algorithm (green boxes denote result data from the algorithm)

The first phase of the algorithm is the retrieval of data sequences. First of all, the proper genes have to be selected whose promoter regions we wish to analyze. These are assumed to be co-regulated genes which may take part in the same physiological process or biochemical pathway (in our case abiotic stress, that is, cold, drought, salt and osmotic stress). Next, the promoter sequences of according length are extracted from the promoterome of the given algorithm. The promoters may be downloaded from the appropriate database, or extracted with a script.

Next, the promoters are partitioned into four different promoter sets: a positive learning set, a negative learning set, a positive test set, and a negative test set. The learning sets are used to train the algorithm. That is, the algorithm "learns" how to discriminate between functionally relevant and irrelevant regulatory elements. Therefore we expect functionally relative dyads to be found with greater occurrence in the positive learning set compared to the negative learning set. The dyads found in the learning phase of the algorithm are therefore expected to occur in the positive test set with greater frequency compared to the negative test set, if the algorithm has been properly trained. In other words, the positive and negative test promoter sets are defined as those promoter sets used to test the validity of the dyads found during the learning phase. Our experience shows that around 100 promoters which we used in the learning sets can be enough for appropriate statistical measurements.

The next phase of the algorithm is the learning phase. Since this is an enumeration method, the occurrence of each possible dyad must be counted in the positive and negative learning sets. Based on the frequency of each dyad one can calculate a *cdr* score for all the dyads and then rank them accordingly. These dyads will be used in different sets in the testing phase of the algorithm.

In the testing phase of the algorithm different sets of dyads are formed according to their minimum *cdr* score and the minimum number of promoters in the positive learning set they occur in are set as dynamic variables, which all correspond to a parametrically specified test run. Furthermore the number of test runs may be increased by allowing the head and tail sequences of the dyads to wobble up to ±n bp, where n can be set. Each dyad set is found back in the test set (which equals the positive test promoter set and the negative test promoter set), and each promoter is scored according to dyad content; afterwards all the promoters are ranked. Afterwards ROC analysis is performed on the set of promoter scores, and then the optimum parameter setup with the highest AUC score is selected for promoterome analysis.

Since the equation which calculates the *cdr* score takes the number of dyad occurences into account (see Eq. 4), this means that we can apply a cutoff value to select those dyads which occur a minimal number of times. A lower cutoff would include a larger set of dyads. If the distance between the head and tail motifs in the dyad are also allowed to wobble, the algorithm thereby picks up more instances of the given dyad. This also influences the *cdr* score of the dyad. Studying the distribution of the TRANSFAC/PLACE elements also influences the promoter's score. Therefore we used these parameters to study a large number of different dyad sets.

The final phase of the algorithm is the promoterome search where the promoterome is filtered with those optimum dyads as defined by the optimum parameters with the highest AUC score calculated during the ROC analysis. The promoters are then scored accordingly with the optimum dyads' scores, and are then ranked afterwards. The highest scoring promoters are then predicted to have the same function as those genes which the algorithm originally started out from.

The present algorithm was first developed at the Agricultural Biotechnology Center as a part of my college thesis work under the supervision of Dr. Miklós Cserző and further improved and refined at the Institute of Plant Biology at the BRC-HAS under the supervision of Dr. Sándor Pongor and Dr. János Györgyey.

5.5. Motifs and sequences used in testing and validating the algorithm

5.5.1. TRANSFAC and PLACE motifs

For the testing of the algorithm we selected 37 motifs known to take part in abiotic stress response based on their annotation from the PLACE (Higo, 1999) and TRANSFAC (Matys, 2006) database. These transcription factor binding sites were selected because of their involvement in abiotic stress (drought, osmotic, salt, cold stress). They were used in the analysis to check whether they could improve the behaviour of the algorithm since they were already known to be involved in stress. These sequences were short oligomers mostly 4-9 bp long each. A similar *cdr* score was calculated for each oligonucleotide according to Equation 4. These motifs were used in the testing of the algorithm in Arabidopsis and Oryza sativa in that they were also found in the test promoter sets and their scores added to the individual promoter scores. We studied the occurence of pairs of these TRANSFAC/PLACE motifs in the stress and non-stress learning promoter sets. That is, we formed dyads out of these motifs and determined their individual *cdr* scores. Overall, 277 TRANSFAC/PLACE dyads were found in the learning sets. Only 10 of these had a *cdr* score less than 0.5, and 265 had a *cdr* score of 1.0. The dyad sequence, the dyad's occurence in the stress and non-stress learning sets as well as a *cdr* score was can be seen in Supplementary Table 1.

5.5.2. Promoter sequences

In Arabidopsis, the 3 Kbp upstream regions were downloaded from the TAIR website at the following ftp website and then truncated to 2Kbp. This corresponds to the average intergenic region for Arabidopsis (Picot, 2010). This file contains 31,128 promoter sequences. All Arabidopsis promoters were BLASTed against each other to check for repetitive elements (parts of sequence with high similarity over 50 bp long).

The exact coordinates for the Oryza sativa promoter sequences were taken from the all.1kUpstream.gz Oryza sativa promoter sequence file from the TIGR/JCVI (J. Craig Venter Institute) website. This file however contained only 1 Kbp sequences for all rice grene sequences, so we had to extract the whole 2 Kbp promoter sequence from the 12 Oryza sativa chromosome sequences (the file all.con) using our own script. The regions maximum 2 kb upstream of the ATG start site, excluding the overlaps with the coding regions of upstream genes were collected.

The promoter regions for 21 Oryza sativa aldo-keto reductase genes were manually extracted from the NCBI database after the mRNA sequences were BLASTed against the Oryza sativa genome. The Oryza sativa homologs of the found wheat glucanase, chitinase, and pathogenesis related genes were BLASTed against the Oryza sativa genome at the Gramene website (Liang, 2007), with some local mismatches allowed. All BLAST hits were selected whose e-scores were below 10^{-60}. Overall we were able to manually retreive the 2Kbp upstream region of 29 glucanase sequences, 19 chitinase sequences, 5 PR1 sequences, 5 PR4 sequences, 13 PR5 sequences, and 20 PR9 sequences at the NCBI database.

5.6. Definition of dyad clusters

Clusters of dyads were made where pairs of dyads were aligned with each other. Two dyads belonged to the same cluster if their alignment had a minimum point score between 7 and 10 (meaning a Hamming distance of 0-3). An ungapped local alignment method was entailed to measure the similarity between two dyads where the two dyad sequences were slid against each other. The two dyads were aligned where the Hamming distance was the smallest. An alignment was given 1 point if the bases matched, and 0.5 points if A was paired with T or C with G. The dyads A<u>CCTG</u>NNN<u>C</u>C<u>AAT</u> and <u>CCTG</u>GNN<u>C</u>G<u>AAT</u> score 8.5 points and therefore belong to the same cluster. Note that we did not make it a requirement for all cluster members to have a Hamming distance less than or equal to 3.

5.7. Definition of regulatory networks

In the case of Arabidopsis we studied networks of regulatory elements. This includes either single TRANSFAC/PLACE motifs, single dyads, or clusters of dyads (that is, a group of dyads which are sequentially similar to each other as defined in Section 5.6.). Therefore, we studied „dyad dyads". We calculated the occurence of a pair of elements within a given set of promoters. We calculated a *cdr* score for each REP as seen in Equation 7:

$$(7)\ cdr = (N_+ - N_-)/N_+$$

where N_+ is the number of positive promoters a given regulatory element occurs in, and N_- is the number of times it occurs in the negative promoter set. This we did to get a picture of how significant a given regulatory element is in its role in abiotic stress.

Besides this, we calculated the occurence of all possible regulatory element pairs (REP) within a given distance (100 bp, which is larger than the longest possible dyad, and still fits into our computational constraints) from each other within a given promoter set (found in a promoterome search). Here a *cdr* score was assigned to each REP according to its occurence in the positive and negative promoter set the same way as was described in Equation 4. Afterwards, we studied the top 1224 REPs which had a minimum *cdr* score of 0.5, since this was used as the minimum *cdr* score value used in the test phase in *Arabidopsis*.

5.8. ROC analysis

ROC analysis (called receiver operating characteristic or relative operating characteristic) is a statistical technique used for visualizing, organizing, and selecting classification models based on their performance. In our case, ROC analysis is used in the testing phase of the algorithm in the test promoter sets to determine the optimum dyad set (the classification model) which gives the best set of parameters for the promoterome search. For a review on ROC analysis, see Fawcett, 2004.

In order to construct a classification model one must set up a test set and then map the elements of the set to an element of the set {p, n} (positive or negative). In our case, the test stress promoters were characterized by a 1, whilst the test non-stress promoters were characterized by a 0 (meaning that their relative expression change during abiotic stress was greater then or equal to 2). In total there are n_+ known positive elements and n_- known negative elements. Afterwards an algorithm may either predict whether each instance of the test set is positive or negative, or in our case it can give a score value to the individual elements of the set indicating the degree to which they are an instance of the class (that is, the sum score of the putative stress dyad elements in the promoter). The set members are then ranked according to their score. ROC analysis measures how successfully a classification system separates positive instances from negative ones.

On the so-called ROC graph the true positive rate (TPR) is plotted on the y-axis against the false positive rate (FPR) shown on the x-axis. The points on the ROC graph are plotted in a stepwise fashion. A score threshold of $+\infty$ corresponds to the point (0,0) on the graph and $-\infty$ corresponds to the point (1,1). As the score threshold is lowered each member of the set is taken into account. For each positive element encountered in the set of ranked scores, a step is taken upwards parallel to the y-axis with an increment of $1/n_+$, while a step is taken with an increment of $1/n_-$, parallel to the x-axis.

In the test example of Figure 13, 20 elements (10 positive, 10 negative) have been scored by a certain classification model. On the ROC graph plot on the right each point corresponds to one of the score values attained by the classification model.

Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

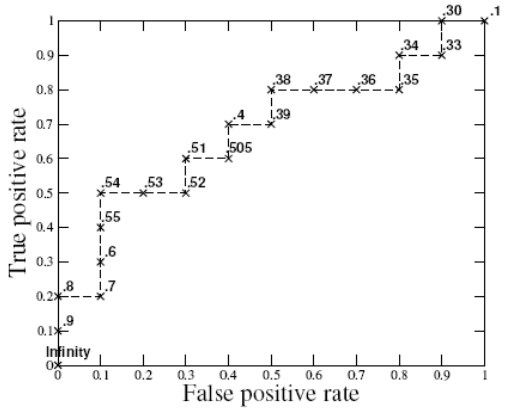


Figure 13. Test example of ranked scores and its corresponding ROC graph (figure taken from Fawcett, 2004)

The larger the number of elements in the test set the finer the curve on the ROC graph is. The better the classification model, the more the curve will reach the upper lefthand corner at the point (0,1). A perfect classification model which completely segregates the positive and negative instances of the test set from each other reaches this point, while a random classification model represents a curve which approaches the curve which represents the equation f(x) = x.

As a measure of the quality of the classification model one can calculate the AUC (area under curve) value for a given ROC graph. An AUC value of 1 corresponds to a perfect classification model, while 0.5 corresponds to a random one. The classification model in Figure 13 has an AUC value of 0.68.

5.9. Calculation of Jacquard coefficient

The Jacquard coefficient is a method of calculating the ratio of elements common to two sets to all elements in both sets. Mathematically, if N_A is the number of elements in set A, N_B is the number of elements in set B, and N_{AB} is the number of elements common to both sets, then the Jacquard coefficient:

$$(8)\ J = \frac{N_{AB}}{N_A + N_B - N_{AB}}$$

The Jacquard coeffient was used in the analysis to calculate the REP content between two given promoters. Here, the distance between two individual promoters is equal to 1-J, which signifies the difference in REP content.

5.10. Determination of expression change for selected genes

In order to check whether a given gene in *Arabidopsis* or *Oryza sativa* from a promoterome search was stress-induced, we determined that the relative gene expression change for such a gene is equal or greater than 2. For this we checked gene expression data from Genvestigator (expression level change for genes involved in cold, drought, osmotic, and salt stress) (courtesy of William Gruissem) and the GEO datasets at NCBI for *Arabidopsis*, namely data sets GDS1620 (cell cultures responding to cold, and hydrogen peroxide), GSE10670 (leaf samples responding to drought), GDS3216 (whole

seedling roots responding to salinity stress), GDS1382 (response to mild dehydration stress), and GSE5620-4 (root and shoot tissues in response to cold, drought, osmotic, and salt stress). For *Oryza sativa* we checked the following GEO datasets: GSE3053 (crown and growing point tissues under salt stress), GSE4438 (Rice Crown and Growing Point Tissue Under Salt Stress imposed during the Panicle Initiation Stage), and GSE6901 (expression profiles of Oryza sativa genes under cold, drought, and salt stress). Here, the expression level for stress experiments were divided by the corresponding control experiments.

5.11. Detection of repetitive elements in learning set promoters

The learning set promoters were blasted against each other and some promoters were removed if their sequences had large stretches of DNA which were similar to each other containing repetitive elements (stretches of DNA at least 50 bp long, with more than 90% similarity). In such cases where there was a match between two such promoters, one of the promoters was discarded randomly.

5.12. Programming environment

The algorithm was implemented in a 64 bit IRIX 6.5 programming environment using a C shell. A number of scripts written in C (GCC 3.4.6), gawk (GNU Awk 3.1.5) (Aho et al., 1988) were used to analyze input data sets and create data files.

6. Results

6.1. Testing of algorithm in Arabidopsis thaliana

The algorithm was first tested in Arabidopsis thaliana, since it is a widely used dicot model organism, and since its compact genome sequence and gene annotation is the most complete.

6.1.1. Selection of promoter sets

For the learning promoter sets 125 stress and 125 non-stress promoters were selected based on their involvement in abiotic stress according to their annotaion (cold, drought, salt, or osmotic stress). The promoter sets were filtered prior to selection so as not to contain promoters containing any repetitive sequences (see Materials and Methods, section 5.11). The test sets were made up of 44 promoters each (this is approximately one third the size of the learning set, a ratio often used in machine learning algorithms). The involvement in abiotic stress for each gene was checked in the Genevestigator database (Hruz, 2008). A list of the Arabidopsis genes used in the positive learning set can be seen in Supplementary Table 2.

6.1.2. ROC analysis and parameter definition for promoterome analysis

The test parameters for ROC analysis were set up for Arabidopsis in the following way: top dyad sets were defined where the given dyad occurred at least 5-20 times (16 sets in total) in the positive learning (stress) promoter set. Furthermore, these dyad sets were split up into subsets where the minimum *cdr* score was 0.6-1.0 in increments of 0.1 (5 subsets per set). A further parameter was the wobbling factor which was set to 0 to ±5bp (6 more parameter combinations). Besides this, the 37 TRANSFAC/PLACE motifs were also found back in the test promoter set in another parametrical test combination, where the *cdr* score of those motifs found in a given promoter were added to the specific individual promoter score. This made $16 \times 6 \times 5 \times 2 = 960$ parameter combinations in total, corresponding to 960 AUC values.

The reason we chose 5 bp as the minimum limit was that under 5 bp the algorithm found too many dyads to be biologically realistic (e.g. 60,302 in *Orzya sativa*), and that

when performing ROC analysis, these dyads saturated the test promoters, covering 1735 bp on average.

ROC analysis was performed for each of these parameter combinations, and an optimum parameter combination was deduced where the AUC value was the highest, which was 0.66736. The p-value for getting such an AUC value in our case was 0.0044, which is highly significant (calculated by MedCalc Version 11.3.6). This parameter combination was where all dyads occurred at least 14 times in the positive learning promoter set. The top 81 dyads were taken without the 37 TRANSFAC/PLACE motifs, and a wobbling factor of ±2 bp was applied. A 3D graph of AUC values can be seen in Figure 14 where the minimum dyad occurrence was 14.

In order to measure the significance of finding such dyads we applied the algorithm to two randomly selected sets of 125 Arabidopsis promoters and 2 sets of randomly selected Oryza sativa promoters. Overall, we found 2 and 3 dyads using the random Arabidopsis promoter sets and 2 and 6 dyads with the Oryza sativa sets. None of these dyads matched any known abiotic stress motifs.

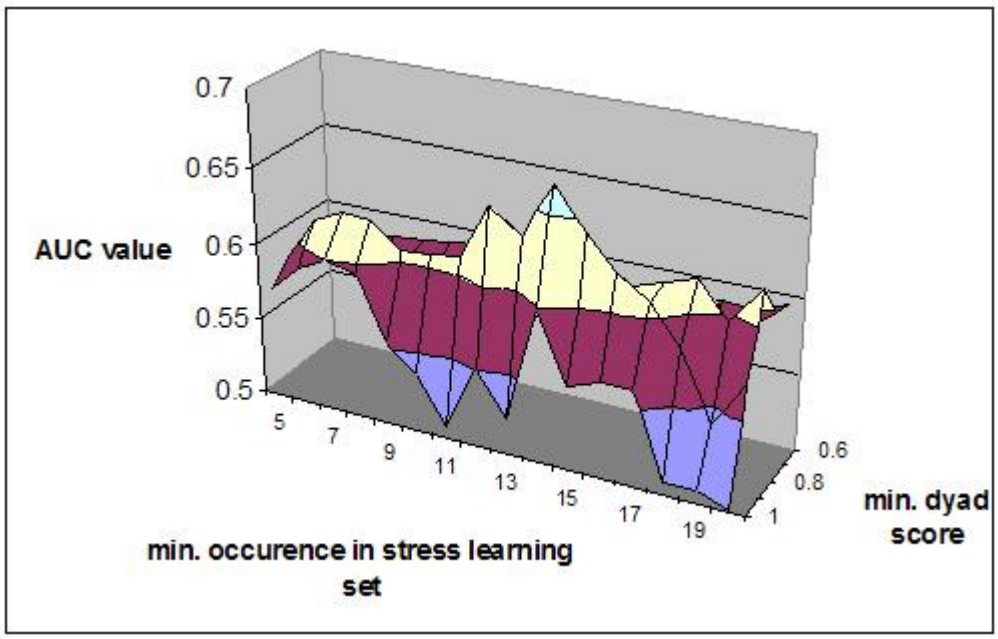


Figure 14. AUC values according to spacer mismatch and top number of dyads in Arabidopsis

The spacer wobbling of ±2 bp indicates that slight spacer wobbling was possible and that the head and tail motifs of the dyads were not too strictly defined. Furthermore, the AUC value was similarly high for similar parameter combinations: above 0.65 at a

minimum occurrence of 14 in the positive promoter set, and a minimum *cdr* score of 0.9 with a wobbling factor of ±2 bp. However, a similar AUC peak could still be discerned for other wobbling factors, meaning that the algorithm was capable of finding a stable dyad parameter combination. A list of the top 81 dyads can be seen in Table 5.

Dyad sequence	*cdr* score	annotation	Dyad sequence	*cdr* score	annotation
AAAAA{N10}GAAGG	1		AAAAA{N21}CACGT	0.933333	
AAAAA{N12}TGGTA	1		AAAAA{N21}GGTAA	0.933333	
AAAAA{N39}TACGT	1		AAAAA{N17}ATGAG	0.933333	
AAAAA{N18}TTGGC	1		ATATG{N1}TTTTA	0.933333	Heat shock element
AAAAA{N11}GAGTT	1		ATATA{N3}AGTTT	0.933333	SEF1 binding site
AAAAA{N14}CTCTA	1		TGTTA{N49}TTATT	0.933333	
AAAAA{N28}GTAGA	1		CCACA{N22}AAAAA	0.933333	
GAAGT{N45}AAAAA	1		TTATA{N2}GTTTT	0.933333	Heat shock element
ACTAA{N10}AGAAA	1		TTATA{N4}TGATT	0.933333	OCS element
CAAGT{N49}TTTTT	1		AGTTG{N46}TTTTT	0.933333	
TATGA{N11}TTTTT	1		AACTA{N28}TAAAA	0.933333	
TTGAA{N7}TATAA	1		ATAAA{N46}ACTAT	0.933333	
TTTTG{N41}GTAAA	1		AAGAA{N43}ATGAT	0.928571	
TTTTT{N44}AAGCA	1		AAAAA{N9}ACTGA	0.928571	Ethylene responsive element
TTTTT{N15}CCTTG	1		AAAAA{N25}TGGGT	0.928571	
TTTTC{N30}AAAAG	1		AAAAA{N5}TCGAA	0.928571	Ethylene responsive element
AACTA{N48}TTAAA	1		AAAAA{N50}CCTTG	0.928571	
GCAAA{N41}AAATT	1		AAAAA{N1}GACAA	0.928571	Ethylene responsive element
GAAGA{N22}TTTTT	0.954545		AAAAA{N2}AGCAT	0.928571	Ethylene responsive element
AAAAA{N7}CGAAT	0.947368		ATATG{N14}TTTAT	0.928571	
AAAAA{N37}TGTAC	0.947368		ATTGT{N2}TTAAA	0.928571	Heat shock element
ATATA{N26}AATGT	0.947368		ATTTT{N23}TAACT	0.928571	
TGATG{N14}AAAAA	0.947368		ATTTT{N3}ACAAG	0.928571	
TATTT{N22}CATTT	0.947368		TAAAA{N4}AAAGC	0.928571	GT-1 binding site
TTTAG{N31}TATTT	0.947368		AAATA{N36}CATTT	0.928571	
AAAAA{N33}ATAGT	0.944444		TATAT{N42}GAAAC	0.928571	
ACTTG{N23}TTTTT	0.944444		TATAT{N28}GTTGA	0.928571	
ATAAA{N45}AACTA	0.944444		TGTTT{N45}AAGAA	0.928571	
AGAAA{N42}ATGAT	0.941176		TATTT{N19}AATCA	0.928571	
AAAAA{N44}TCTAC	0.941176		TATTT{N26}AACTT	0.928571	
AAACT{N29}ATCTT	0.941176		TATTT{N42}ATGAA	0.928571	
AAAAA{N19}TGAGT	0.9375	SEF3 binding site	CTATA{N26}TTTTT	0.928571	
AAAAA{N30}CAACT	0.9375		CTATT{N25}TAAAA	0.928571	
AAAAA{N4}AAGCC	0.9375		TTATT{N29}ACTTT	0.928571	
AAAAT{N10}AGTTT	0.9375	Elicitor-motif	CTTTA{N27}ATATA	0.928571	
ATACT{N22}TTTTT	0.9375		TTCTA{N34}AAATA	0.928571	
TGTGT{N35}AAAAA	0.9375		AAAAA{N29}TAGAT	0.923077	
CATAT{N25}ATATA	0.9375		AAAAA{N24}AGAAT	0.909091	
TTGGC{N41}AAAAA	0.9375		AAAAA{N25}ATTCT	0.9	
CTTTT{N27}TTAAT	0.9375		TTAGA{N16}AAAAA	0.9	
AAAAA{N9}ACTAG	0.933333	Ethylene responsive element			

Table 5. List of top 81 putative dyads used in optimum AUC parameterization in Arabidopsis (occurrence in minimum 14 promoters, minimum *cdr* score of 0.9, wobbling factor of ±2bp)

As we can see, 12 of the 81 (14.8%) top Arabidopsis dyads matched a well-known motif found either in the PLACE or PlantCARE database. This means that our algorithm was capable of finding well-known TFBS motifs. This also means that the

algorithm predicted a number of new TFBS's, which could be involved in abiotic stress. Quite a number of dyads have a spacer length of less than 5-10 bp. Sterical factors would most probably block individual TF's from binding the head and tail motifs belonging to such a dyad. This means that such a dyad might in reality be a single TFBS, where bases between the front and back end of the motif might be inspecific or not important for binding (e.g. in these positions the TF binds to the sugar-phosphate backbone of the DNA).

6.1.3. Promoterome analysis

For the promoterome analysis in Arabidopsis we took the top 81 putative dyads which corresponded to the optimum parameter combination, and searched for them in the Arabidopsis promoterome. Each promoter sequence was scored according to Equation 5 in Section 5.2. and then ranked. We also studied the number of stress learning, non-stress learning, stress test and non-stress promoters from the top 10,000 promoters found in the promoterome search in increments of 100 promoters. This can be seen in Figure 15. We can see that as a general rule more stress learning and stress test promoters were found back than non-stress learning and non-stress test promoters.

We also studied the percentage of non-stress promoters to all promoters found in the promoterome search for the top 10,000 promoters, also in increments of 100 promoters. As we can see in Figure 16, this percentage value declines between the top 1600 and the top 3100 promoters from 3.5% to 2.5%, corresponding to 2 false discoveries among 57 (3.5%) and 81 (2.5%) total promoters found back from both the stress and non-stress learning and test sets.

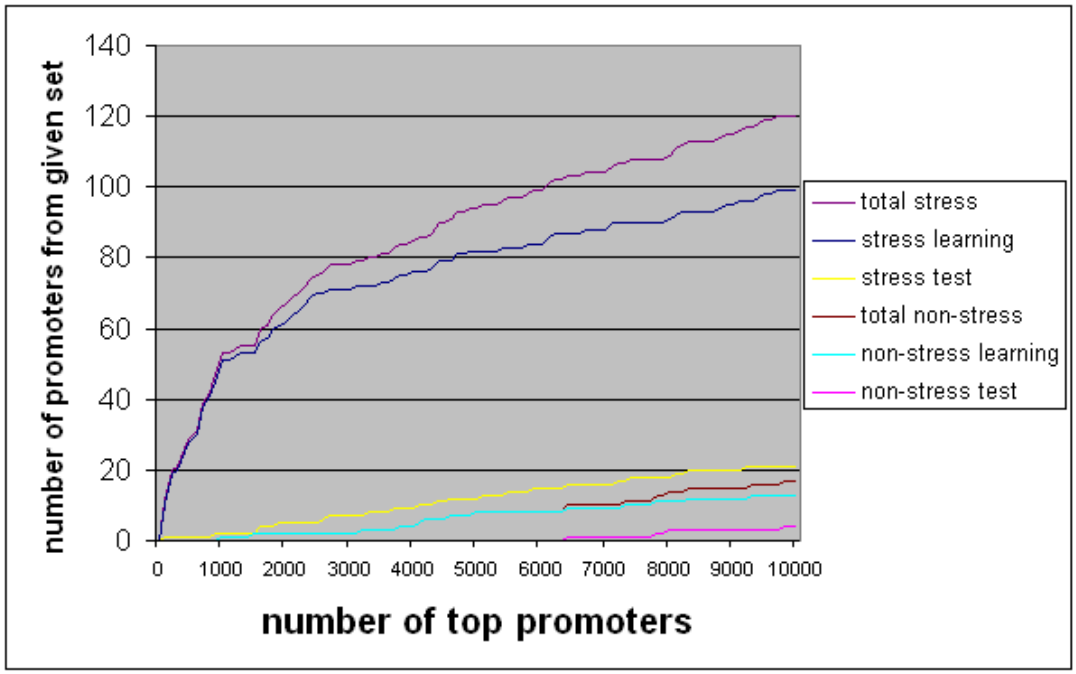


Figure 15. Number of promoters from different sets (stress learning, non-stress learning, stress test, non-stress test) found back in the top 10,000 promoters found in the promoterome search in increments of 100. in Arabidopsis, with the highest scoring promoters at the beginning.

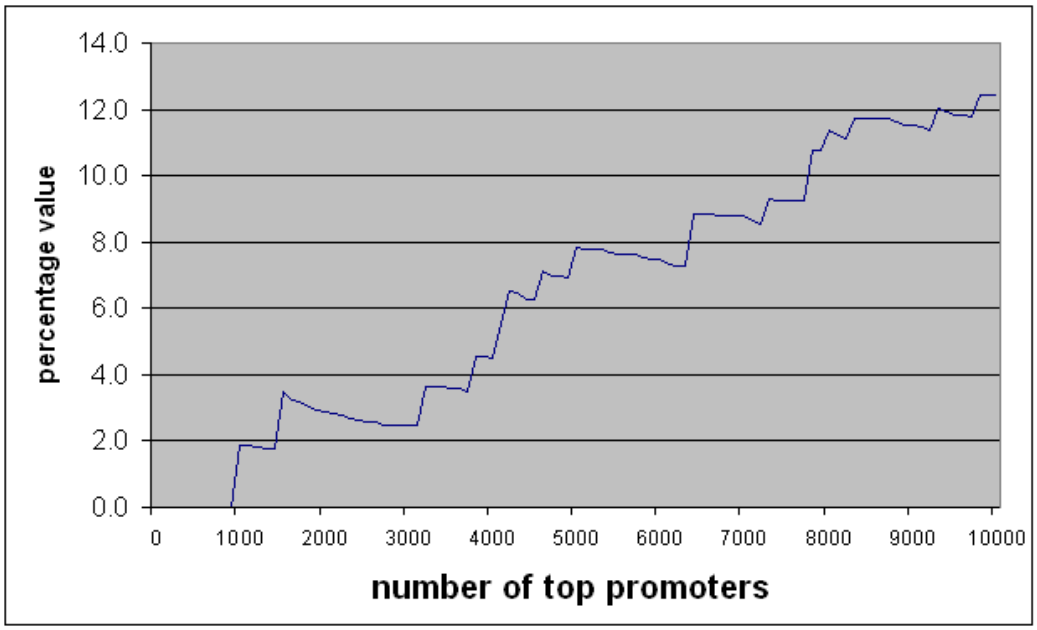


Figure 16. The percentage of non-stress promoters to all promoters found back in the promoterome search for the top 10,000 promoters in increments of 100 promoters in Arabidopsis.

Therefore we used the top 3100 promoters from the promoterome search for further analysis in Arabidopsis. Overall, 1,542 of the 3100 highest scoring genes either had an Affymetrix probe associated with it, or had abiotic stress induction data in either the Genevestigator database or GEO datasets, or belonged to the original stress test or learning promoter set. Overall, 1,212 of these genes were shown to be involved in abiotic stress, meaning a positive prediction rate of 78.6%.

6.1.4. Regulatory element network analysis

In Arabidopsis, analysis was performed in such a way as to get a picture of how our putative dyads formed regulatory networks with each other (as well as the selected TRANSFAC/PLACE elements). Therefore the top 81 putative dyads were clustered according to the description in the Materials and Methods (section 5.6). Overall 11 dyad clusters were created with a pairwise score of at least 7. The cluster sizes varied from 2 to 7 with an average of 3.5 dyads belonging to each cluster. A list of these clusters, their constituent dyads and their definitive consensus sequence are shown in Table 6.

Cluster number	**dyads**	**consensus sequence**
Cluster1	**AAAAA{N9}ACTGA, AAAAA{N9}ACTAG, AAAAT{N10}AGTTT, AAAAA{N7}CGAAT, AAAAA{N10}GAAGG, AAAAA{N12}TGGTA, AAAAA{N11}GAGTT**	**AAAAAAAT{N5}CGRMDRRTWT**
Cluster2	AAAAA{N1}GACAA, AAAAA{N2}AGCAT, TAAAA{N4}AAAGC, AAAAA{N4}AAGCC	TAAAAAARAMAAGCMT
Cluster3	**AAAAA{N25}ATTCT, AAAAA{N29}TAGAT, AAAAA{N25}TGGGT, TATAT{N28}GTTGA, TATTT{N26}AACTT, AAAAA{N28}GTAGA**	**AAAWATWT{N22}WKKSWASWTGA**
Cluster4	**TATTT{N19}AATCA, TATTT{N22}CATTT**	**TATTT{N19}AATCATTT**
Cluster5	**AAAAA{N24}AGAAT, CTATA{N26}TTTTT, CTTTA{N27}ATATA, CTTTT{N27}TTAAT, ATATA{N26}AATGT**	**CYTWWAWA{N25}WDWWTRT**
Cluster6	CTATT{N25}TAAAA, CATAT{N25}ATATA	CMTATT{N24}ATAWAA
Cluster7	**AAAAA{N21}GGTAA, AAAAA{N17}ATGAG, AAAAA{N19}TGAGT, AAAAA{N18}TTGGC**	**AAAAAA{N17}ATKRGYAA**
Cluster8	**ATATA{N3}AGTTT, TTATA{N2}GTTTT**	**ATWTATANAGTTTT**
Cluster9	ATACT{N22}TTTTT, GAAGA{N22}TTTTT	ATAAA{N45}AACTAT
Cluster10	ATACT{N22}TTTTT, GAAGA{N22}TTTTT	GAAKACT{N20}TTTTTTT
Cluster11	**AAAAA{N37}TGTAC, AAAAA{N39}TACGT**	**AAAAA{N35}TGTACGT**

Table 6. List of putative dyads belonging to each of the 11 dyad clusters. Dyad clusters in bold predicted to play key regulatory roles in Arabidopsis

Regulatory element pairs or REPs were formed where either one of the regulatory elements came from either the 37 TRANSFAC/PLACE motifs or from one of the dyad clusters, or was a singleton dyad. These REPs were found back in the top 3100 stress and non-stress promoters found by the algorithm in the promoterome search described in the previous section, and scored according to the description in the Materials and Methods section. Overall there were 1224 such REPs with a minimum *cdr* score of 0.5, which were used in the regulatory element and promoterome analysis of *cor* end *erd* genes.

In the next step we described the regulatory networks of two sets of selected *Arabidopsis* genes known to be involved in abiotic stress. These genes belonged to the *cor* and *erd* families of genes (*co*ld *r*esponsive and *e*a*r*ly *d*ehydration). A list of the selected genes and their annotation can be seen in Table 7. The reason these genes were selected was because their expression profiles are known to be similar to each other, therefore we can suspect that their promoters may contain similar regulatory elements.

Overall, we can see that 22 PLACE/TRANSFAC motifs play key roles in both networks (connected to at least 10 other regulatory elements): tfpl_1, 3, 4, 5, 7, 9, 10, 11, 14, 15, 17, 18, 21, 22, 23, 24, 26, 27, 30, 31, 32, and 34. Amongst these, tfpl_3, 4, and 5 part of the well-known ABRE element, while tfpl_15, 17, 22, 23, 26, 30, 32, and 34 correspond to the MYB binding site within dehydration genes (Abe, 2003). The motifs tfpl_1, 7, and 11 correspond to the DRE element, which is well known to take part in the response to dehydration (Zhang, 2005). The motifs tfpl_9, 10, 18, 21, 24, 27, and 31 correspond to a MYC binding site (Abe, 1997).

cor genes	
gene id	**TAIR description**
At2g42530	COLD REGULATED 15B (COR15B)
At5g52310	cold regulated gene, the 5' region of cor78 has cis-acting regulatory elements that can impart cold-regulated gene expression
At1g29395	encodes a protein similar to the cold acclimation protein WCOR413 in wheat.
At2g15970	encodes an alpha form of a protein similar to the cold acclimation protein WCOR413 in wheat.
At1g20440	Belongs to the dehydrin protein family, which contains highly conserved stretches of 7-17 residues
Erd genes	
At1g08930	encodes a putative sucrose transporter whose gene expression is induced by dehydration and cold.
At1g62320	early-responsive to dehydration protein-related / ERD protein-related;
At4g19120	dehydration-responsive protein, putative;
At4g15430	ERD (early-responsive to dehydration stress) family protein;

Table 7. List of cor and erd genes used in regulatory network analysis

Four of our dyad clusters were also found to play key roles in the regulatory networks for both the *cor* and *erd* genes, clusters 5, 9, 10, and 11, being connected to at

least 10 other elements in the network. A list of the dyad clusters may be seen in Table 3. The dyads CTATA{N26}TTTTT, ATACT{N22}TTTTT, GAAGA{N22}TTTTT, ATACT{N22}TTTTT, and GAAGA{N22}TTTTT from clusters 5, 9, and 10 contain the motif TTTTT, which corresponds to the MYB binding site (Kim, 2010). The tail motif of the dyad AAAAA{N39}TACGT from cluster 11 corresponds to the well known ACGT core motif, which takes part in dehydration stress (Simpson et al. 2003).

The *Arabidopsis* promoterome was screened to see which other promoters contained REPs common to the 5 *cor* promoters found by the algorithm. This was done by listing all REPs from the top 3100 working set and checking how many of them were in common with the 5 *cor* promoters and all other *Arabidopsis* promoters.

Arabidopsis gene	**Functional annotation**
At1g06580	Pentatricopeptide repeat (PPR) superfamily protein
At1g25550	myb-like transcription factor family protein
At1g46480	Encodes a WUSCHEL-related homeobox gene family member with 65 amino acids in its homeodomain.
At1g50040	Unknown protein
At1g67480	Galactose oxidase/kelch repeat superfamily protein
At2g05200	transposable element gene; non-LTR retrotransposon family (LINE)
At2g41060	RNA-binding (RRM/RBD/RNP motifs) family protein
At2g41120	Unknown protein
At2g42470	TRAF-like family protein
At3g08500	Encodes a putative R2R3-type MYB transcription factor (MYB83)
At3g10980	Unknown protein
At3g23805	Member of a diversely expressed predicted peptide family showing sequence similarity to tobacco Rapid Alkalinization Factor (RALF)
At3g23970	F-box family protein
At3g29620	transposable element gene; transposase IS4 family protein
At3g30718	transposable element gene; gypsy-like retrotransposon family
At3g32360	transposable element gene; non-LTR retrotransposon family (LINE)
At3g52490	Double Clp-N motif-containing P-loop nucleoside triphosphate hydrolases superfamily protein
At4g01080	Encodes a member of the TBL (TRICHOME BIREFRINGENCE-LIKE) gene family
At4g06530	Hypothetical protein
At4g14290	alpha/beta-Hydrolases superfamily protein
At4g17410	DWNN domain, a CCHC-type zinc finger
At4g36510	Unknown protein
At5g16740	Transmembrane amino acid transporter family protein
At5g41505	Unknown protein
At5g43300	PLC-like phosphodiesterases superfamily protein

Table 8. List of top 25 new *Arabidopsis* gene with lowest distance (below 0.5) to all of our selected set of *cor* genes

We calculated the Jacquard coefficient (see Materials and Methods) between all promoters between each of the five *cor* promoters based on their REP content to measure the distance between these pairs of promoters. We selected those 25 Arabidopsis promoters whose minimum distance was less than 0.5. This list can be seen in Table 8. Amongst these genes were a hypothetical gene (At4g06530) and 5 genes of unknown function (At1g50040, At2g41120, At3g10980, At4g36510, and At5g41505). We assume these new candidate genes to be regulated similarly to the *cor* genes.

6.1.5. Comparison of the algorithm with YMF and dyad-analysis

We compared the present algorithm to two other motif finding algorithms, which are well-known and widely used, namely the YMF (Yeast Motif Finder) of Sinha and Tompa, and dyad-analysis of Jacques van Helden (van Helden, 2000; Sinha, 2000) on the set of 125 *Arabidopsis* stress learning promoters. For the YMF program we looked for pentamer dyads without any mismatches, with a spacer length of at least 0 bp to a maximum of 52 bp. Both programs were tested on the 125 Arabidopsis learning promoters.

Using the YMF program we were able to find 283 promoters containing more than 1 of the significant dyad motifs found by the program. Of these only 3 belonged to the original 125 stress promoters (1.1%). Of these 283 promoters, 195 could be found in the Genevestigator database, of which only 6 were found to be stress-inducible (3.1%), which is close to the percentage of stress promoters which we find in a search through the Genevestigator database with randomly selected promoters (3.52%).

Using the dyad-analysis program we were able to search pure pentamer dyads with spacers 0 to 52 bp long. With the dyad-analysis program we were able to find 149 promoters with more than 40 dyad motifs. Of these only 1 belonged to the original promoter set (0.77%). We found 110 of the 149 promoters in the Genevestigator database, of which only 4 were stress-inducible (3.6%). The reason the number of found stress-induced genes was so low is because data was used for the verification in their involvement in stress only from the Genevestigator database.

6.2. Usage of algorithm in Oryza sativa

After the algorithm had been verified in Arabidopsis, we tested the algorithm in the analysis of the Oryza sativa promoterome. Oryza sativa was used because its genome sequence is also fairly complete compared to other plants, and also since it is an important agricultural plant species and an important monocot model plant widely used in cereal genomics (Zhang, 2008).

6.2.1. Selection of promoters

We selected 129 stress genes which were partitioned into a positive learning promoter set of 87 sequences, and a positive test promoter set of 42 sequences. 143 non-stress promoters were selected, and 87 were put into the negative learning promoter set, while 56 were put into the negative test promoter set.

Both of the positive and negative promoter sets came from an experimentally verified data set which contained a list of Oryza sativa genes which were induced by drought conditions (Zombori Zoltán, personal communications). According to these experiments, the expression level of 3,137 genes was measured at 8 hrs and 14 hrs at 100% water level, and at the same timepoints at 20% water level, simulating drought conditions. The promoters for the positive learning set were selected based on their expression level increase at 14 hrs at both 20% and 100% water level as well as 8 hrs at 20% water level being greater than 2. The 143 non-stress genes were selected based on their indifference to drought conditions where their expression level change was less than 0.33 at 8 hrs and 14 hrs of 20% water level and 14 hrs under control conditions. A list of these genes can be seen in Supplementary Table 3.

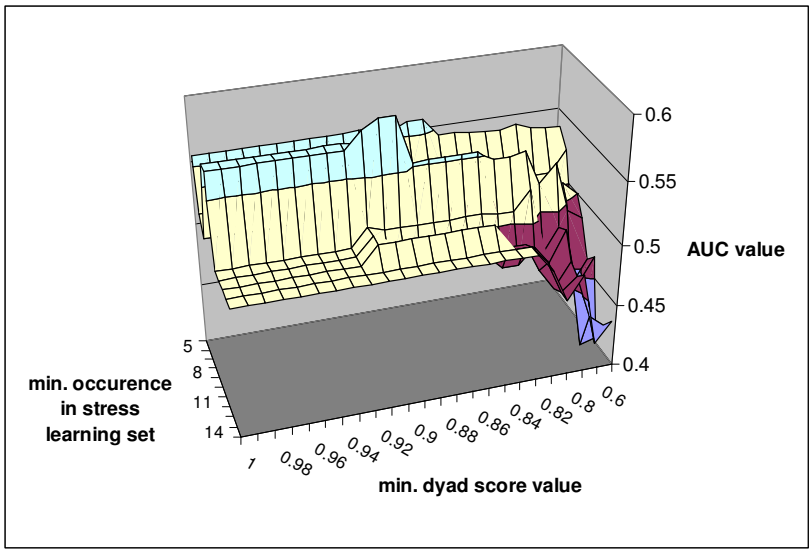


Figure 17. AUC values according to spacer mismatch and top number of dyads in Oryza sativa

6.2.2. ROC analysis and parameter definition for promoterome analysis

In Oryza sativa the minimum occurence of a given dyad in the positive learning promoter set was studied between 5 and 14. 5 was the minimum occurence set in Arabidopsis, and only very few dyads (less than 10) occured at least 14 times in Oryza sativa in the positive learning set. Lastly, instead of selecting the top n number of dyads, we chose those top dyads with a *cdr* score of at least 0.5, 0.6, 0.7, 0.8, 0.9, or 1.0.

The highest AUC value for Oryza sativa was 0.59069. The p-value for getting such an AUC value in our case was 0.0743 (calculated by MedCalc Version 11.3.6). As we can see in Figure 17 this corresponds to a well-defined peak AUC value which is defined by the following parameters: a wobbling factor of 0, a minimum dyad occurence of 9 in the positive learning promoter set, and a minimum *cdr* score of 0.89. After determining a maximum AUC value at *cdr* scores of 0.9, we refined our search by using dyads with a *cdr* score between 0.9 and 1.0 with increments of 0.01.

We then performed the promoterome analysis in Oryza sativa with these parameters, which corresponded to 38 top dyads, which can be seen in Table 9 with their *cdr* score values and PLACE and PlantCARE annotations. Amongst these, pentamers comprising the dyads ATTTG{N43}TTTAT and AAATA{N32}AAATG corresponded

to the well-known MYC TFBS, which is found in a number of cold-responsive gene promoters (Abe, 2003). Out of these, 8 elements were shown to take part in abiotic stress (21.1%).

Dyad	*cdr* score	PLACE annotation	PlantCARE annotation
GAGAA{N40}AATAT	1	ROOTMOTIFTAPOX1	
AGGGA{N3}GGGAG	1	UPRMOTIFIIAT	
AGAGA{N15}ACTTT	1	DOFCOREZM	
AGAAG{N40}TTTAT	1		
AAAGA{N0}AAATT	1	GT1CONSENSUS DOFCOREZM POLLEN1LELAT52	X98521 heat shock element
AAAAG{N32}TTTTG	1	CANBNNAPA DOFCOREZM	
GAGGC{N39}GAGGA	1		
GCGCC{N13}CCGCG	1	CGCGBOXAT	
GTTTG{N35}TTTAT	1	RSRBNEXTA CANBNNAPA	
GTTTA{N38}TTTGA	1	CANBNNAPA	
ATTTG{N43}TTTAT	1	CANBNNAPA MYCCONSENSUSAT	
TAAAA{N21}CACTT	1	CACTFTPPCA1	
TAATT{N47}CAAAA	1		
AAATA{N32}AAATG	1	MYCCONSENSUSAT	
GAGTA{N11}ACACA	1	CACTFTPPCA1	
CATTT{N42}TATAT	1		
TCAAA{N51}ATTTT	1		
TTATT{N29}TATAC	1	P1BS	
TTTGA{N22}ACATG	1		
TTTAA{N50}GATTT	1		
TTTCA{N30}TTTGA	1		
TTCTA{N11}AAAAT	1		U46545 heat shock element
CTCCT{N7}TTCTT	1		
TTCTT{N13}ATTCA	1		U46545 heat shock element
TCTTT{N5}TATTT	1	DOFCOREZM	
GATTT{N15}AATTA	1		
TGTTT{N35}TATTT	0.916667		
GAGAG{N5}GAAAA	0.909091		
TTCTT{N7}TAAAA	0.909091		U46545 heat shock element
AGAGA{N6}AAAGA	0.9	DOFCOREZM	
GTTTT{N51}ATATA	0.9		
TAAAA{N43}TGAAA	0.9		
CAATT{N24}AATTT	0.9	CAATBOX1	
TATGT{N24}TTTTG	0.9	CANBNNAPA	
TCAAA{N20}CAAAA	0.9		D10661 sequence imperative for maximal ELICITOR-mediated activation of PsChs1
TCATT{N30}TAAAT	0.9		
TTTGA{N38}ACTAA	0.9		
AATTC{N19}AAATT	0.9		D10661 sequence imperative for maximal ELICITOR-mediated activation of PsChs1

Table 9. Top 38 putative dyads found in Oryza sativa ROC analysis and annotation of matching motifs in the PlantCARE and PLACE databases.

6.2.3. Promoterome analysis

Promoter analysis was performed with the top 38 dyads defined in the previous section on the Oryza sativa promoterome as described in Section 5.5.2. In order to measure the performance of the promoterome search we counted how many promoters

were found from the stress and non-stress learning, and stress and non-stress test promoter sets. This we did for the top 10,000 promoters in increments of 100. The number of promoters from each set can be seen in Figure 18. In Figure 19 we can see the percentage ratio of promoters from non-stress promoter sets to promoters from stress and non-stress sets. We can see the percentage ratio of non-stress promoters rise to one small and two larger peaks where they drop off. The second larger peak drops off to a percentage value of 7%, corresponding to the top 4,600 Oryza sativa promoters found in the promoterome search.

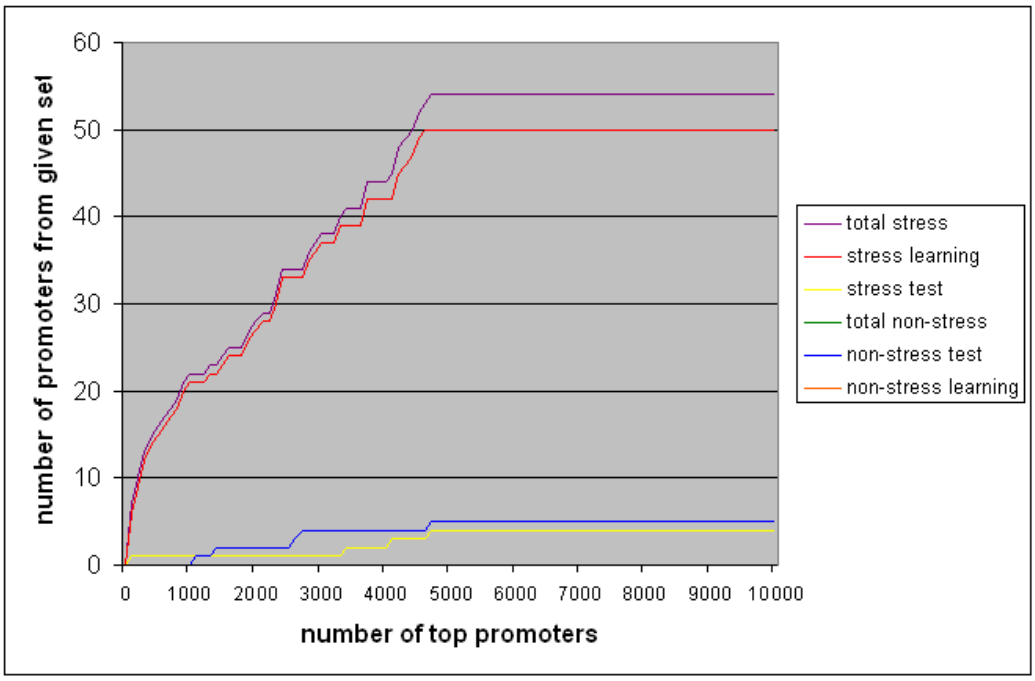


Figure 18. Number of promoters from different sets (stress learning, non-stress learning, stress test, non-stress test) found back in the top 10,000 promoters found in the promoterome search in increments of 100 in Oryza sativa, with the highest scoring promoters at the beginning. No non-stress learning promoters were found, therefore the total number of found non-stress promoters is the same as the number of found non-stress test promoters.

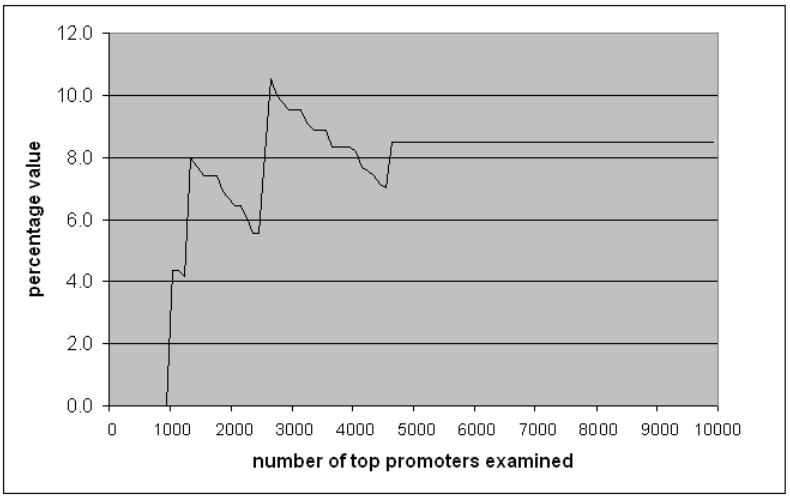


Figure 19. The percentage of non-stress promoters to all promoters found back in the promoterome search for the top 10,000 promoters in increments of 100 promoters in Oryza sativa

In order to calculate a success rate for the algorithm in Oryza sativa besides the Genevestigator data we used GEO (Gene Expression Omnibus) datasets from NCBI which contained abiotic stress data, namely datasets GSE3053 (crown and growing point tissues under salt stress), GSE4438 (Rice Crown and Growing Point Tissue Under Salt Stress imposed during the Panicle Initiation Stage), and GSE6901 (expression profiles of Oryza sativa genes under cold, drought, and salt stress).

GO term	GO function	Number of genes with GO function
GO:0006950	**response to stress**	359
GO:0009719	response to endogenous stimulus	210
GO:0009987	**cellular process**	199
GO:0007165	**signal transduction**	198
GO:0006464	**protein modification process**	170
GO:0009607	response to biotic stimulus	159
GO:0006350	transcription	155
GO:0009058	**biosynthetic process**	152
GO:0009628	**response to abiotic stimulus**	150
GO:0006810	transport	116
GO:0019538	protein metabolic process	111
GO:0006519	cellular amino acid and derivative metabolic process	91
GO:0019748	secondary metabolic process	91
GO:0008150	biological_process	83
GO:0009605	**response to external stimulus**	75
GO:0016043	cellular component organization	72
GO:0008152	metabolic process	63
GO:0006629	lipid metabolic process	60
GO:0009056	catabolic process	56
GO:0007275	multicellular organismal development	45

Table 10. Top 20 GO terms for biological functions for 1028 of the top 4,600 genes found in the Oryza sativa promoterome analysis

Out of 4600 genes, 3144 had an Affymetrix probe id, based on which we could check their expression data in the GEO datasets. 3102 genes showed a minimum twofold relative expression level change compared to the respective control sets according to

these datasets, meaning a positive predictive value (PPV) of 0.9866. Parallel to this, we selected a random set of 4600 *Oryza sativa* genes, from which 3975 had an Affymetrix probe id. Out of these, only 1243 were shown to be stress-induced which is only 31.3% of the total.

Gene Ontology (GO) terms for biological functions were retreived for 1028of the top genes at the Rice Array Database (Jung, 2008). This data can be seen in Table 10. It is interesting to note that 359 (34.9%) of the genes show a response to stress (which is the most common gene function, corresponding to a p-value of 0.0358), and 150 (14.6%) of them respond to abiotic stimuli, which is ninth in the list with a p-value of 0.023. This finding validates the usefulness of our algorithm as it shows that it is capable of finding other promoters involved in abiotic stress. Other important GO terms were found which had a significant p-value, such as transcription (p=0.0312), regulation of transcription, DNA-dependent (p=0.0036), translation (p=0.0008), defense response (p=0.0104), response to freezing (p=0.0010).

GO term	GO function	Number of genes with GO function
GO:0003677	**DNA binding**	364
GO:0003676	**nucleic acid binding**	346
GO:0008270	zinc ion binding	320
GO:0005524	ATP binding	306
GO:0000166	**nucleotide binding**	240
GO:0005515	**protein binding**	192
GO:0046872	metal ion binding	189
GO:0003824	catalytic activity	180
GO:0016491	oxidoreductase activity	178
GO:0004672	protein kinase activity	152
GO:0003723	RNA binding	150
GO:0050825	ice binding	149
GO:0004674	protein serine/threonine kinase activity	146
GO:0004713	protein tyrosine kinase activity	143
GO:0016740	transferase activity	130
GO:0003964	RNA-directed DNA polymerase activity	118
GO:0005488	binding	118
GO:0016301	kinase activity	110
GO:0009055	electron carrier activity	99
GO:0016787	hydrolase activity	99

Table 11. GO and annotation terms for the top 20 molecular functions for 2073 of the top 4,600 genes found in the Oryza sativa promoterome analysis

GO terms and annotation terms were also retreived for molecular functions for 2037 of the top Oryza sativa genes. These terms and the number of genes they occur can be seen in Table 11. Significantly enriched GO terms include the following: DNA binding (p=0.0000), nucleotide binding (p=0.0003), nucleic acid binding (p=0.0000), sequence-specific DNA binding transcription factor activity (p=0.0155), catalytic activity (p=0.0112), protein kinase activity (p=0.0003), protein serine/threonine kinase activity

(p=0.0008), protein tyrosine kinase activity (p=0.0008). All p-values were calculated by GO Enrichment Analysis at the Rice Array Database (Jung et al., 2008). This employs a conditional hypergeometrical test to calculate the significance of the p-values.

6.3. Usage of algorithm in two seperate test cases in Oryza sativa

After the algorithm had been tested and validated in Arabidopsis and Oryza sativa, we applied the algorithm for finding dyads in two seperate test cases: Oryza sativa aldo-keto reductase genes involved in the detoxification of ROS in the cell as well as Oryza sativa orthologues of wheat genes involved in biotic stress response, such as glucanases, chitinases, and four classes of PR genes.

6.3.1. Dyad discovery in Oryza sativa aldo-keto reductase promoters

Rice protein identifier	Annotation
Q7XCS4	Aldo/keto reductase family protein, putative, expressed (Os10g0517400 protein)
Q7XEJ9	Putative polyprotein
Q8GSK5	Aldo/keto reductase family-like protein
Q8H011	Putative NADPH-dependent oxidoreductase
Q8LMU5	Putative NADPH-dependent oxidoreductase
Q94LH8	Putative NADPH-dependent oxidoreductase
Q94LH9	Putative NADPH-dependent oxidoreductase
Q94LN0	Putative NADPH-dependent oxidoreductase
Q94LN1	Putative NADPH-dependent oxidoreductase
Q40648	Probable voltage-gated potassium channel subunit beta; AltName: K(+) channel subunit beta
Q7X7K5	OSJNBa0032I19.4 protein
Q8H4J8	Aldo/keto reductase family-like protein (Os07g0142900 protein)
Q7X8G7	OSJNBb0115I21.2 protein (Os04g0167800 protein) (OSJNBb0089K06.1 protein)
Q7XQ45	OSJNBa0032I19.9 protein (Os04g0339400 protein)
Q7XQ49	OSJNBa0032I19.1 protein (OSJNBa0008A08.10 protein)
Q7XT99	OSJNBa0008A08.11 protein (Os04g0338000 protein) (OSJNBa0032I19.2 protein)
Q7XTA2	OSJNBa0008A08.8 protein
Q7XV15	OSJNBa0064H22.3 protein (Os04g0447500 protein)
Q7XV16	OSJNBa0064H22.4 protein (Os04g0447600 protein)
Q7XV17	OSJNBa0064H22.5 protein
Q7XVS5	OSJNBa0067G20.18 protein
Q7XCS4	Aldo/keto reductase family protein, putative, expressed (Os10g0517400 protein)
Q7XEJ9	Putative polyprotein
Q8GSK5	Aldo/keto reductase family-like protein

Table 12. Identifiers and annotation for 24 AKR genes analyzed by our algorithm

Aldo-keto reductases (AKR) are a highly conserved family of genes which play an important role in the detoxification of toxic aldehydes coming from lipid peroxidation

(Oberschall, 2000). Such genes are activated mainly by salt and osmotic stress (Hideg, 2003). Their conservedness and specific biochemical role makes them an ideal subject for promoter analysis by our algorithm. 6 further TC sequences were found in rice to belong to the AKR gene family through BLAST analysis. In total, the promoter regions of 30 Oryza sativa AKR genes were analyzed (Turóczy, 2011), of which 3 were analyzed in depth experimentally.

6.3.1.1. Dyad selection and analysis

Dyad	***cdr* score**	**PlantCARE references**
AAAAT{N22}ATTTA	1	
AAAAT{N51}TTTCT	1	
AAATT{N11}TTAAA	1	U46545, Z95153, Z95153
AAATT{N45}AAATA	1	
ATATT{N27}AAAGT	1	
ATATT{N3}AAATT	1	U46545, Z95153
ATTTA{N36}AAATT	1	
ATTTT{N6}AACAT	1	S44898, D10661, X76131
CATTT{N42}AAATT	1	
CTATA{N5}ATTTT	1	
CTCTC{N43}GCGGC	1	
GAAAA{N10}TTTCT	1	U46545, Z95153, L41253
GAAAA{N43}TTTAT	1	
TATAT{N17}TAAAT	1	
TATAT{N19}AATTT	1	
TATCA{N36}TAAAA	1	
TCAAA{N35}TTTAT	1	
TCTCT{N44}GCGGC	1	
TTAAA{N0}ATTTT	1	X98521
TTAGA{N20}GAAAA	1	
TTGAA{N10}AATTT	1	
TTTAA{N41}TTTTA	1	
TTTAT{N19}TAAAA	1	
TTTAT{N8}AAATT	1	GT-1 factor binding site
TTTGA{N12}AATTT	1	
TTTGA{N19}TCAAA	1	
TTTTA{N33}AAAAT	1	
TTTAT{N33}ATATT	0.9	

Table 13. Top 28 dyads found in the analysis of the 24 Oryza sativa AKR promoters

A list of the protein identifiers used in the analysis can be seen in Table 12. In addition to the 30 genes identified, Turóczy and colleagues studied 3 more Oryza sativa AKR genes which were also significantly induced by hydrogen peroxide and ABA: Os01g0847600 (AKR1), Os01g0847700 (AKR2), and Os01g0847800 (AKR3) (Turóczy, 2011) (in bold in Table 14). The AKR1 gene was induced already by 20 minutes after

treatment, and showed a 12-16-fold expression level increase, whereas AKR2 and AKR3 showed only a somewhat diminished induction to stress conditions than AKR1. These two genes were induced only after 20-60 minutes of treatment and showed only a 4 and 6-fold expression level increase, respectively. After constructing a negative promoter learning set of 24 randomly selected Oryza sativa promoters, we calculated the *cdr* value for the highest scoring dyads.

We selected the top highest scoring 28 pentamer dyads for further analysis which had a *cdr* score higher than or equal to 0.9, with an occurrence in at least 7 of the input learning promoter set. We searched the PlantCARE database (Lescot, 2002) to check whether our dyads matched any well-known, experimentally verified stress motifs. Out of the 28 dyads listed above, we found a number of heat shock factors as well as a GT-1 factor. A list of these dyads and their *cdr* scores as well as their annotation in either the PlantCARE databases can be seen in Table 13.

Sequence id	**score**
Q7XVS5	42.7
Q7XT99	26.7
Q7XCS4	24.8
Q7X7K5	20.8
Q94LH8	17
Q94LN0	16
Os04g0338000	14.7
Os01g0847600 (AKR1)	12
Q7XV17	10.9
Q8LMU5	10
Os01g0847800 (AKR3)	10
Q8GSK5	9
Q94LN1	9
Q94LH9	8.9
Q40648	8
Q7XV16	8
Os04g0447600	8
Q8H011	7
Q8H4J8	7
Os01g0847700 (AKR2)	7
Q7XQ45	7
Q7X8G7	6
Os04g0338100	5
Os04g0337500	5
Q7XTA2	5
Q7XQ49	4
Os06g0164000	3
Q7XV15	3
Os10g0419100	1
Q7XEJ9	1

Table 14. List of 30 AKR genes and their promoter scores. Genes studied by Turoczy et al. are in bold

6.3.1.2. Promoter set analysis

As a separate verification of the algorithm, thee 28 dyads were found back in the 30 Oryza sativa AKR promoters. A list of sequences and their score can be seen in Table 14. As we can see, the gene Os01g0847600 (AKR1) had a higher score than the other experimentally studied AKR genes, containing 9 dyads: ATATT{N3}AAATT, CTATA{N5}ATTTT,CTCTC{N43}GCGGC, GAAAA{N43}TTTAT, TATCA{N36}TAAAA, TCTCT{N44}GCGGC, TTTAA{N41}TTTTA, TTTAT{N19}TAAAA, and TTTGA{N19}TCAAA.

6.3.2. Dyad discovery in promoters of Oryza sativa glucanase, chitinase, pathogen-related gene orthologues

Our colleagues, Noémi Lukács and Veronika Pós studied the effects of wheat leaf rust on nearly isogenic lines (NIL) of wheat and discovered a number of genes which were responsible for the difference in protein expression patterns in wheat cells. These proteins belonged to the glucanase, chitinase, and pathogen-resistance family of genes. Therefore the question was what could be behind this phenomenon?

Wheat protein homologues	NCBI accession number	Top rice orthologue
Glucanases		
beta-1,3-glucanase	gi|68250406/AAY88778	Os01g71670
glucan endo-1,3-beta-D-glucosidase	gi|68349051/AAY96422	Os01g71670
putative glucan endo-1,3-beta-D-glucosidase	gi|3757682/CAA77085	Os01g71670
beta-1,3-glucanase precursor	gi|61657664/CAI64809	Os01g71380
endo-beta-1,3-glucanase	gi|4741846/AAD28732	Os01g51570
(1,3;1,4) beta glucanase	gi|109150348/BAE96089	Os05g31140
Chitinases		
chitinase 1	gi|18146825/BAB82471	Os10g39680
31.7 kDa class I endochitinase-antifreeze protein precursor	gi|12407647/AAG53609	Os03g30470
chitinase IV precursor	gi|4741848/AAD28733	Os04g41680
Pathogen related 1		
pathogenesis-related protein 1.1	gi|3702663/CAA07473	Os01g28500
pathogenesis-related protein 1	gi|14334165/AAK60565	Os01g28500
pathogenesis related-1	gi|30144637/AAP14676	Os01g28500
Pathogen related 9		
Peroxidase	gi|732974/CAA59486	Os07g48050
peroxidase 2	gi|57635149/AAW52716	Os07g48050
peroxidase 6	gi|57635157/AAW52720	Os02g14430
Peroxidase	gi|732972/CAA59485	Os07g48050
peroxidase precursor (WP2)	gi|730298/Q05855	Os07g48050
Pathogen related 5		
thaumatin-like protein TLP8	gi|14164983/AAK55326	Os03g45960
Barperm1 thaumatin-like protein	gi|2454602/AAB71680; gi|20257409/AAM15877	Os12g43490
thaumatin-like protein	gi|1321999/CAA66278	Os12g43490
thaumatin-like protein	gi|14334171/AAK60568	Os12g43430
Pathogen related 4		
pathogenesis-related protein	gi|1588926/2209398A	Os11g37950
putative vacuolar defense protein - wheatwin 5	gi|45862004/AAS78780	Os11g37950
Wheatwin-1 precursor (PR 4a)	gi|34925030/O64392	Os11g37950

Table 15. Names and accession numbers of wheat genes used in promoterome analysis

Since the whole genome sequence of wheat was not yet available, we performed the promoter analysis on the promoter sequences of the Oryza sativa homologues of the aforementioned wheat genes and see wheat kind of regulatory elements could be in the

background of this difference in gene expression. We used Oryza sativa as a model organism, since it is a relative of wheat. A list of the wheat genes used in this analysis can be seen in Table 15.

Motif	Sequence	Annotation	Sequence number	Pos. upstream ATG start	Reference
Glucanases					
Box-W1	TTGACC	WRKY1 (U48831) protein binding site	Glucanase 1	984 -	U48863
			Glucanase 21	9 +	
			Glucanase 21	1337 -	
			Glucanase 21	1044 +	
EIRE	TTCGACC	elicitor-responsive element	Glucanase 1	485 -	X69794
MBS	TAACTG	MYB binding site involved in drought-inducibility	Glucanase 1	20 -	D13044
			Glucanase 1	1308 +	U14599
			Glucanase 1	822 +	
			Glucanase 1	310 -	
			Glucanase 1	950 -	
PR1					
WUN-motif	TCATTACGAA	wound-responsive element	PR1-1	500 +	X98521
AuxRR-core	GGTCCAT	cis-element involved in auxin responsiveness	PR1-1	806 +	D85911
PR9					
ABRE	TACGTG	cis-element involved in ABA responsiveness	PR9-7	253 -	U01377
	TACGTG	cis-element involved in ABA responsiveness	PR9-7	534 +	U01377
	TACGTG	cis-element involved in ABA responsiveness	PR9-7	458 +	U01377
	CGCACGTGTC	cis-element involved in ABA responsiveness	PR9-7	1193 +	L19119
CCAAT-box	CAACGG	MYBHv1 binding site	PR9-1	454 -	X58339
			PR9-1	1049 +	
			PR9-1	870 -	
PR5					
Box-W1	TTGACC	WRKY1 (U48831) protein binding site	PR5-9	833 -	U48863
			PR5-9	927 -	
			PR5-9	896 -	

Table 16. List of wounding and fungal elicitor related regulatory elements in promoters of Oryza sativa orthologues of wheat gene families (+ = sense, - = antisense)

6.3.2.1. Selection of promoters

The Oryza sativa gene homologue of each wheat protein was found first by tblastn at the NCBI website. The gene sequence was then blasted at the Gramene website (blastn) to find significant paralog genes (Clark, 2007). It was possible that there were multiple gene hits for every wheat gene in Oryza sativa; therefore the 2 Kbp promoter region was extracted for all hits where the e-score was under 10^{-60}.

In such a way 29 promoters were retrieved for the 6 wheat glucanases, 19 promoters for the 3 wheat chitinase genes, 5 promoters for the 3 wheat PR1 genes, 20 promoters for the 5 wheat PR9 genes, 13 promoters for the 4 wheat PR5 genes, and 3

promoters for the 3 wheat PR4 genes. This makes 91 promoter sequences in total. However, in the further analysis, only the best orthologue hit was used for each gene. In total, 13 Oryza sativa orthologue genes' promoters were used in the analysis, since some of the wheat genes had the same Oryza sativa genes as their top orthologue hit which can also be seen in the third column of Table 15.

6.3.2.2. Promoter analysis at the PlantCARE database

All of the promoters were analyzed at the PlantCARE database (Lescot, 2002), and a number of motifs were found which played a role in wounding response and fungal elicitation. These motifs are listed for each gene family in Table 16.

6.3.2.3. Application of our algorithm to the promoter sets

After these preliminary analyses, we analyzed this promoter set with our own algorithm. In order to do this we split up our 91 stress orthologue promoters in the following way: the 13 best orthologues were partitioned into a positive learning promoter set, while the remaining orthologues were partitioned into a negative learning promoter set comprised of 78 sequences. The reasoning behind this was that we wanted to see what kind of regulatory elements were present in the most homologous promoters, which were the ones reasonably responsible for response to wounding and fungal infection.

In the analysis we studied tetramer dyads because of the small size of the input learning promoter sets. Also, since there were 6 times as many negative promoters as positive ones, we had to define a corrected version of the *cdr* score equation:

$$(9)\ cdr = \frac{\lambda \cdot N_{positive} - N_{negative.}}{\lambda \cdot N_{positive}}$$

In Equation 9 λ is a correction coefficient used if the size of the positive learning promoter set is different from the negative one. $\lambda = n_-/n_+$, where n_+ is the number of promoters in the positive learning set, and n- is the number of promoters in the negative learning promoter set. In our case, λ is equal to 6 (78/13). We selected the top 263 dyads which had a *cdr* score value above or equal to 0.9. This list is given in Supplementary Table 4.

Out of these 263 statistically significant dyads we selected those ones which were either present in the promoters of at least four of the 6 gene families (a minimal majority

of the glucanase, chitinase, PR1, 4, 5, or 9 genes), or were found in the PLACE database and matched motifs which were well-known biotic stress motifs. Overall there were 28 such dyads. These results can be seen in Supplementary Table 5. Here we can see the dyad identifier, it's sequence, *cdr* score, which promoter it was found in, and the position upstream of the ATG start, as well as it's annotation in the PLACE database (if applicable).

Gene id	Function	Score
Os01g51570	Glycosyl hydrolases family 17	133.225
Os05g31140	Glycosyl hydrolases family 17	120.744
Os02g14430	bacterial-induced peroxidase precursor	118.604
Os07g48050	peroxidase POC1	118.302
Os04g41680	Chitinase class I, putative	117.263
Os03g45960	putative antifungal zeamatin-like protein	111.513
Os12g43430	thaumatin-like protein precursor	107.5
Os01g28500	SCP-like extracellular protein, putative	101.225
Os12g43490	thaumatin-like protein TLP7	100.959
Os01g74440	SRF-type transcription factor (DNA-binding and dimerisation domain), putative	93.1897
Os01g71670	Glycosyl hydrolases family 17	92.3825
Os05g33970	hypothetical protein	91.45
Os07g31140	3-oxo-5-alpha-steroid 4-dehydrogenase, putative	88.3611
Os03g11010	probable integral membrane protein - rice	85.204
Os06g49660	benzoyl coenzyme A: benzyl alcohol benzoyl transferase	84.1587
Os02g41690	retrotransposon protein, putative, unclassified	83.3516
Os03g63240	putative resistance complex protein	82.6103
Os04g46330	hypothetical protein	78.9016
Os01g74110	ZIP Zinc transporter	77.6698
O*s01g71380*	Glycosyl hydrolases family 17	75.5286
Os06g18960	Similar to embryogenesis transmembrane protein - maize	75.1659
Os05g28090	Similar to At1g70760	75.0921
Os03g03300	hypothetical protein	74.177
Os08g42150	hypothetical protein	73.7738
Os11g43530	Glutaredoxin	73.4476
Os01g41270	F-box domain, putative	72.3397
Os05g49030	Ribosomal L18ae protein family, putative	70.4246
Os08g25490	FAD binding domain, putative	70.3111
Os10g39680	Chitinase	70.0504
Os10g22460	Protein phosphatase 2C, putative	69.9722

Table 17. Top 30 Oryza sativa promoters from the Oryza sativa promoterome search for genes involved in biotic response and pathogen resistance. Genes in bold play a role in biotic stimulus, while genes in italics belong to original positive learning set

Some of these dyads were found to match the WAR motif (wounding activating region) as well as the RSR motif („root specific region" motif) (Elliot, 1998). The H-box motif was also found to correspond to 3 dyads. This motif is known to take part in wounding and abiotic stress (Mhiri, 1997). A heat shock element, HSE was also found in two copies, which also takes part in heat shock and pathogen response (Pfitzner, 1988).

6.3.2.4. Oryza sativa promoterome search for other biotic stress resistant genes

As yet another independent test of our algorithm we ran a promoterome search on the entire Oryza sativa promoterome with the top 263 dyads found in the learning phase (minimum occurence 5, minimum *cdr* score of 0.9, spacer wobbling of ±1bp) in order to find other promoters whose dyad content might be similar and therefore might have a similar function in resistance to pathogens.

In Table 17 we can see the top 30 genes whose promoters had the highest score from the promoterome search. As can be seen, 11 of the 13 promoters (shown in italics) in the positive learning promoter set were found back in the promoterome search amongst the top 30 promoters. Besides this, 4 other Oryza sativa genes were found which play a role in response to biotic stimuli (shown in bold) according to the MSU Osa1 6.1 Annotation. Besides these genes 6 other genes were found which were either poorly annotated or described as hypothetical proteins (Os05g33970, Os04g46330, Os06g18960, Os05g28090, Os03g03300, and Os08g42150). Therefore we may reason that these genes could possibly play a role in response to biotic stress and pathogens.

6.3.2.5 Promoter analysis of biotic stress genes in wheat

As of August, 2010, the 5x coverage of the wheat genome had been achieved, with at least one read for 95% of the genome (ScienceDaily, 2010). Therefore we decided to study the distribution of dyads in the promoters of the 22 gene sequences for 24 glucanase, chitinase, and PR1, PR4, PR5, and PR9 proteins studied by our colleauges, Pós et al. The gene sequences were extracted from the cerealsDB.uk.net database. Here each gene was blasted against the sequence database and a cutoff e-score of 10^{-100} was applied to select sequence hits. The hit with the largest overlap with the gene query was selected. Since the database sequences corresponded to only short sequence reads, the hits were assembled into a longer query sequence which were be re-blasted. Where possible, a 2 Kbp upstream sequence was extracted. These promoter sequences were to be used in the positive training set. A set of 11 wheat promoters was downloaded from the European Promoter Database, which were maximum 2 Kbp long. A list of positive and negative training promoters, their annotation and lengths can be seen in Table 18.

Positive training set		
protein ID	**annotation**	**length**
AAY88778	beta-1,3-glucanase	1163
AAY96422	beta-1,3-glucanase	720
CAA77085	glucan endo-1,3-beta-D-glucosidase	529
CAI64809	putative glucan endo-1,3-beta-D-glucosidase	361
AAD28732	beta-1,3-glucanase precursor	128
BAE96089	endo-beta-1,3-glucanase	888
ABB96917	(1,3;1,4) beta-glucanase	770
BAB82471	chitinase 1	989
AAG53609	31.7 kDa class I endochitinase-antifreeze protein precursor	611
AAD28733	chitinase IV precursor	590
CAA07473	pathogenisis-related protein 1.1	390
AAK60565	pathogenesis-related protein 1	560
AAP14676	pathogenesis related-1	792
CAA59486	peroxidase	994
AAW52716	peroxidase 2	798
AAW52720	peroxidase 6	387
CAA59485	peroxidase	813
Q05855	PER1_WHEAT RecName: Full=Peroxidase; AltName: Full=WP2; Flags: Precursor	no hits
AAK55326	thaumatin-like protein TLP8	773
AAM15877	thaumatin-like protein	686
CAA66278	thaumatin-like protein	490
AAK60568	thaumatin-like protein	524
2209398A	pathogenesis-related protein	no hits
AAS78780	putative vacuolar defense protein	251
O64392	WHW1_WHEAT RecName: Full=Wheatwin-1; AltName: Full=Pathogenesis-related protein 4a; AltName: Full=Protein 0.14; Flags:	no hits
Negative training set		
ID	**annotation**	**length**
EP07001	Ta histone H3	185
EP07002	Ta histone H4	668
EP17004	Ta HMW glutenin	385
EP26035	Ta LMW glutenin 1D1	938
EP14002	Ta a'/b'gliadin 1215	517
EP14003	Ta a'/b'gliadin 8233	2000
EP14004	Ta a'/b'gliadin 8142	845
EP24010	Ta g' gliadin B	428
EP17006	Ta LHC cab-1	1813
EP29007	Ta RuBPCss	653
EP35062	Ta carboxypept. Y	1007

Table 18. ID, annotation and length of positive and negative promoters in training sets used in wheat

Tetramer dyad	PLACE annotation	rice dyad with common motif
AAAA{N12}CCAT		GGTC{N22}AAAA AAAA{N11}GGCC GGAC{N31}AAAA AAAA{N3}CGGC AAAA{N9}AGGC
AAAA{N5}CCTA		GGTC{N22}AAAA AAAA{N11}GGCC GGAC{N31}AAAA AAAA{N3}CGGC GATC{N27}CCTA AAAA{N9}AGGC
AAAC{N5}AAAT		GCTG{N12}AAAT AAAT{N45}CTTC AAAT{N1}GCTG
AACT{N4}AAAT		GCTG{N12}AAAT AAAT{N45}CTTC AAAT{N1}GCTG CGAA{N27}AACT
AAGC{N0}TAGC		AAGC{N13}GTAG AAGC{N34}TGCC
AGCT{N0}AGCT		TGCT{N39}AGCT
AGCT{N1}GCTA		TGCT{N39}AGCT
AGCT{N2}CTAG		TGCT{N39}AGCT
AGTT{N8}CATG		CATA{N44}AGTT AGTC{N28}AGTT CCCC{N42}CATG GGCA{N3}AGTT CCCC{N50}AGTT
ATAT{N0}GCAT		ATAT{N23}CGGA CGGT{N45}ATAT ATAT{N17}GCGG ATAT{N35}CGCG ATAT{N32}GGGG ATAT{N23}GGAG ATAT{N22}TCGG ACGG{N17}ATAT CCCC{N29}GCAT GGAC{N2}ATAT
ATGA{N0}TGAA		ATGA{N43}CGAG ATGA{N33}TTAC TCTG{N44}TGAA
ATGC{N0}ATGC	RYRE repeat	GAAC{N37}ATGC CCCT{N4}ATGC ATGC{N51}TCAG ATGC{N7}CCGA ATGC{N41}GAGT
CACA{N7}AACC		CCCC{N45}AACC CTAT{N9}CACA CTAA{N19}CACA CACA{N21}CAAG CACC{N27}CACA
CCAC{N3}CACA		CCAC{N7}TATA ATGT{N16}CCAC CTAT{N9}CACA CTAA{N19}CACA CACA{N21}CAAG CACC{N27}CACA
CCAC{N8}AATT		CCAC{N7}TATA ATGT{N16}CCAC
CGTA{N0}CGTA	copper resp. element	
CTAG{N0}CTAG	copper resp. element	
CTAG{N1}TAGC		
CTAT{N1}CGTA		TAGT{N43}CTAT CTAT{N9}CACA CTAT{N33}GTGT
CTAT{N2}GTAC		TAGT{N43}CTAT TAAA{N34}GTAC GTAC{N1}ATTC AGTG{N43}GTAC CTAT{N9}CACA TGCA{N1}GTAC TGGA{N43}GTAC CTAT{N33}GTGT GTAC{N39}AATA
GAAA{N0}ATTC	GATA box	GAGT{N50}GAAA ATTC{N42}CATA ATTC{N4}ATCC ATTC{N21}CTTA ATTC{N4}GGAT GTAC{N1}ATTC GTAA{N21}ATTC CTTG{N33}ATTC TACT{N33}ATTC GGGA{N21}GAAA GTTA{N6}ATTC
GAAA{N9}CAAA		GAGT{N50}GAAA TGTC{N17}CAAA GGGA{N21}GAAA ACGG{N26}CAAA CAAA{N45}GAGA
GCAG{N13}AAAT		GCTG{N12}AAAT AAAT{N45}CTTC AAAT{N1}GCTG TAGG{N14}GCAG GCAG{N10}AATA
GCTA{N0}GCTA		
GCTA{N1}CTAG		
GTTC{N1}TGCA		GTTC{N42}TTTT GTTC{N41}TTTC TGCA{N1}GTAC TGCA{N35}GGTA TCTT{N30}TGCA
TAGC{N0}TAGC		
TAGC{N1}AGCT		TGCT{N39}AGCT
TAGC{N2}GCTA		
TATA{N2}TACA		CCAC{N7}TATA AATG{N18}TACA TACA{N27}AGAG TATA{N49}CCTT GGAT{N21}TATA GGGA{N27}TACA TATA{N30}TTCG TATA{N24}CGCG TATA{N32}GGGG AGTC{N2}TATA
TATG{N0}CATG	RYRE repeat	TATG{N19}ATAG ATTG{N7}TATG TATG{N46}AGTG TATG{N35}CGGA CCCC{N42}CATG GTCA{N26}TATG TACC{N50}TATG TATG{N47}GTGT TATG{N23}GAGA TATG{N21}GGAG
TATG{N1}ATGC		TATG{N19}ATAG ATTG{N7}TATG TATG{N46}AGTG GAAC{N37}ATGC TATG{N35}CGGA CCCT{N4}ATGC GTCA{N26}TATG ATGC{N51}TCAG ATGC{N7}CCGA TACC{N50}TATG TATG{N47}GTGT TATG{N23}GAGA TATG{N21}GGAG ATGC{N41}GAGT
TCCA{N12}CATG		AATG{N37}TCCA TTAG{N39}TCCA CCCC{N42}CATG
TGCA{N14}AAAT		GCTG{N12}AAAT TGCA{N1}GTAC TGCA{N35}GGTA AAAT{N45}CTTC AAAT{N1}GCTG TCTT{N30}TGCA
TGCA{N14}CCAT		TGCA{N1}GTAC TGCA{N35}GGTA TCTT{N30}TGCA
TGGA{N0}ATTC	EEC element	ATTC{N42}CATA GAAT{N33}TGGA TGGA{N19}GTTG ATTC{N4}ATCC ATTC{N21}CTTA ATTC{N4}GGAT GCAC{N28}TGGA GTAC{N1}ATTC ACGT{N48}TGGA GTAA{N21}ATTC CTTG{N33}ATTC TACT{N33}ATTC GTTA{N6}ATTC TGGA{N43}GTAC

Table 19. Wheat dyads found by our algorithm with motif from corresponding rice dyads

The Windows version of the Dyadscan program was run with the following parameters: minimum occurence in positive training set: 5, minimum *cdr* score: 1.0, maximum spacer length: 52. Tetrads were searched for, because of the small size of the training promoter set.

A total of 36 dyads were found corresponding to these parameters, and can be seen in Table 19. We performed cluster analysis to whether our dyads form more specific clusters. We found that 13 of the 36 dyads formed 2 clusters with 4 and 9 members, respectively. The consensus sequence for these clusters are ATATGCATGC and GCTAGCTAGCTAG.

	W-box	**TC repeat**	**GCC box**	**ERE motif**	**E-box**	**S-box**	**WUN motif**	**JERE motif**
AAY88778	406 +	490 +						
AAY96422	155 +							
CAA77085		438 -, 329 +						
CAI64809			299 -					
AAD28732								
BAE96089			152 +	172 +				
ABB96917	216 -	509 +			669 +	344 +		
BAB82471							922 +	
AAG53609								
AAD28733					473 +			
CAA07473								
AAK60565								
AAP14676						379 +		
CAA59486	462, 524 -							
AAW52716	524 -							
AAW52720								
CAA59485		654 -	344 -					
AAK55326		511 +						
AAM15877	439, 592 +							548 +
CAA66278		420 +			295 -			
AAK60568	371 -							
AAS78780								

Table 20. Position and orientation (+ = sense, - = antisense) of 8 known biotic stress motifs in the promoters of the 22 glucanase, chitinase, PR1, 4, 5, and 9 promoters

When compared to elements in the PLACE database we found that 6 of our dyads matched known elements, among them two RYRE-repeat elements, two copper responsive elements, a GATA-box, and an EEC element. We also compared the dyads we found with those dyads found in the promoter analysis of the 91 rice genes homologous to the wheat genes studied in the present analysis. We checked whether the head or tail

motif the 36 wheat dyads matched the head or tail motif of one of the rice dyads. As seen in Table 19, 29 of the top 36 (80.6%) wheat dyads had such a matching motif.

The promoter sequences of 22 genes were analyzed at the PlantCARE database. Here the position and orientation of 8 biotic stress motifs were analyzed (W-box, TC repeat, GCC box, ERE motif, E-box, S-box, WUN motif, and JERE motif). Out of 22 promoters, 16 contained 1 or more of these motifs. These results can be seen in Table 20.

7. Discussion of results

The present Ph.D. thesis presents an enumeration-based algorithm for the prediction of putative regulatory element dyads in co-regulated genes. It was tested in and applied to the complete promoterome analysis of two plant species, one, a dicot (*Arabidopsis thaliana*) and the other a monocot (*Oryza sativa*), for genes involved in abiotic stress. Furthermore it was also tested in two seperate cases in the promoter analysis and dyad prediction of two specialized sets of Oryza sativa promoters involved in abiotic and biotic stress. Finally, it was also compared to two well-known motif discovery programs (YMF and dyad-analysis), and was shown to give better results in predicting putative abiotic stress-related dyads.

The algorithm was tested on different sets of genes involved in abiotic stress response. However, the scope of such studies can be widened to involve different sets of genes in all kinds of organisms which regulate different types of physiological processes or biochemical pathways, such as the regulation of the cell cycle, development, or seed maturation, because of the basic concept of analyzing common regulation machinery mirrored in TFBS content. For example, in rice a number of biochemical and physiological pathways overlap with each other (Cooper, 2003), therefore making it possible to compare TFBS's found in one gene set with those found in another.

7.1. Dyad motif lengths, input promoters, spacer wobbling

What makes the algorithm special is that it involved finding motif dyads seperated by a spacer region of a characteristic length. The lengths of the motifs making up the dyads could be adjusted (e.g. tetramers or pentamers) for smaller or larger input promoter sets. For example, since the rice AKR, glucanase, chitinase and PR genes all belonged to smaller families, the algorithm therefore had to be adjusted to take this into account. These gene families were smaller since they all play more specific roles than abiotic stress response in general which involves many more genes.

The spacer region itself could be adjusted to take wobbling into account between the head and tail motifs. This is because in the case of some TF's which are made up of

subunits, the subunits sometimes do not fit perfectly to their respective binding sites, but rather wobble 1-2 bp. As we could see in the promoter analysis of Arabidopsis, the optimum dyad set allowed a spacer wobbling of ±2 bp, which shows that the algorithm is realistic in this sense. The concept of searching for dyad elements and REPs instead of simple oligomers was used, since individual elements pairs found in the promoters of co-regulated genes would tend to form dyads, or be parts of regulatory networks responsible for the regulation of those specific genes whose promoters they were found in, and which played roles in a specific biochemical pathway or physiological process. Should the input learning promoter sets be disproportionate in their number of promoters, this can be accomodated for by using a correction coefficient defined in Equation 9 used for the calculation of the *cdr* score.

Although the p-value for the optimal dyad set in rice was not highly significant, this is only one factor used to validate the algorithm. The AUC peak however was consistently present for all wobbling factors from 0 to 5 bp, indicating that the parameter values of the corresponding dyad set is still statistically robust.

7.2. Experimental verification of finding motifs

The algorithm was verified by finding a number of putative dyads which had significant sequential overlaps with experimentally verified stress motifs as described in the literature. In the case of the glucanase, chitinase, and PR genes, these are the W1-box, the WAR motif, the WUN motif, the EIRE motif (Elicitor Responsive Element), the H-box, the RSR motif, and the HSE. However, when the stress learning promoter set in *Arabidopsis* and *Oryza sativa* were both replaced with a set of randomly selected promoters, only 2 and 6 significant dyads were found, which however did not match any known motifs in the PLACE database.

Furthermore, as mentioned in the subsection „Regulatory element networks", 22 TRANSFAC/PLACE motifs were found to be part of regulatory element pairs found by our algorithm. We looked for other kinds of abiotic stress motifs annotated in the PLACE database which occured in the *cor* and *erd* promoters, and found that 5 variants of the ABRE, MYB, MYC, and AS-1 motifs were missed by our algorithm. This means that in

total, our algorithm is capable of finding 81.5% (22/27) of all motifs. The reason for this is that these 5 motif variants were not present in the initial stress learning promoter set in Arabidopsis.

As a seperate line of support for our approach, we looked through the literature for cases where either certain parts of the promoter sequence of the genes found by the Arabidopsis promoterome search were deleted, and as a subsequent result, the gene lost it's stress-inducibility, or, where certain parts of the promoter could be localized in stress response. This was done to show that at least some of the dyads found by the algorithm have biological function, since without them the gene whose promoter contains them are unable to function properly. For example, Alonso-Blanco (Alonso-Blanco, 2005) defined a 160 bp minimal promoter for DREB1C (At4g25470), otherwise known as CBF2, in which our algorithm found two dyad elements: AAATA{N36}ATCTT and AAACT{N29}CATTT at positions -57, and -19. In the AGRIS database we found that the tail motif of the dyad AAAAA{N21}CACGT contains an ACGT core element, which is known to take part in abiotic stress response in a number of genes in Arabidopsis (Simpson, 2003).

7.3. Parameterization considerations

Since we applied the algorithm to a different plant species we were able to draw certain conclusions as to how to apply the algorithms to different species which have genomes of different sizes and genomic structural characteristics.

One of these main factors is the size of the genome that is being analyzed as well as its repetitive element content, which are somewhat related to each other. Although we had to discard 5 of the original Arabidopsis positive learning promoters, we found that analyzing the rice genome was harder than the Arabidopsis genome because of the differences in genome size and repetitive element content. Arabidopsis has a relatively small and compact genome (125 Mbp) with relatively less repetitive elements (around 10%) than rice (Casacuberta, 2003), which has a genome size of 430 Mbp and a repetitive element content of around 35%.

If it just so happens that some of the input promoters contain a number of repetitive elements, then, since the algorithm we applied is an enumeration method, and since repetitive elements are fairly well conserved, we easily picked up a lot of sequences which turned out to be repetitive, and therefore found other genes whose promoters also contained such repetitive elements, thereby skewing our results. Therefore it is highly advisable to either purge promoters of such repetitive elements, using a sequence mask, or discard promoters with highly repetitive regions. This is especially true, if one wishes to apply the algorithm to a genome similar to that of wheat which is 17,000 Mbp and contains a proportionately high amount of repetitive elements.

Another main factor which must be taken into consideration during the application of the program is the selection of promoters for the test sets, as relatively few of the positive test set promoters were found back in Arabidopsis and rice. Therefore we ran the algorithm on both species with the following setup: the positive test set was swapped with the positive learning set, and the negative test set was also swapped with the negative learning set. The same optimal parameterization was used to find the optimum dyad sets in both species which were deduced in the first run. In Arabidopsis this was using the top 81 dyads which occured at least 14 times in the new positive learning set with a spacer wobbling of ±1 bp and a minimum *cdr* score of 0.9. In rice this corresponded to a dyad set of 38 elements with a *cdr* score of at least 0.89, an occurence of at least 9 times, and a wobbling factor of 0 bp.

According to the switched promoterome search, the top 2,400 Arabidopsis promoters were selected and analyzed. Only 568 of these genes were found to be in common with those found in the original promoterome search (18.3% and 23.7% respectively). In rice, the top 1200 promoters were analyzed in a similar fashion to the original promoterome search. 226 genes were found to be in common to both rice promoterome searches, which is 4.9% and 18.8% respectively. This means that although the algorithm's success rate (PPV) was 78.6% in Arabidopsis and 98.7% in rice, it might mean that different regulatory elements are picked out by the algorithm depend on the promoters originally selected.

This indicates that our method can not give exhaustive results. The use of learning and test sets can provide differentiating dyads but will not give all of the possible ones.

As mentioned in the Results, when the test and the learning promoter sets were switched in both species, we found that less than 25% of the promoters found in both the original and switched promoterome searches were the same. As in the case of the motif searches in the *cor* and *erd* genes, this is because not all stress motifs were present in the original learning sets which we did the dyad search with. This means that some variants of existing dyads or new motifs altogether were present in the test stress promoter set which we used as the new learning set in the switched run. Indeed, in the switched run in *Arabidopsis*, 45 dyads were found by the algorithm, and compared to the 81 dyads of the original run, 37 (82.2%) had a Hamming distance less than or equal to 3, or either their full head or tail motif matched the full head or tail motif of one of the 81 dyads. The algorithm can be developed in such a way that promoters used in the learning and test sets can be switched around in order to get differentiated results. Dyads coming from most or all runs can then be selected for further study.

The rice promoterome search found 4,600 promoters whose putative stress dyad content was significant. Out of these genes, 1,456 had no Affymetrix probe id because these genes have an unknown function. Since the PPV of the rice promoterome search was 98.66%, this means that about 1,436 of these genes could be predicted to play a role in abiotic stress. A further 1,245 genes found by the algorithm in rice which had an Affymetrix probe id were termed as either hypothetical or expressed proteins. However, since these genes showed an at least twofold expression level change in an average of 3.4 experiments per gene according to the GEO datasets, they can also be annotated as stress resistance genes. Overall, this means that these 2,681 (1,436+1,245) genes could be newly annotated as abiotic stress genes, meaning 6.7% of the rice genome.

7.4. Outlook

In the future, candidate plant crop genomes such as wheat, barley, potato, or maize can be targeted for similar promoterome analyses, especially wheat (Feuillet, 2007), since there is a current plan to determine its genome sequence, of which the first draft read is available (cerealsDB.uk.net). Information drawn from the analysis of the rice promoterome could be used in comparative studies with other monocot or grass species

(Sasaki, 2008; Walia, 2009). This is especially the case since it has been shown that many genes are colinear in grass species (Gale, 2001).

Furthermore, differences in dyad/motif content between relative species in the promoters of orthologous genes may throw light on the molecular genetic background which is involved in speciation events (Wang, 2010). This is the case for example between the two dicot species Arabidopsis and *Medicago truncatula*, where only 62% of orthologs predicted through phylogenetic analysis had a similar expression profile (Benedito, 2008). This approach is especially called for since we know that the assumption that orthologous genes should have the same expression profile is false.

7.5. Website

Finally, a website has been constructed where users may download a stand-alone Windows desktop application to find putative regulatory dyad elements in input promoter sequences that they themselves supply. The website can be found at http://bhd.szbk.u-szeged.hu/dyadscan/ (id: totto, password: pwd1). Here the user may give a number of options such as motif length, minimum dyad occurence in the input promoter set, and minimum *cdr* score. The result of the algorithm is a list of dyad sequences which can then be used in further analysis.

8. Major scientific findings

This doctoral thesis presents the development and application of our own putative dyad prediction algorithm. The algorithm is capable of predicting regulatory element pairs, or dyads in a set of co-regulated genes promoters. Based on the putative dyad content found by the algorithm in the learning phase we were capable of predicting a number of new genes to take part in abiotic stress. The algorithm was tested in a set of abiotic stress genes in Arabidopsis and Oryza sativa as well as two different sets of stress genes in this species. Our new scientific results can be summarized as follows:

1. In Arabidopsis, the algorithm predicted 81 new putative regulatory dyad elements. These elements had a minimum occurence of 14 in the input positive learning promoter set, with a minimum *cdr* score of 0.9, and a spacer wobbling of 2 between the individual motifs in the dyad.
2. The algorithm predicted 38 new putative regulatory dyad elements in rice. These elements had a minimum occurence of 9 in the input positive learning promoter set, with a minimum *cdr* score of 0.89, and a spacer wobbling of 0 between the individual motifs in the dyad.
3. A promoterome search was also completed for the 81 putative dyad elements in Arabidopsis. The result is that we predicted 3100 genes in Arabidopsis to be newly involved in abiotic stress. 78.6% of the 1542 genes with Affymetrix ids were correctly predicted to be involved in abiotic stress, and 49 of them were also annotated as hypothetical genes. 1224 of the remaining genes with no Affymetrix id are then predicted to be involved in abiotic stress. Therefore these genes can also be predicted to be new abiotic stress genes in Arabidopsis according to our algorithm. Compared to our algorithm, the well-known YMF and dyad-analysis programs were only capable of predicting dyads of which only 3.1% and 3.6% were involved in abiotic stress.
4. The aforementioned 38 dyad elements were searched back in the entire rice promoterome. Based on the percentage of non-stress promoters to total promoters found by the algorithm, 4600 rice genes were selected as abiotic-stress candidate genes. The algorithm predicted that 98.7% of the 3144 genes with an Affymetrix

id were involved in abiotic stress, which is a rather high positive prediction rate. 1245 of the 3144 genes were termed as hypothetical, while 1437 of the remaining genes without an Affymetrix id were also predicted to be inolved in abiotic stress.

5. The dyad prediction algorithm has its own website at http://bhd.szbk.u-szeged.hu/dyadscan/. Here the user can set the dyad motif length, maximum spacer length, and minimum positive learning promoter set occurence. The user can also upload a positive and negative learning promoter set from which the algorithm will predict putative dyad regulatory elements. The user can then save these results.
6. 28 putative regulatory dyad elements were found in a set of 24 rice AKR genes. These dyad had a minimum occurence of 7 in these genes with a minimum *cdr* score of 0.9. Based on these elements we were able to differentiate between three AKR genes studied in detail. The algorithm correctly predicted that the AKR1 gene (Os07g0671700) contained 9 newly predicted putative dyad elements compared to the two other genes.
7. The algorithm was also performed on a set of rice promoters for orthologs of glucanase, chitinase, PR 1, 4, 5, and 9 genes. Overall, 263 dyad elements were found with a minimum occurence of 5, and a minimum score of 0.9, out of which 28 were present in at least 4 of the 6 gene families.
8. The algorithm was also tested on a set of promoters for glucanase, chitinase, PR 1, 4, 5, and 9 genes in wheat. We found that albeit with small alterations, 18 of the 36 predicted dyads corresponded to a dyad found in the analysis in rice.

9. Acknowledgements

I would hereby like to express my thanks to professor Dénes Dudits for giving me a Ph.D. position at the Biological Research Center. I would like to express my thanks to my supervisors Drs. János Györgyey and Sándor Pongor for their valuable help and advice for defining, developing, and refining the algorithm described in this thesis.

I would like to give thanks to my colleague Zoltán Zombori for providing data for drought stressed rice genes described in this thesis.

I would like also like to give thanks to Mária Sečenji and Krisztina Talpas for their help in learning laboratory techniques for RNA extraction, cDNA synthesis, and doing RT-PCR experiments.

I would furthermore like to give thanks to all the members of our workgroup and all of my colleagues who took part in the work, which was needed to make this Ph.D. thesis possible.

Finally, I would like to give thanks to Almighty God the Father without whose spiritual and supporting grace this thesis could not have been written.

10. References

Abe H, Yamaguchi-Shinozaki K, Urao T, Iwasaki T, Hosokawa D, Shinozaki K. Role of arabidopsis MYC and MYB homologs in drought- and abscisic acid-regulated gene expression. Plant Cell. 1997 Oct;9(10):1859-68.

Abe H, Urao T, Ito T, Seki M, Shinozaki K, Yamaguchi-Shinozaki K. Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling. Plant Cell. 2003 Jan;15(1):63-78.

A. V. Aho, B. W. Kernighan, P. J. Weinberger. The awk Programming Language Addison-Wesley, 1988

Alonso-Blanco C, Gomez-Mena C, Llorente F, Koornneef M, Salinas J, Martínez-Zapater JM. Genetic and molecular analyses of natural variation indicate CBF2 as a candidate gene for underlying a freezing tolerance quantitative trait locus in Arabidopsis. Plant Physiol. 2005 Nov;139(3):1304-12.

Amoutzias GD, Robertson DL, Van de Peer Y, Oliver SG. Choose your partners: dimerization in eukaryotic transcription factors. Trends Biochem Sci. 2008 May;33(5):220-9.

Bailey TL, Williams N, Misleh C, Li WW. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W369-73.

Barta E, Sebestyén E, Pálfy TB, Tóth G, Ortutay CP, Patthy L. DoOP: Databases of Orthologous Promoters, collections of clusters of orthologous upstream sequences from chordates and plants. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D86-90.

Benedito VA, Torres-Jerez I, Murray JD, Andriankaja A, Allen S, Kakar K, Wandrey M, Verdier J, Zuber H, Ott T, Moreau S, Niebel A, Frickey T, Weiller G, He J, Dai X, Zhao PX, Tang Y, Udvardi MK. A gene expression atlas of the model legume Medicago truncatula. Plant J. 2008 Aug;55(3):504-13.

Blanchette M, Tompa M. FootPrinter: A program designed for phylogenetic footprinting. Nucleic Acids Res. 2003 Jul 1;31(13):3840-2.

Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science. 2003 Feb 28;299(5611):1391-4.

Bryne JC, Valen E, Tang MH, Marstrand T, Winther O, da Piedade I, Krogh A, Lenhard B, Sandelin A. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res. 2008 Jan;36(Database issue):D102-6.

Bulyk ML. Computational prediction of transcription-factor binding site locations. Genome Biol. 2003;5(1):201.

Carmack CS, McCue LA, Newberg LA, Lawrence CE. PhyloScan: identification of transcription factor binding sites using cross-species evidence. Algorithms Mol Biol. 2007 Jan 23;2:1.

Casacuberta JM, Santiago N. Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. Gene. 2003 Jun 5;311:1-11.

CerealsDB.uk.net website: [http://www.cerealsdb.uk.net/]

Chaivorapol C, Melton C, Wei G, Yeh RF, Ramalho-Santos M, Blelloch R, Li H. CompMoby: comparative MobyDick for detection of cis-regulatory motifs. BMC Bioinformatics. 2008 Oct 27;9:455.

Chang WC, Lee TY, Huang HD, Huang HY, Pan RL. PlantPAN: Plant promoter analysis navigator, for identifying combinatorial cis-regulatory elements with distance constraint in plant gene groups. BMC Genomics. 2008 Nov 26;9:561.

Chen QK, Hertz GZ, Stormo GD. MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. Comput Appl Biosci. 1995 Oct;11(5):563-6.

Chen X, Guo L, Fan Z, Jiang T. W-AlignACE: an improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/ChIP-chip data. Bioinformatics. 2008 May 1;24(9):1121-8.

Chinnusamy V, Schumaker K, Zhu JK. Molecular genetic perspectives on cross-talk and specificity in abiotic stress signalling in plants. J Exp Bot. 2004 Jan;55(395):225-36.

Cooper B, Clarke JD, Budworth P, Kreps J, Hutchison D, Park S, Guimil S, Dunn M, Luginbühl P, Ellero C, Goff SA, Glazebrook J. A network of rice genes associated with stress response and seed development. Proc Natl Acad Sci U S A. 2003 Apr 15;100(8):4945-50.

Cserháti M. Usage of enumeration method based algorithms for finding promoter motifs in plant genomes. Acta Biol Szeged 2006, 50(3-4):145.

Cserháti M, Turóczy Z, Zombori Z, Cserző M, Dudits D, Pongor S, Györgyey J. Prediction of new abiotic stress genes in Arabidopsis thaliana and Oryza sativa according to enumeration-based statistical analysis, Mol Genet Genomics. 2011.

Das MK, Dai HK. A survey of DNA motif finding algorithms. BMC Bioinformatics. 2007 Nov 1

Defrance M, Janky R, Sand O, van Helden J. Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences. Nat Protoc. 2008;3(10):1589-603.

Defrance M, van Helden J. info-gibbs: a motif discovery algorithm that directly optimizes information content during sampling. Bioinformatics. 2009 Oct 15;25(20):2715-22.

Doi K, Hosaka A, Nagata T, Satoh K, Suzuki K, Mauleon R, Mendoza MJ, Bruskiewich R, Kikuchi S. Development of a novel data mining tool to find cis-elements in rice gene promoter regions. BMC Plant Biol. 2008 Feb 27;8:20.

Dudits, D., Heszky, L. Növényi biotechnológia és géntechnológia. Agroinform publishers, Budapest, 2003.

Dudits, D. A búza nemesbítésének tudománya. MTA Szegedi Biológia Központ – Winter fair Kft., Szeged, 2006.

Eddy SR. A probabilistic model of local sequence alignment that simplifies statistical significance estimation. PLoS Comput Biol. 2008 May 30;4(5):e1000069.

Elliott KA, Shirsat AH. Promoter regions of the extA extensin gene from Brassica napus control activation in response to wounding and tensile stress. Plant Mol Biol. 1998 Jul;37(4):675-87.

European Promoter Database: [http://www.epd.isb-sib.ch]

Fawcett T. ROC Graphs: Notes and Practical Considerations for Researchers: [http://home.comcast.net/~tom.fawcett/public_html/papers/ROC101.pdf]

Fehér A, Ötvös K. The physiology of somatic embryo induction: A stressful start: a review in: Advances in Plant Physiology, Ed.: Hemantaranjan, A. Scientific Publishers, Jodhpur, India, 2005-6.

Feuillet C, Eversole K. Physical mapping of the wheat genome: A coordinated effort to lay the foundation for genome sequencing and develop tools for breeders. Israel Journal of Plant Sciences 2007, 55, 307-313

Fujimori S, Washio T, Tomita M. GC-compositional strand bias around transcription start sites in plants and fungi. BMC Genomics. 2005 Feb 28;6(1):26.

Fujita M, Fujita Y, Maruyama K, Seki M, Hiratsu K, Ohme-Takagi M, Tran LS, Yamaguchi-Shinozaki K, Shinozaki K. A dehydration-induced NAC protein, RD26, is involved in a novel ABA-dependent stress-signaling pathway. Plant J. 2004 Sep;39(6):863-76.

Gale M, Moore G, Devos K. Rice--the pivotal genome in cereal comparative genetics. Novartis Found Symp. 2001;236:46-53; discussion 53-8.

Gao G, Zhong Y, Guo A, Zhu Q, Tang W, Zheng W, Gu X, Wei L, Luo J. DRTF: a database of rice transcription factors. Bioinformatics. 2006 May 15;22(10):1286-7.

Geisler M, Kleczkowski LA, Karpinski S. A universal algorithm for genome-wide in silicio identification of biologically significant gene promoter putative cis-regulatory-elements; identification of new elements for reactive oxygen species and sucrose signaling in Arabidopsis. Plant J. 2006 Feb;45(3):384-98.

Gunewardena S, Zhang Z. A hybrid model for robust detection of transcription factor binding sites. Bioinformatics. 2008 Feb 15;24(4):484-91.

Gómez-Porras JL, Riaño-Pachón DM, Dreyer I, Mayer JE, Mueller-Roeber B. Genome-wide analysis of ABA-responsive elements ABRE and CE3 reveals divergent patterns in Arabidopsis and rice. BMC Genomics. 2007 Aug 1;8:260.

Gramene website: [http://www.gramene.org/multi/blastview]

Grotewold E. Transcription factors for predictive plant metabolic engineering: are we there yet? Curr Opin Biotechnol. 2008 Apr;19(2):138-44.

Hannenhalli S. Eukaryotic transcription factor binding sites--modeling and integrative search methods. Bioinformatics. 2008 Jun 1;24(11):1325-31.

Hertz GZ, Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics. 1999 Jul-Aug;15(7-8):563-77.

Hertz GZ, Hartzell GW 3rd, Stormo GD. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. Comput Appl Biosci. 1990 Apr;6(2):81-92.

Hideg É, Nagy T, Oberschall A, Dudits D, Vass I. Detoxification function of aldose/aldehyde reductase during drought and ultraviolet-B (280-320 nm) stresses. Plant Cell & Environment. 26, 2003 513-522.

Higo K, Ugawa Y, Iwamoto M, Korenaga T. Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Res. 1999 Jan 1;27(1):297-300.

Hong JK, Hwang BK. Promoter activation of pepper class II basic chitinase gene, CAChi2, and enhanced bacterial disease resistance and osmotic stress tolerance in the CAChi2-overexpressing Arabidopsis. Planta. 2006 Feb;223(3):433-48.

Hruz T, Laule O, Szabo G, Wessendorp F, Bleuler S, Oertle L, Widmayer P, Gruissem W, Zimmermann P. Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. Adv Bioinformatics. 2008;2008:420747.

Immink RG, Kaufmann K, Angenent GC. The 'ABC' of MADS domain protein behaviour and interactions. Semin Cell Dev Biol. 2010 Feb;21(1):87-93.

Jakoby M, Weisshaar B, Dröge-Laser W, Vicente-Carbajosa J, Tiedemann J, Kroj T, Parcy F; bZIP Research Group. bZIP transcription factors in Arabidopsis. Trends Plant Sci. 2002 Mar;7(3):106-11.

Janky R, van Helden J. Discovery of conserved motifs in promoters of orthologous genes in prokaryotes. Methods Mol Biol. 2007;395:293-308.

Jin XF, Xiong AS, Peng RH, Liu JG, Gao F, Chen JM, Yao QH. OsAREB1, an ABRE-binding protein responding to ABA and glucose, has multiple functions in Arabidopsis. BMB Rep. 2010 Jan;43(1):34-9.

Guiltinan MJ, Marcotte WR Jr, Quatrano RS. A plant leucine zipper protein that recognizes an abscisic acid response element. Science. 1990 Oct 12;250(4978):267-71.

Joshee N, Kisaka H, Kitagawa Y. Isolation and characterization of a water stress-specific genomic gene, pwsi 18, from rice. Plant Cell Physiol. 1998 Jan;39(1):64-72.

Jung KH, Dardick C, Bartley LE, Cao P, Phetsom J, Canlas P, Seo YS, Shultz M, Ouyang S, Yuan Q, Frank BC, Ly E, Zheng L, Jia Y, Hsia AP, An K, Chou HH, Rocke D, Lee GC, Schnable PS, An G, Buell CR, Ronald PC. Refinement of light-responsive transcript lists using rice oligonucleotide arrays: evaluation of gene-redundancy. PLoS One. 2008 Oct 6;3(10):e3337.

Kechris KJ, van Zwet E, Bickel PJ, Eisen MB. Detecting DNA regulatory motifs by incorporating positional trends in information content. Genome Biol. 2004;5(7):R50.

Kel-Margoulis OV, Kel AE, Reuter I, Deineko IV, Wingender E. TRANSCompel: a database on composite regulatory elements in eukaryotic genes. Nucleic Acids Res. 2002 Jan 1;30(1):332-4.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature. 2003 May 15;423(6937):241-54.

Kim JC, Lee SH, Cheong YH, Yoo CM, Lee SI, Chun HJ, Yun DJ, Hong JC, Lee SY, Lim CO, Cho MJ. A novel cold-inducible zinc finger protein from soybean, SCOF-1, enhances cold tolerance in transgenic plants. Plant J. 2001 Feb;25(3):247-59.

Kim HJ, Kim YK, Park JY, Kim J. Light signalling mediated by phytochrome plays an important role in cold-induced gene expression through the C-repeat/dehydration responsive element (C/DRE) in Arabidopsis thaliana. Plant J. 2002 Mar;29(6):693-704.

Kim SY, Nam KH. Physiological roles of ERD10 in abiotic stresses and seed germination of Arabidopsis. Plant Cell Rep. 2010 Feb;29(2):203-9.

Kolchanov NA, Ignatieva EV, Ananko EA, Podkolodnaya OA, Stepanenko IL, Merkulova TI, Pozdnyakov MA, Podkolodny NL, Naumochkin AN, Romashchenko AG. Transcription Regulatory Regions Database (TRRD): its status in 2002. Nucleic Acids Res. 2002 Jan 1;30(1):312-7.

Kushwaha H, Gupta S, Singh VK, Rastogi S, Yadav D. Genome wide identification of Dof transcription factor gene family in sorghum and its comparative phylogenetic analysis with rice and Arabidopsis. Mol Biol Rep. 2010 Dec 16.

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science. 1993 Oct 8;262(5131):208-14.

Lenhard B, Sandelin A, Mendoza L, Engström P, Jareborg N, Wasserman WW. Identification of conserved regulatory elements by comparative genome analysis. J Biol. 2003;2(2):13.

Lescot M, Déhais P, Thijs G, Marchal K, Moreau Y, Van de Peer Y, Rouzé P, Rombauts S. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res. 2002 Jan 1;30(1):325-7.

Li N, Tompa M. Analysis of computational approaches for motif discovery. Algorithms Mol Biol. 2006 May 19;1:8.

Liang C, Jaiswal P, Hebbard C, Avraham S, Buckler ES, Casstevens T, Hurwitz B, McCouch S, Ni J, Pujar A, Ravenscroft D, Ren L, Spooner W, Tecle I, Thomason J, Tung CW, Wei X, Yap I, Youens-Clark K, Ware D, Stein L. Gramene: a growing plant comparative genomics resource. Nucleic Acids Res. 2008 Jan;36(Database issue):D947-53.

Lichtenberg J, Yilmaz A, Welch JD, Kurz K, Liang X, Drews F, Ecker K, Lee SS, Geisler M, Grotewold E, Welch LR. The word landscape of the non-coding segments of the Arabidopsis thaliana genome. BMC Genomics. 2009 Oct 8;10:463.

Liu J, Zhu JK. A calcium sensor homolog required for plant salt tolerance. Science. 1998 Jun 19;280(5371):1943-5.

Liu L, White MJ, MacRae TH. Transcription factors and their genes in higher plants functional domains, evolution and regulation. Eur J Biochem. 1999 Jun;262(2):247-57.

Liu X, Brutlag DL, Liu JS. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac Symp Biocomput. 2001:127-38.

Loots GG, Ovcharenko I, Pachter L, Dubchak I, Rubin EM. rVista for comparative sequence-based discovery of functional transcription factor binding sites. Genome Res. 2002 May;12(5):832-9.

Mahajan S, Tuteja N. Cold, salinity and drought stresses: an overview. Arch Biochem Biophys. 2005 Dec 15;444(2):139-58.

Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D108-10.

McGloughlin MN. Modifying agricultural crops for improved nutrition. N Biotechnol. 2010 Nov 30;27(5):494-504.

Mhiri C, Morel JB, Vernhettes S, Casacuberta JM, Lucas H, Grandbastien MA. The promoter of the tobacco Tnt1 retrotransposon is induced by wounding and by abiotic stress. Plant Mol Biol. 1997 Jan;33(2):257-66.

Mitsuda N, Ohme-Takagi M. Functional analysis of transcription factors in Arabidopsis. Plant Cell Physiol. 2009 Jul;50(7):1232-48.

Moore G, Devos KM, Wang Z, Gale MD. Cereal genome evolution. Grasses, line up and form a circle. Curr Biol. 1995 Jul 1;5(7):737-9.

Morris RT, O'Connor TR, Wyrick JJ. Osiris: an integrated promoter database for Oryza sativa L. Bioinformatics. 2008 Dec 15;24(24):2915-7.

Nakashima K, Kiyosue T, Yamaguchi-Shinozaki K, Shinozaki K. A nuclear gene, erd1, encoding a chloroplast-targeted Clp protease regulatory subunit homolog is not only induced by water stress but also developmentally up-regulated during senescence in Arabidopsis thaliana. Plant J. 1997 Oct;12(4):851-61.

Narusaka Y, Nakashima K, Shinwari ZK, Sakuma Y, Furihata T, Abe H, Narusaka M, Shinozaki K, Yamaguchi-Shinozaki K. Interaction between two cis-acting elements, ABRE and DRE, in ABA-dependent expression of Arabidopsis rd29A gene in response to dehydration and high-salinity stresses. Plant J. 2003 Apr;34(2):137-48.

Nguyen DH, D'haeseleer P. Deciphering principles of transcription regulation in eukaryotic genomes. Mol Syst Biol. 2006;2:2006.0012.

O'Connor TR, Dyreson C, Wyrick JJ. Athena: a resource for rapid visualization and systematic analysis of Arabidopsis promoter sequences. Bioinformatics. 2005 Dec 15;21(24):4411-3.

Oberschall A, Deák M, Török K, Sass L, Vass I, Kovács I, Fehér A, Dudits D, Horváth GV. A novel aldose/aldehyde reductase protects transgenic plants against lipid peroxidation under chemical and drought stresses. Plant J. 2000 Nov;24(4):437-46.

Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, Orvis J, Haas B, Wortman J, Buell CR. The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res. 2007 Jan;35(Database issue):D883-7.

Palaniswamy SK, James S, Sun H, Lamb RS, Davuluri RV, Grotewold E. AGRIS and AtRegNet. a platform to link cis-regulatory elements and transcription factors into regulatory networks. Plant Physiol. 2006 Mar;140(3):818-29.

Pattanaik S, Xie CH, Yuan L. The interaction domains of the plant Myc-like bHLH transcription factors can regulate the transactivation strength. Planta. 2008 Feb;227(3):707-15.

Pavesi G, Mauri G, Pesole G. In silico representation and discovery of transcription factor binding sites. Brief Bioinform. 2004 Sep;5(3):217-36.

Pesole G, Liuni S, D'Souza M. PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance. Bioinformatics. 2000 May;16(5):439-50.

Picot E, Krusche P, Tiskin A, Carré I, Ott S. Evolutionary analysis of regulatory sequences (EARS) in plants. Plant J. 2010 Oct;64(1):165-76. doi: 10.1111/j.1365-313X.2010.04314.x.

Pfitzner UM, Pfitzner AJP, Goodman HM. DNA sequence analysis of a PR-1a gene from tobacco: Molecular relationship of heat shock and pathogen responses in plants. Mol Gen Genet 1988 211:290-295.

Pilpel Y, Sudarsanam P, Church GM. Identifying regulatory networks by combinatorial analysis of promoter elements. Nat Genet. 2001 Oct;29(2):153-9.

Prestridge DS. SIGNAL SCAN: a computer program that scans DNA sequences for eukaryotic transcriptional elements. Comput Appl Biosci. 1991 Apr;7(2):203-6.

Pérez-Rodríguez P, Riaño-Pachón DM, Corrêa LG, Rensing SA, Kersten B, Mueller-Roeber B. PlnTFDB: updated content and new features of the plant transcription factor database. Nucleic Acids Res. 2010 Jan;38(Database issue):D822-7.

Quandt K, Frech K, Karas H, Wingender E, Werner T. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res. 1995 Dec 11;23(23):4878-84.

Riaño-Pachón DM, Ruzicic S, Dreyer I, Mueller-Roeber B. PlnTFDB: an integrative plant transcription factor database. BMC Bioinformatics. 2007 Feb 7;8:42.

Rombauts S, Florquin K, Lescot M, Marchal K, Rouzé P, van de Peer Y. Computational approaches to identify promoters and cis-regulatory elements in plant genomes. Plant Physiol. 2003 Jul;132(3):1162-76.

Rosa M, Prado C, Podazza G, Interdonato R, González JA, Hilal M, Prado FE. Soluble sugars--metabolism, sensing and abiotic stress: a complex network in the life of plants. Plant Signal Behav. 2009 May;4(5):388-93.

Sand O, van Helden J. Discovery of motifs in promoters of coregulated genes. Methods Mol Biol. 2007;395:329-48.

Sasaki T. The rice genome structure as a trail from the past to beyond. Genome Dyn. 2008;4:131-42.

Schmid CD, Perier R, Praz V, Bucher P. EPD in its twentieth year: towards complete promoter coverage of selected model organisms. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D82-5.

ScienceDaily: Biotechnology and Biological Sciences Research Council (2010, August 27). Wheat's genetic code cracked: Draft sequence coverage of genome to aid global food shortage. *ScienceDaily*.

Sebestyén E, Nagy T, Suhai S, Barta E. DoOPSearch: a web-based tool for finding and analysing common conserved motifs in the promoter regions of different chordate and plant genes. BMC Bioinformatics. 2009 Jun 16;10 Suppl 6:S6.

Shahmuradov IA, Gammerman AJ, Hancock JM, Bramley PM, Solovyev VV. PlantProm: a database of plant promoter sequences. Nucleic Acids Res. 2003 Jan 1;31(1):114-7.

Shen C, Wang S, Bai Y, Wu Y, Zhang S, Chen M, Guilfoyle TJ, Wu P, Qi Y. Functional analysis of the structural domain of ARF proteins in rice (Oryza sativa L.). J Exp Bot. 2010 Sep;61(14):3971-81.

Shinozaki K, Yamaguchi-Shinozaki K, Seki M. Regulatory network of gene expression in the drought and cold stress responses. Curr Opin Plant Biol. 2003 Oct;6(5):410-7.

Silva-Ortega CO, Ochoa-Alfaro AE, Reyes-Agüero JA, Aguado-Santacruz GA, Jiménez-Bremont JF. Salt stress increases the expression of p5cs gene and induces proline accumulation in cactus pear. Plant Physiol Biochem. 2008 Jan;46(1):82-92.

Simpson SD, Nakashima K, Narusaka Y, Seki M, Shinozaki K, Yamaguchi-Shinozaki K. Two different novel cis-acting elements of erd1, a clpA homologous Arabidopsis gene function in induction by dehydration stress and dark-induced senescence. Plant J. 2003 Jan;33(2):259-70.

Sinha S, Tompa M. YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. Nucleic Acids Res. 2003 Jul 1;31(13):3586-8.

Sinha S, Tompa M. Discovery of novel transcription factor binding sites by statistical overrepresentation. Nucleic Acids Res. 2002 Dec 15;30(24):5549-60.

Solovyev VV, Shahmuradov IA, Salamov AA. Identification of promoter regions and regulatory sites. Methods Mol Biol. 2010;674:57-83.

Steffens NO, Galuschka C, Schindler M, Bülow L, Hehl R. AthaMap: an online resource for in silico transcription factor binding sites in the Arabidopsis thaliana genome. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D368-72.

Sudarsanam P, Pilpel Y, Church GM. Genome-wide co-occurrence of promoter elements reveals a cis-regulatory cassette of rRNA transcription motifs in Saccharomyces cerevisiae. Genome Res. 2002 Nov;12(11):1723-31.

Szafranski K, Lehmann R, Parra G, Guigo R, Glöckner G. Gene organization features in A/T-rich organisms. J Mol Evol. 2005 Jan;60(1):90-8.

TAIR FTP website:
[ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/other_datasets/CURRENT/]

Tatusov RL, Altschul SF, Koonin EV. Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. Proc Natl Acad Sci U S A. 1994 Dec 6;91(25):12091-5.

Than C, Ruths D, Nakhleh L. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics. 2008 Jul 28;9:322.

Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouzé P, Moreau Y. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics. 2001 Dec;17(12):1113-22.

Thomas-Chollier M, Sand O, Turatsinze JV, Janky R, Defrance M, Vervisch E, Brohée S, van Helden J. RSAT: regulatory sequence analysis tools. Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W119-27.

Thomashow MF. PLANT COLD ACCLIMATION: Freezing Tolerance Genes and Regulatory Mechanisms. Annu Rev Plant Physiol Plant Mol Biol. 1999 Jun;50:571-599.

TIGR/JCVI website:
[ftp://ftp.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_5.0/all.chrs/]

TIGR Rice Genome Annotation Resource: [http://rice.plantbiology.msu.edu/]

Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Régnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z. Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol. 2005 Jan;23(1):137-44.

Tran LS, Mochida K. Identification and prediction of abiotic stress responsive transcription factors involved in abiotic stress signaling in soybean. Plant Signal Behav. 2010 Mar;5(3):255-7.

Turatsinze JV, Thomas-Chollier M, Defrance M, van Helden J. Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. Nat Protoc. 2008;3(10):1578-88.

Turóczy Z, Kis P, Török K, Cserháti M, Lendvai Á, Dudits D, Horváth GV. Improvement of the oxidative and heat stress tolerance in transgenic tobacco by the overexpression of an ABA induced to reductase from rice. Plant Mol Biol. 2011.

Tuteja N, Mahajan S. Calcium signaling network in plants: an overview. Plant Signal Behav. 2007 Mar;2(2):79-85.

van Helden J, Rios AF, Collado-Vides J. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. Nucleic Acids Res. 2000 Apr 15;28(8):1808-18.

van Helden J. Regulatory sequence analysis tools. Nucleic Acids Res. 2003 Jul 1;31(13):3593-6.

van Helden J. Metrics for comparing regulatory sequences on the basis of pattern counts. Bioinformatics. 2004 Feb 12;20(3):399-406.

Vardhanabhuti S, Wang J, Hannenhalli S. Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation. Nucleic Acids Res. 2007;35(10):3203-13.

Walhout AJ. Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping. Genome Res. 2006 Dec;16(12):1445-54.

Walia H, Wilson C, Ismail AM, Close TJ, Cui X. Comparing genomic expression patterns across plant species reveals highly diverged transcriptional dynamics in response to salt stress. BMC Genomics. 2009 Aug 25;10:398.

Walther D, Brunnemann R, Selbig J. The regulatory code for transcriptional response diversity and its relation to genome structural properties in A. thaliana. PLoS Genet. 2007 Feb 9;3(2):e11.

Wang H, Datla R, Georges F, Loewen M, Cutler AJ. Promoters from kin1 and cor6.6, two homologous Arabidopsis thaliana genes: transcriptional regulation and gene expression induced by low temperature, ABA, osmoticum and dehydration. Plant Mol Biol. 1995 Jul;28(4):605-17.

Wang L, Xie W, Chen Y, Tang W, Yang J, Ye R, Liu L, Lin Y, Xu C, Xiao J, Zhang Q. A dynamic gene expression atlas covering the entire life cycle of rice. Plant J. 2010 Mar;61(5):752-66.

Wang ZY, Kenigsbuch D, Sun L, Harel E, Ong MS, Tobin EM. A Myb-related transcription factor is involved in the phytochrome regulation of an Arabidopsis Lhcb gene. Plant Cell. 1997 Apr;9(4):491-507.

Wasserman WW, Sandelin A. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet. 2004 Apr;5(4):276-87.

Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA. The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol. 2003 Sep;20(9):1377-419.

Yaish MW, El-Kereamy A, Zhu T, Beatty PH, Good AG, Bi YM, Rothstein SJ. The APETALA-2-like transcription factor OsAP2-39 controls key interactions between abscisic acid and gibberellin in rice. PLoS Genet. 2010 Sep 9;6(9). pii: e1001098.

Yamaguchi-Shinozaki K, Shinozaki K. Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses. Annu Rev Plant Biol. 2006;57:781-803.

Yamaguchi-Shinozaki K, Shinozaki K. Organization of cis-acting regulatory elements in osmotic- and cold-stress-responsive promoters. Trends Plant Sci. 2005 Feb;10(2):88-94.

Yamamoto YY, Obokata J. ppdb: a plant promoter database. Nucleic Acids Res. 2008 Jan;36(Database issue):D977-81.

Yoo SD, Sheen J. MAPK signaling in plant hormone ethylene signal transduction. Plant Signal Behav. 2008 Oct;3(10):848-9.

Yu X, Lin J, Masuda T, Esumi N, Zack DJ, Qian J. Genome-wide prediction and characterization of interactions between transcription factors in Saccharomyces cerevisiae. Nucleic Acids Res. 2006 Feb 6;34(3):917-27.

Zarka DG, Vogel JT, Cook D, Thomashow MF. Cold induction of Arabidopsis CBF genes involves multiple ICE (inducer of CBF expression) promoter elements and a cold-regulatory circuit that is desensitized by low temperature. Plant Physiol. 2003 Oct;133(2):910-8.

Zhang Q, Li J, Xue Y, Han B, Deng XW. Rice 2020: a call for an international coordinated effort in rice functional genomics. Mol Plant. 2008 Sep;1(5):715-9.

Zhang W, Ruan J, Ho TH, You Y, Yu T, Quatrano RS. Cis-regulatory element based targeted gene finding: genome-wide identification of abscisic acid- and abiotic stress-responsive genes in Arabidopsis thaliana. Bioinformatics. 2005 Jul 15;21(14):3074-81.

Zhu JK. Salt and drought stress signal transduction in plants. Annu Rev Plant Biol. 2002;53:247-73.

11. Summary in Hungarian

11.1. Bevezető

Az abiotikus stresszhatásokra adott növényi válaszokban a gének nagy számban játszanak szerepet, mivel ez egy olyan komplex alkalmazkodási folyamatok, amelyek kiterjedten érintik a növény teljes szervezetét. A növények általában kétféleképpen válaszolnak; vagy megpróbálják fenntartani az eredeti fiziológiai állapotukat, vagy alkalmazkodnak a megváltozott körülményekhez. Abiotikus stresszhatások körébe sok olyan környezeti tényező tartozik, mely a növény vízháztartását érinti: a szárazság, a talajban előforduló sók, ozmotikus hatású anyagok, sőt bizonyos mértékig a hideg is ilyen stressz-tényező. A stressz-szignált általában a membránban található receptorok közvetítik (sokszor hormonok hatására, pl. ABA, citokininek, vagy etilén), az érzékelt jelet másodlagos hírvivők továbbítják a citoplazmán belül (pl. ROS, Ca^{2+} vagy IP molekulák) (Zhu, 2002; Mahajan, 2005). A sejten, illetve a sejtmagon belül pedig több transzkripciós faktor kölcsönhatása révén alakul ki a stresszre adott génexpressziós válasz.

A vízhiánnyal kapcsolatos abiotikus stresszválaszok szabályozási útjait általában két csoportra különítjük el, az egyikbe az ABA-függő, a másikba az ABA-független utakat soroljuk (Yamaguchi-Shinozaki, 2005). Eközött a két szabályozási hálózat között azonban jelentős átfedések, illetve kölcsönhatások vannak, ez közösen szabályozott transzkripciós faktorokban, illetve mindkettőhöz tartozó transzkripciós faktor kötőhelyekben is megnyilvánul.

11.2. Célok

Mivel az abiotikus stresszválaszban szereplő gének sok esetben hasonló szabályozás alatt állhatnak, így feltételezhetjük azt, hogy ennek alapját legalább részben a hasonló szabályozó elemeket is tartalmazó a promóter régióik adhatják. Mivel összetett molekuláris genetikai folyamatról van szó, így az is feltételezhető, hogy számos gén, transzkripciós faktor, illetve szabályozó motívum együttműködéséről van szó.

Eddig számos motívumkereső algoritmust fejlesztettek ki DNS szekvenciák analízisére és motívumok előrejelzésére. Ezek leginkább rövid oligonukleotid motívumokat képesek felfedezni. Tompa és kollégái vizsgálatai alapján több jól ismert motívumkereső program vagy algoritmus érzékenysége alacsony (Tompa, 2005), így jelentősen javítani lehet ezeket az algoritmusokat. Mivel olyan algoritmus-összeállítás eddig nem volt, amely együtt szabályozott gének promóter régióiban szabályozó elem párokat, azaz diádokat keressen, azt tűztük ki célul, hogy ilyen algoritmust kifejlesszünk. Az algoritmust a jövőben más, megismert genomszekvenciájú eukarióta élőlényre, így többek között különböző fontos gazdasági növény (árpa, búza, rizs, stb.) promótereinek a vizsgálatára lehet majd alkalmazni.

Az algoritmus egyik fontos eredménye az, hogy bemeneti promóter szettekben, azokra jellemző feltételezett diádokat képes optimalizáltan megtalálni. Az így kapott diádokkal pedig a vizsgált élőlény promóteromját lehet tovább elemezni, és megkeresni a többi hasonló diádot tartalmazó promótert, és ezáltal a promóterek génjeiről előrejelzést lehet vele tenni arra vonatkozóan, hogy állhatnak hasonló transzkripciós szabályozás alatt, így jó eséllyel hasonló folyamatokban is vesz részt, mint a kiindulási promóterek génjei.

11.3. Az algoritmus leírása

Az algoritmus DNS szekvencia diádokat keres, amelyeket az $M_1N_nM_2$ formula határoz meg, ahol M_1 a feji, illetve M_2 a farki motívum. Köztük pedig egy jellemző hosszúságú, szekvencia-specificitás nélküli spacer régió van, ami n bázis hosszú, kevés lötyögéssel megengedve. A fej és farok motívumok ugyanolyan hosszúak, és a spacer régió pedig 0-52 bp hosszú lehet.

Az elemzés több fázisban zajlik. Az első fázis abból áll, hogy a megfelelően együtt szabályozott géneket kiválasztjuk, és a hozzájuk tartozó promóter szekvenciákat egy adatsorba gyűjtjük, majd ezeket különböző szettekbe osztjuk be. A promótereket általában 2 Kbp hosszúságig vizsgáltuk, illetve akkor tekintettük rövidebbnek, ha

közelebb már másik gént találtunk. A promótereket pozitív illetve negatív tanuló szettekbe, valamint pozitív illetve negatív teszt szettekbe kell beosztani.

A következő fázis a tanuló fázis, amikor is az algoritmus összeszámolja az összes lehetséges diád előfordulását a pozitív és negatív tanuló promóter szettekben. Ezek után a diádokat súlyozza, kiszámolja az ún, cdr értékét, és a diádokat eszerint rangsorolja. Egy diád cdr értékét így lehet kiszámolni:

$$cdr = \frac{N_{positive} - N_{negative}}{N_{positive}}$$

Itt a $N_{positive}$ azon promóterek száma a pozitív tanuló promóter szettben, amelyben a diád előfordul, míg $N_{negative}$ azon promóterek száma a negatív tanuló szettben, amelyben a diád szintén előfordul. A cdr értéke -∞-től 1-ig terjedhet (csak azokat az eseteket vettük figyelembe ahol $N_{positive}$ nagyobb volt, mint nulla). Minél magasabb egy diád cdr értéke, annál jelentősebb szerepe lehet abban a folyamatban, amit vizsgálunk. Esetünkben ez a vízhiánnyal kapcsolatos abiotikus stressz volt. A kapott diád adatbázis további szűrés, homopolimerek, repetitív szekvenciák eltávolítása után lehet tovább használni.

A tanuló fázist követi a teszt fázis, amelyben különböző paraméterek szerint vizsgáljuk a diádokat, hogy a pozitív és negatív promóter szettek elkülönítésére legoptimálisabb diádokat kiválasszuk. Ezek a paraméterek a következők: küszöbérték a pozitív tanuló szettben való előfordulás gyakoriságára, a spacer régió „lötyögésének" mértéke ±5bp-ig, illetve küszöbérték a diádok cdr értékére. Minden diád szett esetében visszakeressük a találatokat a pozitív és a negatív promóterek teszt szettjeiben. ROC analízis (Fawcett, 2004) során azt az optimális diád szettet választjuk ki, amelyet majd a promóterom szűréséhez lehet használni. A promóterom szűrése alapján az egyes promótereket a bennük előforduló diádok alapján osztályozzuk és rangsoroljuk.

11.4. Eredmények

Az algoritmusunkat először lúdfű (*Arabidopsis thaliana)* esetében alkalmaztuk, és igazoltuk használhatóságát, majd rizs *(Oryza sativa)* esetében is alkalmaztuk; hogy a

kétszikű modellnövény mellett egy egyszikű növény promóteromjának vizsgálatában is teszteljük. A lúdfű esetében az abiotikus stresszhez kapcsolódó és a kontroll (nem stressz) tanuló promóter szettekbe 125-125 promóter került, míg a megfelelő teszt szettekbe 44-44 darab. Rizs esetében 87-87 promóter került a két tanuló szettbe, míg 42 illetve 56 promóter került bele a pozitív illetve negatív teszt promóter szettbe. A pozitív szettekben a promóterek olyan génekhez tartoztak, amelyek annotációjuk alapján, egy microarray expressziós adatbázis (Genevestigator) adatai alapján, vagy saját kísérleti eredményeink alapján indukálhatóak voltak abiotikus stressz hatására.

Lúdfű esetében az optimalizálással 81 feltételezhető diádot sikerült találnunk. Ebben az esetben a minimum 14-szeri előfordulás a pozitív tanuló promóter szettben, minimum 0.9-es cdr értékkel, ±2bp lötyögéssel adta a pozitív és negatív szettek szétválasztását. Rizs esetében 38 diáddal kaptuk a legjobb eredményt, minimum 9-szer előfordulva a pozitív tanuló promóter szettben, minimum 0.89-es cdr értékkel, és lötyögés megengedése nélkül. A 81 lúdfű diád közül 38-at 11 csoportba klasztertük az alapján, hogy mennyire volt hasonló két adott diád egymáshoz képest. A hasonlóság egy Hamming-távolság volt, ahol a maximális hasonlóság 10 volt (mivel pentamer párokat vizsgáltunk). A Hamming távolság küszöbértéke 3 volt.

A promóterom szűrés alapján a lúdfű esetén a legjobb cdr értéket adó 3100 promótert vizsgáltuk, rizs esetében a legjobb 4600-t. Úgy húztuk meg az ezekhez a számokhoz tartozó határértékeket, hogy ezek között csak minimális arányban forduljanak elő fals pozitívok, azaz a program tanításához használt kontroll (nem stressz) promóterek. Lúdfű esetében a promóterom szűréssel talált génekről 78.6% esetében igazolta vissza rendelkezésre álló expressziós adat, hogy stressz indukálható. Rizs esetében ez 98.7% volt. A predikcióval 1273 (49+1224) hipotetikus vagy újonnan stressz indukálhatónak mondható gént jelzett előre az algoritmus. Rizs esetében ez 2682 gén volt (1437 új gén, illetve 1245 hipotetikus gén).

Az lúdfűben megtalált diádok, illetve a belőlük képzett klaszterek előfordulását vizsgáltuk ismert géncsaládokban. A 11-ből 7 cluster, valamint 5 egyedi diád kulcsfontosságú szerepet játszanak 5 *cor*, illetve 4 *erd* gén hálózat-szerű

szabályozásában. Ezen túl 1224 feltételezhető SZEP (szabályozó elem pár) (amelyeknek a módosított cdr értékük 0.5 fölött volt) jelezhető előre, hogy valamilyen módon szerepet játszhat az abiotikus stressz válaszban.

Lúdfűben promóterome keresést végeztünk azért, hogy megnézzük, milyen SZEP-ek fordulnak elő benne. Kiszámoltuk a Jacquard-együtthatót minden promóter és minden *cor* gén promóter között. Ez alapján a SZEP tartalom különbségét számoltuk ki minden egyes promóter párra. A legjobb 25 promótert választottuk ki, amelyeknek a SZEP tartalom különbsége 0.5 alatt volt. Ezek közül 1 hipotetikus és 5 új ismeretlen funkciójú gént jelöltünk meg, amelyről az elemzés alapján feltételezhetjük, hogy hasonló funkcióval rendelkezhet, mint a *cor* gének.

Harminc rizs aldo-keto reduktáz gén vizsgálata során 28 jellemző, feltételezhető diád elemet fedeztünk fel, amelyek legalább 7 bemeneti promóterben előfordulnak, és amelynek legalább 0.9 volt a cdr értéke. A 30 AKR gén közül három expresszióját vizsgálták meg kísérletesen. Ezek közül az AKR1 (Os01g0847600) több jellegzetes diádot tartalmazott, mint az AKR2 és AKR3. Ez egybecseng a génexpressziós kísérleti eredményekkel, melyek szerint az AKR1 erősen, a másik két gén kevésbé indukálhatóak ozmotikus stressz által. Ezzel egy, a kiindulási génszettől független esetben is igazoltuk az algoritmusunk használhatóságát.

Hat rizs géncsaládhoz (glükanázok, kitinázok, PR1, PR4, PT5, és PR9) tartozó 91 promóterrel is végeztünk elemzéseket. Ezen promótereket tartalmazó gének olyan búza génekkel homológok, amelyek szerepet játszanak a biotikus stressz válaszban. Bennük sok olyan motívumot megtaláltunk, amelyeket már ismerünk, és előfordulnak a növényi cisz-regulátor elemeket tartalmazó PlantCARE adatbázisban is (pl. a W1-box, az EIRE, és a WUN-motif). Így azt tételezzük fel, hogy a szekvencia-hasonlóságok révén ezeket, vagy ezekhez hasonló motívumokat meg lehet találni az ortológ búza gének promótereiben is. Ugyanezeket a géneket később megvizsgáltuk búzában is, és azt találtuk, hogy a prediktált diádok fele kis módosulással ugyan, de egyezik a megfelelő rizs diáddal.

Az algoritmusunkat lefuttattuk erre a 91 rizs biotikus stresszhez kapcsolódó promóterre. Közülük 13 szerepelt a pozitív tanuló szettben (mivel ezek voltak a feltételezhető ortológok), míg az összes többi a negatív tanuló promóter szettet alkották. Mivel a vizsgált promóterek száma csekély volt, tetrád diádokat kerestünk. Összesen 263 diádot talált az algoritmus, amelyeknek a cdr értéke legalább 0.9 volt. 28 olyan diád volt közöttük, amely vagy a 6 vizsgált géncsalád közül legalább 4-ben előfordult, és így statisztikailag jelentős volt, vagy megtalálhatóak voltak a PLACE adatbázisban.

Az algoritmust két jól ismert motívumkereső algoritmussal hasonlítottuk össze, a YMF-fel és a dyad-analysis-szel. Ehhez mindkét programmal vizsgáltuk a lúdfű promóteromot, kiindulásként használva ugyanazt a 125 stressz tanuló promótert, amit a saját fejlesztésű módszerrel is használtunk. Az YMF programmal 283 promótert találtunk meg, amely jelentős számú szabályozó elemet tartalmazott. Ezek közül mindössze 3 tartozott az eredeti 125-ös tanuló promóter szetthez, és csak 3.1%-uk volt stressz-indukálható a Genevestigator adatai alapján. A dyad-analysis program 149 promótert talált meg, amely jelentős számú feltételezhető szabályozó szekvencia elemet tartalmazott. Ezek közül azonban csak 1 tartozott az eredeti 125-höz. Ezeknek a 3.6%-a volt indukálható abiotikus stressz által a Genevestigator adatai alapján. Ezek az eredmények azt mutatják, hogy a mi algoritmusunk ehhez a két programhoz képest sokkal hatékonyabban tud új, feltételezhetően stresszben szerepet játszó szabályozó elemeket előre jelezni, valamint az eredetileg vizsgált, együtt szabályozódó génekhez olyan újabb tagokat találni, amelyek sokat tartalmaznak az előre jelzett motívumokból.

Az algoritmust 64 bites IRIX64 programozási környezetben valósítottuk meg awk (GNU Awk 3.1.5.), C shell, és C (GCC 3.4.6.) szkriptek és programok kombinációival. Az algoritmusnak saját weboldala is van rövid leírással, és egy letölthető, PC-n futtatható önálló programmal: http://bhd.szbk.u-szeged.hu/dyadscan/. Bemeneti paraméterként meg kell adni a pozitív és negatív promóter szetteket, a keresett motívumok hosszát, a spacer régió maximális hosszát, a minimális előfordulást a pozitív tanuló promóter szettben, illetve a diádok minimális cdr értékét. A program eredménye egy olyan feltételezhető diád lista, ami megfelel a bemeneti kritériumoknak. A kimenetben láthatók a diádok

szekvenciái, a pozitív és negatív tanuló szettekben való előfordulásuk, illetve a cdr értékeik.

11. 5. Publikációk

11.5.1. A disszertáció alapját képező közlemények:

Turóczy, Z., Kis, P., Török, K., **Cserháti, M.**, Lendvai, Á., Dudits, D., and Horváth, G: Overproduction of a rice aldo-keto reductase increases oxidative and heat stress tolerance by malondialdehyde and methylglyoxal detoxification, *Plant Molecular Biology*, 2011. (kiadás alatt)

Cserháti M., Turóczy, Z., Zombori, Z., Cserző, M., Dudits, D., Pongor, S., Györgyey, J.: Prediction of new abiotic stress genes in Arabidopsis thaliana and Oryza sativa according to enumeration-based statistical analysis, *Molecular Genetics and Genomics*, 2011 (kiadás alatt).

Cserháti, M., Pongor, S. and Györgyey, J: Statistical methods for finding biologically relevant motifs in promóter regions and a few of its implementations, In: 5th International Conference of PhD Students, University of Miskolc, Hungary, 14-20 August 2005, (Eds L. Lehoczky and L. Kalmár) Published by University of Miskolc, Innovation and Technology Transfer Centre, pp. 41-46, 2005

Cserháti, M., Pongor, S., Dudits, D., and Györgyey, J: (2006). „Enumerációs módszereken alapuló algoritmusok használata promóter motívumok keresésére." Tavaszi Szél 2006 conference. Kaposvár. ISBN 963 229 773 3

Cserháti M.: Usage of enumeration method based algorithms for finding promóter motifs in plant genomes. Acta Biol Szeged 2006, 50(3-4):145.

Veronika Pós, Klára Manninger, Krisztián Halász, Éva Hunyadi-Gulyás, Emília Szájli, **Mátyás Cserháti**, Huijun Duan, Katalin Medzihradszky, János Györgyey, Noémi Lukács: Proteomic changes of the wheat apoplast associated with resistance against leaf rust. 15th International Congress of the Hungarian Society for Microbiology: July 18-20, 2007, Eötvös Loránd University (Budapest, Hungary)

Pós Veronika, **Cserháti Mátyás**, Hunyadi-Gulyás Éva, Manninger Sándorné, Györgyey János, Medzihradszky Katalin, Lukács Noémi: KÖZÖS CISZ-REGULÁLÓ elemek LEVÉLROZSDA FERTŐZÉSSEL ASSZOCIÁLT BÚZA APOPLASZTFEHÉRJÉK GÉNEXPRESSZIÓJÁBAN. A Magyar Biokémiai Egyesület 2007. évi Vándorgyûlése 2007. augusztus 26-29. Debreceni Egyetem (Debrecen, Magyarország)

11.5.2. További közlemények:

Cserháti, M. and Györgyey J. 2006. „Génkutatás i*n silico*", könyvfejezet: „Korszakváltás a molekuláris biológiában" c. könyvben. Szerkesztő: Dudits Dénes.

Dudits, D., **Cserháti, M.**, Miskolczi, P., Horváth, G. The growing family of plant cyclin-dependant kinases with multiple functions in cellular and developmental regulation. 2006. Cell cycle control and plant development. Editor Dirk Inzé. Blackwell Publishing, Oxford.

Cserháti, M., Turóczy, Z., Dudits, D., Horváth, G., and Györgyey, J: Bioinformatic analysis of heptamer palindromes in rice stress promóters. 3rd EPSO vonference, Visegrád, poster.

Turóczy, Z., Kis, P., **Cserháti, M.** Dare to bet? –from the *in silico* predictions to the demonstration of stress induced gene expression. 7th Biologist Days, Cluj Napoca, Romania

Turóczy, Z., Kis, P., **Cserháti, M.**, Dudits, D., Horváth, G. Response of rice AKR genes to abiotic stresses: expression profiling and enzyme activity characterization. 3rd EPSO conference, Visegrád, poster.

Dénes, D., **Cserháti, M.**, Miskolczi, P., Fehér, A., Ayaydin, F. and Horváth, G. V.: Use of Alfalfa In Vitro Cultures in Studies on Regulation of Cyclin-Dependent Kinase (CDK) Functions. 2006. Proceedings of the 11th IAPTC&B Congress, Beijing. Editors: Z. Xu, J. Li, I.K. Vasil, Y. Xue and W. Yang.

Cserháti, M., Turóczy, Z., Sečenji, M., Pongor, S., Cserző, M., Dudits, D., Horváth V., G., Györgyey, J. Növényi promóterek analízise abiotikus stressz folyamatok megértésében. 2006. Straub napok előadás, November 15-17.

András Cseri, András Palágyi, **Mátyás Cserháti**, János Pauk, Dénes Dudits, Ottó Törjék: EcoTILLING analysis of drought related candidate genes in barley. Plant Abiotic Stress - from signaling to development, 2nd meeting of INPAS(Interrnational Network of Plant Abiotic Stress), 14-17 May 2009, Tartu, Estonia

12. Summary in English

12.1. Introduction

A large number of genes play a role in abiotic stress response, since this affects a large part of the plant's physiology. Plants respond to stress in two basic ways: they either try to return to their former physiological status, or try to adapt to their changed environments. Abiotic stress is defined as certain environmental conditions, which reduce the plant's water potential, such as cold, drought, salt, or osmotic stress. Stress signals are transmitted through the plasma membrane (many times due to hormones such as ABA, cytokines, or ethylene). The signal is transmitted within the cell by secondary messengers (e.g. ROS, Ca^{2+}, or IP molecules). The complex interplay between many different kinds of transcription factors within the nucleus is responsible for changes in gene expression levels in response to a given form of stress.

Abiotic stress response in plants usually follows two basic pathways, one dependent from the plant hormone ABA, and one which is independent from it. There are significant overlaps between these two pathways, as well interactions between common transcription factors and transcription factor binding sites.

12.2. Objectives

Since genes which take part in abiotic stress response undergo similar regulation, we may assume that common regulatory elements can be found in their promoter regions. Since we are dealing with a complex molecular genetic phenomenon, we may also assume that a number of genes, transcription factors and transcription factor binding sites have an integrated effect on each other.

Until now a number of motif discovery programs have been developed for the analysis of DNA sequences and the prediction of DNA motifs. These are usually capable of finding short oligonucleotide motifs. According to studies done by Tompa et al., the sensitivity of a number of well-known motif discovery programs was defined to be around 0.22, therefore there is room for improving these algortihms. Since until now no

such algorithm existed for the discovery of regulatory elements in co-regulated promoters, we decided to develop this kind of algorithm. The algorithm can be used in the promoter analysis of a number of agricultural crops (e.g. barley, rice, wheat) among other species.

One of the main results of the algorithm is that it is capable of predicting a number of putative optimal dyads in input promoter sets. The studied organism's promoterome can afterwards be analyzed with these optimal dyads in order to find other promoters which contain a large number of these dyads, and therefore undergo similar transcriptional regulation. The genes of these promoters can be predicted to take part in similar processes as the original genes that the analysis started out from.

12.3. Description of the algorithm

The algorithm searches for dyad sequences, which can be described with the formula $M_1N_nM_2$ where M_1 denotes the head motif, and M_2 the tail motif. In between them is a well-defined spacer region which is n bp long, with slight wobbling allowed. The head and tail motif are the same length, while the spacer region can be 0-52 bp long.

The algorithm is made up of a number of phases. The first phase deals with selecting the proper co-regulated genes as well as determining their promoter sequences. Afterwards we split up the promoters into different sets. The promoters are usually 2 Kbp long shorter if they overlap with upstream genes. The promoters are to be divided into a positive and negative learning set and a positive and negative test set.

The next phase is the learning phase, where the algorithm counts the total number of occurences of all possible dyads in the positive and negative learning sets. Afterwards the algorithm calculates the dyads' weight, or cdr value (cumulative difference ratio), described below, and then ranks them for further analysis and use. The cdr score is calculated as follows:

$$cdr = \frac{N_{positive} - N_{negative}}{N_{positive}}$$

Here $N_{positive}$ is the number of promoters from the positive learning set that a given dyad occurs in, and $N_{negative}$ is the number of promoters in which the dyad also occurs. The cdr score takes up a value between $-\infty$ and 1 (only those cases can be taken into account where $N_{positive}$ is greater than 1). The greater the cdr score of a dyad, the more relevant role it plays in the mechanism under study (in our case abiotic stress).

After the learning phase comes the test phase where the dyads are analyzed according to a number of different parameters in order to select the optimal set. These parameters are: occurence in the positive learning set, the possible wobbling of the spacer region (up to ±5bp), and the minimal cdr value for all selected dyads. All dyad sets were found back in the positive and negative test promoter sets. ROC analysis was used to determine the optimal dyad set which was used in a promoterome search. Promoters found during the promoterome search were scored and ranked according to their optimal dyad content.

12.4. Results

The algorithm was first developed and verified in the case of Arabidopsis and then used in rice; in the promoterome study of one dicot and one monocot. In the case of Arabidopsis 125 promoters were put into the positive and negative learning sets, and 44 into the positive and negative test sets. In rice 87 promoters were put into both learning sets, while 42 were put into the positive test set and 56 into the negative test set. These genes were selected based on their invovlement in abiotic stress, which was determined based on either their annotation or data from the Genvestigator database.

In Arabidopsis we found 81 putative optimal dyads with a minimum occurence of 14 in the positive learning set, a minimum cdr value of 0.9, and a ±2bp wobbling. In rice 38 optimal dyads were found with a minimum occurence of 9 in the positive learning set, a minimum cdr value of 0.89, with no wobbling. The 81 optimal Arabidopsis dyads were clustered into 11 groups, based on how similar two given dyads were to each other. A Hamming distance was calculated for each dyad pair, with a maximum similarity of 10 (since we studied pentamer pairs). The maximum Hamming distance was 3.

According to the individual promoterome searches, we studied the top 3100 Arabidopsis promoters and the top 4600 rice promoters. The reason these numbers of top promoters were selected for both organisms was because here the ratio of non-stress promoters to all promoters was the lowest. According to the promoterome search in Arabidopsis, 78.6% of the found promoters were shown to be involved in abiotic stress response. This means that 49 hypothetical genes and 1224 genes (1273 in total) without any chip data were newly predicted to be involved in abiotic stress. In rice, 98.7% of the genes were shown to be involved in abiotic stress, meaning that 1245 hypothetical genes and 1437 genes without an Affymetrix id were also predicted to be involved in abiotic stress (2682 in total).

38 of the 81 dyads found in Arabidopsis were clustered into 11 groups. 7 clusters and 5 individual dyads were found to be play an important role in the network-type regulation of 5 cor and 4 erd genes. 1224 tentative REPs (Regulatory Element Pairs) (with a modified cdr score above 0.5) play a role in abiotic stress response.

In Arabidopsis we performed a promoterome search in order to find promoters with significant REP content. We calculated the Jacquard coefficient between each gene and each *cor* gene. Based on this we calculated the difference in REP content between each gene and each of the *cor* genes. We selected those 25 promoters whose difference in REP content was below 0.5. In this way we found 1 hypothetic gene and 5 genes of unknown function which could then be newly annotated to have a function similar to *cor* genes.

In the case of the 30 rice aldo-keto reductase (AKR) genes we found 28 new putative dyads which occured in at least 7 of the input promoters, with a minimum cdr value of 0.9. We studied three of the 30 AKR genes in detail, and found that one of them (AKR1, or Os01g0847600) contained more dyads than the other ones (AKR2, and AKR3), which were shown experimentally to be induced by osmotic stress to a lesser degree. In this way also we were capable of independently verifying our algorithm.

We studied tetrad dyads in the promoter region of 91 genes belonging to six rice gene familes (glucanases, chtinases, PR1, PR4, PR5, and PR9 genes). These genes were homologous to certain wheat genes which play a role in biotic stress response. We found many motifs in these promoter regions which have a match in the PlantCARE database (e.g. the W1-box, the EIRE, and a WUN-motif). Because of homology we assumed that we would be able to find similar regulatory elements in the promoter regions of the rice genes homologous to the wheat genes. We studied the promoters of the wheat genes and found that half of the predicted dyads match those in rice with slight modifications.

We ran the algorithm on these 91 rice homolog biotic stress promoters. 13 of these were used as a positive learning set since they were the best scoring homologs, while the rest were put into the negative learning set. Overall 263 dyads were found which had a minimum cdr score of 0.9. 28 dyads were found which occured in promoters of at least 4 of the 6 gene families, therefore their occurence was taken to be statistically significant. Motifs matching these dyads were also found in the PLACE database.

We compared our algorithm with two well-known motif finding algorithms, YMF and dyad-analysis. We ran these two programs on the 125 stress learning promoters from Arabidopsis. With YMF we found 283 promoters which contained a substantial amount of putative regulatory elements. Out of these, only 3 belonged to the original 125 positive learning promoters. 3.1% were shown to be stress-inducible based on data from the Genevestigator database. The dyad-analysis program found 149 promoters, which contained a substantial amount of regulatory elements, amongst which only 1 of them belonged to the original set of 125 positive learning promoters. 3.6% of these were shown to be involved in abiotic stress according to data in the Genevestigator database. These results show that our algorithm is much more capable of finding putative regulatory elements which are involved in abiotic stress response, as well as discovering the involvement of further genes in physiological processes in which the original co-regulated genes take part in, and whose promoters contain a large number of such regulatory elements.

The algorithm was run in a 64 bit IRIX64 programming environment using a combination of awk (GNU Awk 3.1.5), C shell, and C (GCC 3.4.6) scripts. The algorithm can be dowloaded from its own website (which a short description) in the form of a desktop application: http://bhd.szbk.u-szeged.hu/dyadscan/. Input parameters include a positive and negative learning promoter set, length of motifs, maximum length of the spacer region, minimum occurence of dyads in the positive learning set, and minimum cdr value. The program's output is a list of dyads which meet the input criteria. In the output the dyads' sequence, occurence in the positive and negative learning sets, and the cdr score is given.

13. Supplementary data

TRANSFAC/PLACE dyad	max. pos.	pos. occ.	neg. occ.	*cdr* score	TRANSFAC/PLACE dyad	max. pos.	pos. occ.	neg. occ.	*cdr* score
AAACCA CATATG	2	2	0	1	CATCTG CCGAC	10	1	0	1
CAAATG CATATG	13	1	0	1	CTGTTG TAACTG	10	1	0	1
CCGAC TACGTGGC	31	1	0	1	CAAATG CAATTG	49	1	0	1
CACTTG CCGTTG	27	1	0	1	CAACGG CGGTTG	9	1	0	1
CACTTG CTGTTA	25	1	0	1	CATCTG TAACTG	27	1	0	1
ACCGACA CACATG	13	1	0	1	ACCGACA TACCGACAT	8	2	0	1
ACGTG CAGATG	25	1	0	1	CAGGTG CTGTTA	43	1	0	1
CACATG CCGAC	10	1	0	1	ACGT CATATG	11	1	0	1
CACGTG TACCGACAT	40	1	0	1	CAGTTA CTGTTA	7	1	0	1
CAAGTG CTGTTA	49	1	0	1	ACGTG CTGTTA	48	1	0	1
CACGTG TAACGG	15	1	0	1	CAAGTG CAGATG	50	1	0	1
ACGT CAATTG	34	1	0	1	CAAATG CGGTTA	50	1	0	1
CACGTG CACTTG	10	1	0	1	CCGTTA TAACTG	2	1	0	1
ACCGACA TAACCA	32	1	0	1	AAACCA CACCTG	10	1	0	1
CAATTG CCGTTA	14	1	0	1	CAGATG CATTTG	26	1	0	1
CAGCTG CATATG	40	1	0	1	CAAATG CACCTG	46	1	0	1
AAACCA CACGTGGC	39	1	0	1	CATTTG TACGTGGC	49	1	0	1
CACGTG CATGTG	25	1	0	1	ACGTG CACATG	21	1	0	1
AAACCA CCGTTG	39	1	0	1	ACGT CCGTTA	25	1	0	1
AAACCA CTGTTA	49	1	0	1	AAACCA CAGATG	15	1	0	1
CAAATG CTGTTA	47	1	0	1	CAGATG CTAACCA	52	1	0	1
CCGAC TAACGG	31	1	0	1	ACGT CCGAC	19	1	0	1
CATTTG CCGAC	41	1	0	1	CAGATG CTGTTA	42	1	0	1
CACATG TAACGG	38	1	0	1	CGGTTA TAACGG	9	2	0	1
CAAGTG CTGTTG	12	2	0	1	AAACCA CGGTTG	34	1	0	1
CACTTG TAACTG	31	1	0	1	CACGTGGC CATTTG	29	1	0	1
CAAATG CCGTTG	45	1	0	1	ACGTG CTGTTG	25	1	0	1
ACGT CACCTG	9	1	0	1	CAATTG CACGTG	16	1	0	1
CACATG CACTTG	5	1	0	1	CAAATG CAGATG	37	1	0	1
CAAGTG TAACCA	16	1	0	1	CAAATG CGGTTG	40	1	0	1
ACGTG CAAGTG	50	1	0	1	CCGTTG TAACCA	25	1	0	1
AAACCA CACATG	23	1	0	1	CTGTTA TAACGG	34	1	0	1
CATATG CATCTG	52	1	0	1	CAATTG CGGTTG	48	1	0	1
CATTTG TAACGG	46	1	0	1	ACGT CTGTTA	37	1	0	1
CACATG CATGTG	8	3	0	1	CACCTG CTGTTG	19	1	0	1
ACGT CAGATG	18	1	0	1	ACGT CCGTTG	49	1	0	1
CACTTG CATTTG	12	1	0	1	CAGATG CTGTTG	51	1	0	1
CACGTGGC CAGTTA	35	1	0	1	ACGTG TAACGG	27	1	0	1
TAACCA TACGTGGC	8	1	0	1	TAACCA TAACTG	40	1	0	1
CAAATG CACATG	18	2	0	1	CACCTG TAACCA	21	1	0	1
AAACCA CTGTTG	35	1	0	1	ACGTG CACTTG	15	1	0	1
CAGTTA CCGAC	51	1	0	1	AAACCA CAAGTG	16	2	0	1
CAAATG CTGTTG	24	1	0	1	CAGTTA CATTTG	4	1	0	1
CAACTG CACGTG	41	1	0	1	CAGGTG CATGTG	39	1	0	1
CAAGTG CATTTG	37	1	0	1	CACTTG CCGAC	49	1	0	1
CAATTG CATATG	26	1	0	1	CAAATG CAAGTG	34	1	0	1
CAAATG TAACCA	16	1	0	1	ACGTG CATGTG	25	1	0	1
CAGCTG CCGTTG	24	1	0	1	ACGT CACATG	33	1	0	1
CAGCTG CTGTTA	3	1	0	1	CAACTG CCGTTG	50	1	0	1
CAACGG CCGAC	50	2	0	1	ACGT CTGTTG	46	1	0	1
CAAATG TAACTG	45	1	0	1	CAGATG TAACGG	2	1	0	1
CAATTG CAGCTG	8	1	0	1	ACGT TAACCA	21	1	0	1
CATGTG CGGTTA	17	1	0	1	CACCTG CATTTG	34	1	0	1
ACGTG CAGTTA	29	1	0	1	CCGTTA TAACGG	41	1	0	1
AAACCA ACGTG	41	2	0	1	CAGTTG CATTTG	36	1	0	1
ACGT CAAGTG	40	2	0	1	CAACGG CAAGTG	24	1	0	1
AAACCA CATTTG	41	1	0	1	CACTTG CAGTTG	28	2	0	1

CAAATG CATTTG	12	1	0	1	AAACCA CAGTTA	10	1	0	1
CAGCTG CTGTTG	8	1	0	1	CAACTG TAACCA	50	1	0	1
CAGCTG TAACCA	21	1	0	1	CAACTG CAGCTG	12	1	0	1
CCGAC CGGTTA	37	1	0	1	CAGTTG CTGTTA	21	1	0	1
ACGTG CAACTG	52	1	0	1	ACGT CATTTG	49	1	0	1
CATATG CGGTTA	19	1	0	1	CACGTGGC CATGTG	2	1	0	1
CAGTTA CAGTTG	26	1	0	1	CAAGTG CAGTTG	18	1	0	1
CAGCTG TAACTG	25	1	0	1	ACGTG CCGAC	8	2	0	1
ACGTG CAGTTG	38	1	0	1	CAAATG CAGTTA	21	1	0	1
CAATTG CTGTTA	37	1	0	1	AAACCA CAACTG	18	1	0	1
CACATG CACGTG	17	1	0	1	CAACTG CATTTG	27	1	0	1
CACATG CATCTG	30	1	0	1	AAACCA CAGTTG	51	1	0	1
CACTTG TAACGG	32	1	0	1	ACGTG CATCTG	51	1	0	1
CAGCTG CATTTG	29	1	0	1	CATGTG TAACCA	11	1	0	1
CCGAC CGGTTG	12	1	0	1	CAGTTA CATGTG	51	1	0	1
CAACTG CACCTG	17	1	0	1	CACGTG CAGCTG	17	1	0	1
CAAGTG TAACGG	1	1	0	1	CAGTTG CTGTTG	43	1	0	1
CACTTG CATGTG	21	1	0	1	CAAATG CAACTG	33	1	0	1
ACGT CAGTTA	30	1	0	1	TAACCA TAACGG	15	1	0	1
CAGATG CAGTTG	52	1	0	1	ACGTG CAAATG	19	1	0	1
CAAGTG CACTTG	40	1	0	1	CAAATG CAGTTG	31	1	0	1
CAATTG CACATG	38	1	0	1	CACGTG CACGTGGC	8	6	0	1
CAACTG CAGATG	23	1	0	1	CACATG CATATG	22	1	0	1
CAATTG CTGTTG	52	1	0	1	CAGTTG TAACGG	43	1	0	1
AAACCA TAACGG	19	1	0	1	CAGATG CATCTG	27	1	0	1
CAGATG CCGAC	43	1	0	1	CAACTG CACATG	22	1	0	1
AAACCA CACTTG	21	1	0	1	CATGTG CATTTG	21	1	0	1
ACGT CAACTG	52	1	0	1	CATATG TAACCA	25	1	0	1
CAAATG TAACGG	46	1	0	1	CACGTG CCGTTG	12	1	0	1
CAATTG TAACTG	45	1	0	1	ACGT TAACGG	41	1	0	1
ACCGACA ACGTG	16	1	0	1	CAACTG CTGTTG	21	1	0	1
ACGT CAGTTG	35	1	0	1	CACGTG CTGTTA	36	1	0	1
AAACCA CATGTG	35	1	0	1	CACATG CCGTTA	24	1	0	1
CAAGTG CCGAC	25	1	0	1	ACGT CACTTG	14	1	0	1
CAAATG CACTTG	26	1	0	1	ACGT CAAATG	2	1	0	1
AAACCA CAAATG	9	1	0	1	ACGT CATGTG	14	1	0	1
ACGT ACGTG	5	29	0	1	CATATG CATTTG	48	1	0	1
AAACCA ACGT	5	1	0	1	CAATTG CACGTGGC	16	1	0	1
ACGTG CATATG	7	1	0	1	CAGGTG CGGTTA	39	1	0	1
CAAATG CATGTG	26	1	0	1	CAAGTG CCGTTA	41	1	0	1
CAGCTG CCGAC	49	1	0	1	CAATTG CAGTTA	7	1	0	1
CACGTG CAGATG	36	1	0	1	CACGTG CCGAC	50	1	0	1
CAATTG CATTTG	6	1	0	1	CACGTG CTGTTG	8	1	0	1
ACCGACA CACGTG	39	1	0	1	CATATG CTGTTA	37	1	0	1
ACCGACA CATCTG	11	1	0	1	CACGTG TAACCA	3	1	0	1
ACGTG CAACGG	23	1	0	1					
ACGT TACGTGGC	7	1	0	1					
CGGTTA CTGTTA	32	1	0	1					
CAACTG CCGAC	36	1	0	1					
CAACGG CACGTG	34	1	0	1					
CAGCTG CATGTG	12	1	0	1					
CAAGTG CACGTG	28	1	0	1					
CATGTG TACGTGGC	19	1	0	1					
CGGTTA TAACCA	0	1	0	1					
AAACCA CACGTG	16	1	0	1					
AAACCA CATCTG	14	1	0	1					
CGGTTG CTGTTA	5	1	0	1					
CGGTTA TAACTG	2	1	0	1					
AAACCA CCGAC	11	1	0	1					
CAAATG CACGTG	18	1	0	1					

CAAATG CATCTG	15	1	0	1					
CAGTTA CATATG	8	1	0	1					
CAATTG TAACGG	45	1	0	1					
CATGTG CCGAC	39	1	0	1					
CTGTTA TAACTG	3	1	0	1					
CAGATG CCGTTG	34	1	0	1					
CAGGTG TAACCA	48	1	0	1					
CAAATG CCGAC	29	1	0	1					
CAAGTG CAATTG	24	1	0	1					
CACGTG CAGTTA	21	1	0	1					
ACGTG TAACCA	10	1	0	1					
ACGTG CAGCTG	42	1	0	1					
CGGTTG CTGTTG	41	1	0	1					
CATATG CCGAC	49	1	0	1					
CAGCTG CATCTG	15	1	0	1					
AAACCA CAATTG	32	1	0	1					
CATCTG CTGTTG	9	2	0	1					
CATCTG TAACCA	2	2	0	1					
CACCTG CAGCTG	23	1	0	1					
CAGTTG CATATG	9	1	0	1					
ACGTG CATTTG	17	1	0	1					
CACGTG CAGTTG	52	1	0	1					

Supplementary Table 1. List of TRANSFAC and PLACE oligomer dyads used in testing the algorithm in Arabidopsis

Arabidopsis gene id	Functional annotation	Type of stress and expression level change
STRESS LEARNING SET		
At1g01470	hypothetical protein contains similarity to 1-phosphatidylinositol-4-phosphate 5-kinase(AtPIP5K1) GI:3702691 from [Arabidopsis thaliana]	C (9.765053362) Os (5.304472223) S (3.617001063)
At1g09530	putative phytochrome-associated protein 3 similar to GB:AAC99771; supported by cDNA: gi_3929585_gb_AF100166.1_AF100166	Os (9.393757394) S (4.125163639)
At1g12950	unknown function	C(21.15) D(6.78) Os (3.14) Ox (2.33) S (13.19)
At1g14540	T5E21.4 anionic peroxidase, putative similar to anionic peroxidase GI:170202 from [Nicotiana sylvestris]	C (2.985601635) S (76.09338207)
At1g16030	heat shock protein hsp70, putative similar to heat shock protein hsp70 GI:1771478 from [Pisum sativum]	Os (4.608790968) Ox (3.207638265)
At1g16150	hypothetical protein contains similarity to wall-associated kinase 4 GI:3355308 from [Arabidopsis thaliana]	S (4.511250879)
At1g17170	putative glutathione transferase One of three repeated putative glutathione transferases. 72% identical to glutathione transferase [Arabidopsis thaliana] (gi|4006934)	C (4.971387207) Os (2.217088194) Ox (2.499852081) S (3.282292866)
At1g17180	putative glutathione transferase Second of three repeated putative glutathione transferases. 72% identical to glutathione transferase [Arabidopsis thaliana] (gi|4006934). Location of ests 191A10T7 (gb|R90188) and 171N13T7 (gb|R65532)	C (2.314708731) S (5.883524133)
At1g20440	unknown function	C (7.670107015) Os (5.041350247) S (2.901521024)
At1g27730	salt-tolerance zinc finger protein identical to salt-tolerance zinc finger protein GB:CAA64820 GI:1565227 from [Arabidopsis thaliana]; supported by cDNA: gi_14334649_gb_AY034998.1_	C (18.96940365) Os (5.628083337) Ox (2.16846888) S (21.8589509)
At1g32640	protein kinase, putative identical to bHLH protein GB:CAA67885 GI:1465368 from [Arabidopsis thaliana]; supported by cDNA: gi_14335047_gb_AY037203.1_	S (8.86385272)
At1g42990	bZIP transcription factor, putative contains Pfam profile: PF00170: bZIP transcription factor; supported by cDNA: gi_15028322_gb_AY045964.1_	C (2.189474281) Os (2.015105008) S (6.633284104)
At1g48000	myb-related transcription factor (cpm10), putative similar to myb-related transcription factor (cpm10) GB:U33915 GI:1002795 from [Craterostigma plantagineum]; supported by cDNA: gi_15375307_gb_AY008377.2_	C (2.200595818) Os (11.02923287) S (5.073045894)
At1g52560	chloroplast-localized small heat shock protein, putative similar to chloroplast-localized small heat shock protein GI:6601536 from [Funaria hygrometrica]	Os (15.87441264) Ox (5.083658482) S (15.87430066)
At1g53540	T3F20.29 17.6 kDa heat shock protein (AA 1-156) identical to GI:4376161 from [Arabidopsis thaliana] (Nucleic Acids Res. 17 (19), 7995 (1989))	D (2.084099597) Os (11.99039281) Ox (6.408432379) S (7.062776915)
At1g53580	T3F20.11 glyoxalase II, putative similar to GI:1644427 from [Arabidopsis thaliana]; supported by cDNA: gi_15450394_gb_AY052298.1_	Os (5.137537223) S (2.419064343)
At1g56650	anthocyanin2, putative similar to anthocyanin2 (An2) GI:7673088 from [Petunia integrifolia]; supported by cDNA: gi_3941507_gb_AF062908.1_AF062908	Os (6.480290304) S (7.12481888)
At1g58340	unknown protein contains Pfam profile: PF01554 uncharacterized membrane protein family UPF0013; supported by cDNA: gi_6520160_dbj_AB028198.1_AB028198	C (3.400683156) S (4.766201563)
At1g59500	T4M14.15 auxin-regulated protein GH3, putative similar to auxin-regulated protein GH3 GI:18590 from [Glycine max]	D (24.22) Os (7.25) Ox (10.82) S (4.13)
At1g59860	heat shock protein, putative similar to heat shock protein GI:19617 from [Medicago sativa];supported by full-length cDNA: Ceres:32795.	C (4.61) Os (7.83) S (22.62)
At1g63440	ATP dependent copper transporter, putative similar to ATP dependent copper transporter GB:AAD29109 GI:4760370 [Arabidopsis thaliana]	Os (2.967133467)
At1g69260	unknown function	C (4.444230608) Os (19.46529026) S (12.18842019)
At1g69920	putative glutathione transferase similar to glutathione transferase GB:CAA09188 [Alopecurus myosuroides]	C (3.630700341) Os (2.319094005) S (4.598509272)
At1g69930	putative glutathione transferase similar to glutathione transferase GB:CAA09188 [Alopecurus myosuroides]	C (2.584146026) Os (2.930454811) S (14.14306346)
At1g78340	glutathione transferase, putative similar to glutathione transferase GI:2853219 from [Carica papaya];supported by full-length cDNA: Ceres:252874.	S (3.436494051)
At2g02390	putative glutathione S-transferase ; supported by cDNA: gi_15450462_gb_AY052332.1_	Os (2.908658631)
At2g02990	ribonuclease, RNS1 identical to ribonuclease SP:P42813, GI:561998 from	D (3.534923066) Os

	[Arabidopsis thaliana]; supported by full-length cDNA: Ceres:27242.	(8.803118874) S (3.986068111)
At2g04040	unknown function	C (2.347500927) Ox (2.685612254) S (8.602782816)
At2g04050	unknown function	Ox (7.843314827) S (83.8064251)
At2g14960	putative auxin-regulated protein	C (2.927111115) S (17.10175917)
At2g17660	unknown function	C (60.30006075) Os (2.018902523) S (5.989887155)
At2g29460	putative glutathione S-transferase ; supported by cDNA: gi_14423533_gb_AF387004.1_AF387004	C (3.038105056) D (4.662885472) Os (24.98442596) Ox (4.766600252) S (17.33455765)
At2g29480	putative glutathione S-transferase ; supported by cDNA: gi_11096001_gb_AF288184.1_AF288184	Os (2.309642038) S (5.601690175)
At2g29500	putative small heat shock protein ;supported by full-length cDNA: Ceres:25828.	Os (7.036160099) Ox (5.826684811) S (2.81737695)
At2g32020	putative alanine acetyl transferase ;supported by full-length cDNA: Ceres:21201.	C (5.910467367) S (27.93902587)
At2g32120	70kD heat shock protein ;supported by full-length cDNA: Ceres:98979.	Os (3.697065688)
At2g33380	putative calcium-binding EF-hand protein ; supported by cDNA: gi_10862967_dbj_AB039924.1_AB039924	C (3.431709202) D (2.17268216) Os (14.91838404) S (11.72873093)
At2g36270	abscisic acid insensitive 5 (ABI5) contains a bZIP transcription factor basic domain signature (PDOC00036); supported by cDNA: gi_13346150_gb_AF334206.1_AF334206	Os (4.143588214) S (2.477123538)
At2g39800	delta-1-pyrroline 5-carboxylase synthetase (P5C1) identical to SP:P54887:P5C1_ARATH; supported by cDNA: gi_1532270_dbj_D32138.1_ATHATP5CS	C (3.372099701) Os (5.443842235) S (4.011966631)
At2g40140	putative CCCH-type zinc finger protein also an ankyrin-repeat protein	C (8.476051284) Os (2.054896766) S (7.363223162)
At2g40170	ABA-regulated gene (ATEM6) ; supported by cDNA: gi_13430489_gb_AF360157.1_AF360157	Os (16.46740387) S (14.88995584)
At2g42530	cold-regulated protein cor15b precursor ;supported by full-length cDNA: Ceres:19221.	C (7.838331191) Os (2.214823833) S (2.497123095)
At2g42540	cold-regulated protein cor15a precursor ; supported by cDNA: gi_14532457_gb_AY039853.1_	C (30.89632336) D (2.079436904) Os (26.31429328) S (17.8146657)
At2g44840	putative ethylene response element binding protein (EREBP) ;supported by full-length cDNA: Ceres:6397.	C (5.00054184) Os (2.612806473) S (141.6652725)
At2g46670	unknown function	C (8.448695445) S (2.328234063)
At2g46680	homeodomain transcription factor (ATHB-7) identical to SP:P46897; supported by cDNA: gi_15027938_gb_AY045826.1_	Os (15.39668442) S (7.298793909)
At2g46790	unknown function	C (8.448695445) S (2.328234063)
At2g47000	putative ABC transporter related to multi drug resistance proteins and P-glycoproteins	S (4.822138268)
At2g47190	MYB transcription factor (Atmyb2)	C (2.283666684) Os (7.251950201) S (18.27700769)
At3g03620	unknown function	Os (5.461716517)
At3g09640	putative ascorbate peroxidase strong similarity to ascorbate peroxidase GB:CAA56340	Os (14.71764047) S (10.48836199)
At3g11020	DREB2B transcription factor identical to dehydration response element binding transcription factor DREB2B GB:BAA33795 [Arabidopsis thaliana]; supported by cDNA: gi_3738231_dbj_AB007791.1_AB007791	C (6.823298218) S (11.21883046)
At3g11410	protein phosphatase 2C (PP2C) identical to protein phosphatase 2C (PP2C) GB:P49598 [Arabidopsis thaliana]	C (2.573772973) Os (6.421914545) S (6.441221214)
At3g12500	basic chitinase identical to basic chitinase GB:AAA32769 GI:166666 [Arabidopsis thaliana] (Plant Physiol. 93, 907-914 (1990)); supported by cDNA: gi_15451095_gb_AY054628.1_	Os (4.045016612) S (2.019476006)
At3g12580	heat shock protein 70 identical to heat shock protein 70 GB:CAA05547 GI:3962377 [Arabidopsis thaliana]; supported by cDNA: gi_15809831_gb_AY054183.1_	C (2.080006951) Os (5.327898161) Ox (2.045675788) S (3.71212872)
At3g14440	9-cis-epoxycarotenoid dioxygenase, putative similar to 9-cis-epoxycarotenoid dioxygenase GB:AAF26356 [Phaseolus vulgaris]; supported by cDNA: gi_15810432_gb_AY056255.1_	C (15.34324585) Os (25.16675281) S (31.88222812)
At3g15500	putative jasmonic acid regulatory protein similar to jasmonic acid 2	C (3.066392786) Os

	GB:AAF04915 from [Lycopersicon esculentum];supported by full-length cDNA: Ceres:109984.	(7.428420589) Ox (2.017326201) S (22.71475661)
At3g19580	zinc finger protein, putative similar to Cys2/His2-type zinc finger protein 2 GB:BAA85107 from [Arabidopsis thaliana]; supported by cDNA: gi_15028256_gb_AY046043.1_	C (3.30070691) Os (5.246128584) S (8.341394928)
At3g22370	alternative oxidase 1a precursor identical to GB:Q39219 from [Arabidopsis thaliana];supported by full-length cDNA: Ceres:116257.	C (2.099883082) Os (2.783168479) S (7.296535978)
At3g22840	early light-induced protein identical to early light-induced protein GB:AAB88391 from [Arabidopsis thaliana];supported by full-length cDNA: Ceres:14490.	C (11.0876051) Os (8.407112373) S (3.482990342)
At3g23220	ethylene responsive element binding protein, putative similar to EREBP-2 GB:BAA07324 from [Nicotiana tabacum]	C (2.55227834) S (39.36506893)
At3g23230	ethylene responsive element binding protein, putative similar to EREBP-4 GB:BAA07323 from [Nicotiana tabacum]	Ox (2.004563967) S (21.67122544)
At3g23240	ethylene response factor 1 (ERF1) identical to ethylene response factor 1 GB:AAD03544 from [Arabidopsis thaliana];supported by full-length cDNA: Ceres:21068.	S (13.29084921)
At3g24500	ethylene-responsive transcriptional coactivator, putative similar to GB:AAD46402 from [Lycopersicon esculentum] (Plant J. 18 (6), 589-600 (1999));supported by full-length cDNA: Ceres:158734.	Os (4.262994) S (3.097016558)
At3g26830	putative cytochrome P450 similar to cytochrome P450 71B2 GB:O65788 [Arabidopsis thaliana]	Os (4.150127109) Ox (3.025061692) S (8.471440115)
At3g28210	zinc finger protein (PMZ), putative identical to putative zinc finger protein (PMZ) GB:AAD37511 GI:5006473 [Arabidopsis thaliana]	C (8.445383922) Os (3.240804296) Ox (3.26546631) S (8.780954073)
At3g28580	unknown function	C (5.313383048) S (16.56609622)
At3g55610	delta-1-pyrroline-5-carboxylate synthetase	C (3.372099701) Os (5.443842235) S (4.011966631)
At3g60140	beta-glucosidase-like protein several beta-glucosidases - different species; supported by cDNA: gi_10834547_gb_AF159376.1_AF159376	D (2.353502) Os (21.09583488) Ox (2.717621771) S (10.13160617)
At3g61890	homeobox-leucine zipper protein ATHB-12 ;supported by full-length cDNA: Ceres:32615.	C (8.409976152) Os (21.10349653) S (9.214743967)
At3g62100	auxin-induced protein homolog auxin-induced protein IAA20 - Arabidopsis thaliana, PIR:T02188	S (9.049765848)
At4g09600	gibberellin-regulated protein GASA3 precursor ; supported by cDNA: gi_15450402_gb_AY052302.1_	Os (125.2732454) S (41.55691989)
At4g09610	gibberellin-regulated protein GASA2 precursor ; supported by cDNA: gi_887936_gb_U11765.1_ATU11765	Os (9.408109008)
At4g10250	F24G24.50 heat shock protein 22.0 ; supported by cDNA: gi_511795_gb_U11501.1_ATU11501	Os (11.5494547) Ox (2.566527017) S (10.64627084)
At4g11280	ACC synthase (AtACS-6) ; supported by cDNA: gi_16226285_gb_AF428292.1_AF428292	C (10.51663851) Os (2.114881715) S (22.41376675)
At4g12410	T4C9.250 putative protein auxin-induced protein 10A -Glycine max,PID:g255579	Os (56.67191621) S (14.92167373)
At4g17090	putative beta-amylase ;supported by full-length cDNA: Ceres:36882.	C (6.557554743)
At4g17490	ethylene responsive element binding factor-like protein (AtERF6) ; supported by cDNA: gi_3298497_dbj_AB013301.1_AB013301	C (7.864044237) Os (2.050937349) S (24.38425368)
At4g24960	F6I7.6 abscisic acid-induced - like protein abscisic acid-induced protein HVA22, Hordeum vulgare, PIR2:A48892;supported by full-length cDNA: Ceres:28535.	C (14.65573491) Os (3.947865005) S (3.814915017)
At4g25200	Arabidopsis mitochondrion-localized small heat shock protein (AtHSP23.6-mito) ; supported by cDNA: gi_1669865_gb_U72958.1_ATU72958	Os (5.797315058) Ox (5.10495808) S (6.216829282)
At4g25470	M7J2.161 DRE CRT-binding protein DREB1C involved in low-temperature-responsive gene expression00; supported by cDNA: gi_3738227_dbj_AB007789.1_AB007789	C (144.6572867) S (31.33907598)
At4g25480	M7J2.150 transcriptional activator CBF1-like protein strong similarity to transcriptional activator CBF1, Arabidopsis thaliana00	C (78.76388632) Os (5.859978195) S (8.288327857)
At4g25490	M7J2.140 transcriptional activator CBF1 CRT CRE binding factor 1 involved in low-temperature-responsive gene expression00; supported by cDNA: gi_1899057_gb_U77378.1_ATU77378	C (142.3906717) Os (5.47341521) S (60.10608211)
At4g26080	protein phosphatase ABI1 ; supported by cDNA: gi_14334799_gb_AY035073.1_	C (3.389714207) Os (6.443328562) S (5.038956954)
At4g27670	heat shock protein 21	Os (16.26399782) Ox (4.88692217) S (3.432086244)
At4g34000	F28A23.230 abscisic acid responsive elements-binding factor(ABF3) identical to abscisic acid responsive elements-binding factor (ABF3)	Os (3.444812777) S (2.738029061)

	GI:6739280 from [Arabidopsis thaliana]; supported by cDNA: gi_15451049_gb_AY054605.1_	
At4g34710	arginine decarboxylase SPE2 ; supported by cDNA: gi_14517491_gb_AY039581.1_	C (2.178313484) Os (4.5751623) S (3.312401022)
At4g35770	F8D20.280 senescence-associated protein sen1 ;supported by full-length cDNA: Ceres:13699.	D (3.536359778) Os (7.870020002) Ox (2.484522891)
At4g36110	putative auxin-induced protein high similarity to auxin-induced protein 15A, soybean, PIR2:JQ1096; supported by cDNA: gi_13194817_gb_AF349524.1_AF349524	C (8.998247413)
At4g36740	homeodomain protein	Os (4.90329603) S (6.555684839)
At4g37390	unknown function	C (2.829046981) S (4.096263444)
At4g38410	putative cold-regulated protein cold-regulated protein cor47 - Arabidopsis thaliana, PIR2:S19226	Os (3.069472048)
At5g05340	peroxidase	D (3.105428334) Os (3.787160628) Ox (2.27321998) S (4.512906605)
At5g05410	DREB2A (dbj|BAA33794.1) ; supported by cDNA: gi_3738229_dbj_AB007790.1_AB007790	C (6.872428672) Os (6.490834353) S (14.74805059)
At5g06730	peroxidase ;supported by full-length cDNA: Ceres:7360.	D (2.039161215) Os (3.451099896) S (3.755796008)
At5g11210	putative protein GLUR3 ligand-gated channel-like protein precursor, Arabidopsis thaliana, EMBL:AF167355	C (3.293798038) Os (2.203925789) S (5.99648723)
At5g12020	heat shock protein 17.6-II ;supported by full-length cDNA: Ceres:2281.	D (2.256367085) Os (7.832308403) Ox (8.431967654) S (4.98949109)
At5g13750	MSH12.22 transporter-like protein ;supported by full-length cDNA: Ceres:27439.	C (2.126333501) Os (2.273998412) S (2.826738505)
At5g15960	cold and ABA inducible protein kin1 ;supported by full-length cDNA: Ceres:2270.	C (6.359021963) Os (6.266282114) S (5.992884229)
At5g17220	glutathione S-transferase-like protein ; supported by cDNA: gi_11096011_gb_AF288189.1_AF288189	Os (2.03658352) S (3.239367869)
At5g19880	peroxidase peroxidase, Lycopersicon esculentum, PIR:S32768;supported by full-length cDNA: Ceres:100990.	S (18.92664553)
At5g20830	sucrose-UDP glucosyltransferase	C (3.616436321) Os (2.54103548)
At5g24470	putative protein contains similarity to two-component response regulator protein; supported by cDNA: gi_10281005_dbj_AB046955.1_AB046955	C (7.818666346)
At5g27420	RING-H2 zinc finger protein-like RING-H2 zinc finger protein ATL6 - Arabidopsis thaliana, EMBL:AF132016;supported by full-length cDNA: Ceres:106078.	C (2.837859086) Os (3.527802956) Ox (2.333563081) S (6.261424915)
At5g37670	low-molecular-weight heat shock protein - like cytosolic class I small heat-shock protein HSP17.5, Castanea sativa, EMBL:CSA9880	Os (9.695451683) Ox (3.275077635) S (2.560129036)
At5g39580	peroxidase ATP24a	C (5.419110375) Os (2.767833034) S (9.282182641)
At5g40645	unknown function	Os (4.952174626)
At5g40990	GDSL-motif lipase/hydrolase-like protein	D (2.096266462) Os (3.417040776) Ox (2.119596786) S (12.39095089)
At5g45890	senescence-specific cysteine protease SAG12 identical to senescence-specific protein SAG12 GI:1046373 from [Arabidopsis thaliana]	Os (7.83304039)
At5g47220	ethylene responsive element binding factor 2 (ATERF2) (sp|O80338);supported by full-length cDNA: Ceres:3012.	C (2.710101732) S (11.33240771)
At5g47230	ethylene responsive element binding factor 5 (ATERF5) (sp|O80341) ; supported by cDNA: gi_14326511_gb_AF385709.1_AF385709	C (7.10987893) S (3.771949628)
At5g49480	K7J8.1 NaCl-inducible Ca2+-binding protein-like; calmodulin-like ; supported by cDNA: gi_13358217_gb_AF325028.2_AF325028	C (5.785472764)
At5g50720	unknown function	C (6.463215164) Os (3.006067145) S (2.091431403)
At5g51440	mitochondrial heat shock 22 kd protein-like ; supported by full-length cDNA: Ceres: 268536.	Ox (3.000907968) S (6.748329841)
At5g52300	low-temperature-induced 65 kD protein (sp|Q04980)	C (64.4433391) Os (361.9111253) S (156.4280397)
At5g52310	low-temperature-induced protein 78 (sp|Q06738) ; supported by cDNA: gi_348691_gb_L22567.1_ATHCOR78A	C (44.36469089) Os (25.77978761) S (20.10508768)
At5g54490	unknown function	C (5.687922123) S

		(13.18383967)
At5g57050	protein phosphatase 2C ABI2 (PP2C) (sp|O04719)	C (5.153631854) Os (11.40095552) S (6.44304951)
At5g59220	protein phosphatase 2C - like ABA induced protein phosphatase 2C, Fagus sylvatica, EMBL:FSY277743; supported by cDNA: gi_15809791_gb_AY054163.1_	Os (33.38885955) S (15.84450559)
At5g59310	nonspecific lipid-transfer protein precursor - like nonspecific lipid-transfer protein precursor, Brassica napus, EMBL:AF101038;supported by full-length cDNA: Ceres:43057.	Os (454.2004012) S (353.1607566)
At5g59720	heat shock protein 18 ;supported by full-length cDNA: Ceres:97197.	Os (19.6776896) Ox (4.559391973) S (11.31058473)
At5g59820	zinc finger protein Zat12 ;supported by full-length cDNA: Ceres:40576.	C (10.91264505) Os (4.689596503) Ox (3.061731858) S (25.21033062)
At5g61900	MAC9.16 copine - like protein copine I, Homo sapiens, EMBL:HSU83246;supported by full-length cDNA: Ceres:146738.	S (6.035368919)
At5g62480	glutathione S-transferase-like protein ; supported by cDNA: gi_11095991_gb_AF288179.1_AF288179	Os (2.806665756) S (11.48542073)
At5g66400	K1L20.1 dehydrin RAB18-like protein (sp|P30185) ; supported by cDNA: gi_16226664_gb_AF428458.1_AF428458	C (3.077336066) D (2.109991264) Os (90.4725501) S (42.91756474)
STRESS TEST SET		
At1g05260	putative peroxidase Strong similarity to Arabidopsis peroxidase ATPEROX7A (gb|X98321);supported by full-length cDNA: Ceres:114862.	C (4.53) D (6.24)
At1g08920	putative sugar transport protein, ERD6 similar to GB:BAA25989; supported by cDNA: gi_14194108_gb_AF367260.1_AF367260	C (3.28194562) Os (3.28194562) S (3.28194562)
At1g08930	zinc finger protein ATZF1, putative identical to GB:BAA25989; supported by cDNA: gi_3123711_dbj_D89051.1_D89051	C (2.569044169)
At1g09070	unknown protein Similar to Glycine SRC2 (gb|AB000130). ESTs gb|H76869,gb|T21700,gb|ATTS5089 come from this gene; supported by cDNA: gi_15010557_gb_AY045580.1_	C (7.997085652) Os (7.997085652) S (7.997085652)
At1g12610	transcriptional activator CBF1, putative similar to transcriptional activator CBF1 GI:1899058 from [Arabidopsis thaliana	C (11.46873487) S (11.46873487)
At1g29395		C (4.432692268) Os (4.432692268) S (4.432692268)
At1g30360		C (3.472335909)
At1g54100	aldehyde dehydrogenase homolog, putative similar to aldehyde dehydrogenase homolog GI:913941 from [Brassica napus]; supported by cDNA: gi_14190390_gb_AF378873.1_AF378873	Os (0.999701064) S (0.999701064)
At1g62320		C (12.48628192)
At1g78240		C (2.753972644)
At2g03850	putative cold-regulated protein ;supported by full-length cDNA: Ceres:13146.	Os (0.576267247) S (0.576267247)
At2g15970	similar to cold acclimation protein WCOR413 [Triticum aestivum] ;supported by full-length cDNA: Ceres:7835.	C (4.317963459) Os (4.317963459) S (4.317963459)
At2g17840	putative senescence-associated protein 12 ;supported by full-length cDNA: Ceres:40806.	C (11.59409215) Os (11.59409215) S (11.59409215)
At2g23680	F27L4.20 similar to cold acclimation protein WCOR413 (Triticum aestivum)	C (4.93) Os (2.67)
At2g24040		Os (0.755094635)
At2g31380	putative CONSTANS-like B-box zinc finger protein ; supported by cDNA: gi_12698721_gb_AF323666.1_AF323666	C (5.04212603)
At2g38330		S (2.604745)
At2g38340	DREB-like AP2 domain transcription factor DRE binding proteins may be involved in dehydration or low temp response	C (3.350605571) Os (3.350605571) Ox (3.350605571) S (3.350605571)
At2g38905		C (2.090201278) Os (2.090201278) S (2.090201278)
At2g40350	AP2 domain transcription factor	C (5.439880637) S (5.439880637)
At3g01100	unknown protein similar to HYP1 GB:CAA55187 from [Arabidopsis thaliana]	Os (0.804968429)
At3g02480	unknown protein similar to pollen coat protein GB:CAA63531 from [Brassica oleracea]; supported by cDNA: gi_14335127_gb_AY037243.1_	C (9.36983275) Os (9.36983275) S (9.36983275)
At3g05880	low temperature and salt responsive protein LTI6A identical to low temperature and salt responsive protein LTI6A GB:AAC97512 from [Arabidopsis thaliana]	C (2.24) D (9.2)
At3g21620	unknown protein similar to HYP1 GB:CAA55187 [Arabidopsis thaliana]	Os (0.710297369) S (0.710297369)

At3g56080		Os (1.019854844)
At3g59350	protein kinase-like protein Pto kinase interactor 1 - Lycopersicon esculentum, EMBL:U28007; supported by cDNA: gi_15451117_gb_AY054639.1_	C (9.268559567)
At4g02200	T10M13.20 drought-induced-19-like 1 similar to drought-induced-19, GenBank accession number X78584 similar to F2P16.10, GenBank accession number 2191179 identical to T10M13.20	C (3.12) S (3.04)
At4g12400	T4C9.240 stress-induced protein sti1 -like protein stress-induced protein sti1 -Glycine max,PID:g872116	Os (0.577898442) S (0.577898442)
At4g15430		C (2.834498174)
At4g15755		S (0.715137306)
At4g15910	drought-induced protein like	C (2.21574616) Os (2.21574616)
At4g19120		C (2.439577822)
At4g25580	putative protein similarity to low-temperature-induced protein 65, Arabidopsis thaliana, PIR2:S30153~contains EST gb:W43419, W4351200	Os (1.593612118) S (1.593612118)
At4g30650	low temperature and salt responsive protein homolog low temperature and salt responsive protein LTI6A - Arabidopsis thaliana,PID:g4039153	C (3.508499143)
At4g35985	F4B14.250 putative protein physical impedance induced protein, Zea mays, gb:AF001635	C (3.357336601)
At4g36020	glycine-rich protein glycine-rich protein 2, wood tobacco, PIR1:KNNT2S;supported by full-length cDNA: Ceres:122924.	C (2.245404353)
At4g36680	salt-inducible like protein	C (3.07)
At4g37220	cold acclimation protein homolog	Os (1.329595553) S (1.329595553)
At4g39070	putative zinc finger protein salt-tolerance protein - Arabidopsis thaliana, PID:e224078	C (2.396645201)
At5g01300		Os (0.983361943) S (0.983361943)
At5g38710	proline oxidase, mitochondrial precursor -like protein PROLINE OXIDASE, MITOCHONDRIAL PRECURSOR, Arabidopsis thaliana, SWISSNEW:PROD	C (2.746489012) Os (2.746489012)
At5g38760	pollen coat -like protein pollen coat protein, wild cabbage, PIR:T14467;supported by full-length cDNA: Ceres:37918.	Os (1.370692325) S (1.370692325)
At5g51990	AP2 domain transcription factor	C (7.092679619) Os (7.092679619) S (7.092679619)
At5g53820	ABA-inducible protein-like ;supported by full-length cDNA: Ceres:24640.	C (7.092679619) Os (7.092679619) S (7.092679619)

Supplementary Table 2. List of 125 Arabidopsis stress genes used in the stress learning promoter set and 44 stress genes used in the stress test promoter set

Ricearray identifier	Gene locus identifier	Annotation
STRESS LEARNING SET		
11667.m00039	Os01g01380	hypothetical protein
11667.m00419	Os01g04920	glycosyl transferase, group 1 family protein, putative
11667.m00641	Os01g07070	transposon protein, putative, CACTA, En/Spm sub-class
11667.m00691	Os01g07550	hypothetical protein
11667.m00831	Os01g08780	Endonuclease/Exonuclease/phosphatase family, putative
11667.m01560	Os01g15600	PurA ssDNA and RNA-binding protein
11667.m01725	Os01g17150	plastid-specific 50S ribosomal protein 5, chloroplast precursor, putative, expressed
11667.m02668	Os01g27740	DnaJ domain, putative
11667.m03415	Os01g35930	hypothetical protein, (retrotransposon protein, putative, unclassified)
11667.m03850	Os01g40090	Protein phosphatase 2C, putative
11667.m04159	Os01g43000	transposon protein, putative, unclassified
11667.m05039	Os01g51570	Glycosyl hydrolases family 17
11667.m05186	Os01g52850	hypothetical protein
11667.m05467	Os01g55420	expressed protein
11667.m06591	Os01g65600	PHD-finger, putative
11667.m06772	Os01g67280	hypothetical protein
11667.m07062	Os01g70460	hypothetical protein
11668.m00900	Os02g09770	expressed protein
11668.m03030	Os02g32230	retrotransposon protein, putative, unclassified
11668.m03399	Os02g35770	Homeobox domain, putative
11668.m03930	Os02g40780	expressed protein
11668.m04545	Os02g46930	jmjC domain, putative
11668.m04566	Os02g47140	Ribosomal protein L11, RNA binding domain, putative
11668.m04678	Os02g48200	Protein kinase domain, putative
11668.m05056	Os02g51750	Annexin, putative
11668.m05377	Os02g54630	hypothetical protein
11669.m00870	Os03g08910	MatE, putative
11669.m01142	Os03g11860	hypothetical protein, (blastp: putative polyprotein from transposon TNT)
11669.m01219	Os03g12570	C-5 cytosine-specific DNA methylase, putative
11669.m01426	Os03g14430	hypothetical protein
11669.m02154	Os03g21040	r40c1 protein - rice
11669.m02485	Os03g24550	hypothetical protein
11669.m03467	Os03g34230	hypothetical protein
11669.m03601	Os03g36540	putative chelatase subunit
11669.m04133	Os03g41680	hypothetical protein
11669.m04963	Os03g49400	putative manganese transport protein
11669.m05261	Os03g52090	putative Ca2+-transporting ATPase, 3'-partial
11669.m05352	Os03g52930	hypothetical protein
11669.m06327	Os03g62010	putative harpin inducing protein
11670.m00345	Os04g04350	hypothetical protein
11670.m03335	Os04g34600	Similar to anth(ABA/WDS induced protein, expressed)
11670.m03698	Os04g38950	glutamine amidotransferase of anthranilate synthase or para-aminobenzoate synthase, putative
11670.m04026	Os04g41870	oxidoreductase, short chain dehydrogenase/reductase family
11670.m04073	Os04g42320	AT hook motif, putative
11670.m04438	Os04g45650	expressed protein, (putative AG-motif binding protein-4 {Oryza sativa (japonica , 70%)
11670.m04508	Os04g46300	NBS-LRR disease resistance protein homologue
11670.m04746	Os04g48400	GMC oxidoreductase, putative
11670.m05080	Os04g51940	YT521-B-like family, putative
11670.m05480	Os04g55650	Similar to oryzain alpha chain precursor (ec 3.4.22.-). [rice]
11673.m01386	Os07g14260	hypothetical protein, (ankyrin repeat protein-like)
11673.m02862	Os07g30140	Phosphoribosyl transferase domain, putative
11673.m02981	Os07g31290	Protein kinase domain, putative
11673.m03220	Os07g33580	Cytochrome P450
11673.m03538	Os07g36580	hypothetical protein
11673.m03989	Os07g40820	hypothetical protein, (pentatricopeptide, putative)
11674.m00089	Os08g01820	hypothetical protein
11674.m00126	Os08g02190	hypothetical protein
11674.m00907	Os08g09610	hypothetical protein
11674.m01746	Os08g17760	hypothetical protein
11674.m03626	Os08g36490	hypothetical protein

11674.m03813	Os08g38260	hypothetical protein
11676.m02463	Os10g28850	hypothetical protein, (retrotransposon protein, putative, unclassified)
11676.m02543	Os10g29620	Protein kinase domain, putative
11676.m02714	Os10g31250	hypothetical protein
11676.m02890	Os10g32910	hypothetical protein
11676.m03192	Os10g35720	putative PKCq-interacting protein
11682.m01158	Os05g12220	hypothetical protein
11682.m02883	Os05g31070	retrotransposon protein, putative, unclassified
11682.m04202	Os05g44120	expressed protein
11682.m04474	Os05g46480	LEA protein - rice
11682.m04516	Os05g46880	hypothetical protein
11682.m05001	Os05g51720	hypothetical protein
11686.m00113	Os12g02050	expressed protein
11686.m01833	Os12g18490	hypothetical protein
11686.m02510	Os12g26060	thymidylate synthase, putative
11686.m03075	Os12g31620	NB-ARC domain, putative
11686.m03624	Os12g36920	calmodulin-binding protein homolog F14M19.80 - Arabidopsis thaliana
11686.m03959	Os12g39970	expressed protein
11686.m04229	Os12g42410	hypothetical protein
11686.m04375	Os12g43780	hypothetical protein
11687.m00208	Os11g02990	hypothetical protein
11687.m03478	Os11g37690	TBC domain, putative
11687.m03499	Os11g37880	NB-ARC domain, putative
gi|32978787|dbj|AK068762.1|		putative zinc finger protein {Oryza sativa (japonica cultivar-group);} ^|^GB|BAD17151.1|46805783|AP004853 putative zinc finger protein {Oryza sativa (japonica cultivar-group);}
gi|32986675|dbj|AK101466.1|		unknown {Zea mays;}
gi|32994814|dbj|AK109605.1|		unknown protein {Oryza sativa (japonica cultivar-group);}
gi|32995870|dbj|AK110661.1|		hypothetical protein similar to CCCH-type zinc finger protein {Oryza sativa (japonica cultivar-group);} ^|^GB|BAB55496.1|14090337|AP002972 hypothetical protein similar to CCCH-type zinc finger protein {Oryza-TRUNCATED-
STRESS TEST SET		
11667.m03867	Os01g40250	hypothetical protein
11667.m03967	Os01g41190	hypothetical protein
11667.m05207	Os01g53060	Mpv17/PMP22 family protein, expressed
11667.m05283	Os01g53790	expressed protein
11667.m05518	Os01g55870	chorismate mutase, putative
11667.m06686	Os01g66550	expressed protein
11667.m06935	Os01g68790	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain), putative
11668.m00135	Os02g02290	SNF2 family N-terminal domain, putative
11668.m00463	Os02g05250	hypothetical protein
9630.m02532	Os02g26800	expressed protein
11668.m04731	Os02g48720	adp,atp carrier protein, mitochondrial precursor (adp/atp translocase)(adenine nucleotide translocator) (ant)
11668.m05182	Os02g52880	AP2 domain, putative
11668.m05241	Os02g53400	Similar to thioredoxin-like 5
11669.m01922	Os03g18950	Similar to C-x8-C-x5-C-x3-H type Zinc finger protein, putative
11669.m05092	Os03g50530	expressed protein
11669.m06288	Os03g61710	putative galactose kinase
11669.m06400	Os03g62620	putative late embryogenesis abundant protein
11670.m01142	Os04g12730	hypothetical protein
11670.m03511	Os04g36750	Hsp20/alpha crystallin family, putative, (22.0 kDa class IV heat shock protein precursor, putative, expressed)
11670.m03831	Os04g40130	expressed protein, (Rf1 protein, mitochondrial precursor, putative, expressed)
11670.m04057	Os04g42170	hypothetical protein
11670.m05004	Os04g51250	hypothetical protein
11674.m00881	Os08g09350	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain), putative
11674.m01570	Os08g16050	senescence-associated protein 5
11674.m03032	Os08g30800	hypothetical protein
11674.m03294	Os08g33330	expressed protein
11674.m03727	Os08g37450	Signal recognition particle, alpha subunit, N-terminal, putative
11676.m02351	Os10g27280	Thaumatin family
11676.m02586	Os10g30040	BTB/POZ domain, putative
11676.m03439	Os10g37970	putative kinase

11676.m03612	Os10g39610	putative zinc-metallothionein
11676.m03864	Os10g41870	hypothetical protein (blstp:putative F-box domain containing protein)
11680.m01143	Os06g11730	RNA recognition motif-containing protein SEB-4
11681.m03078	Os09g33970	expressed protein
11682.m03691	Os05g38950	TBC domain, putative
11682.m04906	Os05g50920	amino acid permease-like protein
11686.m00607	Os12g06670	Protein kinase domain, putative
11686.m01254	Os12g12850	Clp amino terminal domain, putative
11686.m01325	Os12g13550	NB-ARC domain, putative
11686.m02388	Os12g24870	SWIM zinc finger, putative
11686.m03508	Os12g35790	hypothetical protein
gi|37990599|dbj|AK120976.1|		contains ESTs AU031905(R2501),AU164448(E50582),D22996(C1978A),AU031905 (R2501) similar to Arabidopsis thaliana F17A9.6 unknown protein {Oryza sativa (japonica cultivar-group);} ^|^GB|BAB19414.1|1-TRUNCATED-
NON-REGULATED LEARNING SET		
11667.m00008	Os01g01070	hypothetical protein, (GO: DNA binding)
11667.m00307	Os01g03930	hypothetical protein
11667.m00686	Os01g07500	Allinase, C-terminal domain, putative
11667.m00755	Os01g08090	UDP-glucoronosyl and UDP-glucosyl transferase
11667.m00842	Os01g08880	hypothetical protein
11667.m02971	Os01g31640	hypothetical protein
11667.m03019	Os01g32080	Thiamine pyrophosphate enzyme, central domain, putative
11667.m04784	Os01g49190	ATP synthase F1, beta subunit
11667.m04864	Os01g49930	expressed protein
11667.m05945	Os01g59800	Vacuolar protein sorting 36, putative
11667.m06307	Os01g63020	hypothetical protein
11667.m06457	Os01g64450	expressed protein
11667.m06552	Os01g65210	POT family, putative
11668.m00151	Os02g02410	dnaK protein
11668.m00299	Os02g03680	hypothetical protein
11668.m02043	Os02g21220	hypothetical protein
11668.m02881	Os02g30320	expressed protein
11668.m04809	Os02g49410	Histone-like transcription factor (CBF/NF-Y) and archaeal histone, putative
11668.m05210	Os02g53120	hypothetical protein
11668.m05324	Os02g54130	DnaJ domain, putative
11669.m00070	Os03g01660	Skp1 family, dimerisation domain, putative
11669.m01603	Os03g16020	Hsp20/alpha crystallin family, putative
11669.m01639	Os03g16360	hypothetical protein
11669.m01761	Os03g17440	expressed protein
11669.m02178	Os03g21260	2,3-bisphosphoglycerate-independent phosphoglycerate mutase
11669.m02921	Os03g29070	hypothetical protein
11669.m03785	Os03g38350	putative GDSL-like lipase/acylhydrolase
11669.m04604	Os03g46110	expressed protein
11669.m05080	Os03g50420	expressed protein
11669.m05138	Os03g50960	putative root-specific protein
11669.m05427	Os03g53600	expressed protein
11669.m05618	Os03g55770	hypothetical protein
11669.m05868	Os03g57970	putative protease inhibitor
11669.m05896	Os03g58170	hypothetical protein, (stem-specific protein TSJT1, putative, expressed)
11669.m06358	Os03g62280	hypothetical protein
11670.m03256	Os04g33850	hypothetical protein
11670.m03334	Os04g34590	hypothetical protein
11670.m04376	Os04g45080	expressed protein
11670.m04990	Os04g51130	Similar to AT3g12730/MBK21_9
11670.m05094	Os04g52060	Transposable element protein, putative
11670.m05235	Os04g53390	BTB/POZ domain, putative
11670.m05601	Os04g56760	PHF5-like protein
11670.m05605	Os04g56800	Pin1-type peptidyl-prolyl cis/trans isomerase
11670.m05627	Os04g56990	myb-like DNA-binding domain, SHAQKYF class, putative
11670.m05814	Os04g58690	Adenosine-deaminase (editase) domain, putative
11670.m05835	Os04g58890	expressed protein
11673.m00013	Os07g01110	Multicopper oxidase, putative
11673.m01183	Os07g12310	hypothetical protein
11673.m01683	Os07g17180	No apical meristem (NAM) protein, putative
11673.m03464	Os07g35870	Helix-loop-helix DNA-binding domain, putative

11673.m03530	Os07g36500	histone H4 - Arabidopsis thaliana
11673.m03647	Os07g37610	B3 DNA binding domain, putative
11673.m03760	Os07g38670	hypothetical protein
11673.m03964	Os07g40580	translation initiation factor eIF-5A
11673.m04194	Os07g42780	hypothetical protein
11673.m04329	Os07g44060	HAD-superfamily hydrolase, subfamily IA, variant 3, putative
11674.m00131	Os08g02230	plant-specific FAD-dependent oxidoreductase
11674.m02809	Os08g28650	hypothetical protein
11674.m02930	Os08g29800	Leucine Rich Repeat, putative
11674.m03250	Os08g32910	hypothetical protein
11674.m03306	Os08g33420	Agenet domain, putative
11674.m03775	Os08g37910	hypothetical protein
11676.m00234	Os10g03260	hypothetical protein
11676.m03045	Os10g34390	expressed protein
11676.m03241	Os10g36160	putative lipid transfer protein
11676.m03983	Os10g42870	putative peptide transporter
11680.m00388	Os06g04660	oxidoreductase, 2OG-Fe(II) oxygenase family, putative
11681.m02517	Os09g27390	transposon protein, putative, unclassified
11682.m00006	Os05g01060	expressed protein
11686.m00087	Os12g01800	hypothetical protein
11686.m00099	Os12g01920	hypothetical protein
11686.m01060	Os12g10950	hypothetical protein
11686.m01368	Os12g13970	hypothetical protein
11686.m01393	Os12g14220	expressed protein
11686.m01860	Os12g18720	hypothetical protein
11686.m03333	Os12g34110	H+-transporting two-sector ATPase (EC 3.6.3.14) protein 9 - barley mitochondrion
11686.m03720	Os12g37780	calmodulin-binding protein MPCBP
11686.m03755	Os12g38120	osmotin protein homolog - rice (fragment), putative
11686.m03962	Os12g40000	hypothetical protein
11686.m04019	Os12g40540	At3g04650/F7O18_13
11686.m04099	Os12g41260	protein kinase ATMRK1 (EC 2.7.1.-) [imported] - Arabidopsis thaliana
11686.m04147	Os12g41680	salicylic acid-induced protein 19
11686.m04206	Os12g42200	Sodium/hydrogen exchanger family protein, putative
11687.m00228	Os11g03180	hypothetical protein
11687.m03415	Os11g37060	F-box domain, putative
11687.m03497	Os11g37860	RGH1A
9635.m04857	Os07g48160	Similar to alpha-galactosidase-like protein
NON-REGULATED TEST SET		
11667.m00646	Os01g07120	DREB1
11667.m00685	Os01g07490	expressed protein
11667.m03122	Os01g33070	Similar to gene_id:K1F13.16
11667.m05243	Os01g53420	UDP-glucoronosyl and UDP-glucosyl transferase
11667.m06462	Os01g64500	hypothetical protein
11667.m06803	Os01g67570	hypothetical protein
11667.m07097	Os01g70810	Homeobox domain, putative
11667.m07480	Os01g74410	Similar to snapdragon myb protein 305 homolog
11668.m01821	Os02g19110	hypothetical protein
11668.m04080	Os02g42630	hypothetical protein
11669.m00232	Os03g03020	ribosomal protein L11, putative
11669.m00593	Os03g06390	expressed protein
11669.m01590	Os03g15900	SH3 domain, putative
11669.m02804	Os03g27470	hypothetical protein
11669.m04627	Os03g46310	expressed protein
11669.m05325	Os03g52690	expressed protein
11669.m05560	Os03g55260	putative cytochrome P450
11669.m05617	Os03g55760	hypothetical protein
11669.m06065	Os03g59700	putative cyclophilin
11669.m06109	Os03g60110	expressed protein
11670.m01128	Os04g12590	hypothetical protein
11670.m02266	Os04g24270	hypothetical protein
11670.m02537	Os04g26960	Terpene synthase family, metal binding domain, putative
11670.m03022	Os04g31700	expressed protein
11670.m04837	Os04g49200	oxidoreductase, 2OG-Fe(II) oxygenase family, putative
11670.m05485	Os04g55700	exonuclease, putative
11670.m05828	Os04g58830	Ribosome biogenesis regulatory protein (RRS1), putative

11673.m01593	Os07g16290	hypothetical protein
11673.m02953	Os07g31010	hypothetical protein
11673.m04261	Os07g43430	hypothetical protein
11674.m00053	Os08g01480	Cytochrome P450
11674.m00362	Os08g04440	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain), putative
11674.m00719	Os08g07820	hypothetical protein
11674.m03116	Os08g31630	trehalose-phosphatase, putative
11676.m01395	Os10g16970	Cytochrome P450
11676.m02896	Os10g32970	cap-binding protein p28
11676.m03301	Os10g36700	putative indole-3-acetic acid-regulated protein
11680.m00768	Os06g08160	hypothetical protein
11681.m02515	Os09g27370	At5g06680
11681.m02609	Os09g28190	hypothetical protein
11682.m00229	Os05g03040	APETALA2-like protein
11682.m01125	Os05g11900	hypothetical protein
11682.m01308	Os05g14180	AUX/IAA family
11682.m04124	Os05g43360	nadh-ubiquinone oxidoreductase 24 kda subunit, mitochondrial precursor(ec 1.6.5.3) (ec 1.6.99.3)
11682.m04929	Os05g51130	ATPase, AAA family, putative
11686.m01030	Os12g10660	B-box zinc finger, putative
11686.m01043	Os12g10780	DAHP synthetase I family, putative
11686.m01632	Os12g16540	expressed protein
11686.m02692	Os12g27830	Similar to AB023037 11-beta-hydroxysteroid dehydrogenase
11686.m02975	Os12g30620	hypothetical protein
11686.m03461	Os12g35360	At3g27530/MMJ24_7
11687.m03520	Os11g38070	expressed protein
gi|32986734|dbj|AK101525.1|		putative small nuclear ribonucleoprotein D1 polypeptide 16kDa; snRNP core protein D1; Sm-D autoantigen {Oryza sativa (japonica cultivar-group);}
gi|32988823|dbj|AK103614.1|		contains ESTs C23613(S14085),AU032777(S14085) similar to Arabidopsis thaliana chromosome 5, K15I22.11 unknown protein {Oryza sativa (japonica cultivar-group);} ^|^GB|BAA94239.1|7523511|AP001633 unnamed prote-TRUNCATED-
gi|32993536|dbj|AK108327.1|		unknown protein {Oryza sativa (japonica cultivar-group);}
rice|CF329806		N/A

Supplementary Table 3. Promoters used in positive and negative learning and test promoter sets in rice

Dyad number	Dyad sequence	*cdr* score	Dyad number	Dyad sequence	*cdr* score
1	GCGT{N33}GTGA	1	133	TCCG{N0}TTTC	0.916667
2	GTAT{N5}TCTC	1	134	CTCA{N29}TGGT	0.916667
3	ATAC{N39}AGGA	1	135	CTCA{N34}ACAA	0.916667
4	GGAT{N32}TCTC	1	136	CCAC{N7}TATA	0.916667
5	AAAT{N1}GCTG	1	137	TGCT{N49}TTTC	0.916667
6	TATG{N35}CGGA	0.97619	138	CATA{N12}ATGG	0.916667
7	CTTT{N37}CCCC	0.972222	139	CATA{N44}AGTT	0.916667
8	TCTC{N14}GTAG	0.972222	140	GAGT{N50}GAAA	0.916667
9	TTCT{N20}GAAC	0.972222	141	AGAC{N9}TTTA	0.916667
10	GGAC{N31}AAAA	0.972222	142	AGGA{N14}GAAC	0.916667
11	CGAT{N23}AGCC	0.966667	143	GAGG{N15}GAAT	0.916667
12	TAGG{N14}GCAG	0.966667	144	AAGA{N13}CGGA	0.916667
13	ATTA{N36}TCGG	0.966667	145	AGAG{N20}TGAT	0.916667
14	GCAC{N40}TATT	0.966667	146	GGGG{N29}TTAC	0.916667
15	GTAT{N32}ACCA	0.966667	147	ATTC{N42}CATA	0.904762
16	ATAT{N32}GGGG	0.966667	148	ATAT{N23}CGGA	0.904762
17	ATGG{N48}GTCA	0.966667	149	GGTC{N22}AAAA	0.904762
18	GATA{N45}CTGA	0.966667	150	CTTT{N36}CTTC	0.904762
19	CCCC{N42}CATG	0.966667	151	TTCT{N39}CCCC	0.904762
20	CCCT{N4}ATGC	0.966667	152	TATG{N19}ATAG	0.904762
21	CCTT{N47}GATA	0.966667	153	GAAT{N33}TGGA	0.904762
22	CTCT{N34}CAAC	0.966667	154	AAGA{N38}TTGC	0.904762
23	CTCT{N1}GGGG	0.966667	155	GTTC{N42}TTTT	0.902778
24	TCTG{N14}GATA	0.966667	156	CGGT{N45}ATAT	0.9
25	CCCG{N3}AGAG	0.966667	157	TAGT{N43}CTAT	0.9
26	TGCT{N30}GGAA	0.966667	158	CAAC{N3}CAGA	0.9
27	CACC{N48}GTCC	0.966667	159	CAAC{N2}TCAG	0.9
28	CACG{N18}TGGT	0.966667	160	CAGC{N29}TAAA	0.9
29	CATA{N26}CGCG	0.966667	161	TAAA{N34}GTAC	0.9
30	GAAC{N31}TGAT	0.966667	162	TAGA{N16}TAAG	0.9
31	GAAC{N44}CAAC	0.966667	163	TAGA{N18}GATG	0.9
32	GGAT{N14}CTGG	0.966667	164	CAAG{N15}TCTT	0.9
33	GAGA{N25}ACGG	0.966667	165	TGGA{N19}GTTG	0.9
34	AGAA{N5}AGGC	0.966667	166	ATTC{N4}ATCC	0.9
35	AGTA{N1}TGAT	0.952381	167	ATTC{N21}CTTA	0.9
36	TAAC{N39}TAGT	0.944444	168	ATTC{N4}GGAT	0.9
37	GTTA{N6}ATTC	0.944444	169	ATTT{N17}GAGG	0.9
38	ATAT{N23}GGAG	0.944444	170	GCTT{N42}TTTC	0.9
39	GGAT{N21}TATA	0.944444	171	ACCT{N12}TTCA	0.9
40	AAAA{N3}CGGC	0.944444	172	GTTC{N41}TTTC	0.9
41	GGGA{N27}TACA	0.944444	173	ATTG{N7}TATG	0.9
42	TGAC{N26}ACAG	0.933333	174	GCTG{N33}GTTA	0.9
43	CAAT{N48}GAGT	0.933333	175	GCTG{N41}GACT	0.9
44	CGGA{N33}TCTC	0.933333	176	GCTG{N1}AATG	0.9
45	TAAG{N31}GGAA	0.933333	177	GCTG{N12}AAAT	0.9
46	CAGG{N30}TTTT	0.933333	178	GCTG{N24}GGAA	0.9
47	CGAA{N27}AACT	0.933333	179	ATGT{N49}GTGT	0.9
48	TGGA{N43}GTAC	0.933333	180	ATGT{N16}CCAC	0.9
49	GCTC{N46}ACGA	0.933333	181	GCAC{N28}TGGA	0.9
50	ATCC{N12}GTAT	0.933333	182	GTAC{N1}ATTC	0.9
51	GCTG{N7}CGAT	0.933333	183	ACGT{N48}TGGA	0.9
52	GCTG{N21}GTGT	0.933333	184	ATAT{N17}GCGG	0.9
53	GCTG{N21}GAAC	0.933333	185	ATAT{N35}CGCG	0.9
54	GTCA{N26}TATG	0.933333	186	GTGC{N15}TGTT	0.9
55	ATGC{N51}TCAG	0.933333	187	ATGA{N43}CGAG	0.9
56	ATGC{N7}CCGA	0.933333	188	ATGA{N33}TTAC	0.9
57	ATGT{N14}CCCC	0.933333	189	GCAA{N33}TCTT	0.9
58	ATGT{N25}TCCG	0.933333	190	GTAA{N21}ATTC	0.9
59	GCGC{N7}TTCT	0.933333	191	ATAA{N13}AGGA	0.9
60	GCGC{N5}CCTT	0.933333	192	ATAG{N34}AGAG	0.9
61	ATAC{N4}TCTC	0.933333	193	AGTC{N28}AGTT	0.9
62	ATAT{N22}TCGG	0.933333	194	AATC{N4}CCGA	0.9
63	GTGT{N35}ACAA	0.933333	195	AGTA{N6}GGAT	0.9
64	GCAG{N10}AATA	0.933333	196	AGTG{N43}GTAC	0.9
65	ACAA{N24}AGAC	0.933333	197	GATA{N5}TCTC	0.9
66	ACGG{N26}CAAA	0.933333	198	AACG{N25}TGGT	0.9
67	ACGG{N17}ATAT	0.933333	199	AATG{N1}TGGT	0.9
68	ATAG{N22}TAGT	0.933333	200	AATG{N37}TCCA	0.9

69	GATC{N27}CCTA	0.933333	201	AATG{N18}TACA	0.9
70	GATC{N28}CTAA	0.933333	202	AGCA{N43}TATC	0.9
71	GATC{N50}GAAC	0.933333	203	GGTA{N43}CTAA	0.9
72	AGTG{N19}GATA	0.933333	204	CTTT{N29}TAGG	0.9
73	AGTG{N27}CTGT	0.933333	205	CTTT{N0}TCAG	0.9
74	GATA{N43}CTCT	0.933333	206	TCTT{N10}AGAG	0.9
75	GGCA{N3}AGTT	0.933333	207	TTCC{N46}ATAC	0.9
76	AATG{N44}GCTG	0.933333	208	TTCC{N27}ACAA	0.9
77	AATG{N28}AGTA	0.933333	209	CCCC{N45}AACC	0.9
78	TCTT{N39}CCCC	0.933333	210	TTTC{N13}CAGT	0.9
79	TCTT{N30}TGCA	0.933333	211	TTTC{N7}TAGG	0.9
80	CCCC{N29}GCAT	0.933333	212	TTTT{N43}GTCG	0.9
81	CCCC{N50}AGTT	0.933333	213	CTTG{N33}ATTC	0.9
82	CCTT{N40}ACAT	0.933333	214	TCTG{N37}CAAT	0.9
83	CTCT{N19}GTTA	0.933333	215	TCTG{N44}TGAA	0.9
84	TTTA{N35}GGGG	0.933333	216	TTCA{N48}ATCT	0.9
85	CTGT{N36}TGAT	0.933333	217	TTTA{N25}GCTG	0.9
86	TCAT{N38}CTCA	0.933333	218	TCAT{N6}ACAG	0.9
87	CCGT{N21}TTTA	0.933333	219	CTAT{N9}CACA	0.9
88	CTAC{N5}TAGT	0.933333	220	TCAG{N50}GTTG	0.9
89	CTAT{N33}GTGT	0.933333	221	TCGA{N23}ATGT	0.9
90	CTGA{N30}CAGA	0.933333	222	TTAG{N39}TCCA	0.9
91	TTAG{N47}GGGT	0.933333	223	TTAG{N39}GAAG	0.9
92	CGTC{N12}CGTT	0.933333	224	CCAG{N1}TTAA	0.9
93	TACC{N50}TATG	0.933333	225	CCGA{N32}TAAG	0.9
94	CACC{N27}CACA	0.933333	226	CTAA{N19}CACA	0.9
95	CATT{N22}CCTT	0.933333	227	TACC{N51}ATGT	0.9
96	TATA{N30}TTCG	0.933333	228	TACC{N20}TCTC	0.9
97	TATA{N24}CGCG	0.933333	229	TACC{N33}TCAT	0.9
98	TATA{N32}GGGG	0.933333	230	TACT{N33}ATTC	0.9
99	TATG{N47}GTGT	0.933333	231	TGCT{N8}CGAT	0.9
100	TATG{N23}GAGA	0.933333	232	TGCT{N39}AGCT	0.9
101	TATG{N21}GGAG	0.933333	233	TGCT{N15}TTAG	0.9
102	CACG{N47}TTAG	0.933333	234	CATT{N18}GTTA	0.9
103	CGCG{N5}CCTT	0.933333	235	TGTC{N17}CAAA	0.9
104	TGTA{N3}TTGC	0.933333	236	TGTT{N30}ACAC	0.9
105	TGTA{N17}TGCT	0.933333	237	TACA{N27}AGAG	0.9
106	AGGC{N5}TAAA	0.933333	238	TATA{N49}CCTT	0.9
107	GAAC{N36}TCCT	0.933333	239	TATG{N46}AGTG	0.9
108	GGAC{N2}ATAT	0.933333	240	TGCA{N1}GTAC	0.9
109	AAGC{N34}TGCC	0.933333	241	TGCA{N35}GGTA	0.9
110	AGAC{N31}GGAT	0.933333	242	CACA{N21}CAAG	0.9
111	AGAT{N20}GTGG	0.933333	243	CACG{N20}AAGA	0.9
112	AGGA{N33}TTCA	0.933333	244	CGCA{N19}TAGT	0.9
113	GGAA{N19}AGTC	0.933333	245	CGCG{N8}TTCT	0.9
114	AGAA{N42}GAGT	0.933333	246	GAAC{N37}ATGC	0.9
115	TAAC{N50}ATAG	0.928571	247	GAAC{N31}CATC	0.9
116	ACAT{N42}GTTA	0.928571	248	GAAT{N5}TCTC	0.9
117	TGTA{N29}AGAG	0.928571	249	GGAC{N17}CTCT	0.9
118	AAAA{N9}AGGC	0.928571	250	AAAT{N45}CTTC	0.9
119	TAGT{N46}GCTG	0.916667	251	AAGC{N13}GTAG	0.9
120	TAGT{N39}GAAG	0.916667	252	AAGT{N24}CCAA	0.9
121	TGGT{N30}TTGT	0.916667	253	AGGA{N4}TAGA	0.9
122	CAAA{N45}GAGA	0.916667	254	GAGA{N39}AGCG	0.9
123	ATGC{N41}GAGT	0.916667	255	GAGA{N2}TTAT	0.9
124	GTAC{N39}AATA	0.916667	256	GGAA{N30}TTAG	0.9
125	GCAA{N38}CTCT	0.916667	257	GGAA{N37}TACG	0.9
126	AGTC{N2}TATA	0.916667	258	GGAA{N35}AAGT	0.9
127	AGTA{N35}GGAA	0.916667	259	AAAA{N11}GGCC	0.9
128	AGTA{N19}AAGA	0.916667	260	AAGA{N5}GTTA	0.9
129	GATA{N42}ACAA	0.916667	261	AGAG{N51}GGAA	0.9
130	AATA{N36}CCTT	0.916667	262	GGGA{N10}TTCT	0.9
131	TCTC{N36}CAAC	0.916667	263	GGGA{N21}GAAA	0.9
132	TCTC{N2}GTCA	0.916667			

Supplementary Table 4. List of top 263 dyads found in promoters of glucanase, chitinase, PR-1, PR-4, PR-5, and PR-9 genes

Dyad identifier	sequence	*cdr* score	Promoter identifier	Position upstream	PLACE database annotation
13	*GCAC{N40}TATT*	*0.966667*	gl2	319	
			k15	1150	
			pr9-1	1831	
			pr5-5	1731	
			pr4-1	675	
33	**AGAA{N5}AGGC**	**0.966667**	gl1	203	**HBOXCONSENSUSPVCHS**
			gl21	976	
			k1	562	
			pr1-1	154	
			pr5-1	1161	
41	*TGAC{N26}ACAG*	*0.933333*	gl21	709	
			k15	299	
			pr1-1	166	
			pr9-1	1625	
			pr5-9	908	
64	*ACAA{N24}AGAC*	*0.933333*	gl5	638	
			pr1-1	1061	
			pr9-1	324	
			pr5-1	968	
			pr4-1	542	
94	*CATT{N22}CCTT*	*0.933333*	gl5	1896	
			k15	184	
			pr1-1	498	
			pr9-7	498	
			pr5-5	834	
103	**TGTA{N3}TTGC**	**0.933333**	gl21	1904	**HSELIKENTACIDICPR1**
			pr9-7	1812	
			pr5-1	415	
			pr5-1	85	
			pr5-5	6	
105	**AGGC{N5}TAAA**	**0.933333**	k1	522	**HBOXCONSENSUSPVCHS**
			k15	351	
			pr1-1	1077	
			pr9-7	472	
			pr5-5	276	
121	*CAAA{N45}GAGA*	*0.916667*	gl2	1199	
			gl2	1197	
			k15	546	
			pr1-1	504	
			pr5-1	1316	
			pr4-1	519	
125	***AGTC{N2}TATA***	***0.916667***	gl5	1511	-141NTG13
			k15	1228	
			pr1-1	482	
			pr9-1	1617	
			pr5-1	1853	
			pr5-1	1775	
			pr5-1	1635	
			pr5-1	1482	
			pr5-9	1926	
128	*GATA{N42}ACAA*	*0.916667*	gl1	706	
			gl5	1630	
			k15	78	
			pr1-1	100	
			pr9-1	1231	
			pr9-1	517	
			pr5-1	996	
131	**TCTC{N2}GTCA**	**0.916667**	gl21	1031	-141NTG13', OBF5ATGST6
			gl21	1031	
			k1	251	
			pr1-1	855	

			pr5-1	653	
			pr5-5	946	
132	*TCCG{N0}TTTC*	*0.916667*	gl2	636	
			k1	790	
			pr1-1	1229	
			pr9-7	1002	
			pr5-5	768	
			pr5-5	610	
133	*CTCA{N29}TGGT*	*0.916667*	gl1	847	
			k15	433	
			pr1-1	828	
			pr9-1	1376	
			pr5-5	1802	
			pr5-9	1187	
243	*AAGA{N13}CGGA*	*0.916667*	gl1	336	
			gl21	744	
			k1	638	
			pr9-7	1249	
			pr5-1	882	
			pr4-1	49	
148	*GGTC{N22}AAAA*	*0.904762*	gl21	1971	
			gl21	971	
			pr1-1	1846	
			pr9-1	102	
			pr9-7	1423	
			pr5-5	132	
			pr5-5	131	
			pr4-1	559	
149	*CTTT{N36}CTTC*	*0.904762*	gl2	262	
			gl21	245	
			k15	855	
			k15	183	
			pr9-1	1763	
			pr5-9	538	
			pr4-1	237	
158	**CAAC{N2}TCAG**	**0.9**	pr9-1	1968	HDMOTIFPCPR2
			pr9-1	1225	
			pr9-7	1656	
			pr9-7	38	
			pr5-1	790	
160	*TAAA{N34}GTAC*	*0.9*	gl1	1165	
			k15	239	
			pr1-1	134	
			pr9-7	434	
			pr5-5	1813	
165	*ATTC{N4}ATCC*	*0.9*	k15	730	
			pr1-1	227	
			pr9-7	1160	
			pr5-9	534	
			pr4-1	322	
171	*GTTC{N41}TTTC*	*0.9*	gl21	863	
			k15	1617	
			pr1-1	1189	
			pr9-1	731	
			pr4-1	98	
182	*ACGT{N48}TGGA*	*0.9*	gl1	324	
			k1	718	
			pr1-1	192	
			pr9-1	81	
			pr5-1	360	
191	**ATAG{N34}AGAG**	**0.9**	gl2	861	
			k15	1289	
			pr5-1	1868	
			pr5-1	1790	
			pr5-1	354	
			pr5-9	1639	

204	**CTTT{N0}TCAG**	**0.9**	gl1	660	-141NTG13
			gl5	1900	
			k15	549	
			pr9-7	1656	
			pr5-9	63	
216	*TTTA{N25}GCTG*	*0.9*	gl21	799	
			pr1-1	1168	
			pr9-1	667	
			pr5-1	1149	
			pr4-1	430	
239	**TGCA{N1}GTAC**	**0.9**	gl5	678	ABADESI1, ABREATRD22**,** **WARBNEXTA**, **RSRBNEXTA**
			pr1-1	134	
			pr1-1	7	
			pr9-1	744	
			pr5-1	409	
251	*AAGT{N24}CCAA*	*0.9*	gl21	954	
			pr1-1	1737	
			pr9-1	1558	
			pr5-9	1355	
			pr4-1	423	
254	**GAGA{N2}TTAT**	**0.9**	pr9-1	1952	14BPATERD1**,** **HSELIKENTACIDICPR1**
			pr9-1	1952	
			pr5-9	923	
			pr4-1	513	
			pr4-1	510	
259	**AAGA{N5}GTTA**	**0.9**	gl1	122	**HBOXCONSENSUSPVCHS**
			gl5	1467	
			gl21	1735	
			pr1-1	540	
			pr9-1	1289	

Supplementary Table 5. Top 28 dyads found in positive learning promoter set. Dyads in bold have a corresponding motif in the PLACE database. Dyads in italics occur in promoters of at least 5 different gene families.