Mol Genet Genomics (2011) 285:375–391
DOI 10.1007/s00438-011-0605-4

ORIGINAL PAPER

Prediction of new abiotic stress genes in *Arabidopsis thaliana* and *Oryza sativa* according to enumeration-based statistical analysis

Mátyás Cserháti · Zoltán Turóczy · Zoltán Zombori · Miklós Cserző · Dénes Dudits · Sándor Pongor · János Györgyey

Received: 3 December 2010 / Accepted: 31 January 2011 / Published online: 25 March 2011

Abstract Plants undergo an extensive change in gene regulation during abiotic stress. It is of great agricultural importance to know which genes are affected during stress response. The genome sequence of a number of plant species has been determined, among them *Arabidopsis* and *Oryza sativa*, whose genome has been annotated most completely as of yet, and are well-known organisms widely used as experimental systems. This paper applies a statistical algorithm for predicting new stress-induced motifs and genes by analyzing promoter sets co-regulated by abiotic stress in the previously mentioned two species. After identifying characteristic putative regulatory motif sequence pairs (dyads) in the promoters of 125 stress-regulated *Arabidopsis* genes and 87 *O. sativa* genes, these dyads were used to screen the entire *Arabidopsis* and *O. sativa* promoteromes to find related stress-induced genes whose promoters contained a large number of these dyads found by our algorithm. We were able to predict a number of putative dyads, characteristic of a large number of stress-regulated genes, some of them newly discovered by our algorithm and serve as putative transcription factor binding sites. Our new motif prediction algorithm comes complete with a stand-alone program. This algorithm may be used in motif discovery in the future in other species. The more than 1,200 *Arabidopsis* and 1,700 *Orzya sativa* genes found by our algorithm are good candidates for further experimental studies in abiotic stress.

Keywords Abiotic stress · *Arabidopsis thaliana* · Dyad · *Oryza sativa* · Promoter · Transcription factor binding site

Communicated by Y. Van de Peer.

Electronic supplementary material The online version of this article (doi:10.1007/s00438-011-0605-4) contains supplementary material, which is available to authorized users.

M. Cserháti (✉) · Z. Turóczy · Z. Zombori · D. Dudits · J. Györgyey
Biological Research Center, Institute of Plant Biology, Hungarian Academy of Sciences, P.O. BOX 521, Temesvári Krt. 62, 6701 Szeged, Hungary
e-mail: csmatyi@brc.hu

Z. Turóczy
e-mail: turoczy@brc.hu

Z. Zombori
e-mail: zzoli@brc.hu

D. Dudits
e-mail: dudits@brc.hu

J. Györgyey
e-mail: arthur@brc.hu

M. Cserző
Institute of Physiology, Semmelweis University, 1094 Budapest, 37-47 Tűzoltó u, Hungary
e-mail: cserzo@puskin.sote.hu

S. Pongor
ICGEB, Padriciano 99, 34012 Trieste, Italy
e-mail: pongor@icgeb.org

Abbreviations

ABA	Abscisic acid
AUC	Area under curve
PLACE	Plant cis-acting regulatory DNA elements
REP	Regulatory element pair
rev comp	Reverse complement
ROC	Receiver operating characteristic
TC	Tentative consensus
TFBS	Transcription factor binding site
TIGR	The institute for genome research

Introduction

In higher plants, the extensive reprogramming of gene expression patterns is one of the key mechanisms in adaptation to suboptimal environmental conditions. *Abiotic stress* is a summary concept meant to denote external sources of stress, including cold, drought, osmotic, oxidative, and salt stress, with general mechanisms underlying the resistance of plants to the corresponding conditions (Mahajan and Tuteja 2005). Studies toward the identification of genes involved in stress responses are many times based on the experimental identification of genes whose promoters bind specific transcription factors induced by such factors (Gómez-Porras et al. 2007; Wang et al. 2008). Even though expression of stress responsive genes is linked to the presence of common Transcription factor binding site (TFBS's) in their promoters, currently available data are usually restricted to the analysis of individual genes and pathways. Information gleaned from genome-wide identification of stress responsive elements and promoters may be facilitated in plant breeding experiments, which may improve crop harvests (European Plant Science Organization 2005).

Until now, the whole genome sequence has been completed for several plant species, such as *Arabidopsis thaliana*, *Oryza sativa*, *Medicago truncatula*, *Lotus japonicus*, *Populus trichocarpa*, *Zea mays*, and *Brachypodium distachyon*. Yamaguchi-Shinozaki outlined a basic stress response network within *Arabidopsis*, involving ABA-dependent and independent pathways (Yamaguchi-Shinozaki and Shinozaki 2005) in which a number of well-known abiotic stress TFBS's are involved, which in plants are about 5–10 bp long (Solovyev et al. 2010).

Many TFBS's lie in close proximity to one another in the promoter region, therefore different protein–protein interactions occur between individual TF's and other regulatory proteins (Wray et al. 2003). For example, some of the best known abiotic stress response elements in plants are the ABA responsive element (ABRE) element (represented by the motif ACGTGKC) (Hattori et al. 2002), and is found in a number of monocot species, such as barley, rice, and wheat, but has been discovered and characterized in *Arabidopsis* (Gómez-Porras et al. 2007). The drought responsive element (DRE) element (represented by the motif RCCGAC), and MYB and MYC binding sites are also involved in drought and cold stress (Shinozaki et al. 2003). The ABRE element also co-occurs with other abioitic stress motifs, such as the DRE or the coupling element (CE), thereby forming TFBS modules (Gómez-Porras et al. 2007; Zhang et al. 2005).

Walther et al. have shown in a study of *Arabidopsis* that upstream promoters and promoters taking part in multiple stimuli tend to have larger promoters and a larger density of TFBS's within them. 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 (RE) as their main role in a gene cascade or biochemical pathway to produce a protein or enzyme with a specific non-regulatory function. The TATA-box was found by these researchers in promoters of genes connected many times to some forms of abiotic stress response, which were shown to respond to many kinds of external stimuli (Walther et al. 2007). Therefore, we can reason that interactions between TFBS's in abiotic stress promoters are high. 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 et al. 2006; Vardhanabhuti et al. 2007).

In spite of the considerable amount of work done on stress responsive genes, relatively few promoter elements have been discovered and examined thoroughly that are involved in abiotic stress response. As of today there are a number of methods for identifying combinations of motifs within promoters (Sandve and Drabløs 2006). Computational motif discovery has been successfully used in simple organisms such as yeast; however, analysis of more complex genomes of higher organisms represents a challenge. The search for common elements in gene families as diverse as the response to drought, cold, and osmotic stress is a problem that seems difficult for the algorithms primarily designed to analyze individual pathways.

The goal of this work is to find common regulatory element pairs (REP's) in promoters of a representative set of genes involved in the response of drought and related stresses in *A. thaliana* and *O. sativa* as well to predict newer genes in the promoterome of these two plant species which could be implied in abiotic stress. These genes could then be subsequently tested and used to increase abiotic stress tolerance in *O. sativa*.

Since abiotic stress pathways overlap, we focus on finding REs common to all of these pathways. The algorithm is built up in such a way so that those elements are found which are statistically over-represented in a large number of input promoters. We identified a non-redundant set of 169 abiotic stress genes (which were split up into a learning set of 125 promoters, and a tuning set of 44 promoters) in Arabidopsis based on expressed sequence tag (EST) data from the TIGR database (Quackenbush et al. 2001; Lee et al. 2005), and 129 *O. sativa* drought stress genes (which were split up into a learning set of 87 promoters, and a tuning set of 42 promoters) studied by our own group which take part in a wide variety of abiotic stress responses (Online Resources 1 and 2, "selected genes" worksheet). In order to increase the thoroughness of

the analysis we applied an exhaustive enumeration of motifs that we subsequently evaluated with a cumulative statistical method developed by ourselves. Here we report a number of pentamer dyad motifs that are present in a wide variety of stress induced promoters while they are not significantly present in promoters not induced by abiotic stress as well as abiotic stress-induced genes newly discovered by the algorithm whose promoters also contain such elements.

Materials and methods

Dyad definition

The regulatory promoter elements analyzed in this paper were pairs of oligomers called dyads (see Fig. 1). A dyad is made up of a head and a tail motif (specifically in our case 5 bp long each), and a characteristic spacer length. The occurrence of a given dyad was calculated at different spacer lengths (0–52 bp) between the head and the tail motif. The spacer length is calculated for a given dyad where the dyad occurs with the greatest frequency in the positive learning set (see Fig. 2).

Promoter selection

For Arabidopsis, the 3,000 upstream sequences for all genes were downloaded from the TAIR website (ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/OLD/), and were truncated to 2 kbp. This corresponds to the average intergenic region for Arabidopsis (Picot et al. 2010). 169 genes were selected because they corresponded to TC sequences which were comprised of EST sequences, 75% of which came from an EST library produced under abiotic stress conditions. This was split up into a learning set of 125 promoters, and a tuning set of 44 promoters. 125 + 44 non-stress genes were randomly selected from the whole Arabidopsis genome, and their expression profile (relative expression change less than two-fold under stress conditions) was checked in the Genevestigator database, and assigned to the non-stress learning and tuning sets accordingly. For a list of promoters included in the study see the "selected genes" worksheets in Online Resource 1.

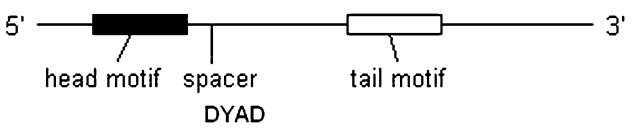

Fig. 1 The dyad is made up of a head and tail motif, which occurs at a specific distance from each other (that is, spacer length). The occurrence of all possible pentamer dyads was enumerated for motif distances of 0–52 bp. A characteristic spacer length is defined for each head and tail pair where the dyad occurs the most frequently within the positive promoter set

For *O. sativa*, 129 drought-stressed and 143 non-stress genes were selected from an experimental drought stress dataset, provided by our colleague, Zombori et al., personal communications. Expression-level change was recorded for each gene under control conditions (100% water content) and drought conditions (20% water content). Stress genes were selected whose expression level change was twofold under drought conditions. The 143 non-stress genes were selected on the basis of their expression level change being between ±0.33. Eight other *O. sativa* genes were selected because they were shown to be induced by abiotic stress (cold, drought, osmotic stress) in our other experiments.

The exact coordinates for the *O. sativa* promoter sequences were taken from the all.1kUpstream.gz promoter sequence file from the TIGR/JCVI website (ftp://ftp.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_5.0/all.chrs/). This file, however, contained only 1 kbp sequences, so we had to extract the whole 2 kbp promoter sequence from the 12 *O. sativa* chromosome (all.con) sequences using our own script.

The *O. sativa* promoter sequences were split up into the four following categories: a stress learning set containing 87 promoters and a stress tuning set containing 42 promoters (34 from the selected 129 genes, with 8 other of our own genes). 87 promoters were put into a non-stress learning set, and a further 57 promoters were put into a non-stress tuning set. One further promoter was put into the non-stress tuning set because our experimental data showed it to be non-stress inducible. The list of promoters used in the learning set and their expression level data may be found in the supplementary Excel worksheet "selected genes" in Online Resource 2.

Dyad selection

The main concept in our approach was to differentiate between dyads which occured mostly in stress promoters (positive set) and non-stress promoters (negative set). For this, we calculated the occurence of a given dyad over all spacer lengths from 0 to 52 bp in both the stress learning set and the non-stress learning set. For each dyad a characteristic spacer length was calculated where the dyad occurs the most in the stress learning promoter set. The occurence of the dyad at a spacer length ±1 bp of the characteristic spacer length was also taken into account as a wobbling factor. We therefore characterized a given dyad by its head and tail motifs, 5 bp long in our case, as well as this specific spacer length.

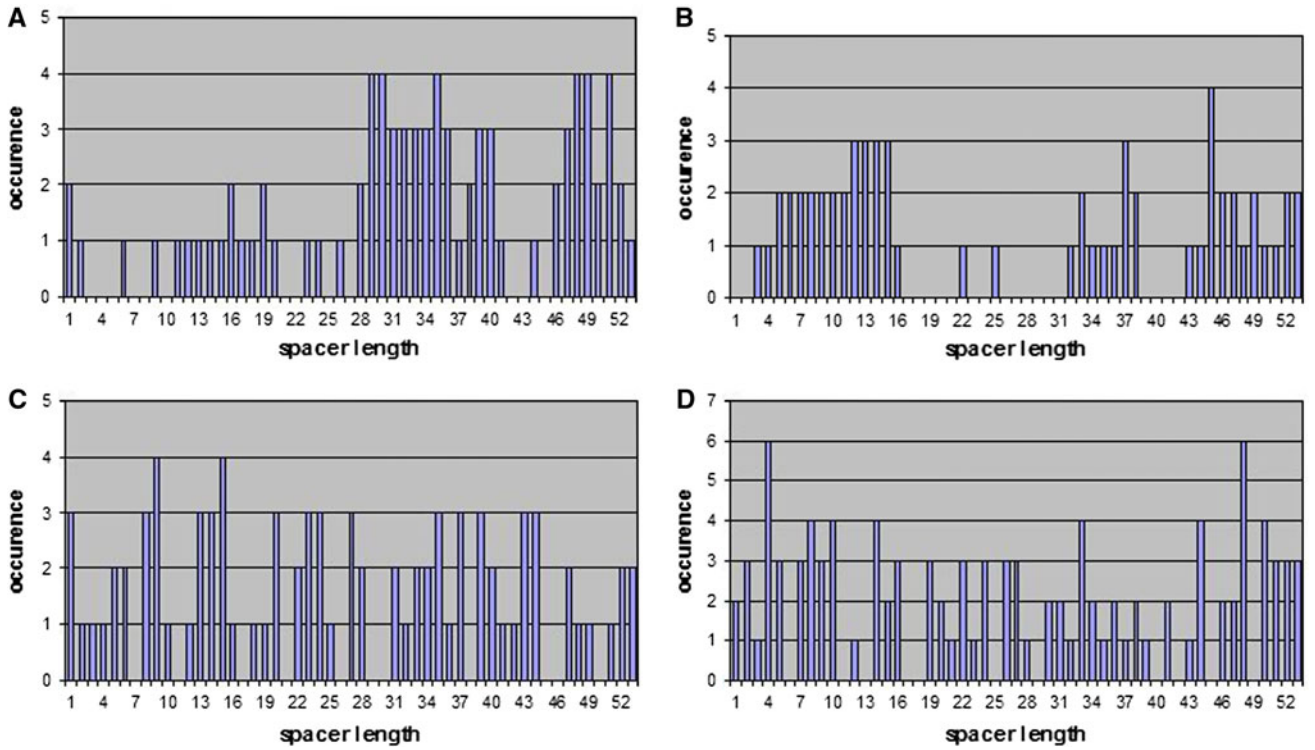

Fig. 2 **a** Dyad distribution of the dyad ACGTGN$_n$TTTTT in the stress learning promoter set in *Arabidopsis*. The consensus sequence for monocot ABRE elements is MGTACGTGKC, of which the core sequence ACGT is called the G-box, which occurs in the promoter of a number of genes regulated by abiotic stress and abscisic acid (Hattori et al. 2002). The dyad occurs with greater frequency in the stress learning set than the non-stress learning set, therefore its numerical measure is high. **b** Dyad distribution of the dyad ACGTGN$_n$TTTTT in the non-stress learning promoter set. **c** Dyad distribution of the dyad ATGATN$_n$TTTAT that is not associated with stress-related promoters in the stress learning promoter set. The head and tail motifs of this dyad occur just about the same number of times within both the positive and the negative promoter set, therefore its numerical measure is low. We can therefore infer from this that this dyad is biologically irrelevant. **d** Dyad distribution of the dyad ATGATN$_n$TTTAT in the non-stress learning promoter set

In order to calculate the statistical significance of a given dyad we scored each one by calculating how many stress learning and non-stress learning promoters each specific dyad occurred in N_{stress} and $N_{\text{nonstress}}$. To obtain the dyad score we calculated the following weight measure:

$$\text{cdr} = \frac{N_{\text{stress}} - N_{\text{nonstress}}}{N_{\text{stress}}}. \quad (1)$$

Here cumulative difference ratio is termed the cdr. This mathematical measure was calculated for all possible dyads in *Arabidopsis* and *O. sativa* with a minimal occurence of five in the stress learning promoter set.

Selection of TRANSFAC and PLACE stress motifs

We studied the distribution of 37 well-known plant stress transcription factor binding sites from the TRANSFAC and PLACE databases. 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.

We studied the occurence of pairs of these TRANSFAC/PLACE motifs in the stress and non-stress learning promoter sets. In other words, we formed dyads out of these motifs and calculated their individual cdr scores similar to the de novo dyad analysis. Here the maximum spacer length was limited to 52 bp. Overall, 277 TRANSFAC/PLACE dyads were found in the learning sets. Only ten 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 its cdr score can be seen in the "TRANSFAC + PLACE motifs" worksheet in Online Resource 1.

Scoring of promoters in the tuning set and calculation of AUC values (ROC analysis)

The dyads we selected were used to search the *Arabidopsis* and *O. sativa* promoteromes. In order to get the best results we analyzed the distribution of the dyads in the tuning set (stress promoters plus non-stress promoters). Since the

equation which calculates the cdr score takes the number of dyad occurences into account (see Eq. 1), 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. This process can be seen in Fig. 3.

In *Arabidopsis* we selected those dyads with a minimal cdr score of 0.6–1.0 with increments of 0.1. In *O. sativa* the minimum cdr score was 0.5–1.0 with the same increment. The dyads' distribution was also calculated where the dyads' head and tail motif were allowed to wobble upstream and downstream of the characteristic spacer length by 0 to ±5 bp. Furthermore, those dyads were selected where each dyad occurred a minimum of 5–20 times in the stress promoter set in *Arabidopsis* and 5–14 times in *O. sativa* (dyads did not occur with frequencies above the upper bounds). 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 *O. sativa*), and that when performing ROC analysis, these dyads saturated the test promoters, covering 1,735 bp on average. The tuning promoter set was also analyzed in such a way that the distribution of the 37 selected TRANSFAC and PLACE motifs were also taken into account. In this case these motifs' cdr score was also added to the promoter weight score if present in the given promoter.

The individual promoters were scored by adding up the individual cdr scores of all of the dyads occuring in them, that is,

$$S_{\text{promoter}} = \sum_{i}^{N} n_i \cdot cdr_i. \quad (2)$$

Similarly, the score for an individual promoter is,

$$S_{\text{promoter}} = \sum_{i}^{N} n_i \cdot cdr_i + \sum_{1}^{37} n_i \cdot cdr_i, \quad (3)$$

where the sum of the *cdr* scores of the 37 TRANSFAC and PLACE stress motifs is also added to the promoter's score in the case where these motifs were also included in the analysis. Here N signifies the number of dyads used in the given dyad set, and n_i and cdr_i indicate the number and cdr score of the ith dyad.

All stress promoters and non-stress promoters in the tuning sets were scored this way. The tuning stress promoters were characterized by a 1, whilst the tuning non-stress promoters were characterized by a 0 (meaning that their relative expression change during abiotic stress was greater then or equal to 2, see "Determination of expression change for selected genes"). A total of $5 \times 16 \times 6 \times 2 = 960$ possible AUC values were calculated for all parameter combinations in *Arabidopsis* (in *O. sativa* this was equal to $6 \times 6 \times 16 \times 2 = 1{,}152$ dyad sets). Those parameters (minimum dyad score, spacer wobbling, minimum number of dyads in stress learning promoter set, and usage or non-usage of TRANSFAC and PLACE motifs) were selected which produced the highest AUC value for the promoterome search. p values for AUC values were calculated by MedCalc for Windows, Version 11.3.6 (MedCalc Software, Mariakerke, Belgium).

Clustering of dyads

In *Arabidopsis*, the sequences of all optimal dyads were compared to one another in a pairwise manner. A local ungapped 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. A perfect base match counted as one point, and a C-G or A-T match counted as half point. Any base matched with an *N* was counted as zero. Two dyads belonged to the same cluster if they had a minimum score of 7.

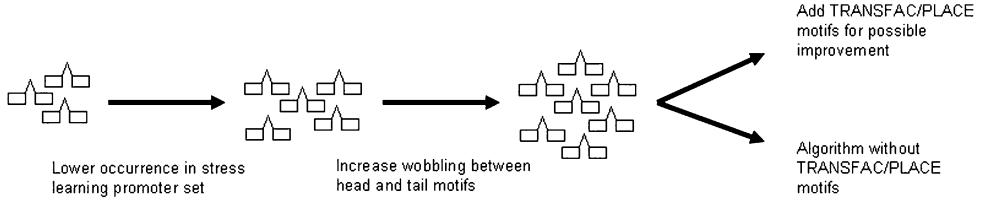

Fig. 3 During the process of parameterizing the dyads for ROC analysis, more dyads can be picked up if the minimum occurrence in the stress learning set is lowered (in increments), as well as allowing the head and tail motifs to wobble relative to each other (0–5 bp). Doing so leads to different dyad sets each differing in size. Also, by taking the occurrence of the TRANSFAC + PLACE motifs into consideration, the behavior of the algorithm can also be altered

Scoring of REs used in RE network analysis

In the analysis of the regulatory network of abiotic stress promoters we checked the frequency of all REs within 100 bp from each other within the top 3,100 candidate promoters found by the algorithm shown to be induced by stress (N_{stress}) as well as in the top 3,100 candidate promoters found by the algorithm shown not to be induced by stress ($N_{nonstress}$). A RE was taken to be either a single dyad found by the algorithm, a dyad cluster, or one of the 37 PLACE or TRANSFAC motifs used all throughout the analysis. In this way we studied dyad dyads. The cdr score value calculated for each regulatory element pair (REP) was taken to be

$$(N_{stress} - N_{nonstress})/N_{stress}, \quad (4)$$

which is similar to the cdr score for simple dyads. We studied the top 1,224 REP's 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*.

Calculation of Jacquard coefficient and promoter distances

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 would be

$$J = \frac{N_{AB}}{N_A + N_B - N_{AB}}. \quad (5)$$

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.

Determination of expression change for selected genes

In order to check whether a given gene in *Arabidopsis* or *O. sativa* from a promoterome search was stress induced, we determined that the relative gene expression change for such a gene is equal or >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 *O. 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 rice genes under cold, drought, and salt stress). Here, the expression level for stress experiments were divided by the corresponding control experiments.

Results

Selection of stress-induced promoters for the analysis

The TIGR database provides a comprehensive collection of *Arabidopsis* tentative consensus (TC) gene sequences, which are made up of EST sequences coming from different libraries. We searched for genes through a keyword search involved in salt, cold, osmotic, and drought stresses. Their abiotic stress expression profiles were checked in the Genevestigator database to make sure that they exhibited an at least twofold increase in the expression in at least one of the abiotic stress experiments in that database. 169 such genes were found and 125 of these promoters were put into the stress learning set, and 44 in the stress tuning set (which are three-fourth and one-fourth the size of the whole set of 169 promoters). The regions maximum 2 kb upstream of the ATG start site, excluding the overlaps with the coding regions of upstream genes were collected as the examples of stress-induced promoters. A matching number of non-induced promoters were randomly selected from the genes that were not represented in stress-induced libraries (see "Materials and methods" for details). The promoters of these genes served as the non-stress learning (125 promoters) set and non-stress tuning set (44 promoters).

In *O. sativa* 87 promoters were put into both stress and non-stress learning sets. 42 and 56 promoters were put into the stress and non-stress tuning sets. These genes were selected because they were shown to be induced by abiotic stress (cold, drought, and osmotic stress) by our own experiments.

Principle of evaluation

Yu et al. (2006) showed that certain experimentally verified motif pairs exhibit a characteristic motif distance between each other. This means that a biologically important motif pair will have a characteristic spacer length while randomly occurring motif pairs will not have any distinguished spacer length that would differ from the average distance. Therefore, if a head and tail motif occur very frequently at a specific distance from each other, we assume that there is

a biologically relevant function involved (van Helden et al. 2000; Cserháti 2006). While it is true that many motifs occur at quite flexible distances from each other, the connection between transcription factors binding these two motifs together at these distances (e.g., many thousands of bp) are very weak. 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, 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. In fact, a whole class of transcription factors, the leucine zippers, binds to sites on the DNA molecule which are separated from one another by a stretch of DNA of unspecified sequence. One such leucine zipper, EmBP-1 binds to the well known ABRE element (CACGTGGC) (Guiltinan et al. 1990). Others include bZIP proteins which regulate morphology in *Arabidopsis*.

Thus, the algorithm entails the enumeration of all possible n-mer motif pairs called dyads (in our case, pairs of pentamers) represented by the formula $M_H\{Ns\}M_T$, where M_H denotes the head motif pentamer, M_T the tail motif pentamer, and s denotes the spacer length between the head and tail motif pentamers in the dyad (Fig. 1). An exhaustive enumeration of very long sequence motifs is prohibitively expensive in terms of computer resources, so we restricted the number of motifs by enumerating dyads of pentamers separated by a maximum of 52 residues. As there are $4^5 = 1{,}024$ possible pentamers, the number of M_H–M_T pairs is $4^{5+5} = 1{,}048{,}576$ and the total possible number of dyad motifs is $53 \times 4^{5+5} = 55{,}574{,}528$ (added up over all spacer lengths from 0 to 52).

We constructed a cumulative measure designed to express the functionality of a given dyad. This is calculated from the comparison of two promoter datasets, the positive learning set representing promoters showing the desired biological function (in our case, abiotic stress responsiveness), and the negative learning set being a collection of promoters showing neither induction or repression to abiotic stress. A comparison measure is then calculated using the two datasets after which the dyads are ranked accordingly. The definition of this measure can be found in "Materials and methods".

Figure 2 shows an example of a biologically relevant and irrelevant dyad. The head motif of the dyad ACGTG{N *n*}TTTTT shown in Fig. 2a, b is a variant of the so-called ABRE element that occurs in promoters regulated by abiotic stress or abscisic acid in a number of monocot crops, and is represented by the core motif ACGTG (Hattori et al. 2002). The occurence of the dyad ACGTG{N *n*}TTTTT is greater in the learning stress promoter set (82 times, Fig. 2a) than in the non-stress learning set (61 times, Fig. 2b). This dyad occurs the most with a spacer length of 29. The specific dyad ACGT G{N29}TTTTT occurs in 11 stress learning promoters, while it occurs in none of the non-stress learning promoters. Therefore, its cdr score is $(11 - 0)/11 = 1.0$, which makes it a good candidate for being a stress dyad.

On the other hand, the dyad ATGAT{N n}TTTAT (Fig. 2c, d) which is not associated with stress-response elements occurs even less in the stress set (85 times, Fig. 2c) as in the non-stress set (107 times, Fig. 2d). This dyad occurs the most with a spacer length of eight. The specific dyad ATGAG{N8}TTTAT occurs in eight stress learning promoters, while it occurs in 11 of the non-stress learning promoters. Therefore, its cdr score is $(8–11)/8 = -0.375$, and is highly likely to be an irrelevant dyad. We note that our numerical measure is of experimental nature and we use them only for ranking the motifs.

Selection of top scoring pentamer dyads

We enumerated all possible dyads within the positive and negative promoter sets described above, and ranked them according to the numerical measure cdr (cumulative difference ratio), which is the ratio of the number of stress promoters minus the non-stress promoters to stress promoters that the dyad was found in. In order to get robust statistics, only pentamer dyads with a minimum occurrence of 5 in the stress learning promoter set were taken into consideration. In total, 995,304 dyads were present in both the stress promoter set and the non-stress promoter set in *Arabidopsis*. Out of these, 62,434 dyads occurred in at least five promoters in the learning set with a minimum cdr score of 0.6. In *O. sativa*, we found 21,639 dyads occuring in at least five promoters in the learning set with a minimum cdr score of 0.5.

Determination of the optimum set of dyads for promoterome search

We ran a series of ROC analyses on a smaller tuning set made up of 44 *Arabidopsis* stress promoters and 44 non-stress promoters in order to find the optimal dyad set. For this, we had to fine-tune the algorithm by finding the optimal set of parameters (minimum cdr score, minimal occurence in stress learning set, wobbling factor, and inclusion of the 37 TRANSFAC and PLACE motifs), which had the highest AUC value found during ROC analysis.

The tuning set is approximately one-third the size of the learning set (a ratio often used in machine learning algorithms), which was made of 125 promoters in *Arabidopsis*.

We selected those dyads with a minumum cdr score of 0.6–1.0 with 0.1 increments (5 parameter combinations) which occured each at least more than 5–20 times (16 parameter combinations) in the stress learning set. Furthermore, we allowed the spacer length to wobble from 0 to ±5 bp (6 parameter combinations), 10 bp being one full turn in the double helix of the DNA. Additionally we selected 37 oligomer motifs 5–9 bp long each from the PLACE and TRANSFAC databases and calculated a cdr score for these motifs by calculating the frequency occurence of each motif in the stress set as well as in the non-stress learning set (Higo et al. 1999; Matys et al. 2003). These motifs were used to study the algorithm's behavior with known stress motifs (thereby giving 2 more parameter combinations) and were selected based on their role in abiotic stress response. A list of these motifs can be seen in Online Resource 1, "TRANSFAC + PLACE motifs" worksheet. In a seperate screen we also took the presence of these motifs into account for each parameter combination.

We ranked the promoters in the *Arabidopsis* and *O. sativa* tuning promoter sets based on the score of each individual promoter and then calculated an AUC value for each parameter combination described above. A three-dimensional diagram of the AUC values according to the minimum number of dyads in the stress learning set and minimum cdr score can be seen in Supplemental Figures 1 and 2 for *Arabidopsis* and *O. sativa* showing the optimum AUC value. The highest AUC value calculated over all parameters from the tuning promoter set in *Arabidopsis* is 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). In this case, those 81 dyads (Table 1) had a minimum score of 0.9, a minimum number of occurences of 14 in the stress learning promoter set, and a spacer wobbling maximum of 2 bp.

We checked to see whether these dyads had anything to do with abiotic stress by matching them with known elements in the PLACE and PlantCARE databases (Higo et al. 1999; Lescot et al. 2002) (see "optimal dyad set" worksheet in Online Resource 1). For example, the dyad ATTGT{N2}TTAAA (rev. comp.) (part of the motif TTCTTCAAGCTTCAAGACAATCCTAGAAATTAC) responds to ABA, while the dyads AAAAA{N9}ACTGA (rev. comp.), AAAAA{N5}TCGAA (rev. comp.), AAAAA{N1}GACAA, AAAAA{N2}AGCAT (rev. comp.), AAAAA{N9}ACTAG (rev. comp.), and ATATG{N1}TTTTA (rev. comp.) all form part of a regulatory complex (the E4-ERE ethylene responsive element: CACAAGTTTGTTTTTGTTTTTACTACCAACAA) which responds to the production of ethylene. Indeed, some of the dyads found to take part in ethylene response formed a cluster of dyads as we shall see later on. ABA and ethylene are both phytohormones which are known to be produced with the onset of abiotic stress (Ludwig et al. 2005; Tuteja 2007). Furthermore, the dyad TTATA{N4}TGATT is known to be part of the OCS-element (ATCTTATGTCATTGATGACGACCTCC), which responds to oxidative stress (Zhang et al. 1995).

In *O. sativa* the maximum AUC value was 0.59069 with a corresponding *p* value of 0.0743 (calculated by MedCalc Version 11.3.6). This corresponds to a wobbling factor of 0, a minimum occurence of nine in the stress learning promoter set, and a minimum cdr score of 0.89. A list of the top 38 *O. sativa* dyads can be seen in Table 2 corresponding to the optimum *O. sativa* AUC parameter scoring scheme described previously. Out of these, eight elements were shown to take part in abiotic stress.

Searching for other stress genes in the *Arabidopsis* and *O. sativa* promoteromes

With the optimal parameters defined above, we did a search of the 31,128 promoters in the *Arabidopsis* promoterome with the top 81 dyads as well as with the 37 PLACE and TRANSFAC motifs. After scoring and ranking the promoters with these dyads and motifs we checked to see how many of the promoters from the original stress learning and tuning set were found back by the algorithm, as well as how many of the promoters found qualified as abiotic stress promoters. This we did by checking whether the expression level difference of the given gene during abiotic stress experiments changed at least twofold. These experimental results can be found in the Genevestigator *Arabidopsis* database (Zimmermann et al. 2004) as well as certain gene expression omnibus (GEO) datasets at NCBI, which studied abiotic stress response in *Arabidopsis* (specifically GEO sets GDS1620, GSE10670, GDS3216, GDS1382, and GSE5620-4).

We performed the promoterome search in *Arabidopsis* and *O. sativa* with the top 81/38 dyads which we found according to the optimum AUC parameters. As an independent measure for showing the accuracy of the algorithm, we counted how many promoters were found from the stress learning, non-stress learning, stress tuning, and non-stress tuning promoter sets from the promoterome search. This we did for the top 10,000 promoters in increments of 100. The number of promoters from each set can be seen in Figs. 4 and 6 for *Arabidopsis* and *O. sativa*, respectively. In Figs. 5 and 7 we can see the percentage of promoters from non-stress promoter sets to all promoters for *Arabidopsis* and *O. sativa*. In Arabidopsis, we can see that this percentage value declines between the top 1,600 and the top 3,100 promoters from 3.5 to 2.5%, corresponding to two false discoveries among 57 (3.5%) and 81 (2.5%) total promoters found back from both the stress and non-stress learning and tuning sets. Therefore, we used the

Table 1 List of top 81 dyads used in optimum AUC parameterization in *Arabidopsis*

Dyad sequence	Dyad score	Dyad sequence	Dyad score
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
AAAAA{N11}GAGTT	1	ATATA{N3}AGTTT	0.933333
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
ACTAA{N10}AGAAA	1	TTATA{N4}TGATT	0.933333
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
TTTTT{N15}CCTTG	1	AAAAA{N25}TGGGT	0.928571
TTTTC{N30}AAAAG	1	AAAAA{N5}TCGAA	0.928571
AACTA{N48}TTAAA	1	AAAAA{N50}CCTTG	0.928571
GCAAA{N41}AAATT	1	AAAAA{N1}GACAA	0.928571
GAAGA{N22}TTTTT	0.954545	AAAAA{N2}AGCAT	0.928571
AAAAA{N7}CGAAT	0.947368	ATATG{N14}TTTAT	0.928571
AAAAA{N37}TGTAC	0.947368	ATTGT{N2}TTAAA	0.928571
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
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	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	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		

top 3,100 promoters from the promoterome search for further analysis in *Arabidopsis*. Overall, 1,542 of the 3,100 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 tuning 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%. A list of promoters found by the promoterome search, their annotation and score can be found in the "promoterome search" worksheet in Online Resource 1.

In *O. sativa*, this ratio rises to one small and two larger peaks from where they drop off. The second larger peak

Table 2 List of top 38 dyads used in optimum AUC parameterization in *Orzya sativa*

Dyad sequence	Dyad score
AAAAG{N32}TTTTG	1
AAAGA{N0}AAATT	1
AAATA{N32}AAATG	1
AGAAG{N40}TTTAT	1
AGAGA{N15}ACTTT	1
AGGGA{N3}GGGAG	1
ATTTG{N43}TTTAT	1
CATTT{N42}TATAT	1
CTCCT{N7}TTCTT	1
GAGAA{N40}AATAT	1
GAGGC{N39}GAGGA	1
GAGTA{N11}ACACA	1
GATTT{N15}AATTA	1
GCGCC{N13}CCGCG	1
GTTTA{N38}TTTGA	1
GTTTG{N35}TTTAT	1
TAAAA{N21}CACTT	1
TAATT{N47}CAAAA	1
TCAAA{N51}ATTTT	1
TCTTT{N5}TATTT	1
TTATT{N29}TATAC	1
TTCTA{N11}AAAAT	1
TTCTT{N13}ATTCA	1
TTTAA{N50}GATTT	1
TTTCA{N30}TTTGA	1
TTTGA{N22}ACATG	1
TGTTT{N35}TATTT	0.9167
GAGAG{N5}GAAAA	0.9091
TTCTT{N7}TAAAA	0.9091
AATTC{N19}AAATT	0.9
AGAGA{N6}AAAGA	0.9
CAATT{N24}AATTT	0.9
GTTTT{N51}ATATA	0.9
TAAAA{N43}TGAAA	0.9
TATGT{N24}TTTTG	0.9
TCAAA{N20}CAAAA	0.9
TCATT{N30}TAAAT	0.9
TTTGA{N38}ACTAA	0.9

drops off to a percentage value of 7%, corresponding to four false discoveries among 57 and 81 total promoters found back from both the stress and non-stress learning and tuning sets found within the the top 4,600 *O. sativa* promoters from the promoterome search. In order to calculate a success rate for the algorithm in *O. sativa* we used GEO datasets from NCBI which contained data on abiotic stress experiments performed in *O. sativa*, namely datasets GSE

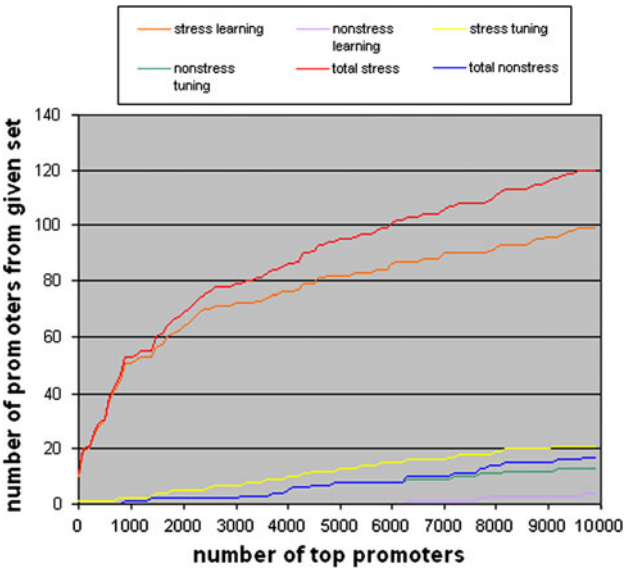

Fig. 4 We studied the number of promoters from the stress learning, non-stress learning, stress tuning, and non-stress tuning promoter sets in the top 10,000 scoring promoters found by our promoterome search in *Arabidopsis* in increments of 100. As we can see, the number of stress learning and stress tuning promoters outnumbered the non-stress promoters by far

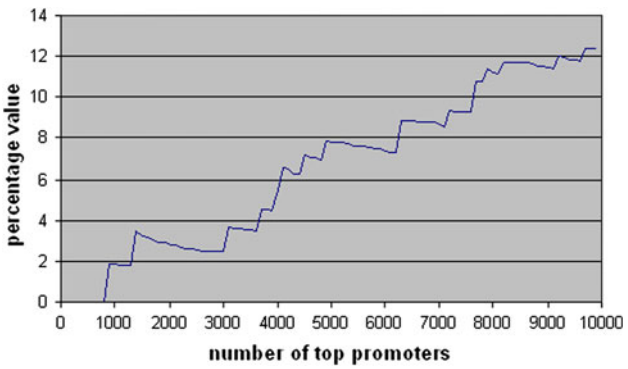

Fig. 5 The percentage ratio of all non-stress promoters to all promoters found in the *Arabidopsis* promoterome search is shown here. We can see that the percentage of non-stress promoters steadily decreases in the top 1,600–3,100 promoters. The percentage of non-stress promoters is 2.5% in the top 3,100 promoters

3053, GSE 4438, and GSE 6901. These datasets contained expression level changes for *O. sativa* genes due to cold, salt, and drought stress. Out of 4,600 genes, 3,144 had an Affymetrix probe id, based on which we could check their expression data in the GEO datasets. 3,102 genes showed a minimum twofold expression level increase according to these datasets. This corresponds to a 98.66% positive prediction rate of finding new abiotic stress genes using our algorithm in *O. sativa*. Parallel to this, we selected a random set of 4,600 *O. sativa* genes, from which 3975 had an Affymetrix probe id. Out of these, 1,243 were shown to be stress-induced which is only 31.3% of the total.

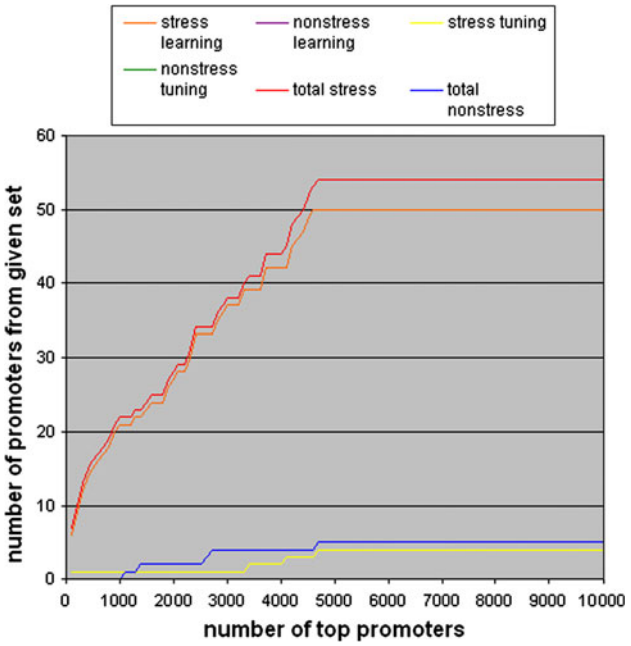

Fig. 6 We studied the number of promoters from the stress learning, non-stress learning, stress tuning, and non-stress tuning promoter sets in the top 10,000 scoring promoters found by our promoterome search in *Oryza sativa* in increments of 100. Not a single non-stress learning promoter was found in the search, therefore the number of non-stress tuning promoters is the same as all non-stress promoters, and thereby these two curves overlap

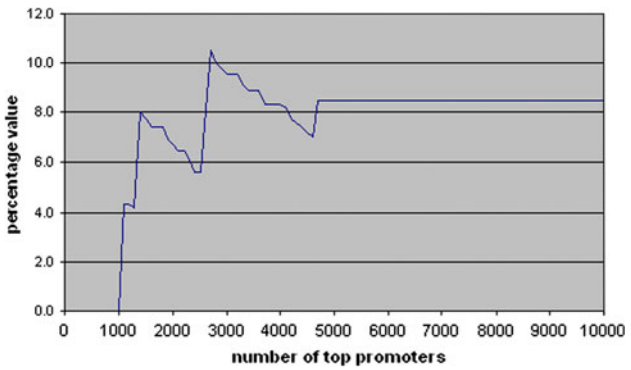

Fig. 7 The percentage ratio of all non-stress promoters to all promoters found in the *Oryza sativa* promoterome search is shown here. One smaller peak and two larger peaks are visible. Based on this information the top 4,600 promoters were examined in further detail, where this ratio was 7%

One of the factors that must be taken into account in the application of the algorithm is selection of the input promoter set. This is because the promoterome searches retrieved only a smaller part of the stress tuning set promoters, even though they were able to find many of the stress learning promoters. In both *Arabidopsis* and *O. sativa* the tuning promoter sets were switched with the learning promoter sets to see whether they were able to retrieve a large number of the same promoters (using the same optimum parameterization as deduced in their respective individual test runs).

In *Arabidopsis* the top 3,100 and 2,400 promoters were examined for both runs. Overall, 568 promoters were found by both the normal run and the switched run (18.3 and 23.7%, respectively). In *O. sativa*, the top 4,600 and top 1,200 promoters of the original and reversed promoterome search were compared and only 226 were found by both promoterome searches (4.9 and 18.8%, respectively). Out of 4,600 *O. sativa* genes found by the algorithm in *O. sativa*, 1,456 did not have an affymetrix identifier; therefore these genes can be counted as new putative stress genes. 534 of these genes were either hypothetical or expressed genes, which are listed in the "new unknown genes" worksheet in Online Resource 2. These genes are good candidates to be analyzed in further experiments.

Comparison of the described algorithm with yeast motif finder and dyad-analysis

We compared the present algorithm with two other well-known motif finding algorithms, namely yeast motif finder (YMF) of Sinha and Tompa, and dyad-analysis of Jacques van Helden (2000; Sinha and Tompa 2003) on the set of 125 *Arabidopsis* stress learning promoters. For the YMF program we looked for pairs of pentamers without any mismatches, with a spacer length of at least 0 bp to a maximum of 52 bp. With this method we were able to find 283 promoters with more than 1 of the significant dyad motifs found by the YMF program. Of these only three belonged to the original promoter set of original 125 stress promoters (1.1%). Of these 283 promoters, 195 could be found in the Genevestigator database, of which only six were found to be stress-inducible (3.1%), which is very low. For the dyad-analysis program we were able to search pure pentamer dyads with spacers 0–52 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 four were stress-inducible (3.6%). Therefore, we can say that our algorithm is much more effective in finding new members of a set of co-regulated genes than these two programs.

GO term analysis

We were able to retrieve Gene Ontology (MSU GOSlim) terms regarding biological functions for 232 of the top *O. sativa* genes found in the promoterome search from the top 4,600 at the Rice Array Database (Jung et al. 2008). This data can be seen in the "GO biol. functions" worksheet in Online Resource 2. It is interesting to note that 87

(37.5%) of the genes had GO terms connected to response to stress (which is the most common gene function), corresponding to a p value of 0.0358, and 40 (17.2%) of them had GO terms connected to abiotic stimuli (fourth in the list) (p value: 0.023). From this GO term analysis we can then draw the conclusion that these 232 gene selection was enriched with genes which take part in abiotic stress. Furthermore, 93 of the genes were stimulated by either stress or abiotic stimuli, meaning corresponding to 40% of the genes. 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 signal transduction ($p = 0.0242$), metabolic processes ($p = 0.0483$), biosynthetic processes ($p = 0.0478$), cellular processes ($p = 0.0039$), protein metabolic processes ($p = 0.0081$), and cellular homeostasis ($p = 0.0155$).

GO terms were also retrieved for molecular functions for 329 of the top 4,600 genes. A number of other molecular functions are listed found in the annotaion of these genes not found among the GO terms. A list of terms and the number of genes they occur can be seen in the "GO mol. functions" worksheet in Online Resource 2. Significantly enriched GO terms include the following: nucleotide binding ($p = 0.0468$), RNA binding ($p = 0.0466$), catalytic activity ($p = 0.0043$), receptor activity ($p = 0.0184$), structural molecule activity ($p = 0.0408$), binding ($p = 0.0393$), protein binding ($p = 0.0457$), transferase activity ($p = 0.0269$), hydrolase activity ($p = 0.0168$), oxygen binding ($p = 0.0305$), carbohydrate binding ($p = 0.309$).

All p values were calculated by GO Enrichment Analysis provided at the Rice Array Database (Jung et al. 2008). This employs a conditional hypergeometrical test to calculate the significance of the p values.

Regulatory dyad network analysis

Cluster analysis of top 81 dyads and scoring of REP's in Arabidopsis

In our further analysis we studied how frequently the top 81 dyads occurred alongside the 37 PLACE and TRANSFAC motifs. The main concept behind the regulatory network analysis was to study pairs of dyads, or "dyad dyads", in order to get a more global view of how dyads interact with each other. Thus, we analyzed the distribution of REP's in the top 3,100 promoters found by our algorithm. Before we did this, we ran a cluster analysis on our dyads to decrease the level of redundancy that might occur between them (see "Materials and methods"). Overall, we were able to assign 38 of the 81 dyads to a dyad cluster (Table 3). 11 clusters were formed in total, and the size of the clusters included from 2 to 7 dyads.

We then calculated the frequency of two given REs occurring together. Here a RE was taken to be either a singleton dyad from the top 81 dyads found by the algorithm in *Arabidopsis* when analyzing the learning set, a dyad belonging to one of the 11 dyad clusters, or one of the 37 PLACE or TRANSFAC motifs. Two different dyads belonging to the same cluster also may be counted as a REP. For each possible REP, we counted how many times they occur within 100 bp from each other in stress-induced promoters from the top 3,100 ranking promoters taken from the promoterome search, as well as how many non-stress-induced promoters from the the top 3100 promoters they occur in. We assigned each REP a cdr score as described in the "Materials and methods". All of this information is available in the worksheet "REP analysis" in Online Resource 1. Overall there were 1224 such REP's with a minimum cdr score of 0.5 (this was the lowest cdr cutoff score used in Arabidopsis during the test phase), which were used in the RE and promoterome analysis of *cor* end *erd* genes described in the following two sections.

RE networks

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*arly *d*ehydration). A list of the selected genes and their annotation can be seen in Table 4. 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 contain similar REs.

Overall 22 PLACE/TRANSFAC motifs play key roles in both networks (connected to at least 10 other REs): tfpl_1, 3, 4, 5, 7, 9, 10, 11, 14, 15, 17, 18, 21, 22, 23, 24, 26, 27, 30, 31, 32, and 34. Their sequences can be found in the worksheet "TRANSFAC + PLACE motifs" in Online Resource 2. Amongst these, tfpl_3, 4, 5 are 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 et al. 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 et al. 2005). The motifs tfpl_9, 10, 18, 21, 24, 27, and 31 correspond to a MYC binding site (Abe et al. 1997).

Four of our dyad clusters were also found to play key roles in the regulatory networks for both the *cor* and *erd* genes, clusters 2, 3, 5, and 11, being connected to at least ten other elements in the network. A list of the dyad clusters may be seen in Table 3. The dyads CAAGT {N49}TTTTT, TATGA{N11}TTTTT, and TTTTT{N15} CCTTG from cluster 2 contain the motif TTTTT, which corresponds to the MYB binding site (Wang et al. 1997).

Table 3 List of dyads belonging to clusters from the top 81 dyads in *Arabidopsis*

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 4 List of *cor* and *erd* genes selected for regulatory element network analysis in *Arabidopsis*

Gene id	TAIR description
Cor genes	
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; involved in: biological_process unknown
At4g15430	functions in: molecular_function unknown; involved in: biological_process unknown

In cluster 3, the tail motif of the dyad TATAT{N28}GTTGA corresponds to the motif CAACTC, which takes part in gibberellin upregulation (Sutoh and Yamauchi 2003). The head motif of the dyad ATATG{N1}TTTTA corresponds to a MYC binding site (Abe et al. 2003). 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).

Promoterome analysis to find other stress promoters with similar REP content

The *Arabidopsis* promoterome was screened to see which other promoters contained REPs common to the five *cor* promoters found by the algorithm. This was done by listing all REPs from the top 3,100 working set and checking how many of them were in common with the five *cor* promoters and all other *Arabidopsis* promoters. We calculated the distance between the five *cor* promoters and all other promoters (as described in the "Materials and methods"). We selected those 25 new promoters (seen in Table 5) whose minimum distance (REP content) from any of the five *cor* promoters were below 0.5. Amongst these genes a hypothetical gene were discovered (At4g06530), and five new unknown genes (At1g50040, At2g41120, At3g10980, At4g36510, and At5g41505). Therefore, we assume that these new genes undergo similar regulation as the *cor* genes.

Table 5 List of top 25 new *Arabidopsis* gene with lowest distance (below 0.5) to all of our selected set of *cor* genes

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

Dyadscan website

A website has been constructed (http://www.bhd.szbk.u-szeged.hu/dyadscan/, login information: id: totto, password: pwd1), where a user may download a stand-alone version of the program (Dyadscan.exe) along with a short help file (README.txt). Here the user may give a number of options such as motif length, minimum dyad occurence in the input promoter sets, and minimum cdr score besides supplying input positive and negative promoter sets. The user can also upload two sequence sets, one being the positive set, the other the negative set. The result of the algorithm is a list of dyad sequences with occurences in the input sequence sets as well as a cdr value. These results can then be used in further analysis.

Discussion

In this work we applied a motif prediction algorithm for finding putative RE dyads in a set of co-regulated promoters. The algorithm was tuned to find pairs of pentamers, which we assumed to represent core motifs of individual TFBS's acting in concert with each other. While we measured the distribution and statistical significance of only pentamer dyads in this study, the algorithm can be adjusted to study dyad motifs of different lengths (tetramers, hexamers). Being an enumeration method, the algorithm predicts motifs with higher precision, since these kinds of algorithms perform a complete statistical analysis on all possible motifs (Rombauts et al. 2003). Furthermore, the present algorithm describes dyads with a unique head and tail motif and a spacer in between with a preferential length rather than a simple oligonucleotide sequence. It also found additional members of co-regulated gene sets with a far higher success rate than dyad-analysis, another well-known motif prediction program.

One of our main findings is that 534 genes were found in the promoterome search in *O. sativa* which did not have an Affymetrix identifier. Still, since the positive predictive rate in *O. sativa* was 98.66%, we can predict that from these genes more than 500 genes are possible candidates in

abiotic stress in *O. sativa*. Furthermore, 1,245 genes were predicted to be induced by abiotic stress in *O. sativa*. These genes are not yet annotated, therefore our predictions may aid the functional annotation process for these genes. Overall more than 1,700 new genes can be designated as candidate abiotic stress-induced genes in *O. sativa* according to our algorithm, which is a substantial part of the *O. sativa* genome. In *Arabidopsis* 1,558 genes from the top 3,100 had no Affymetrix ids, therefore, since 78.6% of these genes were predicted to play a role in abiotic stress, more than 1,200 new genes can be predicted to be stress-induced. Furthermore, 48 genes were annotated as hypothetical proteins, and 1 as an expressed protein. Therefore, in *Arabidopsis* more than 1,200 new genes can be designated as stress genes.

The algorithm was tested on a set of genes involved in abiotic stress response. However, the scope of such studies can be altered to involve different sets of genes which regulate different types of physiological processes or biochemical reactions, such as regulation of the cell cycle, development, biotic stress, or seed maturation, because of the basic idea of analyzing common regulation machinery mirrored in TFBS content. In fact, in the case of *O. sativa*, a number of pathways overlap with each other such as abiotic stress and seed maturation (Cooper et al. 2003), therefore making it possible to compare TFBS's found in one gene set with those found in another.

While it is true that the highest AUC value for *O. sativa* was rather low and its corresponding p value somewhat high, other statistics support the robustness of the algorithm's application in this species. Also, the positive prediction rate for finding stress promoters in the *O. sativa* genome was very high (98.66%). Furthermore, the statistical significance of GO terms for stress processes were also high for genes found in the promoterome search. Since the end goal of the application of the algorithm was to find new stress gene candidates, we should be encouraged by the positive, significant findings following the application of our algorithm.

To test the robustness of the algorithm we ran two tests in both species where we replaced the stress learning promoter set with a set of randomly selected promoters using the same set of parameters which were deduced in the test phase of the algorithm. In *Arabidopsis*, two and three dyads were found compared to 81 found by the original run, and in *O. sativa* six and two dyads were found compared to 38 in the original run. None of these dyads matched any known abiotic stress motifs.

Furthermore, as mentioned in the subsection "Regulatory element networks", 22 TRANSFAC/PLACE motifs were found to be part of REP's 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 five 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 five motif variants were not present in the initial stress learning promoter set in *Arabidopsis*.

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. Our method cannot give exhaustive results. The use of learning and tuning sets can provide differentiating dyads but will not give all of the possible ones. 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 ≤ 3, or either their full head or tail motif matched the full head or tail motif of one of the 81 dyads.

As a separate 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. For example, Alonso-Blanco et al. (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 et al. 2003).

The algorithm was tested in two different kinds of plant species, *Arabidopsis*, a dicot, and *O. sativa*, a monocot. The results we got during the analysis of these two species can give us insight into how to apply the algorithm in the case of other species. One such factor is the size and structure of the genome of a given species, and the presence of repetitive elements and transposons, which are also present in promoters. *Arabidopsis* has a compact genome of 125 Mb, and whose repetitive element content is only around 10%. However, this was a major problem during the analysis of the *O. sativa* genome, as some estimates state that up to 35% or more of the 430 Mbp *O. sativa* genome is made up of repetitive elements. Hence we had to compare promoter sequences with each other in the input

sets and filter out some promoters which were too similar to other sequences. This problem will be even more pronounced if, e.g., the 17,000 Mbp *Triticum aestivum* genome is to be analyzed, which contains a high proportion of repetitive elements.

Although the algorithm is successful in finding new promoters, it is sensitive as to what kinds of promoters are used for dyad definition. Overall, along with the downloadable stand-alone program this algorithm may be useful in the future for finding new, putative regulatory sequences as well as new genes which play similar roles to already studied and annotated genes based on their RE content.

Acknowledgments The first author would hereby like to mention that the underlying idea of the analysis of the occurence distributions of putative transcription factor binding site dyads originated from his college thesis work which was done at the Agricultural Biotechnology Center in Gödöllő, Hungary. This work was funded by the OTKA T046495 grant given by the Hungarian National Science foundation and the Bio-140-KPI given by the National Office for Research and Technology (NKTH) as well as grant number 4-065-2004. We would also like to thank William Gruissem for supplying us data from the Genevestigator *Arabidopsis* database. The authors would like to thank Maria Sečenji and Krisztina Talpas for their assistance in setting up and helping in greenhouse experiments and QT-PCR work.

References

Abe H, Yamaguchi-Shinozaki K, Urao T, Iwasaki T, Hosokawa D, Shinozaki K (1997) Role of arabidopsis MYC and MYB homologs in drought- and abscisic acid-regulated gene expression. Plant Cell 10:1859–1868

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

Alonso-Blanco C, Gomez-Mena C, Llorente F, Koornneef M, Salinas J (2005) 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 139(3):1304–1312

Cooper B, Clarke JD, Budworth P, Kreps J, Hutchison D, Park S, Guimil S, Dunn M, Luginbühl P, Ellero C, Goff SA, Glazebrook J (2003) A network of rice genes associated with stress response and seed development. Proc Natl Acad Sci USA 100(8):4945–4950

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

European Plant Science Organization (2005) European plant science a field of opportunities. J Exp Bot 56(417):1699–1709

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

Guiltinan MJ, Marcotte WR Jr, Quatrano RS (1990) A plant leucine zipper protein that recognizes an abscisic acid response element. Science 250(4978):267–271

Hattori T, Totsuka M, Hobo T, Kagaya Y, Yamamoto-Toyoda A (2002) Experimentally determined sequence requirement of ACGT-containing abscisic acid response element. Plant Cell Physiol 43(1):136–140

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

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 (2008) Refinement of light-responsive transcript lists using rice oligonucleotide arrays: evaluation of gene-redundancy. PLoS One 3(10):e3337

Lee Y, Tsai J, Sunkara S, Karamycheva S, Pertea G, Sultana R, Antonescu V, Chan A, Cheung F, Quackenbush J (2005) The TIGR Gene Indices clustering and assembling EST and known genes and integration with eukaryotic genomes. Nucleic Acids Res, 33:D71–D74 (Database issue)

Lescot M, Dehais P, Thijs G, Marchal K, Moreau Y, Van de Peer Y, Rouze R, Rombauts S (2002) PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res 30(1):325–327

Ludwig AA, Saitoh H, Felix G, Freymark G, Miersch O, Wasternack C, Boller T, Jones JD, Romeis T (2005) Ethylene-mediated cross-talk between calcium-dependent protein kinase and MAPK signaling controls stress responses in plants. Proc Natl Acad Sci USA 102(30):10736–10741

Mahajan S, Tuteja N (2005) Cold, salinity and drought stresses: an overview. Arch Biochem Biophys 444(2):139–158

Matys V, Fricke E, Geffers R, Gössling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Münch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E (2003) TRANSFAC: transcription regulation, from patterns to profiles. Nucleic Acids Res 31(1):374–378

Picot E, Krusche P, Tiskin A, Carré I, Ott S (2010) Evolutionary Analysis of Regulatory Sequences (EARS) in Plants. Plant J, doi: 10.1111/j.1365-313X.2010.04314.x

Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R, White J (2001) The TIGR Gene Indices analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids Res 29(1):159–164

Rombauts SK, Florquin K, Lescot M, Marchal K, Rouzé P, van de Peer Y (2003) Computational approaches to identify promoters and cis-regulatory elements in plant genomes. Plant Physiol 132(3):1162–1176

Sandve GK, Drabløs F (2006) A survey of motif discovery methods in an integrated framework. Biol Direct 1:11

Shinozaki K, Yamaguchi-Shinozaki K, Seki M (2003) Regulatory network of gene expression in the drought and cold stress responses. Curr Opin Plant Biol 6(5):410–417

Simpson SD, Nakashima K, Narusaka Y, Seki M, Shinozaki K, Yamaguchi-Shinozaki K (2003) 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 33(2):259–270

Sinha S, Tompa M (2003) YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. Nucleic Acids Res 31(13):3586–3588

Solovyev VV, Shahmuradov IA, Salamov AA (2010) Identification of plant promoters and regulatory sites. Methods Mol Biol 674:57–83

Sutoh K, Yamauchi D (2003) Two cis-acting elements necessary and sufficient for gibberellin-upregulated proteinase expression in rice seeds. Plant J 34(5):635–645

Tuteja N (2007) Abscisic acid and abiotic stress signaling. Plant Signal Behav 2(3):135–138

van Helden J, Rios AF, Collado-Vides J (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. Nucleic Acids Res 28(8):1808–1818

Vardhanabhuti S, Wang J, Hannenhalli S (2007) Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation. Nucleic Acids Res 35(10):3203–3213

Walther D, Brunnemann R, Selbig J (2007) The regulatory code for transcriptional response diversity and its relation to genome structural properties in *A. thaliana*. PLoS Genet 3(2):216–229

Wang Z-Y, Kenigsbuch D, Sun L, Harel E, Ong MS, Tobin EM (1997) A myb-related transcription factor is involved in the phytochrome regulation of an Arabidopsis Lhcb gene. Plant cell 9:491–507

Wang S, Yang S, Yin Y, Guo X, Wang S, Hao D (2008) An in silico strategy identified the target gene candidates regulated by dehydration responsive element binding proteins (DREBs) in Arabidopsis genome. Plant Mol Biol 69(1–2):167–178

Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA (2003) The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol 20(9):1377–1419

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

Yu X, Lin J, Masuda T, Esumi N, Zack DJ, Qian J (2006) Genome-wide prediction and characterization of interactions between transcription factors in *Saccharomyces cerevisiae*. Nucleic Acids Res 34(3):917–927

Zhang B, Chen W, Foley RC, Büttner M, Singh KB (1995) Interactions between distinct types of DNA binding proteins enhance binding to ocs element promoter sequences. Plant Cell 7(12):2241–2252

Zhang W, Ruan J, Ho TH, You Y, Yu T, Quatrano RS (2005) Cis-regulatory element based targeted gene finding: genome-wide identification of abscisic acid- and abiotic stress-responsive genes in *Arabidopsis thaliana*. Bioinformatics 21(14):3074–3081

Zimmermann P, Hirsch-Hoffmann M, Hennig L, Gruissem W (2004) GENEVESTIGATOR Arabidopsis microarray database and analysis toolbox. Plant Physiol 136(1):2621–2632