

Genome analysis

Cis-regulatory element based targeted gene finding: genome-wide identification of abscisic acid- and abiotic stress-responsive genes in *Arabidopsis thaliana*Weixiong Zhang^{1,2,*}, Jianhua Ruan¹, Tuan-hua David Ho³, Youngsook You³, Taotao Yu¹ and Ralph S. Quatrano³¹Department of Computer Science and Engineering, ²Department of Genetics and ³Department of Biology, Washington University in Saint Louis, Saint Louis, MO 63130, USA

Received on January 20, 2005; revised on May 5, 2005; accepted on May 6, 2005

Advance Access publication May 12, 2005

ABSTRACT

Motivation: A fundamental problem of computational genomics is identifying the genes that respond to certain endogenous cues and environmental stimuli. This problem can be referred to as targeted gene finding. Since gene regulation is mainly determined by the binding of transcription factors and *cis*-regulatory DNA sequences, most existing gene annotation methods, which exploit the conservation of open reading frames, are not effective in finding target genes.

Results: A viable approach to targeted gene finding is to exploit the *cis*-regulatory elements that are known to be responsible for the transcription of target genes. Given such *cis*-elements, putative target genes whose promoters contain the elements can be identified. As a case study, we apply the above approach to predict the genes in model plant *Arabidopsis thaliana* which are inducible by a phytohormone, abscisic acid (ABA), and abiotic stress, such as drought, cold and salinity. We first construct and analyze two ABA specific *cis*-elements, ABA-responsive element (ABRE) and its coupling element (CE), in *A. thaliana*, based on their conservation in rice and other cereal plants. We then use the ABRE–CE module to identify putative ABA-responsive genes in *A. thaliana*. Based on RT–PCR verification and the results from literature, this method has an accuracy rate of 67.5% for the top 40 predictions. The *cis*-element based targeted gene finding approach is expected to be widely applicable since a large number of *cis*-elements in many species are available.

Contact: zhang@cse.wustl.edu

Supplementary information: Supplementary data for this paper are available at *Bioinformatics* online.

1 INTRODUCTION**1.1 Targeted gene finding**

It is fundamentally important, yet difficult, to identify the genes that respond to certain endogenous cues and/or environmental stimuli. For example, it is of great importance to find genes in plants that are responsive to abiotic stress to enhance the genomic makeup of plants to combat harsh stress, such as drought and low temperature.

We call the genes of particular interest target genes and the problem of identifying target genes targeted gene finding.

One possible approach to targeted gene finding is to use the knowledge of experimentally verified target genes in closely related species and utilize gene conservation across different species to identify putative target genes [see Zhang (2002) for a review]. By focusing on the conservation of open reading frames (ORFs) of genes, this strategy has been used to annotate many genomes (The *Arabidopsis* Genome Initiative, 2000; Lander *et al.*, 2001; Yu *et al.*, 2002). Despite its great success, this method is able to discover only a small number of genes of particular functions, partly because the number of experimentally determined genes is limited. As a result, a large portion of the predicted genes of many species does not have any functional annotation at all. For example, about half of the genes of plant *Arabidopsis thaliana* currently do not have any definitive functional annotation.

Furthermore, an ORF-centric gene finding method may not be effective in discovering target genes that express under specific conditions. Although gene functions may be indicative of a gene's responsiveness to certain stimuli, there is no direct correlation between gene function and gene expression. Gene expression is controlled mainly at the transcription level, where the binding between transcription factors (TFs) and *cis*-regulatory DNA sequences (or *cis*-elements) in the upstream regions of genes plays an important role (Brivanlou and Darnell, 2002). In other words, a gene's responsiveness to certain conditions is 'hard-wired' by their *cis*-elements. Therefore, if some *cis*-elements are known to be directly involved in gene transcription regulation in responding to specific stimuli, we should be able to use the *cis*-elements to identify the genes of interest. When combined with experimental verification, this constitutes an effective approach to genome-wide targeted gene finding and function annotation. This approach is supported by the fact that a large number of TFs and their binding *cis*-elements have been identified over the years. For example, we know most of the TFs and their corresponding *cis*-binding elements in the yeast *Saccharomyces cerevisiae* (Harbison *et al.*, 2004). Moreover, TRANSFAC, a database of experimentally verified and computationally predicted *cis*-elements in many species has been established and has been widely used for many years (Matys *et al.*, 2003). There are also databases of plant-specific *cis*-elements, including PLACE (Higo *et al.*,

*To whom correspondence should be addressed.

1999) and PlantCARE (Lescot *et al.*, 2002). Furthermore, there is abundant information on *cis*-elements in literature that has not been achieved as yet.

In this paper, we investigate the targeted gene finding approach and demonstrate its validity and efficacy by identifying the genes inducible by abscisic acid (ABA) and abiotic stress in *A.thaliana*.

1.2 ABA in plants and abiotic stress

ABA is an important phytohormone that is prevalent in a plant's developmental stages. It plays many key roles in the synthesis of seed storage, the promotion of seed desiccation tolerance and dormancy, and the inhibition of the phase transitions from embryonic to germinative growth and from vegetative to reproductive growth (Finkelstein *et al.*, 2002). Many genes have been identified as ABA-responsive; these include the genes encoding seed storage proteins, late embryogenesis abundant (LEA) proteins, and various other proteins and protein families. Examples of these genes are *EM* in wheat (Marcotte *et al.*, 1989), *EM* and *RAB16* in rice (Hattori *et al.*, 1995; Mundy *et al.*, 1990), *HVA1*, *HVA22* and dehydrins in barley (Shen and Ho, 1997; Shen *et al.*, 1996; Xu *et al.*, 1996), and *EM* and *RD29* in *A.thaliana* (Carles *et al.*, 2002; Narusaka *et al.*, 2003).

Most ABA-responsive genes have two eminent characteristics. First, they contain conserved ABA-responsive elements (ABREs) in their promoters (Hattori *et al.*, 1995, 2002; Marcotte *et al.*, 1989; Mundy *et al.*, 1990; Shen and Ho, 1997; Shen *et al.*, 1996; Xu *et al.*, 1996). ABREs are the binding sites of TFs, such as EmBP-1 (Guiltingan *et al.*, 1990), TAF-1 (Oeda *et al.*, 1991) and ABFs (Choi *et al.*, 2000). Second, the ABREs need to be accompanied by some coupling elements (CEs) in order to be functional (Shen and Ho, 1997; Shen *et al.*, 1996; Xu *et al.*, 1996). The sequence specificity of CEs may be lower than that of ABREs. Moreover, other functional elements, such as ABREs themselves and dehydration-responsive element (DRE), can also function as CEs (Guiltingan *et al.*, 1990; Hobo *et al.*, 1999; Narusaka *et al.*, 2003).

Importantly, ABA mediates many aspects of physiological responses to environmental stress, such as drought, cold and salinity. Many experiments have shown that abiotic stress also activate the processes underlying ABA (Finkelstein *et al.*, 2002; Shinozaki and Yamaguchi-Shinozaki, 2000; Zhu, 2002). Specifically, a large number of genes that respond to abiotic stress are also inducible directly by ABA treatment (Seki *et al.*, 2002a,b), providing direct evidences that ABA must be involved in the processes responding to these environmental stress.

In brief, we have gained a substantial amount of knowledge of the critical roles that ABA plays in plant development and stress response. It is clear that ABA is an essential element in gene transcription regulation in responding to abiotic stress.

In this paper, we study the *cis*-element based targeted gene finding method in the context of transcription regulation under ABA and abiotic stress in plants. We are particularly interested in the effectiveness of this method in finding, on a genomic scale, the genes that are inducible by ABA and abiotic stress in *A.thaliana*.

2 MATERIALS AND METHODS

2.1 Plant material, RNA preparation and RT-PCR

For plant material, *A.thaliana* ecotype Columbia seeds were grown at 24°C for 10 days with 16/8 h light/dark period on Murashige and Skoog modified basal medium with Gamborg vitamins (PhytoTechnology Laboratories,

Prod. No. M404). About 20 seedlings were transferred to each MS medium plate with or without 100 μM ABA. After 24 h RNA was extracted from the two groups using TRIzolR reagent (Invitrogen, Cat. No. 15596-026) and further purified by using RNA clean-up columns from RNeasyR Plant Mini kit (Qiagen, Cat. No. 74904). The total RNA was then treated with DNase I (Invitrogen, Cat. No. 18068-015).

RT-PCR analysis was done as follows. First-strand cDNA was synthesized from 1.5 μg of total RNA using ThermoScript™ RNase H-reverse transcriptase (Invitrogen, Cat. No. 12236-014) with Oligo(dT)₁₂₋₁₈ primer following the manufacturer's recommendation. Amplification of the cDNA was optimized using 0.5–2 μl of the cDNA in a total of 25 μl reaction volume and carried out at 94°C for 2 min, 30 cycles of 94°C for 1 min, 60°C for 1 min and 72°C for 1 min, and then 72°C for 5 min. Expression analysis of each gene was confirmed in at least three independent RT-reactions using forward and reverse primers, which are listed in Supplementary Table 1.

2.2 Genomic sequences

Predicting promoters is at least as difficult as predicting genes. The key is the identification of transcription start sites (TSSs). To predict TSSs, we combined an *A.thaliana* cDNA database and a software, TSSP (SoftBerry, <http://www.softberry.com>). As of January, 2004, there were 26 213 predicted *A.thaliana* genes (excluding pseudogenes and RNA genes) in GenBank, among which 12 359 (47%) had cDNA sequences (with annotated 5'-UTR >50 bp). For each gene, we retrieved a segment from the gene's start codon to 1500 bases upstream or its farthest 5' annotated gene boundary. We then applied the TSSP software to each upstream sequence to identify TSS. When multiple TSSs were predicted on a gene, the one closest to the ORF was chosen. The TSSs that were >50 bases downstream of the cDNA start point were considered false positive. Overall, TSSs of 18 343 (70%) genes were predicted and used in further analysis and prediction, whereas 3133 (12%) genes with cDNA and 4737 (18%) genes without cDNA did not have TSSs predicted. Given a TSS, we retrieved the sequence from 350 bp upstream to 50 bp downstream of the TSS as the proximal promoter. The intron and exon regions were retrieved from The *Arabidopsis* Information Resource (TAIR) at <http://www.arabidopsis.org>

2.3 Microarray gene expression data

Data on gene transcription profiling were from Hoth *et al.* (2002), Kreps *et al.* (2002) and Seki *et al.* (2002a,b). We took the datasets in Seki *et al.* (2002a,b) for motif analysis. Using the promoter prediction method mentioned above, we obtained 366 promoters of unique genes upregulated after ABA-treatment and/or under abiotic stress (drought, cold and salinity) for motif analysis.

2.4 Scoring

In order to search for instances of a known degenerate motif in promoters, we scored each short sequences of length w in a given sequence based on the motif and a background model. The degenerate motif W of length w is represented by a PWM $\Theta_W = (q_{i,b})$, where $q_{i,b}$ is the probability of finding base b at position i in the motif. The background model B_m is a m -th order Markov model, which can be estimated from background sequences. The probability that a w mer starting from the j -th position of a sequence is generated from the background model was calculated as $P(j|B_m) = \prod_{l=1}^w P(b_{j+l-1}|b_{j+l-2} \cdots b_{j+l-m-1})$, where b_j is the j -th base of the sequence. The conditional probability $P(b_{j+l-1}|b_{j+l-2} \cdots b_{j+l-m-1})$ is the frequency of observing base b_{j+l-1} following a particular m mer $b_{j+l-2} \cdots b_{j+l-m-1}$ in background sequences. With a 0-th Markov model, the conditional probability is reduced to $P(b_{j+l-1})$, which is the frequency of base b_{j+l-1} in the background sequences. The probability that a w mer was generated from the motif model Θ_W is $P(j|\Theta_W) = \prod_{l=1}^w q_{l,b_{j+l-1}}$, where $q_{l,b_{j+l-1}}$ is the probability of seeing base b_{j+l-1} at the l -th position of the motif. In our final model, we used a 0-th model since it has the highest accuracy for m -th models with m ranging from 0 to 5.

Based on these two probabilities, a log-ratio score A_{j,Θ_W,B_m} was assigned to each position j in the sequence, which is computed as

$$A_{j,\Theta_W,B_m} = \ln \frac{P(j|\Theta_W)}{P(j|B_m)}.$$

To score a sequence S by a motif module consisting of two motifs Θ_M and Θ_N (e.g. ABRE and CE motifs), we considered all their possible positions i and j within S , as long as they did not overlap and were within a certain distance d . The combined score between positions i and j was computed as $A_{S,\Theta_M,\Theta_N,B_m} = \max_{i,j}(A_{i,\Theta_M,B_m} + A_{j,\Theta_N,B_m})$. The highest combined score among all positions was assigned to the sequence.

3 RESULTS

In our research, we began with an analysis of the *cis*-elements, i.e. ABREs and CEs, of the target genes in *A.thaliana*. We then predicted candidate genes, and verified them using RT-PCR experiments and published microarray profiling results.

3.1 ABREs

To reiterate, abiotic stress, such as drought, cold and salinity, can trigger ABA. Many genes can be induced by ABA and/or one of these stress conditions. It is also known that ABA-responsive genes in rice, maize, barley and other cereals typically have ABREs and CEs as their determinant *cis*-elements. An ABRE for *A.thaliana* has also been identified (Zhu, 2002).

Our analysis showed that the ABREs are also significant *cis*-elements for genes responsive to abiotic stress in *A.thaliana*. In our analysis, we used the expression profiling results published in Hoth et al. (2002), Kreps et al. (2002) and Seki et al. (2002a,b). We first considered the promoters of the genes that are upregulated separately under ABA, cold, drought and salinity. MEME (Bailey and Elkan, 1994), which is one of the best motif finding algorithms (<http://meme.sdsc.edu/meme/website/intro.html>), was used to find statistically significant degenerate motifs from these four sets of promoters. They contain many common motifs. Specifically, ACGT-containing motifs, termed as G-box, are over-represented in all these sets of promoters. They are consistently ranked among the top motifs identified: they are the first for ABA-inducible genes, second for drought-responsive genes and third for both cold- and salinity-inducible genes. The ACGT-containing motifs from the three stress-induced genes are very similar to the ACGT-containing motif in ABA-induced genes. They were directly compared by a computer program CompareACE (downloadable at <http://atlas.med.harvard.edu/download/>). The similarity score used by CompareACE is the Pearson correlation coefficient between the nucleotide base frequencies of the alignment of two motifs. The scores vary between -1 and 1 , where 1 means a perfect match. The three ABRE motifs from stress genes have similarity scores >0.99 as compared with those of the ABRE motif from ABA genes. The four ABRE motifs are shown in Supplementary Figure 1. Similar results were obtained using AlignACE motif finding algorithm (Hughes et al., 2000) (data not shown).

There are significant overlaps among the genes induced by ABA and one of the three stress conditions. We further analyzed the *cis*-elements in the set of genes exclusively induced by one of these stress, resulting in three sets of genes not overlapping with ABA-inducible genes. The analysis of the motifs from these three sets of genes led to similar conclusions, except that the ACGT-containing motifs were now slightly degenerate and ranked slightly lower than in the previous analysis. They ranked as the third, second and third for the genes

unique to cold, drought and salinity, respectively. Nevertheless, their similarities to the ACGT-containing motif from ABA genes were still significant, i.e. >0.98 . These ACGT-containing motifs are shown in online Supplementary Figure 2.

This specificity analysis indicated that ABREs are also determinant *cis*-elements for stress-related transcription regulations. This implies that the genes responsive to ABA and abiotic stress are difficult to separate from one another merely based on these *cis*-elements. It is possible that other *cis*-elements cooperate with ABREs to differentiate various stress regulations, a topic beyond the scope of this paper.

3.2 CEs

Cis-elements are generally degenerate. Even though the ACGT-core in ABREs is well conserved, the flanking sequences beyond the ACGT-core vary. For example, a rice ABRE is CGTACGTGTC (Hobo et al., 1999), whereas an ABRE for maize is GACGTG (Busk et al., 1997) and an ABRE for *A.thaliana* is CCACGTGG (Zhu, 2002).

When compared with ABREs, CEs are less conserved (Busk and Pages, 1997; Hobo et al., 1999; Shen and Ho, 1997). The *EM* gene of rice has a CE (CE3) of GACGCGTGTC (Hobo et al., 1999); maize *RAB28* has CE3 of ACGCGCCTCCTC (Busk et al., 1997), and barley *HVA1* has CE3 of ACGCGTGTCCCTC and *HVA22* has CE1 of TGCCACCGG (Shen and Ho, 1997). These CEs are more diverged than ABREs, whereas the CE3s have a relatively conserved CGCG core. To our knowledge, no experiment has been done to characterize the CEs of *A.thaliana*.

To obtain an accurate, as well as degenerate, pair of ABRE and CE for predicting ABA-inducible and stress-inducible genes in *A.thaliana*, we combined computational methods with the knowledge of known motifs from other plants to obtain the following three types of degenerate ABRE and CE motifs. The first type is based on the experimentally identified ABRE and CE motifs. We constructed PWMs (Stormo, 2000) for ABREs based on motifs from rice [*OSEM* (Hattori et al., 2002), *RAB16A/B/D* (Mundy et al., 1990; Ono et al., 1996)], maize [*RAB28* (Busk and Pages, 1997)], barley [*HVA1* (Shen and Ho, 1997)], and *A.thaliana* [*RD29A* (Zhu, 2002)]. The resulting ABRE is MGTACGTGK. To obtain CEs, we considered the ones from monocots (as no CE from *A.thaliana* is known) and resulted in CEs of GMCGCGTGK. The logos for these ABRE and CE are shown in Figure 1a. Since we used information from monocots, we refer to this type of motifs as Monocots-based motifs for convenience. Note that the number of experimentally verified genes is small; hence the accuracy of the motifs may be low.

The second type of degenerate ABRE and CE motif comes from a refinement to the first type through an iterative procedure that combined the experimentally verified motifs, i.e. the first-type motifs and the results from expression profiling. We applied a motif scan program, which we developed (see Section 2), to the 366 upregulated genes that were identified by gene expression profiling in Seki et al. (2002a,b). We explicitly took the first-type motifs as seeds to scan the promoters. Matched sequences were ranked by their scores (Section 2); the top 15 matched motifs were chosen to construct new PWMs. The new PWMs must be similar but not identical to the monocots-based PWMs. This step was repeated until the PWMs did not change or their specificity decreased. The refined motifs are SRTACGTGTC for ABRE and GACRCGTGK for CE, respectively, whose sequence logos are shown in Figure 1b. Compared with

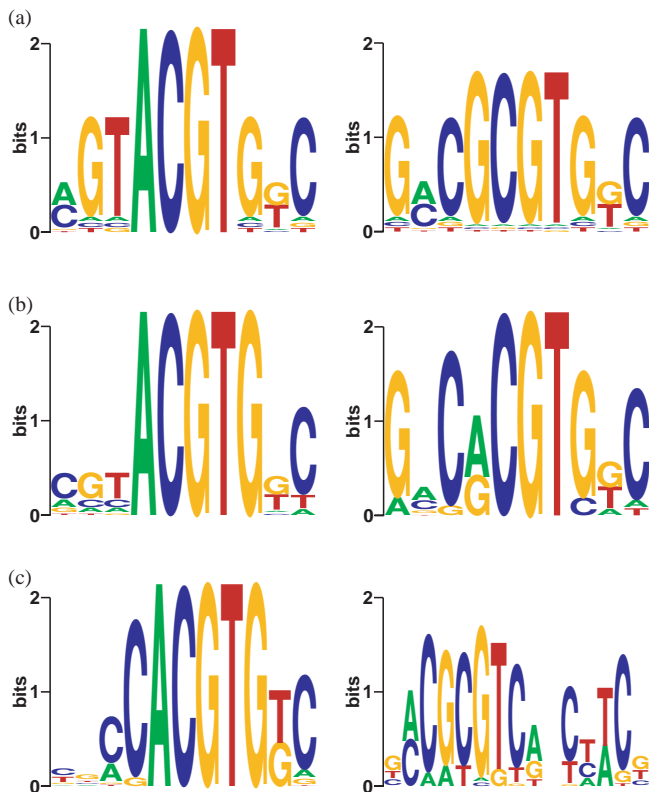


Fig. 1. ABRE–CE modules, where ABREs are to the left and CEs to the right. (a) The monocot-based module was constructed using known ones from monocots; (b) the *Arabidopsis*-specific module was a refinement to the monocot-based module by a repeated search on *A.thaliana* promoters; (c) the MEME-derived module was inferred by MEME motif algorithm.

the monocot-based motifs (Fig. 1a), there are substantial changes in the flanking sequences beyond the ACGT-core. Besides, *A.thaliana* seems to have a less conserved GCGT-core in its CE component, the first G in particular can be A about 30% of the time. As a result, the refined ABREs and CEs are similar to each other and become almost palindromic. We refer to these motifs as *Arabidopsis*-specific motifs.

A third type of ABRE and CE motifs was computationally inferred, as a reference, from the promoters of the 366 stress-responsive genes from microarray experiments under ABA treatment and abiotic stress (Seki *et al.*, 2002a,b). Two motifs, the second and seventh, from the top 10 motifs produced by MEME, appear to be meaningful and the rest seem to be repeats. The motifs for ABREs and CEs are YKMCACGTGKC and MCGCGTCRNYYWCK, respectively, whose sequence logos are shown in Figure 1c. These two motifs are significantly different from the previous two types. For ABREs, the prefix before the ACGT-core is less conserved. The base immediately before the ACGT-core is a relatively conserved C, whereas it is T or A in the previous two types in Figure 1a and b. For CEs, the suffix sequence of the GCGT element does not at all match those in Figure 1a and b.

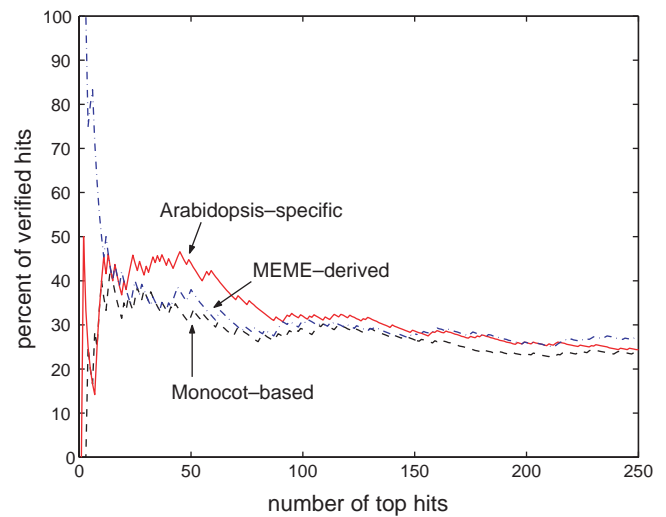


Fig. 2. Prediction accuracies of three ABRE–CE modules in Figure 1. The *x*-axis is the number of motif hits, ordered by their scores matching to the modules; the *y*-axis is the percentage of scored motifs that are expressed in Hoth *et al.* (2002), Kreps *et al.* (2002) and Seki *et al.* (2002a,b).

3.3 Motif modules as target gene indicators

The three types of motifs discussed above were constructed differently and have different degeneracies. One problem with the monocot-based motifs is that they may not be necessarily good indicators of ABA-inducible and stress-inducible genes in dicot *A.thaliana*. One problem with MEME-inferred motifs is that the ABREs and CEs were identified individually rather than as a module.

The *Arabidopsis*-specific motif module seems to be a good indicator for ABA-responsive and stress-responsive genes. To quantify its prediction power, we compared it with the other two modules, measured by the number of predicted genes that were also identified by microarray experiments (Hoth *et al.*, 2002; Kreps *et al.*, 2002; Seki *et al.*, 2002a,b). The results are shown in Figure 2. The monocot-based module has the worst prediction accuracy, and the *Arabidopsis*-specific module is superior to the MEME-derived module, except for the first eight predictions. This result partially supports the interactive approach we used to obtain refined ABRE and CE motifs.

In the rest of this paper, we use the *Arabidopsis*-specific motif module (shown in Figure 1b) for our analysis and prediction.

3.4 ABRE and DRE as coupling elements

A close examination of the CE in the *Arabidopsis*-specific module shows that it is very often a G-box (ACGT-core) or has a GCGT-core. The GCGT-core has a strong conservation in the CEs for many monocots, as shown in Figure 1a. It is also known that ABREs can act as CEs, and so can dehydration responsive elements (DREs) (Guiltingan *et al.*, 1990; Hobo *et al.*, 1999; Narusaka *et al.*, 2003).

We examined the utilities of ABREs, DREs and GCGT-containing motifs as CEs in ABRE–CE module. The results in Figure 3 show that the prediction accuracies using DREs and GCGT-containing motifs as CEs are significantly lower than those with ACGT-core as CE. The figure indicates that unlike most monocots, which very

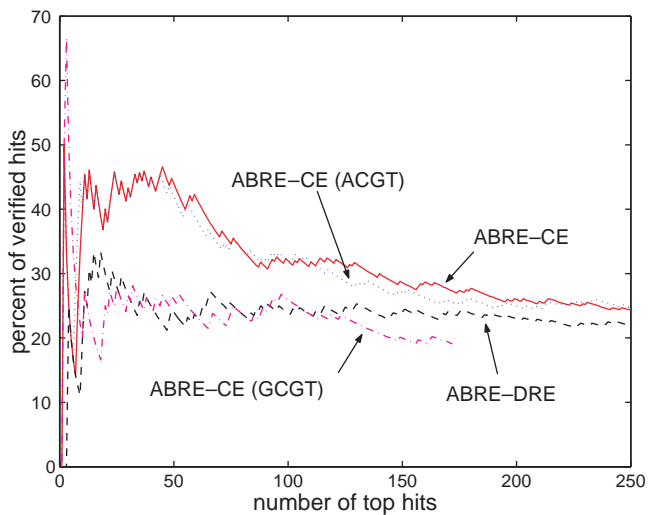


Fig. 3. Prediction accuracies using ABREs (ACGT-core), GCGT-containing motifs and DREs as CEs. ABRE-CE is *Arabidopsis*-specific module. ABRE-CE(ACGT), ABRE-CE(GCGT) and ABRE-DRE use ABREs, GCGT-containing motifs and DREs as CEs, respectively. The *x*-axis and *y*-axis are the same as those of Figure 2.

often have GCGT-core in their CEs, *A.thaliana* tends to have ABREs as CEs. The results also show that DREs are less effective as CEs than ABREs. One reason may be that the DREs have only six bases (RCCGAC), which may give rise to a relatively large number of false positive matches.

3.5 Where ABRE-CE module locate

Before applying the *Arabidopsis*-specific ABRE-CE module, we need to assess if this module is indeed unique to ABA-inducible and stress-inducible genes. For this purpose, we analyzed the distributions of high-quality matches of the module in various regions of *A.thaliana* genome. We considered the promoters, the intron regions and the second exon regions of all genes, and the promoters of the upregulated and downregulated genes under ABA and stress conditions (Hoth *et al.*, 2002; Kreps *et al.*, 2002; Seki *et al.*, 2002a,b). We also included randomly constructed sequences using the frequencies of nucleotides in all promoters, to evaluate the possibility that the module appears by chance. The results in Figure 4 show that the ABRE-CE module occurs more often in the promoters of the target genes than in the coding and other non-coding regions.

3.6 Prediction and experimental verification

Using the *Arabidopsis*-specific ABRE-CE module, we detected a large number of putative ABA-responsive and stress-responsive genes. We closely examined the highest scored 40 predictions, listed in Table 1. We tested 27 genes using RT-PCR on 10-day-old seedlings (Section 2). Among these 27 genes, 17 (63.0%) are verified as upregulated, 3 (11.1%) have no transcripts detected and 7 (25.9%) have no significant expression change. Some of the RT-PCR results are shown in Supplementary Figure 3. In combination with the results from previously published microarray results (Hoth *et al.*, 2002; Kreps *et al.*, 2002; Seki *et al.*, 2002a,b), we found that 27 of the top 40 genes were confirmed as

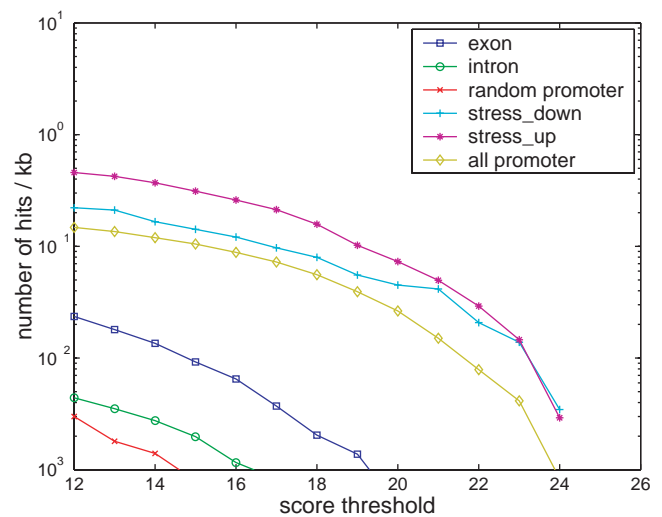


Fig. 4. Location specificities of ABRE-CE module in second exon regions, intron regions, random promoters, promoters of downregulated (stress_down) and upregulated (stress_up) genes identified in Hoth *et al.* (2002), Kreps *et al.* (2002) and Seki *et al.* (2002a,b), and all promoters.

ABA and/or abiotic-stress responsive, giving a prediction accuracy of 67.5%.

Three genes (*At3g33131*, *At5g52290* and *At1g65200*) from the set of 27 genes were not detected by RT-PCR. Based on the GenBank annotation as of May 2004, two of these three genes (*At3g33131* and *At5g52290*) were annotated as hypothetical proteins. In addition, seven genes (*At1g79040*, *At1g77450*, *At1g28530*, *At1g52230*, *At2g18700*, *At4g21270* and *At4g21280*) did not show significant expression changes in the RT-PCR experiment. It is possible that these genes express in different developmental stages or tissues other than the conditions of the RT-PCR experiments. Indeed, one of these genes (*At1g77450*) was verified as upregulated in Hoth *et al.* (2002). One of the differences between the experimental conditions used in Hoth *et al.* (2002) and our RT-PCR is the age of the seedlings [4-week-old seedlings in Hoth *et al.* (2002)] versus 10-day-old seedlings in our experiments). Combining these observations, the 67.5% prediction accuracy is, apparently, a lower bound on the prediction accuracy.

Although we used the degenerate ABRE-CE module, which allows an ACGT component in the CE motif, only one gene among the top 40 predictions actually has the GCGT-core in its CE. In other words, most of the top candidates match to ABRE-ABRE module well, in agreement with our analysis of CEs in Section 3.4.

3.7 Where ABREs and CEs locate

It is worthwhile to know the locations of ABRE and CE within promoters. We examined two position statistics. The first is the gap between the ABRE and CE in a module. Figure 5a shows the results in the promoters of all genes and ABA-inducible and stress-inducible genes identified in (Hoth *et al.*, 2002; Kreps *et al.*, 2002; Seki *et al.*, 2002a,b) whose promoters contain these ABRE-CE module. The gaps are typically <150 bases in these genes, although a few gaps beyond 150 bases exist. The most possible gap between these two elements are ~40–50 bases.

Table 1. Top 40 predicted ABA-inducible and stress-inducible genes in *A.thaliana* and their experimental verification

Assession no.	ABRE	CE	Strand	Function annotation in GenBank	Verification
<i>At5g07920</i>	cctacgtggc	ggcacgtggc	+	Diacylglycerol kinase (ATDGK1)	4
<i>At1g79040</i>	cctacgtggc	gccacgtgtc	+	Photosystem II polypeptide-related	-1
<i>At4g37220</i>	cctacgtggc	gccacgtgtc	+	Cold acclimation protein homolog	1
<i>At3g33131</i>	cgaacgtgtc	gacggtggc	+	Hypothetical protein	-2
<i>At5g24155</i>	cacacgtggc	gccacgtggc	-	Squalene monooxygenase	4
<i>At2g44660</i>	gctacgtggc	gacacgtggc	+	ALG6, ALG8 glycosyltransferase family	4
<i>At4g12680</i>	tgtacgtggc	gacacgtggc	-	Expressed protein	4
<i>At5g66580</i>	agaacgtggc	gccacgtggc	-	Expressed protein	1
<i>At5g50360</i>	cgcacgtggc	gccacgtctc	+	Expressed protein	4
<i>At5g51210</i>	cgtacgtgtc	gacacgtgac	+	Glycine-rich protein oleosin	4
<i>At1g77450</i>	cgaacgtgtc	gccacgtgtc	+	No apical meristem (NAM) protein family	1,-1
<i>At1g17120</i>	cgaacgtggc	gtcacgtggc	+	Amino acid permease family protein	
<i>At5g52290</i>	cgtacgtgtc	gagacgtggc	-	Hypothetical protein	-2
<i>At5g52300</i>	cgtacgtgtc	gagacgtggc	+	Desiccation-responsive protein 29B (RD29B)	2, 3, 4
<i>At5g58650</i>	catacgtggc	gacacgtgtc	+	Expressed protein	
<i>At2g38820</i>	ggtacgtgtc	ggcacgtgtc	-	Expressed protein	2, 4
<i>At1g28530</i>	tgcacgtgtc	gccacgtggc	-	Expressed protein	-1
<i>At1g28540</i>	tgcacgtgtc	gccacgtggc	+	Expressed protein	
<i>At1g54130</i>	gccacgtggc	gacacgtgtc	-	RSH3 (RelA/SpoT homolog)	4
<i>At1g32550</i>	tctacgtggc	gacacgtggc	-	Ferredoxin family protein	4
<i>At1g32560</i>	tctacgtggc	gacacgtggc	+	LEA group 1 protein	3
<i>At1g52220</i>	tccacgtggc	gccacgtggc	-	Expressed protein	1
<i>At1g52230</i>	tccacgtggc	gccacgtggc	+	Photosystem I subunit VI precursor	-1
<i>At4g21270</i>	accacgtgtc	gccacgtggc	+	Kinesin-like protein A (katA)	-1
<i>At4g21280</i>	accacgtgtc	gccacgtggc	-	Oxygen-evolving enhancer protein 3 (PSBQ)	-1
<i>At3g03680</i>	ccgacgtggc	gtcacgtggc	+	C2 domain-containing protein	4
<i>At2g22240</i>	ccaacgtgtc	gccacgtgtc	+	Myo-inositol 1-phosphate synthase-related	2
<i>At3g19590</i>	cccacgtgtc	gccacgtgac	-	Mitotic checkpoint protein	4
<i>At1g58520</i>	acaacgtggc	gacacgtggc	+	Early-responsive to dehydration (ERD4)	4
<i>At3g18290</i>	agcacgtggc	ggcacgtgac	-	Zinc finger protein-related	1
<i>At1g65200</i>	cgtacgtgac	gtcacgtggc	+	Ubiquitin carboxyl-terminal hydrolase-related	-2
<i>At2g18700</i>	gaaacgtggc	gccacgtggc	-	Glycosyltransferase family 20	-1
<i>At2g36270</i>	cacacgtgtc	gacacgtgtc	+	ABA insensitive 5 (ABI5)	4
<i>At1g02660</i>	ctgacgtggc	gccacgtgtc	+	Lipase (class 3) family	1
<i>At1g02670</i>	ctgacgtggc	gccacgtgtc	-	DNA repair protein, putative	4
<i>At1g74450</i>	agcacgtgga	gccacgtggc	-	Expressed protein	4
<i>At5g05220</i>	caaacgtgtc	gacacgtggc	+	Expressed protein	3
<i>At3g62260</i>	gccacgtgtc	gacacgtgtc	+	Protein phosphatase 2C (PP2C)	
<i>At5g65890</i>	tccacgtgtc	gccacgtggc	-	ACT domain-containing protein (ACR1)	
<i>At5g62490</i>	aacacgtgtc	gccacgtggc	-	ABA-responsive protein (HVA22b)	3, 4

1: Hoth *et al.* microarray (Hoth *et al.*, 2002); 2: Kreps *et al.* microarray (Kreps *et al.*, 2002); 3: Seki *et al.* microarray (Seki *et al.*, 2002a,b); 4: Upregulated, verified by RT-PCR; -1: no ABA response detected by RT-PCR; -2: no transcripts detected by RT-PCR.

The second statistic is the start position of an ABRE-CE module from the transcription start site (TSS) of a gene. The results on all genes and ABA-inducible and stress-inducible genes are shown in Figure 5b. ABRE-CE modules are usually within 200 bases from TSSs; the majority of them are <120 bases from TSSs. Note that a few ABRE-CE modules start within 5'-UTRs.

3.8 Functional categories

The function of ABA-responsive genes are diverse, reflected by the large number of functional categories these genes may be involved in. We carried out a functional analysis on two sets of genes, the ABA-inducible and stress-inducible genes reported in Hoth *et al.* (2002), Kreps *et al.* (2002) and Seki *et al.* (2002a,b), and the top

150 putative genes predicted in our study. This was done using the MIPS functional category classification from <http://mips.gsf.de/projects/plants>

Among the 1825 stress-inducible genes from the microarray experiments, 1530 (83.8%) can be assigned to at least one functional category. Among our top 150 predicted genes, 126 (84.0%) have a functional category. Moreover, these two sets of genes have similar distributions across a wide range of functional categories, as depicted in Supplementary Figures 4 and 5. Except the unclassified proteins, the three largest categories are transcription, metabolism and binding proteins. This result suggests that there may be a lot of gene regulation activities after ABA treatment and stress.

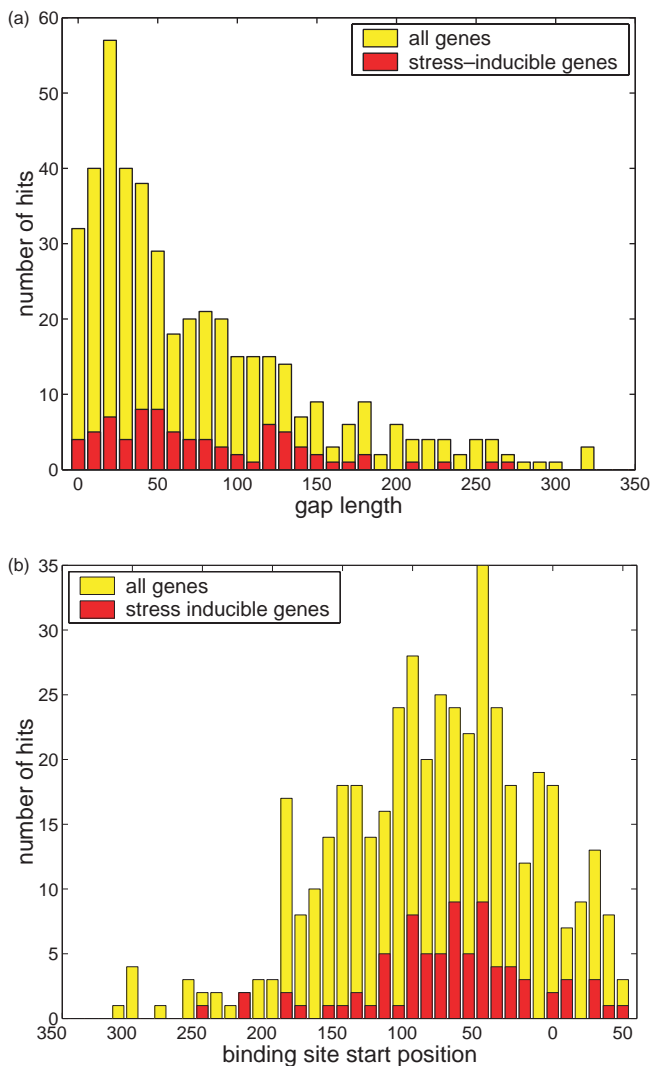


Fig. 5. Position statistics of *Arabidopsis*-specific ABRE-CE modules. (a) Distribution of the gaps between ABREs and CEs; (b) distribution of the start positions of ABRE-CE modules relative to TSSs.

4 DISCUSSION

We advocated and investigated a *cis*-element based method for finding genes that are responsive to certain conditions, which can be referred to as targeted gene finding. This method is orthogonal and complementary to conventional approaches to gene finding and function annotation that are based on the conservation of ORFs. The *cis*-element based targeted gene finding method explicitly utilizes the information of transcription regulations. By exploiting experimentally verified *cis*-elements in a genome-wide screening, it naturally combines the fidelity of gene functions elucidated in experimental analyses with computational efficiency of a genome-scale search.

In this study, we focused on ABA-responsive and abiotic stress-responsive genes in *A.thaliana* and their *cis*-elements, i.e. ABREs and CEs. By employing the experimentally identified *cis*-elements, we are able to leverage genome-wide targeted gene finding with a

huge amount of previous experimental efforts devoted to identifying ABA-responsive and stress-responsive genes and elucidating their regulatory mechanisms.

The idea of using *cis*-elements to identify genes responsive to certain stimuli is intuitive, and was pursued in at least two previous studies. The first study was reported in Markstein *et al.* (2002), on a genome-wide analysis of the binding sites for Dorsal, one of the best-characterized sequence-specific TFs in *Drosophila*. It was known that many Dorsal targeted genes contain a cluster of multiple Dorsal binding sites in a small vicinity of their promoter regions. Using the known Dorsal binding motifs, fifteen promoters that contains clusters of Dorsal binding motifs were identified from *Drosophila* genome. Among the fifteen genes, three are known Dorsal target genes. Using *in situ* localization assays, two other genes were shown to be upregulated in the presumptive mesoderm of early embryos, leading to a total prediction accuracy of ~34% (5 positive of 15 putative ones). The second study was on interneurons called AIY in *Caenorhabditis elegans* (Wenick and Hobert, 2004). Using the newly sequenced *Caenorhabditis briggsae* genome, another nematode diverged from *C.elegans* ~70–100 million years ago. Wenick and Hobert found eight genes in AIY in *C.elegans* that are also conserved in *C.briggsae*. Using a standard promoter-dissection approach, they discovered *cis*-elements that are necessary and sufficient for AIY transcriptome. Moreover, they carried out a genome-wide screening using the discovered *cis*-elements to predict genes in *C.elegans* that may express in AIY. They experimentally tested 15 of the top 26 predictions and confirmed 14 of them expressed in AIY, giving a prediction accuracy of 14/15 or 94%. Overall, the verified AIY hit rate was 41 of 57 or 72%. As a comparison, we achieved a comparable prediction accuracy of 67.5% for the top 40 predictions in our study.

Our approach in this paper and the approach taken in Markstein *et al.* (2002) and Wenick and Hobert (2004) are similar and complement one another. These studies used similar genome-wide search strategies to predict genes that may have certain expression profiles. They did not make strong assumptions about where within promoters the motif matches should be. However, these approaches differ in the way that *cis*-elements were derived. Markstein seemed to use exact known Dorsal binding motifs in the analysis. Wenick *et al.* relied on the conservation of genes in *C.elegans* and *C.briggsae*, to infer *cis*-elements that are characteristic to the target genes. We obtained *cis*-elements based on previous experimental analyses.

In summary, based on the previous studies and the results in this paper, it is evident that the *cis*-element based targeted gene finding approach is effective and general; it has a high prediction accuracy and is applicable to different organisms and different type of genes. With more information of TFs and their DNA binding information becoming freely available, including those in TRANSFAC database, we expect this cost-effective and accurate approach to be widely applied to various targeted gene finding problems in the future.

ACKNOWLEDGEMENTS

This research was supported in part by NSF grant EIA-0113618 and a grant from Monsanto Corporation to W.Z. and in part by a grant from Monsanto Corporation to R.S.Q. We thank the other members of W.Z. and R.S.Q.'s groups for the helpful discussions.

REFERENCES

- Bailey, T. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the 2nd ISMB Conference*, Palo Alto, CA, pp. 28–36.
- Brivanlou, A. and Darnell, J. (2002) Signal transduction and the control of gene expression. *Science*, **295**, 813–818.
- Busk, P. and Pages, M. (1997) Protein binding to the abscisic acid-responsive element is independent of viviparous1 *in vivo*. *Plant Cell*, **9**, 2261–2270.
- Busk, P. *et al.* (1997) Regulatory elements *in vivo* in the promoter of the abscisic acid responsive gene rab17 from maize. *Plant J.*, **11**, 1285–1295.
- Carles, C. *et al.* (2002) Regulation of *Arabidopsis thaliana* Em genes: role of AB15. *Plant J.*, **30**, 373–383.
- Choi, H.I. *et al.* (2000) ABFs, a family of ABA-responsive element binding factors. *J. Biol. Chem.*, **275**, 1723–1730.
- Finkelstein, R. *et al.* (2002) Abscisic acid signaling in seeds and seedlings. *Plant Cell*, (suppl.), S15–S45.
- Gultinan, M. *et al.* (1990) A plant leucine zipper protein that recognizes an abscisic acid response elements. *Science*, **250**, 267–271.
- Harbison, C. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
- Hattori, T. *et al.* (1995) Regulation of the Osem gene by abscisic acid and the transcriptional activator VP1: analysis of *cis*-acting promoter elements required for regulation by abscisic acid and VP1. *Plant J.*, **7**, 913–925.
- Hattori, T. *et al.* (2002) Experimentally determined sequence requirement of ACGT-containing abscisic acid response element. *Plant Cell Physiol.*, **43**, 136–140.
- Higo, K. *et al.* (1999) Plant *cis*-acting regulatory DNA elements (PLACE) database. *Nucleic Acids Res.*, **27**, 297–300.
- Hobo, T. *et al.* (1999) ACGT-containing abscisic acid response element (ABRE) and coupling element 3 (CE3) are functionally equivalent. *Plant J.*, **19**, 679–689.
- Hoth, S. *et al.* (2002) Genome-wide gene expression profiling in *Arabidopsis thaliana* reveals new targets of abscisic acid and largely impaired gene regulation in the abi1-1 mutant. *J. Cell Sci.*, **115**, 4891–4900.
- Hughes, J. *et al.* (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **296**, 1205–1214.
- Kreps, J. *et al.* (2002) Transcriptome changes for *Arabidopsis* in response to salt, osmotic, and cold stress. *Plant Physiol.*, **130**, 2129–2141.
- Lander, E. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
- Lescot, M. *et al.* (2002) PlantCARE, a database of plant *cis*-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences. *Nucleic Acids Res.*, **30**, 325–327.
- Marcotte, W. *et al.* (1989) Abscisic acid-responsive sequence from the Em gene of wheat. *Plant Cell*, **1**, 969–976.
- Markstein, M. *et al.* (2002) Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. *Proc. Natl Acad. Sci. USA*, **22**, 763–768.
- Matys, V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
- Mundy, J. *et al.* (1990) Nuclear proteins bind conserved elements in the abscisic acid-responsive promoter of a rice rab gene. *Proc. Natl Acad. Sci. USA*, **87**, 1406–1410.
- Narusaka, Y. *et al.* (2003) 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.*, **34**, 137–148.
- Oeda, K. *et al.* (1991) A tobacco bZip transcription activator (TAF-1) binds to a G-box-like motif conserved in plant genes. *EMBO J.*, **10**, 1793–1802.
- Ono, A. *et al.* (1996) The rab16b promoter of rice contains two distinct abscisic acid-responsive elements. *Plant Physiol.*, **112**, 483–491.
- Seki, M. *et al.* (2002a) Monitoring the expression pattern of around 7000 *Arabidopsis* genes under ABA treatments using a full-length cDNA microarray. *Funct. Integr. Genomics*, **2**, 282–291.
- Seki, M. *et al.* (2002b) Monitoring the expression profiles of 7000 *Arabidopsis* genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. *Plant J.*, **31**, 279–292.
- Shen, Q. and Ho, T.H. (1997) Promoter switches specific for abscisic acid (ABA)-induced gene expression in cereals. *Physiol. Plantarum*, **101**, 653–664.
- Shen, Q. *et al.* (1996) Modular nature of abscisic acid (ABA) response complexes: composite promoter units that are necessary and sufficient for ABA induction of gene expression in barley. *Plant Cell*, **8**, 1107–1119.
- Shinozaki, K. and Yamaguchi-Shinozaki, K. (2000) Molecular responses to dehydration and low temperature: differences and cross-talk between two stress signaling pathways. *Curr. Opin. Plant Biol.*, **3**, 217–223.
- Stormo, G. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
- The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
- Wenick, A. and Hobert, O. (2004) Genomic *cis*-regulatory architecture and *trans*-acting regulators of a single interneuron-specific gene battery in *C.elegans*. *Dev. Cell*, **6**, 757–770.
- Xu, D. *et al.* (1996) Expression of a late embryogenesis abundant protein gene, HVA1, from barley confers tolerance to water deficit and salt stress in transgenic rice. *Plant Physiol.*, **110**, 249–257.
- Yu, J. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science*, **296**, 79–91.
- Zhang, M. (2002) Computational prediction of eukaryotic protein-coding genes. *Nat. Rev. Genet.*, **3**, 698–709.
- Zhu, J.K. (2002) Salt and drought stress signal transduction in plants. *Annu. Rev. Plant Biol.*, **53**, 247–273.