Genome Informatics 21: 177-187 (2008)

CIS-REGULATORY ELEMENT BASED GENE FINDING: AN APPLICATION IN ARABIDOPSIS THALIANA

YONG LI[a,1] YANMING ZHU[a,1*] YANG LIU[2]
yong@neau.edu.cn ymzhu2001@neau.edu.cn liuyang@cuhk.edu.hk

YONGJUN SHU[1] FANJIANG MENG[3] YANMIN LU[3]
syjun@neau.edu.cn fjmeng@neau.edu.cn luyanmin@neau.edu.cn

BEI LIU[3] XI BAI[1] DIANJING GUO[2*]
liubei@neau.edu.cn maixi@neau.edu.cn djguo@cuhk.edu.hk

[1] *Plant Bioengineering Laboratory, Northeast Agricultural University, Harbin, China*
[2] *State Key Lab for Agrobiotechnology and Department of Biology, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong*
[3] *Department of Computer Science, Northeast Agricultural University, Harbin, China*

[a] *These authors contributed equally to this work*

[*] *Corresponding author*

Abstract

Using cis-regulatory motifs known to regulate plant osmotic stress response, an artificial neural network model was built to identify other functionally releted genes involved in the same process. The rationale behind our approach is that gene expression is largely controlled at the transcriptional level through the interactions between transcription factors and cis-regulatory elements. Gene Ontology enrichment analysis on the 500 top-scoring predictions showed that, 60% of the enriched GO classification was related to stress response. RT-PCR analysis showed that nearly 70% of the top-scoring predictions exhibited altered expression under various stress treatments. We expect that similar approach is widely applicable to infer gene function in various cellular processes in different species.

***Keywords*:** Artificial Neural Network; Gene Expression; Gene Finding; Cis-regulatory element; *Arabidopsis thaliana*

1. Introduction

Gene expression is largely controlled at the transcriptional level, where the interactions between transcription factors (TFs) and cis-regulatory elements in the promoter region of a gene play crucial roles [6]. Previous research suggests that functional related genes tend to be co-regulated by similar sets of transcription factors. Therefore, using cis-regulatory motifs are known to regulate gene expression in certain cellular process, one can identify other functionally relevant genes involved in the same process. When combined with experimental verification, this has been proved to be an effective approach to genome-wide targeted gene identification [28].

Drought, high salinity, and low temperature are three major osmotic stresses that

adversely affect plant growth, development, or productivity. Osmotic stress elicits a dehydration response in plants that shares many common elements and interacting signaling pathways [5, 6, 28], which have been suggested to be Abscisic Acid (ABA) dependent [20]. Subsequent analysis of the ABA-regulated gene promoter region has led to the identification of several ABA-responsive elements (ABREs) [7, 12]. Zhang et al. [28] reported a computational approach to identifying putative ABA responsive genes using conserved ABA-responsive element (ABRE) and its coupling element (CE). Using similar cis-element based approach, promoters that contains known binding motifs were used for targeted gene finding in *Drosophila melanogaster* [13] and *C. elegans* [24]. Despite the proved success, the previous researchers all used one or two specifically defined motifs for gene screening. In fact, a growing body of evidence suggests that functional related genes tend to be regulated by a common set of regulatory proteins to form namely transcription regulatory modules, in order to respond to internal and external signals. By organizing the genome into such modules, a living cell can coordinate the activities of many genes and carry out complex functions [25]. For gene function inference in complex cellular process such as stress response, more sophisticated approaches are required.

Identification of genes that specifically respond to internal and external cues remains one of the most compelling yet elusive areas in computational genomics. Currently the commonly used gene finding approach is consensus-based comparative analysis that relies on sequence homology among genes in closely related species [27]. Such method has limited application because a large portion of those sequenced genomes still remain uncharacterized. Furthermore, such consensus-based method may not be efficient for identification of genes that are induced under specific environmental stimuli. In this study, we applied an Artificial Neural Network (ANN) modeling approach [8, 12, 16, and 17] to plant functional genomics and identified genes respond to osmotic stress in *A. thaliana*. We demonstrate its efficacy by Gene Ontology enrichment analysis as well as by RT-PCR analysis.

2. Materials and Methods

2.1. *Stress Response Genes and Cis-regulatory Elements*

Cis-regulatory elements in the promoter region of drought, salinity, and/or cold stress responsive genes were collected from public database PLACE [9, 29], PlantCARE [18, 32], and DoOP [2]. Other motifs were collected through literature-mining approach. The redundant motifs were eliminated and in total 55 cis-acting elements were collected for further analysis. A bioperl module was used to search for significant motifs occurred in the promoter region. P-value was calculated to confirm the significance of motif detection (Poisson distribution [19]).

2.2. *Promoter Sequences*

Arabidopsis genome sequences were downloaded from TAIR [33]. Transcription start site (TSS) was predicted using TSSP-TCM software from Shahmuradov's group [35].

When multiple TSSs were predicted, the one closest to the ORF was chosen. For each given TSS, we retrieved a segment from 500 bases upstream to 20 bases downstream of the TSS for motif analysis. In total, the TSSs of 18061 ORFs were retrieved.

2.3. *Scoring algorithms*

A Bioperl module was used to search for significant motifs occurred in the promoter region of reported stress responsive genes. P-value was calculated to confirm the significance of motif detection. The ANN toolkit in Matlab was used to establish a feed-forward cascade neural network model. For network training and simulation, we retrieved the promoter region of 362 genes annotated as "response to drought, high salinity, or cold stress" according to Gene Ontology terminology [30, 31] and used these as positive dataset. The promoter sequences of a randomly selected 1086 ORFs (3 fold of positive dataset) from the rest of the gene pool (not annotated as "response to stress or ABA treatment" according to GO) were used as negative dataset. The number of times each cis-regulatory element appears in the promoter region and the ratio of cis-element length to promoter length (we defined it as coverage) were taken as inputs for the network training. Principle component analysis was conducted to eliminate the input node with least effect.

2.4. *Gene Expression Data Analysis and GO Enrichment*

Microarray gene expression data was collected from AtGenExpress [32]. The dataset include global Arabidopsis transcriptome profile change over UV-B light, high salinity, drought and cold stress responses. The raw data was normalized using RMAExpress [32, 33] and differentially expressed genes were detected using BRB ArrayTools [34] ($p<0.05$). The subset of differentially expressed genes contains 3276 gene transcripts identifiers with significantly higher abundance at least once per treatment and per time-point. We performed GO term enrichment analysis using software suite DAVID 2007 (The **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery 2007) [35]. The enrichment analysis is indeed to compare the annotation composition in the analyzed gene list to that of population background genes. DAVID default population background, which is the corresponding genome-wide genes with at least one annotation in the analyzing categories, was used in enrichment calculation.

2.5. *Plant Materials, stress treatment, and RT-PCR analysis*

Four-week-old seedlings grown in a controlled environment growth chamber under a 16 hr light/8 hr dark period, a photo fluency rate of 3000 lux, and a temperature of 22 ºC. For salinity stress treatment, seedlings were subjected to 250 mM NaCl and time series samples were collected at 4hr, 12hr, and 24hr respectively. For cold stress treatment, seedlings were incubated at 4ºC under darkness condition for 24hr, 48hr, and 72hr respectively. For drought treatment, seedlings were subjected to drought for 24hr, 48hr, and 72hr. Total RNA was extracted from whole seedlings of the control plants and

stressed plants using RNeasy Plant Mini Kit (Qiagen,Germany). cDNA was synthesized using super script II Kit (Invitrogen) following the manufacturer's instruction. PCR was performed 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, 29 cycles of 94◦C for 1 min, 58◦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 two independent RT-reactions using forward and reverse primers.

3. Results and Discussions

3.1. *In silico Gene Identification and Gene Ontology Analysis*

Using cis-regulatory motifs known to regulate osmotic stress response, an artificial neural network model was built to identify other relevant genes involved in the same process. The trained model was able to distinguish between genes that do and do not respond to stress, based on the motif patterns of gene promoters in the training dataset. We then applied the model to the candidate dataset to infer the function of the unknown genes.

According to network theory [1, 23], genes within co-expression context often share conserved biological functions. To investigate the significant functional annotation of our predictions, we selected the 500 top ranking genes predicted by ANN model and performed Gene Ontology (GO) enrichment analysis (Table 1). GO provides a controlled vocabulary for describing genes and gene products in living organism [19]. We used terms from "Biological Process" (GOBP), which is one of the three broad GO categories (the other two being "Molecular Function" and "Cellular Component"), to represent gene function. GOBP terms are organized into a directed acyclic graph (DAG) to reflect the hierarchical relationships between the terms. Parent GOBP terms are subdivided into increasingly specific child GOBP terms. This GOBP term: "reponse to stress" has 19 child terms, such as *GO:0009409 [response to cold]*, *GO:0009408 [response to heat]*, and *GO:0009414 [response to water deprivation]*, etc. The GO enrichment analysis is indeed to compare the annotation composition in the analyzed gene list to that of population background genes. We used the DAVID default population background in enrichment calculation, which is the corresponding genome-wide genes with at least one annotation in the analyzing categories. The default background is a good choice for the studies in genome-wide scope or close to genome-wide scope.

According to GO enrichment analysis, except for the un-annotated ORFs (~40%), about 60 %of the significantly enriched GO classification was related to stress response or ABA response. In fact, ABA plays a protective role in plant response to osmotic stress [15, 21, 26] and a large number of genes respond to abiotic stress are also inducible by ABA treatment [15]. The GO enrichment analysis clearly demonstrated that our regulatory motif based computational model is a reliable means for gene function

inference at the genome level.

Table 1. GOBP description of top predictions using ANN model

GOBP term	No. of Genes	P-Value
photosynthesis	9	1.32E-04
cold acclimation	5	2.49E-04
response to abscisic acid stimulus	10	6.45E-04
response to water	7	0.006
response to temperature stimulus	10	0.007
response to cold	7	0.008
response to water deprivation	6	0.013
development	28	0.015
response to hormone stimulus	18	0.015
response to abiotic stimulus	34	0.02
response to salt stress	5	0.034
seed development	7	0.035
response to chemical stimulus	25	0.035
reproductive structure development	7	0.039
actin cytoskeleton organization	4	0.041
response to stimulus	44	0.042
embryonic development	6	0.047
response to stress	24	0.054
cytoskeleton organization	7	0.057
response to endogenous stimulus	20	0.057
response to osmotic stress	5	0.072
reproduction	8	0.078
carbohydrate biosynthesis	8	0.084
cellular response to water	2	0.09

3.2. *Cross Validation Using Gene Expression Profiling Data*

Since many functionally related genes display coordinated transcriptional regulation, large-scale gene expression measurements can therefore serve as a check point for our *in silico* prediction. Comparison of our prediction with stress microarray data from AtGenExpress revealed that about 30% of the top-scoring gene transcripts were

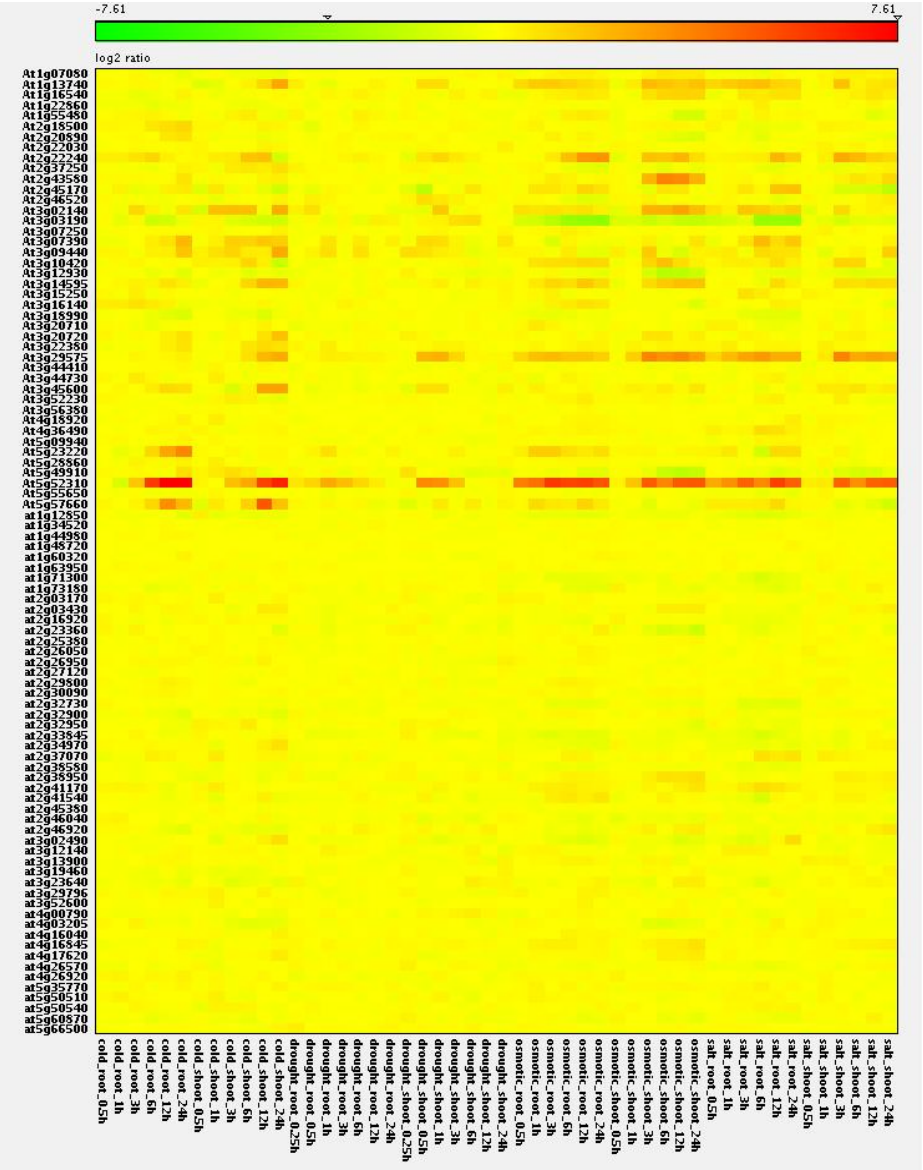

Figure 1. Comparative E-northern analysis of top scoring genes (upper part of the image) vs. randomly selected genes (lower part of the image). High scoring genes and randomly selected genes can be distinguished by different IDs ("At-----" for high scoring genes and "at------" for randomly selected gene). Each row in the heat map is a gene, and each column is an experimental condition, such as high salinity, cold, or drought stress etc. The color at a point represents the log2 of the ratio of the average of replicate treatments relative to the average of corresponding controls.

significantly changed upon stress treatment (p-value ≈ 0). It is well-known that microarray gene expression data, although powerful in providing global transcriptome information, is highly noisy and discontinuous. Some genes that are rapidly and transiently induced by stress may not be detected by microarray analysis. At1g16540,which encodes molybdenum cofactor sulfurase, was predicted as stress response gene by our method. Although direct measurement of its transcript abundance is lacking, previous research indicate that this protein may play important role in regulating many stress relevant genes including *RD29A, COR15, COR47, RD22*, and *P5CS* [20, 21]. However, this gene was not differentially expressed in published Arabidopsis stress microarray data. Similarly, another top-scoring gene At2g23430 (*GO: 0009737*), which encodes a cyclin-dependent kinase inhibitor known to respond to ABA stimulus [22], was also not detected in the stress microarray data. Furthermore, gene expression at the mRNA level does not always reflect the protein function due to post-transcriptional and post-translational regulation. Our method may serve as a complimentary approach to gene expression analysis in stress response gene identification.

To provide visual evidence supporting our computational predictions, we generated a E-Northern heat map (Fig. 1) for 41 top scoring genes (upper part of the image) and 41 randomly selected genes beyond the prediction list (lower part of the image) using the Expression Browser tool of the Botany Array Resource (BAR) [22] and the AtGenExpress Stress data. From E-Northern analysis, we observed significant up-regulation and down-regulation of many of our predictions. The up- and down-regulation were also observed among the genes in the training data (data not shown). In contrast, randomly selected genes show much less change upon stress treatment. Since a gene that shows significant alteration at transcript level under stress conditions is likely to be involved in response to stress, we conclude that regulatory motif in the promoter region is highly indicative of gene function and can therefore be used for *in silico* gene identification. The rapid development of cis-elements databases, such as TRANSFAC [14], PLACE [9, 29] and PlantCARE [18, 32], provides valuable sources that can be used to identify more cis-regulatory motifs relevant to cellular response to various internal and external stimuli.

3.3 *Experimental Validation and Comparison with Other Methods*

To further assess our method, the Arabidopsis seedlings were subjected to various stressful conditions (method) and the transcript abundance of 41 top-scoring predictions were monitored by RT-PCR analysis (Fig. 2). Gene-specific primers were designed based on the cDNA full-length sequence. Overall, 27 of 41 tested genes showed altered level of expression compared to control, giving a prediction accuracy of 65.8% for the top scoring predictions. The full list of the RT–PCR results is provided in supplementary file. Some of the tested genes exhibited varied expression patterns in different types of stress treatment, suggesting possible synergistic effects of multiple transcription factors

involved in various type of osmotic stress. We believe that identification of distinctive class of transcription-factor binding motifs under specific stress condition will facilitate more efficient computational discovery of stress relevant genes. Such knowledge is also important in understanding the cross-talks between distinctive signaling pathways and the underlying regulatory machinery of cellular stress response.

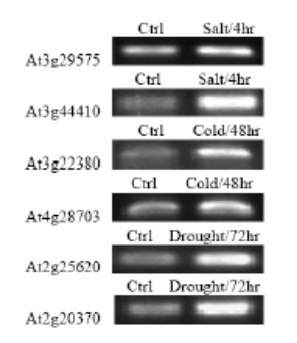

Figure2. RT-PCR analysis of selected predictions. For salinity stress treatment, 4-week old Arabidopsis seedlings were subjected to 250 mM NaCl and samples were collected at 4hr time point. For cold stress and drought stress, seedlings were incubated at 4ºC and 22 ºC respectively in the dark and samples were collected at 48hr, and 72hr repectively. Samples were also collected from control plants grown under the same condition for parallel comparison. Actin was used as loading control and loading was estimated by staining the gel with ethidium bromide. Expression analysis of each gene was confirmed in at least two independent RT-reactions using forward and reverse primers.

Cis-regulatory element based gene identification has been reported previously [13, 20, 24]. These researchers utilized one or two well defined cis-regulatory motifs to search for functional related genes and achieved varied prediction accuracy (ranging from 34% to 72%). As a comparison, we expand the motif list and used a set of diversed regulatory motifs (55 in total) identified from promoters of both experimentally validated and computational predicted stress responsive genes to train an ANN model. The learned model achieved a comparable prediction accuracy of 67.8% for the top scoring predictions in our study. Our results, together with those from other groups, further demonstrate that cis-regulatory motifs are highly indicative of gene function. With more information of transcription factors and their DNA binding information becoming freely available, we anticipate this computational approach can be widely used for gene function inference in different organisms.

4. Conclusions

In this study, we present a cis-regulatory element based ANN modeling method for genome wide gene function prediction in Arabidopsis thaliana. By explicitly utilizing the information of transcription regulation of known gene and cis-regulatory motifs, our method gives reliable result with high prediction accuracy. We demonstrate this is a

practical and compelling means for genome wide gene identification. We anticipate that identification of more condition-specific cis-regulatory motifs and further understanding of the synergistic effects of different regulatory motifs will facilitate more efficient computational discovery of stress relevant genes. One promising aspect of the approach as applied in this study is in its potential use for gene finding in various cellular processes in different organisms. The software codes are available upon request.

Acknowledgments

This project was supported by grant from National Natural Science Foundation of China (30570990), and in part by grants from the Science and Technology Department of Heilongjiang province and Hong Kong UGC/AoE Plant & Agricultural Biotechnology Project AoE-B-07/09.

References

1 Barabasi, A.L. and Oltvai Z.N., Network biology: understanding the cell's functional organization**,** *Nat Rev Genet*, 5(2):101-113, 2004.

2 Barta, E., Sebestyén, E., Pálfy, T.B., Tóth, G., Ortutay, C.P. and Patthy L., DoOP: Databases of Orthologous Promoters, collections of clusters of orthologous upstream sequences from chordates and plants, *Nucl. Acids Res.***,** 33:D86-D90 2005.

3 Bittner, F., Oreb, M. and Ralf, Mendel, R.R., ABA3 Is a Molybdenum Cofactor Sulfurase Required for Activation of Aldehyde Oxidase and Xanthine Dehydrogenase in *Arabidopsis thaliana*, *The Journal of Biological Chemistry*, 276(44):40381-40384, 2001.

4 Bolstad, B.M., Irizarry, R.A., Astrand, M. and Speed, T.P., A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance, *Bioinformatics*, 19(2):185-193, 2003.

5 Bray, E.A., Molecular responses to water deficit, *Plant Physiol.*, 103:1035–1040, 1993.

6 Brivanlou, A. and Darnell, J., Signal transduction and the control of gene expression**.** *Science*, 295:813-818, 2002.

7 Busk, P.K. and Pages, M., Regulation of abscisic acid-induced transcription, *Plant Mol. Biol.*, 37:425-435, 1998.

8 Demeler, B. and Zhou, G., Neural Network Optimization for *E. coli* Promoter Prediction**,** *Nucleic Acids Research*, 19:1593-1599, 1991.

9 Higo, K., Ugawa, Y., Iwamoto, M. and Korenaga, T., Plant cis-acting regulatory DNA elements (PLACE) database, *Nucleic Acids Research*, 27:297-300, 1999.

10 Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B. and Speed, T.P., Summaries of Affymetrix GeneChip probe level data**,** *Nucleic Acids Research*, 31(4):e15, 2003.

11 Kilian, J., Whitehead, D., Horak, J., Wanke, D., Weinl, S., Batistic, O., D'Angel, C., Bauer, E.B., Kudla, J. and Harter K., The AtGenExpress global stress

expression dataset: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses**,** *The Plant Journal*, 50(2):347-363, 2007.

12 Mahadevan, I. and Ghosh, I., Analysis of *E. coli* Promoter Structures Using Neural Networks**,** *Nucleic Acids Research*, 22:2158-2165, 1994.

13 Markstein, M. et al., Genome-wide analysis of clustered Dorsal binding sitesidentifies putative target genes in the Drosophila embryo, *Proc. Natl Acad. Sci. USA*, 22:763–768. 2002.

14 Matys, V., TRANSFAC: transcriptional regulation, from patterns to profiles**,** *Nucleic Acids Res.*, 31:374-378, 2003.

15 Narusaka, Y., Nakashima, K., Shinwari, Z.K., Sakuma, Y., Furihata, T., Abe, H., Narusaka, M., Shinozaki, K. and Yamaguchi-Shinozaki, K., Interaction between two cis-acting elements, ABRE and DRE, in ABA-dependent expression of Arabidopsis rd29A gene in response to dehydration and high-salinity stresses, *Plant J*, 34:137-148, 2003.

16 Pedersen, A. G. and Nielsen H., Neural network prediction of translation initiation sites in eukaryotes, *ISMB*, 5:226-233, 1997.

17 Reese, M.G., Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome, *Comput Che.*, 26(1):51-56. 2001.

18 Rombauts, S., Déhais, P., Van Montagu, M. and Rouz,é P., PlantCARE, a plant cis-acting regulatory element database, *Nucleic Acids Ressearch*, 27(1):295-296, 1999.

19 Shah, N.H., King, D.C., Shah, P.N. and Fedoroff N.V., A tool-kit for cDNA microarray and promoter analysis, *Bioinformatics*, 19(14):1846-1848, 2003.

20 Shinozaki, K., and Yamaguchi-Shinozaki, K., Molecular responses to dehydration and low temperature: difference and cross-talk between two stress signaling pathways, *Current Opinion in Plant Biology*, 3:217-223, 2000.

21 Shinozaki, K., Yamaguchi-Shinozaki, K., and Seki, M., Regulatory network of gene expression in the drought and cold stress responses, *Current Opinionin Plant Biology*, 6:410-417, 2003.

22 Toufighi, K., Brady, S.M., Austin, R., Ly, E. and Provart, N.J., The Botany Array Resource: e-Northerns, Expression Angling, and promoter analyses, *Plant J.*, 43(1):153-63, 2005.

23 Ubramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. and Mesirov, J.P., From the Cover: Gene set enrichment analysis: A knowledge- based approach for interpreting genome-wide expression profiles, *Proceedings of the National Academy of Sciences*, 102(43):15545-15550, 2005.

24 Wenick, A. and Hobert, O., Genomic cis-regulatory architecture and trans-acting regulators of a single interneuron-specific gene battery in *C.elegans*., *Cell*, Dev. 6, 2004.

25 Wu, W.S., Li, W.H. and Chen, B.S., Computational reconstruction of transcriptional regulatory modules of the yeast cell cycle, BMC *Bioinformatics*, 7:421, 2006.

26 Xiong, L. and Zhu, J.K., Molecular and genetic aspects of plant responses to osmotic stress**,** *Plant, Cell and Environment*, 25:131-139, 2002.

27 Zhang, M., Computational prediction of eukaryotic protein-coding genes, *Nat. Rev. Genet.*, 3:698-709, 2002**.**
28 Zhang, W.X., Ruan, J.H., Ho, T.H., You, Y.S., Yu, T.T. and Quatrano, R.S., Cis-regulatory element based targeted gene finding: genome-wide identification of abscisicacid-and abiotic stress-responsive genes in *Arabidopsis thaliana*, *Bioinformatics*, 21(14):3074-3081, 2005.
29 A Database of Plant Cis-acting Regulatory DNA Elements: http://www.dna.affrc.go.jp/PLACE/.
30 BRB ArrayTools: http://linus.nci.nih.gov/BRB-ArrayTools.html.
31 Gene Ontology: http://www.geneontology.org/.
32 PlantCARE: http://bioinformatics.psb.ugent.be/webtools/plantcare/html/.
33 The Arabidopsis Information Resource (TAIR): http://www.arabidopsis.org.
34 The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology, *Nature Genet.*, 25:25-29, 2000.
35 TSSP-TCM software: http://mendel.cs.rhul.ac.uk/mendel.php?topic=genom
36 DAVID: http://david.abcc.ncifcrf.gov/
37 Wang H, Qi Q, Schorr P, Cutler AJ, Crosby W.L., and Fowke LC., ICK1, a cyclin-dependent protein kinase inhibitor from Arabidopsis thaliana interacts with both Cdc2a and CycD3, and its expression is induced by abscisic acid. *Plant Journal*, 15(4):501-10, 1998