Plant Physiology and Biochemistry xxx (2011) 1–4

Contents lists available at ScienceDirect

Plant Physiology and Biochemistry

journal homepage: www.elsevier.com/locate/plaphy

Research article

Prediction of regulatory interactions in *Arabidopsis* using gene-expression data and support vector machines

Xiaoqing Yu [a], Taigang Liu [b], Xiaoqi Zheng [a,c,*], Zhongnan Yang [d], Jun Wang [a,c]

[a] Department of Mathematics, Shanghai Normal University, Shanghai 200034, China
[b] College of Information Sciences and Engineering, Shandong Agricultural University, Taian 271018, China
[c] Scientific Computing Key Laboratory of Shanghai Universities, Shanghai 200234, China
[d] College of Life and Environmental Sciences, Shanghai Normal University, Shanghai 200234, China

ARTICLE INFO

Article history:
Received 23 March 2010
Accepted 2 January 2011

Keywords:
Arabidopsis
Regulatory relationships
Transcription factor
Gene-expression profile
SVM

ABSTRACT

Identification of regulatory relationships between transcription factors (TFs) and their targets is a central problem in post-genomic biology. In this paper, we apply an approach based on the support vector machine (SVM) and gene-expression data to predict the regulatory interactions in *Arabidopsis*. A set of 125 experimentally validated TF-target interactions and 750 negative regulatory gene pairs are collected as the training data. Their expression profiles data at 79 experimental conditions are fed to the SVM to perform the prediction. Through the jackknife cross-validation test, we find that the overall prediction accuracy of our approach achieves 88.68%. Our approach could help to widen the understanding of *Arabidopsis* gene regulatory scheme and may offer a cost-effective alternative to construct the gene regulatory network.

1. Introduction

The regulatory networks are biological interaction networks among the genes in a chromosome and the proteins. Understanding the transcriptional regulatory network is crucial to understand fundamental cellular processes, such as growth control, cell-cycle progression, and development, as well as differentiated cellular function [1]. In eukaryotes, an integrated regulatory network comprises transcription factors (TFs), target genes, and their relationships. Transcription factors are the key regulators of gene expression and play critical roles in the life cycle of higher plants [2,3]. They usually bind to the transcription factor binding sites (TFBSs) which are specific, short DNA sequence motifs of ~5–25 bp in the *cis*-regulatory region of a target gene (promoter and enhancer) to activate or repress its transcription in response to changes in the environment, as well as during development. Determination of the relationships between TFs and their target genes is a key step to construct the transcriptional regulatory networks.

With the advent of availability of large scale genome sequencing and high-throughput gene-expression analysis techniques, many efforts have been made to decipher the regulatory relationships using the computational tools in the past two decades. These methods are mainly classified into the following two categories. Methods in the first category are based on the assumption that coexpression of genes arises mainly from their transcriptional co-regulation. So given a set of coexpressed genes, one could retrieve their corresponding promoter sequences, and then find the statistically over-represented motifs [4–9]. After identification of the binding motif, target genes of a given TF can be obtained by scanning their corresponding promoter regions. Methods in the second category circumvent the motif discovery process. Since gene-expression profile data provide a direct measurement of the transcriptional program in the cell, researchers focus on mining gene-expression data to construct dynamical model to predict the regulatory relationship directly [10–15], such as through sequence similarity and structural comparisons [16–18].

As one of the model organisms, *Arabidopsis thaliana* is widely used for studying plant sciences, including genetics and plant development [19,20]. In this article, we aimed to predict the regulatory relationships between *Arabidopsis* genes and transcription factors using gene-expression profile data and the computer algorithm known as support vector machine. In the first step, 125 positive and 750 negative examples were constructed based on the experimentally validated regulatory relationships and randomly selected TF-target pairs under some strategies. Then, the positive and negative gene pairs characterized by their gene-expression data were put into the SVM to perform the prediction. The jackknife test

Abbreviations: TFs, Transcription factors; TFBSs, Transcription factor binding sites; SVM, Support vector machine; RBF, Radial Basis Function; S_n, Sensitivity; S_p, Specificity; *OA*, Overall accuracy; PCCs, Pearson's correlation coefficients.

* Corresponding author. Department of Mathematics, Shanghai Normal University, Shanghai 200034, China. Fax: +86 021 64324284.
E-mail address: xqzheng@shnu.edu.cn (X. Zheng).

0981-9428/$ – see front matter © 2011 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.plaphy.2011.01.002

showed that our prediction had the overall accuracy of 88.68%, while sensitivity and specificity reached 35.2% and 97.6%, respectively.

2. Materials and methods

2.1. Datasets

We downloaded the microarray expression data of *Arabidopsis* development from the TAIR database [21], and extracted the upstream promoter sequences up to 1000 bp for each gene from the genome-wide of *Arabidopsis*.

2.2. Reconstruction of expression profile data

Expression dataset is based on different tissues and different developmental stages in wild type Columbia (Col-0) and various mutants: flower and pollen, seedling and whole plant, shoots stems, siliques and seeds, leaves, and roots. There are 79 experimental conditions in all, and each condition is repeated 3 times by experiments. In our work, for each condition, the average expression value at 3 experiments was calculated and regarded as the final expression data. Finally, a 79-D vector was constructed to represent each gene.

2.3. Vector representation of gene pairs

In order to meet the requirement of our regulatory relationships prediction problem in a form suitable for training, we constructed TF-target pairs as follows. These pairs consist of a *known* transcription factor *T* and a putative target gene *G* that might be regulated by this factor *T*. To connect this gene pair with expression information, each gene in the pair was characterized by a set of expression experiments data, which was illustrated in above paragraph. Then the putative TF-target pair corresponded to a 158-D = 2 × 79-D gene-expression vector, in which the first 79 vector elements were for the TF while the rest 79 vector elements were for its putative target gene *G*.

2.4. Positive training example

We downloaded all 555 TF-target pairs whose regulatory relationships were confirmed by experiments from the Arabidopsis Gene Regulatory Information Server (AGRIS) database. There are 38 TFs (Table 1) included in these pairs. However, of all the 555 TF-target pairs, numbers of target genes regulated by different TFs vary significantly. For example, there are 216 TF-target pairs for TF HY5, but GL2 has only one confirmed target gene. If we chose all the 216 pairs of the TF HY5 and the one pair of the TF GL2, the training result would be possibly dominated by the HY5, and the effect of the GL2 would be underestimated. In order to tackle this imbalance problem and make sure that each TF-target pair contributes equally to the prediction results, we reduced the number of TF-target pairs for TFs like the HY5, and only retained up to 20 TF-target pairs in the final. In addition, we also removed those gene pairs which have no expression profile data. After the above processes, only 125 TF-target pairs were left as the positive examples.

2.5. Negative training example

As we know, reliable negative examples are important for machine learning. In contrast to the positive example, the negative example would be gene pairs that definitely have no regulatory interaction. However, up to our knowledge, there is no published literature on reporting definite negative regulatory relationships on *Arabidopsis*. In the present work, we used the following strategies to construct the negative examples. For the TF *T* whose TFBS was known, we searched for its binding site in the upstream sequence of all genes. If the promoter sequence of the target gene *G* contained no TFBS for transcription factor *T*, the pair (*T*, *G*) constituted a negative example. Then for the TF *T* whose TFBS was unknown, we randomly selected a gene *G* to construct a gene pair (*T*, *G*). To make sure the gene *G* was not regulated by *T*, we disorganized the expression profile of gene *G* by rearrangement of 79-D vector while keeping the expression profile of TF *T* unchanged. Note that there was a corresponding relationship for the number of gene pairs between positive examples and negative examples. For example, if TF *T* had *n* positive examples, we constructed 6*n* negative examples for TF *T* using the above process. In summary, we constructed 750 negative examples, which is 6 times the number of positive examples.

2.6. Support vector machine

The support vector machine was employed to train and predict our examples. The SVM is a form of supervised machine learning algorithm, which is based on recent developments in statistical learning theory [22]. It had been successfully used to tackle several biological problems, such as functional prediction, membrane gene classification, structural classification [23–25]. It performs binary classification problem by finding maximal margin hyperplanes in terms of a subset

Table 1
TFs in our study and their corresponding positive and negative TF-targets pairs.

Transcription factor	Positive pairs	Negative pairs	Overall pairs	Transcription factor	Positive pairs	Negative pairs	Overall pairs
HY5	14	84	98	FLO10	1	6	7
AT-HSFB2A	6	36	42	ANACOT2	1	6	7
PI	1	6	7	ANAC019	1	6	7
AP3	4	24	28	ATMYB123	1	6	7
ATBZIP53	1	6	7	AGLT	2	12	14
AT-TCP20	4	24	28	TT8	1	6	7
ATBZIP14	1	6	7	ANAC055	1	6	7
GL2	1	6	7	BLH9	1	6	7
ATBPC2	1	6	7	PAP3	5	30	35
ATBPC4	1	6	7	ATGL1	14	84	98
TGA3	1	6	7	PIF1	10	60	70
AHBP-1B	1	6	7	AGL9	15	90	105
LFY	11	66	77	ATWRKY44	1	6	7
AG	8	48	56	FUS3	2	12	14
CO	2	12	14	AGL25	2	12	14
AGL15	2	12	14	BP	2	12	14
PGA6	1	6	7	CCA1	1	6	7
AT-HSFC1	1	6	7	LHY	1	6	7
ATBZIP10	1	6	7	T7B11	1	6	7

of the input data (support vectors) between different classes. If the input data are not linearly separable, SVM first maps the data into a high dimensional feature space, and then classifies the data by the maximal margin hyperplanes. In our study, the LIBSVM package was used to implement the SVM classifier [26]. The Radial Basis Function (RBF) was chosen as the kernel function, which was defined as $K(x,x') = \exp(-\gamma|x-x'|^2)$. Two parameters, the regularization parameter C and the kernel width parameter γ were optimized on the training set using a grid search strategy in the LIBSVM.

3. Results

3.1. Evaluating the performance of our approach

In this work, 875 gene pairs, including 125 positive examples and 750 negative examples were collected to evaluate the performance of our method. We first transformed these gene pairs to fixed-length feature vectors through a map to the gene-expression levels in different experimental conditions. Then, these feature vectors were scaled and fed to the support vector machine to perform the prediction. We selected radial basis kernel function at $\gamma = 0.125$ and $C = 32{,}768$ to build the prediction models and get the optimal prediction accuracy. In such case, 62 examples were predicted as the positive examples and 813 examples were predicted as the negative examples. In detail, 44 of 125 positive examples and 732 of 750 negative examples were correctly predicted.

The prediction quality was examined by the jackknife test. Compared with the independent dataset test and sub-sampling test which are often adopted in statistic validation, the jackknife test is thought to be the most objective and effective one [27,28]. During the process of the jackknife test, each gene pairs in the dataset is singled out in turn as a test sample, and the predictor is trained by the remaining gene pairs. To evaluate the performance of the test, sensitivity (S_n), specificity (S_p) and the overall accuracy (OA) were calculated. Their definitions are as follows: $S_n = TP/(TP+FN)$, $S_p = TN/(TN+FP)$, $OA = (TP+TN)/(TP+\mathrm{TN}+FP+FN)$, where TP, TN, FP, and FN are defined as the number of true positive, true negative, false positive, and false negative obtained from the prediction respectively. We checked the obtained results of our prediction, and found that the different classes are basically equally distributed in the four groups of TP, TN, FP, and FN. After calculating, the overall prediction accuracy of our approach achieved 88.68%, and the sensitivity and specificity achieved 35.2% and 97.6%. It's worth mentioning that the sensitivity value appears relatively low in our prediction result, but this phenomenon is acceptable due to the following reasons: The ratio between positive and negative interactions is set to 1:6 in our experiment. So a higher specificity could give birth to a higher overall prediction accuracy. Moreover, the actual ratio between positive and negative interactions in *Arabidopsis* is much higher than 1:6. So considering the cost of experimental validation of TF-target interactions, it is more reasonable to keep the most likely regulatory gene pairs by setting a relatively higher specificity value.

3.2. Expression correlation coefficient

To further demonstrate the utility of our approach, we also checked the distribution of the Pearson's correlation coefficients (PCCs) between the known positive samples, negative samples and

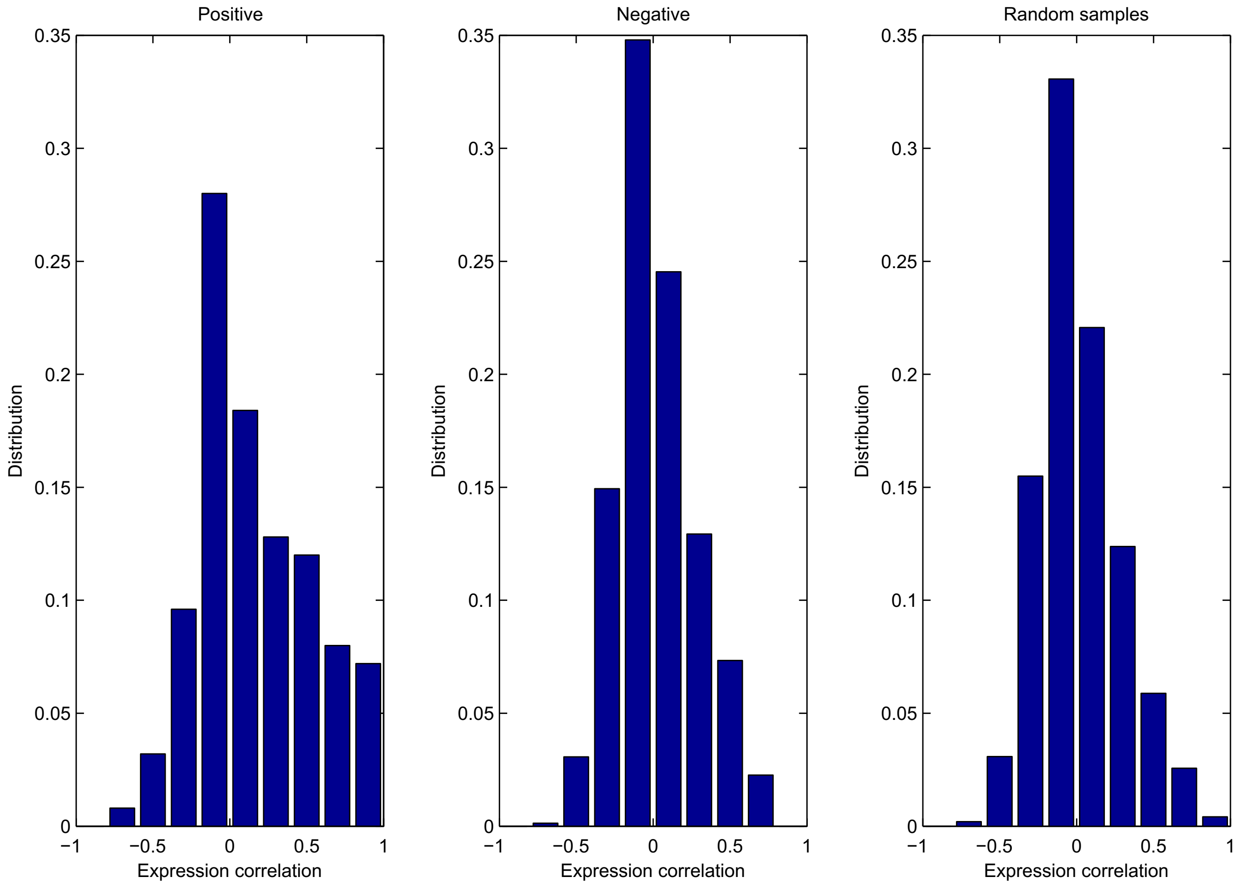

Fig. 1. Correlation coefficient distributions. We calculated the distribution of the PCCs between the known positive samples, negative samples and random selected samples, respectively.

random selected samples (which were composed by random combinations of 38 TFs and 788 targets), respectively. As it is shown in Fig. 1, PCCs of negative examples and randomly selected samples are approximately normally distributed, with the peak value of 0.35 and 0.33 at $-0.2 < \mathrm{PCC} < 0$, but the distribution of the positive examples has a fat tail. For example, only 9.6% negative examples and 8.86% random samples when $\mathrm{PCC} > 0.4$, but the value was 27.2% for positive examples. But on the whole, there is no significant difference between negative examples and randomly selected samples, which indicated that the construction of negative pairs was unbiased. In conclusion, although the PCC distribution of the positive and negative examples have some difference, it was difficult for us to detect the regulatory relationships between TFs and target genes based solely on their expression correlation. In other words, only when employing some other subtle strategies could we obtain a better prediction of the regulation relationships between TFs and their target genes, such as the SVM predictor employed in the present work.

4. Discussion

Inferring transcriptional interactions between TFs and their target genes has utmost importance for understanding the complex regulatory mechanisms in cellular systems. However, due to the lack of enough experimental data about transcription factor, especially for *Arabidopsis*, prediction of the regulatory relationships between transcription factors and their targets remains a complex and wonderful challenge for biologists. In this study, we provided a new method to predict the regulatory relationships between transcription factors and their targets in *Arabidopsis*. The result showed that our method achieved an overall accuracy of 88.68%, which demonstrated that our method is feasible.

However, the analyses and discussions of this study are preliminary and still have some deficiencies. First, at present, the number of the regulatory interactions validated by experiments is limited, which prevents a broader application of our approach. Second, the number of negative samples is much larger than that of the positive samples. Actually, for *Arabidopsis*, the ratio between number of negative TF-target pairs (with no regulatory relationship) and number of regulatory interactions may achieve 1000:1 or larger. But for a positive to negative ratio of 1:1000, an algorithm that always predicts negative will be correct 1000 times and incorrect only once, which will incur the imbalance problem and result in biased prediction in favor of the negative data [29,30]. To avoid this imbalance problem, in our experiment, we employed the under-sampling approach by reducing the size of the over-represented negative samples. We set the ratio of the positive to negative data at 1:6. Third, the gene-expression profile data (79 experiments) are used to represent a TF-target pair. This kind of representation could certainly capture the time-space relationships between TFs and their target genes, but fails to take into account the sequence information, especially the binding activities in the promoter sequences. In the future work, we will try more information to represent the TF-target pairs, e.g., component information of their nucleotide or promoter sequences, structural domain information and so on. We believe that our approach will be more useful with the growing amount of the validated regulatory relationships, microarray data, and knowledge of transcription factors in *Arabidopsis*.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (No.10731040), Shanghai Leading Academic Discipline Project (No. S30405) and Innovation Program of Shanghai Municipal Education Commission (No. 09zz134).

Reference

[1] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, J. Watson, Molecular Biology of the Cell. Garland Publishing, New York, 1994.

[2] W. Gong, Y.P. Shen, L.G. Ma, et al., Genome-wide ORFeome cloning and analysis of *Arabidopsis* transcription factor genes, Plant Physiol. 135 (2004) 773–782.

[3] D.M. Riańo, S. Ruzicic, I. Dreyer, B. Mueller-Roeber, PlnTFDB: an integrative plant transcription factor database, BMC Bioinformatics 8 (2007) 42.

[4] C.E. Lawrence, S.F. Altschul, M.S. Boguski, J.S. Liu, A.F. Neuwald, J.C. Wootton, Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignments, Science 262 (1993) 208–214.

[5] J.L. DeRisi, V.R. Iyer, P.O. Brown, Exploring the metabolic and genetic control of gene expression on a genomic scale, Science 278 (1997) 680–686.

[6] J. van Helden, B. Andre, J. Collado-Vides, Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies, J. Mol. Biol. 281 (1998) 827–842.

[7] J. Vilo, A. Brazma, I. Jonassen, A. Robinson, E. Ukkonen, Mining for putative regulatory elements in the yeast genome using gene expression data, Proc. Int. Conf. Intell. Syst. Mol. Biol. 8 (2003) 384–394.

[8] K. Doi, A. Hosaka, T. Nagata, et al., Development of a novel data mining tool to find *cis*-elements in rice gene promoter regions, BMC Plant Biol. 8 (2008) 20.

[9] X.Q. Zheng, T.G. Liu, Z.N. Yang, J. Wang, Large cliques in *Arabidopsis* gene coexpression network and motif discovery, J. Plant Physiol. (2010). doi:10.1016/j.jplph.2010.09.010.

[10] I. Nachman, A. Regev, N. Friedman, Inferring quantitative models of regulatory networks from expression data, Bioinformatics (Oxford, England) 20 (Suppl. 1) (2004) i248–i256.

[11] D. Pe'er, A. Regev, G. Elidan, N. Friedman, Inferring subnetworks from perturbed expression profiles, Bioinformatics (Oxford, England) 17 (Suppl. 1) (2001) S215–S224.

[12] J. Qian, J. Lin, N.M. Luscombe, H. Yu, M. Gerstein, Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data, Bioinformatics (Oxford, England) 19 (2003) 1917–1926.

[13] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, N. Friedman, Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data, Nat. Genet. 34 (2003) 166–176.

[14] X. Xu, L. Wang, D. Ding, Learning module networks from genome-wide location and expression data, FEBS Lett. 578 (2004) 297–304.

[15] G. Karlebach, R. Shamir, Modelling and analysis of gene regulatory networks, Nat. Rev. Mol. Cell Biol. 9 (2008) 770–780.

[16] E. Wingender, X. Chen, E. Fricke, R. Geffers, R. Hehl, I. Liebich, M. Krull, V. Matys, H. Michael, R. Ohnhäuser, M. Prűss, F. Schacherer, S. Thiele, S. Urbach, The TRANSFAC system on gene expression regulation, Nucleic Acids Res. 29 (2001) 281–283.

[17] J.L. Riechmann, J. Heard, G. Martin, L. Reuber, C.Z. Jiang, J. Keddie, L. Adam, O. Pineda, O.J. Ratcliffe, R.R. Samaha, R. Creelman, M. Pilgrim, P. Broun, J.Z. Zhang, D. Ghandehari, B.K. Sherman, G.L. Yu, *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes, Science 290 (2000a) 2105–2110.

[18] J.L. Riechmann, J. Heard, G. Martin, L. Reuber, C.Z. Jiang, J. Keddie, L. Adam, O. Pineda, O.J. Ratcliffe, R.R. Samaha, R. Creelman, M. Pilgrim, P. Broun, J.Z. Zhang, D. Ghandehari, B.K. Sherman, G.L. Yu, *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes, Science 290 (2000b) 2105–2110.

[19] W.A. Rensink, C.R. Buell, *Arabidopsis* to rice. Applying knowledge from a weed to enhance our understanding of a crop species, Plant Physiol. 135 (2) (2004) 622–629.

[20] S.M. Coelho, A.F. Peters, B. Charrier, D. Roze, C. Destombe, M. Valero, J.M. Cock, Complex life cycles of multicellular eukaryotes: new approaches based on the use of model organisms, Gene 406 (1–2) (2007) 152–170.

[21] http://www.arabidopsis.org/portals/expression/microarray/ATGenExpress.jsp.

[22] V. Vapnik, Statistical Learning Theory. Wiley, New York, 1998.

[23] C.Z. Cai, L.Y. Han, Z.L. Ji, X. Chen, Y.Z. Chen, SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence, Nucleic Acids Res. 31 (2003) 3692–3697.

[24] Y.D. Cai, X.J. Liu, X.B. Xu, K.C. Chou, Prediction of protein structural classes by support vector machines, Comput. Chem. 26 (2002) 293–296.

[25] L. Zhang, B. Liao, D.C. Li, W. Zhu, A novel representation for apoptosis protein subcellular localization prediction using support vector machine, J. Theor. Biol. 259 (2009) 361–365.

[26] C.C. Chang, C.J. Lin, LIBSVM: A Library for Support Vector Machines Software Available at: (2001) http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[27] P.E. Jupp, K.V. Mardia, Maximum likelihood estimators for the matrix von Mises-Fisher and Bingham distributions, Ann. Stat. 7 (1979) 599–606.

[28] Y.H. Zeng, Y.Z. Guo, R.Q. Xiao, L. Yang, L.Z. Yu, M.L. Li, Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach, J. Theor. Biol. 259 (2009) 366–372.

[29] J.N. Song, K. Burrage, Z. Yuan, T. Huber, Prediction of cis/trans isomerization in proteins using PSI-BLAST profiles and secondary structure information, BMC Bioinformatics 7 (2006) 124.

[30] L.J. Wee, T.W. Tan, S. Ranganathan, SVM-based prediction of caspase substrate cleavage sites, BMC Bioinformatics 7 (Suppl. 5) (2006) S14–S15.

Please cite this article in press as: X. Yu, et al., Prediction of regulatory interactions in *Arabidopsis* using gene-expression data and support vector machines, Plant Physiology and Biochemistry (2011), doi:10.1016/j.plaphy.2011.01.002