BMC Bioinformatics

BioMed Central

Research

Computational identification of condition-specific miRNA targets based on gene expression profiles and sequence information

Je-Gun Joung[1] and Zhangjun Fei*[1,2]

Address: [1]Boyce Thompson Institute for Plant Research, Cornell University, Ithaca, NY 14853, USA and [2]USDA Robert W. Holley Center for Agriculture and Health, Ithaca, NY 14853, USA

Email: Je-Gun Joung - jj294@cornell.edu; Zhangjun Fei* - zf25@cornell.edu

* Corresponding author

from The Seventh Asia Pacific Bioinformatics Conference (APBC 2009)
Beijing, China. 13–16 January 2009

Published: 30 January 2009

BMC Bioinformatics 2009, **10**(Suppl 1):S34 doi:10.1186/1471-2105-10-S1-S34

This article is available from: http://www.biomedcentral.com/1471-2105/10/S1/S34

Abstract

Background: MicroRNAs (miRNAs) are small and noncoding RNAs that play important roles in various biological processes. They regulate target mRNAs post-transcriptionally through complementary base pairing. Since the changes of miRNAs affect the expression of target genes, the expression levels of target genes in specific biological processes could be different from those of non-target genes. Here we demonstrate that gene expression profiles contain useful information in separating miRNA targets from non-targets.

Results: The gene expression profiles related to various developmental processes and stresses, as well as the sequences of miRNAs and mRNAs in *Arabidopsis*, were used to determine whether a given gene is a miRNA target. It is based on the model combining the support vector machine (SVM) classifier and the scoring method based on complementary base pairing between miRNAs and mRNAs. The proposed model yielded low false positive rate and retrieved condition-specific candidate targets through a genome-wide screening.

Conclusion: Our approach provides a novel framework into screening target genes by considering the gene regulation of miRNAs. It can be broadly applied to identify condition-specific targets computationally by embedding information of gene expression profiles.

Background

MicroRNAs (miRNAs) are small RNAs that play important regulatory roles in animals and plants [1]. They cause transcriptional cleavage or translational repression through binding their target mRNAs. miRNAs affect a variety of cellular processes such as development, cell proliferation, apoptosis, and stress response [2-4]. Thus identification of mRNA targets is an essential step to understand miRNA functions.

Currently several miRNA target prediction tools have been developed [1,5-10]. The majority of these algorithms are based on the sequence alignment or the minimum free energy of the hybridization. The sequence alignment or the binding energy of miRNA/mRNA pairs can sometimes

hold definitive information in screening target genes. However, a number of candidate targets could be false positives due to the omission of gene expression information in the screening process.

Microarray analysis allows us to observe a number of target mRNAs down-regulated by overexpressing miRNAs [11]. Expression profiles may be useful in identifying miRNA targets that have been missed or mis-identified by the sequence analysis [12]. However, it is labor intensive to generate miRNA over-expression lines and gene expression profiles in these lines. Furthermore, it is difficult to generate gene expression profiles in diverse tissues, stages, and environments of transgenic lines due to the high cost. For these reasons, currently available gene expression profiles generated without performing the transfection experiment may also be useful sources for identifying target genes.

In this paper, we propose a novel approach for screening miRNA targets by considering gene expression profiles. Our approach is based on the model combining a machine learning tool, SVM, which uses the datasets of gene expression profiles, and a scoring method, which uses the sequences of miRNAs and mRNAs. SVM can identify unknown targets by using a kernel function that describes the similarity between given input examples. SVM was developed by Vapnik for classification of data based on statistical learning theory [13]. It has provided a number of applications in biological data analysis, including the classification of cancers, splice site identification, and the classification of protein folding [14-16]. In the present study, by employing the classifier, we first investigate whether the expression profiles in specific biological processes contain enough information for the prediction of miRNA targets. Then the properties of the combined model are analyzed and the model is applied to the genome-wide target screening.

Our method was analyzed with a validated target set, gene expression profiles and gene sequences in *Arabidopsis*. The validated target sets were collected from several literature sources that describe the experimentally verified target genes. The gene expression dataset was generated with a total of 211 conditions including different developmental series and stress treatments [17]. The ability of the SVM classifier to discriminate between target and non-target genes was analyzed using only the gene expression dataset, and then several major conditions relevant to the classification were extracted using a feature selection method. Finally, we performed the target prediction using the method combining both express profiles and sequence information. Our study suggests that gene expression profile information can be combined with other miRNA target prediction algorithms to identify targets involved in specific biological processes.

Methods

SVM classifier

A supervised machine-learning algorithm, support vector machine (SVM), was used to classify miRNA targets from non-targets. Recently SVM has been successfully applied to miRNA predictions as well as miRNA target predictions [18,19]. Given a kernel and a set of labeled training examples belonging to positives or negatives, SVM learns a linear decision boundary in the feature space defined by the kernel function in order to discriminate between the two classes. Then, given any unlabeled example, SVM determines whether it is positive or negative, depending on the position of its image in the feature space relative to the linear boundary. In our case, using a training set containing known verified targets and non-targets, SVM builds a model for the prediction of the test set, i.e., the unknown set. In this study we used LIBSVM, a library for support vector machines [20]. The input features of SVM are expression profiles. A training or test set is represented by $D \in \{x_i, y_i\}_{i=1}^{N}$, $x_i = (x_{i1},..., x_{im})$ = and $y_i \in \{-1, 1\}$, where x_i is a vector of expression ratios under different conditions from a gene i. If $y_i = 1$, then the i-th gene represents a target gene, otherwise it represents a non-target gene.

Dataset construction

A number of putative targets have been predicted from sequence analysis in previous studies. However, the predicted targets should contain a small portion of false positives. Therefore, in the present study, we used only a list of ~100 experimentally validated targets as the true positive set. Nevertheless, it is challenging to make a proper training dataset for the construction of a SVM model because of the imbalance issue in machine learning [21]: the size of the validated target set is much smaller than that of the set containing all the genes excluding the validated targets. To overcome this imbalance problem, we increased the size of the validated target set through random resampling. After we increased the size of the positive dataset by a predefined number, which we set to 1,000, we constructed the negative dataset of which the size is the same as the size of the positive data through random sampling.

Dataset of gene expression profiles

Two expression datasets were used for miRNA target prediction. The first microarray dataset contains 79 different conditions derived from several developmental series in *Arabidopsis*. The second dataset contains 132 conditions from ten different stress treatments including light, cold,

drought, genotoxic, heat, osmotic, oxidative, salt, UV-B, and wound. Affymetrix CEL files of the gene expression datasets were obtained from the Nottingham Arabidopsis Stock Centre (NASC; [22]). Both datasets were generated using the ATH1 genome array containing ~22,800 probe sets. The CEL files were processed and normalized at the probe level using the GC content based robust multi-array algorithm (GCRMA; [23]). After normalization, the average of the triplicate values was calculated for each sample. In the development dataset, the relative expression level of each gene was calculated by taking the log ratio between each expression level and the mean expression level across all the samples. The stress dataset was processed by taking the log ratio between the expression level of treatments and that of the corresponding normal cell types.

Binding scoring between miRNA and mRNA

The most recent collection of *Arabidopsis* miRNAs in miRBase (Release 11.0; [24]) and mRNA sequences from the TAIR database [25] were obtained. Given a miRNA, the sequence alignment of the miRNA against all mRNAs was performed. The binding scoring function between miRNA and mRNA is based on the weighted summation of the numbers of mismatches, wobbles and indels described in Jones-Rhoades and Bartel [26].

Combining gene expression profiles and binding information

Our target prediction strategy is based on the gene expression profiles and the binding scores between miRNA and mRNA sequences. Figure 1 shows the overall procedure of computational prediction of condition-specific miRNA targets. The prediction system consists of two parts: the SVM classifier and the binding scoring function. The expression profiles of the validated miRNA targets were used as the training dataset for modeling SVM. Then the test set is predicted by making a decision between the output of SVM and that of the scoring function. When an input gene in both outputs is indicated as a positive, it is predicted as a miRNA target.

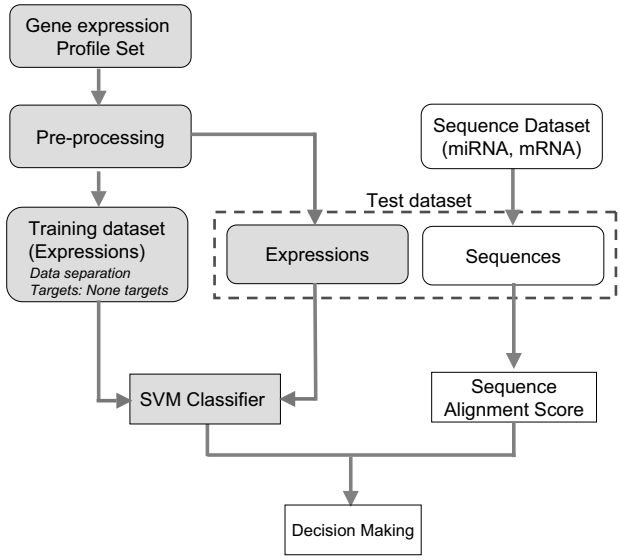

Figure 1
The procedure of computational prediction of miRNA targets. After the training dataset of gene expression files is trained by SVM, the test set is predicted by the decision making of SVM classifiers and the scoring method based on the sequence alignment.

Results

Classification of miRNA targets using gene expression profiles

Our prediction model classifies the targets by combining gene expression profiles and sequence information (Figure 1). Before testing the prediction model, we first investigated whether gene expression profile information can be used to discriminate the target genes from non-target genes. We applied SVM to classify target genes from non-target genes. The procedure is highlighted in gray in Figure 1. The classification is only based on patterns of gene expression between the target set and the non-target set in specific conditions. The type of SVM used is C-SVM and the type of kernel used is a linear kernel function. The gene expression dataset contains a total of 211 conditions, including 79 conditions derived from several developmental series and 132 conditions from diverse stress treatments. It has been reported that miRNAs affect the expression of a number of target genes involved in different developmental processes and stresses. We expect that both the developmental series dataset and the stress dataset are informative enough to discriminate targets from non-targets.

To achieve a good classification, it is important to define true miRNA target genes. We collected the experimentally validated miRNA targets to construct a highly accurate training dataset. The true target genes were extracted from several literature sources describing experimentally validated miRNA targets [12,27-30]. A total of 101 non-redundant target genes were collected (Additional file 1). Eighty-nine of them overlap with those in the expression dataset. 60% of these genes (53 genes) were used as the positive examples of the training dataset and the rest (36 genes) were used as the positives of test dataset. 1,000 negative examples were randomly selected from all the genes on the array excluding the validated target genes. The positive examples were increased by the number of negative examples through random re-sampling in order

to keep a balance (1:1 ratio) between the size of the positive dataset and that of the negative dataset.

We investigated the prediction accuracies of using target datasets with different qualities for classification: validated, putative, and random sets (Figure 2). The validated dataset is the same as the dataset described above. The putative dataset contains 378 targets collected from several reports which were identified through computational screening [7,8,26,31], of which 328 overlapped with those in the expression dataset. The positive and training and test sets were generated using the expression profiles of these 328 putative target genes while the negative training and test sets were generated by randomly selecting genes excluding those 328 target genes. The dataset of random targets was generated by random assignment of positive or negative labels in order to observe the baseline of prediction. The size of these three datasets is identical through random re-sampling of positive examples. As expected, the target genes could be classified by prediction using only gene expression dataset. The prediction accuracy is lower when the putative target dataset is used than when the validated target dataset is used (Figure 2).

We then performed the analysis to determine which expression datasets can be used to classify the genes more accurately. Our results indicated that no significant difference regarding the specificity and the sensitivity was found between the two datasets: the developmental dataset and the stress dataset, as well as the combined dataset (Figure 3).

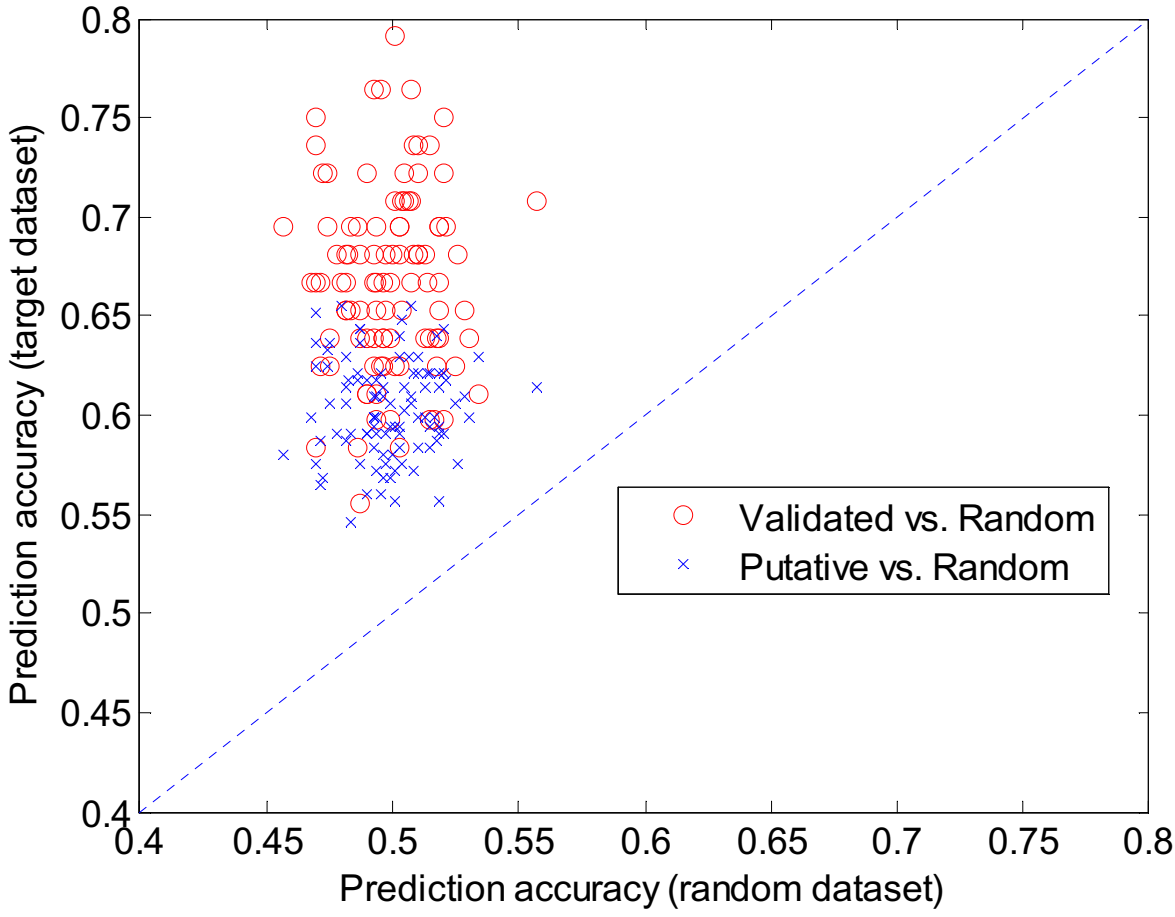

Figure 2
The miRNA target prediction with SVM using the gene expression dataset. Three datasets with different qualities, which were the validated target dataset, the putative target dataset, and the random dataset, were compared in terms of the prediction accuracy.

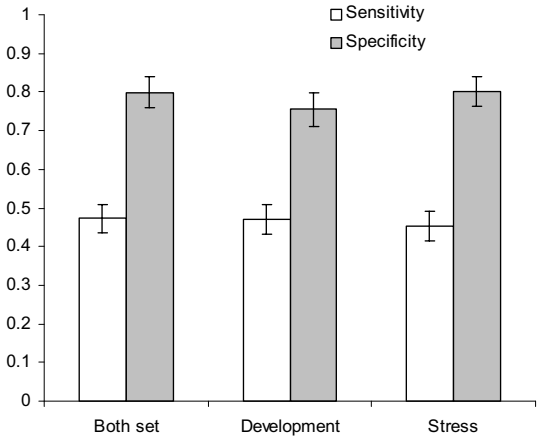

Figure 3
The performance of target prediction with SVM using different gene expression sets. The performance of target prediction with SVM using developmental- or stress-related gene expression profiles, or combined expression profiles from the two datasets.

Table 1: Feature selection in the gene expression dataset.

Rank	Sample ID	Type	Tissue
1	ATGE26	D	Leaf
2	Heat-Shoots-1.0 h	S	Shoot
3	UV-B-Roots-1.0 h	S	Root
4	ATGE73	D	Pollen
5	ATGE91	D	Leaf
6	ATGE34	D	Flower
7	Drought-Roots-0.25 h	S	Root
8	Drought-Shoots-0.25 h	S	Shoot
9	UV-B-Shoots-3.0 h	S	Shoot
10	Cold-Roots-24 h	S	Root
11	ATGE97	D	Seedling
12	Drought-Roots-24.0 h	S	Root
13	UV-B-shoots-0.5 h	S	Shoot
14	ATGE55	D	Flower
15	ATGE101	D	Seedling
16	Drought-Shoots-3.0 h	S	Shoot
17	Wounding-Shoots-6.0 h	S	Shoot
18	Osmotic-Shoots-1.0 h	S	Shoot
19	Oxidative-Roots-6.0 h	S	Root
20	UV-B-Roots-6.0 h	S	Root

The top 20 ranked features for miRNA target classification. Each feature corresponds to a condition in the two expression datasets (D: developmental process and S: stress treatment).

We then determined which features in the expression datasets are important for the classification. The ranker search method using SVM was used to select the features. The list of the top ranked 20 features is shown in Table 1. The rank was determined by 10-fold cross validation with the training dataset, which is consisted of the validated targets (positive) and the randomly selected genes excluding the validated targets. The features from the developmental dataset and the stress dataset are highly ranked without significant disproportion, further confirming that there is no significant difference of performance between the two datasets. The full list of ranked features is shown in Additional file 2.

Classification of miRNA targets using gene expression profiles and sequence information

We then compared the efficiencies of target prediction between two different methods: the method using the combined information of expression profiles and sequence information (SVM+SC) and the method using the sequence information alone (SC). The results are shown in Table 2. SVM+SC_3 indicates our method combining SVM classifier and SC, the scoring method based on the weighted summation of the numbers of mismatches, as well as wobbles and indels between miRNA and mRNA as described in Jones-Rhoades and Bartel [26], with 3.0 as the cutoff score. SC_1 indicates the scoring method with a cutoff score of 1.0. TP, FP, TN and FN are the true positive, false positive, true negative, and false negative, respectively. The precision is a positive predictive value calculated by TP/(TP+FP). The sensitivity and the specificity are calculated as TP/(TP + FN) and TN/(TN + FP), respectively. The sensitivity of SVM+SC_3 is higher than that of SC_1, whereas its specificity is higher than that of SC_3. Although the false positive rate of SC_1 achieves zero, which is the same as that of SVM+SC_3, the true positive rate is much lower. SC_3 can predict more true positives than SVM+SC_3, but it contains more false positives. These results suggest that the information of gene expression profiles can be utilized to increase the efficiency of miRNA target gene prediction when combined with sequence information.

Table 2: Comparison of predictions using different methods.

	SVM+SC_3	SC_1	SC_3
TP (True Positive) rate	0.36	0.20	0.83
FP (False Positive) rate	0.00	0.00	0.03
TN (True Negative) rate	1.00	1.00	0.97
FN (False Negative) rate	0.64	0.80	0.17
Sensitivity (TP/(TP+FN))	0.36	0.20	0.83
Specificity (TN/(TN+FP))	1.00	1.00	0.97
Precision (TP/(TP+FP))	1.00	1.00	0.97

SVM+SC_3, the method combining the SVM classifier and the scoring method based on the sequence matches. SC_τ indicates the score cutoff, τ. The results were obtained with 100 test sets.

Genome-wide identification of miRNA target genes associated with developmental processes and stress responses

We extracted the target genes identified by our classifier ($SVM+SC_3$) excluding those that have been validated in *Arabidopsis*. The training dataset was generated as described in the previous section. Since the classification is dependent on the expression dataset, these targets may be involved in the corresponding biological process. The top 20 ranked genes predicted as the development-related and stress-related targets are listed in Tables 3 and 4, respectively. A number of genes retrieved by the classifier have reported roles in the corresponding developmental processes and stress responses, while the functions of most targets we identified are not clear.

Developmental-related miRNA targets

AGO7/ZIPPY (At1g69440), a member of the Argonaute family, plays a role in the *TAS3* ta-siRNA pathway. *TAS3* ta-siRNAs are required for proper leaf development through the action of AGO7 [32]. SPL5 (At3g15270) and SPL9 (At2g42200) are the members of the SQUAMOSA PROMOTER BINDING PROTEIN-LIKE (SPL) family of transcription factors. Increased expression of *SPL5*, together with two other genes from the same family, *SPL3* and *SPL4*, promotes vegetative phase change and flowering, and the decreased level of *miR156* during juvenile-to-adult transition is responsible for this increase [33]. *SPL3* and *SPL4* are the validated targets that belong to our training dataset. *SPL9* is also regulated by *miR156* and acts redundantly with *SPL15* in controlling shoot maturation [34]. *AtREM1* (At4g31610) encodes a protein with features of transcriptional activators and its deduced protein contains three repetitions of a B3-related DNA-binding domain. It may play a role in the organization of reproductive meristems, as well as during flower organ development [35]. *NTT* (*NO TRANSMITTING TRACT* ; At3g57670) encodes a C2H2/C2HC zinc finger transcription factor specifically expressed in the transmitting tract. Mutations in *NTT* cause reduced fertility by severely inhibiting pollen-tube movement [36].

Stress-related miRNA targets

At1g74840 encodes a protein belonging to the myb family of transcription factors and responds to the $CdCl_2$ and NaCl treatments [37]. BIT1 (At2g36890), also a MYB transcription factor, plays an important role in controlling blue light responses [38].

Discussion

In this study we presented a novel method for screening miRNA targets that are likely to be involved in specific biological processes. Currently, several computational algorithms for miRNA target prediction have been implemented and the majority of them use properties such as the hybridization based on sequence base pairing between miRNA and mRNA or the minimum free energy. Although computational screening has identified a large number of putative miRNA targets, only a small portion of the targets can be validated. In addition, these computational tools do not imply which biological processes might be correlated with the targets. One advantage of our

Table 3: Top 20 target genes associated with the developmental series.

Locus ID	miRNA	Rate	Description
At1g69440*	miR854	0.97	Encodes ARGONAUTE7
At1g62930	miR400	0.83	Similar to pentatricopeptide (PPR) repeat-containing protein
At5g47250	miR472	0.82	Disease resistance protein
At3g15270*	miR156	0.78	Squamosa promoter-binding protein-like 5
At5g59000	miR414	0.77	Zinc finger family protein
At4g31610*	miR414	0.77	REM1 (Reproductive Meristem 1) transcription factor
At5g58980	miR396	0.77	Ceramidase family protein
At5g43730	miR472	0.76	Disease resistance protein
At4g15430	miR855	0.72	Similar to early-responsive to dehydration protein-related
At5g08430	miR414	0.70	SWIB complex BAF60b domain-containing protein/plus-3 domain-containing protein
At2g28510	miR829	0.69	Dof-type zinc finger domain-containing protein
At5g48560	miR778	0.69	Basic helix-loop-helix (bHLH) family protein
At1g27360	miR156	0.68	Squamosa promoter-binding protein-like 11
At3g53310	miR414	0.65	Transcriptional factor B3 family protein
At2g42200*	miR156	0.64	Squamosa promoter-binding protein-like 9
At1g63130	miR400	0.62	Transacting siRNA generating locus
At3g20910	miR169	0.62	CCAAT-binding transcription factor
At2g34960	miR157	0.61	Encodes a member of the cationic amino acid transporter
At1g62670	miR161	0.61	Pentatricopeptide (PPR) repeat-containing protein
At3g57670*	miR854	0.57	Similar to zinc finger

The targets were predicted with the expression dataset of the developmental series. The rate indicates the fraction of runs in which the gene was predicted as a positive in 200 runs. * indicates the gene reported to be involved in the developmental process.

Table 4: Top 20 target genes associated with stress responses.

Locus ID	miRNA	Rate	Description
At5g43760	miR854	0.88	A member of the 3-ketoacyl-CoA synthase family involved in the biosynthesis of VLCFA
At5g47250	miR472	0.79	Disease resistance protein
At3g20710	miR859	0.79	F-box/Kelch-repeat protein
At2g36890*	miR847	0.60	Myb-like transcription factor MYB38
At4g28310	miR837-5p	0.60	Unknown protein
At5g41410	miR414	0.55	Homeodomain protein required for ovule identity
At2g25980	miR846	0.53	Jacalin lectin family protein
At5g57590	miR396	0.52	Mutant complemented by E coli Bio A gene encoding 7,8-diaminopelargonic acid aminotransferase
At1g49750	miR854	0.47	Leucine-rich repeat family protein
At5g39710	miR400	0.47	Similar to pentatricopeptide (PPR) repeat-containing protein
At3g13690	miR419	0.47	Protein kinase family protein
At3g18980	miR859	0.45	F-box family protein
At5g43730	miR472	0.45	Disease resistance protein
At2g32760	miR414	0.43	Unknown protein
At1g74840*	miR863-5p	0.43	Myb family transcription factor
At1g80340	miR835-5p	0.42	Encodes a protein with gibberellin 3 β-hydroxylase activity
At1g26210	miR414	0.41	unknown protein
At2g17830	miR859	0.41	F-box family protein
At4g14680	miR395	0.40	ATP sulfurylase
At5g61480	miR870	0.38	Leucine-rich repeat transmembrane protein kinase

The targets were predicted with the expression dataset of stress treatments. The rate indicates the fraction of runs in which the gene was predicted as a positive in 200 runs. * indicates the gene reported to be involved in the stress responses.

method, by using gene expression profile information, is that it can suggest which target genes have highest priorities to be involved in a specific biological process.

If gene expression profiles of transgenic lines with increased miRNA expression are available, it is possible to do high-throughput and more accurate screening of targets [39]. As the under expressed genes are extracted, putative targets can be defined and the set overlapped with computationally predicted targets can be obtained.

Unfortunately, this kind of high-throughput expression profile dataset is difficult to generate due to the high cost and the labor-intensive experimental process. However, currently many expression profile datasets, which were generated without the context of miRNA are available in the public domains for several organisms. This expression profile information could be a valuable source for miRNA target prediction. Although exclusively using gene expression profiles for prediction does not show very good performance, our results indicate that utilization of expression profiles combined with sequence information can identify condition-specific targets and compensate for the limitations of current sequenced based methods.

We identified miRNA target genes associated with the developmental processes and stress responses at the genomic scale using our proposed method. Our results are supported by previous reports indicating that several genes we identified are involved in the corresponding biological processes. However, the biological functions of most target genes are still largely undetermined. The genes ranked with high priorities in developmental processes or stress responses could be the candidates for further studies in terms of gene regulation. We expect that our application alleviates experimental efforts as it suggests novel candidates with high confidence.

Our method provides a framework for identifying miRNA targets involved in specific conditions. It can be applied to diverse gene expression datasets including cancers, diseases, and other species of which the validated target information is sufficient for training the SVM classifier. Since the free energy for miRNA-target duplex is important to predict the targets in animals, it is possible to combine our method with the method using the minimum free energy of hybridization to improve target prediction and to identify condition-specific targets. Consequently, our approach could contribute to elucidation of gene regulatory programs related to miRNAs and their target genes in diverse biological processes.

Conclusion

Our results suggested that the gene expression profiles related to specific conditions have the potential to discriminate miRNA targets from non-targets. The combination of gene expression and sequence-based methods ensures retrieval of true targets and targets related to specific biological process. We have shown that in *Arabidopsis* the targets related to the biological processes of develop-

ments and stresses were successfully extracted by the proposed method. The same framework can be applied to other biological processes or species.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JGJ proposed the idea, organized overall procedure, built the dataset for computational experiments and carried out the analysis. ZF developed the idea, provided intellectual guidance and mentorship. All authors read and approved the final manuscript.

Additional material

Additional file 1

List of validated and putative targets.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-S1-S34-S1.xls]

Additional file 2

Ranked list of features selected by the ranker search method using SVM.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-S1-S34-S2.xls]

Acknowledgements

This work was supported by the National Science Foundation (DBI-0501778 to ZF).

This article has been published as part of *BMC Bioinformatics* Volume 10 Supplement 1, 2009: Proceedings of The Seventh Asia Pacific Bioinformatics Conference (APBC) 2009. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/10?issue=S1

References

1. Bartel DP: **MicroRNAs: genomics, biogenesis, mechanism, and function.** *Cell* 2004, **116(2):**281-297.
2. Hwang HW, Mendell JT: **MicroRNAs in cell proliferation, cell death, and tumorigenesis.** *Br J Cancer* 2007, **96(Suppl):**R40-44.
3. Jovanovic M, Hengartner MO: **miRNAs and apoptosis: RNAs to die for.** *Oncogene* 2006, **25(46):**6176-6187.
4. Sunkar R, Zhu JK: **Novel and stress-regulated microRNAs and other small RNAs from Arabidopsis.** *Plant Cell* 2004, **16(8):**2001-2019.
5. Kruger J, Rehmsmeier M: **RNAhybrid: microRNA target prediction easy, fast and flexible.** *Nucleic Acids Res* 2006:W451-454.
6. Krek A, Grun D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M, *et al.*: **Combinatorial microRNA target predictions.** *Nat Genet* 2005, **37(5):**495-500.
7. Zhang Y: **miRU: an automated plant miRNA target prediction server.** *Nucleic Acids Res* 2005:W701-704.
8. Wang XJ, Reyes JL, Chua NH, Gaasterland T: **Prediction and identification of Arabidopsis thaliana microRNAs and their mRNA targets.** *Genome Biol* 2004, **5(9):**R65.
9. Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB: **Prediction of mammalian microRNA targets.** *Cell* 2003, **115(7):**787-798.
10. Rhoades MW, Reinhart BJ, Lim LP, Burge CB, Bartel B, Bartel DP: **Prediction of plant microRNA targets.** *Cell* 2002, **110(4):**513-520.
11. Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM: **Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs.** *Nature* 2005, **433(7027):**769-773.
12. Jones-Rhoades MW, Bartel DP, Bartel B: **MicroRNAS and their regulatory roles in plants.** *Annu Rev Plant Biol* 2006, **57:**19-53.
13. Vapnik V: **Statistical Learning Theory.** *Wiley, New York*; 1998.
14. Shamim MT, Anwaruddin M, Nagarajaram HA: **Support Vector Machine-based classification of protein folds using the structural properties of amino acid residues and amino acid residue pairs.** *Bioinformatics* 2007, **23(24):**3320-3327.
15. Baten AK, Chang BC, Halgamuge SK, Li J: **Splice site identification using probabilistic parameters and SVM classification.** *BMC Bioinformatics* 2006, **7(Suppl 5):**S15.
16. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D: **Support vector machine classification and validation of cancer tissue samples using microarray expression data.** *Bioinformatics* 2000, **16(10):**906-914.
17. Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU: **A gene expression map of Arabidopsis thaliana development.** *Nat Genet* 2005, **37(5):**501-506.
18. Sheng Y, Engstrom PG, Lenhard B: **Mammalian microRNA prediction through a support vector machine model of sequence and structure.** *PLoS ONE* 2007, **2(9):**e946.
19. Kim SK, Nam JW, Rhee JK, Lee WJ, Zhang BT: **miTarget: microRNA target gene prediction using a support vector machine.** *BMC Bioinformatics* 2006, **7:**411.
20. Fan R-E, Chen P-H, Lin C-J: **Working set selection using the second order information for training SVM.** *Journal of Machine Learning Research* 2005, **6:**1889-1918.
21. Japkowicz N, Stephen S: **The Class Imbalance Problem: A Systematic Study.** *Intelligent Data Analysis* 2002, **6(5):**429-450.
22. Craigon DJ, James N, Okyere J, Higgins J, Jotham J, May S: **NASCArrays: a repository for microarray data generated by NASC's transcriptomics service.** *Nucleic Acids Res* 2004:D575-577.
23. Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo FM, Spencer F: **A Model-Based Background Adjustment for Oligonucleotide Expression Arrays.** *Journal of the American Statistical Association* 2004, **99(468):**909-917.
24. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ: **miRBase: tools for microRNA genomics.** *Nucleic Acids Res* 2008:D154-158.
25. Rhee S, Beavis W, Berardini T, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, Miller N, Mueller LA, Mundodi S, Reiser L, Tacklind J, Weems D, Wu Y, Xu I, Yoo D, Yoon J, Zhang P: **The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community.** *Nucleic Acids Res* 2003, **31:**224-228.
26. Jones-Rhoades MW, Bartel DP: **Computational identification of plant microRNAs and their targets, including a stress-induced miRNA.** *Mol Cell* 2004, **14(6):**787-799.
27. Addo-Quaye C, Eshoo TW, Bartel DP, Axtell MJ: **Endogenous siRNA and miRNA targets identified by sequencing of the Arabidopsis degradome.** *Curr Biol* 2008, **18(10):**758-762.
28. Fahlgren N, Howell MD, Kasschau KD, Chapman EJ, Sullivan CM, Cumbie JS, Givan SA, Law TF, Grant SR, Dangl JL, *et al.*: **High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of MIRNA genes.** *PLoS ONE* 2007, **2(2):**e219.
29. Lu C, Kulkarni K, Souret FF, MuthuValliappan R, Tej SS, Poethig RS, Henderson IR, Jacobsen SE, Wang W, Green PJ, *et al.*: **MicroRNAs and other small RNAs enriched in the Arabidopsis RNA-dependent RNA polymerase-2 mutant.** *Genome Res* 2006, **16(10):**1276-1288.
30. Allen E, Xie Z, Gustafson AM, Carrington JC: **microRNA-directed phasing during trans-acting siRNA biogenesis in plants.** *Cell* 2005, **121(2):**207-221.
31. Adai A, Johnson C, Mlotshwa S, Archer-Evans S, Manocha V, Vance V, Sundaresan V: **Computational prediction of miRNAs in Arabidopsis thaliana.** *Genome Res* 2005, **15(1):**78-91.
32. Adenot X, Elmayan T, Lauressergues D, Boutet S, Bouche N, Gasciolli V, Vaucheret H: **DRB4-dependent TAS3 trans-acting siRNAs**

control leaf morphology through AGO7. *Curr Biol* 2006, **16(9):**927-932.
33. Wu G, Poethig RS: **Temporal regulation of shoot development in Arabidopsis thaliana by miR156 and its target SPL3.** *Development* 2006, **133(18):**3539-3547.
34. Schwarz S, Grande AV, Bujdoso N, Saedler H, Huijser P: **The microRNA regulated SBP-box genes SPL9 and SPL15 control shoot maturation in Arabidopsis.** *Plant Mol Biol* 2008, **67(1–2):**183-195.
35. Franco-Zorrilla JM, Cubas P, Jarillo JA, Fernandez-Calvin B, Salinas J, Martinez-Zapater JM: **AtREM1, a member of a new family of B3 domain-containing genes, is preferentially expressed in reproductive meristems.** *Plant Physiol* 2002, **128(2):**418-427.
36. Crawford BC, Ditta G, Yanofsky MF: **The NTT gene is required for transmitting-tract development in carpels of Arabidopsis thaliana.** *Curr Biol* 2007, **17(13):**1101-1108.
37. Yanhui C, Xiaoyuan Y, Kun H, Meihua L, Jigang L, Zhaofeng G, Zhiqiang L, Yunfei Z, Xiaoxiao W, Xiaoming Q, *et al.*: **The MYB transcription factor superfamily of Arabidopsis: expression analysis and phylogenetic comparison with the rice MYB family.** *Plant Mol Biol* 2006, **60(1):**107-124.
38. Hong SH, Kim HJ, Ryu JS, Choi H, Jeong S, Shin J, Choi G, Nam HG: **CRY1 inhibits COP1-mediated degradation of BIT1, a MYB transcription factor, to activate blue light-dependent gene expression in Arabidopsis.** *Plant J* 2008.
39. Wang X, Wang X: **Systematic identification of microRNA functions by combining target prediction and expression profiling.** *Nucleic Acids Res* 2006, **34(5):**1646-1652.