Discovery of Functional Genes for Systemic Acquired Resistance in Arabidopsis Thaliana through Integrated Data Mining

Pan, Youlian; Pylatuik, Jeffery D.; Ouyang, Junjun; Famili, A. Fazel; Fobert, Pierre R.

Publisher's version / Version de l'éditeur:

Journal of Bioinformatics and Computational Biology, 2, 4, 2004

Journal of Bioinformatics and Computational Biology
Vol. 2, No. 4 (2004) 639–655

Imperial College Press
www.icpress.co.uk

DISCOVERY OF FUNCTIONAL GENES FOR SYSTEMIC ACQUIRED RESISTANCE IN *ARABIDOPSIS THALIANA* THROUGH INTEGRATED DATA MINING

YOULIAN PAN

Institute for Information Technology, National Research Council Canada
1200 Montreal Road, Bldg M-50, Ottawa, Ontario, Canada K1A 0R6
Youlian.Pan@nrc.ca

JEFFREY D. PYLATUIK

Plant Biotechnology Institute, National Research Council Canada
110 Gymnasium Place, Saskatoon, Saskatchewan, Canada S7N 0W9
Jeffrey.Pylatuik@nrc.ca

JUNJUN OUYANG* and A. FAZEL FAMILI†

Institute for Information Technology, National Research Council Canada
1200 Montreal Road, Bldg M-50, Ottawa, Ontario, Canada K1A 0R6
**Junjun.Ouyang@nrc.ca*
†Fazel.Famili@nrc.ca

PIERRE R. FOBERT

Plant Biotechnology Institute, National Research Council Canada
110 Gymnasium Place, Saskatoon, Saskatchewan, Canada S7N 0W9
Pierre.Fobert@nrc.ca

Received 23 December 2003
Revised 30 January 2004
Accepted 29 March 2004

Various data mining techniques combined with sequence motif information in the promoter region of genes were applied to discover functional genes that are involved in the defense mechanism of systemic acquired resistance (SAR) in *Arabidopsis thaliana*. A series of K-Means clustering with *difference-in-shape* as distance measure was initially applied. A stability measure was used to validate this clustering process. A decision tree algorithm with the *discover-and-mask* technique was used to identify a group of most informative genes. Appearance and abundance of various transcription factor binding sites in the promoter region of the genes were studied. Through the combination of these techniques, we were able to identify 24 candidate genes involved in the SAR defense mechanism. The candidate genes fell into 2 highly resolved categories, each category showing significantly unique profiles of regulatory elements in their promoter regions. This study demonstrates the strength of such integration methods and suggests a broader application of this approach.

Keywords: Integrated data mining; motif identification; classification; systemic acquired resistance; microarray.

1. Introduction

One of the greatest challenges in modern biology is to understand how the expression pattern of thousands of genes in a living organism is regulated and how expression patterns are related to one another. High throughput determination of expression profiles has been prevalent in the past decade, particularly with the advent of microarray technology. This has motivated researchers to utilize tools, techniques, and algorithms, developed through many years of data mining and knowledge discovery research, to search for useful patterns in the gene expression data. This is exemplified by the abundance of computerized data analysis tools that have become available to perform clustering, pattern recognition, and motif identification in genes.

Generally, none of these individual data analysis tools are able to completely reveal the true nature of gene regulation and co-expression in a living cell. Microarray gene expression data is subject to multiple sources of variation. These include biological variation, which may be influenced by environmental, developmental, or genetic factors; technical variation, which may be influenced by sample preparation, hybridization, array platform, or probe design; and measurement variation, which can be influenced by the array scanner or label fluorescence.[1] To cope with such instability in the data, many normalization techniques have been developed, but these techniques can only ease rather than solve the problems completely. As a consequence, the confidence in knowledge derived from the data by a single analysis tool is dependent on the extent of noise and bias.

One of the most important questions in data mining is how to understand the scope and minimize the impact of such noise and bias within the data. In this paper, we describe an integrative approach in mining microarray gene expression data by using a software package developed in-house called BioMiner (http://iit-iti.nrc-cnrc.gc.ca/projects-projets/biomine_e.html), which contains a suite of tools for functional genomics. The domain problem we studied was the regulation of a defense response in *Arabidopsis thaliana*, a small flowering mustard plant, using data generated by microarray analysis.

The microarray analysis addressed two key variables: the first was the effect of salicylic acid (SA), a key elicitor of pathogen-induced systemic acquired resistance (SAR) in plants; and the second variable was the effect of mutating the *NPR1* (Non-expresser of Pathogenesis Related) gene, a key regulator of SAR.

The establishment of SAR, an inducible defense response that leads to broad-spectrum systemic resistance, requires an endogenous increase in SA levels.[2] However, the exogenous application of low concentrations of SA, as used in this study, can also trigger a SAR response. In *Arabidopsis*, the *NPR1* gene is essential for SA-mediated SAR.[3] Plants with *npr1* mutations are therefore compromised in their ability to launch an SAR response. Currently there is no evidence to suggest that NPR1 binds DNA directly to regulate transcription. Rather, all research to date suggests that NPR1 indirectly regulates the expression of genes involved in SAR

by interacting with DNA-binding transcription factors in the nucleus such as those of the TGA family of bZIP transcription factors.[4–9]

Other than its interaction with the TGA family of bZIP transcription factors, NPR1 has not been shown to interact with any other transcription factors. However, it has been shown that the expression of several members of the WRKY family of transcription factors is dependent on NPR1.[10] Furthermore, many disease-related genes contain promoters that are highly enriched with the W box, the DNA binding site for WRKY transcription factors.[11] It is therefore possible that NPR1 also mediates the SAR response through other transcription factors such as those of the WRKY family.

In this study, our objective was to demonstrate the advantage of an integrated data mining approach in knowledge discovery from the microarray gene expression profiles and to identify genes that are regulated by NPR1 in response to SA during the onset of SAR. This can be achieved by examining the expression profiles of genes altered in wild type (WT) and mutant (*npr1-3*) plants in response to SA. We first explain the material used and the method applied. The paper then continues with a detailed description of our knowledge discovery process that is given in Sec. 3. Section 4 contains discussion and Sec. 5 is future directions of our research.

2. Materials and Methods

2.1. *Array design and hybridization*

Double-stranded amplicons averaging 600 bp in length were designed by PBI's annotation project for *Arabidopsis thaliana* (http://bioinfo.pbi.nrc.ca/bioinfo/Current_projects/Arabidopsis_Annotation/index.html) and generated from *Arabidopsis thaliana* ecotype Columbia genomic DNA. These amplicons were quantified, and purified to produce 12662 reporters representing the predicted loci of 9899 genes based on the TAIR sequence viewer dataset last annotated on Nov. 6, 2003 (ftp://ftp.arabidopsis.org/home/tair/Maps/seqviewer_data/). An amount of 0.1 to 0.2 μg/μl dilutions of the amplicons were spotted (100 μm diameter) in duplicate on CMT-GAPS slides (Corning, Cat# 40004) using a ChipWriterPro (Virtek) equipped with quill pins (ArrayIt).[12]

Wild-type *Arabidopsis thaliana* (L.) Heynh. ecotype Columbia and the *npr1-3* mutant were grown with a 10-h-light/14-h-dark photoperiod at 22°C/18°C. Three to four week-old plants were either sprayed with water or with 0.5 mM salicylic acid and samples were collected at two time points afterwards (2 or 8 hour for SA, 8 hour for water) (Fig. 1). Spraying of SA/water for these time points was arranged such that tissue collection was always performed at the same time of day, thereby minimizing any circadian or temperature-regulated effects. Rossette leaf tissue was collected and immediately frozen in liquid N_2. Total RNA was extracted from frozen tissue using RNeasy mini columns with an on-column DNase treatment (Qiagen).

Forty to 70 μg of total RNA was used as template to directly label cDNA with either cy3 or cy5 fluorophores and hybridized for 16 to 18 hours at 37°C in DIGeasy

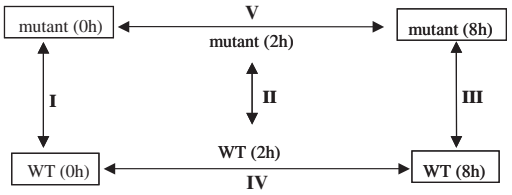


Fig. 1. Experimental design. The Roman numerals are the experiment IDs. The number in brackets refers to the number of hours plants were treated with salicylic acid prior to tissue collection. (0 h) refers to plants treated with water.

Hyb (Roche). Arrays were washed 3 times in 1XSSC, 0.1% SDS at 50°C. Four hybridizations, representing biological replicates, were performed for each experiment (Fig. 1), 2 hybridizations for which the control was labeled with cy3, and 2 hybridizations for which the control was labeled with cy5 (i.e. reciprocal labeling).

2.2. *Data collection and pre-processing*

Arrays were scanned at 10 μm resolution in a ScanArray 4000 scanner (PerkinElmer). Spot location and intensity quantitation were performed using QuantArray version 3.0. Adaptive spot quantitation was employed and median intensity values were used for subsequent analysis. Localized background subtraction was performed using the BASE software platform[13] and resulting signals from each channel were normalized using the intensity-dependent LOWESS method.[14] Paired channel intensity values (background subtracted and normalized) from replicate hybridizations were analyzed to identify statistically significant up- or down-regulated genes using SAM (Significance Analysis of Microarrays) software.[15] Delta values were adjusted to achieve a false discovery rate of 5% ± 0.60% for each experiment (Table 1). A total of 738 target amplicons (reporters) representing 685 annotated loci (TAIR) were identified by SAM as showing significant change in expression between the control and experimental samples in at least one experiment (Table 1).

Expression values approaching saturation (> 50,000) were considered unreliable and therefore filtered. Log(2) ratios of all the eight replicates (4 biological replicates

Table 1. Summary of results for significance analysis of microarrays.

Experiment	Significant Reporters	Significant Up-regulated	Significant Down-regulated	Predicted False Positives	False Discovery Rate (%)	Delta Value
I	19	3	16	0.87	4.56	1.04
II	43	4	39	2.41	5.60	1.21
III	344	174	170	16.89	4.91	1.32
IV	404	271	133	20.32	5.03	1.38
V	64	64	0	3.21	5.01	1.42

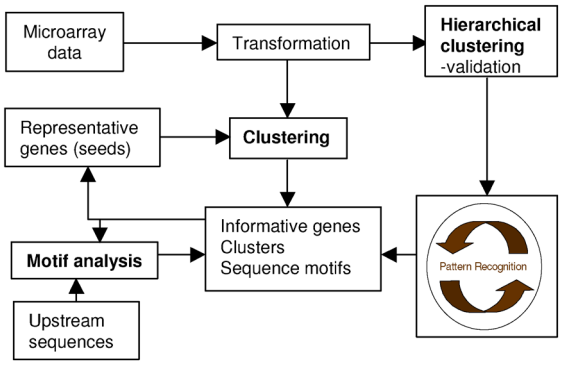


Fig. 2. Knowledge discovery process.

with 2 technical replicates each) were used in the subsequent knowledge discovery process.

3. Knowledge Discovery Process

The knowledge discovery process is illustrated in Fig. 2. Groups of gene reporters were determined using K-Means clustering (unsupervised learning) methods. A group of informative genes were identified from the entire dataset through pattern recognition (a supervised learning method detailed below) and compared to interesting clusters generated by K-Means. Interesting motifs in the upstream promoter region were identified for each gene and compared with other genes in the same cluster. A combination of results of informative genes, gene expression profiles and motif information constituted a representative gene reporter for each cluster. These representative gene reporters were used as seeds for regrouping the data through K-Means to determine more refined clusters.

3.1. *Search for informative genes through pattern recognition*

The input data (Fig. 3) for pattern recognition consists of a matrix containing p attributes (gene reporters) for n cases (samples) and an attribute vector containing labels for all cases. From the perspective of data-mining, the gene reporters in this study are considered as attributes and each replicate of an experiment condition is considered as a case ($p = 738$ and $n = 16$). The labels correspond to the case identifications, i.e. experimental conditions that are used for discrimination by pattern recognition. Based on the design of microarray experiments, we conducted two sets of classifications (Table 2). One was the ratio based on the mutant over wild type (RA = mutant/WT) that resulted in two classes, 0 h and 8 h. This classification was to identify gene reporters whose ratios were significantly changed after treatment with SA for 8 hours. The other classification was the ratio of treatment with SA for 8 hours over non-treated (RB = 8 h/0 h). The two classes were mutant

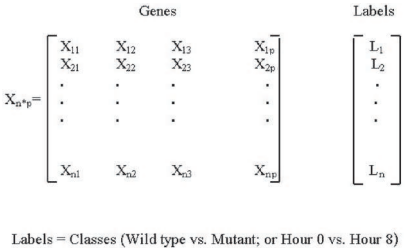


Fig. 3. Format of input data for classification.

Table 2. Two classification experiments. Both experiments were done based on ratios of expression values indicated in the row labeled as "Values". The J4.8 decision tree algorithm was used to find a classifier that distinguishes the two classes based on their ratio values.

Experiments	RA	RB
Values	mutant/WT	h8/h0
Classes	h0	WT
	h8	mutant

and wild type. This classification was to identify gene reporters that responded to SA treatment differently between the two classes.

The agglomerative hierarchical clustering[16] with the Euclidean distance measure was used to ensure the validity of the partitioning properties between the classes (Fig. 4). These results showed that the two classes in each process were truly distinguishable, and our design of the supervised learning process was feasible.

A search for the informative genes among the 738 gene reporters was done using the J4.8 decision tree algorithm.[17] A "*discover-and-mask*" technique[18] was applied in this process. Figure 5 shows an example of a decision tree and its conversion to rules (see legend of Fig. 5) that separate the two classes (0 h and 8 h, see RA in Table 2). The corresponding gene reporter (e.g. At2g31880-1025 in Fig. 5) was identified as an informative gene of this dataset and then masked (removed). The same analyses were performed for the remainder in the dataset until (*i*) a drop in the percentage accuracy of the resulting model (decision tree) in the prediction of participating cases; or (*ii*) no more identification of informative genes that are able to distinguish the two classes. The same technique has been used to identify informative genes in other applications.[18,19]

In this study, we exhaustively applied the "*discover-and-mask*" technique on both datasets for classification. Namely, all potential models that were able to distinguish the two classes in the dataset were identified and the discovered genes were ranked (Table 3). Sometimes, a decision tree involved more than one gene. We denote the single gene model as a *simple* model and a multi-gene model as a *complex* model.

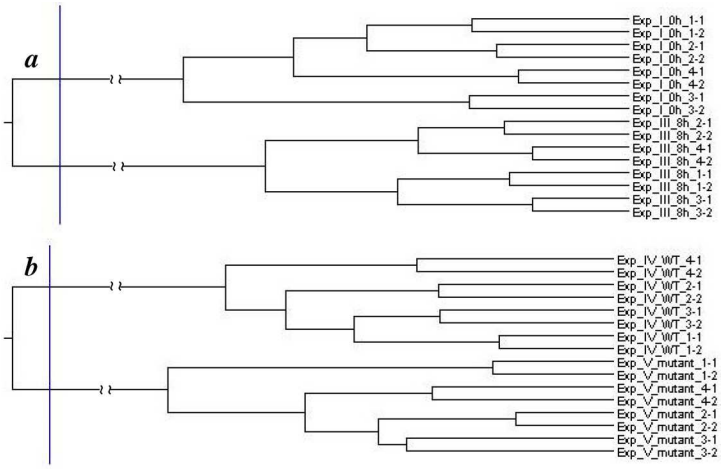


Fig. 4. Hierarchical clustering of RA (***a***) and RB (***b***). See text for details. The vertical lines on the left locate the separating branches in the tree for the two clusters.

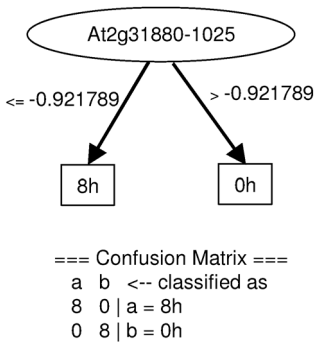


Fig. 5. A decision tree. Total number of instances = 16, all correctly classified. The rule is "if At2g31880-1025 is larger than −0.921789, then this plant was untreated, otherwise this plant was treated with SA for 8 hours".

Table 3. Number of informative gene reporters identified using the "*discover-and-mask*" technique. "Simple" = models involve only one gene. "Complex" = models involve more than one gene. "Total models" = simple + complex, "Simple and 100%" = number of gene reporters (and models) identified through simple models with 100% training accuracy.

Experiment	Total Models	Total Reporters	Simple Models	Simple and 100%
RA	293	484	106	40
RB	392	604	181	68

A total of 40 gene reporters were identified through the RA classification experiment, and 68 through the RB that were involved in simple models with 100% training accuracy. A potential group of gene reporters that are related with both the mutation and SA treatment can be identified through intersection of RA and RB results with 100% training accuracy. Consequently, we were able to identify 15 gene reporters belonging to this category and denote this group of gene reporters as the most informative.

3.2. *Search for significant gene clusters*

The input data in this section was formatted differently from the earlier section. Individual gene reporters were the cases and the experiment conditions were the attributes. There is no attribute vector of labels for unsupervised learning. Replicates of the same attributes were pooled and a mean of such replicates was used in the clustering process.

Due to the nature of these five experiments, the data were categorized into two groups. The first group was the time series data that contains experiments I, II and III. The second group was the effect of SA within each of the two strains of *Arabidopsis thaliana* and contained experiments IV and V. We also considered all five experiments together in the following clustering processes.

Initially, a series of K-Means clustering processes was conducted with K ranging from 2 to 40.[20] For a distance measure, we used *difference-in-shape* (Eq. 1),[21] for time series data and the entire five experiments, but *coefficient of divergence* (Eq. 2),[22] for the second group.

$$d_{ik} = \sqrt{\left| \left[\sum_{j \in A} (X_i[j] - X_k[j])^2 \right] - \left[\sum_{j \in A} (X_i[j] - X_k[j]) \right]^2 / c \right| / (c-1)}. \quad (1)$$

$$d_{ik} = \sqrt{\left[\sum_{j \in A} ((X_i[j] - X_k[j])/(X_i[j] + X_k[j]))^2 \right] * (p/c)}. \quad (2)$$

- $A = \{j \mid j \in \{1 \ldots n\} \wedge$ attribute value $X_i[j]$ is not missing$\}$;
- $p = 1$ or $n, c = [1 \ldots n]$;
- where both $X_i[j]$ and $X_k[j]$ are not missing value, 'c' is the number of variables for which neither $X_i[j]$ nor $X_k[j]$ is missing and 'n' is the total number of variables for certain attribute. 'p' is the denormalized coefficient that could be 1 or n.

The quality of each separated cluster was computed based on the Silhouettes value[23] and cluster stability[24] over various clustering processes. Silhouettes value is determined based on the comparison of tightness of each cluster and its separation from the others,[23] while a stability value is determined by the repeatability of a

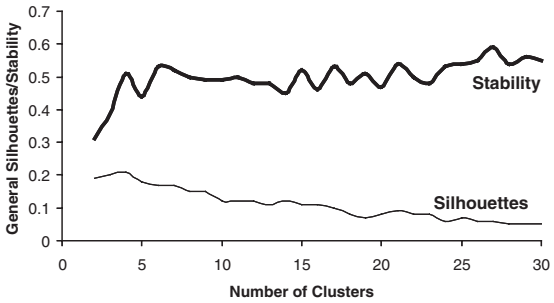


Fig. 6. Variation of the general Silhouettes and general stability values with the number of clusters in a clustering process.

clustering over a series of clustering processes.[24] The Silhouettes value and stability value of individual clusters were used to compute the general Silhouettes and general stability values of each clustering process, respectively. The general Silhouettes value usually decreases as the number of clusters exceeds a certain threshold. However the general stability does not have this property, but have scattered maxima (Fig. 6). By considering both general Silhouettes and general stability values, we were able to identify an optimal number of clusters for each dataset. We were particularly interested in the clusters of genes whose mean expression levels (centroids) were significantly up- or down-regulated for the mutant (experimental) as compared to the wild type (control) through the progression of the SA treatment time course, or the SA treated (8 h) vs. non-treated (0 h). The clusters that contain genes that have no significant change in the expression levels were not considered even though they have higher values of Silhouettes and/or stability values. As a result, we selected two clusters to be further considered in the subsequent analysis. They are labeled as *up* and *down* with respect to the time series experiments.

In comparing the gene reporters and stability among all *up* clusters from the three datasets (1: I, II, III; 2: IV, V; and 3: I, II, III, IV, V), we noticed a conserved list of genes between the clusters generated by dataset 1 and dataset 3. A non-redundant list of reporters was combined from the two clusters. A list for *down* clusters was generated similarly (Table 4).

3.3. *Search for interesting sequence motifs in the promoter region*

A general survey of 368 plant transcription factor consensus binding sites was performed using GenericBioMatch, a modified pattern match algorithm for biological sequences,[25] for the promoter region (1000 bp upstream of putative transcription start site) of the genes appearing in each cluster. We found 193 plant motifs that commonly appeared in this set of genes (data not shown). Since the *NPR1* gene has been found to interact with and enhance the binding of members of the TGA-bZIP transcription factors[4–6] and the expression of several members of WRKY family of transcription factors is dependent on NPR1[10], we are particularly interested in

Table 4. Appearances of W, Wy and ASF-1 motif elements in 1000 bp upstream sequence (both strands) in all genes represented by the microarray, significant genes, and clustered genes. The integer on the motif line indicates the number of sequences that contain the specific motif. In the "*up*" and "*down*" columns, "A" is for the clusters generated before application of selected seeds, while "B" is for the clusters generated afterward. "SEF" = statistical expected frequency, which is calculated based on nucleotide distribution probabilities in the promoter sequences. "All" = entire 9899 genes used in the microarray, "All significant" = dataset used in this paper. "RA × RB" = genes in both RA and RB lists (Table 3) with 100% training accuracy.

				Up		*Down*		
	SEF	All	All significant	A	B	A	B	RA × RB
Number of Sequences	N/A	9899	685	25	12	15	12	15
TTGAC	N/A	8366	592	20	9	13	10	14
TTGACY	N/A	6363	470	13	7	11	8	11
TGACG	N/A	5276	389	12	7	9	7	10
Motif index								
W box	2.04	2.28	2.32	1.72	1.83	3.20	3.33	2.73
Wy box	1.01	1.22	1.29	0.84	1.00	1.93	2.00	1.47
ASF-1 motif	1.02	0.89	0.91	0.60	0.75	0.93	1.00	1.07

the *cis*-acting elements bound by these two families of transcription factors. The TGA-bZIP transcription factors bind to a TGACG sequence element called the ASF-1 motif.[26] The WRKY transcription factors bind to a TTGAC sequence element called a W box.[27] Most W boxes analyzed so far are followed with either C or T (represented as Y).[27] We denote these more stringent W boxes as a Wy box. The number of genes whose promoter regions contain such motifs is listed in Table 4.

In Table 4, the **W box index** is defined as the mean number of W boxes that appear in the 1000 bp upstream region (both sense and anti-sense strands) of a group of genes. The **Wy box index** and **ASF-1 motif index** are defined similarly.

3.4. *Identification of representatives for co-expressed genes as seeds for re-clustering*

Based on motif results (Table 4), and published information regarding the possible roles of these elements in regulating SAR-related genes (see introduction), a strategy was devised for selecting representative genes as seeds to re-cluster. From each of the three datasets (1: I, II, III; 2: IV, V; and 3: I, II, III, IV, V), we identified two clusters: *up* and *down*. Gene reporters that were present in the *up* clusters generated by all three datasets were chosen as candidate seeds for the *up* cluster. Candidate seeds for the *down* cluster were identified similarly. These candidate seeds were further screened based on appearance of W box and ASF-1 motif, and ranking in the list of informative reporters. Genes that contained both W box and ASF-1 motif and ranked the highest were selected (Table 5).

A series of clustering, which generated 2 to 40 clusters, were performed for the three datasets using the two representative genes as seeds. We denote these two

Table 5. Representative genes. "RA × RB" stands for intersection of RA and RB with 100% training accuracy (Table 3), namely the list of most informative genes.

Cluster	Reporter	W Box	ASF-1 Motif	Wy Box	Informative Gene List
Up	At2g43560-1813	1	1	0	RA × RB
Down	At3g11340-5623	2	2	2	RA × RB

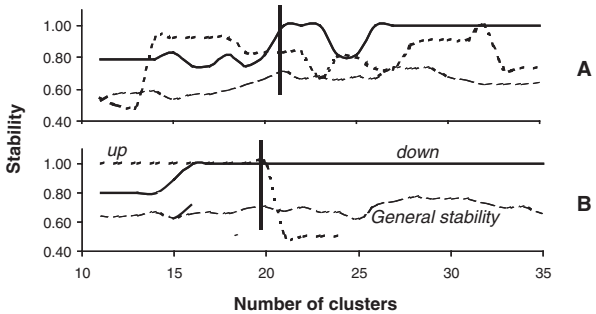


Fig. 7. Stability of *up, down* cluster and general stability during re-clustering process of two datasets. **A**: time series dataset (I, II, III); **B**: entire dataset (I, II, III, IV, V). Vertical bars indicate where the *up* and *down* clusters were taken for the final two clusters (Fig. 8). Labels in panel **B** also apply to panel **A**.

seeds as master seeds to distinguish them from additional seeds. In order to perform clustering for more than two clusters, additional seeds were selected from those first occurred in each dataset. Both master seeds represent respective clusters with the highest stability in a series of clustering processes (Fig. 7). This demonstrates the robustness of the chosen seeds in distinguishing their cluster of genes from other genes in the dataset. A union of genes appearing in the *up* clusters of datasets one (I, II, III) and three (I, II, III, IV, V) produced a non-redundant list of genes. The *down* cluster was generated similarly. The resulting size of both new clusters decreased (Table 4), while all of the member genes in each respective cluster had co-clustered prior to the use of master seeds. Figure 8 shows the expression profiles of genes in these two clusters.

3.5. *Analysis of motif results*

Although there is significant overlap between the W boxes and ASF-1 motif, their frequencies of occurrence, as indicated by their index values, were markedly different from each other among the various populations of genes. We found the majority (10 out of 12) of genes in the *down* cluster contained at least one W box in their promoter region. WRKY proteins bind to sequences with an invariant TGAC core, which is often preceded by a T.[11] Before the application of the selected seeds, this TTGAC motif appears 48 times on both strands of the 13 promoters, with an average of 3.7 copies per promoter (not considering the promoters that do not have this motif). The statistical expectation for a random distribution of the pentamer

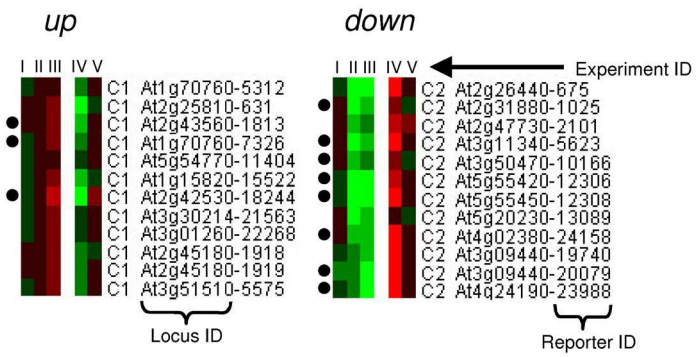


Fig. 8. Expression profile of genes in *up* and *down* cluster. Red indicates up-regulation. Green indicates down-regulation. Black dots on the left denote that these genes were also identified as the most informative genes (RA × RB, Table 3) through pattern recognition.

is 2.04 copies per 1000 bp (on both strands). Thus, W boxes were highly enriched in the promoters of the *down* cluster genes. Using the most stringent definition of WRKY binding site, a TTGACY (Y = C/T, named Wy box in Table 4) hexamer motif,[27] we still detected a significant over representation of 29 potential binding sites on 11 of the 15 promoters, an average of 2.6 (as compared to 1.02 by random distribution) copies per promoter. This property did not change after application of selected seeds. Because the *down* cluster represents genes that were down regulated in the mutant in response to SA, our data would indicate that WRKY transcription factors, to a certain degree, mediate the regulatory control exerted by NPR1 on SAR-responsive genes in the wild type plant.

Conversely, the ASF-1 motif appears under-represented in the promoters of the genes from *up* cluster, with an average frequency of 0.75 copies per promoter. Furthermore, even though it shares the same TGAC core with W box, this motif appears to be under-represented throughout the genome; the calculated index for all genes represented on the array is 0.89, when its statistically expected frequency (SEF) is 1.01 copies per promoter (Table 4). However, the frequency of this motif in the promoters of genes from *down* cluster appears to be similar to SEF.

3.6. *Integration of results from various stages of the knowledge discovery process*

Through repeated clustering, we were able to identify and confirm 12 genes that were markedly down-regulated in the mutant as compared to the wild type following SA treatment and 12 genes that were up-regulated. Using the "*discover-and-mask*" technique, we were able to identify 15 highly informative genes (both in RA and RB with 100% training accuracy, Table 3), the majority of which were down-regulated. In these 15 most informative gene reporters, 8 also appeared in the *down* cluster (Fig. 8) and had higher **W box** and **ASF-1 motif indices** when compared to

those not in this group (Table 6). Among the remaining 7 genes, only 3 were in the *up* cluster (Fig. 8) and had a low **W box index** (1.00). This is consistent with the results in Table 4. All reporters in the final list of the two clusters were also identified as informative by the "*discover-and-mask*" technique and appeared in both RA and RB list (Table 3) except for At1g15820-15522 that appeared only in RA list.

4. Discussion and Conclusion

We have demonstrated the strength of integrated data mining in the knowledge discovery process of microarray gene expression data of *Arabidopsis thaliana*. We have shown that the combination of motif information with both unsupervised and supervised learning methods in knowledge discovery, and the use of representative genes as seeds for re-clustering are advantageous and novel approaches to mine biological data. The use of representative genes as seeds improved the quality of clustering analysis. This re-clustering process refined the two clusters.

This study has identified genes whose expression patterns were specifically altered in the mutant in response to the SAR defense mechanism as initiated by SA treatment. These genes represent candidates that could function as indirect targets for regulation by NPR1, and our study warrants further investigation into their roles for the SAR response. The fact that the promoters of down-regulated genes are highly enriched in W boxes suggests that they could be activated indirectly by NPR1 through WRKY proteins.

This study also showed that the ASF-1 motif is under-represented in the *up* clusters of genes identified in this study. Of further interest, the **ASF-1 motif index** of our entire dataset (representing 9899 genes) is lower than the statistically predicted frequency (0.89 vs. 1.01), suggesting that ASF-1 motifs are generally under represented in the promoter regions of the genome. However, in the *down* cluster, the **ASF-1 motif index** appears to be slightly higher than the genome wide average. Furthermore, 8 genes in the *down* cluster, which were also identified as the most informative by the "*discover-and-mask*" technique, are further enriched for ASF-1 motif (1.38, Table 6). The higher **ASF-1 motif index** in the most informative down-regulated genes and its under-representation in the *up* cluster suggests that NPR1 may mediate transcriptional activation through the TGA family of bZIP transcription factors during the SAR response. This finding is supported by earlier work.[28]

Table 6. The difference in **motif indices** between two groups of the most informative genes. Eight reporters appeared in *down* cluster, while the other seven did not.

In *Down*?	W Box Index	ASF-1 Motif Index	Wy Box Index
Yes	3.88	1.38	2.13
No	1.43	0.71	0.71

We initially considered a transcription factor (TF) in our seed selection as suggested by Zhu *et al.*[29] This attempt was unsuccessful due to the fact that, in this experiment, the expression change of most TFs was not very significant. Among the 1411 TFs from AGRIS (http://arabidopsis.med.ohio-state.edu/AtTFDB/, Nov. 6, 2003), 14 appear in the 685 genes used in this study, but none are among the final list of either *up* or *down* clusters. This implies that a group of co-expressed genes, if they are also co-regulated, do not necessarily co-express with the TF that regulates this group of genes. There could be a delay between the expression of TFs and the clusters of genes that the TFs regulate, which cannot be resolved in this study. The regulatory module networks identified by Sagel *et al.* in yeast clearly demonstrate this phenomenon.[30]

In conclusion, we have demonstrated the advantages of an integrative data mining approach that consists of clustering with various distance measure algorithms, pattern recognition, and motif information in the promoter region of a group of genes. This study has shown the strength of such combined approaches in knowledge discovery related to the SAR defense mechanism in *Arabidopsis thaliana.* Through this integration, we were able to identify 12 informative genes that were down regulated and 12 up-regulated genes in the *npr1-3* mutant plants in response to treatment with SA. The promoter regions of the down-regulated genes are highly enriched with W-box and Wy-box motifs, while that of up-regulated genes are deprived of W-box. Throughout the entire genome of *Arabidopsis thaliana*, the ASF-1 motif appears to be under represented. In the *up*-regulated group of genes, this motif is even further deprived. However, a group of 8 most informative down-regulated genes are enriched with the ASF-1 motif as compared with its genome wide distribution.

5. Future Direction

This study has clearly shown the strength of integrating various data mining techniques. Further study may include the development of new computational approaches to discover potential, currently unknown common motifs in the promoter region of a cluster of genes. Integration of gene expression profiles and certain known functionality of co-expressed genes and discovery of common motifs would promote the annotation of previously un-described genes. Our current study is on functional genes related with SAR defense mechanism in *Arabidopsis thaliana*, this same technique can be applied to other domains including biomedical research.

Acknowledgments

We are thankful to two anonymous reviewers for their constructive comments and criticism. Funding for this research was provided by Genome and Health Initiative phases 1 and 2 of National Research Council of Canada and a Visiting Fellowship awarded to J.D.P from Natural Sciences and Engineering Research Council of Canada. We would like to thank Bill Crosby (University of Saskatchewan) for

the development of the *Arabidopsis* amplicon collection and Jacek Nowak, Kevin Koh (PBI/NRC) and Mark Wilkinson (University of British Columbia) for assistance with the BASE software, Ganming Liu (IIT/NRC) for assistance at an early stage of clustering quality measurement. This is publication NRC 46550 of National Research Council of Canada.

References

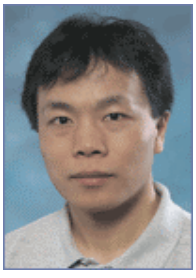
1. Churchill GA, Fundamentals of experimental design for cDNA microarrays, *Nat Genet* **32**:490–495, 2002.
2. Ryals JA, Neuenschwander UH, Willits MG, Molina A, Steiner HY, Hunt MD, Systemic acquired resistance, *Plant Cell* **8**:1809–1819, 1996.
3. Delaney TP, Friedrich L, Ryals JA, *Arabidopsis* signal transduction mutant defective in chemically and biologically induced disease resistance, *Proc Natl Acad Sci USA* **92**:6602–6606, 1995.
4. Zhang Y, Fan W, Kinkema M, Li X, Dong X, Interaction of NPR1 with basic leucine zipper protein transcription factors that bind sequences required for salicylic acid induction of the PR-1 gene, *Proc Natl Acad Sci USA* **96**:6523–6528, 1999.
5. Zhou JM, Trifa Y, Silva H, Pontier D, Lam E, Shah J, Klessig DF, NPR1 differentially interacts with members of the TGA/OBF family of transcription factors that bind an element of the PR-1 gene required for induction by salicylic acid, *Mol Plant Microbe Interactions* **13**:191–202, 2000.
6. Després C, DeLong C, Glaze S, Liu E, Fobert PR, The *Arabidopsis* NPR1/NIM1 protein enhances the DNA binding activity of a subgroup of the TGA family of bZIP transcription factors. *Plant Cell* **12**:279–290, 2000.
7. Kinkema K, Fan W, Dong X, Nuclear localization of NPR1 is required for activation of PR gene expression, *Plant Cell* **12**:2339–2350, 2000.
8. Subramaniam R, Desveaux D, Spickler C, Michnick SW, Brisson N, Direct visualization of protein interactions in plant cells, *Nat Biotech* **19**:769–772, 2001.
9. Johnson C, Boden E, Arias J, Salicylic acid and NPR1 induce the recruitment of *trans*-activating TGA factors to a defense gene promoter in *Arabidopsis*, *Plant Cell* **15**:1846–1858, 2003.
10. Yu D, Chen C, Chen Z, Evidence for an important role of WRKY DNA binding proteins in the regulation of NPR1 gene expression, *Plant Cell* **13**:1527–1539, 2001.
11. Maleck K, Levine A, Eulgem T, Morgan A, Schmid J, Lawton KA, Dangl JL, Dietrich RA, The transcriptome of *Arabidopsis thaliana* during systemic acquired resistance, *Nat Genet* **26**:403–410, 2001.
12. Tessier DC, Benoit F, Rigby T, Hogues H, Van het Hoog M, Thomas DY, Brousseau R, A DNA Microarrays Fabrication Strategy for Research Laboratories, http://bri-irb.nrc-cnrc.gc.ca/pdf/Microarray-Chapter.pdf, 2003.
13. Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg Å, Peterson C, BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data, *Genome Biol* **3**: software0003.1-0003.6, 2002.
14. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP, Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Res* **30**:e15, 2002.
15. Tusher V, Tibshirani R, Chu C, Significance analysis of microarrays applied to ionizing radiation response, *Proc Natl Acad Sci* **98**:5116–5121, 2001.
16. Orloci L, "An agglomerative method for classification of plant communities," *J Ecol* **55**:193–206, 1967.

17. Witten I, Eibe F, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, San Mateo, CA, 1999.
18. Famili A, Ouyang J, Data mining: understanding data and disease modeling, in *Proceedings of the 21st IASTED International Conference on Applied Informatics.* ACTA Press, Anaheim, pp. 32–37, 2003.
19. Walker PR, Smith B, Liu QY, Famili AF, Valdes JJ, Liu Z, Data mining of gene expression changes in Alzheimer brain, *Artificial Intelligence in Medicine* **31**: 137–154, 2004.
20. Anderberg MR, *Cluster Analysis for Applications.* Academic Press, New York, 1973.
21. Sokal RR, Sneath PH, *Principles of Numerical Taxonomy*, Freeman WH, San Francisco, 1963.
22. Clark PJ, An extension of the coefficient of divergence for use with multiple characters, *Copeia* **1952**:61–64, 1952.
23. Rousseeuw PJ, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *J Comp Appl Math* **20**:53–65, 1987.
24. Famili AF, Liu G, Liu Z, Evaluation of optimization of clustering in gene expression data analysis, *Bioinformatics* **20**:1535–1545, 2004.
25. Pan Y, Famili AF, GenericBioMatch: a novel generic pattern match algorithm for biological sequences, in: *Proceedings of the 2003 IEEE Bioinformatics Conference (CSB2003)*, IEEE Computer Society, Los Alamitos, CA, pp. 562–563, 2003.
26. Lebel E, Heifetz P, Thorne L, Uknes S, Ryals J, Ward E, Functional analysis of regulatory sequences controlling PR-1 gene expression in *Arabidopsis*, *Plant J* **16**: 223–233, 1998.
27. Eulgem T, Rushton PJ, Robatzek S, Somssich IE, The WRKY super-family of plant transcription factors, *Trends Plant Sci* **5**:199–205, 2000.
28. Fan W, Dong X, *In vivo* interaction between NPR1 and transcription factor TGA2 leads to salicylic acid-mediated gene activation in *Arabidopsis*, *Plant Cell* **14**: 1377–1389, 2002.
29. Zhu Z, Pilpel Y, Church GM, Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm *J Mol Biol* **318**:71–81, 2002.
30. Segal E, Shapira M, Gegev A, Pe'er D, Botstein D, Koller D, Friedman N, Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data, *Nat Genet* **34**:166–176, 2003.

Youlian Pan holds his Ph.D. in Biology and his Master of Computer Sciences, both from Dalhousie University, Halifax, Canada. He is currently a Research Officer at the Institute for Information Technology of National Research Council of Canada. His research interest is in data mining with biological application, specifically in biological sequence signature.

Jeffrey Pylatuik received his B.Sc. in Biochemistry and his Ph.D. in Molecular Biology at the University of Saskatchewan, Saskatoon, Canada, in 1996 and 2001 respectively.

Since 2001, he has been at the Plant Biotechnology Institute as a visiting scientist. His research interests focus on the application of microarray analysis to study the transcriptional regulation of plant disease resistance.



Junjun Ouyang received his M.S. degree in Botany from Miami University, Oxford, USA, 1998, and M.S. in Information and Computer Science from University of California, Irvine, USA, 2000. He is currently working at the Institute for Information Technology of National Research Council of Canada. He is interested in the development and application of data mining technologies for biomedical research.

Fazel Famili is a Senior Research Officer and Project Leader working at the Institute for Information Technology of the National Research Council of Canada, where he has been for the past 19 years. His research has been on data mining, machine learning and bioinformatics and their applications to real world problems.

Fazel has published or has been co-author of over 35 articles in data mining and AI. He has a data mining US patent (with two other colleagues from IIT). He has organized many workshops and has been involved in a number of data mining conferences.

Pierre Fobert received his Ph.D. in Plant Molecular Biology from Carleton University, Ottawa, Canada. After post-doctoral stages at the John Innes Centre and the Canadian Forestry Service, he joined the Plant Biotechnology Institute as a Research Officer in 1994. His group makes use of molecular biology, genomics and genetic approaches to study transcription factors.