Towards a temporal modeling of the genetic network controlling systemic acquired resistance in Arabidopsis thaliana

Tchagang, Alain; Shearer, Heather; Phan, Sieu; Famili, Fazel; Fobert, Pierre; Pan, Youlian

Publisher's version / Version de l'éditeur:
https://doi.org/10.1109/CIBCB.2010.5510589
2010 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), pp. 1-8, 2010-05-05

Towards a Temporal Modeling of the Genetic Network Controlling Systemic Acquired Resistance in *Arabidopsis thaliana*

Alain B. Tchagang, *Member, IEEE*, Heather Shearer, Sieu Phan, Hugo Bérubé, Fazel Famili, Pierre Fobert, and Youlian Pan, *Member, IEEE*

***Abstract*—We studied defense mechanism of the *Arabidopsis thaliana* subjected to Salicylic Acid (SA) treatment for 0, 1, and 8 hours using a broader application of the frequent itemset approach. Four genotypes of the plant were used in this study, Columbia wild type, mutant *npr1*, double mutant *tga1 ga4* and triple mutant *tga2 ga5 ga6*. We defined the major patterns of transcription regulation governing pathogen defense mechanism, thereby creating a model of the Systemic Acquired Resistance (SAR) at three time points. The temporal model describes the relationships among the regulators and defines groups of genes that are subject to similar regulation. The results obtained offered a first glimpse into the temporal pattern of the gene network controlling SAR in plant. We found that most of the genes that responded to SA challenge are in fact dependent on one or more of the NPR1 and TGA transcription factors tested in this study.**

I. Introduction

Systemic acquired resistance (SAR) is a general plant immune response that is induced after a local infection. The onset of SAR requires the endogenous increase of salicylic acid (SA) and the coordinated expression of Pathogenesis-Related (PR) genes, which encode small secreted or vacuole-targeted proteins that have antimicrobial activities [1]-[5]. Exogenous application of SA can also trigger a SAR response and activate PR gene expression in plants in the absence of pathogen infection [6]-[7]. In *Arabidopsis*, NPR1 (Non-expressor of Pathogenesis-Related genes 1) is essential for SA-mediated SAR [8]-[10]. 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 PR genes through interaction with DNA-binding transcription factors (TFs) in the nucleus, such as TGA family of bZIP factors [7]. The actual contribution of individual TGA factors has been difficult to discern because of functional redundancy among the factors, and possible dual function for some single factors [11]. There are 10 TGA TFs in *Arabidopsis* [12] of which seven (TGA1-TGA7) have been shown to interact with NPR1 [11], [13]. These seven TGAs can be divided into three groups based on sequence homology [14]. Group I consists of TGA1 and TGA4, both of which contain two Cys residues that do not appear in other TGA factors; reduction of the two Cys residues is responsible for the SA-dependent interaction with NPR1 in *Arabidopsis* leaves [15]; Group II consists of TGA2, TGA5 and TGA6; and group III consists of TGA3 and TGA7.

Most of what is known about the topology of the defense signaling network is based on comparing the effects of various defense signaling mutants on a few particular phenotypes [16]. Considering the large number of genes induced or repressed in response to pathogen attack and the apparent complexity of the signaling network, a comparison of mutant phenotypes on a larger scale is desirable. One method for obtaining system-wide information about mutant phenotypes is expression profiling using DNA microarray technology. Expression profiling has been used in studies of responses to pathogens; for example, to describe the response to pathogens-associated molecular patterns [17], to discern the defense-suppressing activities of effectors [18], to characterize particular defense-signaling mutants, such as *mpk4* [19] and *pmr4* [20], and to model the genetic network controlling the *Arabidopsis* response to *Pseudomonas syringae* pv. *maculicola* at 24h after inoculation [16]. Microarray technology has been prevailing over the past decade mainly because of its high throughput nature. Thousands of genes are examined concurrently under the same conditions. This allows identification of groups of genes exhibiting similar responses to different experimental conditions, hence of, group of genes that are likely to be controlled by similar regulatory mechanisms. Therefore, identifying genes with similar behaviors is important in order to identify and to understand the underlying machinery driving the biological process. From a computational perspective, this is a classification problem.

In this paper, as a follow-up to an expansion of our earlier study on *npr1* mutant [21], we studied the concerted effect of NPR1 and TGA factors at the onset of SAR. We used four genotypes, the Columbia wild type, the mutant *npr1,* double mutant of group I TGA factors *tga1 tga4* and the triple mutant of group II TGA factors *tga2 tga5 tga6*. Our objective is to understand the role and behavior of NPR1 and

Manuscript received December 15, 2009. This work is supported by Genomics and Health Initiative and Institute for Information Technology of the National Research Council Canada. This is publication NRCXXXXX of the National Research Council.

A. B. Tchagang, S. Phan, H. Bérubé, F. Famili, and Y. Pan are with the Institute for Information Technology of the National Research Council of Canada, 1200 Montreal Road, Ottawa, ON K1A 0R6, Canada: 613-993-7899; fax: 613-952-0215; (e-mails: {alain.tchagang, sieu.phan, hugo.berube, fazel.famili, youlian.pan} @ nrc-cnrc.gc.ca).

H. Shearer and P. Fobert are with the Plant Biotechnology Institute of the National Research Council Canada, 110 Gymnasium Place, Saskatoom, SK S7N 0W9, Canada (e-mail: {heather.shearer, pierre.fobert} @ nrc-cnrc.gc.ca).

various TGA factors and their target genes at the onset of SAR.

We cast the problem of mining the huge amount of gene expression data collected into the classical frequent itemset mining problem and define a brute force algorithm to tackle it. Frequent itemset mining [22]-[23] is a key technique for the analysis of binary matrices. In the binary representation, a frequent itemset corresponds to a submatrix of 1s containing a sufficiently large set of rows. Although frequent itemset mining was originally developed to discover association rules [24]-[25], its broader application provides the basis for subspace clustering and for building classifiers [26]-[27]. In these applications the ultimate goal is to discover interesting associations between object and attribute sets, rather than associations among attributes alone. In gene expression data analysis, for example, the joint discovery of both the set of conditions that significantly affect gene regulation and the set of coregulated genes is of great interest. Here, unlike in the classical frequent itemset mining where one is interested in the largest *all-1* submatrix, we exploited the fact that we are dealing with few experimental conditions and used a brute force approach to identify all the *all-1* submatrices. These submatrices are then used to predict and establish relationship between NPR1 and TGA transcription factors on one hand and their target genes on the other. Results obtained showed that most of the genes that responded to SA challenge are in fact dependent on one or more of the NPR1 and TGA transcription factors tested in this study. The rest of this paper is organized as follows. We first describe the materials used and give some definitions. Then, we present the frequent itemset mining algorithm; some application results in III, which is followed by a discussion and conclusion.

II. Materials and Definitions

A. Definitions

We define a gene expression matrix using either an N × M matrix (**Eq. 1a**), or using a set (**Eq. 1b**).

$$A = \begin{bmatrix} a(1,1) & a(1,2) & \cdots & a(1,m) & \cdots & a(1,M) \\ a(2,1) & a(2,2) & \cdots & a(2,m) & \cdots & a(2,M) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a(n,1) & a(n,2) & \cdots & a(n,m) & \cdots & a(n,M) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a(N,1) & a(N,2) & \cdots & a(N,m) & \cdots & a(N,M) \end{bmatrix} \quad (1a)$$

$$S = \{G, C\} \quad (1b)$$

where $G = \{g(1), g(2), \ldots, g(n), \ldots, g(N)\}$ represents the set of genes corresponding to the rows of the gene expression matrix. $C = \{c(1), c(2), \ldots, c(m), \ldots, c(M)\}$ represents the set of experimental conditions, time points, or tissue samples corresponding to the columns of the gene expression matrix. The entry $a(n,m)$ of the gene expression matrix (**Eq. 1a**) corresponds to the expression level of the nth gene, under the mth condition. $a(n,:) = [a(n,1)\ a(n,2)\ \ldots\ a(n,m)\ \ldots\ a(n,M)]$ is a $1 \times M$ vector corresponding to the expression level of $g(n)$ under the M conditions. $a(:,m) = [a(1,m)\ a(2,m)\ \ldots\ a(n,m)\ \ldots\ a(N,m)]^T$ is an $N \times 1$ vector corresponding to the expression level of the N genes under $c(m)$ or the mth condition.

B. Gene Expression Data

The gene expression data used in this study was obtained using Affymetrix *Arabidopsis* genechip consisting of 22810 probes. The Columbia wild-type, mutant *npr1*, double mutant *tga1 tga4,* and triple mutant *tga2 tga5 tga6* were treated with *salicylic acid* for 0, 1, and 8 hours, each time point has three biological replicates. After data preprocessing and normalization, we ended up with 2613 genes with significant expression level. We took the mean of their replicates, set the Columbia wild-type as our baseline and take the $\log_2$ ratio of the mutant gene expression levels over the wild-type at respective time points. We then discretized our gene expression matrix into three numbers (-1, 0, and 1) for a given threshold δ, corresponding to down-regulation, constant, and up-regulation relative to the baseline (wild-type) respectively. In other words, if the $\log_2$ ratio $a(n,m) \geq \delta$ then we set its value to 1. If $a(n,m) \leq -\delta$, we set its value to -1, and the value is set to 0 if $-\delta < a(n,m) < \delta$. It is important to assess the effects of the δ on the discretization procedures. This is done by performing a simple sensitivity analysis in which the parameter δ is perturbed about its selected value [31]. It is enough to consider one or two values for δ below and above its selected numerical. In this study we used $\delta = 0.2$. It was inferred based on the expected level of noise generated during microarray experiment. Hence, we ended up with three N × M matrices, each corresponding to one of the three time points: 0h, 1h, and 8h, with N = 2613 rows (genes) and M = 3 columns, corresponding to the three mutant sets: *npr1, tga1 tga4*, and *tga2 tga5 tga6*.

Discretization of gene expression data [28] is widely used in computational biology and bioinformatics as preprocessing step to several reverse engineering methodologies of the genetic regulatory network from the observed gene expression data [29]-[31]. For example, it is common to model the behavior of a gene using logical function such as in Boolean networks [31], where 1 means that the gene is *"on"* representing active, and 0 means that the gene is *"off"* representing inactive.

Given the N × M discretized gene expression matrix $D = [d_{nm}]$ with set of genes $G = \{g_1, \ldots, g_N\}$ and set of experimental conditions $C = \{c_1, \ldots, c_M\}$, our goal is to identify the set of genes that are control by the TFs tested in this study at a given time point, to study similarities and differences between them, and to infer a temporal transcriptional regulatory network controlling SAR in *A. thaliana*. Here, we cast the problem into a frequent itemset mining problem and seek associations between transcription factors and their target genes at each time point using a brute force algorithm.

III. METHODOLOGY

The approach used in this study has two main parts: matrix decomposition and identification of all the *all-1* submatrices.

A. Matrix Decomposition

The first step of the approach used in this study consists of expressing the discretized gene expression matrix as the sum of its distinct three values (-1, 0, 1) and their corresponding binary matrices (D_{-1}, D_0, D_1) as: $D = -1D_{-1} + 0D_0 + 1D_1$. For example,

$$D = \begin{bmatrix} -1 & 0 & -1 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ -1 & 0 & -1 \end{bmatrix} = -1\begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix} + 0\begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} + 1\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \quad (2)$$

This is the first important step of our algorithm. Because, after this decomposition, we have a clear picture of which genes may be downregulated, stay constant or upregulated by a group of TFs. For example, all the entries in D_{-1} that are 1 correspond to the set of genes that are downregulated in the corresponding mutants thus the corresponding TFs have an upregulation function on these genes. Furthermore, the entries in D_0 that are 1 correspond to the genes that stay constant in the corresponding mutants. In other words, the corresponding TFs do not have an effect on these genes. Finally, the entries in D_1 that are 1 correspond to the genes upregulated in the corresponding mutants, thus the corresponding TFs have a downregulation function on these genes. In this study, we will call D_{-1}, D_0, and D_1 the upregulated, constant, and downregulated matrices respectively. This is relative to the action of the TFs.

NB. Note here that downregulation (upregulation) in the mutants means that the corresponding set of TFs has the opposite function on the target genes that is upregulation (downregulation).

B. All-1 Submatrices Identification

The second part of the algorithm consists of identifying all the *all-1* submatrices from D_{-1}, D_0, and D_1. An *all-1* submatrix $B_k = [b_{ij}]$ is any submatrix of D_{-1}, D_0, or D_1 whose entries are all 1. Finding such matrices from a binary matrix is well known in the data mining community and it is referred to as the frequent itemset mining problem. That is, given an N-by-M (0,1)-matrix, of all its submatrices of all 1's which is the largest? By *largest* we mean that the submatrix has the most entries. Here, unlike in the classical frequent itemset, we are not only interested in the largest *all-1* submatrix but in all the *all-1* submatrices. The ultimate goal is to discover all the interesting associations between objects (genes) and attribute sets (transcription factors), rather than associations among attributes alone. Identification of the entire *all-1* submatrices from a binary matrix is shown in the data mining literature to be *NP*-complete when the set of attributes becomes very large [25]. Here, we take advantage of the fact that we are dealing with few experimental conditions (three in this case) and apply a brute force technique to identifying all such submatrices. This is done using **Eq. 3**.

$$u_k \,.*\, r(n,:) = u_k \quad (3)$$

The operator '.*' corresponds to the element wise product of two vectors. For example, [0 1 1].*[1 0 1] = [0 0 1]. Since we are dealing with only three experimental conditions that is *npr1, tga1 tga4,* and *tga2 tga5 tga6*, the set U whose elements are u_k corresponds to the following seven ($7 = 2^3-1$) row vectors: U = {[0 0 1]; [0 1 0]; [0 1 1]; [10 0]; [1 0 1]; [1 1 0]; and [1 1 1]}. The entries of each u_k correspond to NPR1, TGA2 TGA4 and TGA2 TGA5 TGA6 respectively. For example u_k = [1 0 0] is used to identify the set of genes that are upregulated by NPR1 in D_{-1}, downregulated by NPR1 in D_1 and set of genes with no NPR1 effect in D_0. u_k = [1 1 1] is used to identify set of genes that may be downregulated, upregulated by the three set of TFs (NPR1, TGA2 TGA4 and TGA2 TGA5 TGA6) in D_1 and D_{-1} respectively, and set of genes that are independent of all the three sets of TFs (NPR1, TGA2 TGA4 and TGA2 TGA5 TGA6) in D_0.

The $r(n,:)$ corresponds to a row of either D_{-1}, D_0, or D_1. More precisely, U corresponds to the set of memberships and **Eq. 2** the condition to verify by any other genes, *i.e.* rows of D_{-1}, D_0, and D_1, in order to be a member of the corresponding *all-1* submatrix $B_k = [b_{ij}]$. For example, if we consider the toy example define above (**Eq. 2**), u_k = [1 0 1] will tell us that genes 1 and 4 are simultaneously downregulated under conditions 1 and 3, thus upregulated by the corresponding TFs. Therefore, genes 1 and 4 under conditions 1 and 3 form an *all-1* submatrix. This is easily shown by proving that only row 1 and row 4 verify **Eq. 3** above: ([1 0 1] .* $r(1,:)$ = [1 0 1] and [1 0 1] .* $r(4,:)$ = [1 0 1], where $r(1,:)$ and $r(4,:)$ are the 1st and 4th rows of the toy example respectively). The brute force approach that we defined and used in this study is guaranteed to identify all the *all-1* submatrices ($B_k = [b_{ij}]$) from D_{-1}, D_0, or D_1.

The cardinality of the set of memberships U corresponds to the maximum number of all the *all-1* submatrices that can be found in a binary matrix, which is 7 in the current application case. In general, given that we are dealing with binary numbers, the elements u_k of U can be chosen as the binary representation of numbers from *1* to 2^M-1 on M bits, where M is the number of columns of the binary matrix. The drawback of the *brute force* approach used here is that, the complexity of the algorithm will grow exponentially as M becomes very large $\sim O(2^M \times N \times M \times L)$ (Section D). In this study, because of the relatively small number of columns, taking this combinatorial approach does not incur expensive computations.

C. Algorithms

Algorithms 1 and **Algorithm 2** are used for matrix decomposition and identification of all the *all-1* submatrices respectively. Recall that we have defined an *all-1* submatrix above as: $B_k = [b_{ij}]$, It can also be defined using the set notation that is: $B_k = \{I_k, J_k\}$, with I_K subset of genes ($I_K \subseteq G$), and J_K a subset of conditions ($J_K \subseteq C$), with $i \in I_k$ and $j \in J_k$. $k = 1$ to K, where K is the maximum number of *all-1* submatrices that can be found in a binary matrix: $K = 2^M-1$. L, N, and M are the number of distinct elements, rows, and columns of D respectively.

1) Algorithm 1: matrix decomposition.

Input:

- D = discretized gene expression matrix
- $\alpha = [-1\ 0\ 1]$ set of discrete values

Output:

- D_{-1}, D_0, and D_1 = binary matrices associated with upregulation, constant, and downregulation respectively

```
Begin,
    [N,M] = size (D);
     D-1 = zeros(N,M); D0 =zeros(N,M);  D1 = zeros(N,M);
    For n = 1 to N
        For m = 1 to M
            If D(n,m) == -1
                D-1(n,m) = 1
            Elseif D(n,m) == 0
                D0(n,m) = 1
                Elseif D(n,m) == 1
                D1(n,m) = 1
            End
        End
    End
  End Begin
```

---

2) Algorithm 2: All-1 submatrices identification.

Input:

- D_{-1}, D_0, and D_1 (from Algorithm 1)
- $U = \{u_k\}$ = set of memberships
- $C = [c(1)\ c(2)\ \ldots\ c(m)\ \ldots\ c(M)]$ = set of columns
- $G = [g(1)\ g(2)\ \ldots\ g(n)\ \ldots g(N)]^T$ = set of genes

Output:

- I = set of genes in *all-1* submatrices
- J = set of conditions in *all-1* submatrices

```
Begin,
   Z(:,:,1) = D-1; Z(:,:,2) = D0; Z(:,:,3) = D1;
   [N,M,L] = size (Z); I = []; J = [];
     For l = 1 to L
        For k = 1 to K
           J{k,l} =C(find(b(k) == 1);
           For n = 1 to N
              If U(k,:). .* Z(n,:,l) == U(k,:)
                  I{k,l} = [I{k};g(n)];
              End
           End
        End
     End
End Begin.
```

D. Complexity Analysis

We can easily estimate the complexity of the proposed approach. The matrix decomposition algorithm requires about $(N\times M)$ operations. Algorithm for *all-1* submatrices identification uses $O((N\times M+N+P+P\times M) \times L\times (2^M-1))$ operations because we perform $\sim(2^M-1)N\times M$ binary multiplications, N comparisons, and P assignments $\sim L\times (2^M-1)$ times. Here, 2^M-1 is the maximum number of *all-1* submatrices and P the number of times the membership equation (**Eq. 3**) is verified. Thus the complexity of our brute force approach is $\sim O(N\times M\times L\times 2^M)$. Note that L = 3 in this case and that we are dealing with matrices that have few columns, with M = 3 in this case. Thus the complexity here will be $\sim O(2613\times 3\times 3\times(2^3-1))=164619$. As we mentioned earlier, when M becomes very large, the complexity increases exponentially (2^M) and we end up with large sets of *all-1* submatrices. Observe that: $M = 5 \rightarrow 2^5 = 32$; $M = 7 \rightarrow 2^7 = 128$; $M = 10 \rightarrow 2^{10} = 1024$; $M = 20 \rightarrow 2^{20} = 1048576$.

IV. Results

Using the above described materials and methodologies, we obtained the following results.

A. Potential Transcription Factor Gene Interactions

Fig. 1-3 show the number of genes that are controlled by NPR1, TGA1 TGA4, and TGA2 TGA5 TGA6 respectively at different time point experiments: 0h, 1h, 8h, and the overlap between them. For example, at 0h, 537 and 580 genes are downregulated and upregulated by NPR1 respectively. 0h_1h_8h shows the number of genes (7 downregulated and 9 upregulated) that stay under the influence of NPR1 during the three time points. Thus different set of genes are triggered by NPR1 at different time points. The number of genes that are upregulated or downregulated by NPR1 is an increase function of time compared to the ones that are NPR1 independent which is a decrease function of time. Unlike the effect of NPR1, the number of genes regulated by TGA1 TGA4 is a parabolic function of time, with the maximum (upregulation and downregulation) or minimum (constant genes) at 1h.

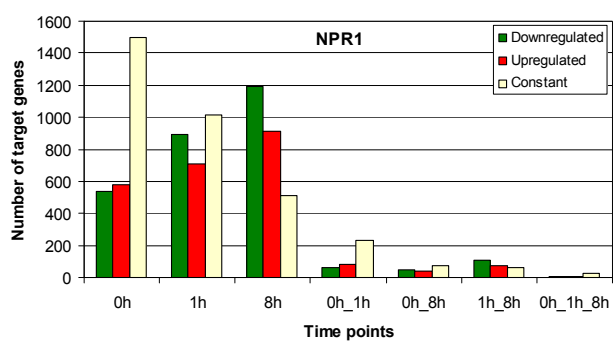

Fig. 1 NPR1 regulated genes. The interaction could be direct or indirect.

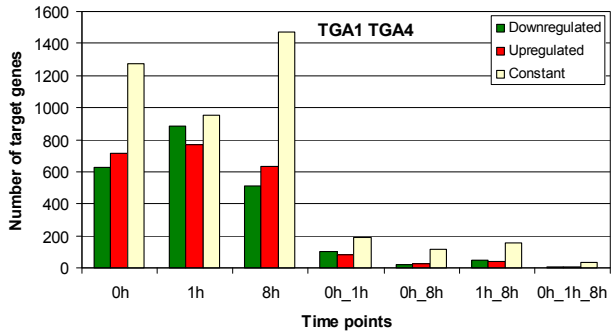

Fig. 2 TGA1 TGA4 regulated genes. The interaction could be direct or indirect.

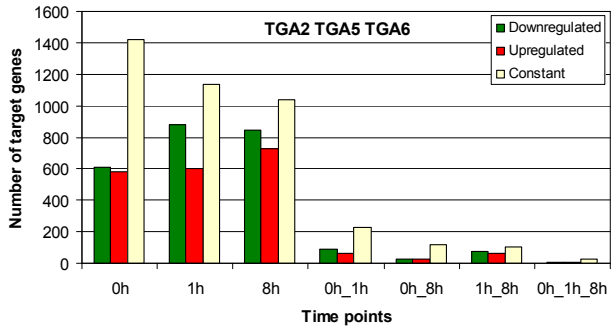

Fig. 3 TGA2 TGA5 TGA6 regulated genes. The interaction could be direct or indirect.

The number of genes that is TGA2 TGA5 TGA6 independent is similar to that of NPR1 but decrease slowly compared to the NPR1 ones. This indicates that larger number of genes is regulated by NPR1 than that by TGA2 TGA5 TGA6 and TGA1 TGA4. On the other hand, the number of genes that is downregulated by TGA2 TGA5 TGA6 increases from 0h to 1h and stay constant up to 8h. Whereas, the ones that are upregulated stay constant from 0h to 1h and increase after that.

In all three cases, results obtained showed that there are few overlaps between the set of genes controlled by a given set of TFs at different time points. This observation suggests that several of the genes that participate in SA challenge have an impulse behavior and that different set of genes are triggered at different time point to participate in the defense mechanism in response to a pathogen infection.

B. Similarities and Differences between Transcription Factors

Fig. 4-7 show the set of genes controlled by at least two of the TFs tested in this study at different time points: 0h, 1h, 8h, and the overlap between them.

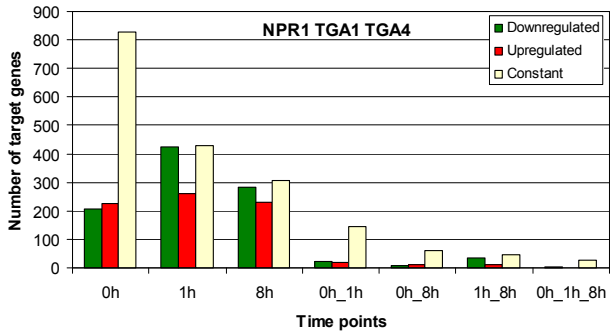

Fig. 4 NPR1 and TGA1 TGA4 regulated genes. The interaction could be direct or indirect.

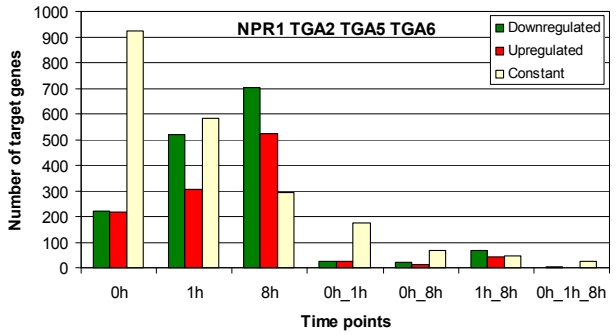

Fig. 5 NPR1 and TGA2 TGA5 TGA6 regulated genes. The interaction could be direct or indirect.

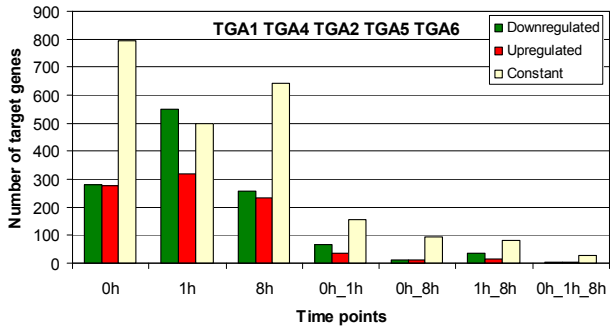

Fig. 6 TGA1 TGA4 and TGA2 TGA5 TGA6 regulated genes. The interaction could be direct or indirect

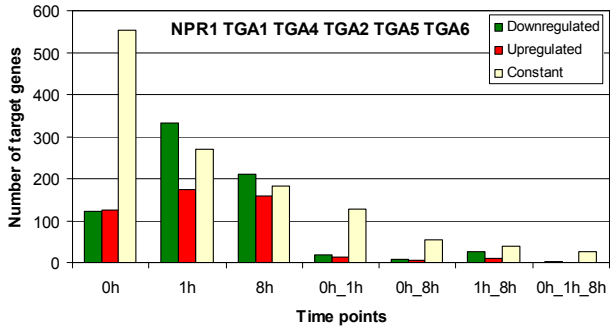

Fig. 7 NPR1, TGA1 TGA4, and TGA2 TGA5 TGA6 regulated genes. The interaction could be direct or indirect.

The number of genes upregulated or downregulated by NPR1 and TGA1 TGA4 TFs simultaneously, is a parabolic function of time, with the maximum at 1h, whereas the ones that may not be under both their influence is a decrease function of time. The number of genes upregulated or downregulated by NPR1 and TGA2 TGA5 TGA6 TFs simultaneously, is an increase function of time, whereas the ones that may not be under both their influences is a decrease function of time. Like in the TGA1 TGA4, the number of genes that are under the influence of both TGA1 TGA2 and TGA2 TGA5 TGA6 is a parabolic function of time, with the maximum (upregulation and downregulation) or minimum (constant genes) at 1h.

Finally, the number of genes that are upregulated or downregulated by all the three sets of TFs tested in this study NPR1, TGA2 TGA4 and TGA2 TGA5 TGA6 simultaneously is a parabolic function of time, with the maximum at 1h, and the ones that might not be under their influence is a decrease function of time. Here again, in all cases, results obtained showed that there are few overlaps between the set of genes regulated by the considered group of TFs at different time points.

C. Time Varying Transcriptional Network Model

Combining the above results, we built a preliminary wiring diagram of the genetic network of SAR in *Arabidopsis thaliana* at 0h, 1h, and 8h: **Fig. 8-10**. In the following diagrams, ▬ (action could be inclusive or exclusive), ▼ (direction of regulation unknown), ▼ (up regulation), ▲ (down regulation), and D (AND gate: combined action of the inputs). At 0h for example, only 121 and 125 genes are downregulated and upregulated respectively by the combined action of the three TFs. Furthermore, at 1h only 333 and 178 genes are down and up regulated, respectively, by the combined action of the three TFs, whereas at 8h only 211 and 158 genes are down and up regulated, respectively, by the combined action of the three TFs. The number of NPR1 targeted genes is less than that of TGA1 TGA4 and TGA2 TGA5 TGA6 at 0h. But at 8h, it is the reverse situation where the number of NPR1 targeted genes is higher than those regulated by TGA1 TGA4 and TGA2 TGA5 TGA6, respectively. This is consistent with the fact that NPR1 gene expression in the Columbia wild type was initially moderate but drastically increased after 0 hour and continued increasing until 8 hours after SA treatment (**Fig. 11**).

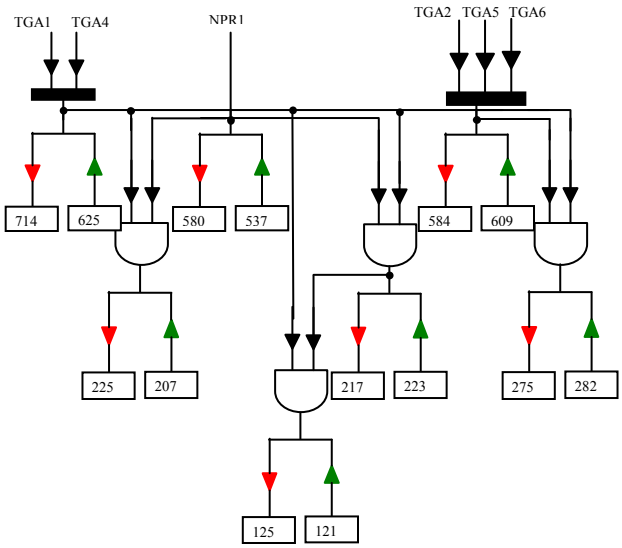

Fig. 8 SAR transcriptional network at 0h

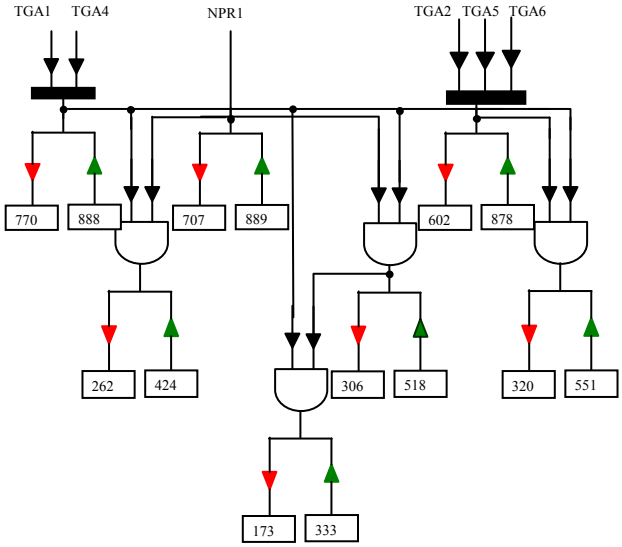

Fig. 9 SAR transcriptional network at 1h

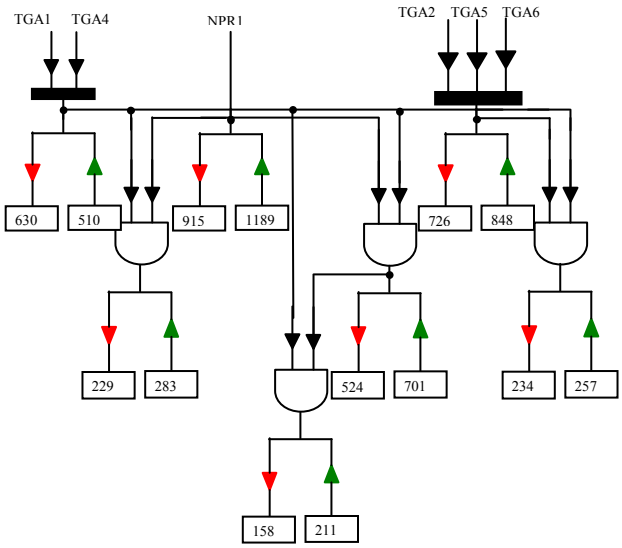

Fig. 10 SAR transcriptional network at 8h

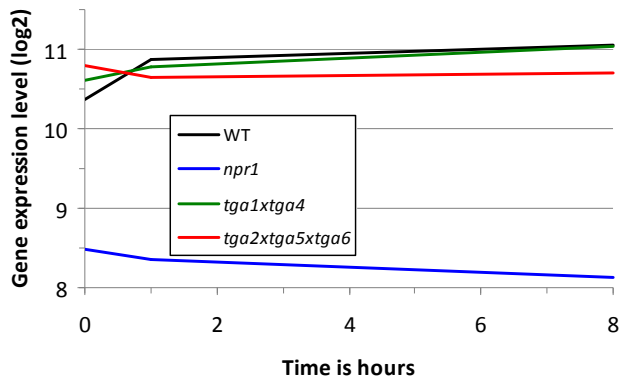

Fig. 11 Expression profile of NPR1 in Columbia wild type (WT), mutant *npr1,* double mutant (*tga1 tga4*), and triple mutant (*tga2 tga5 tga6*).

Gene ontology (GO) analysis using the GOAL software [32] [http://bioinfo.iit.nrc.ca/GOAL/] reveals that several of the genes that are regulated by the three set of TFs (NPR1, TGA1 TGA4, and TGA2 TGA5 TGA6) at 8h (**Fig.10**) for example show that these genes belong to response to stimulus (GO:0050896; p-value = 2.2e-06), stress (GO:0006950; p-value = 1.8e-05), abiotic stimulus (GO:0009628; p-value = 5.7e-04), biotic stimulus (GO:0009607; p-value = 2.3e-03), and defense mechanism (GO:0006952; p-value =6.9e-03). These responses are expected because the plant is under pathogen attack. They also confirm the fact that the TFs tested in this study are known to play major roles in plant defense mechanism. Another interesting observation is that significant number of genes that stay constant across the three experimental time points are responsible for photosynthesis (GO:0015979; p-value = 3.1e-10, and KEEG pathway ath00195 p-value = 3.5e-04). These results confirmed that during pathogen attacks, the plant is mobilized for defense.

V. Discussion

In this work, we used similarities among expression profiles of plants with multiple mutations in key transcription factors in the defense signaling network. We study the network dynamics over the time series after treatment of salicylic acid (SA), which mimic a pathogen infection. We found that most SA-responsive genes were affected by at least one mutation and that most affected genes fit one of a few simple patterns of regulation. We then provided a first glimpse into the temporal pattern of the genetic network controlling systemic acquired resistance in *Arabidopsis*.

Fig. 1-7 show that there are few overlaps between the number of genes controlled by a given group of TFs at different time points. This observation suggests that several of the genes that participate in SAR have an impulse behavior and that different set of genes are targeted by the corresponding transcription factors at different time point during a response to a pathogen infection. On the other hand these results show that the true behavior of the underlying biological process is capture at different time point, with each time point containing a unique piece of information that should be integrated in order to get the whole picture underlying the signaling pathway during SAR. Therefore, studies, such as [16], that had only focus on a single time point to infer genetic information and generalize the results to describe the biological system under study could miss out important and interesting chunk of information.

The models describe in **Fig. 8-10** show the different regulatory mechanism depicted using our combinatorial brute force approach. We observed that, at each time point, fraction of genes are regulated by only one of the three sets of TFs, while others are subject to the collective regulation of two of the three sets of TFs; and yet another set of gene are subject to regulation of all three sets of the TFs. We also found only few genes were independent of one, two, or all sets of the TFs. This observation suggests that most of the genes that responded to SA challenge are in fact dependent on one or more of NPR1 or the TGA factors. This notion is consistent with the fact that NPR1 binds to specific TGA to regulate SA genes [10].

Our analysis further revealed that the number of triggered NPR1 genes is less than that of TGA1 TGA4 and TGA2 TGA5 TGA6 at 0h. But at 8h, it becomes the reverse situation where the number of NPR1 genes is higher than the ones regulated by TGA1 TGA4 and TGA2 TGA5 TGA6. This indicates that NPR1 kicks off slowly during a pathogen attack and increases its action exponentially to sustain defense in plant against pathogen attacks. This is consistent with the fact that NPR1 play major role in plant defense mechanism [9]-[10].

In this study, we used a broader application of frequent itemset mining to tackle our problem. We exploited the fact that we were dealing with matrices with only few columns less than 9 (3 in this study), and used a deterministic combinatorial approach to identify all the *all-1* submatrices. In real data applications a “1” or “-1” can be accidentally recorded as “0” and vice versa. In a gene expression data the noise can arise from measurement, stemming from the underlying experimental technology and the stochastic nature of the studied biological behavior. In addition, uncertainty involved in choosing the proper thresholds when imputing discrete observations from the continuous gene expression values can introduce error. While frequent itemsets and the algorithms that generate them have been well studied, the difficulties that arise from noise have not been adequately addressed [22]-[27]. In general, the noise present in real applications undermines the ultimate goal of traditional frequent itemset algorithms. In fact, when noise is present, classical frequent itemset algorithms discover multiple small fragments of the true itemset, but miss the true itemset itself. The problem is worse for the most interesting, longer itemsets as they are more vulnerable to noise. Therefore, proper gene expression data pre-processing and normalization algorithms should be used to reduce the level of noise in the dataset before its analysis using our combinatorial deterministic frequent itemsets approach.

VI. Conclusion

In this paper, we studied the defense mechanism of *Arabidopsis thaliana* using a deterministic broader frequent itemset mining approach. We used expression profiling of wild-type and mutants of three sets of transcription factors to define the major patterns of regulation governing the response to salicylic acid treatment, thereby creating a wiring diagram to reveal the number of genes that are regulated by one or more sets of the transcription factors at various stages of defense responses. The temporal model describes the relationships among the transcription factors, and defines groups of genes that are subject to similar regulation. We found that most up and down regulated genes fit one of a small number of regulatory patterns defined by the effects of these mutations, demonstrating that most of the genes that responded to salicylic acid challenge are in fact dependent on one or more of transcription factors, NPR1 or the TGA factors tested in this study.

References

[1] J. Malamy, J. P. Carr, D. F. Klessig, and I. Raskin, "Salicylic acid: A likely endogenous signal in the resistance response of tobacco to viral infection." Science 250:1002-1004, 1990.

[2] J. P. Métraux, H. Signer, J. Ryals, E. Ward, M. Wyss-Benz, J. Gaudin, K. Rascdorf, E. Schmid, W. Blum, and B. Inverardi, "Increase in salicylic acid at the onset of systemic acquired resistance in cucumber." Science 250:1004-1006, 1990.

[3] T. Gaffney, L. Friedrich, B. Vernooij, D. Negrotto, G. Nye, S. Uknes, E. Ward, H. Kessmann, and J. Ryals, "Requirement for salicylic acid for the induction of systemic acquired resistance." Science 261: 754-756, 1993.

[4] L. C. Van Loon, E. A. Van Strien, "The families of pathogenesis-related proteins, their activities, and comparative analysis of PR-1 type proteins." Physiological and Molecular Plant Pathology 55: 85-97, 1999.

[5] T. Eulgem , " Regulation of the Arabidopsis Defense Transcriptome," TRENDS in Plant Science, Vol.10 No.2 February 2005

[6] J. D. Jones and J. L. Dangl, "The plant immune system," *Nature,* vol. 444, pp. 323-9, Nov 16 2006.

[7] C. Després, C. DeLong, S. Glaze, E. Liu, P. Fobert, "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.

[8] T. P. Delaney, L. Friedrich, and J. A. Ryals, *Arabidopsis* signal transduction mutant defective in chemically and biologically induced disease resistance. Proc. Natl. Acad. Sci. USA 92:6602-6606, 1995

[9] C. M. J. Pieterse, and L. C. Van Loon , "NPR1: the spider in the web of induced resistance signaling pathways," *Current Opinion in Plant Biology*, 7:456-464, 2004

[10] C. Johnson , E. Boden, and J. Arias, Salicylic Acid and NPR1 Induce the Recruitment of trans-Activating TGA Factors to a Defense Gene Promoter in Arabidopsis," *The Plant Cell*, Vol. 15, 1846-1858, August 2003.

[11] M. Kesarwani, J. Yoo, and X. Dong "Genetic Interactions of TGA Transcription Factors in the Regulation of Pathogenesis- Related Genes and Disease Resistance in *Arabidopsis thaliana*", Plant Physiol. 44:336-346, 2007

[12] M. Jakoby, B. Weisshaar, W. Droge-Laser, J. Vicente-Carbajosa , J. Tiedemann , T. Kroj, F. Parcy, "bZIP transcription factors in Arabidopsis." Trends Plant Sci 7: 106-111, 2002

[13] X. Dong, "NPR1, all things considered," *Curr Opin Plant Biol,* vol. 7, pp. 547-52, Oct 2004.

[14] C. Xiang, Z. Miao, E. Lam, "DNA-binding properties, genomic organization and expression pattern of TGA6, a new member of the TGA family of bZIP transcription factors in Arabidopsis thaliana." Plant Mol Biol 34: 403-415, 1997.

[15] C. Després, C. Chubak, A. Rochon, R. Clark, T.Bethune, D. Desveaux, and P. R. Fobert, "The Arabidopsis NPR1 Disease Resistance Protein Is a Novel Cofactor That Confers Redox Regulation of DNA Binding Activity to the Basic Domain/Leucine Zipper Transcription Factor TGA1", *Plant Cell* 15:2181-2191, 2003

[16] L. Wang, R. M. Mitra, K. D. Hasselmann, M. Sato, L. Lenarz- Wyatt, J. D. Cohen, F. Katagiri, and J. Glazebrook, "The genetic network controlling the Arabidopsis transcriptional response to Pseudomonas syringae pv. maculicola: roles of major regulators and the phytotoxin coronatine," *Mol Plant Microbe Interact,* vol. 21, pp. 1408-20, Nov 2008.

[17] B. Schwessinger and C. Zipfel, "News from the frontline: recent insights into PAMP-triggered immunity in plants," *Curr Opin Plant Biol,* vol. 11, pp. 389-95, Aug 2008.

[18] W. Truman, M. T. Zabala, and M. "Type III effectors orchestrate a complex interplay between transcriptional networks to modify basal defence responses during pathogenesis and resistance." *Plant J.* 46:14-33, 2006.

[19] M. Petersen, P. Brodersen, H. Naested, E. Andreasson, U. Lindhart, B. Johansen, H. B. Nielsen, M. Lacy, M. J. Austin, J. E. Parker, S. B. Sharma, D. F. Klessig, R. Martienssen, O. Mattsson, A. B. Jensen, , and J. Mundy. "*Arabidopsis* map kinase 4 negatively regulates systemic acquired resistance." *Cell* 103:1111-1120., 2000.

[20] M. T. Nishimura, M. Stein, B. H. Hou, J. P. Vogel, H. Edwards, and S. C. Somerville. "Loss of a callose synthase results in salicylic acid-dependent disease resistance." *Scienc*e 301:969-972, 2003.

[21] Y. Pan, J. D. Pylatuik, J. Ouyang, F. Famili, P. Fobert, "Discovery of functional genes for systemic acquired resistance in Arabidopsis thaliana through integrated data mining", JBCB, Vol. 2, No. 4 , pp. 639-655, 2004.

[22] C. Creighton, S. Hanash. Mining gene expression databases for association rules. Bioinformatics. 2003 Jan;19(1):79-86.

[23] J. Liu, S. Paulsen, X. Sun, W. Wang, A. Nobel, J. Prins, "Mining Approximate Frequent Itemsets In the Presence of Noise: Algorithm and Analysis," *SIAM Conference on Data Mining* (SDM), 2006

[24] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases In SIGMOD 1993.

[25] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A Verkamo. Fast discovery of association rules. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, Advances in Knowledge Discove and Data Mining, chapter 12, pages 307328. AAAI Pre 1996.

[26] C. Becquet, S. Blachon, B. Jeudy, J.F. Boulicaut, O. Gandrillon. Strong-association-rule mining for largescale gene-expression data analysis: a case study on humaMining gene expression databases for association rules in SAGE data. Genome Biol. 2002.

[27] C. Gao , Anthony K. H. Tung, X. Xin, P. Feng , J. Yang. "FARMER: Finding Interesting Rule Groups in Microarray Datasets".

[28] R. G. Pensa, C. Leschi, J. Besson, and J. F. Boulicaut, "Assessment of discretization techniques for relevant pattern discovery from gene expression data," In proceedings, 4th Workshop on Data Mining in Bioinformatics, 2004.

[29] J. Faith, T. Gardner, "Reverse-engineering transcription control networks." *Phys Life Rev* **2**: 65–88, 2005.

[30] R. Laubenbacher, and B. Stigler, "A computational algebra approach to the reverse engineering of gene regulatory networks." *Journal of Theoretical Biology*," **229**, 523-537, 2004.

[31] I. Shmulevich, E. R. Dougherty, S. Kim, W. Zhang, "Probabilistic Boolean Networks: A Rule-based Uncertainty Model for Gene Regulatory Networks," *Bioinformatics*, Vol. 18, No. 2, pp. 261-274, 2002.

[32] A. Tchagang, A. Gawronski, H. Bérubé, S. Phan, F. Famili, and Y. Pan. "GOAL: A Software Tool for Assessing Biological Significance of Genes group**."** *BMC Bioinformatics*. ***under revision***, 2009