Electronic Journal of Biotechnology 17 (2014) 79–82

Contents lists available at ScienceDirect

Electronic Journal of Biotechnology

Data mining and influential analysis of gene expression data for plant resistance gene identification in tomato (*Solanum lycopersicum*)

Francisco Torres-Avilés [a,*], José S. Romeo [a], Liliana López-Kleine [b]

[a] *Departamento de Matemática y Ciencia de la Computación, Universidad de Santiago de Chile, Av. Libertador Bernardo O'Higgins 3363, Santiago, Chile*
[b] *Departamento de Estadística, Universidad Nacional de Colombia, Edificio 405, Oficina 325, Ciudad Universitaria, Bogotá, Colombia*

ARTICLE INFO

Article history:
Received 19 March 2013
Accepted 14 January 2014
Available online 5 February 2014

Keywords:
Classification
Data mining
Functional gene prediction
GEE models
Gene expression data
Plant immunity genes

ABSTRACT

Background: Molecular mechanisms of plant–pathogen interactions have been studied thoroughly but much about them is still unknown. A better understanding of these mechanisms and the detection of new resistance genes can improve crop production and food supply. Extracting this knowledge from available genomic data is a challenging task.
Results: Here, we evaluate the usefulness of clustering, data-mining and regression to identify potential new resistance genes. Three types of analyses were conducted separately over two conditions, tomatoes inoculated with *Phytophthora infestans* and not inoculated tomatoes. Predictions for 10 new resistance genes obtained by all applied methods were selected as being the most reliable and are therefore reported as potential resistance genes.
Conclusion: Application of different statistical analyses to detect potential resistance genes reliably has shown to conduct interesting results that improve knowledge on molecular mechanisms of plant resistance to pathogens.

1. Introduction

Recent advances in genomic and post-genomic technologies have provided the opportunity to analyze genomic data that is publicly available in databases. Several molecular mechanisms can be better understood through the analysis of genomic data such as gene expression data. To reduce losses caused by plant pathogens, more understanding is needed about plant immunity mechanisms [1]. Losses caused by plant pathogens represent one of the most important limitations in crop production, and these losses can compromise food supply.

Plant immunity depends on the recognition of conserved Microbial Associated Molecular Patterns (MAMPs) or strain-specific effectors by Pattern Recognition Receptors (PPRs) or resistance (R) proteins, triggering MTI (MAMP Triggered-Immunity) and ETI (Effector Triggered-Immunity), respectively. Upon recognition, plants activate a complex network of responses that include signal transduction pathways, novel protein interactions and coordinated changes in gene expression. Detailed information concerning specific and punctual interactions between effector and resistance proteins has been accumulated in the past few years, and in some cases, a global picture for some of these interactions has been established [2,3,4,5]. Immunity networks have been described for model plants such as *Arabidopsis* and rice, primarily using yeast-two hybrid experiments [6,7]. Nevertheless each plant–pathogen interaction has its specificities [8] and can lead to the activation of resistance genes that were previously unknown.

Gene expression data, for example microarray data, are used in two ways: i) for detection of differentially expressed genes (comparing two conditions) and ii) for construction of clusters of genes with similar gene expression profile. The second objective is addressed by using multivariate and data mining methods in order to detect genes with similar expression patterns in a tomato time course experiment. On the other hand, resistance is modeled (a binary variable indicating if it is previously known if genes participate in immunity processes or not) considering the time measurement and the most influential genes of this model are identified. The influence analysis allows the detection of genes that produce an effect in the estimation process in order to identify crucial immunity genes. The aim of this study is to predict with original data analysis, a set of genes potentially participating in a functional group. Moreover, clustering and regression methods are compared and differences in predictions are discussed.

User friendly and reliable methods for functional gene prediction are needed in order to get a better knowledge of molecular mechanisms. Multivariate analysis is very simple to apply but often considered unreliable due to their descriptive nature. Therefore, we compared

[*] Corresponding author.
E-mail address: francisco.torres@usach.cl (F. Torres-Avilés).

Peer review under responsibility of Pontificia Universidad Católica de Valparaíso.

Production and hosting by Elsevier

0717-3458/$ – see front matter © 2014 Pontificia Universidad Católica de Valparaíso. Production and hosting by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ejbt.2014.01.003

these methods to classification and regression methods, which allow verifying the quality of predictions.

In a previous study a set of virulence factors was defined based on literature [9] and used to predict potential novel virulence factors based on linear and non-linear clustering methods combining all microarray experiments. Additionally, the GEE modeling and influential analysis were applied to these data [10] separately on both conditions and were able to confirm some virulence factors through influence analysis. Here, clustering, data mining and GEE regression coupled with influential analysis methods were applied to microarray data available at the Tomato Expression Database (TED) at http://ted.bti.cornell.edu [11]. These data sets compare two conditions in the field: healthy cherry tomato and *Phytophthora infestans* inoculated cherry tomato at four time points. Predictions of participation in immunity processes of gene products were performed separately for each condition in all analyses.

Conducting predictions separately allowed determining strong differences in prediction indicating that the shift in gene expression of genes participating in immunity processes is strong. Moreover, it was possible to cluster together or predict correctly more known resistance genes when conducting the analysis on inoculated tomatoes, which indicates that their activation is needed to detect similar co-expression patterns, and that information on most of these genes contained in healthy tomatoes is not enough either for de novo prediction through clustering nor for SVM prediction or identification using influential analysis after regression. In conclusion, the presented methodology, especially SVM, can be used for the prediction of new resistance genes, namely those that will be clustered together with genes that are not known to be resistance genes.

2. Materials and methods

2.1. Microarray data

The data sets were obtained from the Tomato Expression Database (http://ted.bti.cornell.edu/). In this study experiments were carried out using the same platform, namely the TOM1 DNA chip available at http://ted.bti.cornell.edu/cgi-bin/TFGD/miame/experiment.cgi?ID=E022. The experiments carried out by Christine Smart and collaborators [12] (accession number E022) were focused on, where gene expression profiling of infection of tomato by *P. infestans* in the field was studied. The goal of this experiment was to gain insight into the molecular basis of the compatible interaction between *P. infestans* and its hosts (with a major emphasis on the role of gene suppression). The data from that experiment was used separately for inoculated plants (condition I) vs. non-inoculated plants (condition NI) in the field. For this comparison four time points were available at 0, 12, 36 and 60 h with 8 replicates of each condition (32 experiments). Measurements on 13,440 genes were available.

Canonical immune protein domains (WRKY, TIR, NBS, kinase and LysM) from Pfam (http://pfam.sanger.ac.uk/) were searched in the tomato genome annotation file available at TED. We detected 174 genes coding for proteins with these domains and considered them as a set of “known resistance genes”.

2.2. Data pre-processing

As intensities are not directly comparable, data were calibrated by the method proposed by Huber et al. [13] by means of the vsn R package [14] available from http://www.bioconductor.org. More than calibrating the microarray data in terms of the mean, Huber's method allows the stabilization of variance. The variances of the transformed intensities become approximately independent of their expected values.

The transformation is similar to the logarithm in the high intensity range, but does not affect the differences between conditions at low intensity values, as the logarithm transformation does [13]. The used transformation and calibration are done as follows: $h(x) = \text{arsinh}(x)/s$, where x is the measured intensity and s the standard deviation of intensities in the replicates of each experiment.

In order to characterize time, it was necessary to add a new variable, called *period* which assumes the value 0 if *time* = 0 and 1 otherwise. The latter allows considering the case *period* = 0 as if it was the control condition before the inoculation of the bacterium. Given this data structure, the analysis was performed separately for inoculated and non-inoculated, respectively.

Therefore, the response variables for each of the fitted models were,

- First model: Y_{ijkt} represents the median value of the conditions of the *i*-th *gene* in inoculated plants, according to the *j*-th stage of *resistance* and the *k*-th period in *time t*, for $\text{i} = 1,...,13{,}440$, $\text{j} = 0,1$, $\text{k} = 0,1$ and $\text{t} = 0,1,2,3$.
- Second model: Y_{ijkt} denotes the median value of the conditions of the *i*-th *gene* in non-inoculated plants, according to the *j*-th stage of *resistance* and the *k*-th period in *time t*, for $\text{i} = 1,\ldots,13{,}440$, $\text{j} = 0,1$, $\text{k} = 0,1$ and $\text{t} = 0,1,2,3$.

The behavior of the median expression in each condition is then to be explained, separately, in order to identify those genes that produce an influence on estimation and inference.

2.3. Classification methods

Several unsupervised classification methods were applied: k-means cluster analysis using the centroids obtained from a hierarchical cluster analysis [15,16], Agglomerative Nesting — AGNES and Divisive Analysis clustering — DIANA [16] in order to cluster genes in several groups. For all these methods we used two different distance measures: Euclidean and Manhattan or Taxicab metric [17]. We also applied Kohonen's self organizing map [18], which is a highly appreciated algorithm for its ability to classify data into two dimensions and is based on neural networks. To define the best number of clusters, a visual inspection was performed using the dendrogram (not shown) obtained from the hierarchical clustering method using both distance measures. Despite the total number of clusters, the interest was in obtaining a main group of known resistance genes and concentrating in the main resistance cluster for validation and prediction.

Predictions of “resistance” based on a non-linear supervised classification kernel method called support vector machine (SVM), developed by Schölkopf and Smola [19] were also obtained. This method is a supervised method for which part of the data is needed to be used for training a classifier. In order to train the SVM classifier a sample of 1/3 of the genes (4480) was used. Predictions were therefore obtained only for 2/3 of the genes (8960). In order to apply this method, microarray data is needed to be mapped into a feature space by constructing a kernel. A Gaussian kernel was constructed and tuned the sole parameter of this kernel (Sigma) by leave-one-out cross validation on the training set optimizing the correct classification of known resistance genes into a homogeneous group.

2.4. Quasi-likelihood regression

A model using quasi-likelihood regression was fitted and an influence analysis was preformed, in order to identify genes that have outlying behavior and could therefore be implicated in immunity processes. Inference and estimation related to this class of models are specifically treated by Liang and Zeger [20] and Nelder and Pregibon [21].

In order to reach the main goal, given the selected model, four measures were computed to detect influential genes, which were obtained from residual analysis study. Below, the subscript $i[I]$ denotes the estimates considering all data without the cluster of genes i. Then, the influence of the i-th cluster of observations will be computed deleting the gene over the complete period of time, that is, four times. The measures computed for this analysis were,

- Pearson residuals deleting the l-th cluster, denoted by $r_p(l)$, is obtained from the expression

  $$E_l = B_{i[l]}\left(Y_{i[l]} - \hat{\mu}_{i[l]}\right),$$

  where $B_{i[l]}$ is an appropriate matrix of variability and $\hat{\mu}_{i[l]}$ is the location parameter, both estimated without the l-th cluster.
- The Leverage $Le(l)$ of the l-th cluster of observations is represented by the

  $$tr\left(H_{i[l]}\right) = tr\left(Q_{i[l]}W_{i[l]}\right),$$

  where $tr(\cdot)$ represents the trace of the matrix, $W_{i[l]}$ is a matrix of weights without the l-th observation and $Q_{i[l]}$ is the quadratic form associated to the i-th cluster without the l-th observation, with

  $$Q_i = X_i(X'WX)^{-1}X'_i.$$

- Cook's Distance. The measure is defined as

  $$\text{DCook}(l) = QF_l/p\hat{\varphi},$$

  with $QF_l = E_{l'}(W_{l'} - Q_{l'})^{-1}Q_{l'}(W_{l'} - Q_{l'})^{-1}E_l$. The parameter $\hat{\varphi}$ represents the dispersion of the model.
- DfBeta measure can be defined as the effect of deleting cluster i on the estimated parameter vector β. This measure is computed from

  $$\text{DfBeta}(l) = (X'WX)^{-1}X_{l'}\left(W_l^{-1} - Q_l\right)^{-1}E_l.$$

The effect on the specific k-th parameter of β is obtained from this expression.

The analysis was performed in order to study the possible influence of genes. After this analysis, a visual criterion was established for each study, specifically, influence is detected when $r_p(l) < -3$ or $r_p(l) > 10$, $Le(l) < -0.01$, DCook(l) > 10 and DfBeta(l) > |10|, which corresponds to the extreme cases. The same criteria are applied for males and females. For the influence diagnostic details, it is possible to review the work developed by Preisser and Qaqish [22] and Hammill and Preisser [23].

2.5. Proposal of candidates for biological validation

For all unsupervised clustering methods, we identified the two clusters that grouped most of the 174 known resistance genes and considered them to represent resistance gene clusters (RGCs). The most reliable predictions (intersection of all chosen methods) were chosen to assemble a list of potential resistance genes. This list was filtered for the genes with dissimilar sequences to human genes. To achieve this, a BLAST using the predicted protein sequence of the candidate genes was done using the NCBI BLAST-P tool. Only E-values and identities obtained from the best BLAST hit were retained. Proteins with E-values above 10^{-5} were retained.

2.6. Software

The data analysis for this paper was generated using SAS/STAT software, Version 9.2 of the SAS System for Windows [24] for the regression analysis and R-gui software, Version 2.15.1 [14]. Available methods in each software are used, so that no additional methods are necessary.

3. Results

3.1. Validation of known resistance genes

Different clustering methods, the classification method and the influential analysis were used to predict participation in resistance of unknown resistance genes. The first original proposal of this study is to obtain predictions separately for each condition based on the hypothesis that gene expression of inoculated tomatoes should reflect better the behavior of resistance genes involved in immunity processes against the pathogen than healthy tomatoes, because less genes implicated in immunity will be active. Moreover, conducting the analysis separately led to the consideration that gene expression of resistance genes in non-inoculated tomatoes reflects a basal expression, and not necessarily a behavior due to the presence of the pathogen. Therefore, results obtained separating the databases are more reliable and reflect better the biological condition.

Classification results and resistance prediction using regression and influence analysis were different for each of the tested methods but all methods identified more known resistance genes in inoculated plants than in non-inoculated plants. The cluster number was chosen to be 55 for all clustering methods based on a visual inspection of a dendrogram. Differences between the two distances (Euclidean and Manhattan) were small, therefore either one can be used. All methods coincide in grouping together the same RGC cluster 5/174 known resistance genes for non-inoculated tomato measures and 15/174 for inoculated ones. Most of the methods (K-means, AGNES and Kohonen) grouped resistance genes into two main clusters (Table 1). For the complete results, see Supplementary data. The DIANA method grouped most known resistance genes in one cluster, but this method identified a large cluster with almost all genes (over 11,000), therefore it was discarded. For the SVM classification no clusters were constructed, but resistance and non-resistance were considered as response variables for training a non-linear classifier. This classification method turned out to be more accurate, because classification of non-resistance genes into the RGC was very small (*i.e.* the RGC is homogeneous). The influence analysis of the GEE fitted models was not satisfactory despite the Cook's distance method that identified almost all known resistance genes (170/174) as influential in the model fitted on inoculated plants; only two of them turned out to be influential in non-inoculated tomatoes. The fact that SVM provides the most accurate classification could be due to the presence of non-linear patterns in gene expression that cannot be detected by other traditional methods. Nevertheless, some genes predicted to be resistance genes only by SVM, could be due to noise and high variability of microarray data. Therefore, we advise using the predictions of all methods combined (discarding DIANA and all influence analysis methods except the Cook's distance method). These genes are listed in Table 2.

Table 1
Number of gene products predicted to be in an RGC or predicted to participate on plant resistance processes using clustering, classification and influential analysis. E: Euclidean distance, M: Manhattan distance.

Method	NI: non-resistance	NI: resistance	I: non-resistance	I: resistance
K-means (E)	4818	58	4681	73
K-means (M)	4760	57	4681	73
AGNES (E)	6463	87	6863	99
AGNES (M)	6598	86	6756	99
DIANA (E)	11,128	145	12,632	158
DIANA (M)	4760	57	12,325	156
Kohonen	5120	105	4323	75
SVM	98	69	0	105
Pearson	36	0	0	0
Leverage	1	165	0	170
Cook	0	2	1	170
DfBeta	0	0	4	0
DfBeta for period	17	0	0	10
DfBeta for resistance	1	9	13	7
DfBeta for period * resistance	1	8	0	0

Table 2
List of common genes predicted as resistance genes by clustering methods, SVM classification and influential analysis on GEE regression models.

Gene ID	Function
1-1-2.1.13.17	Avr9/Cf-9 rapidly elicited protein 146 *Nicotiana tabacum*
1-1-2.4.18.3	AvrPto-dependent Pto-interacting protein 3 *Lycopersicon esculentum*
1-1-3.3.11.4	WIZZ *Nicotiana tabacum*
1-1-3.4.10.21	Protein kinase-coding resistance protein *Nicotiana repanda*
1-1-4.1.17.2	Putative disease resistance protein RGA4, identical *Solanum bulbocastanum*
1-1-4.3.20.3	Disease resistance protein RGA2, putative *Ricinus communis*
1-1-7.4.19.21	WRKY transcription factor 26 *Populus tomentosa* × *Populus bolleana* × *P. tomentosa*
1-1-8.2.15.5	Disease resistance protein RPS5, putative *Ricinus communis*
1-1-8.2.16.9	Avr9/Cf-9 induced kinase 1 *Nicotiana tabacum*
1-1-8.4.11.17	WRKY *Solanum lycopersicum*

The genes listed in Table 2 are all well known to participate in immunity processes, which conducts to the conclusion that predictions using the intersection of all applied methods are reliable and could therefore be used for predicting new potential resistance genes.

3.2. Prediction of potential resistance genes

When comparing the genes predicted by the seven methods, 5508 genes were identified as resistance genes by four or five methods. No genes were identified by all seven methods as happened for the validation. Therefore, a prediction based on the intersection of all predictions is not possible for these data. Genes that could be considered as potential resistance genes are the 105 genes predicted by the SVM method (see Supplementary data).

4. Discussion

The results suggest that conducting a separate analysis of both conditions is crucial because differences in prediction were found. Prediction of new resistance genes should be considered as those genes grouped together in the inoculated condition and not in the healthy tomatoes. This can be extended to other biological conditions of interest, in which predictions should be analyzed on the altered condition and not on the reference condition. Therefore, it is advised to analyze conditions separately not only for prediction, but also for other analysis, because results can be strikingly different.

Taking into account the results, these methodologies can be applied for the functional prediction of resistance genes through the selection of clusters with the highest frequency of known resistance genes, when no classification is used. GEE regression and influence analysis seem to be very sensitive to the method and are therefore not recommended. The most accurate and reliable method is SVM classification. Nevertheless, the disadvantage is that a part of the data needs to be used for training.

The methodology presented consists of applying several simple methods that are available in R or SAS for prediction on a microarray data set and selecting the most reliable predictions based on the accordance between methods. It is a general approach that can be applied to different organisms for which gene prediction of a certain function needs to be carried out, and for which microarray data sets of two different conditions are available.

Acknowledgments

This collaborative work was partially supported by the CONICYT, Departamento de Relaciones Internacionales, and “Programa de Cooperación Científica Internacional” Folio PCCIC12012.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ejbt.2014.01.003.

References

[1] Dodds PN, Rathjen JP. Plant immunity: Towards an integrated view of plant–pathogen interactions. Nat Rev Genet 2010;11:539–48. http://dx.doi.org/10.1038/nrg2812.

[2] Atias O, Chor B, Chamovitz DA. Large-scale analysis of *Arabidopsis* transcription reveals a basal co-regulation network. BMC Syst Biol 2010;3:86. http://dx.doi.org/10.1186/1752-0509-3-86.

[3] Mutwil M, Usadel B, Schütte M, Loraine A, Ebenhöh O, Persson S. Assembly of an interactive correlation network for the *Arabidopsis* genome using a novel heuristic clustering algorithm. Plant Physiol 2010;152:29–43. http://dx.doi.org/10.1104/pp.109.145318.

[4] Pop A, Huttenhower C, Iyer-Pascuzzi A, Benfey PN, Troyanskaya OG. Integrated functional networks of process, tissue, and developmental stage specific interactions in *Arabidopsis thaliana*. BMC Syst Biol 2010;4:180. http://dx.doi.org/10.1186/1752-0509-4-180.

[5] Pritchard L, Birch P. A systems biology perspective on plant-microbe interactions: Biochemical and structural targets of pathogen effectors. Plant Sci 2011;180:584–603. http://dx.doi.org/10.1016/j.plantsci.2010.12.008.

[6] Lee H, Chah OK, Sheen J. Stem-cell-triggered immunity through CLV3p–FLS2 signalling. Nature 2011;473:376–9. http://dx.doi.org/10.1038/nature09958.

[7] Mukhtar MS, Carvunis AR, Dreze M, Epple P, Steinbrenner J, Moore J, et al. Independently volved virulence effectors converge onto hubs in a plant immune system network. Science 2011;333:596–601. http://dx.doi.org/10.1126/science.1203659.

[8] López-Kleine L, Smart CD, Fry WE, Restrepo S. Identification of key molecular components of the resistance of cherry tomato against *Phytophthora infestans*. Acta Biol Colomb 2012;17:537–50.

[9] López-Kleine L, Torres-Avilés F, Tejedor FH, Gordillo LA. Virulence factor prediction in *Streptococcus pyogenes* using classification and clustering based on microarray data. Appl Microbiol Biotechnol 2012;93:2091–8. http://dx.doi.org/10.1007/s00253-012-3917-3.

[10] Romeo JS, Torres-Aviles F, López-Kleine L. Detection of influent virulence and resistance genes in microarray data through quasi likelihood modeling. Mol Genet Genomics 2013;288:49–61. http://dx.doi.org/10.1007/s00438-012-0730-8.

[11] Fei Z, Tang X, Alba R, Giovannoni J. Tomato Expression Database (TED): a suite of data presentation and analysis tools. Nucleic Acids Res 2006;34:766–70.

[12] Cai G, Restrepo S, Myers K, Zuluaga P, Danies G, Smart C, et al. Gene profiling in partially resistant and susceptible near-isogenic tomatoes in response to late blight in the field. Mol Plant Pathol 2013;14:171–84. http://dx.doi.org/10.1111/j.1364-3703.2012.00841.

[13] Huber W, Von Heydebreck A, Sueltmann H, Poustka A, Vingron M. Parameter estimation for the calibration and variance stabilization of microarray data. Stat Appl Genet Mol Biol 2003;2:1544–6115. http://dx.doi.org/10.2202/1544-6115.1008.

[14] R Development Core Team R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing. URL http://www.R-project.org/; 2012.

[15] Hartigan JA, Wong MA, Algorithm AS. 136: a K-means clustering algorithm. Appl Stat 1979;28:100–8. http://dx.doi.org/10.2307/2346830.

[16] Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. New York: Wiley; 1990.

[17] Krause EF. Taxicab geometry: An adventure in non-Euclidean geometry. New York: Dover; 1986.

[18] Kohonen T. Self-organizing maps. 3rd ed. Springer-Verlag; 2000.

[19] Schölkopf B, Smola A. Learning with kernels: Support vector machines, regularization, optimization, and beyond. Cambridge: The MIT Press; 2002.

[20] Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13–22. http://dx.doi.org/10.2307/2336267.

[21] Nelder JA, Pregibon D. An extended quasi-likelihood function. Biometrika 1987;74:221–32. http://dx.doi.org/10.1093/biomet/74.2.221.

[22] Preisser JS, Qaqish BF. Deletion diagnostics for generalized estimating equations. Biometrika 1996;83:551–62. http://dx.doi.org/10.1093/biomet/83.3.551.

[23] Hammill BG, Preisser JS. A SAS/IML software program for GEE and regression diagnostics. Comput Stat Data Analyst 2006;51:1197–212. http://dx.doi.org/10.1016/j.csda.2005.11.016.

[24] SAS Institute Inc. What's new in SAS 9.2. Cary. SAS Institute Inc.; 2008.