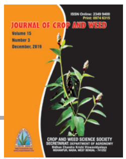
Journal of Crop and Weed, 16(3): 01-08 (2020)
ISSN- O : 2349 9400 ; P : 0974 6315



http://cwssbckv.org
www.cropandweed.com

An ensemble approach for gene regulatory network study in rice blast

C. SARKAR, [1] R. PARSAD, [2] D. C. MISHRA AND [3]A. RAI

Division of Bioinformatics, ICAR- Indian Agricultural Research Institute, New Delhi
[1]Division of Agricultural Statistics, [2,3]Division of Bioinformatics
ICAR- Indian Agricultural Statistics Research Institute, New Delhi

Received : 11.01.2020 ; Revised : 08.12.2020 ; Accepted : 17.12.2020

DOI : https://doi.org/10.22271/09746315.2020.v16.i3.1358

ABSTRACT

Gene Regulatory Network (GRN) is important to understand complex biological process like disease. An ensemble approach for network construction leads to more accurate and consistent GRN than individual methods. In this study we have combined the networks obtained from correlation, partial least square, principal component regression and ridge regression techniques using Fisher's weighted method. The network was ensemble considering the significant edges selected from individual methods. The method was applied to gene expression data of rice blast disease to construct GRN. The stability of the obtained ensemble network was higher than the networks obtained from individual methods.

Keywords: Ensemble approach, Fisher's weighted method, GRN and rice blast disease

Study of gene regulatory network (GRN) is very important for understanding the biological processes in cellular system. For complex biological processes like disease it is very crucial to understand the disease mechanism which can be feasible by GRN study. The basic assumption of GRN studies is that genes function through complex networks. Gene regulatory network represented as graph where nodes represent the genes and the edges represent the pair-wise interaction of genes. The existence of nodes (*i.e.* genes) in network depends on the significance of Edges *i.e.* interaction between genes. It is very important to determine the statistically significant Edges in the network. So the main challenge is to measure the pair-wise connectivity of genes to construct the GRN. Though constructing of gene regulatory network is very challenging task but availability of computational resources makes it possible to analyze using high dimensional gene expression data. The expression profiles of the genes functioning in biological pathways are available in databases. The gene expression profiling involves a large set of gene (p) with few samples (n) for each gene because the generation of gene expression data is very expensive. So the statistical methods used for constructing GRN based on analysis of gene expression dataset using statistical methods suffers from np hard problem (p>n). The gene expression data also suffers from multi-co-linearity problem because of the functional dependencies of a gene in cellular system. Several methods already have been proposed in literature for inferring the construction of GRN like principal component regression (PCR) (Gill *et al*., 2010), partial least square (PLS) regression (Pihur *et al*., 2008), ridge regression (Gill *et al*., 2010), etc. which overcome multi-co-linearity and np hard problem. The pair-wise connectivity was measured based on the connectivity score of each pair of genes in these four methods. But these four methods vary in the resulting significant Edges based on the connectivity scores which make it difficult to determine the appropriate method for inferring GRN. Due to the occurrence of large amounts of false positives and false negatives none of these methods can infer biologically consistent network (Gill *et al*., 2010). This problem can be resolved by using integrating the result of these methods based on metadata analysis. In this present study we have proposed an Ensemble method which combines the outcomes obtained from different methods.

In this study, we have proposed an ensemble method based on meta-analysis approach named Fisher's weighted method (Hedges and Olkin, 1985) to combine the results of different methods based on using the probability values of each Edge. In the proposed method, we have combined the results obtained from Correlation, PCR, PLS and Ridge Regression.

These four scoring methods infer the GRN which vary in the significant Edges but the significant Edges obtained from these methods can be complement of each other because of the difference in measuring connectivity score. Correlation measures pair-wise connectivity between genes which captures the linearity of expression data for computing the connectivity scores. On the other hand PCR, PLS and Ridge regression based scoring methods measures the non-linearity of expression data. Ridge Regression gives a biased estimator to reduce the multi-co-linearity. So in the proposed Ensemble approach the complementary

Email: cschiranjib9@gmail.com

combination of linearity and non-linearity of expression data explains the variability in different directions.

MATERIALS AND METHODS

For this present study we have used gene expression dataset of Blast fungus (*Magneporthe oryzae*) infected leaf of Rice. The gene expression data of rice blast disease have been downloaded from National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO) database (*https://www.ncbi.nlm.nih.gov/geo/*). The microarray data of four different experiments are used for gene regulatory network constructions which are normalized data by Guanine Cytocine Robust Multi-array Average (GCRMA). The dataset contains 47 experimental samples on 57381 probe sets generated using Affymetrix Rice Genome Array. The datasets are of microarray experiments of blast fungus infected rice leaves. The data details are given in table 2.1.

The chromosomal loci information of the probes are downloaded from Rice 57K Rice Data (*http://www.ricechip.org/*) and the QTL information for Blast disease collected from QTL Genome Viewer (*http://qtaro.rd.naro.go.jp/cgi-bin/gbrowse*).

Table 1: Data details downloaded from NCBI GEO database

Accession No.	Platform	Number of samples	Experiments
GDS3441	GPL2025	8	Rice leaves response to blast fungus infection: time course
GSE61952	GPL2025	6	Expression data from rice lesion mimic mutant spotted leaf 5 (spl5)
GSE41798	GPL2025	18	Transcriptome study of rice early response to rice blast fungus
GSE30941	GPL2025	15	Rice gene global expression analysis upon inoculation with different Magnaporthe isolates

Differentially expressed gene

The gene expression dataset of blast fungus infected leaves of rice contains 57381 probe sets has been used. In order to identify the differentially expressed probes between control and *Magnoporthe oryzae* inoculation which also reduce the dimensionality of the dataset based on the differential expression of probes during *Magneporthe oryzae* inoculation, it was analyzed using one way ANOVA (Analysis of Variance) technique. The analysis was performed using the dataset with accession number GDS3441 containing 8 samples among which 4 samples with blast fungus inoculation and 4 samples without inoculation (control), each of these samples were taken at 2 time points. One way ANOVA was performed using SAS software for each of the genes separately by taking the expression levels in control and test samples. The Least significant difference (LSD) and Fold change (FC) *i.e.* difference between means of control and inoculated samples for each probe was computed as -

$$LSD = t_{\alpha,df} \times \sqrt{\frac{2\times mse}{4}} \qquad (1)$$

where,

$t_{\alpha,df}$ is table value of t distribution with α level of significance with given degrees of freedom, here we have considered α at 0.1% with 6 error degrees of freedom.

$$\mathrm{FC} = \bar{x}_{control} - \bar{x}_{inoculated}$$

where, $\bar{x}_{control}$ is the mean of the expression values under control condition (not inoculated) and $\bar{x}_{inoculated}$ inoculated is the mean of the expression values under inoculated condition.

$$mse = \frac{\sum_{j=1}^{2}\sum_{i=1}^{4}\left(x_{ij} - \bar{x}_j\right)^2}{a},$$

a = degrees of freedom *i.e.* 6 in this study,

x_{ij} is the gene expression values of i^{th} of j^{th} group (here dataset contains two group control and inoculated) and $\bar{x}_j$ is the mean of the gene expression values of j^{th} group. The differentially expressed probes were identified at 0.1% level of significance with 6 error degrees of freedom. The significant differentially expressed probe sets were used for further analysis in this study. The significant differentially expressed genes were 536 out of 57381 genes.

Ensemble approach for GRN construction

In this study we propose Fisher's weighted method of meta-analysis for as an Ensemble method for GRN construction. We have combined the probability values (*p*-values) of each Edge obtained from four independent methods *i.e.* Correlation, PCR, PLS and Ridge

Regression. In Fisher's method the *p*-values obtained from *k* independent study are combined as follows -

$$F_w = -2\sum_{i=1}^{k} w_i \ln(P_i) \tag{2}$$

where, the F_w statistic follows chi-square distribution with 2*k* degrees of freedom and P_i follows the uniform distribution. In this study the Edge values *i.e.* the connectivity score for each pair of genes were computed based on four methods Correlation, PCR, PLS and Ridge Regression. We have followed the procedure to compute pair-wise connectivity score of genes as derived in Gill *et al*. (2010) and Dutta (2001). The pair-wise connectivity score (S_{ik}) computed as follows -

i) Correlation coefficient (Gill *et al.*, 2010):

The connectivity score S_{ik} based on correlation coefficient computed as -

$$S_{ik} = \frac{x_i^T x_k}{\sqrt{(x_i^T x_i)(x_k^T x_k)}} \tag{3}$$

where, x_j and x_k are the standardized expression values of ith and kth gene respectively and S_{ik} is the connectivity score between i^{th} and k^{th} gene.

ii) Principal Component Regression (PCR) (Pihur *et al.*, 2008) :

The PCR scoring based connectivity score computed as -

$$[s_{g1},\ldots,s_{g,g-1},s_{g,g+1},\ldots s_{gp}]^T = V\hat{\beta}_g \tag{4}$$

where S_{gp} is the connectivity score between g^{th} and p^{th} gene and V is the matrix of Eigen vectors computed from gene expression values.

iii) Partial least square regression (PLS) (Datta, 2001) :

$$\hat{s}_{ik} = \frac{\sum_{l=1}^{v}\hat{\beta}_{il}c_{ik}^{(l)} + \sum_{l=1}^{v}\hat{\beta}_{kl}c_{ik}^{(l)}}{2} \tag{5}$$

where, $\hat{\beta}_{il} = (t_i^{(l)^T} t_i^{(l)})^{-1} t_i^{(l)^T} x_i$

latent variable $t_i^{(l)} = \sum_{k\neq i}^{p} c_{ik}^{(l)} X_k^{(l)}$

$$c_{ik}^{(l)} = \frac{X^{(l)^T} x_i}{\sqrt{x_i^T X^{(l)} X^{(l)^T} x_i}}$$

PLS measures S_{ik} is obtanied by regressing the expression values of a given gene i on the latent variable $\left(t_i^{(l)}\right)$ derived from expression values of (*i*-1) genes.

iv) Ridge regression (Gill *et al.*, 2010) :

$$\left[s_{g,1},\ldots,s_{g,g-1},s_{g,g+1},\ldots,s_{g,p}\right]^T = (\tilde{X}_g^T\tilde{X}_g + \lambda I)^{-1}\tilde{X}_g x_g \tag{6}$$

where S_{gp} is the connectivity score between gth and pth gene

For each gene pair S_{ik} computed based on a penalized $\hat{\beta}_{ridge}$ estimate where λ is the penalty parameter. 'dna' R package has been used to compute the connectivity score of all pairs of genes.

To compute the mean and Standard Error (SE) of the connectivity score S_{ik} for each Edge we have drawn 250 Bootstrap samples from the gene expression values of the dataset. Each Bootstrap sample consists of same genes with different expression values. For each Bootstrap sample we computed the pair-wise connectivity score using the four methods *i.e.* Correlation, PCR, PLS and Ridge Regression. Based on the 250 Bootstrap samples we computed 250 connectivity score $\left(S_{ik_1}, S_{ik_2}, \ldots, S_{ik_{250}}\right)$ for each pair of genes and for each method. For each pair of gene the mean and Standard Error (SE) computed as follows -

$$\bar{s}_{ik} = \frac{\sum_{i\neq k}^{n}\sum_{j=1}^{B} s_{ik_j}}{B} \tag{7}$$

$$Se = \frac{1}{\sqrt{B-1}}\sqrt{\sum_{i\neq k}^{n}\sum_{j=1}^{B}(s_{ik_j} - \bar{s}_{ik})^2} \tag{8}$$

where, B is the number of Bootstrap samples *i.e.* 250.

The R code for drawing Bootstrap samples and computing the pair-wise connectivity score for 250 samples are provided in the Supplementary Information. We have computed t-test statistic for each Edge based on the mean and Standard Error of connectivity score obtained from PCR, PLS and Ridge Regression. Under the null hypothesis -

$H_0 : s_{ik} = 0$ *i.e.* there is no interaction between the gene pairs,

$H_1 : s_{ik} \neq 0$, there is interaction between the gene pairs.

The computed t-test statistic is as follows:

$$t = \frac{\bar{s}_{ik}}{Se} \tag{9}$$

For Correlation Based scoring method the t-test statistic computed as follows -

$$t = \frac{\overline{s}_{ik}\sqrt{n-2}}{\sqrt{1-\overline{s}_{ik}^{\,2}}} \tag{10}$$

where, n is the number of samples for each gene *i.e.* 47 and $\overline{s}_{ik}$ is the mean computed from 250 connectivity score from Bootstrap samples.

For each connectivity scoring method we fit the Empirical mixture distribution based on the t-test statistic. As proposed by Efron (2004) the mixture distribution $f(t)$ empirically estimated from *t* statistic as follows -

$$f(t) = \eta_0 f_0(t) + (1-\eta_0) f_1(t) \tag{11}$$

where $f_0(t)$ and $f_1(t)$ are the null (H_0) and alternate (H_1) density respectively. η_0 is the proportion of null values. and the false discovery rate is computed as

$$fdr(t) = \frac{f_0(t)}{f(t)} \tag{12}$$

where $f_0(t)$ and $f_1(t)$ are the null (H_0) and alternate (H_1) density respectively.

The empirical *p*-values were computed from cumulative distribution function (cdf) of $f(t)$ as follows -

$$p = 2\times(1-\int_{-\infty}^{t} f(t)) \tag{13}$$

R package 'fdrtool' (Klaus *et al.*, 2015) has been used to fit the Empirical mixture distribution and to compute the fdr values from the mixture distribution.

The significant Edges were selected for each case of four scoring methods based on the fdr values. We have considered the significant edges from each method at fdr value 0.1. The selected significant Edges from four methods were then integrated based on the *p*-values computed from the Empirical mixture distribution. For each edge, the *p*-values follow the Uniform distribution under null hypothesis $H_0 : s_{ik} = 0$. The connectivity score were computed based on four independent methods (Eq. 3-6) which came up with four *p*-values for each significant Edge. The Edges which showed significance in four cases at 0.1 fdr were further used for combing *p*-values. The significant *p*-values obtained from four methods were combined by calculating statistic using Fisher's weighted method $\left(F_w = -2\ln\ln(p_i) \sim \chi_2^2\right)$. The F_w statistics computed for each Edge is as follows -

$$F_w = -2\ln\ln(p_1 \times p_2 \times p_3 \times p_4) \tag{14}$$

In our present study the basis of Ensemble approach depends on computing the F_w statistics as given in (14). The F_w statistics follows chi-square distribution with 8 degrees of freedom. The Edges for the Ensemble network was selected based on the table value of chi-square at 0.1% level of significance with 8 degrees of freedom. As the weighted form for Fisher's method consider (Eq. 2) a weight for each *p*-values obtained from independent methods, we have multiplied a weight value for each *p*-values for each Edge. For each Edge the weight was computed based on the corresponding Edge scoring method. We have computed the weights for each Edge obtained from each method as follows -

$$w_{ie} = \begin{cases} m/N \\ \mathit{Accuracy}\ \text{of scoring method} \\ 1 \end{cases}$$

where, w_{ie} is the weight of e^{th} Edge obtained from i^{th} scoring method,

m is the no. of significant Edges

and N is the total no. of Edges.

The Accuracy of each method computed as (TP+TN)/(TP+TN+FP+FN),

TP = True positive; FP = False positive; TN = True negative; FN = False negative.

For three cases of weights we have computed the F_w statistics for each Edge and the Ensemble GRN was constructed in three cases considering table value of chi-square at 0.1% level of significance with 8 degrees of freedom as the cut-off value of F_w statistics.

Validation of ensemble network

In silico evaluation of the ensemble network of the genes has been done using three procedures i.e. Hamiltonian distance based criteria (Rajapakse and Mundra, 2011), Hub genes in the network and QTL analysis.

Stability of network

The stability of Ensemble network evaluated using the Hamiltonian distance based criteria proposed by Rajapakse and Mundra (2011). B bootstrap samples denoted as $\{X^b\}_{b=1}^{B}$ drawn from original dataset. For each bootstrap samples an Ensemble network (S^b) generated and represented as the connectivity matrix $c^b = \{c_{ij}^b\}_{I\times I}$

where $c_{ij}^b = \begin{cases} 1, \ \mathit{for\ significant\ edges} \\ 0, \ \mathit{otherwise} \end{cases}$ For any two network s^b and $s^{b'}$ obtained from b-th and b-th sub-samples, the Hamming distance of pair-wise network obtained as follows -

$$\rho(s^b, s^{b'}) = 1 - \frac{1}{|s^b| + |s^{b'}|} \sum_i \sum_j d(c^b_{i,j}, c^{b'}_{i,j}) \qquad (15)$$

where d denotes the hamming distance and $|s^b|$ denotes the number of significant edges in the network s^b and the stability $\rho \in [0,1]$.

The overall stability performance of network was measured by averaging over all B samples -

$$\rho_{structure} = \frac{2}{B(B-1)} \sum_{b=1}^{B} \sum_{b'=b+1}^{B} \rho(s^b, s^{b'}) \qquad (16)$$

Each edge stability measure based on total number of times the connectivity occurs for each pair of genes over B samples. Edge stability was computed as -

$$\rho_{edge}(i,j) = \frac{\sum_{b=1}^{B} C^{s^b}_{i,j}}{B} \qquad (17)$$

where, $C^{s^b}_{i,j}$ is the connectivity score between i^{th} and j^{th} genes in s^b network and $\rho_{edge}(i,j)$ is the Edge stability score for edge between i^{th} and j^{th} genes.

In order to compute the stability we have drawn bootstrap samples of size 50, 100 and 150 from the dataset containing 536 probes with 47 samples of expression values. For each sample S_{ik} were computed based on scoring method *i.e.* Correlation, PCR, PLS, Ridge Regression and the *p*-values computed as described in Methodology (Eq. 7-10). The F_w statistic for each Edge computed by combing the *p*-values using Fisher's method. To form the connectivity matrix for each sample we have considered the significant Edges as 1 and 0 otherwise. The pair-wise distance of connectivity matrix was computed as Eq. 15 for each case of sub-samples drawn *i.e.* 50, 100 and 150 and the overall stability performance of network computed based on the sub-samples drawn from the original dataset. The Edge stability score was computed by summing over the Edge value *i.e.* 1 or 0 for the connectivity matrices obtained from 50, 100 and 150 sub-samples. The Precision, Recall, Accuracy and F-measure were computed based on the Edge stability score obtained from Hamiltonian distance based criteria and the significant Edges obtained from Ensemble approach.

Recall = TP/(TP+FN),

Precision = TP/(TP+FP),

Accuracy = (TP+TN)/(TP+FP+FN+TN),

F= 2*(Recall * Precision) / (Recall + Precision).

where, TP = True Positives, FP = False Positives, TN = True Negatives, and FN= False Negatives.

Hub gene identification

We have identified the hub genes in the Ensemble Network based on the weighted gene scoring method (Das *et al.*, 2017) to validate the GRN obtained from our proposed method. In weighted gene scoring method hub genes are identified depending on the connection degree with other nodes in the network. The genes in the Ensemble network connected by the significant Edges were considered for the Hub gene Identification analysis. We have computed the degree of connectivity for each gene in the Ensemble network and the genes with connectivity score more than average score of the Ensemble network was considered as the hub of the network. These hub genes based on the connectivity in the Ensemble network validated with the hub genes obtained from the weighted gene score (WGS) method (Das *et al.*, 2017). The hub gene based on the WGS method was identified using the R package dhga based on the gene expression values of the genes in the Ensemble network.

QTL analysis of the significant genes

Quantitative Trait Loci (QTL) hits in the selected gene set were performed for validation of the significant genes obtained in the Ensemble network. The information of QTLs responsible for rice blast disease resistance was collected from QTL Genome Viewer (http://qtaro.rd.naro.go.jp/ cgi-bin/ gbrowse). The chromosomal location of the significant genes of the Ensemble network was obtained from NCBI (*https:// www.ncbi.nlm.nih.gov/*). The genes showing significantly overlapped with the blast resistance QTLs were identified using GSAQ R package (Das S., 2016).

RESULTS AND DISCUSSION

We have analyzed the gene expression dataset with accession number GDS3441 (Table 1) of rice leaf containing 57381 genes using one-way ANOVA to identify the differentially expressed genes as discussed. We obtained 536 genes which showed significantly differential expression during blast fungus inoculation at 0.1% level of significance with 6 error degrees of freedom. To construct the GRN based on our proposed method we have used the 47 expression values of 536 DEGs from 4 experimental dataset listed in table 1. We have computed the connectivity score for all pairs of genes *i.e.* 143380 ($^{536}C_2$) pairs of genes. Out of 143380 edges the Ensemble network contains 74, 70 and 74 significant edges and 40, 36 and 40 nodes (at 1% level of significance with 8 degrees of freedom of chi-square distribution) with weight m/n, accuracy of the scoring method and 1 for combining the *p*-values (Table 4). We have used three validation criteria to validate the performance of the Ensemble network and the network obtained from Correlation, PLS, PCR and Ridge regression scoring methods. Out of these nodes in the Ensemble Network 16, 15 and 16 genes showed significant overlapping with 7, 6 and 7 blast resistant QTLs. The stability score of the Ensemble network was more than the GRN obtained based on the Correlation, PLS, PCR and Ridge regression scoring methods (Table 3 and Table 5).

The performance measures of the Ensemble network were also higher in three weighted criteria than the individual scoring methods. The hub genes in the Ensemble network based on the connectivity with other genes was 19 in three weighted criteria. The hub genes in the Ensemble network were validated with the WGS scoring method which resulted in 16 genes. The WGS based 16 hub genes were same with the 16 out of 19 hub genes in the Ensemble network. The connectivity score, t-test statistic, MSE and *p*-values of 143380 edges obtained from Correlation, PLS, PCR and Ridge regression scoring methods and the F-statistic value obtained from Fisher's weighted method are given in the Supplementary information. The network obtained from above mentioned methods are visualized using the Cytoscape software (Fig. 1). The numbers of genes significantly overlapped with the QTLs are shown in table 2 and the graphs (Fig. 2) of QTL wise QTLhit genes are generated using GSAQ R package.

The proposed Ensemble approach in this study has higher performance in three weighting criteria *i.e.* w=m/ n , w= accuracy and w=1. As we have only considered the significant Edges (fdr<0.1) obtained from Correlation, PCR, PLS and Ridge regression the significant Edges in Ensemble network reduced to 74 and 70 than the individual methods. Our proposed method showed biologically consistent result in QTL analysis which is more efficient than the other methods. In QTL analysis 16 genes showed significant

Table 2: The results obtained from correlation, PLS, PCR and Ridge regression scoring methods

Method	Significant Nodes (Genes)	No. of QTLs	Q/N ratio
Correlation	461	8	0.0173
PLS	234	8	0.0342
PCR	354	6	0.0169
Ridge Regression	436	7	0.0161

Table 3: Performance of PLS, PCR, Ridge regression, correlation and Fisher's method

Methods	No. of Samples	Stability	Recall*	Precision*	Accuracy*	F-measure*
Pearson Correlation	50	0.597	1	0.704	0.704	0.826
	100	0.598	1	0.777	0.777	0.874
	150	0.602	1	0.791	0.791	0.883
PCR	50	0.786	1	0.255	0.255	0.407
	100	0.750	1	0.312	0.312	0.476
	150	0.744	1	0.318	0.318	0.483
PLS	50	0.602	1	0.471	0.471	0.641
	100	0.593	1	0.489	0.489	0.656
	150	0.471	1	0.504	0.504	0.670
Ridge regression	50	0.370	1	0.374	0.374	0.545
	100	0.365	1	0.452	0.452	0.622
	150	0.366	1	0.475	0.475	0.644

* Recall = TP/(TP+FN), Precision = TP/(TP+FP), Accuracy = (TP+TN)/(TP+FP+FN+TN), F= 2*(Recall * Precision) / (Recall + Precision)

overlapping with 7 blast resistant QTLs *i.e. Pi-4t, Pi-35t, qBFR 11, qDLA 123, qNBL.5, rbr 2* and unnamed1. So in this present study, in silico validation of the ensemble network and the QTL information based validation suggest that our proposed methodology better results than the individual one. The proposed method resulted less genes which may make easy to implement in the molecular breeding program.

Table 4: The results obtained from Fisher's weighted method

Weights	Numbers of Nodes	No. of QTLs	Q/N ratio
w_i= m/n	40	6	0.150
w_i= accuracy	36	5	0.138
w_i= 1	40	6	0.150

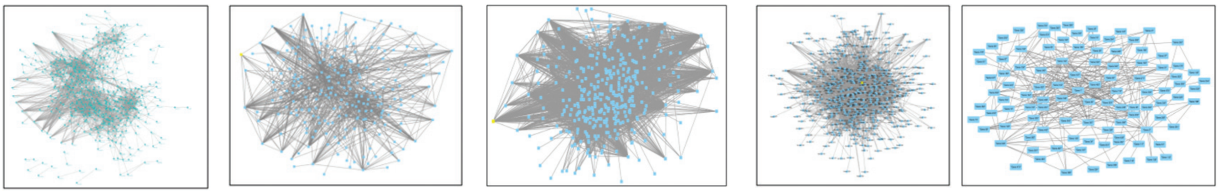
* m= number of significant nodes, n= number of total nodes in the GRN

Table 5: Performance of ensemble approach (Fisher's weighted method)

Weights	No. of Samples	Stability	Recall	Precision	Accuracy	F-measure
w_i= m/n	50	0.891	0.908	0.922	0.889	0.915
	100	0.887	0.903	0.890	0.890	0.896
	150	0.888	0.917	0.892	0.902	0.904
w_i= accuracy	50	0.885	0.900	0.918	0.859	0.909
	100	0.881	0.930	0.881	0.926	0.905
	150	0.881	0.910	0.871	0.853	0.890
w_i= 1	50	0.891	0.907	0.921	0.889	0.915
	100	0.887	0.902	0.890	0.890	0.896
	150	0.888	0.916	0.891	0.902	0.904

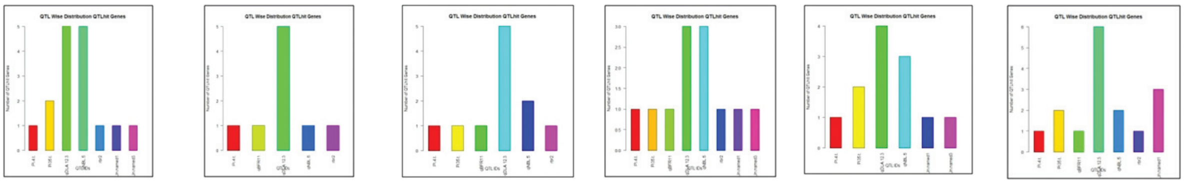
Table 6: The probe IDs of hub genes in the ensemble network

Probe IDs of hub genes	WGS score	Degree of connectivity
Os.10023.1.S1_at	16.14	23
Os.18076.1.S1_at	15.78	14
Os.26989.1.S1_a_at	14.28	9
Os.17334.1.S1_s_at	10.77	8
Os.26784.1.S1_a_at	8.91	7
Os.53099.1.S1_at	12.12	7
Os.16914.1.S2_at	14.04	6
Os.52598.1.S1_at	13.81	6
Os.8441.1.S1_at	12.84	5
Os.49512.1.S1_s_at	12.75	5
Os.33614.1.S1_at	14.06	5
Os.18674.1.S1_at	12.97	5
Os.1005.1.S1_at	3.99	4
Os.17864.1.S1_at	12.09	3
Os.11301.1.S1_x_at	9	3
Os.23574.1.S1_at	11.15	3
Os.12688.1.S1_at	9.27	3
Os.39020.1.S1_at	12.16	3
Os.41883.1.S1_at	4.39	3



(a) (b) (c) (d) (e)

Fig. 1: Visualization of the GRN using Cytoscape Software based on different scoring methods (a) Correlation, (b) Partial Least Square, (c) Principal Component Regression, (d) Ridge Regression and (e) Fisher's weighted method



(a) (b) (c) (d) (e) (f)

Fig. 2: QTL wise distribution of QTLhit genes based on the significant genes obtained from (a) Differential gene expression analysis, (b) Ridge Regression, (c) PLS Regression, (d) PCR, (e) Correlation, (f) Fisher's weighted method

REFERENCES

Das, S. 2016. Package 'GSAQ' CRAN. *https://cran.r-project.org/web/packages/GSAQ/index .html.*

Das, S., Meher P.K., Rai, A., Bhar, L.M. and Mandal, B. N. 2017. Statistical Approaches for Gene Selection, Hub Gene Identification and Module Interaction in Gene Co-Expression Network Analysis: An Application to Aluminum Stress in Soybean (*Glycine max* L.). *PLoS ONE*, **12**(1): 1-24.

Dudoit, S., Shaffer, J.P. and Boldrick, J.C. 2003. Multiple hypothesis testing in microarray experiments. *Statistical Science*, **18(1)**, 71-103.

Efron, B. 2004. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, **99**(465): 96-104.

Gill, R., Datta, S. and Datta, S. 2010. A statistical framework for differential network analysis from microarray data. *BMC Bioinformatics*, **11**: 95.

Hedges, L.V., and Olkin, I. 1985. *Statistical Methods for Meta-Analysis* (Academic press, San Diego).

Klaus, B., Strimmer, K. and Strimmer, M.K. 2015. Package 'fdrtool'. CRAN. *http://http1. debian*. or. *jp/pub/CRAN/web/packages/fdrtool/fdrtool.pdf.*

Pihur, V., Datta, S. and Datta, S. 2008. Reconstruction of genetic association networks from microarray data: a partial least squares approach. *Bioinformatics*, **24**(4): 561–568.

Rajapakse, J.C. and Mundra, P.A. 2011. Stability of building gene regulatory networks with sparse autoregressive models. *BMC bioinformatics*. **12**(13): S17.