A computational system biology approach to construct gene regulatory networks for salinity response in rice (*Oryza sativa*)

SAMARENDRA DAS1, PRIYANKA PANDEY2, ANIL RAI3 and CHINMAYEE MOHAPATRA4

Indian Agricultural Statistics Research Institute, Pusa, New Delhi 110 012

Received: 11 February 2015; Accepted: 26 June 2015

ABSTRACT

Salinity is one of the most common abiotic stress which limits agricultural crop production. Salinity stress tolerance in rice (*Oryza sativa* L.) is an important trait controlled by various genes. The mechanism of salinity stress response in rice is quite complex. Modelling and construction of genetic regulatory networks is an important tool and can be used for understanding this underlying mechanism. This paper considers the problem of modeling and construction of Gene Regulatory Networks using Multiple Linear Regression and Singular Value Decomposition approach coupled with a number of computational tools. The gene networks constructed by using this approach satisfied the scale free property of biological networks and such networks can be used to extract valuable information on the transcription factors, which are salt responsive. The gene ontology enrichment analysis of selected nodes is performed. The developed model can also be used for predicting the gene responses under stress condition and the result shows that the model fits well for the given gene expression data in rice. In this paper, we have identified ten target genes and a series of potential transcription factors for each target gene in rice which are highly salt responsive.

Key words: Gene regulatory network, Multiple linear regression, Singular value decomposition, Target gene, Transcription factor

Rice (*Oryza sativa* L.) is the main staple food for more than half of the world’s population and accounts for supply of 30-80% of daily calories consumed in Asia (Narisco et al. 2002). The production of rice is severely affected by salinity stress in soil or water due to the nature of rice plants to grow in swamp and fresh water marshes (Akbar and Ponnamperuma 1980). The detrimental effects of high salinity on plants can be observed as the death of plants and/or decrease in its productivity (Joseph and Mohanan 2013). Hence, salinity in soil or water is of much importance to agriculture as it limits rice production in about 30-50% of the rice growing area worldwide (Rahimi and Biglarifard 2011). The genes respond to different stresses at the transcriptional level and the products of these genes play role in the stress response and tolerance (Bray et al. 2000). The mechanism of salinity response in rice is quite complex and till now not known completely. Hence, the gene regulatory networks (GRN) can be used to understand the mechanism of salinity stress in rice.

In the post genomics era, modelling and construction of GRN and gene interactions are important tasks in functional genomics and systems biology. The genome encodes thousands of genes which transcribed in proteins responsible for cell development and various cellular functions in response to diverse extracellular signals. These proteins interact with each other to comprise a structured regulatory network known as GRN (Wang et al. 2013). In system biology, GRN is a graphical representation in which nodes consist of genes and edges connecting between them show regulatory relationship. The regulation of gene expression by transcription factors (TF) is of keen importance as it binds to the promoter regions of their target genes (TG), thereby activating or inhibiting their expression. Deciphering and understanding TF-TG interactions are vital in the field of biology and medicines, ranging from the *in silico* modelling and construction of the GRN to the identification of new potential drug targets (Haury et al. 2012). Thus, mathematical modelling and computational algorithms for inferring the regulatory relations are indispensable. Infact, many mathematical models and algorithms have been proposed for the inference of GRN (de Jong 2002). For the discrete gene expression data (expressed: 1, unexpressed: 0), many models including Boolean Network model, Probabilistic Boolean Network model, multivariate Markov model etc. have been proposed (Akutsu et al. 2000, Boros et al. 1998, Ching et al. 2005, Hirose et al. 2006, Idekar et al. 2000, Noda et al. 1998, ...
These methods of discrete network construction fail, when there is large number of regulators and can’t be applied to continuous gene expression data directly. For the continuous expression data, clustering algorithms, Bayesian, Dynamic Bayesian networks and Ordinary Differential Equations (ODE) based methods have been proposed for the network inference (Bansal et al. 2007). Bayesian method of network construction fails to consider the dynamic nature of the genes and Dynamic Bayesian method is limited to small networks due to high computational complexity. Yeung and Ruzzo (2002) used a linear model and singular value decomposition (SVD) to generate a family of candidate networks that are consistent with a given dataset, thus compensating for this deficiency in time points.

In this study, interest is in the continuous gene expression microarray data of salinity stress in rice. The difficulty of applying the ODE based models in such data sets lies in estimation of the interaction coefficients and the networks constructed based on this method may not satisfy the scale-free property of real biological networks (Barabasi and Albert 1999). These methods are often very complex and slow, do not consider the feature (TFs) selection as well as do not provide insight into the biological meanings of each parameter (Huang et al. 2010). In order to address the limitations of existing approaches for modeling and construction GRN for cross sectional gene expression data, an attempt has been made to develop an approach for construction and modelling of GRN which is based on multiple linear regression (MLR) and SVD. This technique also considers gene selection approach. This approach is applied to gene expression data in rice for construction and modeling of GRNs for salinity stress response.

MATERIALS AND METHODS

For this study, the microarray data available at the NCBI Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/GEO) generated by using Affymetrix platform is used, as it shows high degree of reproducibility and homogeneity across different laboratories when compared with other platforms. Affymetrix microarray data sets for salt responsive genes (Walia et al. 2009) are available in the above database (Cotsafitis et al. 2011, Pandit et al. 2010 and Jain et al. 2008) with accession number GSE14403, GSE16108, GSE6901, GSE13735. These data were generated from Affymetrix Rice Genome Array platform (GPL 2025), which contains 57381 probes and each probe corresponded to an individual gene. As there are 123 probe sets designed for control in this microarray. Therefore, after exclusion of these probe set, 57258 valid probe sets are obtained for further analysis.

Data analysis is performed using R (Ihaka and Gentleman 1995) with Bioconductor packages (Gentleman et al. 2004). Raw CEL files are processed using the Robust Multichip Average (RMA) algorithm available in the affy package (Gautier et al. 2004) with steps including background correction, quantile normalization and summarization by the median polish approach (Irizarry et al. 2003). The log2 scale data from RMA is used in statistical analysis. The expression level of the control conditions are taken as baseline. The gene expression level at each stressed conditions are subtracted from the baseline to get the relative expression level. The genes having at least 3-fold change in gene expression are selected and this procedure is continued for each experiment.

The set of genes whose expression level have at least 3-fold change and are commonly expressed in all the experiments, over different stages of growth, different tissues of rice cultivars are selected as TGs. For this study, TGs are identified which are commonly expressed in 4 out of 3 experiments over different tissues and growth stages of rice cultivars. The identified TGs having highly salt responsive nature are validated by published literature.

RiceSRTF data base, which is a good source of rice TFs, is explored to facilitate gene function analysis for abiotic stress responses in rice (Priya and Jain 2013) (http://www.nipgr.res.in/RiceSRTFDB.html). Total 934 TFs are selected from this data base, whose expressions are differentially expressed under salinity stress in rice. These TFs could be considered as the regulators of the TGs. One TG may also be considered as a regulator for another TG and vice versa (Huang et al. 2010). Hence, for a TG, the 934 TFs and the remaining TGs are considered to be regulators and this procedure is repeated iteratively for all TGs.

The candidate regulators from the RiceSRTF database are selected by correlation threshold of gene expression data, which is based on the assumption that there are possible correlations between TG and their upstream regulators (Chen et al. 2008). The Pearson’s correlation co-efficient is computed between the TGs and their own regulatory TFs separately and the correlation value is then used to identify the candidate regulators according to the assumption that the regulatory genes and TGs have a positively (or negatively) correlated with temporal relationship, if, the TG’s expression profile is positively (or negatively) correlated with the regulatory genes profile.

A proper threshold limit for correlations between TFs and TG is chosen to select candidate transcription regulatory genes for each TG. For this purpose, we selected a random sample of 5000 genes from 57258 genes and computed their correlations by the Pearson’s correlation co-efficient and as ranked in Fig 1. Further, by using the step of data randomisation procedure, the probability density function of parameter estimation showed the parameter estimations of regulatory genes by our threshold of correlation has a probability value less than 0.001.

\[ P \left( r_1 \geq r_0 \right) \leq 0.001 \] where \( r_1 \) correlation co-efficient and \( r_0 \) threshold value

According to the Fig 1, 50% (0.7767 in correlation value) is selected as threshold value and the correlation values between the TGs and their regulators greater than threshold value are significant at 0.001 level of significance. In the next step, the candidate regulators are selected from
the pool of regulators. This implies that, if, the correlation between a regulatory TF and the TG is more than threshold value, then the following regulatory TF is selected as a candidate regulator for that TG. After selecting candidate regulators by correlation threshold, these TGs and their candidate regulators are combined to fit the model, which is based on the MLR approach and the regression coefficients are estimated.

Zhang et al. (2010) considered the following model for time series data:

\[ X_{i+1} = AX_i + \epsilon_i \]

Here, \( X_i \) is a vector describing the expression level of \( n \) different genes at time \( t \). \( A \) is an \( n \times n \) matrix, where each entry of \( A \) models the regulatory ability of \( j \)-th gene to \( i \)-th gene. \( \epsilon_i \) is used to model the noise at time \( t \).

The equivalent model of the above model for the cross sectional time-series gene expression can be written as

\[ Y \cdot B \cdot X + \epsilon \]  

(1)

Let’s denote \( y'_i \) be the \( i \)-th row of \( Y \) and \( b'_j \) be the \( i \)-th row of \( B \). Then by simple observation, we have

\[ y'_j = b'_j \cdot x_i \text{ or } y_i = x'_j \cdot b_j \]  

(2)

Adding error term in (2), this looks like the form of standard MLR equation and is given by: model

\[ y_i = x'_i \cdot b_i + \epsilon_i \]

\[ i = 1, 2, \ldots, m \]

\[ j = 1, 2, \ldots, n_t \]

\[ k = 1, 2, \ldots, l \]

where, \( y_i = l \times 1 \) vector of expression values for \( i \)-th TG, \( x_i = n_t \times l \) matrix of expression values of TFs for \( i \)-th TG, \( b_i = n_t \times l \) vector of regulatory strengths, \( \epsilon_i = l \times 1 \) vector of random errors and \( \epsilon_i \sim N(0, \sigma^2I) \).

If the standard estimation procedure like Maximum Likelihood Method is applied to the above model, the estimate of \( b_i \) can be obtained as,

\[ b^*_i = (x'_i X'_i)^{-1} X'_i y_i \]

\[ i = 1, 2, \ldots, m \]  

But the normal regression equation does not work as \( x'_i \) has rank smaller than or equal to \((l-1)\) and is therefore singular. Therefore, the estimation of \( b_i \) we considered SVD of the matrix, \( x'_i = U \Sigma V' \), where \( U \) and \( V \) have \( u_i \) orthonormal columns (left singular vectors) and \( v_j \) (right singular vectors) and \( \Sigma \) is a diagonal matrix with diagonal entries are assumed to be arranged in descending order. Here \( U \) and \( V \) satisfy with

\[ UU' = UU^{-1} = \Sigma V' = VV^{-1} = I \]  

(4)

Applying SVD in the normal Equation (3), it becomes:

\[ b^*_i = V \Sigma^{-1} U' y_i \]  

(5)

This procedure is repeatedly done for all TG. This model suits well when the TG have equal or unequal number of regulators. In equation (5), \( b_{ij} \) is the impact value of influence from gene \( j \) to gene \( i \). The property of regression co-efficient can be used to construct networks with directivity.

\[ j - th \text{ TF acts as a upregulator for } i - th \text{ target gene} \]

\[ b_{ij} > 0 \]

\[ j - th \text{ TF acts as a downregulator for } i - th \text{ target gene} \]

\[ b_{ij} < 0 \]

Based on the above regulatory index, the GRN are constructed by using RCytoscape package (Shannon et al. 2013) of R.

The regulatory strengths of all the possible interactive candidate regulators on TG are found by using regression and SVD approach. But it is still unknown, exactly how significantly the regulatory ability can be regarded as a true potential regulator. In order to decide this, t- test is used for evaluating the significance of our model parameters to prune the unnecessary gene interactions.

\[ t = \frac{b_j}{SE(b_j)} \sim t_{(n-2),a} \]

where, \( b_j \): regulatory strength for \( i \)-th target gene and S. E. \( (b_j) \): standard error of \( b_j \).

The significant interactions between the regulatory TFs and TG were selected at 1% level of significance. The GRN is constructed by retaining all those significant interactions.

RESULTS AND DISCUSSION

Description of target genes

For this study, ten genes are identified as TGs. Among the selected TGs, ‘LOC_Os09g31019’ encoding ubiquitin protein, plays an important role for resistant to salt stress in rice seedling (Liu et al. 2012), ‘LOC_Os06g36850’ encoding Pyridoxal-phosphate dependent enzyme family protein responsible for salt tolerant in a halophytic bacteria Exiguobacterium acetylicum (Rajendran et al. 2009), ‘LOC_Os10g39360’ which encodes Eukaryotic aspartyl protease family protein (EDPG) responsible to salt stress in Arabidopsis thaliana (Wang et al. 2013) and ‘LOC_Os11g18870’ represents a phosphatidylethanolamine
binding family protein that was up regulated at different time points in *Hordeum vulgare* L. (Walia et al. 2005). The other TGs like ‘LOC_Os04g27980’, ‘LOC_Os07g12240’ and ‘LOC_Os01g36560.1’ has regulatory roles in salinity stress in rice (Walia et al. 2005).

**Construction of gene regulatory networks**

Based on the 10 TGs and their potential candidate regulators, GRN is constructed at 1% levels of significance for salinity stress in rice. In these networks, directed graphs are considered with genes, represented by nodes and the edges are represented by different regulation types.

GRN is laid out by considering significant interactions between the genes at 1% level of significance, which has 110 nodes and 173 edges and has clearly 12 gene hubs (which includes ten TGs along with LOC_Os08g09900 and LOC_Os05g14370), these nodes are highly interconnected with each other. Further, to validate the genes (nodes) in *in silico* condition, gene ontology (GO) enrichment analysis is performed by Database for Annotation, Visualization and Integrated Discovery (DAVID), a biological knowledgebase and analytic tool aimed at systematically extracting biological meaning from large gene lists (Huang et al. 2008). The gene enrichment results are listed in Table 1.

In molecular functions, the chosen genes are over represented in the categories of DNA binding, Transcription regulator activity, Transcription factor activity, Sequence-specific DNA binding and also another chosen gene sets are represented in the categories of zinc ion binding, Transition metal ion binding, metal ion binding, cation binding and ion binding all of which may be active due to the high ion concentration in salinity. In biological processes, regulation of transcription, regulation of RNA metabolic process, transcription are over represented followed by auxin mediated signaling pathway, response to auxin stimulus, cellular response to hormone stimulus, hormone mediated signaling, response to hormone stimulus, response to endogenous stimulus, response to organic substance.

**Testing for scale free property of biological network**

The Power law distribution is fitted in the in the above network to test the scale free property of biological networks and the results were represented in Table 2. The high value of co-efficient of determination ($R^2$) as well as the slope value ($b= -1$) indicated that the scale free property of biological network is satisfied to the given network data (Zhang and Horvath 2005). From the facts, it can be concluded that the GRN constructed by taking TGs and their potential regulators for salinity stress in rice satisfied the scale free property of biological networks.

**Construction of refined gene regulatory networks**

In a scale-free network, small-degree nodes are the most abundant, but the frequency of high-degree nodes decreases relative slowly (Barabasi and Albert 1999). The nodes which have degrees much higher than average degree are called as hubs in the network (Albert 2005). The node-degree distribution for the constructed GRN is

<table>
<thead>
<tr>
<th>GO Terms</th>
<th>Functions</th>
<th>Count</th>
<th>Ontology</th>
<th>P-Value</th>
<th>Benjamini</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0006355</td>
<td>Regulation of transcription</td>
<td>50</td>
<td>BP</td>
<td>5.50E-42</td>
<td>1.20E-40</td>
</tr>
<tr>
<td>GO:0003677</td>
<td>DNA binding</td>
<td>49</td>
<td>MF</td>
<td>1.90E-26</td>
<td>6.00E-25</td>
</tr>
<tr>
<td>GO:0006355</td>
<td>Regulation of transcription, DNA-dependent</td>
<td>34</td>
<td>BP</td>
<td>2.50E-25</td>
<td>2.70E-24</td>
</tr>
<tr>
<td>GO:0051252</td>
<td>Regulation of RNA metabolic process</td>
<td>34</td>
<td>BP</td>
<td>2.90E-25</td>
<td>2.10E-24</td>
</tr>
<tr>
<td>GO:0030528</td>
<td>Transcription regulator activity</td>
<td>36</td>
<td>MF</td>
<td>4.70E-24</td>
<td>7.60E-23</td>
</tr>
<tr>
<td>GO:0003700</td>
<td>Transcription factor activity</td>
<td>29</td>
<td>MF</td>
<td>1.50E-21</td>
<td>1.60E-20</td>
</tr>
<tr>
<td>GO:0006351</td>
<td>Transcription</td>
<td>30</td>
<td>BP</td>
<td>3.40E-21</td>
<td>1.90E-20</td>
</tr>
<tr>
<td>GO:0000156</td>
<td>Two component response regulator activity</td>
<td>3</td>
<td>MF</td>
<td>1.40E-02</td>
<td>8.90E-02</td>
</tr>
<tr>
<td>GO:0000160</td>
<td>Two component signal transduction system (Phosphorelay)</td>
<td>3</td>
<td>BP</td>
<td>1.90E-02</td>
<td>8.30E-02</td>
</tr>
<tr>
<td>GO:0008270</td>
<td>Zinc ion binding</td>
<td>15</td>
<td>MF</td>
<td>2.00E-02</td>
<td>1.00E-01</td>
</tr>
<tr>
<td>GO:0046914</td>
<td>Transition metal ion binding</td>
<td>15</td>
<td>MF</td>
<td>5.40E-01</td>
<td>9.60E-01</td>
</tr>
<tr>
<td>GO:0046872</td>
<td>Metal ion binding</td>
<td>15</td>
<td>MF</td>
<td>8.50E-01</td>
<td>1.00E-1</td>
</tr>
<tr>
<td>GO:0043169</td>
<td>Cation binding</td>
<td>15</td>
<td>MF</td>
<td>9.10E-01</td>
<td>1.00E-1</td>
</tr>
<tr>
<td>GO:0043167</td>
<td>Ion binding</td>
<td>15</td>
<td>MF</td>
<td>9.10E-01</td>
<td>1.00E-1</td>
</tr>
</tbody>
</table>

*GO term* is the Gene ontology term. Ontology “BP” indicates Biological Process, Ontology “MF” indicates Molecular Function, and Ontology “CC” indicates Cellular Component. “Count” is the number of genes in set with annotation. “P-value” is the Fisher Exact P-value enrichment test. The smaller P-value is more enriched.
obtained to identify perturbed hub gene interactions that induce salinity stress in rice. The nodes whose connection degree is greater than the average connection degree of the network may be considered as possible molecular target for breeding rice cultivar which may be tolerant in salinity stress condition. By taking the list of nodes whose connection degrees are greater than or equal to 4 (average connection degree), the refined GRN is constructed for salinity stress in rice.

In Fig 2, refined GRN at 1% level of significance contained 19 nodes with 48 edges and there were 8 gene hubs (LOC_Os01g36560.1, LOC_Os10g39360, LOC_Os11g18870, LOC_Os07g12240, LOC_Os04g19740, LOC_Os06g36850, LOC_Os07g33320, LOC_Os05g14370) which were highly connected.

Predicting the target gene responses

To check the validity of the developed model, the responses of ten TGs are predicted by using information on regulators in microarray data with accession number GSE 21651 under salt tolerant and salt sensitive conditions and are represented in Fig 3.

The model based on multiple regression approach is used to generate the TG expression value via the gene expression profiles of the TFs from GSE 21651. The Fig 4 successfully showed that the above model predicted the responses of TGs under salt tolerant condition and salt sensitive with mean squared error 0.4578 and 0.6588 respectively.

In this work, the main aim is the construction and modelling of GRN of salinity stress response in rice. As there are 51 microarrays collected over four different microarray experiments in contrast to tens of thousands of probe sets, it is a great challenge to determine feature selection on this small sample, but high dimension data. Combining the Pearson’s correlation selection algorithm and the t-test, a novel algorithm is created to select potential TFs from the pool of candidate TFs without loss of biological meaning. The small sample and large dimension problem narrows down the applicability of classical statistical techniques like regression in large genomic data sets. This problem was tackled by using SVD along with MLR for estimating the regulatory strengths. The value of mean square error showed that the developed model can be successfully used for predicting the target gene responses under salt stress condition in rice. For testing the scale free property of biological networks, the Power law distribution is fitted to the constructed networks. The results showed that this approach of constructing GRN satisfied this property. For getting a clear idea about the salinity stress response mechanism in rice, the refined GRN is constructed by taking hub nodes. The refined networks did not satisfy the scale free property of biological networks, because the sub-networks of scale free networks are not scale free (Stumpf et al. 2005). From the GO analysis, it is found that most of the hub genes are involved in biological processes like transcription, regulation of transcription, DNA-dependent process, and regulation of RNA metabolic process. For the molecular function category, they are involved in DNA binding, transcription factor activity, transcription regulator activity, sequence-specific DNA binding and followed by DNA binding, zinc ion binding, ion binding, cation binding, metal ion binding, transition metal ion binding, which are due to high ion concentration in soil and water due to salt stress.

In this paper, a method is proposed for construction and modelling of gene regulatory network based on MLR and SVD approach. It is based on database information to construct salinity induced gene regulatory networks in rice.
It is also called a system biology approach for construction of GRN as we process the complex gene network of numerous genes and regulators from different sources at the same time. This approach of construction of gene network considers the reduction of the redundancy of the complex networks step by step. This approach of gene regulatory network construction satisfies the condition of scale free property of biological networks. The validity of the model over other salinity stress rice microarray data set is quite satisfactory. The hubs of gene interactions present in the stress induced gene networks in rice at different levels of significance need to be validated.

Our study may be helpful to set up hypotheses for researchers to design experiments for studying salinity stress response and some guidance for molecular breeders to improve traits. Since some target genes and their potential regulatory TFs have been identified in our study, which means researchers could conduct experiments to clone and validate these genes. The developed model may be used to predict the gene responses in case of missing observations. In our future research work, we will compare the performance our model with some proposed models such as Bayesian, Dynamic Bayesian and Boolean network model by using indices like AUC (Area under curve). Further, the developed computational approaches will be used to study the various stress mechanism in other plant species.

ACKNOWLEDGMENTS

Authors acknowledge for the help from ICAR and NIBMG (Government of India).

REFERENCES


