Plant Gene 1 (2015) 29–34

Contents lists available at ScienceDirect

Plant Gene

journal homepage: www.elsevier.com/locate/plant-gene

In silico identification of regulatory motifs in co-expressed genes under osmotic stress representing their co-regulation

Sanchita, Ashok Sharma *

Biotechnology Division, CSIR-Central Institute of Medicinal and Aromatic Plants, Post Office CIMAP, Lucknow 226015, India

ARTICLE INFO

Article history:
Received 5 September 2014
Received in revised form 7 January 2015
Accepted 12 January 2015
Available online 20 January 2015

Keywords:
Differential gene expression
Coexpression
Coregulation
Solanum tuberosum
Position weight matrix
Transcription factor binding site

ABSTRACT

Osmotic stress is one of the abiotic conditions for plants responsible for osmotic imbalance. The genes have the capability to show differential expression in response to such conditions. In order to understand the role of genes towards multiple stresses simultaneously, a coexpression study is required. In our analysis, the coexpressed genes of *Solanum tuberosum* showed a positive correlation (0.91) in salt and drought stresses. The genes showing similar expression were grouped into five sub groups. Sub group 2 revealed the highest number of genes and formed a network. The genes of this network were found to be coding for different stress related proteins. The largest portion (25%) of genes was found to be coding for lipoxygenase revealing its role in jasmonic acid pathway responsible for abiotic stresses. The coexpressed genes were further analyzed for their regulation by the same regulatory factor. The results suggest that the coexpressed genes were regulated due to presence of similar binding sites for EREBP/AP2. The gene expression, their coexpression, functional annotation and coregulation were integrated to form a network. This approach could yield that the coexpressed genes are under the control of the same regulatory system thus are coregulated and form a network.

1. Introduction

Plants, being sessile, have the ability to dramatically alter the expression patterns of their genes in response to environmental changes (Hazen et al., 2003). These environmental factors affect plants as different abiotic stresses. Several reviews have been published until date on the understanding of abiotic stress responses in plants (Mittler, 2006; Cramer et al., 2011). Among abiotic stresses, salt and drought stresses together are known as osmotic stress. In nature, stress does not generally come in isolation and many stresses act hand in hand with each other so, the plants have to deal with a multitude of challenges (Mittler, 2006). Therefore, there must be the occurrence of a combined effect of more than one stress. The osmotic stress is a common consequence of many of these factors. The exploration of thousands of gene expression data in one experiment is a routine exercise. Many techniques like microarrays have become a standard tool for high throughput omics data analysis. The expression data analyzed from these techniques are available in various public repositories, such as GEO (Edgar et al., 2002), ArrayExpress (Parkinson et al., 2007), Stanford Microarray Database (SMD) (Gollub et al., 2003). The gene expression data provide the functional information of the desired genes to the biologist. The identification of stress related genes by different approaches of genomics, transcriptomics, proteomics and metabolomics has revealed the effect of stress responses in signal transduction pathways (Gechev and Hille, 2012). The genes have the capability to show response to many stress factors imposed simultaneously and participate in a process, related to each other. Different correlation calculations have been widely used for grouping of the omics data with similar expression profiles. The clustering analysis eg. hierarchical (Eisen et al., 1998) uses different correlation calculations for groupings of genes based on expression values. The co-expression analysis thus reveals the response of genes towards more than one stress. The genes involved in co-expression analysis have been underlying in molecular network formation. The co-expressed genes might be validated by their regulation, having similar cis-regulatory elements for a transcription factor. The discovery of co expression and co regulation is an important goal of analyzing gene expression data (Zhang et al., 2004). The co-regulation study validates the relationship of correlated genes. To quantify the similarity of gene expression patterns, various statistical correlation calculations can be performed using expression values of genes. To easily analyze these data without programming skills, several co-expression tools have been constructed based on measures of correlation coefficient. These tools are devoted to the analysis of expression data of model plants only. Most of these tools have their own database linking facility to explore the expression data for analysis. There are other tools available for analyzing co-

* Corresponding author.
E-mail address: ashoksharma@cimap.res.in (A. Sharma).

http://dx.doi.org/10.1016/j.plgene.2015.01.001

expression. Gene Co-expression Analysis Toolbox (GeneCAT) (Mutwil et al., 2008) forms the co-expression network of genes from plants having known genomes such as *Arabidopsis thaliana*, *Populus* sp., *Hordeum vulgare*, and *Oryza sativa*. The *Arabidopsis* Co-expression Tool (ACT) (Jen et al., 2006; Manfield et al., 2006) analyzes data of *A. thaliana* from both single and multi experiments. CressExpress (Co-expression analysis for Arabidopsis) (Srinivasasainagendra et al., 2008) is used to compute patterns of correlated expression among genes of *A. thaliana*. The tools, each with their own advantages offer a range of different features but for our dataset of non-model plants, CoExpress tool (Nazarov et al., 2013) was found to be suitable. In addition to the tools, different databases are also available: the databases PLAnt co-EXpression database (PLANEX) (Yim et al., 2013), CoP (Ogata et al., 2010) and PLEXdb (Dash et al., 2012). These databases have the coexpression data of known genomic sequences of *A. thaliana* (thale cress) and seven crops, *Glycine max* (soybean), *H. vulgare* (barley), *O. sativa* (rice), *Populus trichocarpa* (poplar), *Triticum aestivum* (wheat), *Vitis vinifera* (grape) and *Zea mays* (maize). These databases can be used to explore further the functional as well as regulatory analyses. The genes showing similar expression profiles across many experiments represent co-expression. The prediction of function of unknown genes might be facilitated by co-expression analysis (Horan et al., 2008). The involvement of co-expressed genes in the same biological processes can also be studied (Loraine, 2009). In the present study, commonality of cis-regulatory elements in co-expressed genes has been studied. The genes were from *Solanum tuberosum*, showing change in their expression in *Nicotiana tabacum*. For predicting the cis-regulatory elements, the upstream sequences have been obtained from *Solanum phureja*, a cultivar of *S. tuberosum*. The genome sequences of this plant have been used because the genome of *S. tuberosum* is not available till date. The co-expressed genes can be further explained by their involvement in molecular networks. Lee has explained modeling of complex systems through network analysis (Chae et al., 2012). Various biological systems, including protein–protein interaction (Fukao, 2012), metabolic, gene co-function, co-expression (Mao et al., 2009) and regulatory networks in plants have already been investigated (Chae et al., 2012). In this study, network analyses of proteins supposed to be encoded from co-expressed genes were also performed.

2. Material and methods

2.1. Retrieval of gene expression data

The genes of *S. tuberosum* showing differential expression in *N. tabacum* under salt and drought stress were retrieved from GEO (http://www.ncbi.nlm.nih.gov/geo/) database of NCBI. The expression data were retrieved from the series IDs GSE8158 http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8158 and GSE8161 http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8161. The experimental details have been provided in our earlier paper (Sanchita et al., 2013). Each gene has its corresponding intensity values analyzed at different time points after the first application of the stress. The different values obtained for each gene under the abiotic stress in different time periods show the differential expression of the gene with respect to the control condition. The mean of intensity values of different time periods were analyzed.

2.2. Filtering of data

The data obtained from the GEO database, were filtered to obtain highly up and down regulated genes for both the stress conditions. The genes having fold change value greater than or less than 1 were considered as differentially expressed genes. To obtain highly up and down regulated genes, a cut-off of +1.5 and −1.5 fold change was taken. The log base 2 value of fold change was taken as intensity value. All the samples represent their expression values in the form of their intensity value. To filter the data, the mean of intensity values of each sample was calculated. As 1.5 fold change value is equal to 0.585 intensity value, the genes having mean intensity values less than +0.585 and greater than −0.585 were filtered out.

2.3. Co-expression and network formation

CoExpress v.1.5 (http://www.bioinformatics.lu/CoExpress/) was used for co-expression analysis (Nazarov et al., 2013). The co-expressed genes form sub-networks of related genes. The string tool (http://string-db.org/) was used to analyze and visualize the network connection of genes (Szklarczyk et al., 2011). The genes that belong to the same sub-network were interconnected to each other.

2.4. Upstream sequence mining

For finding, the upstream regions for co-expressed genes, the genome of *S. phureja*, a cultivar of *S. tuberosum* has been used. Approximately, 5 kb upstream sequence, showing an exact alignment with the sequences was taken further.

2.5. Prediction of promoters

The Neural Network Promoter Prediction (NNPP) (http://www.fruitfly.org/seq_tools/promoter.html) tool (Reese, 2001) was used for searching promoter regions in the upstream sequences. A threshold value of 0.9 was used for promoter finding which lies between 0.1 and 1.

2.6. Formation of position weight matrix (PWM)

A position weight matrix (PWM) is a matrix that specifies the frequency distribution of nucleotide at each position. It is also known as position-specific scoring matrix (PSSM) and is commonly used for the representation of the occurrence of motifs (TFBSs) in the biological sequences (Sinha, 2006). The known TFBSs were used to predict TFBSs in differentially expressed genes through PWMs formation. The TFBSs were first converted into consensus sequence using Regulatory Sequence Analysis Tool (RSAT-consensus) (http://rsat.ulb.ac.be/consensus_form.cgi) (van Helden et al., 1998; Thomas-Chollier et al., 2011). The resulting consensus sequence was used as input to RSAT-convert matrix (http://rsat.ulb.ac.be/convert-matrix_form.cgi). RSAT-convert matrix converts the consensus sequence into position weight matrices (PWM). The resulting PWM was cross validated by the program D-matrix (http://203.190.147.116/dmatrix/home.aspx) (Sen et al., 2009).

2.7. Prediction and similarity search of TFBSs

The analyzed promoter regions of differentially expressed genes were used for prediction of TFBSs. The RSAT-patser tool (http://rsat.ulb.ac.be/patser_form.cgi) was used to scan these promoter sequences by PWM generated from known TFBSs. Multiple putative TFBSs for EREBP/AP2 transcription factor were obtained. The alignment between TFBSs of coexpressed genes was done using the sequence alignment tool, EMBOSS water-pairwise sequence alignment (http://www.ebi.ac.uk/Tools/psa/emboss_water/nucleotide.html). EMBOSS water uses the Smith–Waterman algorithm to do the pairwise sequence alignment.

3. Results and discussion

3.1. Analysis of gene expression data

17,453 genes with intensity values in different time periods were retrieved. A sample data set of genes with corresponding intensity values for both the stress conditions has been listed (Tables 1 and 2). In the tables, the first column depicts the gene IDs. The five successive