Genomics Data 7 (2016) 214–221

Contents lists available at ScienceDirect

Genomics Data

journal homepage: www.elsevier.com/locate/gdata

Computational gene expression profiling under salt stress reveals patterns of co-expression

Sanchita, Ashok Sharma *

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

ARTICLE INFO

Article history:
Received 20 October 2015
Received in revised form 11 January 2016
Accepted 14 January 2016
Available online 15 January 2016

Keywords:
Clustering algorithms
Gene expression
Co-expression
Functional annotation
Abiotic stress

ABSTRACT

Plants respond differently to environmental conditions. Among various abiotic stresses, salt stress is a condition where excess salt in soil causes inhibition of plant growth. To understand the response of plants to the stress conditions, identification of the responsible genes is required. Clustering is a data mining technique used to group the genes with similar expression. The genes of a cluster show similar expression and function. We applied clustering algorithms on gene expression data of *Solanum tuberosum* showing differential expression in *Capsicum annuum* under salt stress. The clusters, which were common in multiple algorithms were taken further for analysis. Principal component analysis (PCA) further validated the findings of other cluster algorithms by visualizing their clusters in three-dimensional space. Functional annotation results revealed that most of the genes were involved in stress related responses. Our findings suggest that these algorithms may be helpful in the prediction of the function of co-expressed genes.

1. Introduction

Heat, cold, drought and salinity are common abiotic stress conditions for the plants. These stresses affect both plant growth and production [1]. Salt stress presents an increasing threat to agriculture due to its effect on plant growth [2]. Reduced availability of water, increased respiration rate, altered mineral distribution, membrane instability, failure in the maintenance of turgor pressure are some of the events that prevails during salt stress [3]. The response of plants to salinity consists of numerous processes that function in coordination to reduce the cellular hyper-osmolarity and ion disequilibrium [4]. It is important to analyze the function of stress-inducible genes not only to understand the molecular mechanisms of stress tolerance but also to improve the stress tolerance of crops by gene manipulation. Gene expression analysis or expression profiling refers to the study of response of an organism against environmental changes [5]. The expression profiles of thousands of genes results from microarray technology, a preferred method to identify genes involved in abiotic stress responses [6,7]. In addition, analysis of relationship between genes, their functions and classification is also required. Because of the large number of genes and the complexity of biological networks, clustering algorithms are found to be useful exploratory technique for the analysis of gene expression data [8]. To make some meaningful biological inference from large number of genes or samples, they are required to be clustered together to obtain co-expressed genes. The co-expressed genes may have similarity in their expression levels. Cluster analysis is one of the primary statistical tool for data analysis [9,10]. The clustering algorithms have been proven useful for identifying biologically relevant groups of genes or samples and can be applied for gene function discovery process. Clustering algorithms can be classified mainly as unsupervised like Hierarchical clustering (HCL) [11], K-means clustering (KMC) [12], Principal Component Analysis (PCA) [13] and Self Organizing Map (SOM) [14] or supervised like Support Vector Machine (SVM) [15]. These clustering techniques are based on the hypothesis that the genes in a cluster might share common function and regulatory elements [16]. The cluster analysis routinely run as a first step of data analysis and separates a set of objects into several subsets based on their similarity [17,18]. None of the clustering algorithms available provides evidence to support that clustering of genes having similar expression patterns is more likely to have similar biological function [19]. The genes with similar expression profiles are said to be co-expressed genes [20]. Co-expressed genes may help in revealing useful biological information as they are functionally related [11,21]. Hence, grouping of genes with similar expression levels can reveal the function of those genes, which were previously uncharacterized. To find out the utility of clustering techniques, the gene expression profiles of the genes of *Solanum tuberosum* showing differential expression under salt stress at six different time points in *Capsicum annuum* were retrieved and analyzed. The clusters, obtained from different algorithms showed

---

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

http://dx.doi.org/10.1016/j.gdata.2016.01.009

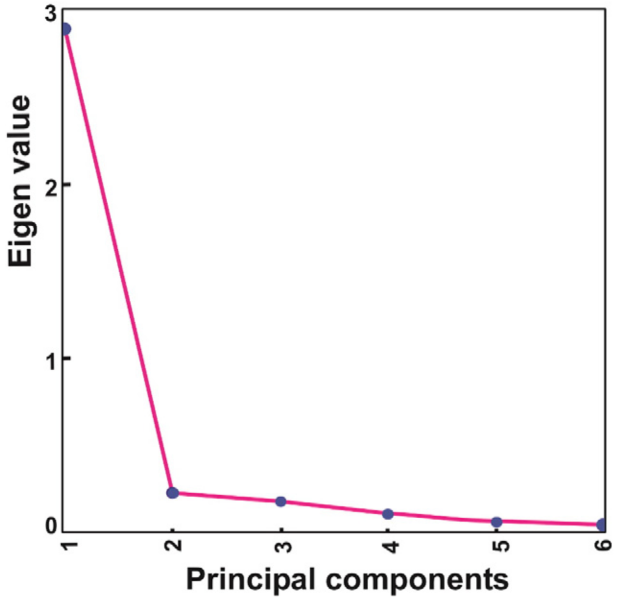

Fig. 1. Eigenvalue plot of the principal components. The six principal components (PC) represent six samples of the gene expression data. The PC containing large eigenvalue represents high variability among its genes.

common patterns of genes having similarity in their expression and function.

2. Materials and methods

2.1. Retrieval of gene expression data

The normalized gene expression data of *S. tuberosum* showing differential expression in *Capsicum annuum* were retrieved. As per experimented details provided in GEO database, the ESTs of *S. tuberosum* were used as spotted PCR amplified cDNA array. After the application of 150 mM NaCl at different time periods in *C. annuum*, the leaf tissue samples were collected including control plants. The RNA from these tissues were isolated and hybridized with cDNA arrays. The genes, showing response under salt stress at six different time points were considered as six samples and were downloaded from Gene Expression Omnibus (GEO) database at NCBI (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8158) [22]. The sample ids were GSM201905, GSM201906, GSM201907, GSM2019058, GSM201909 and GSM201910 (Supplementary Table 1). The filtering of dataset was done in order to obtain highly expressed and significant genes.

2.2. Clustering of genes

Clustering of the genes having similar expression profile was done using the clustering tool, Genesis 1.7.6 [20]. The genes expression data

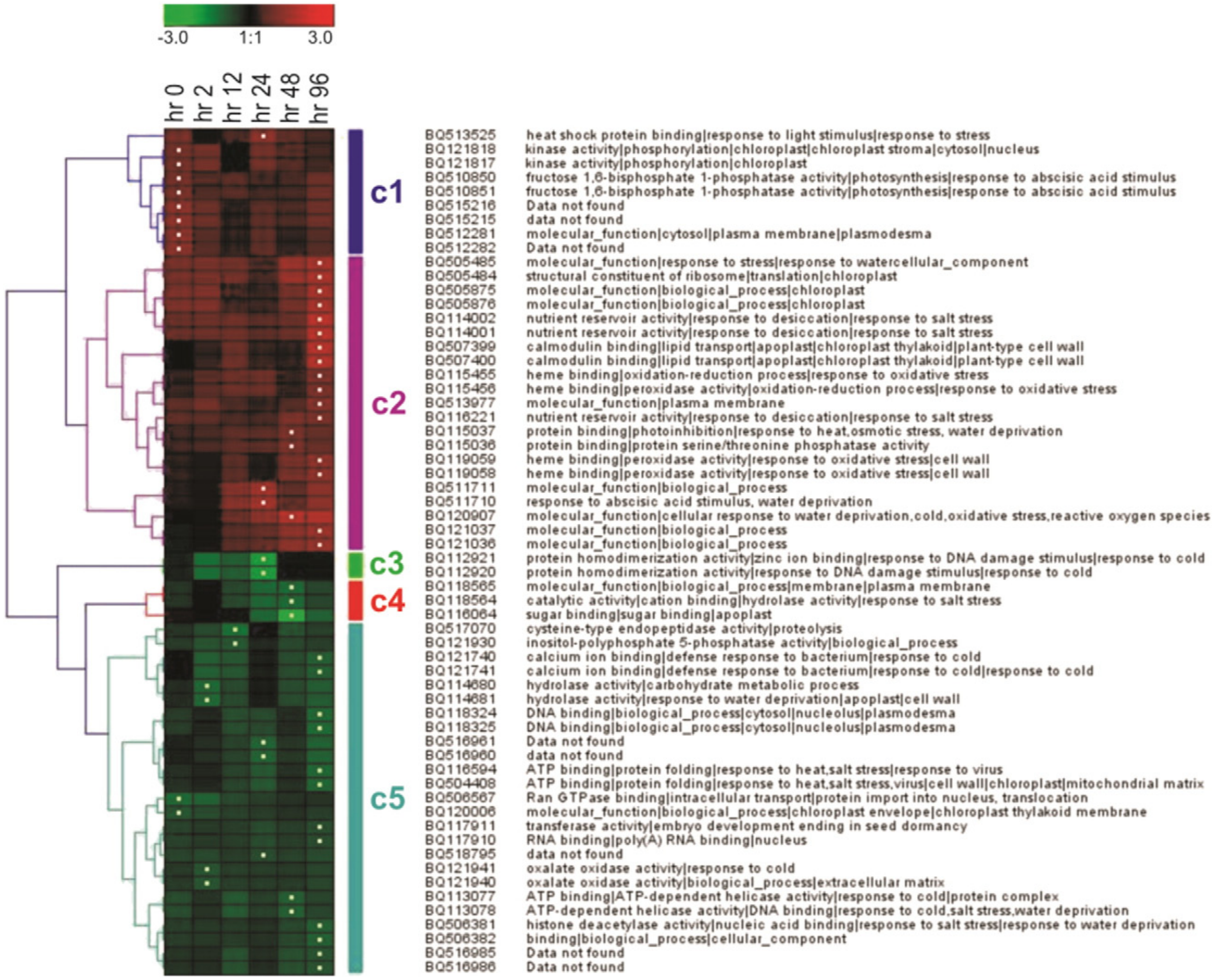

Fig. 2. Heatmap of clustered genes showing five clusters. The hierarchical clustering of the genes is represented in heatmap. The red and green color shows the up and down expression respectively. The range of expression values is −0.3 to 0.3. The genes are represented in form of their accession number followed by their function.