Mining of light-related genes in grapes based on a time-frequency analysis

Longlong Liu[1,a] , Jie Zhou[1,b*] , Zichen Lu[2,c] and Meng Ma[1,d]

[1]School of Mathematical Sciences, Ocean University of China, Qingdao 266100, China

[2]Cloud Computing Center, Chinese Academy of Sciences

[a]xinxijishu@ouc.edu.cn, [b]zhoujie7226@126.com, [c]luzichen@163.com, [d]15254228126@163.com

Keywords: differentially expressed genes, correlation network, Grape gene, degree number, regulatory relationship.

Abstract. In order to better research the Grape gene expression data of temporal series, first of all, we built an undirected network by data preprocessing, screening for differential expressed genes, wavelet transform, correlation network analysis, and then carried out a correlation degree number analysis. Finally, inputting the differential degree number into Mapman software to annotate gene function, we could find out those differentially expressed genes playing a role in cell function, metabolism and transcription etc; differentially expressed genes of both genotypes show an overall upward trend under the long photoperiod than short photoperiod; the long light causes cell function of SV genotype grapes to be enhanced and may accelerate the grape's growth; Genes of SV grapes are more active than that of VR grapes in long and short light conditions. The model is very effective for mining the differential expression of genes comparatively related to the light and genotype.

Introduction

The growth and quality of Grape is often affected by various environmental factors, such as strong light, high and low temperatures, freezing, drought, salinity, acid soils and so on. For the analysis on differentially expressed genes correlating with growth condition, commonly used methods are the multiples factor analysis [1] , t-test in statistical analysis [2] and analysis of variance [3] , significance analysis of chip[4]. Correlation analysis is an important technique of studying regulation relationships between genes for the use of gene expression profile data, however, if correlation analysis of gene expression data is performed directly, it only study the regulatory relationship between two genes and cannot provide complex analysis for comparing and ordering across multiple genes [5]. The complexity of the calculation is another risk. We built an undirected network by data preprocessing, screening for differential expressed genes, wavelet transform, correlation analysis, and then carried out differentially expressed genes' function through differences of correlation degree numbers analysis.

Data Preprocessing

The NCBI database provides probe names and gene expression data for VR grapes (Vitis riparia) and SV grapes (V.spp.Seyval) in the GSE17502 platform [6].We used 84 samples encompassing 14446 gene expression data that were divided into two groups according to the different experimental conditions. One group was maintained under constant long photoperiod (LD) and the other group was maintained under ambient short photoperiod (SD) cycles. Because there are 2 genotypes of grape (SV and VR), LD group was divided into SVLD group and VRLD group (21 samples in each group), and similarly SD group was divided into SVSD group and VRSD group (21 samples in each group).

To remove the impact of the obtained data under the different experimental conditions on the model, each gene record was normalized to [0, 1].

Time-frequency Analysis Model

Euclidean Distance Discrepant Analysis

To explore the effect of light on gene expression in the 2 types of grape, key genes with large expression differences were extracted. It also focuses on the level of differential gene expression between the two types of grape in the same light conditions. For LD group and SD group, each gene has two sets of expression data consisting of two 21-dimensional vectors. Genes with large differences between the LD group and the SD group were considered to be key genes which are significantly affected by light. For further analysis, 1000 genes with large Euclidean distance of two 21-dimensional vectors were selected.

Wavelet Analysis

Affected by various factors, gene expression data at discrete time points are non-stationary signals containing a variety of frequency components. To further mine information and obtain detailed signals with high precision and integrated approximate signals, different time-frequency windows were adopted to transform gene expression data into different time-frequency domains [7]. The Mallat algorithm in the MATLAB toolbox was used to achieve wavelet transforms for each gene record. The approximate signals and the detailed signals of each gene record were acquired.

In each group, approximate signals reflect general characteristics of the original signals, whereas detailed signals show the fluctuation information of the original signals at different frequencies, the detailed signals of gene expression data, which reflect changes in gene expression, are most important. The frequency vector of each gene is composed of an approximate signal and a detailed signal is a 21–dimensional vector. The frequency vectors of gene i in the LD and SD groups respectively are $X_i^L = (x_{i1}^{LA}, \mathbf{L}\, x_{ij}^{LA}, \mathbf{L}\, x_{i11}^{LA}, x_{i1}^{LH}, \mathbf{L}\, x_{ij}^{LH}, \mathbf{L}\, x_{i10}^{LH})$ and $X_i^S = (x_{i1}^{SA}, \mathbf{L}\, x_{ij}^{SA}, \mathbf{L}\, x_{i11}^{SA}, x_{i1}^{SH}, \mathbf{L}\, x_{ij}^{SH}, \mathbf{L}\, x_{i10}^{SH})$. where the superscripts SA and SH represent the low- and high-frequency coefficients of the gene expression data in the SD group, and the superscripts LA and LH represent the low- and high-frequency coefficients of the gene expression data in the LD group.

Pearson's Correlation Network Analysis in Frequency Domain

A correlation analysis quantitatively analyzes the similarity and dependence of signals in the frequency domain between two differentially expressed genes. Because signals in the frequency domain are composed of approximate signals and detailed signals, the correlation degrees of both the high- and low-frequency coefficients of the gene signal are taken into account during correlation analysis. A correlation analysis for each pair of the 1000 key genes in each group was performed and a correlation network was built. The degree number of a gene in a correlation network indicates the gene's degree of importance in some sense. The 300 genes were detected according to this principle in our model. The specific algorithm is as follows:

(1) Calculate the covariance of the frequency vectors X_i of gene i and X_j of gene j.

$$c_{ij} = E((X_i - E(X_i))(X_j - E(X_j))). \tag{2-1}$$

(2) Calculate the correlation coefficient between the frequency vectors X_i and X_j.

$$r_{ij} = \frac{c_{ij}}{\sqrt{c_{ii} \cdot c_{jj}}}. \tag{2-2}$$

Here, the correlation matrix $R = (r_{ij})_{1000\times1000}$ was obtained. The correlation matrix R is a real symmetric $M_i = l_i^L - l_i^S$ matrix whose element r_{ij} represents the correlation coefficient between X_i and X_j.

(3) In the matrix R, if gene i and j are highly correlated ($0.7 \le |r_{ij}| \le 1$), there is an edge between gene i and j, otherwise, there is not an edge. Here, the correlation network is built.

(4) Do not consider autocorrelation, i.e. $r_{ii} = 0$.

(5) Calculate the degree number of every gene in correlation network built by LD group (X_i^L) and the SD group (X_i^S) respectively.

(6) The degree differences of gene i between the LD group and the SD group was obtained by subtracting the degree number l_i^S in the SD correlation network from the degree number l_i^L in the LD correlation network. The first 300 genes with big degree differences were accepted.

$$M_i = l_i^L - l_i^S \quad . \tag{2-3}$$

After getting a correlation degree number of the 300 key genes, the square ratios of negative degree differences and positive degree differences of both the same genotype differentially expressed genes were calculated so as to analyze the effects of the different genotype in the different light conditions [8].

The square ratios of negative and positive degree differences are defined as follows:

$$h_P = \sum_{j=1}^{n_2} M_j^2 \Big/ \sum_{k=1}^{300} M_k^2 \quad . \tag{2-4}$$

$$h_N = \sum_{i=1}^{n_1} M_i^2 \Big/ \sum_{k=1}^{300} M_k^2 \quad . \tag{2-5}$$

Here, i and j represent genes with negative degree differences and positive degree differences, respectively. $i = 1,\mathrm{K}\ ,n_1$, $j = 1,\mathrm{K}\ ,n_2$, $k = 1,\mathrm{K}\ ,300$.

Numerical Experiments and Result

The Excavation of Differentially Expressed Genes in SV Grapes under the Long and Short Photoperiod

The differentially expressed genes in the SV grapes under the long and short photoperiod were analyzed, and the genes comparatively related to the light were found out. In other words, the gene expression data of SVLD group and SVSD group was took to the model (section 2) of this article for degree analysis.

It was discovered that there were 181 genes whose correlation degree numbers in the SVLD group were greater than those in the SVSD group; however, there were also 119 genes whose correlation degree numbers in the SVSD group were greater than those in the SVLD group. By (2-4) and (2-5), we can get the square ratios of negative degree differences and positive degree differences are 30.76% and 69.24%.This shows that the expression level of differentially expressed genes in 300 SV grapes under the long light is higher than the short light.

The Excavation of Differentially Expressed Genes in VR Grapes under the Long and Short Photoperiod

To do the same thing to VR grapes .It was indicated that there were 195 genes whose correlation degree numbers in the VRLD group were greater than those in the VRSD group; however, there were also 105 genes whose correlation degree numbers in the VRSD group were greater than in the VRLD group. The square ratios of negative degree differences and positive degree differences are 19.52% and 80.48%. This shows that the expression level of differentially expressed genes in 300 VR grapes under the long light is higher than the short light.

The Excavation of Differentially Expressed Genes in SV and VR Grapes under the Long Photoperiod

The differentially expressed genes in the SV and VR grapes under the long photoperiod were analyzed, and the genes comparatively related to the light were found out. The gene expression data of SVLD group and VRLD group was took to the part 3 model for degree analysis.

There were 195 genes whose correlation degree numbers in the SVLD group were greater than those in the VRLD group; however, there were also 105 genes whose correlation degree numbers in the VRLD group were greater than in the SVLD group. The square ratios of negative degree differences and positive degree differences are 19.99% and 80.01%.This displays that the expression level of SV grapes is higher than VR grapes under the long light.

The Excavation of Differentially Expressed Genes in SV and VR Grapes under the Short Photoperiod

To do the same for SV and VR grapes under the short photoperiod .It was showed that there were 234 genes whose correlation degree numbers in the SVSD group were greater than those in the VRSD group; there were also 66 genes whose correlation degree numbers in the VRSD group were greater than in the SVSD group. The square ratios of negative degree differences and positive degree differences are 7.68% and 92.32%.This shows that the expression level of SV grapes is higher than VR grapes under the short light.

Conclusion

By the time-frequency analysis model, this paper reaches that the square ratios of positive degree differences are far greater than the negative degree differences in two groups. Compared with the short light, SV genes and VR grapes show the overall up-regulated trend under the long light conditions. Similarly, under the long light and the short light conditions, contrasting with two kinds of genotype, the result shows that the SV grape's genes are more active than the VR grape's genes. These results are in accordance with the researches of biologist. The model is very effective for mining the differential expression of genes comparatively related to the light and genotype.

Acknowledgments

The authors would like to thank all of the researchers who made publicly available data used in this study. The authors would like to thank the National Natural Science Foundation of China (No: 61303145) and the University Basic Research Foundation (No: 201362031) for the support to this work.

References

[1] Gerhold D, Lu M, Xu J, et al. Monitoring expression of genes involved in drug metabolism and toxicology using DNA microarrays. Physiol Genomics, 2001, 5: 161-170.

[2] Baldi P, Long A D. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics, 2001, 17: 509-519.

[3] Pavlidis P. Using ANOVA for gene selection from microarray studies of the nervous system. Methods, 2003, 31: 282-289

[4] Marshall A, Hodgson J. DNA chips: An array of possibilities. Nature Biotechnology, 1998, 16: 731

[5] Sun-Goo Hwang, Dong Sub Kim, et al. Identification of rice genes associated with cosmic-ray response via co-expression gene network analysis. Gene 541 (2014) 82-91.

[6] Lekha, S., Je´roˆme, G., Julie, AD: Differential floral development and gene expression in grapevines during long and short photoperiods suggests a role for floral genes in dormancy transitioning. Plant Mol Biol 2010, 73:191-205.

[7] Shaoxiong Wu. Intelligence statistical process control in cellular manufacturing based on wavelet transform and probabilistic neural network. Journal of computational information systems 6:10 (2010) 3463-3470.
[8] Longlong Liu, Jieqiong Qu, Xilong Zhou, et al. Discovery of a strongly-interrelated gene network in corals under constant darkness by correlation analysis after wavelet transform on complex network model. Plos One, 2014, 3:1-7.