International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor (2018): 7.426

Simple Analysis of the Gene Function of Saccharomyces Cerevisiae in Different Phases

Shuyue Zhu[1]

[1]Hebei University of Technology, School of Science, Tianjin300401, China

Abstract: ***In this paper, the Saccharomyces cerevisiae genome was used as the research object. The yeast gene expression data (8 groups) measured by different laboratories, they were used to construct a complete cell cycle Saccharomyces cerevisiae gene expression database. The aim is to study the functional characteristics of genes in four different phases of the cell cycle.***

Keywords: Saccharomyces cerevisiae; Cell cycle; Gene function; Co-expression network

1. Introduction

Saccharomyces cerevisiae is the first complete genome sequencing of eukaryotic organisms. Compare with the complex human genome, S. cerevisiae has a relatively simple gene composition and structure, it is often used as a model organism for functional genomics research [1] . In addition, Saccharomyces cerevisiae is widely used in the fields of food, medicine and health care, brewing, energy and chemical industry, life science and environmental protection. As the very important industrial microbial resource, Saccharomyces cerevisiae contributes its strength to the continuous development of human beings.

All the genetic characteristics of organisms are regulated by genes, which are essential to all organisms. When a gene in an organism mutates in the course of its growth, it will result in inability to express or lethal phenotypes. These genes are called essential genes [2], and the products of essential genes are also necessary to maintain the basic life activities of organism cells [3]. Saccharomyces Genome Deletion Project was used to classify the genetic importance of Saccharomyces cerevisiae (http://www-sequence.stanford.edu/group/yeast_deletion_project/).

The functional classification data of Saccharomyces cerevisiae genes are derived from the Clusters of Orthologous Groups of proteins (COGs: http://www.ncbi.nlm.nih.gov/COG/), which are organized according to different structures or cell evolutionary relationships. According to the function of genes encoding proteins, it can be divided into four major categories and 25 sub-categories.

2. Materials and Methods

In 2007, Gauthier et al. [4]integrated high-throughput cell cycle gene expression data and established an online database on cell cycle experiments - Cyclebase.org (http://www.cyclebase.org/). In Cyclebase database (version 3.0) [5], the gene cycle expression data and the cell cycle type of Saccharomyces cerevisiae from several laboratories were integrated. Cyclebase database integrates gene cell cycle expression data from eight different experiments, including 6717 gene expression data. Because the data of gene expression come from eight different laboratories, the experimental conditions of measuring gene expression and the starting point of cell cycle are different. Aim to use and compare, it is necessary to process these data. Processing process : (a) All experiments were placed on a common time scale, eliminating the differences caused by different experimental conditions. (b) Align the time scales of all experiments, eliminating the differences caused by the different starting points of cell cycle measurements. (c) Using normalized processing, eliminating the differences in gene expression values.

In Cyclebase database, the number of complete cell cycles(about 2-3 complete cycles) and the number of samples measured is different. Considering the integrity and comparability of data, we should choose a complete cycle of data as the basis for further analysis. The rank values of genes are given in the original database. Here we select 100 genes (rank < 101), then classify the genes into four categories: G1, S, G2 and M based on the peaktime, mapping gene expression profiles at last. Through many analyses and calculations, the time of a complete cell cycle were 105 min to 204 min. And the time of each period was 25% of the complete cell cycle time. In this database, 1312 genes were expressed in all 8 groups of experiments, while 11 genes were expressed in only one group of experiments. Considering the need of statistical analysis and the integrity of the data, we selected the genes that expressed data at least in seven groups of experiments as the genes for further analysis. At the same time, these genes are required to match the corresponding nucleotide (gene) sequence. Finally, 5378 genes were screened to meet the above conditions. In the data downloaded, there are 1156 genes necessary for Saccharomyces cerevisiae cell life activities. Comparing with gene expression data, we conclude that 1000 of 5378 genes are essential genes.

Gene co-expression network is a network structure map based on the similarity of gene expression data. It can be used to study the interaction between biological molecules. Express the regulation relationship between genes through graph model [6]. It has been used to identify cell modules [7]and predict the function of genes encoding unknown proteins [8]. In this paper, we use WGCNA[9] to construct a

S. cerevisiae gene co-expression network, in which nodes in the network represent Saccharomyces cerevisiae genes, while genes with similar expression values are linked to form the edges of the network.

The network construction process is as follows: (a) Prepare to establish data of gene co-expression network, calculate variance of gene expression value and select genes for further analysis. (b) Determine the β, then use the "soft threshold" to calculate the adjacency matrix between genes. (c) Computing topological overlapping matrix. (d) Construct clustering tree, static clipping method is used to clip the clustering tree. (e) Analysis of gene modules.

Constructing the co-expression network in four phases of the cell cycle (G1, S, G and M), selected β were 9, 26, 8, and 26. When performing gene module, 2000 genes were selected for further cluster analysis according to the maximum number of genes. When using static shear method, the shear heights of four phases are 0.97, 0.997, 0.995 and 0.995.

3. Results and Discussion

In the WGCNA, the size of the clustering module is represented by different color, where grey indicates genes that are not clustered together. The colors of the modules from big to small are: turquoise, blue, brown, yellow, green and red. Fig.1 is a clustering tree for G1 phase of cell cycle. Seven gene modules were obtained by using static shear method (shear height was 0.97).

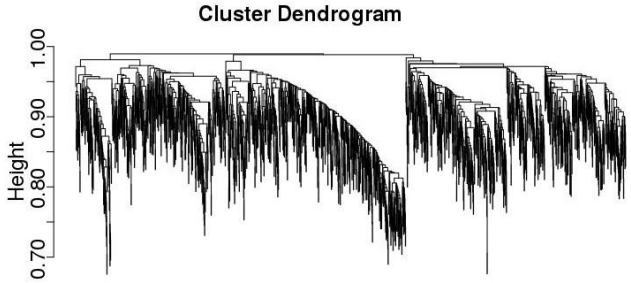

Figure 1: Cluster Dendrogram in G1 phase

Essential genes play an important role in cells. Therefore, it is important to analyze the distribution of essential genes in the gene module, that is, the content ratio of the necessary genes in the gene module. Fig.2 shows the significance of Gene Module in G1 phase.

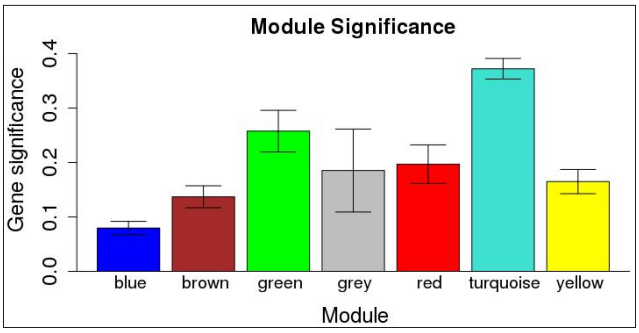

Figure 2: The significance of gene module in G1 phase

It can be concluded that the proportion of essential genes in turquoise module is the highest, suggesting that this module has important biological functions. In order to more accurately describe the biological functions of genes in turquoise module, this paper uses STRING 10.5 (https://string-db.org/) [10]to analyze The Gene Ontology (GO) of module genes [11]. GO analysis can be divided into three categories: biological process, molecular functions and cell component. The gene or protein can find the corresponding GO number by the method corresponding to ID or sequence annotation, and the GO number can correspond to Term, ie cell localization or functional category. The turquoise module GO enrichment analysis results in a co-expression network consisting of 653 nodes and 19,866 edges (p-value: < 1.0e-16). After analyzing the gene function in turquoise module, the following results were found: there are many genes in turquoise module which have the functions of DNA transcription, translation, protein folding, modification, cell metabolism and so on. We use the same method to analyze the other three phases.

In S phase, the proportion of essential genes in brown module is the highest, GO analysis of gene function in this module shows that there are 319 nodes and 4059 edges (p-value: < 1.0e-16). The enrichment analysis shows that there are many genes in brown module that can store and process information such as DNA transcription, RNA processing and modification; there are a large number of genes related to protein folding and modification and so on.

In this paper, we constructed a gene co-expression network in G2 phase of cell cycle. Six gene modules were obtained by clipping the cluster tree using static clipping method (the height of clipping is 0.995). The proportion of essential genes in brown module is the highest. GO analysis of the gene function in this module, resulting in a co-expression network consisting of 235 nodes and 961 edges (p-value: < 1.0e-16). The function of genes in brown module was found by enrichment analysis: the number of genes with RNA processing and modification, amino acid transport and metabolism, energy production and conversion was higher in this module than in other related functions.

At last, we constructed a gene co-expression network in M phase of cell cycle. The clustering results of cell cycle M phase were significantly different from those of other three phases. In M phase, about 75% of the genes were in grey module, and only 455 genes were distributed in the other six gene modules. Statistical analysis showed that the proportion of essential genes in turquoise module and yellow module was 0.54 and 0.59, which was much higher than other gene modules. There are 110 genes in the turquoise module, A co-expression network (p-value: < 1.0e-16) consisting of 110 nodes and 1939 edges is obtained by enrichment analysis. The module has a large number of genes with functions related to DNA transcription, RNA processing and modification.

Constructing co-expression network based on different phases of cell cycle, more gene co-expression information

can be obtained, which is helpful for gene module mining and gene function analysis. The gene cycle expression data used in this paper are from different experiments, and the data samples are relatively small. We look forward to more data on gene cycle expression to make the statistical results more accurate.

References

[1] Botstein, D. and G. R. Fink . "Yeast: An Experimental Organism for 21st Century Biology." Genetics 189.3(2011):695-704.
[2] Chen, Lei, et al. "Prediction and analysis of essential genes using the enrichments of gene ontology and KEGG pathways." Plos One 12.9(2017):e0184129.
[3] Lin, Yan, et al. "Identifying bacterial essential genes based on a feature-integrated method." IEEE/ACM Trans Comput Biol Bioinform PP.99(2017):1-1.
[4] Gauthier, Nicholas Paul, et al. "Cyclebase.org—a comprehensive multi-organism online database of cell-cycle experiments." Nucleic Acids Research 36.Database issue(2008):D854-D859.
[5] Santos, Alberto, R. Wernersson and L. J. Jensen. "Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes." Nucleic Acids Research 43.Database issue(2015):1140-1144.
[6] Luo, Feng, et al. "Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory." BMC Bioinformatics 8.1(2007):299-300.
[7] Sharan, R, I. Ulitsky, and R. Shamir. "Network-based prediction of protein function." Molecular Systems Biology 3.1(2014)
[8] Wren, J. D. "A global meta-analysis of microarray expression data to predict unknown gene functions and estimate the literature-data divide." Bioinformatics 25.13(2009):1694-1701.
[9] LANGFELDER P, HORVATH S. "WGCNA: an R package for weighted correlation network analysis. " Bmc Bioinformatics(2008) 9(1): 559.
[10] Szklarczyk, Damian, et al. "STRING v10: protein–protein interaction networks, integrated over the tree of life." Nucleic Acids Research 43. Database issue(2015):D447.
[11] Thomas, P. D. "The Gene Ontology and the Meaning of Biological Function." Methods in Molecular Biology 1446(2017):15.