Biochemical and Biophysical Research Communications 448 (2014) 473–479

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

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

Characterization of essential genes by topological properties in the perturbation sensitivity network

Lei Yang [a], Jizhe Wang [a], Huiping Wang [a], Yingli Lv [a], Yongchun Zuo [b,*], Wei Jiang [a,*]

[a] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, PR China
[b] The National Research Center for Animal Transgenic Biotechnology, Inner Mongolia University, Hohhot 010021, PR China

ARTICLE INFO

Article history:
Received 11 April 2014
Available online 4 May 2014

Keywords:
Yeast
Essential gene
Statistical test
Biological network

ABSTRACT

Genes that are indispensable for survival are called essential genes. In recent years, the analysis of essential genes has become extremely important for understanding the way a cell functions. With the advent of large-scale gene expression profiling technologies, it is now possible to profile transcriptional changes in the entire genome of *Saccharomyces cerevisiae*. Notwithstanding the accumulation of gene expression profiling in recent years, only a few studies have used these data to construct the network for *S. cerevisiae*. In this paper, based on the transcriptional profiling of the *S. cerevisiae* genome in hundreds of different gene disruptions, the perturbation sensitivity (PS) network is constructed. A scale-free topology with node degree following a power-law distribution is shown in the PS network. Twelve topological properties are used to investigate the characteristics of essential and non-essential genes in the PS network. Most of the properties are found to be statistically discriminative between essential and non-essential genes. In addition, the *F*-score is used to estimate the essentiality of each property, and the core number demonstrates the highest *F*-score among all properties.

1. Introduction

Essential genes are genes that are indispensable to support the survival of an organism [1]. These genes encode foundational functions required for a living cell under certain conditions and constitute a minimal gene set required for a living cell [2]. The deletion of only one of the essential genes is sufficient to result in infertility or lethality. Essential genes are involved in most survival-related housekeeping functions and tend to encode more hubs in the protein–protein interaction (PPI) network [3]. Some research results suggest that essential genes are correlated with disease genes [4], and essential genes of bacteria are attractive drug targets for new antibiotics. Therefore, the analysis of essential genes has been a major focus in genomic research and in drug design.

The behavior of a cell is a consequence of the complex interactions among its numerous constituents that constitute a series of complicated networks. The use of these networks to study biological systems has attracted emphasis among biologists in the last decade. However, most of the networks are too complex to be easily understood. Using graph theoretic concepts to investigate the topological properties of the networks, this problem can be overcome. The topological analysis in the networks has also become a useful tool for studying the social networks in social sciences [5], the characterization of drug-targets [6–8], the human disease genes [9,10], toxin-targets [11], and so on. With the advent of whole-genome expression profiling technologies, such as DNA microarrays, the transcriptional activity of thousands of genes in different biological conditions are simultaneously measured. Microarray experiments comparing expression levels of all genes in *Saccharomyces cerevisiae* for hundreds of mutants allow us to observe not only phenotypic changes of genes but also the up- or down-regulated genes in response to gene disruptions [12]. However, until recently, only a few studies have used these data to construct the network for *S. cerevisiae* [13,14].

In this study, the PS network is constructed from a transcriptional profiling study in response to 276 different gene disruptions in the yeast genome [12–14]. Nodes represent genes, and links are made between nodes if the expressions of the target genes are significantly altered by the disruption of the source gene. Twelve topological properties are calculated for each node in the PS network. Significant differences are found between the topological properties of essential genes and those properties of non-essential genes. We also employ the *F*-score to study the essentiality of each

* Corresponding authors. Fax: +86 471 5227683 (Y. Zuo). Fax: +86 451 8661 9617 (W. Jiang).
E-mail addresses: yczuo@imu.edu.cn (Y. Zuo), jiangwei@hrbmu.edu.cn (W. Jiang).

http://dx.doi.org/10.1016/j.bbrc.2014.04.136
0006-291X/

property and provide suggestions for a novel essential index for further study. The workflow of our study is shown in Fig. 1.

2. Materials and methods

2.1. Construction of the PS network

To construct a PS network from the genome-wide transcriptional profiling study for all genes in *S. cerevisiae*, 300 perturbation experiments, such as gene deletions or drug treatments, were downloaded from the supplemental data of Hughes et al. [12]. Among 300 perturbation experiments, 276 experiments were deletion mutants. Because drug treatments perturb more than one gene and could increase heterogeneity in the experiments, thus, drug treatments were not investigated in our study. We construct the PS network only based on the deletion mutants. In the work of Hughes et al., each pair of gene and deletion mutants was assigned a *P*-value according to the 'error model', correcting for gene measurement error and for biological noise [12]. A link is made between gene *i* and deletion mutant *j* if the expression of gene *i* is significantly changed ($P < 0.05$) in deletion mutant *j*. In our PS network, these links are represented as edges, and nodes are significantly regulated genes in deletion mutants or in the deleted genes. The *S. cerevisiae* PS network has 5638 genes as nodes, which are connected by 36841 edges.

2.2. Datasets of essential and non-essential genes

The lists of essential and non-essential genes of *S. cerevisiae* were obtained from the MIPS database [15] on February 20, 2014. The open reading frame (ORF) of an essential gene was regarded as an essential gene. In total, 949 essential genes and 4505 non-essential genes with unique ORF symbols were collected from the MIPS database. Among these genes, 849 essential genes were mapped into the PS network, and 4059 non-essential genes were mapped into the PS network. In total, 849 non-essential genes were randomly selected from 4059 non-essential genes. This dataset was defined as the control dataset.

Random sampling was also performed. In total, 849 non-essential genes were randomly selected from the 4059 non-essential genes. This procedure was repeated 1000 times to generate 1000 different balanced datasets containing different sets of non-essential genes.

2.3. Topological properties

In this study, the NetworkAnalyzer software [16] is used to calculate the degree, clustering coefficient [17], topological coefficient [18], average shortest path (ASP) and closeness centrality in the constructed network. The core number [19,20] is calculated using the MatlabBGL package, which is implemented in Matlab R2008a software. The Java plug-in cytoHubba [21] is used to explore important nodes or hubs in the network. This plug-in calculates four topological properties: betweenness, edge percolation component (EPC), maximum neighborhood component (MNC) and maximal clique centrality (MCC). The average distance to essential genes (ADEG) is defined as the average shortest distance between a gene and all essential genes in the PS network.

In this study, the common function index (CFI) [22] is defined to measure the amount of common Gene Ontology annotations of adjacent nodes in the PS network. The Gene Ontology annotations for *S. cerevisiae* were retrieved from the *Saccharomyces* Genome Database [23] on February 20, 2014. The CFI is defined as follows:

$$\mathrm{CFI}(i) = \sum_j d_{ij}^B + d_{ij}^c + d_{ij}^M$$

where *j* represents any node adjacent to node *i*. d_{ij}^B, d_{ij}^C and d_{ij}^M are the deepest ontology depth of common Gene Ontology annotations

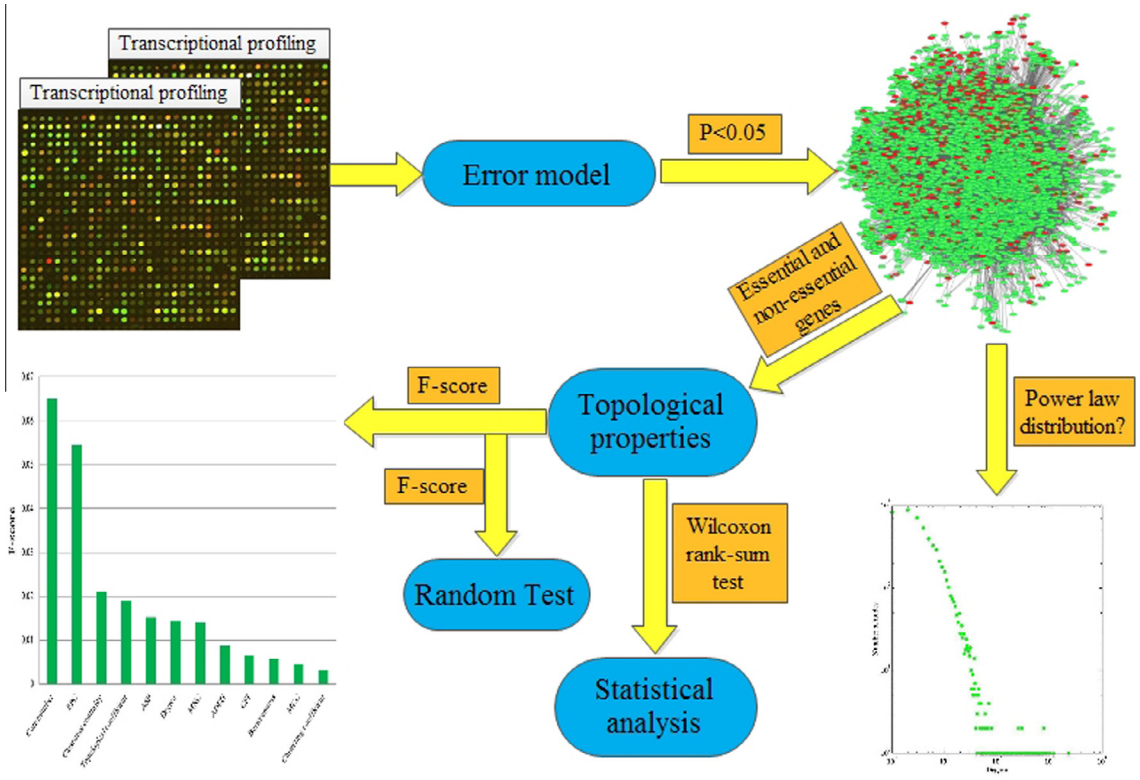

Fig. 1. The workflow of our study.