JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 28, Number 1, 2021

Pp. 1–18
DOI: 10.1089/cmb.2019.0120

Research Articles

Predicting Protein Functions Based on Differential Co-expression and Neighborhood Analysis

JAEL SANYANDA WEKESA,[1,2] YUSHI LUAN,[3] and JUN MENG[1]

ABSTRACT

Proteins are polypeptides essential in biological processes. Protein physical interactions are complemented by other types of functional relationship data including genetic interactions, knowledge about co-expression, and evolutionary pathways. Existing algorithms integrate protein interaction and gene expression data to retrieve context-specific subnetworks composed of genes/proteins with known and unknown functions. However, most protein function prediction algorithms fail to exploit diverse intrinsic information in feature and label spaces. We develop a novel integrative method based on differential Co-expression analysis and Neighbor-voting algorithm for Protein Function Prediction, namely CNPFP. The method integrates heterogeneous data and exploits intrinsic and latent linkages via global iterative approach and genomic features. CNPFP performs three tasks: clustering, differential co-expression analysis, and predicts protein functions. Our aim is to identify yeast cell cycle-specific proteins linked to differentially expressed proteins in the protein–protein interaction network. To capture intrinsic information, CNPFP selects the most relevant feature subset based on global iterative neighbor-voting algorithm. We identify eight condition-specific modules. The most relevant subnetwork has 87 genes highly enriched with cyclin-dependent kinases, a protein kinase relevant for cell cycle regulation. We present comprehensive annotations for 3538 *Saccharomyces cerevisiae* proteins. Our method achieves an AUROC of 0.9862, accuracy of 0.9710, and *F*-score of 0.9691. From the results, we can summarize that exploiting intrinsic nature of protein relationships improves the quality of function prediction. Thus, the proposed method is useful in functional genomics studies.

Keywords: differential co-expression, function prediction, gene expression profile, protein–protein interaction.

[1]School of Computer Science and Technology, Dalian University of Technology, Dalian, China.
[2]School of Computing and Information Technology, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya.
[3]School of Life Science and Biotechnology, Dalian University of Technology, Dalian, China.

1. INTRODUCTION

PROTEINS ARE THE MOST VERSATILE MACROMOLECULES in living systems. Knowledge of protein functions is crucial in downstream analysis and applications such as disease analysis and drug development (Yu et al., 2017). However, accurate and comprehensive assignment of biological functions to proteins is a challenging task in functional genomics. This is due to availability of many labels for each protein and because biological knowledge rapidly evolves (Yu et al., 2015). Although wet-lab techniques including gene knockout are capable of determining annotations, they are expensive and time-consuming. Therefore, the quest to unravel annotations for uncharacterized genes or proteins using computational models is an increasingly important task for bioinformatics researchers in the post-genome era.

The huge influx of information generated by advanced sequencing technologies has led to label deficiency problem. This has invigorated interest in computational models that can identify hypothetical functions of proteins. Several bioinformatics tools such as Panther (Mi et al., 2013) and InterProScan (Jones et al., 2014) exploit semantic similarity between proteins, protein functional interactions, pathway enrichment analysis, and phylogenetic tree (Pesquita et al., 2009; Zhou et al., 2014) for the discovery of novel functional annotations. To address the problem of low reliability and limited coverage, comparative analysis is of vital significance (Gligorijević and Pržulj, 2015). It is based on the overlap of information among data sets for synergistic predictions. For instance, in eukaryotic organisms, gene expression data such as cell cycle and protein–protein interaction (PPI) data can be analyzed. The analysis results can be used to investigate biological systems across diverse conditions for assigning protein functions. The cell cycle is a repetitive process of cell growth and sequential duplication in eukaryotic organisms. The cycle is stimulated by environmental factors through different phases.

The four main cell cycle phases are G1, DNA synthesis (S), G2, and mitosis (M). S and M are the main phases separated by gap phases G1 and G2. Cyclin-dependent protein kinases (Cdks) are regulatory proteins, which signal the cell to transition to the next stage of the cell cycle depending on cyclins. For example, cdc28 is a Cdk complex implicated in cell cycle control in yeast; it promotes transition between the different cell cycle phases (Bertoli et al., 2013).

Researchers have proposed several computational methods for protein function prediction. For multilabel classification based on functional hierarchy, clusDCA (Wang et al., 2015) was proposed. The method predicted protein functions by integrating protein networks and functional hierarchy using PageRank and low-rank matrix approximation. PILL (Yu et al., 2015) combines hierarchical and flat taxonomy similarity between function labels to replenish missing labels and predict functions of completely unlabeled proteins. AptRank (Jiang et al., 2017) and Bi-TMF (Meng et al., 2016) predicted protein functions based on functional interrelationships. They used a birelational graph to propagate from annotated to unannotated proteins. Recently, OGN (Zhang et al., 2018), a centrality measure-based method, has been proposed. It integrates orthologous information, gene expression, and protein interaction data to predict essential proteins. Similarity ensemble approach (O'Meara et al., 2016) predicted gene functions based on functional genomic networks and ligand similarity networks.

Gene expression profiling has vastly been used in cancer to classify tumors and in the discovery of pathway alterations across phenotypes (Ma et al., 2011). To exploit functional and other gene–gene correlation characteristics, ''guilt-by-profiling'' is often employed for functional inference. The past decade has seen an extensive proposal of algorithms for the construction of gene co-expression networks. They are based on the similarity between gene expression profiles and are mainly applied for functional annotation. Weighted gene co-expression network analysis (WGCNA) is a widely used network analysis method (Horvath and Langfelder, 2008). It constructs co-expression networks to determine potential biomarkers, functional module prediction, and discovery of important elements of disease-related genes (Gibbs et al., 2013). WGCNA has been applied for function annotation tasks, such as in rice genes (Childs et al., 2011), in the construction of gene co-expression networks (Langfelder et al., 2008; Kadarmideen et al., 2011) and for differential analysis (Liu et al., 2017). WGCNA has also been implemented to analyze transcription modules associated with tumor in colon cancer (Liu et al., 2017).

Biological systems are highly dynamic depending on the environment, tissue type, disease, or development (Ideker and Krogan, 2012). Various studies have applied differential co-expression analysis and differential co-expression networks to identify modules associated with specific environmental stress and response to genetic changes (Lai et al., 2004; Watson, 2006; Hu et al., 2015). Unlike studies that do

comparative analysis between related organisms (Ihmels et al., 2005), in this study, we focus on one organism. We employ WGCNA to identify specific modules and hub genes related to yeast cell cycle. Additionally, the widely used neighborhood- and module-based protein function prediction methods have the limitation of high dimensionality of the search space for densely connected neighborhood regions (Jiang et al., 2017). To counter these problems, exploring and embedding other features of a protein in the interaction network improve the accuracy level. For instance, discovering functional associations among proteins in a bottom-up level-to-level approach (Prasad et al., 2017).

In this study, we set up a novel integrative network-based approach that predicts protein functions from differential co-expression analysis and neighbor-voting (NV) algorithm. Our approach constructs two networks from yeast cell cycle gene expression and PPI data sets (Zhang and Horvath, 2005). Biweight midcorrelation (BWM) is used to identify highly connected genes and biologically relevant modules. BWM performs better in comparison to two other correlation methods, Pearson's correlation coefficient and Spearman's rank correlation coefficient, which have been commonly used for the construction of weighted co-expression networks (Ma and Wang, 2012). Measures for scrutinizing our method include: number of clusters identified and relationships between genes and biological significance of cluster membership.

Our contributions are mainly threefold. First, cluster analysis by integrating gene expression profile and PPI data sets using ''guilt-by-profiling'' technique. Second, we exploit module differential co-expression based on an adaptive parameter tuning mechanism and enrichment analysis. We identify modules with proteins of specific environmental stress conditions and their functional significance. Finally, exploiting label correlation, intrinsic information from co-expression analysis, genomic features such as transcription factor binding and global iterative approach, we assign novel annotations to proteins. We quantify the interrelationship using semantic similarity. Our method achieves an AUROC of 0.9862, accuracy of 0.9710, and *F*-score of 0.9691. From the results, we can summarize that exploiting intrinsic nature of protein relationships improves the quality of function prediction. Thus, the proposed method is useful in functional genomics studies.

2. METHODOLOGY

2.1. Data sets

Eukaryotic organism *Saccharomyces cerevisiae* (Baker's yeast) species are well studied, and proteins are characterized by knockout experiments. Therefore, most network-based prediction methods have used its data sets for testing performance due to its reliability among various species. We use yeast cell cycle gene expression data set of Zhang and Horvath (2005), which includes 44 samples under various times during the cell cycle and a total of 4000 genes. After preprocessing and filtering, 1264 genes are retained for analysis. Steps for pre-processing the data included (1) filter out genes with missing values, (2) impute the missing values, (3) standardize the data, mean is 0 and the standard deviation is 1. The PPI data are a 2292×2292 matrix of pairs of interacting proteins downloaded from Zhang and Horvath (2005). Of 2292 proteins, 2274 are retained after omission of missing data. The yeast protein annotation data set contains 3469 Gene Ontology (GO) functions belonging to biological process (BP), molecular function (MF), and cellular component terms for 5775 proteins (28.03.13 release) from BioGRID (Stark et al., 2006). We also use an Affymetrix Genechip platform (org.SC.sgd.db Version 3.5.0) for annotation. Analysis of our data sets is performed in R statistical programming environment version 3.4.2, R studio version 1.1.383, and Bioconductor programs.

2.2. Network characteristics for training Co-expression analysis and Neighbor-voting algorithm for Protein Function Prediction

Network construction is essential for the identification of modules and defining intramodular connectivity. Corresponding genes (nodes) that are significantly co-expressed are connected. Furthermore, genes with expression levels that are highly correlated (hubs) participate in similar BPs and tend to encode essential genes (Lai et al., 2012). WGCNA provides a function named *pickSoft-threshold*, which automatically selects threshold value β to determine the number of modules for differential co-expression analysis.

In this study, we explore topological features and expression profiling to solve protein function prediction problem. Existing PPI network has a huge number of false positives, which reduce reliability of the

interactions. To build a more reliable protein function prediction network, we establish a weighted PPI network and perform functional diffusion. Furthermore, we apply iterative procedure to reduce false positives.

2.3. Biweight midcorrelation

The BWM, a robust form of correlation, elucidates pairwise bivariate relationship (Zheng et al., 2014). It generates significantly enriched co-expression modules with coherent expression profiles using topological overlap matrix as dissimilarity. Although Pearson's correlation has been widely preferred for cluster analysis due to its ability to derive information on expression levels such as global linear relationships, BWM is more robust to outliers (Zheng et al., 2014). To define the BWM (bicor) of two numeric vectors $x=(x_1, x_2, \ldots, x_m)$ and $y=(y_1, y_2, \ldots, y_m)$ according to Zheng et al. (2014), we first define u_i, and v_i (where $i=1, 2, \ldots, m$) as follows:

$$u_i = \frac{x_i - \text{med}(x)}{9\text{mad}(x)}, \tag{1}$$

$$v_i = \frac{y_i - \text{med}(y)}{9\text{mad}(y)}, \tag{2}$$

where med(x) is the median and mad(x) is the median absolute deviation. This leads to the definition of weight w_i for x_i as follows:

$$w_i^{(x)} = \left(1-u_i^2\right)^2 l(1-|u_i|), \tag{3}$$

where $l(1-|u_i|)$ takes 1 if $1-|u_i|>0$ and 0 otherwise. Given the weights, we can define BWM of x and y as:

$$\text{bicor}(x, y) = \frac{\sum_{i=1}^{m}(x_i \text{med}(x))w_i^{(x)}(y_i - \text{med}(y))w_i^{(y)}}{\sqrt{\sum_{j=1}^{m}\left[\left(x_j - \text{med}(x)\right)w_j^{(x)}\right]^2}\sqrt{\sum_{k=1}^{m}\left[(y_k - \text{med}(y))w_k^{(y)}\right]^2}}. \tag{4}$$

3. PROPOSED METHOD: CO-EXPRESSION ANALYSIS AND NEIGHBOR-VOTING ALGORITHM FOR PROTEIN FUNCTION PREDICTION

We propose a novel integrative method based on differential Co-expression analysis and Neighbor-voting algorithm for Protein Function Prediction called CNPFP. In this algorithm, co-expression networks are constructed for functional classification and analysis of yeast cell cycle-associated proteins, as shown in Figure 1.

Feature selection through genomic features and global iterative similarity computation is done to improve the prediction performance of NV algorithm and reduce false positives. CNPFP measures semantic similarity of GO annotation terms T. Protein p with a function F can be assigned a function F' if F and F' are semantically similar. F' is an existing function predicted and assigned to protein p based on semantic similarity. The sequential steps taken by CNPFP are presented in Figure 1.

3.1. Construction of co-expression network

We constructed two co-expression networks from PPI and yeast cell cycle gene expression data sets. The networks are based on weighted connectivity of correlated gene expression following the method in Zhang and Horvath (2005). Preprocessing operations including variation filtering were done to remove systematic variation between microarray experiments to bring upregulated and downregulated genes to the same scale. The network was constructed from yeast normalized $\log_2$-transformed matrix of genes. To construct the gene co-expression networks, we calculate the correlation matrix using BWM.

Clusters of genes with similar connection strengths are identified to determine network connectivity. To demonstrate that our co-expression networks manifest complex network properties (power-law degree distribution), we explore topological characteristics of the network such as degree distribution. An input $m \times n$ gene expression matrix is denoted as $X=(x_{il})$. Where column indices ($l=1, 2, \ldots, n$) correspond to

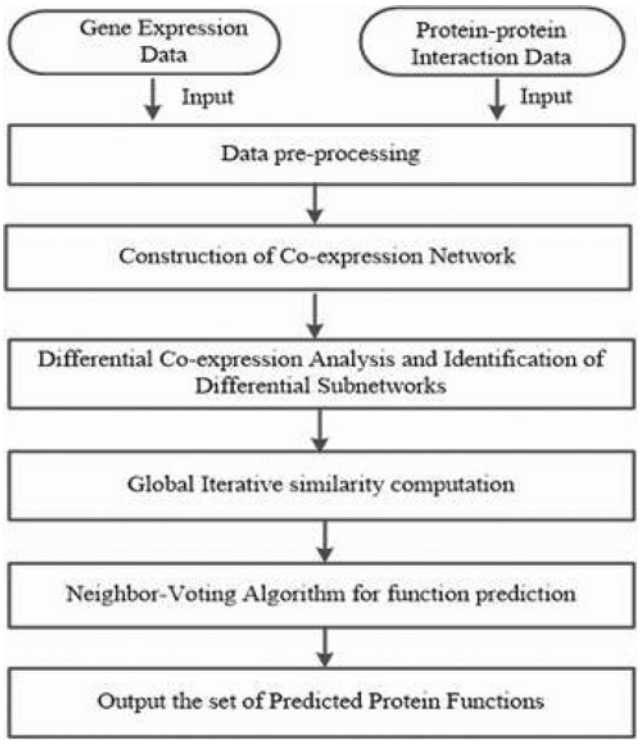

FIG. 1. Flowchart of the proposed method.

samples under specific conditions and the row indices ($i = 1, 2, \ldots, m$) correspond to genes. The size of the network in terms of the number of modules is directly impacted by changing the soft-threshold value β. The β value effectively adjusts how smoothly the connection strengths transition from their lowest to their highest values. Inspired by CEMiTool (Russo et al., 2018), we choose a lower threshold for R_2 ($R_2 \geq 0.77$) compared with WGCNA default threshold ($R_2 \geq 0.85$). This results into a lower value of $\beta = 5$ (the R_2 reached the peak for the first time when $\beta = 5$) in the interval (1, 10) in 1 increments and (12, 20) in 2 increments. β is selected based on scale-free topology criterion maximized with a ($R_2 \geq 0.77$) fit while maintaining high mean connectivity where R_2 is the linear regression model fitting index between log transformation of $p(k)$ and (k) with k as the measure of connectivity.

We use topological overlap dissimilarity measure to determine the number of modules, which are distinguished using different colors. A hierarchical clustering algorithm is then implemented to identify modules of densely interconnected genes to form a dendrogram. We obtain 9 modules for gene expression network and 20 modules for PPI network (Fig. 2). For each network, gray module is reserved for genes that are not co-expressed among each other, which makes up 8 and 19 distinct modules, respectively. We use BWM to implement hierarchical clustering to cluster genes coupled with topological overlap dissimilarity measure between gene expression data vectors. The topology overlap measure w_{ij} between two nodes i and j is calculated as follows:

$$w_{ij} = \frac{l_{ij} + a_{ij}}{\min\{k_i, k_j\} + 1 - a_{ij}}, \tag{5}$$

where $a_{ij} = (1 + cor(y_i, y_j)/2)^{\beta}$ is the adjacency measure between gene i and gene j. $cor(y_i, y_j)$ is the correlation between gene expression profiles y_i and y_j. y_i is a vector of the i-th gene expression profile (where $i = 1, 2, \ldots, n$). y_j is a vector of the j-th gene expression profile (where $j = 1, 2, \ldots, n$). $l_{ij} = \sum_{u \neq i,j} a_{iu} a_{uj}$, a_{iu} and a_{uj} represent the number of nodes i and j are connected to; $k_i = \sum_{u \neq i} a_{iu}$ represents the number of connections of a node; $a_{ij} \in$ provides a measure of connectivity between a pair of genes viewed as network nodes where 1 indicates a connection, whereas 0 indicates no connection. The power parameter 1 in a_{ij} is the threshold that is specified depending on whether it is a signed or unsigned network. A hybrid signed network is preferred because it produces biologically meaningful modules. To ensure cluster stability, we compare the different WGCNA clustering methods including dissimilarity based

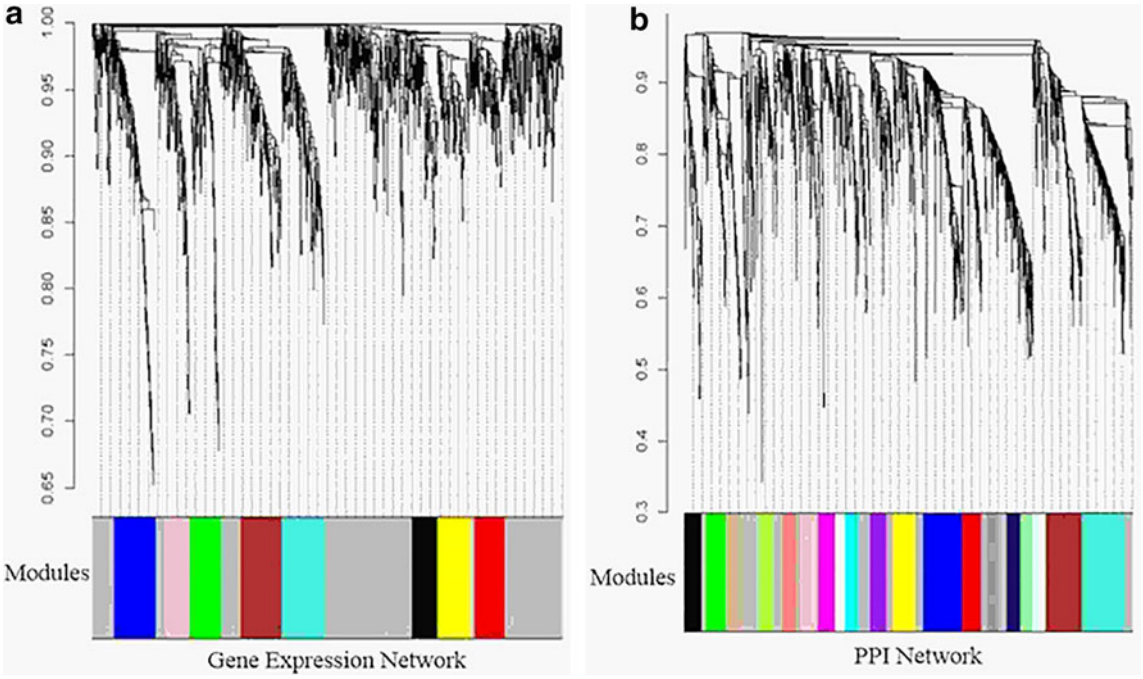

FIG. 2. Dendrogram and module colors. **(a)** Network modules generated from gene expression data. **(b)** Network modules from PPI data. PPI, protein–protein interaction.

on adjacency and topology overlap matrix (TOM). We choose to use Topological overlap dissimilarity measure to define modules since it generates more cohesive modules, which are larger and more robust (Cheng et al., 2013).

$$\text{dissTOM}_{ij} = 1 - \left(\sum_{u \neq i} a_{iu} a_{uj} + a_{ij} \right) / \left(\min(k_i, k_j) + 1 - a_{ij} \right), \tag{6}$$

where $k_i = \sum_{u \neq i} a_{ui}$ and $k_j = \sum_{u \neq j} a_{uj}$ denotes the network connectivity. High robust network interconnectedness is arrived at by TOM combining connection strength between a pair of genes with their connections to other genes. For each module, module significance is the average absolute gene significance for all the genes in the module. We plot cluster coefficient against connectivity to illustrate cliquishness for cluster analysis of the two networks (Fig. 3a, b).

3.2. Differential co-expression with BWM

Key gene expression regulators, including transcription factors and microRNA, cause gene pairs to exhibit robustness in perturbation and co-expression across conditions. Yeast relies on specific internal conditions for optimal growth. However, external environmental stress disrupts the normal processes. Yeast cell cycle data in this article are from different time courses. The samples are from three experimental conditions: long-term exposure to alpha factor, elutriation (elu), and temperature sensitivity mutant cell division cycle (cdc). Given a pair of genes i, j, we calculate BWM coefficients between expression levels represented as $\rho 1$ and $\rho 2$. $\Delta ij = \rho 1 - \rho 2$ is the increase or decrease in correlation between two groups. For both data sets, we select genes with a density threshold of $\Delta ij > 0.7$ or $\Delta ij < -0.7$. Genes within this range exhibit highly reliable differential correlation.

Classical gene selection techniques, t-test, and F-score require a distributional assumption about data and rely on a parameter such as λ. λ is a scoring function used to measure the condition-specific changes of genes and gene–gene co-expression. Our method on the other hand is a gene clustering method with statistical differential expression test based on an adaptive threshold. The threshold is a user-defined value used to generate statistically overrepresented patterns of gene expression profiles among multiple conditions.

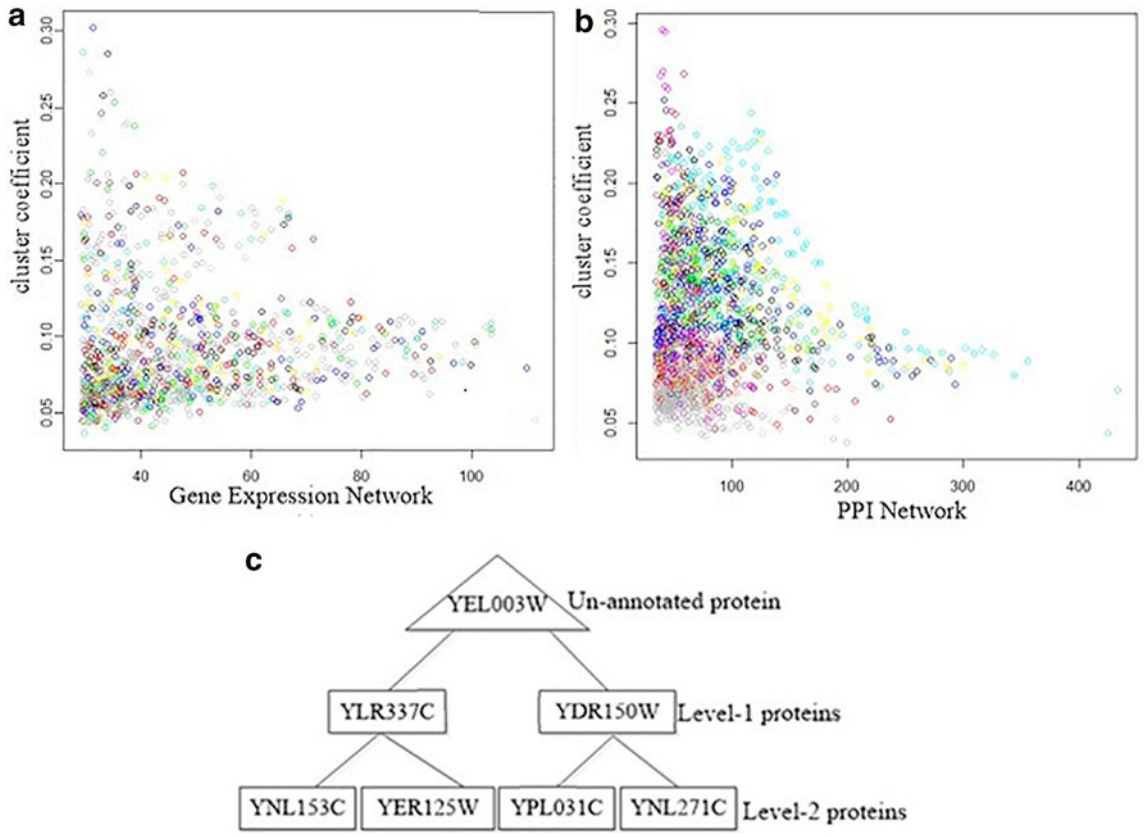

FIG. 3. Network connectivity to illustrate cliquishness for cluster analysis of the gene expression and PPI networks and protein interaction network. **(a)** The plot of cluster coefficient against connectivity of gene expression network. **(b)** The plot of cluster coefficient against connectivity of PPI network. **(c)** Sample protein interaction network of protein YEL003W.

3.3. CNPFP algorithm

We implement NV algorithm (Ballouz et al., 2016) and iterative inference for prediction of annotations. Pseudocode for the algorithm is summarized in Algorithm 1. Given a set of hidden gene labels, we determine if the remaining genes in the annotation set can predict the annotations for the hidden genes. Each annotation receives its confidence score by receiving neighbors' votes; thus, annotations are determined by counting neighbors' votes. Each gene is scored as a fraction of its number of connections with function-associated genes and the total number of connections of that gene in the network.

For each protein u, each function x is given a score based on the frequency of its occurrence in the neighbors of u.

$$f_x(u) = \sum_{v \in N_u} \delta(v, x), \tag{7}$$

$(v,x) = 1$ if v has function x, 0 otherwise. Where N_u refers to the interaction neighbors of protein u. The function k with the largest score $f_k(u)$ is predicted for protein u. Multiple annotations are assigned to u by sorting the functions associated with the neighbors of u based on decreasing $f_x(u)$ and their rank.

Feature selection involves the selection of a subset of features that are relevant for predicting target variables (Navot et al., 2006). For instance, a new binary feature selection method has been proposed by Guan et al. (2017) for predicting extracellular matrix proteins. We implement the global iterative approach, a multilabel learning method that maximizes dependence between functional similarities. It captures intrinsic information of input data for label prediction. We select the most relevant feature subset for prediction based on an iterative process of determining an alignment score in the neighborhood of a protein. The score is based on topological importance of nodes and edges starting from one and stopping at R. The

range for the minimum and maximum number of neighbors is set between 50 (minimum) and 950 (maximum). Following the study of Hashemifar and Xu (2014), we set cutoff score (ε) used to select alignments to 0.2 as it yields biologically meaningful alignment.

The idea is to iteratively count the contribution of functions of the neighbors of a protein to determine the final predicted functions. The contribution of a function to the prediction depends on the number of neighbors and the similarities between their functions. The details of the iterative process are given as follows.

Let $N(p)$ be the set of neighbors of protein p. Where the neighbor proteins of p are those with direct and/or indirect interactions with p in the PPI network. In this study, both level-1 and level-2 neighbor proteins of unannotated protein p are considered (Fig. 3c). Level-1 neighbor proteins are those directly interacting with unannotated protein p. Level-2 are proteins that directly interact with level-1 proteins of p but not directly interacting with p. Suppose F is the set of all functions in the PPI network, we denote F as $F = \{f^1, f^2, \ldots, f^k\}$ where f^i $(i = 1, 2, \ldots, k)$ are the functions in the PPI network. k is the number of functions. We denote the set of functions of two proteins p and p' as $F(p)$ with size m and $F(p')$ with size n. The similarity between p and p' is defined as:

$$sim(p, p') = \frac{1}{\max(m, n)} \sum_{f \in F(p)} \sum_{f' \in F(p')} \delta_{f,f'}, \tag{8}$$

where $\delta_{f,f'}$ is an indicator function, such that if f and f' are the same, its value is 1, otherwise, 0. Given any two functions f and f', they can be represented as two vectors $\vec{f}$ and $\overrightarrow{f'}$. The element values of the two vectors indicate the occurrence of the GO terms that annotate the functions. If the number of GO terms is t, the dimension of each function vector $\vec{f}$ is t. GO is a directed acyclic graph, therefore, each GO term may have multiple parent GO terms also known as ancestors. Thus, a function is annotated not only by a GO term but also by the ancestors of the term. Therefore, the vector element values that correspond to the ancestors are set to 1 otherwise set to 0. For instance, given five GO terms for annotating a protein, a function f is annotated by the fourth term and its parent terms, the second and third terms. Another function f' is annotated by the fifth term and its parent terms, the third and fourth terms. The two functions f and f' can be represented as two vectors $\vec{f} = (0, 1, 1, 1, 0)$ and $\overrightarrow{f'} = (0, 0, 1, 1, 1)$, respectively. The similarity between functions, $fsim(f, f')$ is defined as:

$$fsim(f, f') = \vec{f} \,.\, \overrightarrow{f'} \;/\; \left\|\vec{f}\right\| \,.\, \left\|\vec{f'}\right\|, \tag{9}$$

where $\vec{f} \,.\, \overrightarrow{f'}$ is the dot product of two vectors and $\left\|\vec{f}\right\|$ is the norm of the vector $\vec{f}$. From the above definition, the similarity between the two functions is within the range $0 \le fsim\,(f, f') \le 1$. For the function vectors $\vec{f} = (0, 1, 1, 1, 0)$ and $\overrightarrow{f'} = (0, 0, 1, 1, 1)$ for instance, $\vec{f} \,.\, \overrightarrow{f'} = 2$, $\left\|\vec{f}\right\| = \left\|\vec{f'}\right\| = \sqrt{3}$, and the similarities between the two functions is $fsim(f, f') = 2/3$. From this example of function similarities, the score of an unannotated protein p annotated by function $f \in FN(p)$, that is, the contribution of function f to the final prediction results, is defined as:

$$score(p, f) = \sum_{p' \in N(p)} \left[sim(p, p') \times \left(\sum_{f' \in F(p')} fsim(f, f') \times \log \frac{N}{n_{f'}} \right) \right], \tag{10}$$

where $fsim(f, f')$ is the local influence of functions in the local domain $N(p)$, $log\frac{N}{n_{f'}}$ shows the global influence of available functions on the prediction results. For each function $f \in FN(p)$, its initial score is:

$$score^{(0)}(p, f) = \sum_{p' \in N(p)} \sum_{p' \in N(p')} \left[fsim(f, f') \times log \frac{N}{n_{f'}} \right]. \tag{11}$$

We set the neighborhood score threshold of initial function selection using the following formula:

$$\varepsilon = \frac{1}{size(FN(p))} \sum_{f \in FN(p)} score^{(0)}(p, f), \tag{12}$$

where $size(FN(p))$ is the number of functions in the set $FN(p)$. The function with the score calculated in Equation (11) over a threshold in Equation (12) is selected as initial functions of unannotated protein p.

We denote the functions of protein p as a set $F(p^k)$ and functions of neighbor proteins as $FN(p) = \cup_{pk \in N(p)} F(p^k)$. For an unannotated protein p_x, its predicted functions are recorded as a vector $F_{tp_x} = \left[f_{x,t}^1, f_{x,t}^2, \ldots, f_{x,t}^K\right]^T$, $f_{x,t}^j = 1$ $(j = 1, 2, \ldots, k)$ if the predicted functions of protein p_x from the t-th iteration contain function $f^j (f^j \in F)$, otherwise $f_{x,t}^j = 0$. F_{1p_x} denotes the vector of initial functions of p_x. After M-th iterations, the iteration will have reached the stable status and a matrix AF_{p_x} is formed for all predicted functions of p_x generated from all iteration rounds. The final predicted functions of unannotated protein p_x are selected based on the frequency of their occurrences in the whole iteration process calculated as follows:

$$AF_{p_x} = F_{1p_x}, F_{2p_x}, \ldots, F_{Mp_x}, \tag{13}$$

where AF_{p_x} is the matrix of all predicted functions of p_x generated from all rounds of the iteration.

Algorithm 1

Inputs: Unannotated protein p_x, interaction neighbors of p: $N(p)$, Protein function annotation matrix $F(p)$, Function annotations of $N(p)$: $FN(p)$, the preferred number of predicted functions: k, t, and M = number of iterations,

Procedure:

1: **for** $i = 1$ to t **do**
2: for each test protein p_x
3: Set cutoff range R for $N(p)$
4: Initialize $F(p)$ by Equations (11) and (12)
5: Calculate similarity between proteins [Eq. (8)]
6: Calculate similarities between functions in $FN(p)$
7: Calculate the function scores using Equation (10)
8: Record predicted functions in matrix $F(p)$ and $F'(p)$, before and after next iteration respectively
9: **end for**
10: $Rank(F(p))$: the set of ranked functions of $F(p)$
11: $Rank(F'(p))$: the set of ranked functions of $F'(p)$
12**: for** $i = 1$ to M **do**
13: **for** each test protein p_x
14: Initialize $F(p)$ by Equations (11) and (12)
15: Calculate similarities between proteins by Equation (8)
16: Calculate the function scores using Equation (10)
17: $F'(p)$ = selected k functions in $FN(p)$ with the highest k scores, record results in matrix $F'(p)$
18: **end for**
19: **for** each test protein **do**
20: preliminary classification by $F = [F_1, F_2, \ldots, F_k]^T$
21: If $F'(p) = F(p)$ and the function order in $Rank$ $F'(p)$ = function order in $Rank(F(p))$
22: Output preliminary results $F = [F_1, F_2, \ldots, F_k]^T$
23: else go back to step 4
24: **end for**
25: **return** $AF_{px} = [F_{1\ px}, F_{2\ px}, \ldots, F_{M\ px}]$

3.4. Semantic similarity

The dynamic nature and mutual interaction between pairwise proteins stimulate the need for dynamic function voting based on semantic functional similarity. We measure functional similarity by semantic similarity between GO terms of proteins within the modules. The functional similarity is calculated based on information content. First, we compare GO mappings of gene pairs. Given two proteins X and Y with M and N sets of GO terms, respectively, a similarity matrix S is calculated. Semantic scores S_{ij} using Schlicker's method (Schlicker et al., 2006) is defined as follows:

$$s_{ij} = \max_{c \in \left(GO_i^X,\ GO_j^Y\right)} \left[\frac{2\log p(c)}{\log p\left(GO_i^X + GO_j^Y\right)} \times (1 - p(c))\right], \tag{14}$$

where c is the set of common ancestors of the GO terms and $p(c)$ is the probability that c is equal to its frequency in the annotations. $1 - p(c)$ is used to give less importance to a frequently occurring term. Log_p is

the logarithm of the probability p, used to make the semantic similarity score a nonnegative value. The functional similarity of X and Y is calculated using best match average (BMA) (Ravasz et al., 2002) of matrix S as follows:

$$funsim_{BMA}(X, Y) = \frac{\sum_{i=1}^{m} \max_{1 \le j \le n} s_{ij} + \sum_{j=1}^{n} \max_{1 \le i \le m} s_{ij}}{m+n}, \tag{15}$$

where ''max'' calculates the maximum semantic similarity score over all pairs of terms in the similarity matrix $S = [s_{ij}]_{m \times n}$.

4. EXPERIMENTAL RESULTS

4.1. Co-expression analysis

Co-expression analysis identifies genes significantly associated with specific environmental stress response in yeast. WGCNA applied in this study is a module-assisted method that covers statistical and computational aspects based on ''guilt-by-profiling'' technique. It extracts co-expressed genes from large heterogeneous data sets (Clarke et al., 2013). The principle of ''guilt-by-profiling'' in co-expression analysis states that genes with expression relationship share biological functions. Thus, the correlation between gene expression levels quantitatively assesses gene co-expression and cluster genes of similar functions across a variety of experimental perturbations (Eisen et al., 1998). We analyze *S. cerevisiae* gene expression and PPI networks separately to identify modules of genes with highly correlated expression patterns, as shown in Figure 2. We use signed hybrid functions so that genes within modules are positively correlated.

We further identify modules with similar expression profiles using dynamic height branch cutting method. Dynamic tree cut method (Langfelder et al., 2008) finds clusters that are highly and significantly enriched with known GO terms resulting in biologically meaningful clustering results (Dong and Horvath, 2007). Modules with highly co-expressed genes are then merged to reduce the number of modules with constitutive genes and slightly expressed variations.

We used 1264 highly co-expressed probe sets (genes) of the yeast gene expression cell cycle data set to construct the gene co-expression network. The experiment was run 10 times, from which 8 distinct co-expression modules were identified. The number of identified modules on each experiment instance varied between 2 and 9. The module size range between 50 and 350 genes on each run with changing height cut value. Since yeast is one of the well-studied species, most clustered genes within seven modules were highly enriched with protein domain, biological pathway, transcription unit, or protein complex.

We calculate the mean connectivity for all genes as a function of soft threshold power β and a max-Poutlier of 0.1. Then, we calculate TOM from the transformed correlation adjacency matrix and convert it into a dissimilarity matrix. We then create a hierarchical cluster tree using a dynamic tree cut method. The coefficient of variation filter value is arrived at based on the number and quality of co-expression modules identified. Through co-expression analysis, we identify biologically relevant clusters, which may contain novel co-expression relationships.

4.2. Module connectivity and gene expression analysis

More modules were identified in the condition-independent data set (PPI), which pose a challenge in biological interpretation. Genes from condition-dependent modules are split into multiple modules in the PPI network modules. Correlation between genes is weakened because of the split of genes into many new gene modules. We found that a total of 677 genes were expressed in both data sets. In the analysis of modules, underexpressed (downregulated) module genes are represented with the color red and the over-expressed (upregulated) represented by green, as shown in Figure 4a. Through differential module analysis, we find that expression clusters have intersections with specific PPI clusters (overlap), as shown in Figure 4b. Note that the values represent the number of genes overlapping in the modules and false discovery rate (FDR)-adjusted p-values (e.g., turquoise is module number 1 in both networks, has 31 overlapping genes with FDR-adjusted p-value 0.24). The existence of overlapping modules is an indication of confident results from our clustering method. Related nondistinct modules in PPI data set are merged by considering consensus dissimilarity between eigengenes.

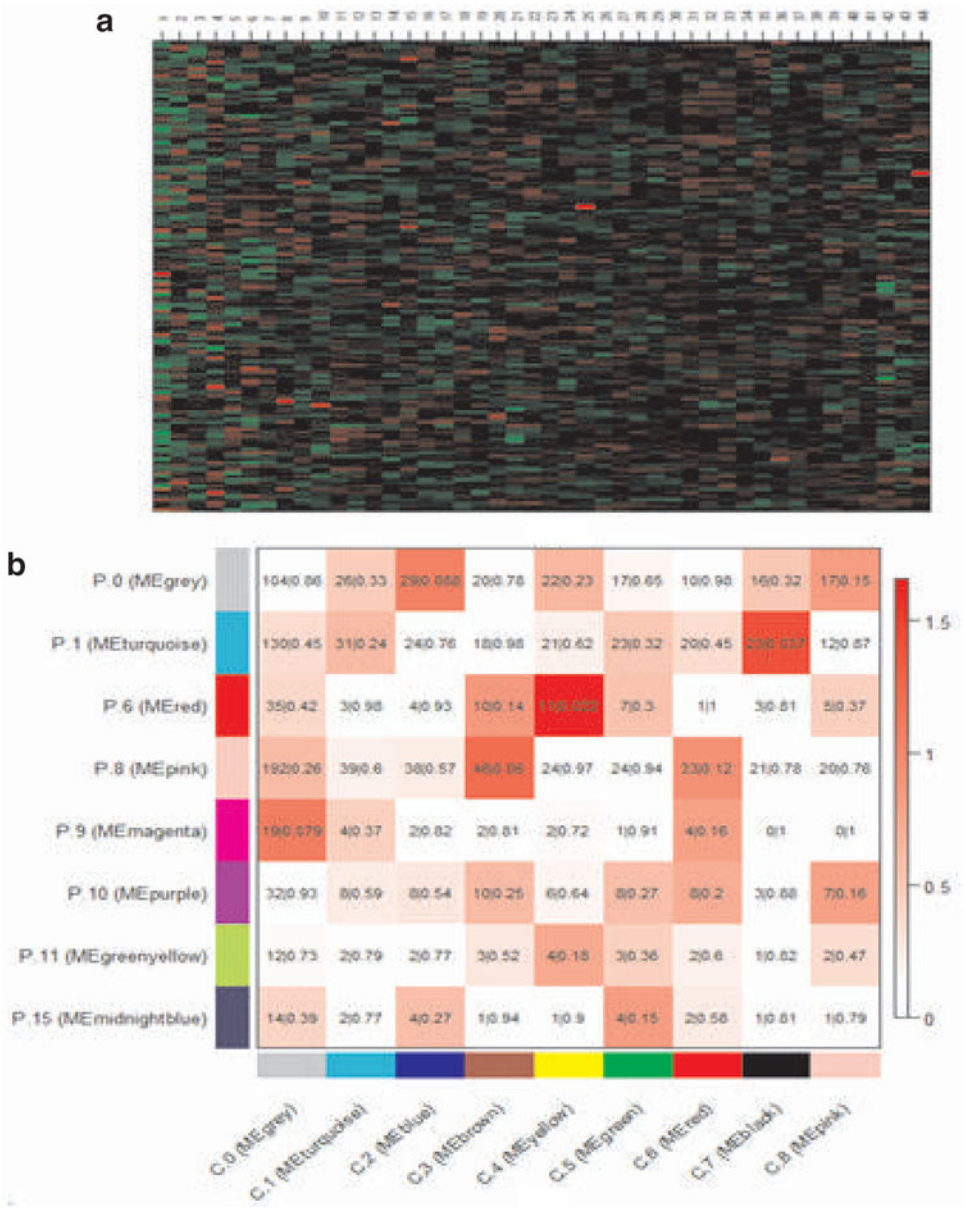

FIG. 4. Module gene expression analysis and overlap between modules. **(a)** Rows correspond to module genes and columns to samples (1–44) with colors red representing low expression and green highly expressed genes. **(b)** Overlap of PPI (*y*-axis) represented by P and yeast cell cycle (*x*-axis) represented by C and associated module numbers.

Fisher's exact test based on the hypergeometric distribution was used to assess memberships within PPI and cell cycle cluster overlap. $-\text{Log}_{10}(p)$ value score was used to define the degree of overlap between clusters, where p is the FDR-adjusted p-values. The heat map colors indicated in the color scale bar represent the $-\log_{10}(p)$-transformed p-values. Negative values indicating under-enrichment. The rows and columns correspond to PPI and gene expression network modules represented as P for PPI and C for gene expression.

4.3. Identification of condition-specific modules and hub genes

In each module, genes are highly co-expressed and co-regulated, hence resulting in module functional homogeneity. We identify gene sets that are differentially regulated with respect to different biological conditions. To accomplish this, we conduct module differential co-expression analysis based on an adaptive

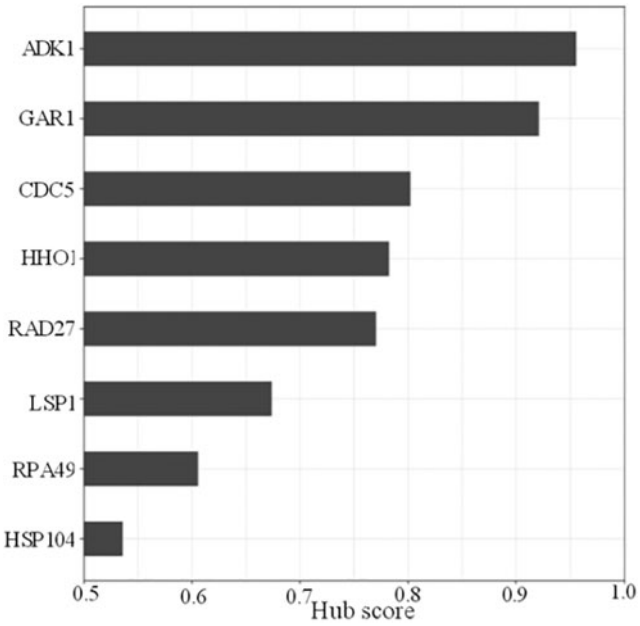

FIG. 5. Hub scores of most highly influential genes.

user-defined density threshold to mine dense subgraphs. The differentially expressed modules are enriched with a specific BP for both condition-dependent or condition-independent data sets.

Gene expression profiling unveils functional heterogeneity distinct in differential sensitivity to stress of genetically identical cells. The blue module of cell cycle data set has the highest coefficient of variation of connectivity (heterogeneity) with a value 0.654323. Highly conserved modules are significantly enriched with condition-specific genes associated with BPs and are dynamically expressed. In this study, hub genes were defined by modular connectivity, measured by their degree of interactions and a recursive relationship, based on the concept of hub and authority score by Kleinberg (1999). A hub score is obtained between the range of 0 to 1. Hub score = 1 is the most influential among the query genes, and the other genes with values relative to the score of the highest one. YHR089C(GAR1), YKL113C (RAD27), YNL248C (RPA49), YPL004C (LSP1), YMR001C (CDC5), YDR226W (ADK1), YLL026W (HSP104), and YPL127C (HHO1) were the cell cycle intramodular hub genes with high biological relevance (Fig. 5).

Hub genes were validated by assessing their repeatability in the two networks from our data sets and their functions in relevant pathways and cell cycle phases. We investigated their transcriptional regulators based on the YEASTRACT database (Teixeira et al., 2017). For instance, M/G1 phase regulator Yox1p/Mcm1p, a potential Cdc28p substrate, and Yhp1p regulate YKL113C (RAD27), YLL026W (HSP104), YPL127C (HHO1), YMR001C (CDC5), and YNL248C (RPA49). YPL004C (LSP1) shows activation of Pkc1p/Ypk1p stress resistance pathways and is regulated by Sok2p and Tec1p.YHR089C (GAR1) is a protein component of H/ACA snoRNP pseudouridylase complex involved in the modification and cleavage of the18S pre-rRNA. YDR226W (ADK1) is regulated by Ste12p transcription factor that activates genes involved in pseudohyphal/invasive growth pathways. YKL113C had a high hub score in both data sets, 0.8702 in PPI data set and 0.7710 in gene expression.

Table 1. Reported Function Versus Function Predicted by Our Method

Protein	*Reported function*	*Predicted function*
YAL003W	Protein synthesis elongation (SGD)	Translational elongation
YAL019W	DNA-dependent ATPase (BioGRID)	DNA synthesis
YAR028W	Putative integral membrane (SGD)	ER-nuclear membrane
YAR003W	Histone methyltransferase activity—H3-K4 specific (SGD)	RB binding protein 5

ATP; BioGRID; ER; RB; SGD.

Table 2. Molecular Function Similarity Between Two Hub Genes YMR001C and YLL026W

	GO:0019237	*GO:0051219*	*GO:0004672*	*GO:0044877*	*GO:0004672*
GO:0043531	0.173	0.144	0.046	0.17	0.046
GO:0005524	0.166	0.124	0.039	0.148	0.039
GO:0042623	0.04	0.093	0.126	0.119	0.126
GO:0051087	0.201	0.661	0.113	0.481	0.113
GO:0051082	0.201	0.661	0.113	0.481	0.113

GO, Gene Ontology.

4.4. Pathway enrichment analysis and functional annotation

Our goal is to provide functional annotations for 2274 and 1264 genes. Our focus is mainly on 677 common probes in both data sets obtained from differential co-expression analysis. Co-expression analysis has been used to predict functions of uncharacterized genes in tomato (*Solanum lycopersicum*) plant model (Ozaki et al., 2010) as also proved useful in this study. Since we are using off-the-shelf data sets obtained from publicly available sources, we perform annotation by mapping probe sets from identified modules to genome annotations.

We present existing functions and novel functions predicted by our method, as shown in Table 1. Our method predicted translation elongation as a function for YAL003W (Xiong et al., 2006). DNA synthesis was predicted to be the function of YAL019W (Pellegrini et al., 1999). ER-nuclear membrane was the function for YAR028W (Hitchcock et al., 2003). For YAR003W, our method predicted the GO annotation, RB binding protein (Zhao et al., 2016). The reported annotations are collected from BioGRID and SGD databases (Dwight et al., 2002), whereas the predicted functions from our method are validated from the literature as referenced.

To verify and validate the predicted annotations, we use semantic similarity and hypergeometric test to determine the reliability of our predictions. For instance, genes YMR001C and YLL026W are annotated by MF term sets {GO:0019237, GO:0051219, GO:0004672, GO:0044877, GO:0004672} and {GO:0043531, GO:0005524, GO:0042623, GO:0051087, GO:0051082}. We measure the MF similarity between them using Equations (14) and (15). We obtain the semantic similarities and list results in Table 2. We get sim(YMR001C,YLL026W) = 0.236. The pairwise similarity semantic values represent how informative the terms are based on the ontology topology. The values indicate what is common between the entities, a higher value indicating the probability of high similarity between the annotation terms. Hypergeometric probability distribution test is used in this work to test GO terms with $p < 0.01$ as a filter value for determining term significance.

Enrichment analysis facilitates annotation through identification of enriched biochemical pathways to obtain co-expression modules that are biologically significant. More conclusive information about the potential functions of proteins is derived from gene expression clusters with highly enriched functions. We conduct enrichment analysis for each module using DAVID (http://david.abcc.ncifcrf.gov) classification system to confirm functional significance and validate annotations, as shown in Table 3.

The module numbers and colors correspond to those in yeast cell cycle network in Figure 4b. Furthermore, we find that the yellow module is highly enriched with proteasome, a protein complex responsible for both cell cycle and responses to oxidative stress. The green module is highly enriched with

Table 3. Protein Enrichment Analysis for Cell Cycle Response Modules

Modules no.	*Module color*	*Pathways*	p
1	Turquoise	Ribosome	4.0E-19
2	Blue	Metabolic pathways	6.3E-2
3	Brown	Ribosome biogenesis in eukaryotes	4.7E-13
5	Green	Cell cycle-yeast	2.9E-7
		Sulfur metabolism	2.2E-2
		Meiosis-yeast	5.7E-2
6	Red	Metabolic pathways	2.6E-5

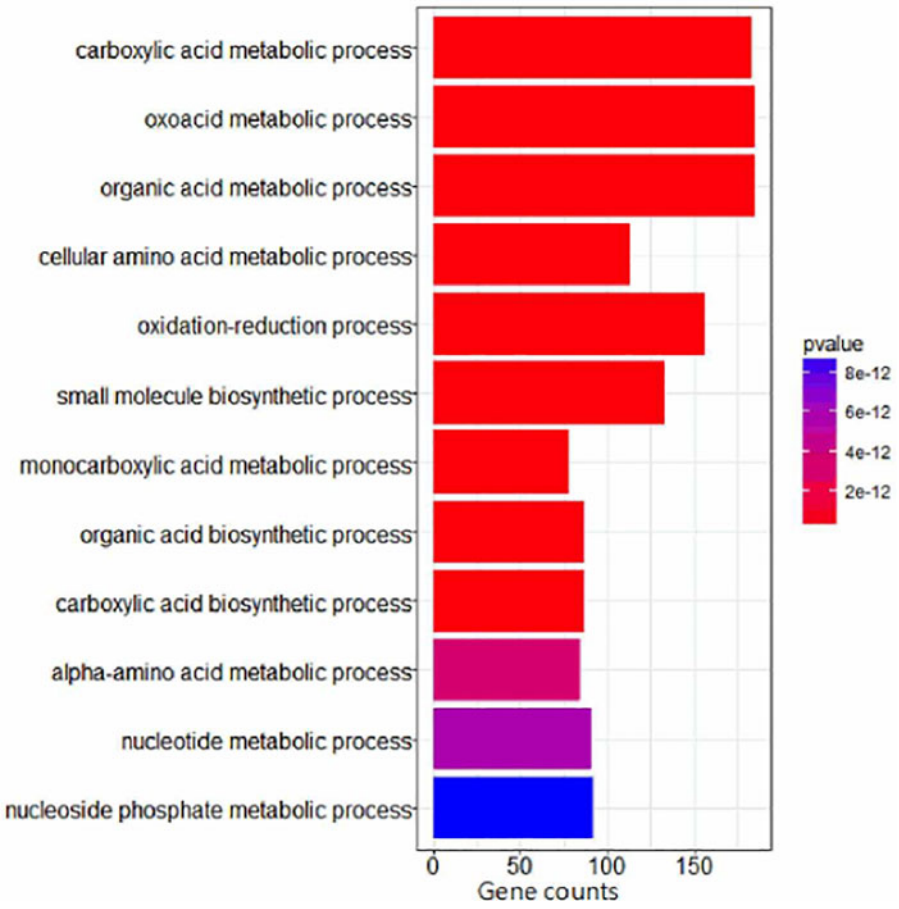

FIG. 6. Enrichment analysis with cluster profiler *p*-value adjustment method to reduce redundancy.

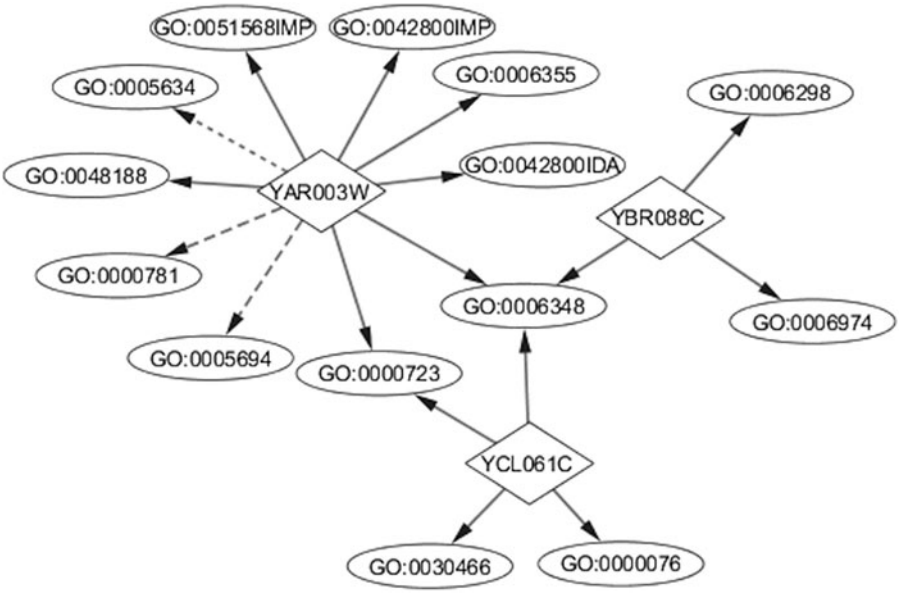

FIG. 7. A directed network diagram for YAR003W, YCL061C, and YBR088C proteins. The dotted line indicates an annotation predicted by our method.

Table 4. Performance Comparison on Yeast Data set for Three Neighbor-Voting-Based Methods, NV, GBA, and CNPFP

Performance metric	*NV*	*GBA*	*CNPFP*
AUROC	0.7241	0.8502	0.9862
Accuracy	0.8879	0.9386	0.9710
Precision	0.8895	0.9393	0.9593
Recall	0.9979	0.9999	0.9792
F-score	0.9406	0.9683	0.9691

AUROC; CNPFP, Co-expression analysis and Neighbor-voting algorithm for Protein Function Prediction; GBA, guilt-by-association; NV, neighbor-voting.

cell cycle and cyclin-dependent kinases, a protein kinase relevant for cell cycle regulation. Figure 6 shows the enrichment analysis results for yeast cell cycle data set. We use Cytoscape, a standalone visualization program, to visualize the results of network module annotation as shown in Figure 7. YAR003W is associated with some go terms, such as GO:0005634, GO:0000781, and GO:0005694. These GO terms have not yet been reported in the common yeast annotation databases such as SGD.

4.5. *Evaluation of predictive performance*

The predictive performance of this study was evaluated using a fivefold cross-validation technique. The data set was randomly divided into five sets. Then, for each iteration, onefold is withheld as the test set, and the remaining fourfolds are used as the training set. Each fold has the same ratio of positive and negative training instances, a property known as fold balance. Fold balance ensures that each classifier trained during cross-validation behaves as closely as possible to the final classifier trained on all the folds. This assumes that the training instances have the same distribution of positive and negative training instances.

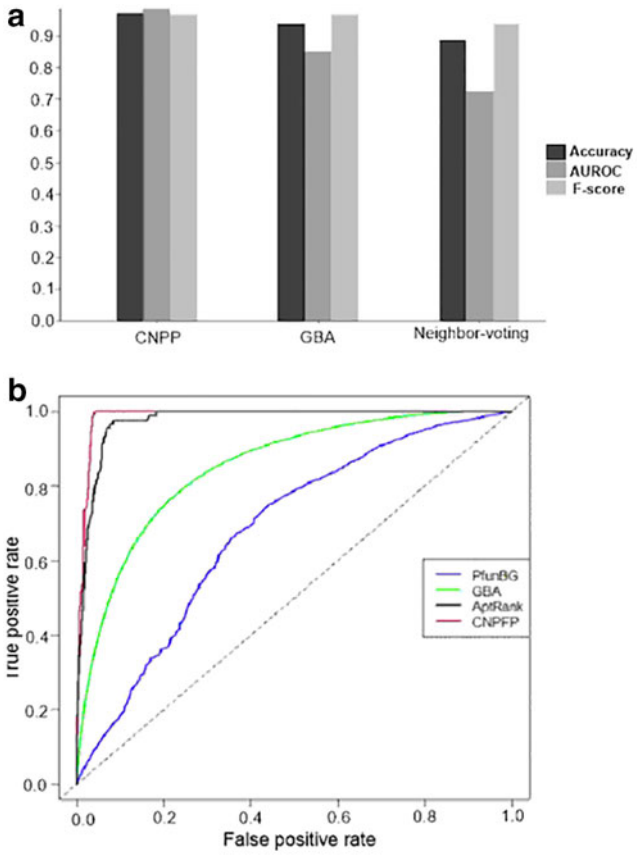

FIG. 8. Performances comparison of related studies. **(a)** Neighbor-voting, GBA, and CNPFP. **(b)** ROC curves of PfunBG, GBA, AptRank, and CNPFP. CNPFP, Co-expression analysis and Neighbor-voting algorithm for Protein Function Prediction; GBA, guilt-by-association; ROC.

Table 5. Prediction Performance of the Proposed Methodology

Alignment score	*No. of target proteins*	*AUROC*
0.3953	2283	0.8588
0.2568	1483	0.9127
0.1558	900	**0.9862**

To have greater coverage of protein functions, we used two annotation sources including GO and an Affymetrix Genechip platform. Therefore, we take into account topological properties and exploit the correlation structure of the respective functional categories. We improve the performance of the NV algorithm via feature optimization. This is achieved through global iterative similarity computation as described in Section 2. We obtained an AUROC of 0.9862, accuracy of 0.9710, and *F*-score of 0.9691. Table 4 and Figure 8a show the results obtained from three NV-based methods, our method performs better. NV refers to neighbor-voting, whereas GBA refers to guilt-by-association by degree method in EGAD R package (Ballouz et al., 2016). We also compared CNPFP with two state-of-the-art function prediction algorithms, diffusion-based method AptRank (Jiang et al., 2017) and a network-based method known as protein functions from birelational graph (PfunBG) (Jiang, 2011). Our method performs better with AUROC values: 0.6801, 0.8502, 0.9753, and 0.9862 for PfunBG, GBA, AptRank, CNPFP, respectively, as shown in Figure 8b. We use an alignment score to exploit the usefulness of global information in terms of relative position of a protein with respect to specific other proteins (Table 5). The alignment score is used to determine the number of target proteins by pruning less significant interactions. The prediction of functions iteratively helps to get the most consistent agreement and hence reduce false positives. From the results, using global iterative approach improves prediction performance.

5. CONCLUSION

In this study, we presented a co-expression network-based approach called CNPFP to detect differentially expressed modules in PPI data and predict protein functions. The proposed method builds highly connected subgraphs (cliques) based on hierarchical aggregation and discovers biologically informative relationships between genes. Then predict protein functions based on co-expression pattern mining across networks, semantic and intrinsic relationships between interacting proteins. To improve prediction performance, the proposed method combines network structure and attribute information contained in the graphs.

Experimental results obtained are better than other algorithms. The method predicts some GO terms that have not yet been reported in the common annotation databases such as SGD, they may be included in the future. For future work, we intend to identify putative potential new cell cycle regulators and novel protein annotations from statistical and network analysis.

AUTHORS' CONTRIBUTIONS

J.S.W. and J.M. designed the methods and conceived the study. J.S.W. and J.M. implemented the entire method and performed the experiments. J.S.W., J.M., and Y.L. analyzed the prediction results. J.S.W. and J.M. wrote the article. All authors read and approved the final article.

AVAILABILITY OF DATA AND MATERIALS

The proposed method is available at https://github.com/Mjwl/CNPFP

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no conflicting financial interests.

FUNDING INFORMATION

This study was supported by the National Natural Science Foundation of China (Grant Nos. 61872055, 31872116).

REFERENCES

Ballouz, S., Weber, M., Pavlidis, P., et al. 2016. EGAD: Ultra-fast functional analysis of gene networks. *Bioinformatics* 33, 612–614.

Bertoli, C., Skotheim, J.M., and De Bruin, R.A. 2013. Control of cell cycle transcription during G1 and S phases. *Nat. Rev. Mol. Cell Biol*. 14, 518.

Cheng, C., Fu, Y., Shen, L., et al. 2013. Identification of yeast cell cycle regulated genes based on genomic features. *BMC Syst. Biol*. 7, 70.

Childs, K.L., Davidson, R.M., and Buell, C.R. 2011. Gene coexpression network analysis as a source of functional annotation for rice genes. *PLoS One* 6, e22196.

Clarke, C., Madden, S.F., Doolan, P., et al. 2013. Correlating transcriptional networks to breast cancer survival: A large-scale coexpression analysis. *Carcinogenesis* 34, 2300–2308.

Dong, J., and Horvath, S. 2007. Understanding network concepts in modules. *BMC Syst. Biol*. 1, 24.

Dwight, S.S., Harris, M.A., Dolinski, K., et al. 2002. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res*. 30, 69–72.

Eisen, M.B., Spellman, P.T., Brown, P.O., et al. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A*. 95, 14863–14868.

Gibbs, D.L., Baratt, A., Baric, R.S., et al. 2013. Protein co-expression network analysis (ProCoNA). *JCBI* 3, 11.

Gligorijević, V., and Pržulj, N. 2015. Methods for biological data integration: Perspectives and challenges. *J. R. Soc. Interface* 12, 20150571.

Guan, L., Zhang, S., and Xu, H. 2017. BAMORF: A novel computational method for predicting the extracellular matrix proteins. *IEEE Access* 5, 18498–18505.

Hashemifar, S., and Xu, J. 2014. HubAlign: An accurate and efficient method for global alignment of protein–protein interaction networks. *Bioinformatics* 30, i438–i444.

Hitchcock, A.L., Auld, K., Gygi, S.P., et al. 2003. A subset of membrane-associated proteins is ubiquitinated in response to mutations in the endoplasmic reticulum degradation machinery. *Proc. Natl. Acad. Sci. U. S. A*. 100, 12735–12740.

Horvath, S., and Langfelder, P. 2008. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* 9, 559.

Hu, W., Lin, X., and Chen, K. 2015. Integrated analysis of differential gene expression profiles in hippocampi to identify candidate genes involved in Alzheimer's disease. *Mol. Med. Rep*. 12, 6679–6687.

Ideker, T., and Krogan, N.J. 2012. Differential network biology. *Mol. Syst. Biol*. 8, 565.

Ihmels, J., Bergmann, S., Berman, J., et al. 2005. Comparative gene expression analysis by a differential clustering approach: Application to the *Candida albicans* transcription program. *PLoS Genet*. 1, e39.

Jiang, B., Kloster, K., Gleich, D.F., et al. 2017. AptRank: An adaptive PageRank model for protein function prediction on bi-relational graphs. *Bioinformatics* 33, 1829–1836.

Jiang, J.Q. 2011. Learning protein functions from bi-relational graph of proteins and function annotations. International Workshop on Algorithms in Bioinformatics, Springer, 128–138.

Jones, P., Binns, D., Chang, H.-Y., et al. 2014. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* 30, 1236–1240.

Kadarmideen, H.N., Watson-Haigh, N.S., and Andronicos, N.M. 2011. Systems biology of ovine intestinal parasite resistance: Disease gene modules and biomarkers. *Mol. Biosyst*. 7, 235–246.

Kleinberg, J.M. 1999. Authoritative sources in a hyperlinked environment. *J. ACM* 46, 604–632.

Lai, Y., Wu, B., Chen, L., et al. 2004. A statistical method for identifying differential gene–gene co-expression patterns. *Bioinformatics* 20, 3146–3155.

Lai, Y.-H., Li, Z.-C., Chen, L.-L., et al. 2012. Identification of potential host proteins for influenza A virus based on topological and biological characteristics by proteome-wide network approach. *J. Proteomics* 75, 2500–2513.

Langfelder, P., Zhang, B., and Horvath, S. 2008. Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R. *Bioinformatics* 24, 719–720.

Liu, R., Zhang, W., Liu, Z.-Q., et al. 2017. Associating transcriptional modules with colon cancer survival through weighted gene co-expression network analysis. *BMC Genomics* 18, 361.

Ma, C., and Wang, X. 2012. Application of the Gini correlation coefficient to infer regulatory relationships in transcriptome analysis. *Plant Physiol*. 160, 192–203.

Ma, H., Schadt, E.E., Kaplan, L.M., et al. 2011. COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method. *Bioinformatics* 27, 1290–1298.

Meng, J., Wekesa, J.S., Shi, G.L., et al. 2016. Protein function prediction based on data fusion and functional interrelationship. *Math. Biosci*. 274, 25–32.

Mi, H., Muruganujan, A., Casagrande, J.T., et al. 2013. Large-scale gene function analysis with the PANTHER classification system. *Nat. Protoc*. 8, 1551.

Navot, A., Shpigelman, L., Tishby, N., et al. 2006. Nearest neighbor based feature selection for regression and its application to neural activity. Advances in Neural Information Processing Systems, 996–1002.

O'Meara, M.J., Ballouz, S., Shoichet, B.K., et al. 2016. Ligand similarity complements sequence, physical interaction, and co-expression for gene function prediction. *PLos One* 11, e0160098.

Ozaki, S., Ogata, Y., Suda, K., et al. 2010. Coexpression analysis of tomato genes and experimental verification of coordinated expression of genes found in a functionally enriched coexpression module. *DNA Res*. 17, 105–116.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., et al. 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A*. 96, 4285–4288.

Pesquita, C., Faria, D., Falcao, A.O., et al. 2009. Semantic similarity in biomedical ontologies. *PLoS Comput. Biol*. 5, e1000443.

Prasad, A., Saha, S., Chatterjee, P., et al. 2017. *Protein Function Prediction from Protein Interaction Network Using Bottom-up L2L Apriori Algorithm*. Springer, Singapore, 3–16.

Ravasz, E., Somera, A.L., Mongru, D.A., et al. 2002. Hierarchical organization of modularity in metabolic networks. *Science* 297, 1551–1555.

Russo, P.S., Ferreira, G.R., Cardozo, L.E., et al. 2018. CEMiTool: A Bioconductor package for performing comprehensive modular co-expression analyses. *BMC Bioinformatics* 19, 56.

Schlicker, A., Domingues, F.S., Rahnenfuhrer, J., et al. 2006. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics* 7, 1.

Stark, C., Breitkreutz, B.J., Reguly, T., et al. 2006. BioGRID: A general repository for interaction datasets. *Nucleic Acids Res.* 34 (Database issue), D535–D539.

Teixeira, M.C., Monteiro, P.T., Palma, M., et al. 2017. YEASTRACT: An upgraded database for the analysis of transcription regulatory networks in *Saccharomyces cerevisiae. Nucleic Acids Res*. 46 (Database Issue), D348–D353.

Wang, S., Zhai, C., et al., 2015. Exploiting ontology graph for predicting sparsely annotated gene function. *Bioinformatics* 31, i357–i364.

Watson, M. 2006. CoXpress: Differential co-expression in gene expression data. *BMC Bioinformatics* 7, 509.

Xiong, J., Rayner, S., Luo, K., et al. 2006. Genome wide prediction of protein function via a generic knowledge discovery approach based on evidence integration. *BMC Bioinformatics* 7, 268.

Yu, G., Zhu, H., and Domeniconi, C. 2015. Predicting protein functions using incomplete hierarchical labels. *BMC Bioinformatics* 16, 1.

Yu, G., Zhao, Y., Lu, C., et al. 2017. HashGO: Hashing gene ontology for protein function prediction. *Comput. Biol. Chem*. 71, 264–273.

Zhang, B., and Horvath, S. 2005. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol*. 4, Article17.

Zhang, X., Xiao, W., and Hu, X. 2018. Predicting essential proteins by integrating orthology, gene expressions, and PPI networks. *PLoS One* 13, e0195410.

Zhao, B., Sai, H., Xueyong, L., et al. 2016. An efficient method for protein function annotation based on multilayer protein networks. *Hum. Genomics* 10, 33.

Zheng, C.-H., Yuan, L., Sha, W., et al. 2014. Gene differential coexpression analysis based on biweight correlation and maximum clique. *BMC Bioinformatics* 15, S3.

Zhou, H., Gao, S., Nguyen, N.N., et al. 2014. Stringent homology-based prediction of *H. sapiens*-M. tuberculosis H37Rv protein-protein interactions. *Biol. Direct*. 9, 5.

Address correspondence to:
Jael Sanyanda Wekesa
School of Computer Science and Technology
Dalian University of Technology
Dalian
Liaoning 116023
China

E-mail: jael@mail.dlut.edu.cn