Centrality Values of Yeast Proteins in a PPI Network Are Related to Their Essentiality and Functions

Md. Altaf-Ul-Amin (1,*), Sony Hartono Wijaya (1,2), Dodi Fitra Chandra (1), and Shigehiko Kanaya (1)

(1) Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0192 Japan

(2) Department of Computer Science, Bogor Agricultural University, Jl. Meranti Wing 20 Level 5 Kampus IPB Dramaga, Bogor 16680, Indonesia

(Received June 1, 2017; Accepted June 13, 2017)

It has long been investigated and understood that centrality of proteins in the context of protein-protein interaction (PPI) networks are related to their essentiality. In the present work, we validate the relations between essentiality of yeast proteins and their centrality measures in a PPI network by following a different approach using the concept of the receiver operating characteristic (ROC) curve. We found that all centrality measures are related to essentiality. However, the degree centrality performed better in case of the data we used. By deeply examining different centrality values of yeast proteins we find that they are not highly correlated, which has leaded us to hypothesize that centralities might have some relations with gene/protein functions. Indeed, we found that many of the clusters generated based on the pattern of centrality values are rich with similar function proteins. Different types of centrality values imply different types of importance of a node in a network and the functions of genes are of various types. In the present work, we hypothesized that important genes of different functions may tend to show different patterns of centralities and here we show some preliminary links between groups of similar function genes and profiles of centrality values. The concepts of network biology discussed in this paper are applicable to other networks including networks of chemical compounds.

Key Words: PPI network, yeast, essential genes, centrality measures, ROC curve

1. Introduction

Network analysis has become popular for systems level analysis of various phenomena in versatile fields including bioinformatics and chemoinformatics. Both micro- and macro-level networks of various types facilitate the development of theory and methodology for solving core scientific issues. The application ranges from studies of ecological structure, biodiversity and environment, evolution and extinction of species. It is also very widely used in studies on networks of chemical
compounds and metabolites. In the present work we have applied the concept of node centralities to characterize essential genes/proteins.

Essential genes are the genes that are must required for the survival of a living organism. Identifying and analyzing different aspects of essential genes are necessary for understanding life as a system. One of the targets of synthetic biology is to create simple life forms having required phenotypic traits dedicated to specific applications [1]. Comprehensive information on essential genes is useful for such synthetic biology research because that can pave the way to determine the minimal sets of essential genes required for a simple life form. Furthermore, essential genes and proteins encoded by them might be potential targets of antimicrobial agents and pesticides [2]. Research therefore has been made to assess the characteristics of essential genes using wet lab experiments as well as computational techniques.

In most cases, proteins do not work individually. They interact with other proteins and form complex of different sizes and shapes to perform specific biological tasks. Recent invention of high throughput experiments such as yeast two hybrid assay (Y2H) [3], mass spectrometry of purified complexes, say tandem affinity purification (TAP) [4] and high-throughput mass-spectrometric protein complex identification (HMS-PCI) [5] have enabled us to construct interaction networks involving global set of proteins of an organism. It has been shown in a number of studies that high-density clusters of PPI networks correspond to protein complexes. Thus, it is obvious that the PPI network contains biologically meaningful information. It is reasonable to consider that the proteins that are placed at structurally important positions of a PPI network are also biologically important. The importance of a node in a network is determined by using the concept of centrality measures. A global PPI network can allow ranking of the proteins by using different centralities. Therefore, it is rational to think that highly central genes are essential and a number of researches have been conducted to find links between essentiality of genes and their centrality in PPI networks.

There are many ways of exposing positive correlation between centrality of proteins in PPI networks and their essentiality. Jeong et al. showed that the degree of nodes in a yeast PPI network correlates with the phenotypic effect of its deletion [6]. Specifically they observed that high degree nodes are three times more likely to be essential than nodes with a few interactions. Another work [7] analyzed a PPI network of yeast and showed that the higher proportion of genes are essential if a set of gene is selected based on higher centrality values. This work showed that the subgraph centrality measure performed better in identifying essential genes compared to several other centrality measures. Elena et al. [8] also investigated relations between the hubs in yeast PPI network and their essentiality. They also determined the fraction of essential genes in fractions of selected nodes based on higher degree or centralities and quantified the relations between essentiality and centrality by calculating Kendall’s tau and Spearman’s rank correlations.

Previously, the ROC curve has been implemented as a powerful tool to measure the quality of classifiers [9] and similarity measures [10]. In the present work we established the relations between centrality and essentiality of genes using ROC analysis by means of determining the area under the ROC curves and the minimum distance of the ROC curve from the optimum point. Our approach assesses the relation between centrality and essentiality in a global context. However, we further find that pattern of centralities of genes can be linked to their functions.

Different centralities indicate different types of importance of a node in the network. In other words different types of information is depicted in different types of centralities. Thus we hypothesize that different centrality measures of proteins in a PPI network may have relations with the different functions of proteins. In this study, we discovered that certain group of similar function genes attain specific pattern of centrality values. In other words, in multi dimensional space of centralities some groups of similar function genes are very close to each other.

2. Material and Methods

a. Collection and Preprocessing of Data

Initially, in this work we focused on Saccharomyces cerevisiae, which is a species of yeast. This species of yeast has been instrumental to baking, and brewing since ancient times. We collected protein-protein interaction datasets from the BioGRID database [11] corresponding to several publications [12–16]. After removing the duplicate interactions using an in-house script, we finally obtained 13,127 interactions. For meaningful assessment of certain centrality measures, it is recommended that the network be connected i.e. there exist at least one path from a node to any other node in the network. So we determined the largest connected component of the aforementioned network using an in-house tool based on depth first search [17]. The largest connected component consisted of 13,071 interactions and we used this network for the next steps of analysis. Notice that most of the interactions are included in the largest component, which consists of 2,983 nodes. This network is shown in Supplementary Figure 1 visualized using Cytoscape [18].

The list of essential genes of yeast was downloaded from the Saccharomyces Genome Database (SGD) [19]. SGD provides comprehensive integrated biological
information for the budding yeast (Saccharomyces cerevisiae) along with search and analysis tools to explore these data, enabling the discovery of functional relationships between sequences and gene products in fungi and higher organisms. We have collected a list of 1,280 essential genes of yeast from the following location: http://www.yeastgenome.org/observable/inviable/overview. Out of these 1,280 genes 835 are included in the network we considered in this work.

b. Centrality Measurements

It is easy to realize that in the context of a network all nodes are not equally important. A node with very high degree for example is obviously more important compared to a node having few degrees. The importance of a node in a network is precisely and mathematically determined by centrality measures. Not only the degree but also several other measures based on different logics can be used to determine the importance of a node in a network and hence there are a number of categories of centrality measures [20].

In the following we discuss about the five types of centrality measures utilized in the present work in the context of a simple graph \( G = (N, E) \) where \( N \) is the set of nodes and \( E \) is the set of edges of the graph. Let \( G \) be a connected graph and \( A \) is the adjacency matrix of the graph \( G \).

**Degree Centrality**

Degree centrality (DC) of a node \( i \) in the context of \( G \) is the degree of the node in the graph which can be mathematically defined as Eq. 1.

\[
C_{\text{degree}}(i) = \sum_{j=1}^{N} A_{ij} \tag{1}
\]

Degree centrality is often interpreted in terms of the immediate risk of the node for catching whatever is flowing through the network (such as a virus, or some information).

**Closeness Centrality**

The farness of a node is the sum of the shortest-path distances from the node to all other nodes in the graph. The reciprocal of farness is the closeness centrality (CC). The closeness centrality of a node \( i \) is defined as Eq. 2.

\[
C_{\text{closeness}}(i) = \frac{1}{\sum_{u\in N} d(i, u)} \tag{2}
\]

Here, \( d(i, u) \) is the shortest distance between node \( i \) and node \( u \). Notice that \( d(i,i)=0 \) in Eq.2. To calculate closeness centrality of the nodes of \( G \) it requires that \( G \) is a connected graph. Closeness centrality can be viewed as the efficiency of a node in spreading information to all other nodes.

**Centrality Measurements**

The betweenness centrality (BC) of a node \( i \) is defined as Eq. 3 [21].

\[
C_{\text{betweenness}}(i) = \sum_{u\in N} \sum_{w\in N, w \neq i} \frac{\sigma_{uw}(i)}{\sigma_{uw}} \tag{3}
\]

Here \( \sigma_{uw} \) is the total number of shortest paths between node \( u \) and \( w \) and \( \sigma_{uw}(i) \) is the number of shortest paths between node \( u \) and \( w \) that pass the node \( i \). Nodes that occur on many shortest paths between other nodes have higher betweenness than those that do not. Nodes of high betweenness centrality are important for transport. If they are blocked, transport becomes less efficient and on the other hand if their capacity is improved transport becomes more efficient.

**Eigenvector Centrality**

Let \( \lambda \) be the largest eigenvalue of \( A \) and \( x \) be the corresponding eigenvector. Here, \( \lambda \) is the largest root of the equation \( |A-\lambda I|=0 \) (where \( I \) is an identity matrix). According to the definition of eigenvector, \( Ax = \lambda x \) (A is a \( |N| \times |N| \) matrix, \( x \) is a \( |N| \times 1 \) vector and \( \lambda \) is a scalar). The \( i \)-th component of the eigenvector gives the eigenvector centrality (EC) score of the \( i \)-th node in the network. From \( Ax = \lambda x \) we can write \( x = \frac{1}{\lambda} Ax \). Therefore, the eigenvector centrality of the \( i \)-th node can be calculated by Eq. 4 [22].

\[
C_{\text{eigenvector}}(i) = x_i = \frac{1}{\lambda} \sum_{j=1}^{N} A_{ij} x_j \tag{4}
\]

For any node, the eigenvector centrality score is proportional to the sum of the scores of all nodes, which are connected to it. Consequently, a node has high value of EC either if it is connected to many other nodes or if it is connected to others that themselves have high eigenvector centrality.

**Subgraph Centrality**

Subgraph centrality (SC) of a node in a network is the weighted sum of the numbers of closed walks of different lengths that start and end at that node. The number of closed walks of length \( k \) starting and ending on node \( i \) in the network is given by the local spectral moments \( \mu_k(i) \), simply defined as the \( i \)-th diagonal entry of the \( k \)-th power of the adjacency matrix, \( A \), i.e. \( \mu_k(i) = (A^k)_{ii} \). The subgraph centrality of node \( i \) is defined as Eq.5 [23].

\[
c_{\text{subgraph}}(i) = \sum_{k=0}^{\infty} \frac{\mu_k(i)}{k!} \tag{5}
\]

c. Calculation of Centralities

Cytoscape is open-source software for integration, visualization and analysis of biological networks. It can
be extended through Cytoscape plug-ins. In this work we use Cytoscape ver.3.4.0 integrated with CytoNCA plug-ins [24] to calculate SC. In case of DC, CC, BC, and EC, we calculate the centrality values by utilizing CentriBin [20]. As an example, we show a small network in Fig. 1 and Table 1 shows the 5 types of centrality values of the nodes of this network.

![Network Diagram](image)

**Fig. 1.** An example of a small network constructed from 10 nodes.

**Table 1. Centrality values from the network in Fig. 1.**

| Node | DC   | CC    | BC    | EC    | SC    |
|------|------|-------|-------|-------|-------|
| n1   | 1    | 0.040 | 0     | 0.215 | 1.774 |
| n2   | 1    | 0.040 | 0     | 0.215 | 1.774 |
| n3   | 1    | 0.040 | 0     | 0.215 | 1.774 |
| n4   | 5    | 0.059 | 21    | 0.572 | 5.507 |
| n5   | 2    | 0.050 | 0     | 0.395 | 3.027 |
| n6   | 3    | 0.063 | 20    | 0.476 | 3.860 |
| n7   | 3    | 0.056 | 18.5  | 0.297 | 3.271 |
| n8   | 2    | 0.042 | 3.5   | 0.156 | 2.441 |
| n9   | 2    | 0.033 | 0.5   | 0.118 | 2.388 |
| n10  | 2    | 0.042 | 3.5   | 0.156 | 2.441 |

Node n4 and n1 are neighbors, but have quite different centrality values. Here, we discuss the calculation process of the centrality values of n4. A can be represented in adjacency matrix A as follows:

\[
A = \begin{pmatrix}
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0
\end{pmatrix}
\]

DC of node n4 is 5, which can be easily verified by counting the edges connected to n4 in Fig. 1 or by counting the number of 1s in the row corresponding to n4 in the adjacency matrix A as indicated in Eq. 1.

\[
C_{\text{Closeness}}(n4) = \frac{1}{\sum_{i=1}^{10} d(n4,n_i)} = \frac{1}{1+1+1+1+1+1+2+3+4+3} = \frac{1}{17} = 0.059
\]

There are 9 other nodes, other than n4 in Fig. 1, which correspond to 36 node pairs. In connection to Eq. 3, \(\sigma_{uw}\) and \(\sigma_{uw}(n4)\) corresponding to these 36 pairs are shown in Table 2. According to Eq. 3, the betweenness centrality of node n4 is the sum of the column “\(\sigma_{uw}(n4) / \sigma_{uw}\)” of Table 2, which is 21.

**Table 2. Detail calculation process of the betweenness centrality of n4.**

| Node 1 | Node 2 | \(\sigma_{uw}\) | \(\sigma_{uw}(n4)\) | \(\sigma_{uw}(n4)/ \sigma_{uw}\) |
|--------|--------|-----------------|-------------------|-----------------------------|
| n1     | n2     | 1               | 1                 | 1                           |
| n1     | n3     | 1               | 1                 | 1                           |
| n1     | n5     | 1               | 1                 | 1                           |
| n1     | n6     | 1               | 1                 | 1                           |
| n1     | n7     | 1               | 1                 | 1                           |
| n1     | n8     | 2               | 2                 | 1                           |
| n1     | n9     | 2               | 2                 | 1                           |
| n1     | n10    | 1               | 1                 | 1                           |
| n2     | n3     | 1               | 1                 | 1                           |
| n2     | n5     | 1               | 1                 | 1                           |
| n2     | n6     | 1               | 1                 | 1                           |
| n2     | n7     | 1               | 1                 | 1                           |
| n2     | n8     | 1               | 1                 | 1                           |
| n2     | n9     | 2               | 2                 | 1                           |
| n2     | n10    | 1               | 1                 | 1                           |
| n3     | n5     | 1               | 1                 | 1                           |
| n3     | n6     | 1               | 1                 | 1                           |
| n3     | n7     | 1               | 1                 | 1                           |
| n3     | n8     | 1               | 1                 | 1                           |
| n3     | n9     | 2               | 2                 | 1                           |
| n3     | n10    | 1               | 1                 | 1                           |
| n5     | n6     | 1               | 0                 | 0                           |
| n5     | n7     | 1               | 0                 | 0                           |
| n5     | n8     | 1               | 0                 | 0                           |
| n5     | n9     | 2               | 0                 | 0                           |
| n5     | n10    | 1               | 0                 | 0                           |
| n6     | n7     | 1               | 0                 | 0                           |
| n6     | n8     | 1               | 0                 | 0                           |
| n6     | n9     | 2               | 0                 | 0                           |
From Table 3, the vector component of the eigenvector related to the highest eigenvalue and corresponding to n4 is 0.57. Therefore, the eigenvector centrality of n4 is 0.57. In term of adjacency matrix, Eq. 5 can be written as follows:

$$C_{subgraph}(i) = \sum_{k=0}^{n} \frac{[a^k]_{ii}}{k!}$$

We calculated the SC values in Table 1 using k=0 to 10 because the higher terms are very small. For n4, i.e. for i=4, this value is 5.507.

Table 3. Eigenvalues and eigenvectors of the adjacency matrix A.

| Values | n1    | n2    | n3    | n4    | n5    | n6    | n7    | n8    | n9    | n10   |
|--------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| [1]    | 2.7E  | 2.0E  | 7.4E  | 2.4E  | -2.4E | -8.8E | -1.3E | -1.1E | -2.0E | -2.3E |
|       | +00   | +00   | +00   | +17   | +17   | +16   | +15   | +00   | +00   | +00   |
| Vectors | n1    | -0.21 | -0.15 | 0.33  | 0.00  | 0.00  | 0.00  | 0.13  | 0.82  | 0.02  | 0.28  | 0.19  |
|        | n2    | -0.21 | -0.15 | 0.33  | 0.03  | -0.10 | -0.79 | -0.23 | 0.02  | 0.28  | 0.19  | 0.19  |
|        | n3    | -0.21 | -0.15 | 0.33  | -0.19 | 0.54  | 0.42  | -0.43 | 0.02  | 0.28  | 0.19  | 0.19  |
|        | n4    | -0.57 | -0.30 | 0.24  | 0.00  | 0.00  | 0.00  | 0.00  | -0.02 | -0.57 | -0.43 | 0.01  |
|        | n5    | -0.39 | -0.15 | -0.32 | 0.15  | -0.44 | 0.23  | -0.16 | 0.57  | 0.28  | 0.01  | 0.01  |
|        | n6    | -0.47 | 0.00  | -0.48 | 0.00  | 0.00  | 0.00  | 0.00  | -0.61 | 0.00  | 0.39  | 0.19  |
|        | n7    | -0.29 | 0.45  | -0.27 | -0.15 | 0.44  | -0.23 | 0.16  | 0.14  | 0.18  | -0.47 | 0.34  |
|        | n8    | -0.15 | 0.45  | 0.14  | -0.66 | -0.23 | 0.00  | 0.00  | 0.22  | -0.28 | 0.34  | 0.34  |
|        | n9    | -0.11 | 0.45  | 0.38  | 0.15  | -0.44 | 0.23  | -0.16 | -0.40 | 0.28  | -0.30 | 0.34  |
|        | n10   | -0.15 | 0.45  | 0.14  | 0.66  | 0.23  | 0.00  | 0.00  | 0.22  | -0.28 | 0.34  | 0.34  |

d. ROC Analysis

We evaluated the capabilities of different centrality measures to identify the essentiality of genes using the concept of ROC analysis. The ROC curve was created by selecting a series of the centrality measures as threshold values to generate True Positive Rate (TPR) and False Positive Rate (FPR). TPR is the proportion of true positive predictions out of all the positive data and FPR is the proportion of false positive predictions out of all the negative data [25,26] and can be expressed by the following equations:

$$FPR = \frac{FP}{FP+TN} \quad TPR = \frac{TP}{TP+FN}$$

Corresponding to a certain threshold centrality value \( th \), false positive (FP), true positive (TP), false negative (FN) and true negative (TN) are defined as follows: TP is the number of essential genes having centrality values > \( th \), FP is the number of non-essential genes having centrality values > \( th \), TN is the number of non-essential genes having centrality values < \( th \), and FN is the number of essential genes having centrality values < \( th \).

We compared the performance of different centrality measures using the minimum distance of (FPR, TPR) points to the theoretical optimum point and by using the Area Under the ROC Curve (AUC) analysis. In term of AUC analysis, we used R package named ROCR [27]. Also we evaluated the correlation between the threshold centrality measures and the corresponding precision.

3. Results and discussion

a. Determination and normalization of the centrality values

Initially we determined 5 types of centrality values of all the proteins in the PPI network. We used a plugin to Cytoscape called "cytoNCA" [24] and CentriBin [20] to determine the centrality values. The distributions of the centrality values are shown in Fig. 2a. From Fig. 2a, it is evident that the range of the centrality values is quite different. Therefore, we normalized the centrality values such that they span between 0 and 1 using the following linear transformation:

$$x' = \frac{x-min}{max-min}$$

(7)

The shape of the distribution of the centrality values remains the same after this linear transformation as it is shown in Fig. 2b.

b. Essential genes tend to be more central

Our objective is to assess how the centrality measures of the genes are related to their essentiality. In other words, we tried to assess which centrality measure can be used for better classification of genes into essential and non-essential groups. To do so first we plotted the probability density masses of essential and non-essential genes with respect to 5 types of centrality values. These distributions are shown in Fig. 3. The blue and red histograms respectively correspond to essential and non-essential genes. The enlarged view of the right side part of each histogram is shown in the inset. These distributions show that essential genes tend to produce higher centrality values corresponding to all five types of centralities. To systematically assess the relations
between essentiality and centrality of yeast genes, we then constructed ROC curves corresponding to each type of centrality values.

For ROC analysis, we divided the range of the centrality measure into 100 intervals and considered the highest value of each interval as a threshold. Corresponding to every threshold, TP and FP were determined from the distribution of essential genes and FN and TN were determined from the distribution of non-essential genes. Here, TP and FN are the number of genes for which centrality values are greater than the threshold whereas FN and TN are the number of genes for which centrality values are less than the threshold. TPR and FPR were calculated for every threshold and the ROC curve was created by plotting the FPR along the x-axis and the corresponding TPR along the y-axis. The five ROC curves associated to five types of centrality values are shown in Fig. 4.

In the case of perfect and ideal classification, the ROC curve follows the vertical line from (0,0) to (0,1) and then horizontal line up to (1,1). In the case of random data, the ROC curve follows the diagonal line from (0,0) to (1,1). In the case of real data, the ROC curve usually follows above diagonal line. The (0,1) is the optimum classification point where FPR is zero and TPR is one and hence the (0,1) point will be referred to as ‘optimum point’. The performance of a classifier can be assessed either by measuring the minimum distance from the optimum point to the curve or by measuring the area under the ROC curve (AUC) [28]. In the case of the minimum distance, the lower the value of the minimum distance the better the classification. In the case of the AUC, the bigger the AUC value the better the classification.

We used the ROC curves of Fig. 4 and related information to evaluate the performance of different centrality measures for indicating essential genes. For this purpose we applied the following three approaches: (1) The Area under the ROC curve, (2) The minimum distance of the ROC curve from the optimum point. (3) The correlation between the thresholds and the precision. The third approach is very similar to the approaches adopted by other works [6-8]. The plots of precisions against the corresponding thresholds are shown in Fig. 5. Table 2 shows the AUC, the minimum distance and the correlation values associated to 5 types of centrality measures.
Based on all three approaches we found that the performance of DC was the best to classify essential and non-essential genes. Our results supported that of Jeong et al., which showed that high degree nodes were three...
times more likely to be essential. Estrada & Rodriguez-Velazquez however showed that subgraph centrality was better related to essentiality [23]. One reason behind these differences in results is that each work uses a different PPI network. All works support that all types of centrality measures are strongly related to essentiality of genes.

c. Relation between different centrality measures

Different centrality measures of genes can express different type of importance of a node in a network. Hence a gene that is ranked the highest by one type of centrality measure may not be ranked the same according to another type of centrality measure. To investigate this fact, we selected 100 top ranked genes based on each type of centrality measure. Fig. 6 shows the Venn diagram of the selected genes corresponding to each pair of centrality measure types. Fig. 7 shows the hierarchical clustering based on correlation between the centrality measures. Fig. 7 implies that correlations between all pairs of the centrality measures are not so high. However EC and SC are highly correlated which is consistent with the Venn diagrams in Fig. 6.

Based on these results, we hypothesize that the pattern of centrality measures of genes may have relations with their functions. To investigate this hypothesis, we performed hierarchical clustering of the genes based on five types of centrality measures and evaluated the richness of similar function genes in the clusters based on hypergeometric distribution about which we discuss in detail in the next section.

Table 2. The AUCs, minimum distances and correlations corresponding to 5 types of centralities.

| Type of centrality | AUC  | Minimum distance | Correlation |
|-------------------|------|------------------|-------------|
| BC                | 0.622| 0.914            | 0.892       |
| CC                | 0.625| 0.566            | 0.791       |
| DC                | 0.685| 0.551            | 0.948       |
| EC                | 0.639| 0.753            | 0.598       |
| SC                | 0.639| 0.917            | 0.434       |
d. Relation between centrality and gene functions

A high centrality measure of a node usually indicates elevated importance of the node in the context of the network. However different centrality measures are related to different types of importance. In this work we determined 5 types of centrality measures of 2,983 yeast proteins in the context of a PPI network. We observed that correlations between the different centrality values are not so high. This tempted us to hypothesize that proteins of different functions may have different pattern of centrality measures. To investigate this, we clustered the proteins based on their centrality measures.
Hierarchical clustering was performed based on Ward method [29] using 'hclust' function of R. We tentatively selected 100 clusters by cutting the dendrogram at an appropriate height. To assess the richness of similar function genes in individual clusters, we determined their p-values based on hypergeometric distribution. The R package GOStats was used for automatic generation of p-values in the context of all 3 types of Gene Ontologies namely Biological Process (BP), Molecular Function (MF) and Cellular Compartment (CeC). The best p-values and corresponding GO terms related to each cluster were used for subsequent analysis. A statistically significant p-value of a cluster indicates that group of similar function genes attain very similar pattern of centrality values.

To confirm that the similar patterns of centrality values of similar function genes is not a random phenomena, we prepared five sets of synthetic random data by shuffling existing centrality values (each type of centrality values were shuffled separately). Using the process as in case of the real centrality values, we generated 100 clusters of proteins from each set of random data. Also, we determined their p-values using GOStats [30]. The distribution of the clusters corresponding to real and random data with respect to -log(p-value) is shown in Fig. 8. Fig. 8 clearly implies that certain groups of similar function genes correspond to similar pattern of centrality values. In other words the pattern of centrality values of genes have relations with their functions.
To determine statistically significant clusters, we set empirical thresholds shown by the vertical dotted lines in Fig. 8 to select clusters most of which have p-values lower than the clusters generated based on random data. In Fig. 8, the p-values corresponding to empirical thresholds have been determined based on gamma distribution. For three types of primary gene ontology terms i.e. BP, MF and CeC, we selected 28, 10, and 33 statistically significant clusters. The centrality profiles of the genes belong to each statistically significant cluster corresponding to GO term biological process (BP) are shown in Fig. 9. Notice that the patterns of centrality values of the genes belonging to individual clusters are very similar and at the same time they are rich with similar function genes. This implies that certain group of genes belonging to certain functions has centrality values that follow specific type of patterns. We also obtained very similar results corresponding to ontology terms MF and CeC illustrated in Supplementary Figures 2 and 3. These links between similar function gene groups and profiles of centrality measures is a novel finding to our consideration.
4. Conclusions

Centrality measures of nodes in a network can indicate the importance of the nodes in the context of a network. Similarly, the centrality measures of proteins in a PPI network of an organism are supposed to indicate the importance of the proteins for the survival or healthy living of the organism. A number of research works have elucidated relations between essentiality of proteins with their centrality measures in PPI networks. In this work we validate those findings by following a different experimental approach based on ROC analysis. We further notice that different centrality measures are not highly correlated which inspired us to hypothesize that the pattern of centrality measures may have some relations with their functions. We cluster the genes based on their centrality values and find that certain gene clusters having similar pattern of centrality measures are rich with similar function genes. This implies that pattern of centrality values are related to their functions. We established by proper experiments that the similar patterns of centrality values of similar function genes are not random phenomena.

Acknowledgements

This work was supported by the National Bioscience Database Center in Japan; the Ministry of Education, Culture, Sports, Science, and Technology of Japan (16K07223 and17K00406) and NAIST Big Data Project.

References

1. Acevedo-Rocha, C.G., Fang, G., Schmidt, M., et al. (2013) From essential to persistent genes: A functional approach to constructing synthetic life. Trends Genet., 29 (5), 273–279.
2. Juhas, M., Eberl, L., and Church, G.M. (2012) Essential genes as antimicrobial targets and cornerstones of synthetic biology. Trends Biotechnol., 30 (11), 601–607.
3. Ito, T., Chiba, T., Ozawa, R., et al. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. U. S. A., 98 (8), 4569–4574.
4. Gavin, a C., Bösche, M., Krauser, R., et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature, 415, 141–147.
5. Ho, Y., Gruhler, a, Heilbut, a, et al. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature, 415 (January), 180–183.
6. Jeong, H., Mason, S.P., Barabási, a L., and Oltvai, Z.N. (2001) Lethality and centrality in protein networks. Nature, 411 (6833), 41–42.
7. Estrada, E. (2006) Virtual identification of essential proteins within the protein interaction network of yeast. Proteomics, 6 (1), 35–40.
8. Zotenko, E., Mestre, J., O’Leary, D.P., and Przytycka, T.M. (2008) Why do hubs in the yeast protein interaction network tend to be essential: Reexamining the connection between the network topology and essentiality. PLoS Comput. Biol., 4 (8).
9. Demšar, J. (2006) Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res., 7, 1–30.
10. Wijaya, S.H., Afendi, F.M., Batubara, I., et al. (2016) Finding an appropriate equation to measure similarity between binary vectors: case studies on Indonesian and Japanese herbal medicines. BMC Bioinformatics, 17 (1), 520.
11. Chatr-Aryamontri, A., Breitkreutz, B.J., Heinicke, S., et al. (2013) The BioGRID interaction database: 2013 Update. Nucleic Acids Res., 41 (D1), 816–823.
12. Krogan, N.J., and Greenblatt, J.F. (2001) Characterization of a Six-Subunit Holo-Elongator Complex Required for the Regulated Expression of a Group of Genes in Saccharomyces cerevisiae Characterization of a Six-Subunit Holo-Elongator Complex Required for the Regulated Expression of a Group of Gen. Mol. Cell. Biol., 21 (23), 8203–8212.
13. Krogan, N.J., Kim, M., Ahn, S.H., et al. (2002) RNA Polymerase II Elongation Factors of Saccharomyces cerevisiae: a Targeted Proteomics Approach. Society, 22 (20), 6979–6992.
14. Jansen, R., Yu, H., Greenbaum, D., et al. (2003) A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. Science (80– ), 302 (5644), 449–453.
15. Krogan, N.J., Peng, W.T., Cagney, G., et al. (2004) High-Definition Macromolecular Composition of Yeast RNA-Processing Complexes. Mol. Cell, 13 (2), 225–239.
16. Krogan, N.J., Cagney, G., Yu, H., et al. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature, 440 (7084), 637–643.
17. Thulasiraman, K., and Swamy, M.N.S. (1992) Graphs: theory and algorithms.
18. Smoot, M.E., Ono, K., Ruscheinski, J., et al. (2011) Cytoscape 2.8: New features for data integration and network visualization. Bioinformatics, 27 (3), 431–432.
19. Cherry, J.M., Adler, C., Ball, C., et al. (1998) SGD: Saccharomyces genome database. *Nucleic Acids Res.*, **26** (1), 73–79.
20. Junker, B.H., Koschützki, D., and Schreiber, F. (2006) Exploration of biological network centralities with CentiBiN. *BMC Bioinformatics*, **7**, 219.
21. Freeman, L.C. (1977) A Set of Measures of Centrality Based on Betweenness. *Sociometry*, **40** (1), 35–41.
22. Newman, M.E.J. (2008) The mathematics of networks, in *The New Palgrave Dictionary of Economics* (eds. Durlauf, S.N., and Blume, L.E.), Nature Publishing Group, Basingstoke, pp. 465–470.
23. Estrada, E., and Rodriguez-Velazquez, J. a. (2005) Subgraph Centrality in Complex Networks. *Phys. Rev. E*, **71** (5), 29.
24. Tang, Y., Li, M., Wang, J., et al. (2015) CytoNCA: A cytoscape plugin for centrality analysis and evaluation of protein interaction networks. *Biosystems*, **127**, 67–72.
25. Sonego, P., Kocsor, A., and Pongor, S. (2008) ROC analysis: Applications to the classification of biological sequences and 3D structures. *Brief. Bioinform.*, **9** (3), 198–209.
26. Li, M., Chen, J., Wang, J., et al. (2008) Modifying the DPClus algorithm for identifying protein complexes based on new topological structures. *BMC Bioinformatics*, **9** (398), 1–16.
27. Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005) ROC: Visualizing classifier performance in R. *Bioinformatics*, **21** (20), 3940–3941.
28. Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recognit. Lett.*, **27** (8), 861–874.
29. Ward, J.H. (1963) Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, **58** (301), 236–244.

30. Falcon, S.Ã., and Gentleman, R. (2007) Using GOstats to test gene lists for GO term association. *Bioinformatics*, **23** (2), 257–258.

**Supplementary Materials**

Supplementary Figure 1. The interaction network of 2983 yeast proteins utilized in this work. (DOCX 3.8 MB)

Supplementary Figure 2. Centrality profiles of the genes belong to each statistically significant cluster corresponding to GO term molecular function (MF). (DOCX 100 KB)

Supplementary Figure 3. Centrality profiles of the genes belong to each statistically significant cluster corresponding to GO term cellular compartment (CelC). (DOCX 159 KB)
Supplementary Figure 1. The interaction network of 2983 yeast proteins utilized in this work.
Supplementary Figure 2. Centrality profiles of the genes belong to each statistically significant cluster corresponding to GO term molecular function (MF).
Supplementary Figure 3. Centrality profiles of the genes belong to each statistically significant cluster corresponding to GO term cellular compartment (CeC).