

LARGE-SCALE COMPLEX NETWORK ANALYSIS TO INVESTIGATE ASSOCIATIONS OF DISORDERED PROTEINS AND SCALE-FREE NETWORK

Satanat Kitsiranuwat¹, Kitiporn Plaimas^{1*}, and Apichat Suratanee²

Received: March 03, 2018; Revised: April 18, 2018; Accepted: May 14, 2018

Abstract

In biomedical engineering studies, a protein-protein interaction network is widely used to identify biological processes and characterize associations among proteins. Typically, its structure forms a scale-free network which contains a few hub proteins and a lot of low-degree proteins. Thus, its degree distribution follows the power-law distribution. Investigating such a protein that might affect this structure may reveal an important protein for the whole network. In this research, we investigated disordered proteins whose removal results in a non-scale-free structure that may cause severe diseases like cancers. Later, a simple measure to identify these proteins was developed as an alternative method avoiding fitting a power-law distribution. A high performance for identifying disordered proteins was achieved by using the area under the ROC curve value greater than 0.9. In addition, the measure yielded a superior performance to the random selection procedure.

Keywords: Protein-protein interaction network, disordered protein, scale-free network, degree distribution, power-law form

Introduction

In biomedical engineering, the relationships among proteins are represented in a network, the so-called protein-protein interaction (PPI) network. The most outstanding topological feature of PPI networks is the scale-free (SF) network. The SF network is described by having a very high number of low-interaction proteins in the, whereas the number of high-

¹ Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Bangkok, 10330, Thailand. E-mail: kitiporn.p@chula.ac.th

² Department of Mathematics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, 10800, Thailand. E-mail: apichat.s@sci.kmutnb.ac.th

* Corresponding author

interaction proteins is very low. The high-interaction proteins are called hub proteins (Plaimas *et al.*, 2010). On the other hand, the property of an SF network shows the most distinguishable property of hub proteins. Technically, examining the property of an SF network can be done using the power-law of degree distribution. The motivation of this work is to study the relationships among SF networks, hub proteins, and disordered proteins.

Haynes *et al.* (2006) studied the relationship between disordered proteins and hub proteins. They compared the feature of disordered proteins between hub proteins and end proteins using analysis of various disorder parameters, the amino acid composition, and gene ontology annotations (biological process, molecular function, and cellular component). The hub proteins were nodes having degrees more than 10, and the end proteins were proteins with a single interaction. Haynes *et al.* mentioned that the characteristics between the hub proteins and disordered proteins are the same. Therefore, the assumption of this work is that disordered proteins might affect the SF network's property.

In this study, we developed a new measure on big data of a protein-protein interaction network. This measure is to identify disordered proteins affecting the SF network ($M_{SF.DISP}$). The measure was developed based on the node attributes of each protein. The performance of the measure was validated by the area under the ROC curve (AUC) according to the receiver operating characteristics (ROC) curve identifying disordered proteins affecting the SF network. In addition, it was validated with random selection experiments.

Background Knowledge

Data

A disordered protein is a protein in which some of the regions of its polypeptide chain, or combination of amino acid residues, are missing some structures such as folding (Sickmeier *et al.*, 2007). Disordered proteins result in some biological functions losing their roles. Therefore, disordered proteins cause some serious diseases

such as cancer, neurodegenerative disease, cardiovascular disease, and amyloidosis (Sickmeier *et al.*, 2007; Li *et al.*, 2015). Thus, disordered proteins are involved with human diseases. In this research, the data of disordered proteins were collected from the database of Online Mendelian Inheritance in Man (OMIM). The OMIM database combines the data of human genetic disorders using the characteristic of the phenotype (Amberger *et al.*, 2015).

A PPI network characterizes the associations among proteins. Mathematically, a network can be represented as a graph in which the nodes represent the proteins and the edges represent the relationship between 2 proteins. Several PPI networks serve the property of the SF network (Province, 2015). The PPI network of *Homo sapiens* (humans) was constructed by collecting data from the STRING database (version 10) (Szklarczyk *et al.*, 2015). The database provides a list of 2 proteins that are related by the confidence of the combined score. In this study, the high STRING confidence score of 950 was chosen to reconstruct the human network. We analyzed a large-scale human PPI network consisting of 8,208 proteins and 45,553 interactions. The slope of the degree distribution in power-law form is 2.75. That means the human PPI network has the property of the SF network.

Mathematical Background

A network is represented as a graph consisting of a non-empty set of nodes and edges. For a given graph, $G = (V, E)$ indicates the relationships among the nodes V and edges E . In this research, the PPI network was represented as an undirected graph with no loop. A node in the PPI network is designated as a protein and an edge is designated as the relationship between 2 proteins.

A well-known property of a graph is the degree. The degree of node i (k_i) is the number of interactions between node i and other nodes in the network. If we consider the network's degree, called the average degree $\langle k \rangle$, the average degree is defined as $\langle k \rangle = \frac{\sum_{i=1}^N k_i}{N}$ in which N is the total number of nodes in the

network (Plaimas *et al.*, 2010; Province, 2015). The degree distribution $p(k)$ describes the probability of randomly choosing a node having degree k . The SF network can be observed by plotting the degree distribution in the y-axis against the degree in the x-axis in a logarithm scale. If the slope of the tendency plot is between 2 and 3, then the network has the property of the SF network. In the same manner, the SF network is followed by the γ of degree distribution in power-law form:

$$p(k) \sim k^{-\gamma}, \quad (1)$$

where $2 < \gamma < 3$ (Suratanee and Plaimas, 2014; Province, 2015).

The correlation measures consisting of the Point-Biserial (r_{pb}) correlation and Pearson correlation coefficient (r_{pcc}) were investigated. The Point-Biserial correlation considers the correlation between 1 continuous variable and 1 dichotomous variable which contains binary values. It is defined by:

$$r_{pb} = \frac{(\bar{X}_1 - \bar{X}_0) \sqrt{\Pi(1-\Pi)}}{S_x}, \quad (2)$$

where $\bar{X}_1, \bar{X}_0$ are the means of continuous variables that are a matching group of the first and second dichotomous variable, respectively, Π is the proportion of 1 dichotomous variable, and S_x is the standard deviation of the continuous variable. In this research, we performed the Point-Biserial measure to identify the correlation between the attribute $\frac{k_i}{<k>}$ and the classes. There are 2 classes (positive and negative classes) in this study; the positive class represented disordered proteins that affect the SF network and the negative class was disordered proteins that do not affect the SF network. Furthermore, the Pearson correlation coefficient is defined by the fraction between the covariance of variables x and y the multiple of the standard deviations of variables x and y . The variables x and y can be dichotomous variables. This can be rewritten by:

$$r_{pcc} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (3)$$

where n is the total list of the data. The value of r_{pcc} is between -1 and +1. There is a negatively linear correlation when the value of r_{pcc} is equal to -1. There is no linear correlation when the value of r_{pcc} equals to 0. There is a positively linear correlation when the value of r_{pcc} is equal to +1. In this study, r_{pcc} was employed to identify the correlation between the attribute $sign(R_i)$ and the classes.

The r_{pcc} analysis can be explained by the degree correlation. R measured with x_i, y_i is the degree of the connected nodes following the edge i , and n is the number of edges. The degree correlation indicates the tenor of the degree in the pair of connected nodes. If 2 connected nodes show a similar number of the degree, called an assortativity network, then $R > 0$. If 2 connected nodes show a different number of the degree, called a disassortativity network, then $R < 0$.

The characterization of a network can be determined by network features, e.g., the local clustering coefficient, the global clustering coefficient, and the eigenvector centrality. The local clustering coefficient (c_i) of node i identifies the probability that neighbors of randomly chosen nodes are connected (Soffer and Vázquez, 2005). It is defined by:

$$c_i = \frac{t_i}{\binom{k_i}{2}}, \quad (4)$$

where t_i is the number of neighbors of node i that are connected, and k_i is the degree of node i (Suratanee and Plaimas, 2014; Province, 2015). The network analysis of the local clustering coefficient is called the global clustering coefficient ($\langle c \rangle$), which is the average of the local clustering coefficient as follows:

$$\langle c \rangle = \frac{\sum_{i|k_i > 1} c_i}{N}, \quad (5)$$

where N is the number of nodes.

The eigenvector centrality identifies the important or influential nodes in the network. The centrality score of node i (X_i) is determined by the centrality of its neighbors, as follows:

$$X_i = \frac{1}{\lambda} \sum_{j \in N(i)} X_j, \quad (6)$$

where λ is a constant and $N(i)$ is all the neighbors of node i (Plaimas *et al.*, 2010). It can be rewritten by

$$X_i = \frac{1}{\lambda} \sum_{j=1}^N a_{i,j} X_j, \quad (7)$$

where $a_{i,j}$ is equal to 1 if node i and j are connected, otherwise it is equal to 0, and N is the number of nodes. According to the eigenvector equation following:

$$A\bar{X} = \lambda\bar{X}, \quad (8)$$

the eigenvector $\bar{X}$ satisfies many eigenvalues λ . The real positive number of the eigenvector centrality is considered. With the Perron Frobenius theorem, $\bar{X}$ is the eigenvector centrality satisfying the largest eigenvalue λ (Kosch and Schreiber, 2004; Bonacich, (2017).

Later on, we present the confusion matrix that visualizes the performance with each column showing the actual classes (positive/negative) and each row showing the predicted classes. The confusion matrix identifies the true positive, false negative, false positive, and the true negative for a cut-off threshold of calculated values from the model. For the threshold, the accuracy, precision, and recall were calculated. The performance of our measure can be calculated by the AUC when the cut-off threshold is varied over the range of values of the developed model. The AUC is the area of the ROC curve. The ROC is the graph of the false positive rate in the x-axis and the recall (true positive rate) in the y-axis visualizing the performance of the model in the varied threshold (Plaimas *et al.*, 2010). The value of the AUC is between 0 and 1. An ideal performance presents the value of the AUC equal to 1.

With our investigation, we found that the number of disordered proteins that affect an SF network's property (positive class) is very different from the number of disordered proteins that do not affect an SF network's property (negative class). To solve this imbalanced data, we employed the method of the synthetic minority over-sampling technique (SMOTE). The algorithm of the SMOTE method is to

synthesize the positive instances by determining the original positive instances (Chawla *et al.*, 2002).

Methodology

Investigation of the Effect of Disordered Protein in an SF Network

To investigate the impact of disordered proteins in the human PPI network with the property of the SF network, all of the disordered proteins from the OMIM database were discarded from the human PPI network. Then, the slope of the degree distribution or *gamma* in power-law form was calculated. Later, the property of the SF in the human PPI network was examined using the value *gamma* in power-law form of the degree distribution. If *gamma* is between 2 and 3, then the human PPI network is the SF network. On the other hand, if the value of *gamma* is not in the range of 2 to 3, then the human PPI network lost the property of the SF network. This situation can imply that disordered proteins affect the property of the SF network. Finally, in this work, the number of proteins in the network is 8,208 proteins and the number of disordered proteins is 2,335 proteins. The number of disordered proteins is quite large compared to the number of all the analyzed proteins.

To determine the effect of the disordered proteins in an SF network, we performed 2 experiments by randomly removing proteins and disordered proteins and then the performances of these experiments were compared. These random experiments were performed 100 times and the results were presented in a scattering plot with the percentage of the number of removed proteins (random proteins and random disordered proteins) in the x-axis and the percentage of the number of the human PPI network lost from the property of the SF network in the y-axis.

Analysis of the Network Properties

Firstly, the human PPI network was reconstructed with the property of the SF network. We discarded a protein one by one to find proteins affecting the property of the SF

network. If we discarded a protein and its neighbors that had a higher eigenvector centrality than the protein and the value of *gamma* disobeyed the power-law form degree distribution, then we inferred that the protein affects the property of the SF network. The eigenvector centrality was determined for reducing the influence of the disordered proteins in the network. Then the properties of each protein were investigated. The properties were calculated by a degree of a protein divided by the average degree $\frac{k_i}{\langle k \rangle}$, the local clustering coefficient divided by the global clustering coefficient $\frac{c_i}{\langle c \rangle}$, and the value sign of the degree correlation $sign(R_i)$. These 3 properties were used to construct the model of our simple measurement. The first attribute, $\frac{k_i}{\langle k \rangle}$, indicates the number of interactions with a comparison in the network. If the value $\frac{k_i}{\langle k \rangle}$ is greater than 1, then the protein *i* has a high interaction. Therefore, considering the attribute $\frac{k_i}{\langle k \rangle}$ is more efficient than considering only the degree (k_i). The second attribute is $\frac{c_i}{\langle c \rangle}$ identifying the groups of proteins that were connected. The last attribute is $sign(R_i)$. It is used to consider the tenor of the interaction of any 2 connected proteins in the network when protein *i* and some of its neighbors, that had a higher value of eigenvector centrality than protein *i*, were removed from the network. The degree correlation R_i can be computed by discarding protein *i* and its neighbors and calculating the degree of correlation *R*. If the value of R_i is greater than 0, then 2 connected proteins tend to be close to the number of interactions.

Development of the New Measure

We developed a new measure for identifying disordered proteins affecting the SF network, $M_{SF,Disp}$. There are 2 processes to generate our measure. Firstly, 2 influential attributes were selected from all 3 attributes, $\frac{k_i}{\langle k \rangle}$, $\frac{c_i}{\langle c \rangle}$, and $sign(R_i)$, using the correlation measurement. We chose 2 influential attributes with high values of correlation between the attributes and the classes. The correlation measurements depend on the types of

variables, e.g., r_{pcc} and r_{pb} (see section headed Mathematical Background). The 2 selected attributes were combined as our measure. The coefficients (or weights) of these 2 attributes were calculated by the correlations between the attributes and the class of disordered proteins affecting the SF network divided by the minimum value of the correlations between the attributes and the class. The format of the measurement for identifying the class of disordered proteins affecting the SF network, $M_{SF,Disp}$, can be performed in Equation (9):

$$M_{SF,Disp} = w_1x_1 + w_2x_2, \quad (9)$$

where x_1 and x_2 are the attributes and w_1 and w_2 are the weights or coefficients, with $w_i = \frac{cor(x_i, class)}{\min_{j \in \{1,2\}} cor(x_j, class)}$.

An optimal threshold was selected by considering the cut-off of the calculated measured value yielding high precision to identify the score of the class of the disordered proteins affecting the property of the SF network. Then, the performance of our measurements was calculated using the value of the AUC.

Results

The Characteristics of the Human PPI Network

The human PPI network was reconstructed and its characteristic properties, such as the number of nodes, the number of edges, the average degree, the global clustering coefficient, the slope or *gamma* in power-law form of the degree distribution, and the correlation of degree, are shown in Table 1. The analysis of our human PPI network characteristics showed that there are 8,208 proteins and 45,553 edges. There are a lot of low-degree nodes with an average degree of 11. The probability of connected neighbors is explained in the global clustering coefficient of 0.29. The small value of the degree correlation of -0.01 was presented. Moreover, the slope of the degree distribution or *gamma* in power-law form was 2.75 (between 2 and 3). Obviously, our human PPI network has the property of the SF network. The slope

Table 1. Basic network characteristics

Species	Characteristics of properties					
	#Nodes	#Edges	$\langle k \rangle$	$\langle c \rangle$	γ	R
Homo Sapiens	8,208	45,553	11.10	0.29	2.75	-0.01

Table 2. The number of nodes, edges and parameter gamma of original and mutated network

Network	Characteristics of properties		
	#Nodes	#Edges	Gamma
Original human SF network	8,208	45,553	2.75
Mutated network after removing disordered proteins	5,873	21,698	6.26

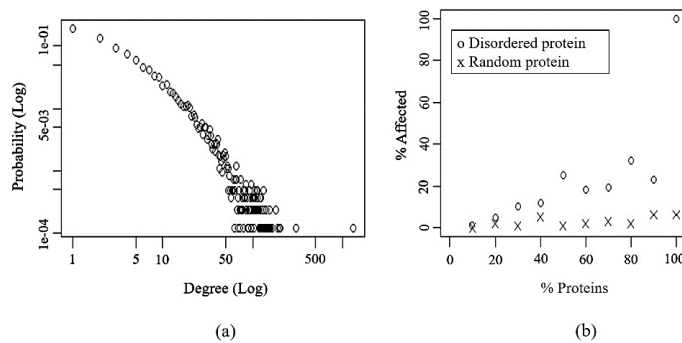


Figure 1. a) The slope of the degree distribution in the logarithm scale and b) the comparison of knocking out the disordered proteins and random proteins in various ranges

of degree distribution in the logarithm scale is shown in Figure 1(a). The graph shows a lot of low-degree proteins and a small number of high-degree proteins.

The Effect of Disordered Proteins in SF Network

The human PPI network lost the property of the SF network when all of the 2,335 disordered proteins were removed from the network. This can be observed by the slope of the degree distribution or *gamma* in power-law form of about 6.26 compared with the original network of about 2.75, as shown in Table 2. Therefore, the loss of the disordered proteins has an effect on the SF network property. However, this might have occurred because of the large number of disordered proteins that were removed from the network. To be fair, the

impact of the disordered proteins on an SF network was investigated by randomly removing proteins with the same number of disordered proteins and also by calculating the *gamma* of the power-law form. The random removal experiments were performed 100 times. The percentages of the impacts are illustrated in a scattering plot as shown in Figure 1(b). It shows that most of the disordered proteins removed from the network affected the SF network property. On the other hand, the disordered proteins are more significant in the property of the SF network.

The Performance of New Measure

Two influential attributes ($\frac{k_i}{\langle k \rangle}$ and $sign(R_i)$) were investigated and found to be related to the measure of the disordered

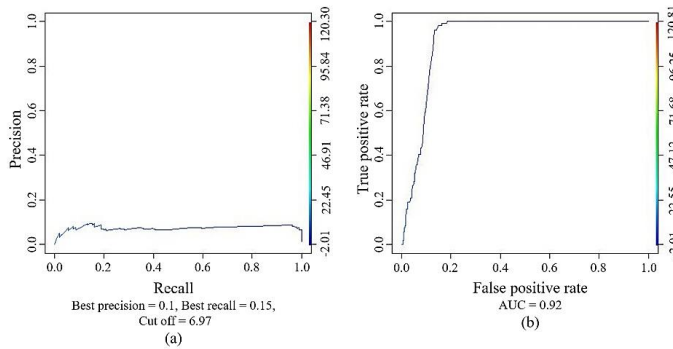
Table 3. The number of proteins relating to class imbalance

Class imbalance data	Binary	The number of proteins
Disordered proteins affecting SF network property	Disordered proteins affecting SF network (Class 1)	106
	Otherwise (Class 0)	8,102

Table 4. a) The confusion matrix of $M_{SF.Disp}$ in the imbalanced data and b) the confusion matrix of $M_{SF.Disp}$ in the balanced data

(a) Confusion matrix			
Predicted classes	Actual classes		
	1	0	
	1	16	152
	0	90	7,950

(b) Confusion matrix			
Predicted classes	Actual classes		
	1	0	
	1	7,824	1,004
	0	278	7,098

**Figure 2. a) The precision-recall and b) the ROC curve of $M_{SF.Disp}$ in the imbalanced data**

proteins affecting the SF network. The class of the disordered proteins affecting the SF network (class 1) consists of 106 proteins and otherwise (class 0) consists of 8,102 proteins. This situation is concerned with the imbalanced dataset, as shown in Table 3. Avoiding the problem of the imbalanced dataset, the method of SMOTE is used. The SMOTE method adjusts the number of proteins in class 1 and class 0 to be equal to obtain a balanced dataset.

The measure of $M_{SF.Disp}$ was generated by 2 influential attributes; $\frac{k_i}{\langle k \rangle}$ and $sign(R_i)$ using correlations. The correlations r_{pb} and r_{pcc} were also calculated for the weight of the measure. The measure $M_{SF.Disp}$ of the imbalanced data can be rewritten in Equation (10):

$$M_{SF.Disp}(i) = \frac{k_i}{\langle k \rangle} + 2.10 \cdot sign(R_i). \quad (10)$$

The threshold yielding the highest precision was 6.97 that yielded an accuracy of 97%, a precision of 10%, and recall of 15%, as shown in Figure 2(a). The confusion matrix of the measure $M_{SF.Disp}$ is shown in Table 4(a). In addition, the performance of the measure $M_{SF.Disp}$ in the imbalanced data can be observed by the value of the AUC in the ROC curve as shown in Figure 2(b). It should be noticed that the value of the AUC was 0.92 which in the high value of AUC was better than in the random selection.

Moreover, the SMOTE method was used to obtain balanced data. The measure of the

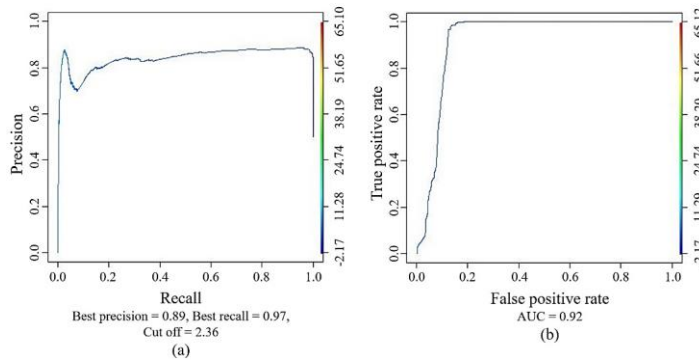


Figure 3. a) The precision-recall and b) the ROC curve of $M_{SF.Disp}$ in the balanced data

balanced data $M_{SF.Disp}$ can be rewritten in Equation (11):

$$M_{SF.Disp}(i) = \frac{k_i}{\langle k \rangle} + 2.22 \cdot \text{sign}(R_i). \quad (11)$$

Selecting the threshold at 2.36, we obtained the best result with an accuracy of 92%, a precision of 89%, and recall of 97%, as shown in Figure 3(a). With this balanced data, the classification result with our developed measures was presented in the confusion matrix, as shown in Table 4(b). The ROC curve yielded the value for the AUC of 0.92 as shown in Figure 3(b). It showed that the performance of $M_{SF.Disp}$ is good enough for classification between the disordered and non-disordered proteins.

We found that examining the measure from the training data is unfair (by a self-consistency test). Therefore, we constructed our measure with the balanced data from the SMOTE method and validated it by the imbalanced data.

With this process, we obtained the high value for the AUC of 0.92 with the threshold at 2.36 for the measure $M_{SF.Disp}(i) = \frac{k_i}{\langle k \rangle} + 2.22 \cdot \text{sign}(R_i)$. The ROC curve is shown in Figure 4.

Validation of the Measure

To validate our developed measure, we performed random experiments by shuffling only the class label but with the attributes still

the same. We compared the measure $M_{SF.Disp}$ in the group of balanced data and imbalanced data with the same weight, influential attributes, and threshold. The results are shown in Figure 5. They consist of (a) the AUC of the ROC curve of our measure $M_{SF.Disp}$ in the imbalanced data of 0.92, (b) the AUC of the random class label measure $M_{SF.Disp}$ in the imbalanced data of 0.51, (c) the AUC of our measure $M_{SF.Disp}$ in the balanced data of 0.92, and (d) the AUC of the random class label measure $M_{SF.Disp}$ in the balanced data of 0.50.

We found that the values of the AUC in our measure (on the left) are greater than in the random class label measure (on the right). Therefore, the performances of our measure are better than the performances of the random experiments in both the balanced and imbalanced data.

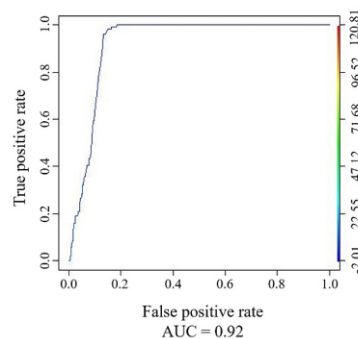


Figure 4. The ROC curve of our measure $M_{SF.Disp}$

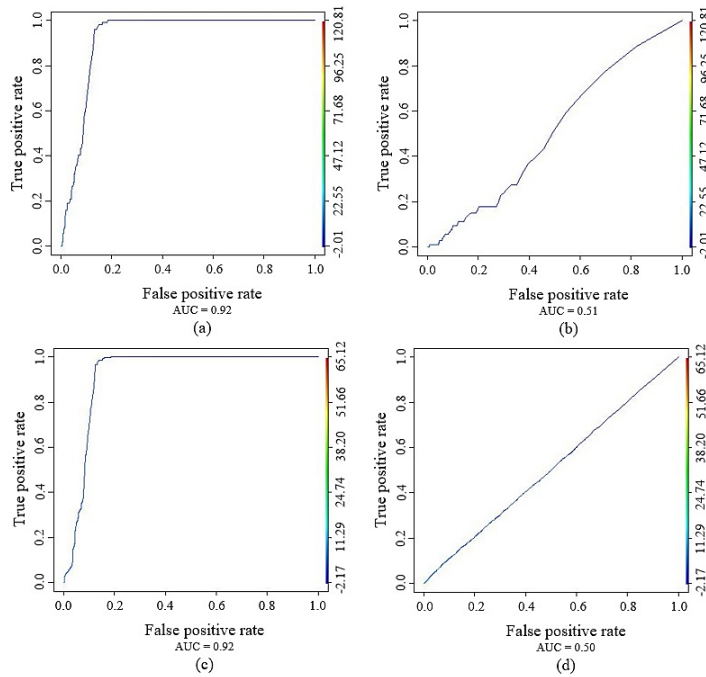


Figure 5. a) The ROC curve of our measure in the imbalanced data; b) the ROC curve of the random class label measure in the imbalanced data; c) the ROC curve of our measure in the balanced data; and d) the ROC curve of the random class label measure in the balanced data

Conclusion and Discussion

In this study, a large-scale human PPI network was constructed and the network properties, such as the average degree, global clustering coefficient, and degree of correlation, were investigated for developing a measure. Moreover, we analyzed the γ in power-law form of the degree distribution to identify the property of the SF network. The impact of the disordered proteins in the property of the SF network was analyzed and a new simple measure, $M_{SF.Disp}$, was developed in this study. We found that removing the disordered proteins from the network provides a significantly different impact on the SF network property in comparison to removing any proteins with the same number from the network. In addition, we found that the $\frac{k_i}{\langle k \rangle}$ and $sign(R_i)$ attributes were

significantly related to the property of the disordered proteins affecting the SF network property by comparing with the random class label. Thus, these attributes were chosen for developing the new measure of $M_{SF.Disp}$. Using $M_{SF.Disp}$, we obtained a high AUC value of 0.92 for identifying disordered proteins affecting the SF network property. We also performed experiments by selecting other pairs of attributes, for example, $\frac{k_i}{\langle k \rangle}$ and $\frac{c_i}{\langle c \rangle}$, to develop the measure. The results showed an inferior performance with the AUC of 0.72 in comparison to our measure.

In conclusion, we showed that disordered proteins affect the scale-free structure of the PPI network and our developed measure is simple and efficient to identify novel disordered proteins and to investigate the degree distribution.

Acknowledgment

This research is supported by the Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Thailand, under grant number: 5942103.

References

- Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, 43:789-798.
- Bonacich, P. (2017). Power and centrality: a family of measures. *Am. J. Sociol.*, 92(5):1,170-1,182.
- Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. (2002). SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, (16):321-357.
- Haynes, C., Oldfield, C.J., Ji, F., Klitgord, N., Cusick, M.E., Radivojac, P., Uversky, V.N., Vidal, M., and Iakoucheva, L.M. (2006). Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput. Biol.*, 2(8):0890-0901.
- Kosch, D. and Schreiber, F. (2004). Comparison of centralities for biological network. *Proc. Ger. Conf. Bioinform.*, p. 199–206.
- Li, J., Feng, Y., Wang, X., Liu, W., Rong, L., and Bao, J. (2015). An overview of predictors for intrinsically disordered proteins over 2010-2014. *Int. J. Mol. Sci.*, 16(1):23,446-23,462.
- Plaimas, K., Eils, R., and König, R. (2010). Identifying essential genes in bacterial metabolic networks with machine learning methods. *BMC Syst. Biol.*, 4(4):56.
- Province, J. (2015). Yeast protein-protein interaction network model based on biological experimental data. *Appl. Math. Mech.*, 36:827-834.
- Sickmeier, M., Hamilton, J.A., LeGall, T., Vacic, V., Cortese, M.S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V.N., Obradovic, Z., and Dunker, A.K. (2007). DisProt: The database of disordered proteins. *Nucleic Acids Res.*, 35(Database Issue):D786-D793.
- Soffer, S.N. and Vázquez, A. (2005). Network clustering coefficient without degree-correlation biases. *Phys. Rev. E*, 71(5):2-5.
- Suratane, A. and Plaimas, K. (2014). Identification of inflammatory bowel disease-related proteins using a reverse k-nearest neighbor search. *J. Bioinf. Comput. Biol.*, 12(4):1450017.
- Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., Kuhn, M., Bork, P., Jensen, L.J., and Von Mering, C. (2015). STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, 43(D1):D447-D452.