CHAPTER 2
BACKGROUND KNOWLEDGE

2.1 Biological background

2.1.1 Disordered proteins

Disordered proteins are three-dimensional structures of proteins, which lack some unique in form of the structures [2]. For example, a part of polypeptide chain, combination of amino acid residues bonded together in form of a protein molecule, is not folding. There is a current study in predictors to predict the disordered protein such as VL3 [4], [7]. In 2015, there was a research about collecting and discussing the predictor of disordered proteins in 2010-2014 [3], which are well-known in bioinformatics for many years. It can be concluded that disordered proteins make some crucial diseases such as cancer, cardiovascular disease, neurodegenerative disease and amyloidosis [3]. Thus, disordered proteins are related to the human crucial diseases.

2.1.2 Protein-protein interaction network

Several protein-protein interaction (PPI) networks, which characterize the relationship between proteins, can be represented as graphs. Another implication is the edge which displays that two proteins are related. On the other hand, PPI network is described the physical interaction between proteins. The protein consists of the amino acid residues establishing in polypeptide chain with physical interaction which can be described by confident of the combined score. One interesting studies is characterizing the features of network such as scale-free network and disassortativity network. Many biological of PPI networks have the properties of scale-free network and disassortativity network [8].

2.2 Mathematical background

2.2.1 Graph and its definition

In mathematics, networks represent in graph consisting the set of non-empty nodes and set of edges. Let a graph $G=(V,E)$ characterizes the association between nodes V with representing in edges E. There are many types of graph such as undirected graph, directed graph and simple graph. The undirected graph is the graph that edges of pair nodes have no orientation. Another implication is the edge displaying that two nodes are related. In contrast, the directed graph is the graph of any edges having the direction, indegree and outdegree. The simple graph is the graph that contains the properties of no duplicated edges or no loop, which is edge connecting to itself of each node.

2.2.2 Degree distribution and power-law form

In analysis of PPI network, we characterized the number of connections in each protein. The degree of node i (k_i) is the number of edges connected to it, that is the number of connections that the node i connects to other nodes. If we considered the degree of the network, we used the average degree, $<k>$. We describe the average degree by $<k>=\frac{\sum_{i=1}^{N}k_i}{N}$ when N is the total number of nodes [5], [8].

Moreover, we considered the degree distribution, $p(k)$ which is the chance that we choose randomly the node having degree k. We plotted the degree distribution, $p(k)$ in y-axis and degree, k in x-axis in form of logarithm scale. Later, we determined the slope or gamma γ. If gamma is in the range of 2 and 3, then we can conclude that this network has the property of scale-free network. Furthermore, scale-free network displays the meaning that there are a few number of high-degree nodes. In contrast, there are high number of low-degree nodes. We can describe that the scale-free network observes the small number of high-degree nodes, that are hub nodes.

On the other hand, scale-free network is the network that follows the parameter gamma of a degree distribution in power-law form

$$p(k) \sim k^{-\gamma}, \tag{2.1}$$

where $2 < \gamma < 3$ [8], [9].

In this thesis, we attempted to figure out the nodes that affect to the property of scale-free network by developing the new measure from the properties of nodes in the network.

2.2.3 Correlation measure

The most famous correlation measure to identify the correlation of two variables is Pearson correlation coefficient (PCC). PCC is a correlation identifying a linear relationship between variable x and variable y. The definition of PCC is the proportion between covariance of variables x, y by the product of two standard deviations of x and y [4], as follows,

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{2.2}$$

where n is the total number of data.

The value of PCC is in the range of -1 and +1. If the value PCC equals to -1, it means that there is negatively linear correlation. If the value of PCC equals to 0, it means that there is no linear correlation. Besides, if the value of PCC equals to +1, it means that there is positively linear correlation. Figure 2.1 shows the plots of PCC in various ranges.

2.2.4 Degree correlations

Degree of correlation (R) is the measure of Pearson correlation coefficient (PCC) to identify the correlation of degree in a pair of connected nodes [8]. The measure of degree correlation (R) was explained by using PCC equation, that means x_i, y_i are the number of degrees in a pair of connected nodes with the edge i and n is the total number of edges in the network.

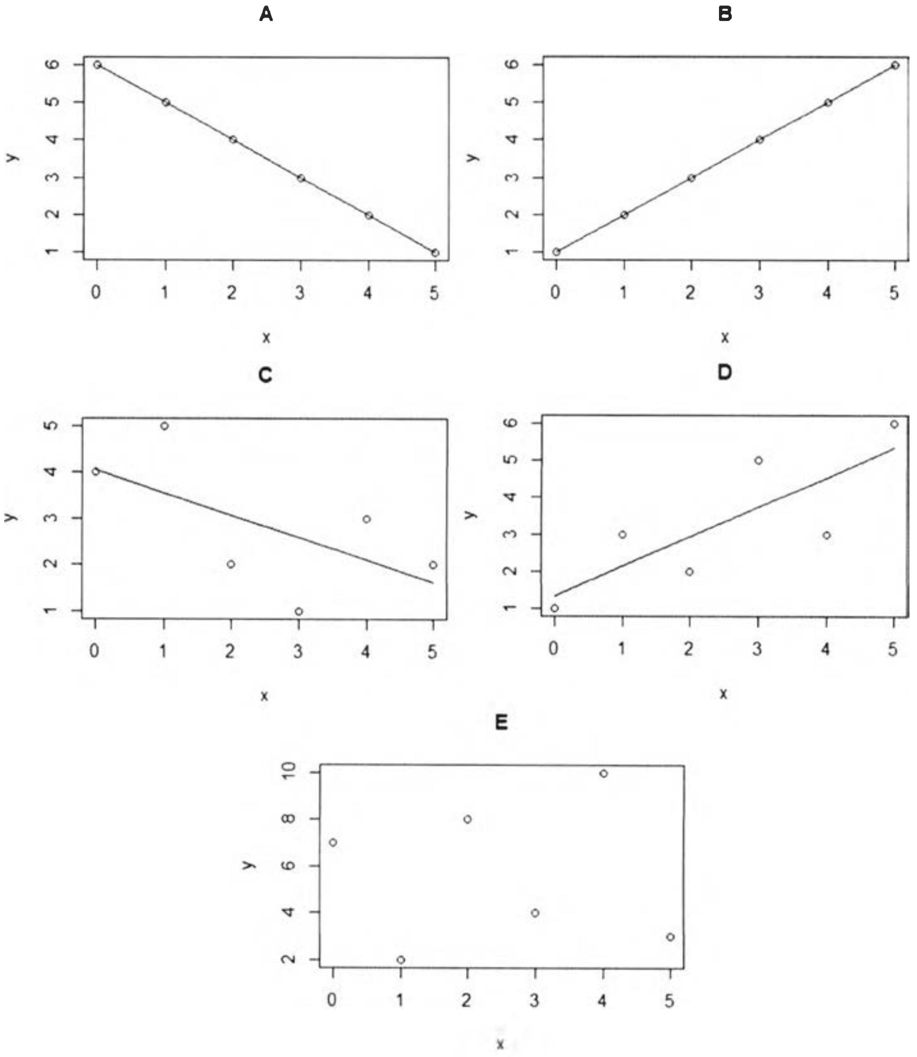

Figure 2.1 The graphs of several ranges in the value of Pearson correlation coefficient (PCC): (A) The graph of PCC equals to -1, (B) The graph of PCC equals to +1, (C) The graph of PCC equals to -0.6173321, (D) The graph of PCC equals to 0.803837 and (E) The graph of PCC equals to 0

Moreover, the interesting property of PPI network is disassortativity network. It represents the connection approach between a pair of proteins in the network with analysis by the measure of degree of correlation (R). If $R > 0$, then the network is assortativity, that means high-degree of node connects to high-degree of node. If $R < 0$, then the network is disassortativity, that means high-degree of node connects to low-degree of node. If $R = 0$, then the network is independent. The assortativity network identifies that two connected nodes are likely similar number of interactions. The disassortativity network shows the two connected nodes in the network are likely different number of interactions. This type of network characterizes the correlation between two connected nodes in the network [8].

2.2.5 Interesting node properties

There is fundamental characterizing to identify the type of network by using the properties of each nodes in the network such as clustering coefficient, global clustering coefficient and eigenvector centrality.

First of all, we mentioned the property of PPI network by the measure of clustering coefficient. The clustering coefficient of node i (c_i) is the measure identifying the probability that we choose randomly node having the interaction between neighbors. The clustering coefficient can be expressed as the fraction of the total number of edges between neighbors of node i and the possible number of edges between neighbors of node i. It is defined by

$$c_i = \frac{t_i}{\binom{k_i}{2}}, \tag{2.3}$$

where t_i is the total number of edges between neighbors of node i and k_i is the number of degree of node i [8], [10].

If we considered the clustering coefficient of the network, then we used the average of clustering coefficient. We described as the global clustering coefficient. It is defined by

$$< c >= \frac{\sum_{i|k>1}^{N} c_i}{N}, \tag{2.4}$$

where c_i is the clustering coefficient of node i and N is the number of nodes.

Next, eigenvector centrality is the measure identifying the influential node, that is the important node in the network. The X_i is the score of centrality in node i, by considering the centrality of all its neighbors. That means the relative centrality score of node is the proportional to the summation of all centrality score of its neighbors. In Mathematics, we can describe as

$$X_i = \frac{1}{\lambda} \sum_{j \in N(i)} X_j, \tag{2.5}$$

where $N(i)$ is the group of neighbors of node i and λ is a constant [5].

Or, it can be rewritten in form of

$$X_i = \frac{1}{\lambda} \sum_{j=1}^{N} a_{i,j} X_j, \tag{2.6}$$

where $a_{i,j}$ equals to 1 if node i connects to node j, equals to 0 otherwise and N is the total number of nodes. In Mathematics, it can be rewritten as an eigenvector equation

$$A\overline{X} = \lambda \overline{X}, \tag{2.7}$$

In general, there are many eigenvalues λ satisfying an eigenvector solution. In this case, the positive real number of eigenvector is considered [11]. By the Perron-Frobenius theorem, there exists the positive real number of eigenvector associated with the maximal eigenvalue for a non-negative matrix. Thus, $\overline{X}$ is the eigenvector centrality with the largest eigenvalue λ [12].

2.3 Performance measure

2.3.1 Accuracy, precision and recall

A confusion matrix or an error matrix is a table that visualizes the performance of model. In confusion matrix, each column indicates the actual classes of positive/negative and each row indicates the predicted classes of positive/negative. We performed the confusion matrix to easily understood the confusing of two classes.

A confusion matrix reports the number of true positive (*TP*), the number of false positive (*FP*), the number of false negative (*FN*) and the number of true negative (*TN*). In addition, we analyzed the value of accuracy, recall or sensitivity or true positive rate (*TPR*), precision, specificity or true negative rate (*TNR*), false positive rate (*FPR*) and F-score (*F*) from the confusion matrix for the binary classifier as described in Table 2.1.

Table 2.1 The analysis of confusion matrix

Confusion matrix	Actual classes		
Predicted classes		1	0
	1	*TP*	*FP*
	0	*FN*	*TN*

Accuracy is the measure of validity of two classes. It can be defined by

$$\text{accuracy} = \frac{(TP + TN)}{(TP + FP + FN + TN)}. \tag{2.8}$$

Recall or sensitivity or true positive rate (*TPR*) is the measure of validity of predicted positive class in actual positive class. It can be defined by

$$\text{recall} = \frac{TP}{(TP + FN)}. \tag{2.9}$$

Precision is the measure of validity of actual positive class in predicted positive class. It can be defined by

$$\text{precision} = \frac{TP}{(TP + FP)}. \tag{2.10}$$

Specificity or true negative rate (TNR) is the measure of validity of predicted negative class in actual negative class. It can be defined by

$$TNR = \frac{TN}{(FP + TN)}. \tag{2.11}$$

False positive rate (FPR) is the measure of invalidity of predicted positive class in actual negative class. It can be defined by

$$FPR = \frac{FP}{(FP + TN)}. \tag{2.12}$$

F-score (F) is the measure of computing both of the precision and recall. It can be defined by

$$F = 2\frac{\text{precision} \cdot \text{recall}}{(\text{precision} + \text{recall})}. \tag{2.13}$$

2.3.2 ROC curve and AUC

Receiver Operating Characteristic (ROC) curve is the graph plotting between true positive rate (TPR) and false positive rate (FPR). The ROC curve analyzes the performance of model in various threshold [5], to visualize the performance of a binary classifier [5]. In addition, we investigated the Area Under the Curve (AUC) of the ROC curve to estimate and visualize the performance of model. The value of AUC is in the range of 0 and 1. The case of AUC equals to 1, that means the classifier is perfect performance.

The imbalance dataset is the number of group in positive or minority instances and the number of group in negative or majority instances are very different [6]. There are some methods to filter the imbalance data to balance data. One of the most popular methods is Synthetic Minority Over-sampling (SMOTE). The SMOTE is the method to synthesize the positive instances, by considering only the positive instances [6]. That means we choose randomly one of *k* nearest neighbors, the value *k* is assigned by user. In general, *k* is 5 and then the synthetic positive instance is generated between the distance of the chosen neighbor and the positive instance. The synthetic positive instances are generated and then the number of positive instances are closed to the number of negative instances. Therefore, we get the balance data from the SMOTE method. There are some researches about the imbalance dataset. They used the AUC to evaluate the performance of model [13].