CHAPTER 3
DATA AND METHODS

3.1 Data

3.1.1 Protein-protein interaction network database

In the first part, we constructed the network by gathering the real human (*Homo sapiens*) proteins interactions from STRING database (version 10) [14]. Only human PPI interactions with high confident of combined score 950 were selected. With this data, we obtained the reconstructed network which contains 8,208 nodes and 45,553 edges. The slope or gamma in power-law form of degree distribution of approximately 2.75. This value showed that the reconstructed PPI network has the property of scale-free network.

3.1.2 Disordered proteins

In this work, we collected the list of disordered proteins from Online Mendelian Inheritance in Man (OMIM) database. OMIM database consists of human genes and genetic disordered by analyzing the phenotype in each chromosome [15]–[17]. In general, there are two steps, transcription and translation, for changing from gene to protein. The detail of gene into protein, first gene's DNA transfers to messenger ribonucleic acid (mRNA) in transcription process and then changes to protein in the process of translation. Thus, we can interpret disordered protein from disorder of gene in OMIM database by mapping gene's name to protein's name. The class of property disordered protein in each node, the case of node that is disordered protein, then the value is 1 and the case of node that is non-disordered protein, then the value is 0.

3.2 Methods

3.2.1 Overview

In this section, we proposed the information of thesis work's process, step by step. It was divided into four parts. First is analysis of network properties, to characterize

the properties of each node in the network. Second is degree distributions for identifying the nodes that affect to the property of scale-free network. Next, it is the development of our new measures, the measure of disordered proteins affecting to the property of scale-free network and the measure of proteins affecting to scale-free network. The last part is the investigation of the association of disordered proteins in scale-free network as illustrate in Figure 3.1.

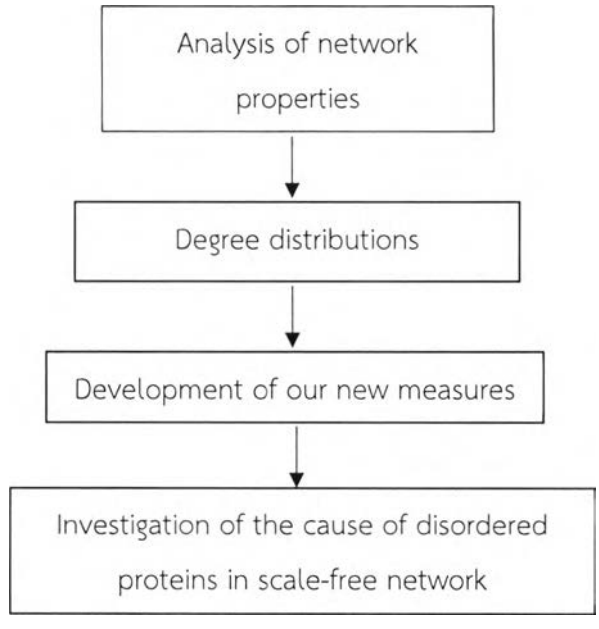

Figure 3.1 The flow chart of proposed methodology

3.2.2 Analysis of network properties

In the beginning, we constructed the connected human (*Homo sapiens*) protein-protein interaction network with the property of scale-free. Then, we analyzed the attributes of each node in the network. First, the interesting attribute was degree of protein i divided by the average degree, $\frac{k_i}{<k>}$ to identify the number of interaction between proteins which compared with the interaction of all proteins in the network. If the protein was high interaction, then the attribute $\frac{k_i}{<k>}$ was greater than 1, otherwise, less than 1. Thus, the attribute $\frac{k_i}{<k>}$ was more efficient than the attribute degree. Second, it was the clustering coefficient of protein i divided by the

global clustering coefficient, $\frac{C_i}{< C >}$ to consider the group of interaction in proteins, compared with all proteins in the network and the last was the sign of the degree correlation in protein i, $sign(R_i)$ to determine the correlation of interaction in two connected nodes. Considering the value degree correlation of the protein i, it was calculated by cutting the protein i and some neighbors of protein i, that were higher in value of eigenvector centrality than protein i and then computed the degree of correlation in the network. The attribute degree of correlation indicated the association of degree between connected nodes. If the attribute degree of correlation was greater than 0, then two connected nodes tended to have the same number of degree interactivity. Otherwise, two connected nodes tended to have the different number of degree interactivity. Thus, we would investigate the attribute sign of degree correlation replacing the attribute degree correlation.

Moreover, the property of disordered protein interprets from OMIM database. If protein i is disordered protein, then it has value 1. Otherwise, it has value 0.

3.2.3 Degree distributions

Degree distribution, $p(k)$ is a function to identify the character of degree in the network. The value of $p(k)$ is calculated from the proportion between the total number of nodes that have k interactions with other nodes and divided by the total number of nodes. On the other hand, the $p(k)$ is the chance to choose randomly the node in the network and has exactly k interactions. The property of scale-free network can be explained by using the analysis of degree distribution. The case of the degree distribution of $k1$ interactions, $p(k1)$, is greater than the degree distribution of $k2$ interactions, $p(k2)$, when $k1$ is less than $k2$, then this case identifies the property of scale-free network. That means there are a lot of low-interaction nodes and there are a few of high-interaction nodes. In general, if the degree distribution $p(k)$ observes the power-law form $p(k) \sim k^{-\gamma}$ where $2 < \gamma < 3$, then the network has the property of scale-free. In analysis of power-law form $p(k) = c \cdot k^{-\gamma}$, where c is a constant, and then we get

$$\log(P(k)) = -\gamma \log(k) + \log(c). \quad (3.1)$$

It can be described as the linear equation in the form of $y = mx + c$ which slope is $-\gamma$. After that, we plotted the degree distribution, $p(k)$ in y-axis against the degree in x-axis in logarithm scale and investigated the slope of this graph by fitting the least square method. The case of slope γ is between 2 and 3, then we get the network has the property of scale-free network, otherwise, the network is not a scale-free. The graph of degree distribution can be explained that the value degree increases while the value degree distribution decreases, with according to the negative slope, $-\gamma$.

In this work, we investigated the node that is important to the property of scale-free network by discarding each node, one-by-one and determined the eigenvector centrality of its neighbors. We also discarded its neighbors that were higher the value of eigenvector centrality than the determining node to reduce the influential of the centrality property in each node. After that, we investigated the degree distribution following the gamma in power-law form. In the case of discarding the determining node and some neighbors of the determining node from the scale-free network, we investigated the gamma, which does not follow the power-law form degree distribution. Then, the determining node which was discarded, was called the node that was important to the property of scale-free. That means the determining node affects to scale-free network. On the contrary, the case of discarding the determining node and some neighbors of the determining node from the scale-free network, we investigated the gamma, which follows the power-law form degree distribution. Then, the determining node which was discarded, was called the node that was not important to the property of scale-free network. That means the determining node does not affect to scale-free network. The class of property scale-free network, the case of node that affected to scale-free network, then the value was 1, and otherwise was 0. Furthermore, the class of property that disordered protein in scale-free network, the case of disordered protein that affected to scale-free network, then the value was 1, and otherwise was 0.

3.2.4 Development of our new measures

In this section, we propose the method of Pearson correlation coefficient (PCC) to develop new measures. The first one is the measure of disordered proteins that affect to scale-free network, denoted as M_{SF_Disp}. The second one is the measure of proteins that affect to scale-free network, denoted as M_{SF}. In the beginning, we investigated the appropriate influential attributes to develop new measure of M_{SF_Disp}. The attributes were $\frac{k_i}{<k>}$, $\frac{c_i}{<c>}$ and $sign(R_i)$ of each node in the network. We figured out the two influential attributes that were related to class of property that disordered protein affecting to scale-free network. We used the technique of calculating the correlation measure between attributes and class of property that disordered protein affecting to scale-free network. For calculating the correlation measure, we used the PCC analysis to choose two influential attributes with high the value of PCC. For this step, we got the two influential attributes: $\frac{k_i}{<k>}$ and $sign(R_i)$ that were related to the property of disordered protein affecting to scale-free network.

After that, we investigated the weight or coefficients of the two influential attributes by using the proportion of value in PCC between attribute and class of property that disordered protein affecting to scale-free network divided by the minimum of value in PCC between these attributes and class of property that disordered protein affecting to scale-free network. If the case of PCC between attribute x_1 and class of disordered protein affecting to scale-free network, we defined as $PCC(x_1, \text{class})$ was greater than the $PCC(x_2, \text{class})$, then the coefficient of attribute x_1 was $PCC(x_1, \text{class})$ divided by the $PCC(x_2, \text{class})$ and the coefficient of attribute x_2 equaled to one. Otherwise, the coefficient of attribute x_2 was the proportion of the $PCC(x_2, \text{class})$ divided by the $PCC(x_1, \text{class})$, and the coefficient of attribute x_1 equaled to one. The meaning of this process showed that the correlation between attributes and the class of disordered protein affecting to scale-free network. The correlation was computable by using the proportion of PCC, and also can be

represented by coefficient or weight. This process, we generated the new measure of disordered proteins that affect to scale-free network, M_{SF_Disp}. It can be defined by

$$M_{SF\_Disp} = w_1 x_1 + w_2 x_2, \quad (3.2)$$

where w_1 and w_2 were coefficients of attributes x_1 and x_2 respectively, with $w_i = \frac{PCC(x_i, \text{class})}{\min(PCC(x_i, \text{class}))}$.

Later on, we chose the suitable cutoff to identify the score of disordered proteins that affect to scale-free network. We chose the cutoff that yielded high values of precision for more efficient of predicting disordered proteins that affect to scale-free network. Furthermore, we applied the method of correlation measure, PCC to develop the measure of proteins that affect to scale-free network, M_{SF}. We used the same manner for investigating the two influential attributes that were related to class of property scale-free network. In addition, we investigated the coefficients of two influential attributes by using the same technique of developing new measure of M_{SF_Disp}.

3.2.5 Investigation of the cause of disordered proteins in scale-free network

In this part, we figured out the association of disordered proteins in scale-free network by discarding all nodes which were disordered proteins from OMIM database. After we eliminated all disordered proteins from the human scale-free network, we determined the architecture of scale-free network by using the gamma in power-law form degree distribution. If the gamma value did not follow the power-law form (gamma was less than or equal to 2 or gamma was greater than or equal to 3), then the human PPI network lost the property of scale-free. If all of disordered proteins were cut from scale-free network and then network lost the property of scale-free. This case can be explained that disordered proteins were important to the scale-free network. That means they affected to the property of scale-free.

Moreover, we compared the case of randomly choosing the disordered proteins and the case of randomly choosing proteins to discard from the human scale-free network for clearly understanding the behavior of disordered proteins in the property of scale-free network. We discarded randomly two cases of proteins in 100 times, for identifying the number of times that network lost the property of scale-free. It was compared in form of scatter plot that percentage of the number of proteins (disordered proteins and random proteins) is in x-axis against with the percentage of the number of proteins that affected to the property scale-free network is in y-axis.