CHAPTER 1
INTRODUCTION AND OVERVIEW

1.1 Introduction

One of the most interesting networks in biology is a protein-protein interaction (PPI) network which characterizes the interactions between proteins. This network can be rewritten in a graph form in mathematics which each node represents a protein and an edge represents an interaction between two proteins in the network. The property of network is very important to describe and characterize each node in the network. One of the network's properties that is discussed widely in biological field is a scale-free network. This property can be characterized by degree distribution that is the probability distribution of the number of connections of a node to other nodes in the network. Usually in a scale-free network, the number of high-degree nodes is small while the number of low-degree nodes is large. If a node has high interactions, it is called "hub" node. Any networks with this property show visible hub nodes to identify the feature of a scale-free network. The property of hub node plays the significant role in biological processes. Especially, there was the report of identifying the common characteristic in hub proteins and disordered proteins [1]. Disordered protein describes the protein that some regions of polypeptide chain in protein absence the three-dimensional structure such as folding [2]. This situation makes some functions of genes have lost their work. Disordered proteins have been reported to cause some crucial diseases such as cancer and heart attack disease [2], [3].

In this thesis, we studied the association of disordered proteins using the property of scale-free in the *Homo Sapiens* (Human) network to investigate the nodes that were disordered proteins and also important to the property of scale-free network. To achieve these aims, we developed a new measure of disordered proteins affecting to scale-free network and applied it to find nodes that affected to scale-free network. The two measures were developed based on the network properties. We proposed the measure of identifying disordered proteins affecting to the property of scale-free,

we described as M_{SF_Dsp}, and the measure of identifying proteins affecting to scale-free network, we described as M_{SF}. These measures are very significant and useful to identify the property of scale-free network by using the properties of each node in the network and more helpful to reduce the complexity processes in discarding the node one by one to investigate the parameter gamma in power-law degree distribution.

1.2 Motivation and objective

The motivation of this thesis was from the study of characterizing the property of scale-free network. In 2006, Schnell and his coworkers determined the association between disordered proteins in the property of scale-free network [4]. They found that degrees of each node were not related to the disorder parameters (disorder scores and disorder regions) by using the correlation measure. They implied that disordered proteins did not affect the scale-free architecture of the protein-protein interaction network. However, this conclusion is not clear because only the degree might not suitable to identify the property of scale-free network. Later on, there was the report of disordered proteins that were related to hub proteins, which had more than 10 interaction proteins [1]. Since the scale-free networks were known to be related to hub nodes [5], therefore, the assumption in this work is to study the effect of disordered proteins to property of scale-free network. Identification of nodes effect to the scale-free property could be done by discarding one by one node in the network and observes parameter gamma in power-law form of a degree distribution. If the gamma value is not in the range between 2 and 3, the nodes are determined to be the effect to the scale-free property. Otherwise, the nodes are determined to be not the effect to the scale-free property.

The objective of this thesis is to investigate the association of disordered proteins in scale-free network. Moreover, we tried to develop the measure of identifying disordered proteins that affect the property of a scale-free in a protein-protein interaction network of human. Furthermore, we developed the new measure of identifying proteins that affect to property of scale-free network.

1.3 Scope of this thesis

In this thesis, we investigated the nodes that affect to scale-free network in *Homo Sapiens* (Human) network. The human network has the property of scale-free network that it is apparently visible of hub nodes. This property is determined by a gamma value in power-law form of a degree distribution of the network. In mathematics, we can investigate the network in form of a simple graph. The simple graph has no weight and no loop, which is edge between the same node. The graph is also no direction, called undirected graph. To simplify the network analysis and reduce calculation time, all isolate nodes were removed from the network. Finally, we obtained a network with the total number of nodes less than 10,000 and the total number of edges less than 50,000 for our investigation.

1.4 Literature review

In 2006, Schnell and his coworkers investigated the cause of disordered proteins using the property of scale-free network [4]. They investigated correlation between degrees of proteins in the network and disorder parameters (disorder score and disorder region) using the Pearson-correlation coefficient (PCC). The assumption of their work was that the degree value implied the property of scale-free network and the disorder parameters implied the disordered protein. Therefore, they tried to figure out the relationship between degree and disorder parameters using PCC. The disorder parameters, disorder score and disorder region are obtained from the disordered predictor, VL3 [4]. The VL3 predictor is an algorithm employing an artificial of neural network to predict disordered proteins. Its input is the sequence of amino acid residues in a protein and its output is disorder scores. The results of this work showed the value of PCC was closed to zero and then they concluded that the disordered proteins were not related to scale-free network.

Later on, Haynes and his colleagues examined a feature, which is the interaction between disordered proteins and hub proteins and compared with end proteins from four eukaryotic interactomes (Worm (*Caenorhabditis Elegans*), Yeast

(*Saccharomyces cerevisiae*), Fly (*Drosophila melanogaster*), and Human (*Homo sapiens*) [1]. The end proteins were defined as the proteins having the number of degree equaled to 1. They also defined hub proteins which were proteins having the number of degree greater than or equal to 10. They used their method to characterize of disorder predictions, analyze the various disorder parameters, analyze of the amino acid composition in group of disordered and ordered in hub and end proteins, and also analyze the GO annotations (biological process, molecular function, and cellular component) to find the correlation between GO annotations and disordered proteins and ordered proteins in group of hub and end proteins. Their results showed that hub proteins were more common in disordered proteins than end proteins. Moreover, the ratio of predicted disorder score (disordered residues and disordered regions) were greater in hub proteins than end proteins. In addition, disordered proteins might be as a characteristic determinant of proteins' interaction.

1.5 Overview of the contents

In this thesis, we investigated a node whether is important or effective to a scale-free network by discarding each node in the network and cutting neighbors that have higher value of eigenvector centrality than the node that we considered. We analyzed the value of eigenvector centrality in neighbors of a node for reducing the significant nodes that close to the node we are interested in. Later on, an investigation has started. If we eliminate the node and some neighbors, then slope of network is disobeyed power-law form degree distribution. When this situation emerges, the node will be called 'node that is important or effective to scale-free network'. Thus, we developed the measure to describe node affecting to scale-free using the technique of recognizing attributes in each node.

Moreover, disordered proteins can be described as the caused proteins of the loss of associated function between proteins and related with crucial diseases in human cells such as cancer and heart attack [3]. The application to build an appropriate measure can be applied to other relevant researches. Thus, in this thesis, we tried to investigate the measure of disordered proteins that affect to scale-free

network. In addition, we figured out the new measure to identify proteins that affect to property of scale-free. Besides, we studied the relationship between disordered proteins in the property of scale-free network.

Furthermore, the number of proteins that affect the property of scale-free network is very small. This means, we were focusing on the imbalance data problem. Thus, we managed the imbalance data problem by using the method of Synthetic Minority Over-sampling (SMOTE). The SMOTE algorithm is widely popular in the field of imbalance data problem [6]. Therefore, we received the balance data to develop the measures. These measures were evaluated by using the confusion matrix (accuracy, precision, recall and F-score) including the Receiver Operating Characteristic (ROC) curve to estimate the Area Under the Curve (AUC) for visualizing the performance of the measures.

In the contents, Chapter 1 consists of introduction, motivation and objective, scope of this thesis, literature review, and the overview of this work. Chapter 2 is divided into biological background, mathematical background and performance measure. Chapter 3 consists of data and methods. Chapter 4 shows the experimental results. Finally, Chapter 5 contains conclusion and discussion of this work.