CHAPTER 4
EXPERIMENTAL RESULTS

In this chapter, it was divided into five parts. The first part is about basic network characteristics. The second part explains the attributes that were selected as influential attributes used to design a novel measurement. The third part is the performance of our new measures. The forth part shows the analysis of the validation score. The last part represents the cause of disordered proteins in its scale-free network.

4.1 Basic network characteristics

The constructed real *Homo sapiens* (Human) protein-protein interaction network was examined by network properties such as the number of nodes, the number of edges, the average degree, global clustering coefficient, the gamma of power-law degree distribution and the correlation of degree in the network as shown in Table 4.1.

Table 4.1 Basic network characteristics of our human protein-protein interaction network

Species	#Nodes	#Edges	Average degree	Global clustering coefficient	Gamma in power-law form	Correlation of degree
Homo Sapiens	8,208	45,553	11.10	0.29	2.75	-0.01

After we constructed the real human PPI network, we got 8,208 proteins and 45,553 interactions (or edges). It contained a lot of low-degree nodes which in average each node may have 45,553/8,208 $\approx$ 5 interactions. However, the average degree of this network was about 11 and the global clustering coefficient was 0.29. This means the probability that neighbors were connected was 0.29 and we got such a small value of correlation of degree about -0.01. Obviously, this network is a scale-free network when we looked at the degree distribution and fit the curve into the power-law form,

we obtained gamma parameter equaled to 2.75 which is between 2 and 3. The plot of degree distribution in logarithm scale can be found in Figure 4.1. It shows that there were a large number of low-degree nodes and a few number of high-degree nodes.

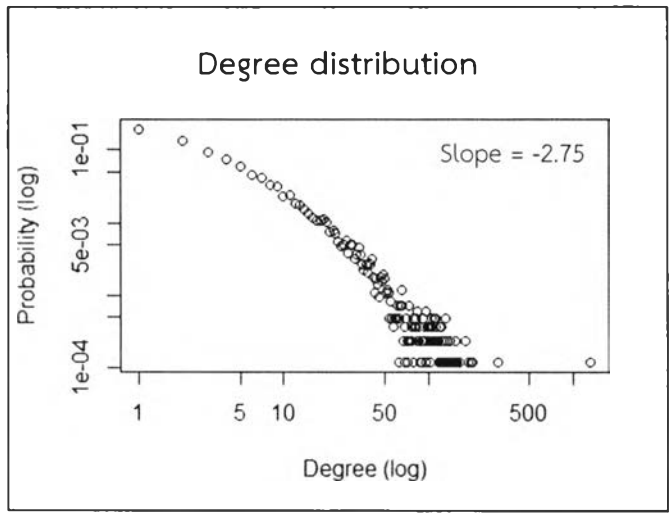


Figure 4.1 The degree distribution of our human protein-protein interaction network

4.2 Influential attributes

To investigate the influential attributes of node in the network, we observed the correlation between each of these attributes: the degree divided by average degree $\frac{k}{<k>}$, the clustering coefficient divided by global clustering coefficient $\frac{c}{<c>}$, the sign of degree correlation *sign*(*R*) and the class of disordered proteins affecting the scale-free network property by using the Pearson-correlation coefficient (PCC) as shown in Table 4.2. Notice that the influential attributes were $\frac{k_i}{<k>}$ and $sign(R_i)$. Since the value of its PCC showed more than or equal to 10% related to the disordered proteins affecting the scale-free network property. As well as, the correlation between these three attributes and the proteins which affected to the scale-free network were also computed as shown in Table 4.2. Still, both $\frac{k_i}{<k>}$ and $sign(R_i)$ showed higher correlation to the scale-free network property.

Table 4.2 The correlation measure (PCC) between each attribute and class labels

Class	Attributes	Correlation (PCC)
Disordered protein affecting the scale-free network property	$\frac{k_i}{<k>}$	0.10
	$\frac{c_i}{<c>}$	0.05
	$sign(R_i)$	0.21
Proteins affecting the scale-free network property	$\frac{k_i}{<k>}$	0.29
	$\frac{c_i}{<c>}$	0.13
	$sign(R_i)$	0.35

The total number of disordered proteins that affect to scale-free (class 1) was 106 proteins and otherwise (class 0) was 8,102 proteins. In addition, the total number of proteins that affect to scale-free (class 1) was 395 proteins and otherwise (class 0) was 7,813 proteins. In this case, we had imbalance data set. The number of proteins in class 1 were significantly smaller than the number of proteins in class 0, as shown in Table 4.3. To avoid this imbalance data problem, we applied the Synthetic Minority Over-sampling Technique (SMOTE) to balance data. Finally, we got the total number of disordered proteins that affect to scale-free (class 1) was 8,102 proteins that is also equal to class 0. In addition, the total number of proteins that affect to scale-free (class 1) was 7,813 proteins that is also equal to class 0, as shown in Table 4.4.

Table 4.3 The number of proteins related to class imbalance

Class imbalance data	Binary	The total number
Disordered protein affecting the scale-free network property	Class 1 (disordered proteins affecting to scale-free)	106
	Class 0 (otherwise)	8,102
Proteins affecting the scale-free network	Class 1 (proteins affecting to scale-free)	395
	Class 0 (otherwise)	7,813

Table 4.4 The number of proteins related to class balance

Class balance data	Binary	The total number
Disordered protein affecting the scale-free network property	Class 1 (disordered proteins affecting to scale-free)	8,102
	Class 0 (otherwise)	8,102
Proteins affecting the scale-free network	Class 1 (proteins affecting to scale-free)	7,813
	Class 0 (otherwise)	7,813

4.3 The performance of our new measures

In this section, we determined the coefficients of these attributes and evaluated the performance of our measures. It was divided into two parts: first is the measure of disordered proteins that affect to scale-free network, M_{SF_Disp} and second is the measure of proteins that affect to scale-free, M_{sf}.

4.3.1 The measure of $M_{SF\ Dsp}$

The attributes $\frac{k_i}{<k>}$ and $sign(R_i)$ were selected to create a measure in imbalance data for predicting disordered proteins affecting the property of scale-free network (see Data and Methods). These two attributes were then compared to the class labels to calculate the appropriate coefficients to the formula by calculating the proportion of the value of the correlation coefficient of each attribute as explained in Data and Methods. After that, we got the coefficient of the attribute $\frac{k_i}{<k>}$ was $0.10/0.10 = 1.00$ and the coefficient of attribute $sign(R_i)$ was $0.21/0.10 = 2.10$, as shown in Table 4.5.

Then, our developed measure of $M_{SF\ Dsp}$ can be rewritten as

$$M_{SF\ Dsp}(i) = \frac{k_i}{<k>} + 2.10 \cdot sign(R_i). \tag{4.1}$$

Table 4.5 The characterizing coefficients in the measure of $M_{SF\ Dsp}$ in imbalance data

Class	Attribute	Correlation (PCC)	Coefficient
Disordered protein affecting the scale-free network property	$\frac{k_i}{<k>}$	0.10	$w_1 = 1.00$
	$sign(R_i)$	0.21	$w_2 = 2.10$

The performance of $M_{SF\ Dsp}$ in imbalance data was observed by plotting a ROC curve and a precision-recall curve as shown in Figures 4.2 and 4.3. With the ROC curve, we yield an AUC of 0.92 which means higher better than randomly selection as well. After that, to figure out which criteria should be a good threshold to make a prediction with this measure, along the precision-recall curve the threshold that yield the highest precision was at 6.97. With this threshold, we obtained an accuracy of 97%, a precision of 10% and a recall of 15%. The confusion matrix of this threshold is shown in Table 4.6. Additionally, the F-score of our measure was 0.12.

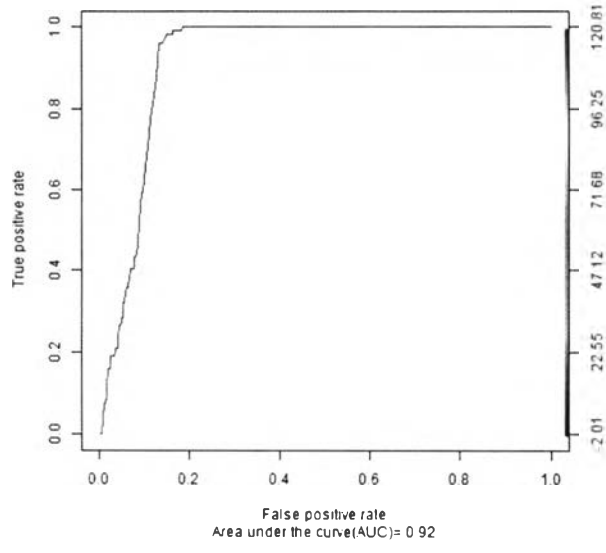


Figure 4.2 The ROC (TPR/FPR) curve in the measure of $M_{SF\ Dsp}$ and the value of AUC in imbalance data

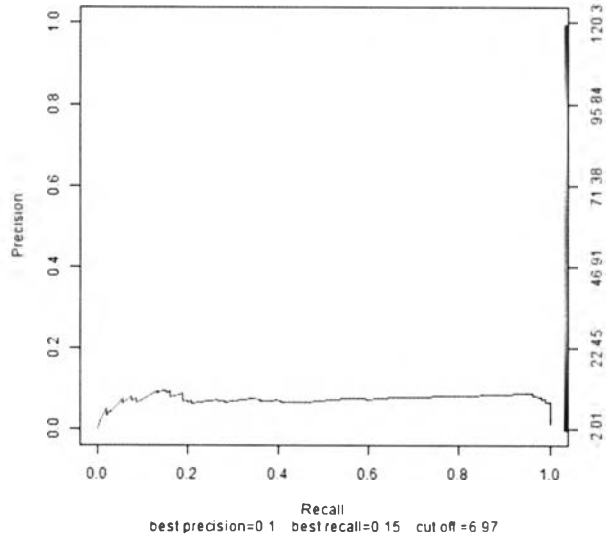


Figure 4.3 The precision-recall curve of $M_{SF\ Dsp}$ in imbalance data

Table 4.6 The confusion matrix of the imbalanced data of $M_{SF.Dsp}$

Confusion matrix	Actual		
Predict		1	0
	1	16	152
	0	90	7,950

In addition, we showed the measure of $M_{SF.Dsp}$ in balance data using SMOTE method for adjusting imbalance data to balance data. The attributes $\frac{k_i}{<k>}$ and $sign(R_i)$ were selected and they were related to disordered protein affecting to the property of scale-free. We got the coefficient of the attribute $\frac{k_i}{<k>}$ was $0.32/0.32=1.00$ and coefficient of attribute $sign(R_i)$ was $0.71/0.32=2.22$ in balance data, as shown in Table 4.7.

Then, our developed measure of $M_{SF.Dsp}$ can be rewritten as

$$M_{SF.Dsp}(i)=\frac{k_i}{<k>}+2.22\cdot sign(R_i). \tag{4.2}$$

Table 4.7 The characterizing coefficients in the measure of $M_{SF.Dsp}$ in balance data

Class	Attribute	Correlation (PCC)	Coefficient
Disordered protein affecting the scale-free network property	$\frac{k_i}{<k>}$	0.32	$w_1=1.00$
	$sign(R_i)$	0.71	$w_2=2.22$

The performance of $M_{SF.Dsp}$ was observed by plotting a ROC curve and a precision-recall curve as shown in Figures 4.4 and 4.5. With the ROC curve, we yield an AUC of 0.92 which means higher better than randomly selection as well. After that, to

figure out which criteria should be a good threshold to make a prediction with this measure, along the precision-recall curve the threshold that yield the highest precision was at 2.36. With this threshold, we obtained an accuracy of 92%, a precision of 89% and a recall of 97%. The confusion matrix of this threshold is shown in Table 4.8. Additionally, the F-score of our measure was 0.92.

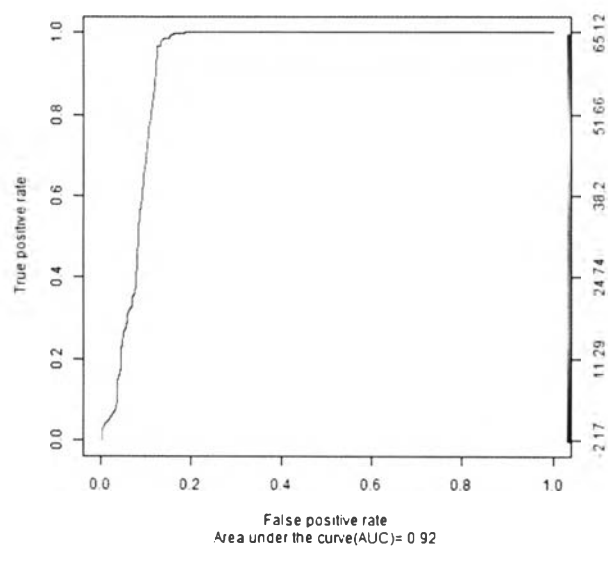


Figure 4.4 The ROC (TPR/FPR) curve in the measure of $M_{SF.Disp}$ and the value of AUC in balance data

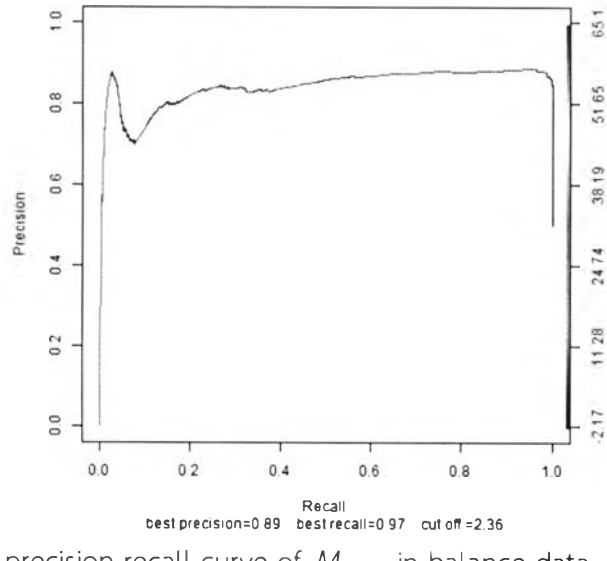


Figure 4.5 The precision-recall curve of $M_{SF.Disp}$ in balance data

Table 4.8 The confusion matrix of the balanced data of M_{SF_Disp}

Confusion matrix	Actual		
Predict		1	0
	1	7,824	1,004
	0	278	7,098

4.3.2 The measure of M_{SF}

In the same manner, to develop the measure of M_{SF} in imbalance data, the coefficients of the influential attributes; $\frac{k_i}{<k>}$ and $sign(R_i)$ were calculated. After that, we obtained the coefficient of the attribute $\frac{k_i}{<k>}$ is $0.29/0.29=1.00$ and the coefficient of the attribute $sign(R_i)$ is $0.35/0.29=1.21$ as shown in Table 4.9. Our developed measure of M_{SF} can be rewritten as

$$M_{SF}(i)=\frac{k_i}{<k>}+1.21\cdot sign(R_i). \tag{4.3}$$

Table 4.9 The characterizing coefficients in the measure of M_{SF} in imbalance data

Class	Attributes	Correlation (PCC)	Coefficient
Scale-free network	$\frac{k_i}{<k>}$	0.29	$w_1=1.00$
	$sign(R_i)$	0.35	$w_2=1.21$

The ROC curve and a precision-recall curve of this measure are shown in Figures 4.6 and 4.7. In this case, we yield an AUC of 0.93 and along the precision-recall curve the threshold that yield the highest precision was at 7.08. With this threshold, we obtained an accuracy of 95%, a precision of 53% and a recall of 20%. The confusion

matrix of this threshold is shown in Table 4.10. Additionally, the F-score of our measure was 0.29.

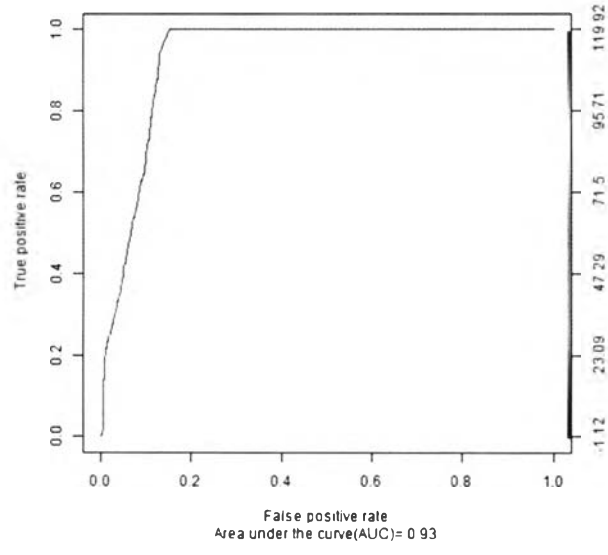


Figure 4.6 The ROC (TPR/FPR) curve in the measure of M_{SF} and the value of AUC in imbalance data

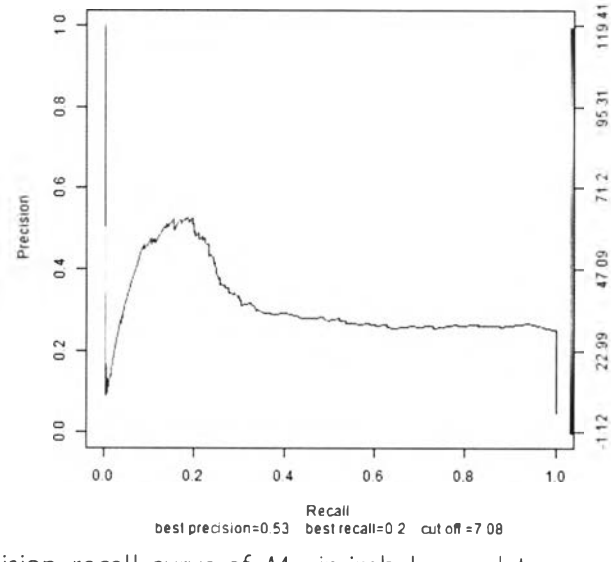


Figure 4.7 Precision–recall curve of M_{SF} in imbalance data

Table 4.10 The confusion matrix of the imbalance data of the measure M_{SF}

Confusion matrix	Actual		
Predict		1	0
	1	78	70
	0	317	7,743

In addition, to develop the measure of M_{SF} in balance data, the coefficients of the influential attributes; $\frac{k_i}{<k>}$ and $sign(R_i)$ which were related to the class of scale-free network were calculated (see Data and Methods). We obtained the coefficient of the attribute $\frac{k_i}{<k>}$ is $0.36/0.36=1.00$ and the coefficient of the attribute $sign(R_i)$ is $0.63/0.36=1.75$ as shown in Table 4.11. Our developed measure of M_{SF} can be rewritten as

$$M_{SF}(i)=\frac{k_i}{<k>}+1.75\cdot sign(R_i). \tag{4.4}$$

Table 4.11 The characterizing coefficients in the measure of M_{SF} in balance data

Class	Attributes	Correlation (PCC)	Coefficient
Scale-free network	$\frac{k_i}{<k>}$	0.36	$w_1=1.00$
	$sign(R_i)$	0.63	$w_2=1.75$

The ROC curve and a precision-recall curve of this measure are shown in Figures 4.8 and 4.9. In this case, we yield an AUC of 0.93 and along the precision-recall curve the threshold that yield the highest precision was at 1.89. With this threshold, we obtained an accuracy of 89%, a precision of 90% and a recall of 87%. The confusion

matrix of this threshold is shown in Table 4.12. Additionally, the F-score of our measure was 0.88.

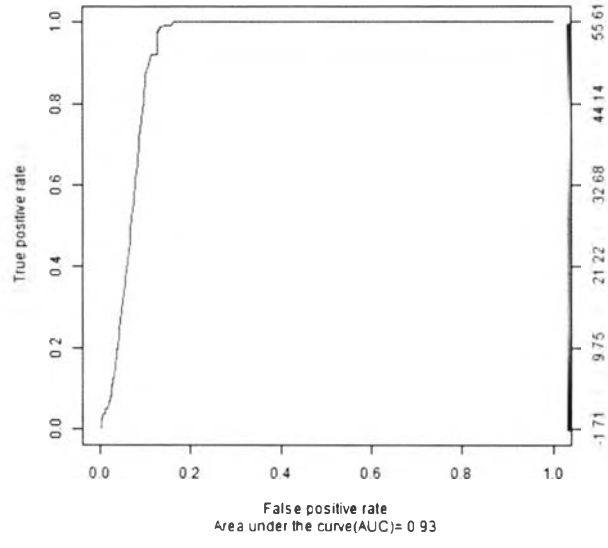


Figure 4.8 The ROC (TPR/FPR) curve in the measure of M_{SF} and the value of AUC in balance data

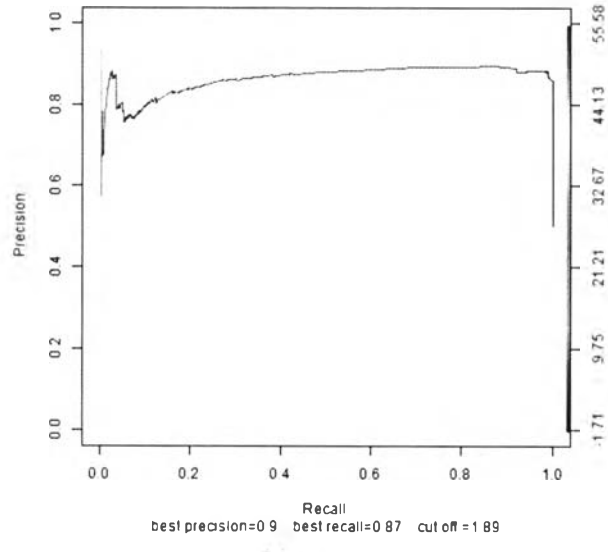


Figure 4.9 Precision–recall curve of M_{SF} in balance data

Table 4.12 The confusion matrix of the balance data of the measure M_{SF}

Confusion matrix	Actual		
Predict		1	0
	1	6,800	781
	0	1,013	7,032

This process is called self-consistency test. The self-consistency test is when the data for creating and testing a measure are the same. Thus, the developed measure might be overfitted to the data. It showed that the value AUC of our two measures are so high. Therefore, the split test is considered. The split test is divided into two parts. First is 90% of data using the method of SMOTE to create a balance data for fitting the measure. Second is 10% of data for testing. In this work, the split test is used for evaluating the measure in 10 times with the value of AUC. The average AUC in the measure of disordered proteins affecting scale-free property, $M_{SF\ Disp}$ is 0.918 and the average AUC of the measure for identifying proteins affecting scale-free property, M_{SF} is 0.937. The average AUC of two measures are more than 90%, this means that the method for developing our measures is good performance. Next, we showed our measures, $M_{SF\ Disp}$ and M_{SF} by fitting the measure using SMOTE to adjust in balance data and test performance in imbalance data with AUC. The value of AUC in the measure of disordered proteins affecting scale-free property, $M_{SF\ Disp}(i) = \frac{k_i}{<k>} + 2.22 \cdot sign(R_i)$ is 0.92 and the value of threshold is 2.36 as shown in Figure 4.10. The value of AUC in the measure of identifying proteins affecting scale-free property, $M_{SF}(i) = \frac{k_i}{<k>} + 1.75 \cdot sign(R_i)$ is 0.94 and the value of threshold is 1.89 as shown in Figure 4.11. Thus, our measures, $M_{SF\ Disp}$ and M_{SF} are good performance.

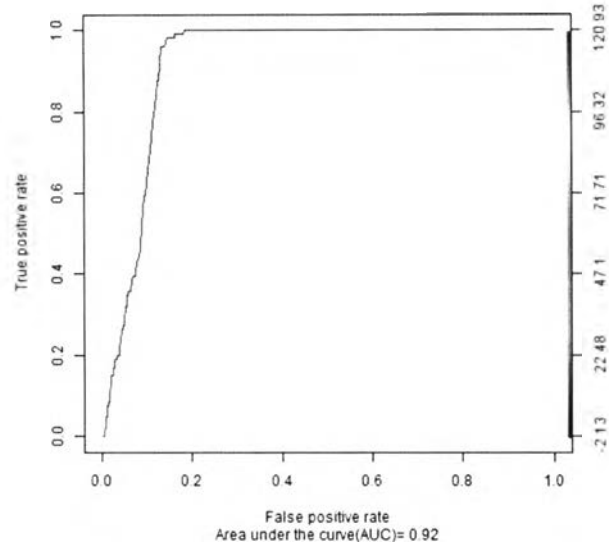


Figure 4.10 The ROC (TPR/FPR) curve in our measure of $M_{SF.Dsp}$

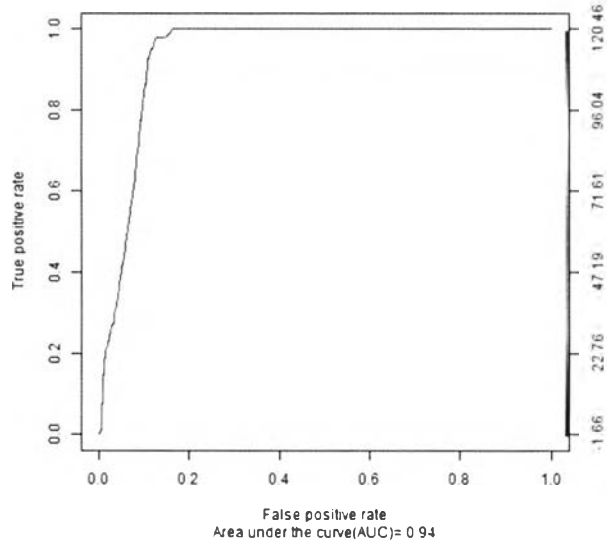
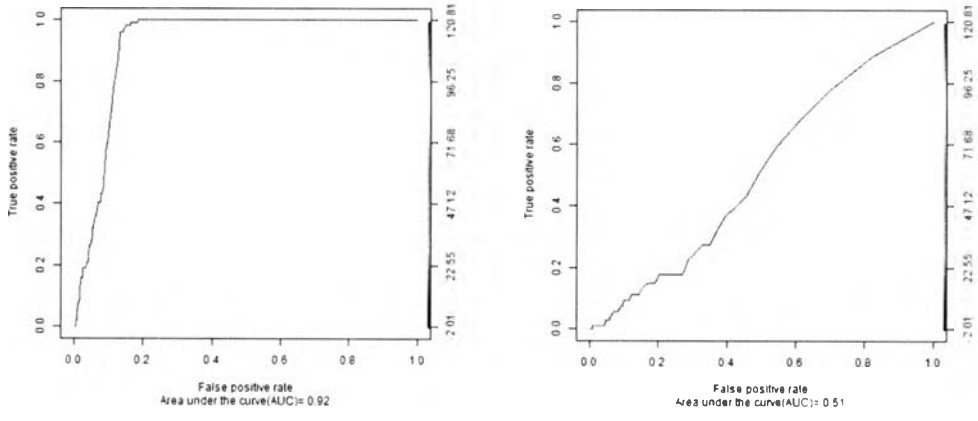


Figure 4.11 The ROC (TPR/FPR) curve in our measure of M_{SF}

4.4 Analysis of the validation score

In this section, we compared the score of performance, the AUC between random class label measures and our measures. The class label in the measure of $M_{SF.Dsp}$ equals to 1, which means disordered proteins affecting to property of scale-free

network, otherwise equals to 0. In addition, the class label in the measure of M_{SF} equals to 1, which means proteins affecting to property of scale-free network, otherwise equals to 0. The random class label described as the shuffle only the class label and still the same of these influential attributes, the attribute of degree divided by the average degree, $\frac{k}{<k>}$ and the other attribute is sign of the degree correlation, $sign(R)$. We compared the value AUC in group of imbalance data and balance data of two measures, $M_{SF\ Disp}$ and M_{SF} with the same coefficients and threshold for validating the significance of these attributes and class labels.



A B

Figure 4.12 The comparison of performance (AUC): (A) The graph plot of ROC and the value AUC of our measure $M_{SF\ Disp}$ in imbalance data, (B) The graph plot of ROC and the value AUC of the random class labels measure $M_{SF.Disp}$ in imbalance data

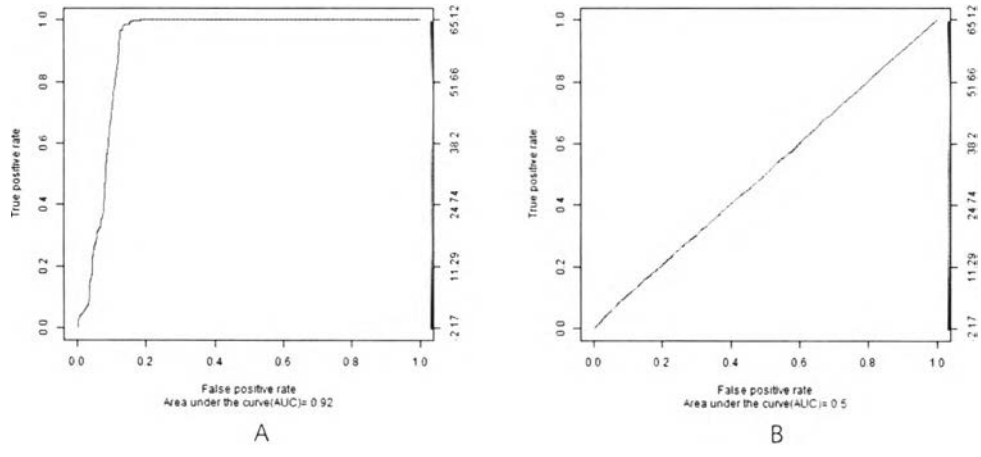


Figure 4.13 The comparison of performance (AUC): (A) The graph plot of ROC and the value AUC of our measure $M_{SF\ Disp}$ in balance data, (B) The graph plot of ROC and the value AUC of the random class labels measure $M_{SF\ Disp}$ in balance data

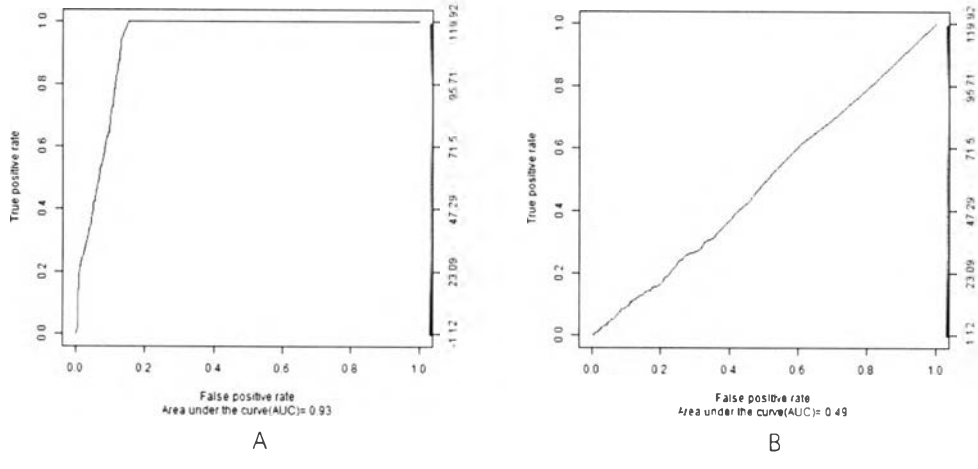
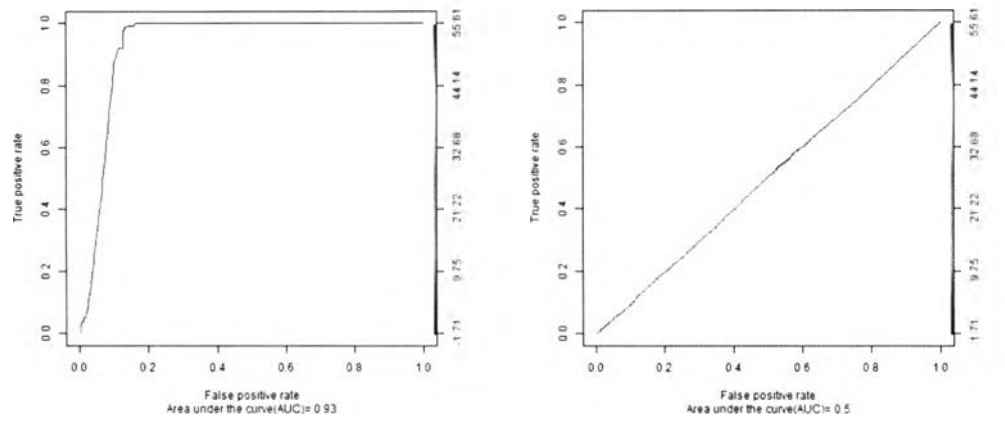


Figure 4.14 The comparison of performance (AUC): (A) The graph plot of ROC and the value AUC of our measure M_{SF} in imbalance data, (B) The graph plot of ROC and the value AUC of the random class labels measure M_{SF} in imbalance data



A B

Figure 4.15 The comparison of performance (AUC): (A) The graph plot of ROC and the value AUC of our measure M_{SF} in balance data, (B) The graph plot of ROC and the value AUC of the random class labels measure M_{SF} in balance data

We can conclude this part in Figures 4.12, 4.13, 4.14 and 4.15, in comparison of the performance between the random class and our measures in both balance and imbalance data. Figure 4.12 (A) showed the graph plot of ROC and the value AUC of our measure $M_{SF\ Dsp}$ in imbalance data was 0.92. Figure 4.12 (B) showed the graph plot of ROC and the value AUC of the random class labels measure $M_{SF\ Dsp}$ in imbalance data was 0.51. Moreover, Figure 4.13 (A) showed the graph plot of ROC and the value AUC of our measure $M_{SF\ Dsp}$ in balance data was 0.92. Figure 4.13 (B) showed the graph plot of ROC and the value AUC of the random class labels measure $M_{SF\ Dsp}$ in balance data was 0.50.

In addition, Figure 4.14 (A) showed the graph plot of ROC and the value AUC of our measure M_{SF} in imbalance data was 0.93. Figure 4.14 (B) showed the graph plot of ROC and the value AUC of the random class labels measure M_{SF} in imbalance data was 0.49. Furthermore, Figure 4.15 (A) showed the graph plot of ROC and the value AUC of our measure M_{SF} in balance data was 0.93. Figure 4.15 (B) showed the graph plot of ROC and the value AUC of the random class labels measure M_{SF} in balance data was 0.50.

Therefore, we implied that the performance of our two measures, M_{SF} and $M_{SF\ Dis}$ better than the random selection as we could see that the value of AUC in part (A) is greater than the value of AUC in part (B).

4.5 The impact of disordered proteins in its scale-free network

To investigate the effect of disordered proteins in scale-free network, all 2,335 disordered proteins were removed from the network and the parameter gamma in the power-law form of the degree distribution was observed. We found out that the mutated network which had the value of gamma was 6.26 while the original one was 2.75. Therefore, this shows that the lack of the disordered proteins obviously affected to the scale-free structure. However, this might be because a lot of number of proteins were deleted from the network (the number of proteins before and after removing the disordered proteins as shown in Table 4.13). We then tried to remove the same number of disordered proteins randomly out of the network and observed the value of gamma again. This random selection was performed 100 times and we found that about 20% of the times, the scale-free free property was affected. To be fair, the comparison of removing randomly selected proteins and randomly selected disordered proteins 100 times were performed as shown in Figure 4.16. Notice that mostly when removing disordered proteins, the mutated networks lost the scale-free property. This implied that the disordered proteins were more crucial in the scale-free network.

Table 4.13 The number of nodes, edges and parameter gamma for the original network and the mutated network

Network	#Nodes	#Edges	Gamma in power-law form
Original human scale-free network	8,208	45,553	2.75
Mutated network after removing disordered proteins	5,873	21,698	6.26

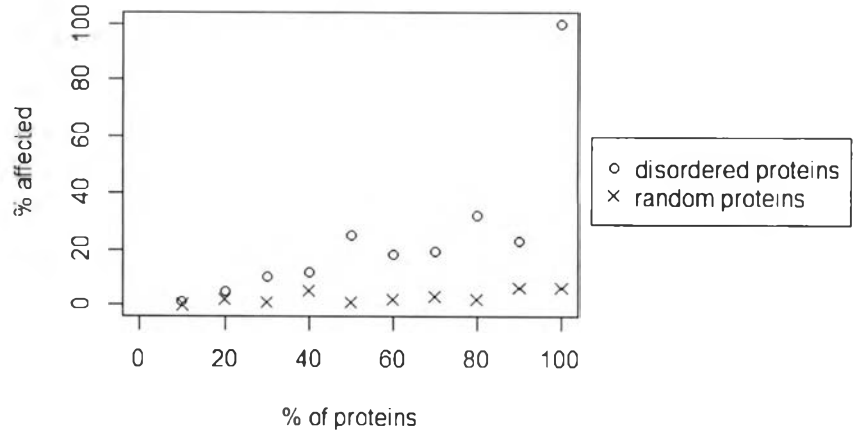


Figure 4.16 The comparison of discarding disordered proteins and random proteins in various range of proteins