

ตัววัดสำหรับโปรตีนที่ผิดปกติที่มีผลกระทบต่อโครงข่ายแบบสเกลฟรี

นางสาวศาดานาฏ กิจศิริานุกัฏ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาคณิตศาสตร์ประยุกต์และวิทยาการคณนา ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2560
ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย



5872061223

MEASUREMENT FOR DISORDERED PROTEINS AFFECTING SCALE-FREE NETWORK

Miss Satanat Kitsiranuwat

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Applied Mathematics and
Computational Science

Department of Mathematics and Computer Science

Faculty of Science

Chulalongkorn University

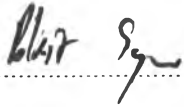
Academic Year 2017

Copyright of Chulalongkorn University



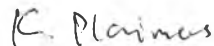
Thesis Title MEASUREMENT FOR DISORDERED PROTEINS
AFFECTING SCALE-FREE NETWORK
By Miss Satanat Kitsiranuwat
Field of Study Applied Mathematics and Computational Science
Thesis Advisor Kitiporn Plaimas, Dr.rer.nat.
Thesis Co-Advisor Assistant Professor Apichat Surataneer, Dr.rer.nat.

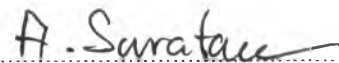
Accepted by the Faculty of Science, Chulalongkorn University in Partial
Fulfillment of the Requirements for the Master's Degree


..... Dean of the Faculty of Science
(Associate Professor Polkit Sangvanich, Ph.D.)

THESIS COMMITTEE

..... Chairman
(Assistant Professor Khamron Mekchay, Ph.D.)

..... Thesis Advisor
(Kitiporn Plaimas, Dr.rer.nat.)

..... Thesis Co-Advisor
(Assistant Professor Apichat Surataneer, Dr.rer.nat.)

..... Examiner
(Assistant Professor Ratinan Boonlurb, Ph.D.)

..... External Examiner
(Assistant Professor Treenut Saithong, Ph.D.)



ศาตนาฏ กิจติรานูวัตร : ตัววัดสำหรับโปรตีนที่ผิดปกติที่มีผลกระทบต่อโครงข่ายแบบสเกลฟรี (MEASUREMENT FOR DISORDERED PROTEINS AFFECTING SCALE-FREE NETWORK) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: อ. ดร. กิติพร พลายมาศ, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม: ผศ. ดร. อภิชาติ ศุภธณี, หน้า.

โครงข่ายปฏิสัมพันธ์ระหว่างโปรตีนเป็นโครงข่ายที่กล่าวถึงกันมากในเชิงชีววิทยา ซึ่งเป็นโครงข่ายขนาดใหญ่แสดงปฏิสัมพันธ์ระหว่างโปรตีน โดยสมบัติที่โดดเด่นของโครงข่ายที่น่าสนใจคือโครงข่ายแบบสเกลฟรี โดยโหนดที่มีดีกรีน้อยจะมีจำนวนมาก และโหนดที่มีดีกรีมากจะมีจำนวนน้อย ในวิทยานิพนธ์ฉบับนี้สนใจศึกษาโครงข่ายแบบสเกลฟรีของปฏิสัมพันธ์ระหว่างโปรตีนในมนุษย์ เพื่อหาความสัมพันธ์และผลกระทบของโปรตีนที่ผิดปกติในโครงข่ายแบบสเกลฟรี และประยุกต์หาโปรตีนที่ผิดปกติซึ่งมีความสำคัญต่อโครงสร้างของโครงข่ายแบบสเกลฟรี โปรตีนที่ผิดปกติเป็นโปรตีนที่สำคัญต่อการเกิดโรคต่างๆ เช่น โรคมะเร็งและโรคหัวใจ การหาตัววัดสำหรับระบุโปรตีนที่มีลักษณะผิดปกติและมีความสำคัญต่อโครงข่ายแบบสเกลฟรีจึงมีความสำคัญและเป็นประโยชน์ในการวิเคราะห์เชิงลึกทางชีววิทยา ตัววัดที่นำเสนอสองตัวนั้นคือ $M_{SF\ Disp}$ และ M_{SF} ที่พัฒนาขึ้นในวิทยานิพนธ์ฉบับนี้พัฒนาบนพื้นฐานของความสัมพันธ์ระหว่างสมบัติของโหนด (ดีกรีของโหนดและสหสัมพันธ์ของดีกรี) กับกลุ่มของตัววัด (โปรตีนที่ผิดปกติที่สำคัญของโครงข่ายแบบสเกลฟรีและโหนดที่สำคัญของโครงข่ายแบบสเกลฟรี) ตัววัดทั้งสองนี้สามารถนำมาประยุกต์เพื่อหาโปรตีนที่มีความสำคัญต่อโครงข่ายแบบสเกลฟรีได้ และยิ่งไปกว่านั้น ตัววัดทั้งสองนี้มีประโยชน์เพื่อลดขั้นตอนในการพิจารณาค่าเกมมาในการแจกแจงดีกรีแบบกฏพาวเวอร์ ในการหาสมบัติแบบสเกลฟรี

ภาควิชา	คณิตศาสตร์และวิทยาการ	ลายมือชื่อนิสิต	ศาสตราจารย์ กิจติรานูวัตร
	คอมพิวเตอร์	ลายมือชื่อ อ.ที่ปรึกษาหลัก	กิติพร พลายมาศ
สาขาวิชา	คณิตศาสตร์ประยุกต์และวิทยาการ	ลายมือชื่อ อ.ที่ปรึกษาร่วม	อภิชาติ ศุภธณี
	คณนา		

ปีการศึกษา 2560



5872061223 : MAJOR APPLIED MATHEMATICS AND COMPUTATIONAL SCIENCE

KEYWORDS: PROTEIN-PROTEIN INTERACTION (PPI) NETWORK / DISORDERED PROTEIN / SCALE- FREE NETWORK

SATANAT KITSIRANUWAT: MEASUREMENT FOR DISORDERED PROTEINS AFFECTING SCALE-FREE NETWORK. ADVISOR: KITIPORN PLAIMAS, Dr.rer.nat., CO-ADVISOR: ASST. PROF. APICHAT SURATANEE, Dr.rer.nat., pp.

Protein-protein interaction network is well-known in biology in which it is a large network displaying the association between proteins. One of the most widely used network's properties is scale-free property which means the network contains many nodes having small numbers of interactions and few nodes having large numbers of interactions. In this thesis, we are interested in characterizing the scale-free network of *Homo Sapiens* (Human) protein-protein interactions for investigating the association of disordered proteins in scale-free network and identifying the disordered proteins which are important to the scale-free structure. The disordered proteins are crucial and essential to cause many serious diseases such as cancer and heart attract diseases. Finding a measure for identifying disordered proteins that are important to the scale-free network is significant and useful to analyze in deep biological processes. The two proposed measures, $M_{SF,Disp}$ and M_{SF} , in this thesis, were developed based on the relationship between the properties (degree of nodes and the degree of correlation) and class of the measures (nodes that are important to the scale-free property and disordered proteins that are important to the scale-free property). They can be applied to investigate a protein that is important to the scale-free network. Furthermore, they are useful to reduce the processes of investigating the value of gamma in the power-law degree distribution in identifying the scale-free property.

Department: Mathematics and
Computer Science

Student's Signature Satanat Kitsiranuwat
Advisor's Signature K. Plaimas

Field of Study: Applied Mathematics and Computational Science
Co-Advisor's Signature A. Suratae

Academic Year: 2017



ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my advisor, Dr. Kitiporn Plaimas and my co-advisor, Assistant Professor Dr. Apichat Surataneer for their valuable guidance, suggestions and encouragement in every process of this thesis work. They were beside me in any rough time and under pressure situations. This thesis would not have been possible without their grateful helps.

Furthermore, I wish to express my sincerely grateful thanks to my thesis committees, Assistant Professor Dr. Khamron Mekchay, Assistant Professor Dr. Ratinan Boonklurb, Assistant Professor Dr. Treenut Saithong, for their time to give valuable advice and every comment in my thesis.

Finally, I would like to thank my supporter from the Science Achievement Scholarship of Thailand (SAST), and also the Applied Mathematics and Computational Science Program, Department of Mathematics and Computer Science, Faculty of Science, Chulaongkorn University, Thailand, for providing the opportunity, all of the knowledge and all resources throughout my graduate study in master's degree.



CONTENTS

	Page
THAI ABSTRACT	iv
ENGLISH ABSTRACT	v
ACKNOWLEDGEMENTS	vi
CONTENTS	vii
CONTENT OF TABLES	x
CONTENT OF FIGURES	xi
CHAPTER 1 INTRODUCTION AND OVERVIEW	1
1.1 Introduction	1
1.2 Motivation and objective	2
1.3 Scope of this thesis	3
1.4 Literature review	3
1.5 Overview of the contents	4
CHAPTER 2 BACKGROUND KNOWLEDGE	6
2.1 Biological background	6
2.1.1 Disordered proteins	6
2.1.2 Protein-protein interaction network	6
2.2 Mathematical background	7
2.2.1 Graph and its definition	7
2.2.2 Degree distribution and power-law form	7
2.2.3 Correlation measure	8
2.2.4 Degree correlations	9
2.2.5 Interesting node properties	10



	Page
2.3 Performance measure	12
2.3.1 Accuracy, precision and recall	12
2.3.2 ROC curve and AUC.....	13
CHAPTER 3 DATA AND METHODS	15
3.1 Data	15
3.1.1 Protein-protein interaction network database.....	15
3.1.2 Disordered proteins	15
3.2 Methods	15
3.2.1 Overview	15
3.2.2 Analysis of network properties.....	16
3.2.3 Degree distributions.....	17
3.2.4 Development of our new measures.....	19
3.2.5 Investigation of the cause of disordered proteins in scale-free network..	20
CHAPTER 4 EXPERIMENTAL RESULTS.....	22
4.1 Basic network characteristics	22
4.2 Influential attributes	23
4.3 The performance of our new measures.....	25
4.3.1 The measure of $M_{SF\ Disp}$	26
4.3.2 The measure of M_{sf}	30
4.4 Analysis of the validation score	35
4.5 The impact of disordered proteins in its scale-free network.....	39
CHAPTER 5 CONCLUSION AND DISCUSSION	41
REFERENCES	43



751435775

	Page
APPENDICES.....	45
VITA.....	69



CONTENT OF TABLES

Table 2.1 The analysis of confusion matrix	12
Table 4.1 Basic network characteristics of our human protein-protein interaction network.....	22
Table 4.2 The correlation measure (PCC) between each attribute and class labels...	24
Table 4.3 The number of proteins related to class imbalance	25
Table 4.4 The number of proteins related to class balance	25
Table 4.5 The characterizing coefficients in the measure of $M_{SF, Disp}$ in imbalance data.....	26
Table 4.6 The confusion matrix of the imbalanced data of $M_{SF, Disp}$	28
Table 4.7 The characterizing coefficients in the measure of $M_{SF, Disp}$ in balance data ...	28
Table 4.8 The confusion matrix of the balanced data of $M_{SF, Disp}$	30
Table 4.9 The characterizing coefficients in the measure of M_{SF} in imbalance data..	30
Table 4.10 The confusion matrix of the imbalance data of the measure M_{SF}	32
Table 4.11 The characterizing coefficients in the measure of M_{SF} in balance data....	32
Table 4.12 The confusion matrix of the balance data of the measure M_{SF}	34
Table 4.13 The number of nodes, edges and parameter gamma for the original network and the mutated network	40



CONTENT OF FIGURES

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	9
Figure 3.1 The flow chart of proposed methodology.....	16
Figure 4.1 The degree distribution of our human protein-protein interaction network	23
Figure 4.2 The ROC (TPR/FPR) curve in the measure of $M_{SF\ Disp}$ and the value of AUC in imbalance data.....	27
Figure 4.3 The precision-recall curve of $M_{SF\ Disp}$ in imbalance data.....	27
Figure 4.4 The ROC (TPR/FPR) curve in the measure of $M_{SF\ Disp}$ and the value of AUC in balance data	29
Figure 4.5 The precision-recall curve of $M_{SF\ Disp}$ in balance data	29
Figure 4.6 The ROC (TPR/FPR) curve in the measure of M_{SF} and the value of AUC in imbalance data.....	31
Figure 4.7 Precision–recall curve of M_{SF} in imbalance data	31
Figure 4.8 The ROC (TPR/FPR) curve in the measure of M_{SF} and the value of AUC in balance data	33
Figure 4.9 Precision–recall curve of M_{SF} in balance data.....	33
Figure 4.10 The ROC (TPR/FPR) curve in our measure of $M_{SF\ Disp}$	35
Figure 4.11 The ROC (TPR/FPR) curve in our measure of M_{SF}	35
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... 36	36



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..... 37

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..... 37

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..... 38

Figure 4.16 The comparison of discarding disordered proteins and random proteins in various range of proteins..... 40

