โครงการ

การเรียนการสอนเพื่อเสริมประสบการณ์

ชื่อโครงการ การค้นหาเพื่อนบ้านใกล้ที่สุดผกผันสำหรับการระบุยีนที่สำคัญบนโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีน

Reverse nearest neighbors search for identification of bacteria's essential genes in protein-protein interaction network

ชื่อนิสิต นางสาวนัยขวัญ ไชยศรี

ภาควิชา คณิตศาสตร์และวิทยาการคอมพิวเตอร์

สาขาวิชา คณิตศาสตร์

ปีการศึกษา 2561

คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

การค้นหาเพื่อนบ้านใกล้ที่สุดผกผันสำหรับการระบุยีนที่สำคัญบนโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีน

นางสาวนัยขวัญ ไชยศรี

โครงงานนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต
สาขาวิชาคณิตศาสตร์ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2561

Reverse nearest neighbors search for identification of bacteria's essential genes in protein-protein interaction network

Miss Naiyakwan Chaisri

A Project Submitted in Partial Fulfillment of the Requirements
for the Degree of Bachelor of Science Program in Mathematics
Department of Mathematics and Computer Science
Faculty of Science
Chulalongkorn University
Academic Year 2018

หัวข้อโครงงาน	การค้นหาเพื่อนบ้านใกล้ที่สุดผกผันสำหรับการระบุยีนที่สำคัญบนโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีน
โดย	นางสาว นัยขวัญ ไชยศรี
สาขาวิชา	คณิตศาสตร์
อาจารย์ที่ปรึกษาโครงงานหลัก	ผู้ช่วยศาสตราจารย์ ดร. กิติพร พลายมาศ

ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับโครงงานฉบับนี้ เป็นส่วนหนึ่ง ของการศึกษาตามหลักสูตรปริญญาบัณฑิต ในรายวิชา 2301499 โครงงานวิทยาศาสตร์ (Senior Project)

.....
(ศาสตราจารย์ ดร.ภฤษณะ เนียมมณี) หัวหน้าภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์

คณะกรรมการสอบโครงงาน

.....
(ผู้ช่วยศาสตราจารย์ ดร. กิติพร พลายมาศ) อาจารย์ที่ปรึกษาโครงงานหลัก

.....
(รองศาสตราจารย์ ดร. พิมพ์เพ็ญ เวชชาชีวะ) กรรมการ

.....
(รองศาสตราจารย์ ณัฏฐนาถ ไตรภพ) กรรมการ

นางสาวนัยขวัญ ไชยศรี: การค้นหาเพื่อนบ้านใกล้ที่สุดผกผันสำหรับการระบุยีนที่สำคัญบนโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีน. (Reverse nearest neighbors search for identification of bacteria's essential genes in protein-protein interaction network) อ.ที่ปรึกษาโครงงานหลัก: ผู้ช่วยศาสตราจารย์ ดร. กิตติพร พลายมาศ, 55 หน้า.

ยีนที่สำคัญเป็นหน่วยพันธุกรรมขั้นพื้นฐานสำหรับการดำรงอยู่ของสิ่งมีชีวิต การค้นหาหรือระบุยีนที่สำคัญจึงมีประโยชน์อย่างมากในการศึกษาและพัฒนายาต้านเชื้อแบคทีเรียและเชื้อโรคต่าง ๆ และปัจจุบันวิธีการทางการคำนวณมีส่วนช่วยให้สามารถระบุยีนหรือโปรตีนที่สำคัญทำได้รวดเร็วขึ้นและประหยัดค่าใช้จ่ายได้มากขึ้นด้วย วิธีการต่าง ๆ เหล่านี้รวมถึงการเรียนรู้ด้วยเครื่อง (Machine Learning) และการศึกษาคุณสมบัติทางโทโปโลยีโครงข่าย (Network Topology) ของโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีนอีกด้วย ดังนั้นในโครงงานผู้จัดทำได้ศึกษาและประยุกต์ใช้การค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน (Reverse Nearest Neighbor Search) เข้ามาช่วยหายีนที่สำคัญในโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีนของแบคทีเรีย *E. coli* โดยนำข้อมูลโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีนในแบคทีเรีย และข้อมูลยีนที่สำคัญ มาวิเคราะห์ร่วมกันเพื่อขยายผลหายีนหรือโปรตีนที่สำคัญตัวใหม่ ซึ่งจะเป็นประโยชน์ต่อการพัฒนายาต้านเชื้อแบคทีเรียต่อไปได้ในอนาคต และนำผลที่ได้จากการระบุยีนที่สำคัญโดยวิธีการค้นหาเพื่อนบ้านผกผันมาเปรียบเทียบประสิทธิภาพกับวิธีการพิจารณาของดีกรีของโหนดยีน และวิธีการค้นหาเพื่อนบ้านที่ใกล้ที่สุด (Nearest Neighbor Search) พบว่าการระบุยีนที่สำคัญโดยใช้การค้นหาเพื่อนบ้านใกล้ที่สุดผกผันมีประสิทธิภาพดีที่สุด โดยมีค่าความถูกต้อง (accuracy) 74.02% ค่าความแม่นยำ (precision) 41.01%

ภาควิชา คณิตศาสตร์และวิทยาการคอมพิวเตอร์ ลายมือชื่อนิสิต นัยขวัญ ไชยศรี

สาขาวิชา คณิตศาสตร์ ลายมือชื่อ อ.ที่ปรึกษาโครงงานหลัก กิตติพร พลายมาศ

ปีการศึกษา 2561

5833528723: MAJOR MATHEMATICS

KEYWORDS: REVERSE NEAREST NEIGHBORS SEARCH/ ESSENTIAL GENES / PROTEIN – PROTEIN INTERACTION NETWORK.

Miss Naiyakwan Chaisri: Reverse nearest neighbors search for identification of bacteria's essential genes in protein-protein interaction network. ADVISOR: ASST. PROF. Kitiporn Plaimas, Ph.D., 55 pp.

Essential genes are basic genes of an organism that are important for its survival. Therefore, an identification of bacteria's essential genes is very important for the study and development of antibiotic drugs. Nowadays, a computational method makes the identification of essential genes much simpler and more economical; including, machine learning and network topology of a protein-protein interaction network. The objectives of this project are to study and to apply reverse nearest neighbor search to identify bacteria's essential genes in a protein-protein interaction network. The protein-protein interaction network and the list of known essential genes were collected from various public database. Reverse nearest neighbor search was used for predicting essential genes for the development of antibiotic drugs in the future. The performances of this method were compared to the degree of connectivity and the nearest neighbor search. The results shows the reverse nearest neighbor search is the best with the accuracy of 74.02% and the precision of 41.01%.

Department: Mathematics and Computer Science Student's Signature Naiyakwan.

Field of Study: Mathematics Advisor's Signature K. Plaimas

Academic Year: 2018

กิตติกรรมประกาศ

การวิจัยในหัวข้อเรื่อง “การค้นหาเพื่อนบ้านใกล้ที่สุดผกผันสำหรับการระบุยีนที่สำคัญบนโครงข่าย ปฏิสัมพันธ์ระหว่างโปรตีน” ได้รับการสนับสนุนอย่างเต็มที่ ผู้ช่วยศาสตราจารย์ ดร. กิติพร พลายมาศ อาจารย์ที่ปรึกษาโครงงานฉบับนี้ ตั้งแต่การเลือกหัวข้อวิจัย การให้องค์ความรู้และกระบวนการต่าง ๆ ที่ใช้ในการดำเนินงานโครงงานครั้งนี้ รวมไปถึงรองศาสตราจารย์ ดร.พิมพ์เพ็ญ เวชชาชีวะ และรองศาสตราจารย์ ณัฏฐนาถ ไตรภพ คณะกรรมการสอบโครงงานฉบับนี้ ซึ่งทำให้โครงงานได้รับคำแนะนำหรือข้อเสนอแนะเพื่อให้ผู้เขียนนำสิ่งเหล่านั้นกลับไปคิด และปรับปรุงโครงงานฉบับนี้ให้มีความถูกต้องสมบูรณ์ยิ่งขึ้น และต้องขอขอบคุณนางสาวชื่นชนก หนูแสง นางสาวศาตนาฎ กิจศิรานุวัตร และนายภีมพัฒน์ วงศ์ศรีพิสันต์ ซึ่งคอยเป็นที่ปรึกษาคอยให้คำแนะนำให้ความรู้ที่ใช้ในการดำเนินงานโครงงานครั้งนี้ ยิ่งไปกว่านั่น ขอขอบคุณทางภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย ที่ได้ให้การสนับสนุนด้านต่าง ๆ รวมถึงงบประมาณในการทำวิจัย จึงทำให้งานวิจัยนี้สำเร็จลุล่วงไปได้ด้วยดี และขอขอบคุณครอบครัวที่เป็นกำลังใจในการทำวิจัย ข้าพเจ้าจึงใคร่ขอขอบพระคุณเป็นอย่างยิ่งสำหรับความช่วยเหลือในทุก ๆ ด้าน และหวังว่าผลการวิจัยนี้ จะเป็นประโยชน์ในการพัฒนาองค์ความรู้และนำไปใช้กับสิ่งมีชีวิตอื่นๆได้ต่อไป

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

สารบัญกราฟ

หน้า

บทที่ 1
บทนำ

ในบทนี้จะกล่าวถึงความเป็นมาและเหตุผลของโครงงาน วัตถุประสงค์ของโครงงาน ขอบเขตของโครงงาน ขั้นตอนในการดำเนินโครงงาน ประโยชน์ที่คาดว่าจะได้รับ และโครงสร้างของรายงาน โดยมีรายละเอียดดังต่อไปนี้

1.1 ความเป็นมาและเหตุผล

ยีนที่สำคัญ (Essential genes) คือ ยีนที่สำคัญต่อการดำรงชีวิตของสิ่งมีชีวิต ยีนที่สำคัญถือเป็นรากฐานของเซลล์สิ่งมีชีวิต เซลล์ทุกเซลล์จึงมียีนที่สำคัญอยู่จำนวนหนึ่ง การระบุยีนที่สำคัญไม่เพียงแต่จะเป็นความต้องการขั้นพื้นฐานสำหรับการอยู่รอดของสิ่งมีชีวิต แต่ยังสำคัญสำหรับการค้นหายีนของมนุษย์และยาต้านเชื้อแบคทีเรียและเชื้อโรคต่างๆ อีกด้วย แต่ด้วยวิธีการทดลองในการระบุยีนที่สำคัญมีค่าใช้จ่ายสูง มีหลายขั้นตอน และใช้เวลานาน การเก็บข้อมูลของลำดับยีนและข้อมูลการทดลองที่มีปริมาณมาก จึงมีการนำเสนอวิธีการจำนวนมากสำหรับการระบุโปรตีนที่สำคัญซึ่งเป็นประโยชน์สำหรับการคัดกรองเลือกยีนสำหรับศึกษาเพิ่มเติมในห้องทดลอง เป็นการลดค่าใช้จ่ายในการทำการทดลองต่อไป โดยวิธีการที่ทันสมัยที่สุดสำหรับการระบุโปรตีนที่สำคัญโดยอาศัยการเรียนรู้ด้วยเครื่องและคุณสมบัติโทโปโลยีของโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีน ซึ่งชี้ให้เห็นถึงความก้าวหน้าและข้อจำกัดของวิธีการปัจจุบัน ในปัจจุบันมีการสร้างฐานข้อมูลของยีนที่สำคัญ (a database of essential genes: DEG) [1] ทั้งหมดที่มีอยู่ในปัจจุบันไว้ เพราะฉะนั้นการวิเคราะห์หายีนที่สำคัญจึงสามารถช่วยในการตอบคำถามเกี่ยวกับสิ่งที่เป็นหน้าที่พื้นฐานที่จำเป็นต่อการดำรงชีวิตของเซลล์ได้

จากการศึกษาเกี่ยวกับการระบุยีนที่สำคัญในแบคทีเรียเพื่อระบุเป้าหมายยาที่มีประสิทธิภาพในการยับยั้งการเจริญเติบโตของแบคทีเรีย โดยใช้คุณสมบัติลักษณะทางโทโปโลยีของโครงข่ายทางชีววิทยา [2] พบว่าคุณสมบัติลักษณะทางโทโปโลยีหนึ่งที่นำมาใช้คือ ความเป็นศูนย์กลาง (Centrality) ประกอบด้วย ความเป็นจุดศูนย์กลางดีกรีหรือความเป็นจุดศูนย์กลางโดยวัดจากระดับ (Degree centrality), ความเป็นจุดศูนย์กลางโดยวัดจากความใกล้ชิด (Closeness centrality), ความเป็นจุดศูนย์กลางโดยวัดจากการคั่นกลาง (Betweenness centrality) และความเป็นจุดศูนย์กลางโดยวัดจากเวกเตอร์ลักษณะเฉพาะ (Eigenvector centrality) ของโหนดยีนในโครงข่ายทางชีววิทยาที่ศึกษา ซึ่งจากการศึกษาด้วยความเป็นจุดศูนย์กลางโดยวัดจากระดับ เป็นการค้นหาว่ายีนใดบ้างที่เป็นจุดศูนย์กลางของการเชื่อมโยง (Hub)

ซึ่งถือเป็นตำแหน่งที่มีอิทธิพลสูงสุดในโครงข่าย วัดได้จากจำนวนเส้นเชื่อมโยงทั้งหมดที่โยงเป็นโหนดยีนนั้น โดยพบว่าหากยีนตัวใดที่เป็นจุดศูนย์กลางของการเชื่อมโยง เป็นไปได้ว่ายีนนั้นเป็นยีนที่สำคัญ [2] และ ความเป็นศูนย์กลาง อื่นๆมีแนวโน้มในลักษณะเดียวกันด้วย [2] จะเห็นได้ว่าการใช้คุณลักษณะทางโทโปโลยีของโหนดในโครงข่ายมีประสิทธิภาพเพียงพอที่จะใช้ระบุยีนที่สำคัญ โดยเฉพาะความเป็นจุดศูนย์กลางดีกรี อย่างไรก็ตามโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีนสามารถจำลองได้ด้วยกราฟถ่วงน้ำหนัก ทำให้การพิจารณาดีกรีของโหนดสามารถพิจารณาด้วยดีกรีแบบถ่วงน้ำหนัก และสามารถขยายการประยุกต์ใช้กับการค้นหาเพื่อนบ้านใกล้ที่สุด ($\boldsymbol{k}$-Nearest Neighbors Search ($\boldsymbol{kNN}$)) และการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน (Reverse $\boldsymbol{k}$-Nearest Neighbors Search ($\boldsymbol{RkNN}$)) ได้

วิธีการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน [3] เป็นวิธีการหาโหนดเพื่อนบ้านที่ได้รับอิทธิพลจากโหนดของข้อมูลที่นำมาวิเคราะห์ในโครงข่าย สำหรับโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีนมีการศึกษาและประยุกต์ใช้ในการอนุมานความสัมพันธ์ระหว่างโปรตีนและโรค โดยอาศัยการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน ใช้เพื่อระบุผลกระทบของโปรตีนที่สนใจต่อโปรตีนอื่นในโครงข่าย จากนั้นความสัมพันธ์ระหว่างโปรตีนและโรคจะถูกอนุมานได้โดยใช้วิธีทดสอบข้อมูลทางสถิติ [4]

ในงานวิจัยนี้ ผู้ดำเนินการจึงสนใจที่จะศึกษาและประยุกต์ใช้การค้นหาเพื่อนบ้านใกล้ที่สุดผกผันเข้ามาช่วยหายีนที่สำคัญในโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีนในแบคทีเรีย โดยจะนำข้อมูลโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีนในแบคทีเรีย และข้อมูลยีนที่สำคัญ มาวิเคราะห์ร่วมกันเพื่ออนุมานหายีนหรือโปรตีนที่สำคัญตัวใหม่ ซึ่งจะเป็นประโยชน์ต่อการพัฒนายาต้านเชื้อแบคทีเรียต่อไปได้ในอนาคต

1.2 วัตถุประสงค์

เพื่อประยุกต์ใช้การค้นหาเพื่อนบ้านใกล้ที่สุดผกผันบนโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีนในการระบุยีนที่สำคัญในแบคทีเรีย

1.3 ขอบเขตของงานวิจัย

1.3.1 สนใจศึกษาการระบุยีนที่สำคัญของแบคทีเรีย Escherichia coli เท่านั้น

1.3.2 ข้อมูลยีนที่สำคัญได้มาจาก DEG: a database of essential genes.

1.3.3 ข้อมูลโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีน STRING database

1.4. ขั้นตอนการดำเนินงาน

1.4.1 ศึกษาค้นคว้าทฤษฎี

1.4.2 ศึกษางานวิจัยที่เกี่ยวข้อง

1.4.3 กำหนดแนวทางในการดำเนินการค้นหาข้อมูลและวิเคราะห์ข้อมูล

1.4.4 สร้างโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีนของแบคทีเรีย

1.4.5 เขียนโปรแกรมเพื่อระบุยีนที่สำคัญโดยวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน

1.4.6 ประเมินประสิทธิภาพกับวิธีการอื่น

1.4.7 ปรับปรุงแก้ไขและอภิปรายผล

1.4.8 จัดทำเอกสารรายงานฉบับสมบูรณ์

ตารางการดำเนินงาน

ขั้นตอนการดำเนินการ	ปี 2561					ปี 2562			
	ส.ค.	ก.ย.	ต.ค.	พ.ย.	ธ.ค.	ม.ค.	ก.พ.	มี.ค.	เม.ย.
1.ศึกษาค้นคว้าทฤษฎี									
2.ศึกษางานวิจัยที่เกี่ยวข้อง									
3.กำหนดแนวทางในการดำเนินการค้นหาข้อมูลและวิเคราะห์ข้อมูล									
4.สร้างโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีนของแบคทีเรีย									
5.เขียนโปรแกรมเพื่อระบุยีนที่สำคัญ โดย วิธีการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน									
6.ประเมินประสิทธิภาพกับวิธีการอื่น									
7.ปรับปรุงแก้ไขและอภิปรายผล									
8. จัดทำเอกสารรายงานฉบับสมบูรณ์									

1.5. ประโยชน์ที่คาดว่าจะได้รับ

ประโยชน์ต่อนิสิตที่ทำโครงงาน

1. ได้ความรู้และความเข้าใจเกี่ยวกับวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดผกผันมากยิ่งขึ้น
2. ได้ความรู้และความเข้าใจเกี่ยวกับการระบุยีนที่สำคัญบนโครงข่ายโปรตีนมากยิ่งขึ้น
3. ได้เพิ่มประสบการณ์การเขียนโปรแกรม R สำหรับการทำงานวิเคราะห์ข้อมูลซึ่งเป็นที่นิยมในปัจจุบัน

ประโยชน์ที่ได้จากโครงงานที่พัฒนาขึ้น

1. ได้วิธีการระบุยีนที่สำคัญบนโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีนโดยการวิเคราะห์ข้อมูลผ่านวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน
2. สามารถนำแนวคิดที่ได้ไปพัฒนาต่อในการระบุยีนที่สำคัญของสิ่งมีชีวิตอื่นๆต่อไป

1.6. โครงสร้างของรายงาน

รายงานฉบับนี้จะกล่าวถึงภาพรวมของการพัฒนาโครงงานนี้ตามลำดับดังนี้

บทที่ 2 ความรู้พื้นฐานและทฤษฎีที่เกี่ยวข้อง ที่นำมาประยุกต์ใช้ในการทำโครงงาน

บทที่ 3 จะกล่าวถึงขั้นตอนการสร้างโปรแกรมเพื่อระบุยีนที่สำคัญโดยวิธีการค้นหาเพื่อนบ้านที่สุดผกผัน

บทที่ 4 เป็นการเปรียบเทียบประสิทธิภาพกับวิธีการอื่น

บทที่ 5 จะกล่าวถึงข้อสรุป และข้อเสนอแนะระหว่างการทำโครงงาน

บทที่ 2
ความรู้พื้นฐานและทฤษฎีบทที่เกี่ยวข้อง

ความรู้พื้นฐานและทฤษฎีบทที่เกี่ยวข้องที่ผู้วิจัยนำมาประยุกต์ใช้กับการดำเนินงานนั้น ได้แก่ ความรู้พื้นฐานทางคณิตศาสตร์ ยีนที่สำคัญ ความสัมพันธ์ระหว่างยีนและโปรตีน ปฏิสัมพันธ์ระหว่างโปรตีน ทฤษฎีกราฟ ความเป็นศูนย์กลาง การค้นหาเพื่อนบ้านใกล้ที่สุด การค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน และ คอนฟิวชันเมทริกซ์ ซึ่งจะได้กล่าวไว้ในบทนี้

2.1 ยีนที่สำคัญ

ยีนที่สำคัญคือยีนของสิ่งมีชีวิตที่มีความสำคัญต่อการอยู่รอด อย่างไรก็ตามการเป็นสิ่งจำเป็นนั้นขึ้นอยู่กับสถานการณ์ที่สิ่งมีชีวิตอาศัยอยู่ ตัวอย่างเช่นยีนที่สำคัญในการย่อยแป้งเป็นสิ่งจำเป็นเฉพาะถ้าหากแป้งเป็นแหล่งพลังงานเท่านั้น ซึ่งในปัจจุบันได้มีการพยายามในการระบุยีนที่สำคัญในการย่อยแป้งเหล่านั้นในการรักษาชีวิตโดยมีเงื่อนไขว่าสารอาหารทั้งหมดนั้นยังคงมีอยู่ [1] การทดลองดังกล่าวนำไปสู่ข้อสรุปว่าจำนวนยีนที่สำคัญสำหรับแบคทีเรียอยู่ในลำดับประมาณ 250–300 ยีน สำคัญเหล่านี้เข้ารหัสโปรตีนเพื่อรักษาเมตาบอลิซึมส่วนกลางทำซ้ำดีเอ็นเอแปลยีนเข้าสู่โปรตีนรักษาโครงสร้างเซลล์พื้นฐานและทำหน้าที่เป็นสื่อกลางในกระบวนการขนส่งเข้าและออกจากเซลล์ ยีนส่วนใหญ่ที่พบเป็นยีนไม่สำคัญ แต่เป็นยีนที่ได้เปรียบในการเลือกถ่ายทอดและความแข็งแรงที่เพิ่มมากขึ้น

วิธีการทดลองสำหรับการระบุยีนที่สำคัญในแบคทีเรีย เป็นการศึกษาบนจีโนมของแบคทีเรียโดยการใช้การยับยั้งยีน ยีนที่ต้องการศึกษาจะถูกลบออกจากจีโนมอย่างเป็นระบบ จากนั้นสังเกตการเจริญเติบโตของแบคทีเรียปกติ (Wild type) และแบคทีเรียที่ถูกลบยีนออก (Mutant) หากแบคทีเรียที่ถูกลบยีนออกไม่สามารถที่จะอยู่รอดหรือเติบโตได้จะจัดเป็นยีนที่สำคัญ

2.2 ความสัมพันธ์ระหว่างยีนและโปรตีน

ยีนคือส่วนหนึ่งของสายดีเอ็นเอซึ่งเป็นโมเลกุลที่ประกอบด้วยนิวคลีโอไทด์สี่ชนิดเชื่อมต่อกันเป็นสายยาว ลำดับนิวคลีโอไทด์สี่ชนิดนี้คือข้อมูลทางพันธุกรรมที่ถูกเก็บและมีการถ่ายทอดในสิ่งมีชีวิต ดีเอ็นเอตามธรรมชาติอยู่ในรูปเกลียวคู่ โดยนิวคลีโอไทด์บนแต่ละสายจะเป็นคู่สมซึ่งกันและกันกับนิวคลีโอไทด์บนสายดีเอ็นเออีกสายหนึ่ง แต่ละสายทำหน้าที่เป็นแม่แบบในการสร้างสายคู่ขึ้นมาได้ใหม่ นี่คือกระบวนการทางกายภาพที่ทำให้ยีนสามารถจำลองตัวเอง และถ่ายทอดไปยังรุ่นลูกได้ ลำดับของนิวคลีโอไทด์ในยีนจะถูกแปลออกมาเป็นสายของกรดอะมิโน ประกอบกันเป็นโปรตีน ซึ่งลำดับของกรดอะมิโนที่มาประกอบกันเป็นโปรตีนนั้นถ่ายทอดออกมาจากลำดับของนิวคลีโอไทด์บนดี

เอ็นเอ ความสัมพันธ์ระหว่างลำดับของนิวคลีโอไทด์และลำดับของกรดอะมิโนนี้เรียกว่ารหัสพันธุกรรม กรดอะมิโนแต่ละชนิดที่ประกอบขึ้นมาเป็นโปรตีนช่วยกำหนดว่าสายโซ่ของกรดอะมิโนนั้นจะพับม้วนเกิดเป็นโครงสร้างสามมิติอย่างไร โครงสร้างสามมิตินี้กำหนดหน้าที่ของโปรตีนนั้น ๆ ซึ่งโปรตีนมีหน้าที่ในกระบวนการเกือบทั้งหมดของเซลล์สิ่งมีชีวิต การเปลี่ยนแปลงที่เกิดกับดีเอ็นเอในยีนยีนหนึ่ง อาจทำให้เกิดการเปลี่ยนแปลงลำดับกรดอะมิโนในโปรตีน เปลี่ยนโครงสร้างโปรตีน เปลี่ยนการทำหน้าที่ของโปรตีน ซึ่งอาจส่งผลต่อเซลล์และสิ่งมีชีวิตนั้น ๆ ได้อย่างมาก

2.3 ปฏิสัมพันธ์ระหว่างโปรตีน (Protein-protein interaction)

ความสัมพันธ์ของยีนสามารถถูกแสดงออกในรูปแบบของปฏิสัมพันธ์กันของยีนในโครงข่าย ลักษณะของโครงข่ายที่นิยมใช้กัน คือ ปฏิสัมพันธ์ระหว่างโปรตีน กำหนดให้ $G(V,E)$ แทนกราฟ ซึ่งประกอบด้วย V แทนเซตของโหนด และ $\mathrm{E} \subseteq \mathrm{V} \times \mathrm{V}$ แทนเซตของเส้นเชื่อม ในที่นี้หาก $G(V,E)$ แทนโครงข่ายปฏิสัมพันธ์ของโปรตีน โหนดของโครงข่ายจะหมายถึงโปรตีนและ เส้นเชื่อมของโครงข่ายหมายถึงปฏิสัมพันธ์ระหว่างโปรตีนนั่นเอง โดยโครงข่ายปฏิสัมพันธ์นี้จะมีลักษณะเป็นกราฟไม่ระบุทิศทาง (Undirected graph) เพราะเส้นเชื่อมนั้นแสดงเพียงปฏิสัมพันธ์ระหว่างโปรตีน ยกเว้นบางงานวิจัย (Suratanee และคณะ, 2014) ที่ได้มีการกำหนดสถานะของเส้นเชื่อมไว้แตกต่างกันเช่น กำหนดให้เป็นสถานะ activation และ inhibition ข้อมูลปฏิสัมพันธ์ระหว่างโปรตีนปัจจุบันได้มาจากข้อมูลทางการทดลอง เช่น yeast two-hybrid screening (Stelzl และคณะ, 2005) และ affinity capture mass spectrometry (Krogan และคณะ,2006). เป็นต้น ข้อมูลเหล่านี้ได้ถูกรวบรวมและเก็บไว้เป็นฐานข้อมูล เช่น HPRD (Prasad และคณะ, 2009), BIND (Liu และคณะ, 2007), STRING (Franceschini และคณะ, 2013) เป็นต้น โดยมีการแจกจ่ายให้ใช้โดยไม่เสียค่าใช้จ่ายเพื่อการศึกษา โดยพบว่าฐานข้อมูลส่วนใหญ่นี้จะมีขนาดใหญ่ กล่าวคือจะมีจำนวนโหนดมากกว่า 10,000 โหนดและเส้นเชื่อมมากกว่า 60,000 เส้น

2.4 ทฤษฎีกราฟ

กราฟ (Graph) เป็นแบบจำลองทางคณิตศาสตร์คิดค้นโดยนักคณิตศาสตร์ชาวสวิตเซอร์แลนด์ เลออนฮาร์ดออยเลอร์ (Leonhard Euler) กราฟสามารถใช้แทนปัญหาในโลกของความเป็นจริง โดยจำลองปัญหาด้วยแผนภาพที่ประกอบด้วยจุด (Point) หรือเรียกว่าโหนด (Node) และเส้นที่เชื่อมระหว่างจุด 2 จุด หรือเส้นเชื่อม (Edge) ตัวอย่าง เช่น แผนภาพแสดงเส้นทางการบิน แผนภาพแสดงเส้นทางรถไฟฟ้าและวงจรไฟฟ้า อีกทั้งปัจจุบันยังมีการนำไปประยุกต์ใช้ในด้านต่าง ๆ อย่างกว้างขวาง เช่น ปัญหาด้านจิตวิทยา ภาษาศาสตร์ เทคโนโลยีคอมพิวเตอร์และเศรษฐศาสตร์ซึ่ง

ในงานวิจัยฉบับนี้ กราฟถูกนำมาใช้ในการจำลองความสัมพันธ์ของคำ ในข้อความ โดยกราฟ (G) ประกอบด้วยโหนดสมาชิก (V) และเส้นเชื่อมระหว่างโหนด (E) แสดงดังสมการ

$$G = (V, E) \quad (2.1)\ [5]$$

โดยที่ V เป็นเซตจำกัดที่ไม่เป็นเซตว่างของสมาชิกที่เรียกว่าจุดยอด หรือ โหนด ; E เป็นเซตของเส้นเชื่อมระหว่างโหนด

2.5 ความเป็นศูนย์กลาง (Centrality)

การวัดค่าความเป็นศูนย์กลางมีสามประเภท ได้แก่ ความเป็นจุดศูนย์กลางโดยวัดจากระดับ และความเป็นจุดศูนย์กลางโดยวัดจากการคั่นกลาง

2.5.1 ความเป็นจุดศูนย์กลางโดยวัดจากระดับ (Degree Centrality)

ความเป็นจุดศูนย์กลางโดยวัดจากระดับ อยู่บนพื้นฐานความคิดที่ว่าศูนย์รวมกิจกรรมซึ่งจัดว่าเป็นศูนย์กลางในเครือข่ายนั้น คือเหล่าศูนย์รวมกิจกรรมที่มีความเชื่อมโยงเป็นส่วนใหญ่กับศูนย์รวมกิจกรรมอื่นๆ กำหนดได้จากการนับ รวมจำนวนทิศทางของเส้นเชื่อมโยงที่เข้ามาสู่ศูนย์รวมกิจกรรมนั้นๆ จากศูนย์รวมกิจกรรมอื่นทั้งหมดใน เครือข่าย ศูนย์รวมกิจกรรมที่มีค่าความเป็นศูนย์กลางสูงย่อมถือว่าเป็นศูนย์รวมกิจกรรมที่มีระดับของ กิจกรรมในเครือข่ายมากตามไปด้วย (Wasserman & Faust 1994) ศูนย์รวมกิจกรรมซึ่งมีความเชื่อมโยง กับศูนย์รวมกิจกรรมอื่นๆ ในเครือข่ายเป็นจำนวนมากอาจอยู่ในตำแหน่งที่เอื้ออำนวยประโยชน์ให้แก่ศูนย์ รวมกิจกรรมต่างๆ ได้ ขณะเดียวกันอาจจะมีการพึ่งพาต่อศูนย์รวมกิจกรรมอื่นๆ ค่อนข้างน้อยเพราะ สามารถเข้าถึงทรัพยากรในเครือข่ายได้ดีกว่า (Hanneman & Riddle 2005) โดยสมการทางคณิตศาสตร์ของความเป็นจุดศูนย์กลางโดยวัดจากระดับ คือ

$$d(i) = \sum_j m_{ij} \quad (2.2)\ [6]$$

โดยที่ $d(i)$ คือ ความเป็นจุดศูนย์กลางโดยวัดจากระดับของศูนย์รวมกิจกรรม i โดย $m_{ij} = 1$ ถ้ามีการเชื่อมต่อระหว่างศูนย์รวมกิจกรรม i และ j หรือ $m_{ij} = 0$ ถ้าไม่มีการเชื่อมต่อระหว่างกัน

2.5.2 ความเป็นจุดศูนย์กลางโดยวัดจากการคั่นกลาง (Betweenness Centrality)

ความเป็นจุดศูนย์กลางโดยวัดจากการคั่นกลาง อยู่บนสมมติฐานที่ว่าศูนย์รวมกิจกรรมใดก็ตามซึ่งอยู่ระหว่างกลางการเชื่อมโยงของศูนย์รวมกิจกรรมอื่นๆ ก็เป็นเสมือนศูนย์รวมกิจกรรมศูนย์กลางของเครือข่าย เพราะศูนย์รวมกิจกรรมในตำแหน่งนี้สามารถควบคุมการมีปฏิสัมพันธ์ของศูนย์รวมกิจกรรมต่างๆ ที่มาเชื่อมโยงผ่านตัวเองได้ โดยสามารถสวมบทบาทผู้ควบคุมประตูแห่งความสัมพันธ์ หรือทำการกีดขวางการติดต่อจากสิ่งที่ไม่พึงปรารถนา (Wasserman & Faust 1994) ซึ่งทำการวัดจากความเชื่อมโยงทางอ้อมระหว่างศูนย์รวมกิจกรรมต่างๆ ในเครือข่าย การที่มีค่าความเป็นจุดศูนย์กลางโดยวัดจากระดับสูงอาจหมายถึง "ภาวะซ่อนเร้นจากผิวนอกของเครือข่าย" (Durland & Fredericks 2005) ศูนย์รวมกิจกรรมเหล่านี้ทาหน้าที่คล้ายกับตัวกลางในการจัดสรรอำนาจเพราะอยู่บนวิถีซึ่งให้โอกาสแก่ศูนย์รวมกิจกรรมอื่นๆ แม้ว่าจะไม่มีการเชื่อมต่อโดยตรงก็ตาม (Durland & Fredericks 2005) โดยสมการทางคณิตศาสตร์ความเป็นจุดศูนย์กลางโดยวัดจากการคั่นกลาง คือ

$$b(i) = \sum_{j,k} \frac{g_{jk}}{g_{jik}} \qquad (2.3)\ [6]$$

โดยที่ $b(i)$ คือ ความเป็นจุดศูนย์กลางโดยวัดจากการคั่นกลางของศูนย์รวมกิจกรรม i

g_{jk} คือจำนวนเส้นทางที่สั้นที่สุดจากศูนย์รวมกิจกรรม i ไปสู่ศูนย์รวมกิจกรรม $j (j, k \neq i)$

g_{jik} คือจำนวนเส้นทางที่สั้นที่สุดจากศูนย์รวมกิจกรรม i ไปยังศูนย์รวมกิจกรรม j ที่ต้องผ่าน i

2.6 การค้นหาเพื่อนบ้านใกล้ที่สุด (k-Nearest Neighbors Search (kNN))

ขั้นตอนวิธีการค้นหาเพื่อนบ้านใกล้ที่สุด เป็นวิธีที่ใช้ในการจัดแบ่งคลาส โดยเทคนิคนี้จะตัดสินใจว่า คลาสใดที่จะแทนเงื่อนไขหรือกรณีใหม่ๆ ได้บ้าง โดยการตรวจสอบจำนวนบางจำนวน (“k” ในขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด) ของกรณีหรือเงื่อนไขที่เหมือนกันหรือใกล้เคียงกันมากที่สุด โดยจะหาผลรวม (Count Up) ของจำนวนเงื่อนไข หรือกรณีต่างๆ สำหรับแต่ละคลาส และกำหนดเงื่อนไขใหม่ๆ ให้คลาสที่เหมือนกันกับคลาสที่ใกล้เคียงกันมากที่สุด

ขั้นตอนวิธีการค้นหาเพื่อนบ้านใกล้ที่สุด

การนำเทคนิคของขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุดไปใช้นั้น เป็นการหาระยะห่างระหว่างแต่ละตัวแปร (Attribute) ในข้อมูล จากนั้นก็คำนวณค่าออกมา ซึ่งวิธีนี้จะเหมาะสำหรับข้อมูลแบบตัวเลข แต่ตัวแปรที่เป็นค่าแบบไม่ต่อเนื่องนั้นก็สามารถทำได้ เพียงแต่ต้องการการจัดการแบบพิเศษเพิ่มขึ้น อย่างเช่น ถ้าเป็นเรื่องของสี เราจะใช้อะไรวัดความแตกต่างระหว่างสีน้ำเงินกับสีเขียว ต่อจากนั้นเราต้องมีวิธีในการรวมค่าระยะห่างของ Attribute ทุกค่าที่วัดมาได้ เมื่อสามารถคำนวณระยะห่างระหว่างเงื่อนไขหรือกรณีต่างๆ ได้ จากนั้นก็เลือกชุดของเงื่อนไขที่ใช้จัดคลาส มาเป็นฐานสำหรับการจัดคลาสในเงื่อนไขใหม่ๆ ได้แล้วเราจะตัดสินได้ว่าขอบเขตของจุดข้างเคียงที่ควรเป็นนั้นควรมีขนาดใหญ่เท่าไร และอาจมีการตัดสินใจได้ด้วยว่าจะนับจำนวนจุดข้างเคียงตัวมันได้อย่างไร โดยขั้นตอนวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดมีขั้นตอนโดยสรุป ดังนี้

1. กำหนดขนาดของ k (ควรกำหนดให้เป็นเลขคี่)
2. คำนวณระยะห่าง (Distance) ของข้อมูลที่ต้องการพิจารณากับกลุ่มข้อมูลตัวอย่าง
3. จัดเรียงลำดับของระยะห่าง และเลือกพิจารณาชุดข้อมูลที่ใกล้จุดที่ต้องการพิจารณาตามจำนวน k ที่กำหนดไว้
4. พิจารณาข้อมูลจำนวน k ชุด และสังเกตว่ากลุ่ม (class) ไหนที่ใกล้จุดที่พิจารณาเป็นจำนวนมากที่สุด
5. กำหนด class ให้กับจุดที่พิจารณา โดยจัดให้เป็น class เดียวกับ class ที่ใกล้จุดพิจารณามากที่สุด

สำหรับโครงข่ายปฏิสัมพันธ์โปรตีนระหว่างโปรตีนการค้นหา kNN ถูกใช้ในการค้นหาโปรตีนที่มีอิทธิพลต่อโปรตีนที่สนใจ น้ำหนักของโครงข่ายปฏิสัมพันธ์โปรตีนระหว่าง โปรตีนคือคะแนนความเชื่อมั่นจากฐานข้อมูล STRING ดังนั้นระยะทางที่เชื่อมต่อระหว่าง 2 โปรตีนจึงเป็นการกลับกันของคะแนนความมั่นใจระหว่าง 2 โปรตีน สูตรของการค้นหา kNN มีดังนี้

โดย $distance(q,p)$ เป็นระยะทางระหว่างโปรตีน q และ โปรตีน p

$distance(q,\bar{p})$เป็นระยะทางระหว่างโปรตีน q และ โปรตีน $\bar{p}$

P เป็นเซตของโปรตีนทั้งหมดในโครงข่าย

$kNN(q)$ เป็นเซตของโปรตีนเพื่อนบ้านที่ใกล้โปรตีน q ที่สุด k ตัว โดยที่

$$\forall p \in kNN(q), \forall \bar{p} \in (P - kNN(q))\{distance(q,p) < distance(q,\bar{p})\} \quad (2.4)\ [7]$$

2.7 การค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน (Reverse k-Nearest Neighbors Search ($RkNN$))

การค้นหาเพื่อนบ้านใกล้ที่สุดผกผันเป็นวิธีที่ใช้ในการจัดกลุ่มข้อมูล โดยการตรวจสอบความเป็นเพื่อนบ้านใกล้เคียงของจุดที่สนใจที่มีต่อข้อมูลอื่นๆ และกำหนดเงื่อนไขใหม่ให้คลาสที่เหมือนกันกับคลาสที่ใกล้เคียงกันมากที่สุด โดยปกติจะมีพารามิเตอร์ k เพื่อระบุจำนวนเพื่อนบ้านที่ใกล้เคียงที่สุดที่พิจารณาของโหนดที่สนใจ ดังนั้นเราจึงเรียกว่า $RkNN$ search โดยทั่วไป แนวคิดของการค้นหาคือการค้นหาโหนดใกล้เคียงที่ได้รับอิทธิพลจากโหนดที่สนใจ สำหรับโครงข่ายปฏิสัมพันธ์โปรตีนระหว่างโปรตีนการค้นหา $RkNN$ ถูกใช้ในการค้นหาโปรตีนที่ได้รับอิทธิพลจากโปรตีนที่สนใจน้ำหนักของโครงข่ายปฏิสัมพันธ์โปรตีนระหว่าง โปรตีนคือคะแนนความเชื่อมั่นจากฐานข้อมูล STRING ดังนั้นระยะทางที่เชื่อมต่อระหว่าง 2 โปรตีนจึงเป็นการกลับกันของคะแนนความมั่นใจระหว่าง 2 โปรตีน สูตรของการค้นหา $RkNN$ ของโปรตีนที่สนใจ q มีดังนี้

$$RkNN(q) = \{p \in \mathrm{P} | q \in kNN(p)\} \quad (2.5)\ [7]$$

ซึ่ง p ในเซตของ $RkNN(q)$ เป็นโปรตีนที่ได้รับอิทธิพลจากโปรตีน q ดังนั้นด้วยพารามิเตอร์ k เดียวกัน $RkNN$ และ kNN ของโปรตีนที่สนใจจึงมีโปรตีนหลายชุด แทนที่จะค้นหาโปรตีน k ที่ใกล้เคียงที่สุดกับโปรตีนที่สนใจ วิธีการ $RkNN$ พยายามระบุชุดของโปรตีนที่โปรตีนที่สนใจเป็น kNN ของชุดโปรตีนเหล่านั้น ดังนั้น $RkNN$ จะให้โปรตีนที่มีขนาดเล็กลงเสมอในขณะที่ kNN จะให้โปรตีนที่ใกล้เคียงที่สุด (หรือใกล้เคียง) กับโปรตีนที่สนใจ ด้วยวิธีการค้นหาเพื่อนบ้านที่ใกล้ที่สุด kNN ดังนั้นเพื่อนบ้านที่ไม่เกี่ยวข้องที่อาจไม่ส่งผลกระทบต่อโปรตีนที่สนใจอาจรวมอยู่และทำให้ความแม่นยำในการทำนายลดลงได้

2.8 คอนฟิวชันเมทริกซ์ (Confusion Matrix)

ตารางที่ 2.1 คอนฟิวชันเมทริกซ์

		Predicted Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

ความหมายของคำต่าง ๆ ในตาราง มีดังนี้

True Positive (TP) คือ จำนวนของข้อมูลจริงที่เป็น Positive และแบบจำลองจำแนกว่าเป็น Positive

True Negative (TN) คือ จำนวนของข้อมูลจริงที่เป็น Negative และแบบจำลองจำแนกว่าเป็น Negative

False Positive (FP) คือ จำนวนของข้อมูลจริงที่เป็น Negative และแบบจำลองจำแนกว่าเป็น Positive

False Negative (FN) คือ จำนวนของข้อมูลจริงที่เป็น Positive และแบบจำลองจำแนกว่าเป็น Negative

โดยสามารถคำนวณค่าต่างๆ [8] เพื่อนำมาเปรียบเทียบประสิทธิภาพดังนี้

Accuracy คือ ค่าความถูกต้อง เป็น % คำนวณจาก (TP+TN) / (TP+TN+FP+FN) * 100

Precision คือ ค่าความแม่นยำ เป็น % คำนวณจาก TP / (TP+FP) * 100

Sensitivity คือ ค่าความไว เป็น % คำนวณจาก TP / (TP+FN) * 100

Specificity คือ ค่าความจำเพาะ เป็น % คำนวณจาก TN / (TN+FP) * 100

โดยค่าความแม่นยำจะบอกประสิทธิภาพของการจำแนกแต่ละคลาสว่า ข้อมูลที่แบบจำลองจำแนกนั้นมีความถูกต้องแม่นยำมากน้อยเพียงใด ความไวจะมีประโยชน์ในการวินิจฉัยแยกกันผลลบปลอม (false negative) เพราะว่าการทดสอบยิ่งไวเท่าไร โอกาสการได้ผลลบ ที่ไม่เป็นจริงก็น้อยลงเท่านั้น และดังนั้นถ้าความไวอยู่ที่ 100% โอกาสได้ผลลบปลอมก็อยู่ที่ 0% และความจำเพาะจะมีประโยชน์ในการยืนยันภาวะที่มี โดยกันผลบวกปลอม (false positive) เพราะว่าการทดสอบยิ่งจำเพาะเท่าไร โอกาสการได้ผลบวกที่ไม่เป็นจริง ก็น้อยลงเท่านั้น

บทที่ 3
ขั้นตอนการดำเนินงาน

ในบทที่ 3 จะอธิบายถึงรายละเอียดของวิธีการดำเนินงานในการทำวิจัยครั้งนี้ ตั้งแต่การรวบรวมข้อมูล การเตรียมข้อมูลก่อนการประมวลผล การสร้างโมเดลการจำแนกประเภท รวมไปถึงการเขียนคำสั่งภาษา R บางคำสั่งที่สำคัญต่อการเปรียบเทียบวิธีการระบุยีนที่สำคัญบนโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีนแต่ละวิธี

3.1 การรวบรวมข้อมูล

ผู้จัดทำได้เก็บรวบรวมข้อมูลโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีนของแบคทีเรีย *Escherichia coli* str. K-12 substr. MG1655_511145 จาก STRING database ซึ่งเป็นฐานข้อมูลหลักที่เป็นทางการที่เก็บข้อมูลความสัมพันธ์ระหว่างโปรตีนแต่ละชนิด ซึ่งความสัมพันธ์ระหว่างโปรตีนแต่ละรายการที่เก็บไว้ใน STRING database จะมีการให้คะแนน คะแนนเหล่านี้เป็นคะแนนที่แสดงถึงความเชื่อมั่น ซึ่งมีค่าตั้งแต่ 1 ถึง 1000 ซึ่งหากโปรตีนสองชนิดมีค่าความเชื่อมั่นมากจะแสดงถึงการที่โปรตีนสองชนิดมีความสัมพันธ์กันมาก ซึ่งผู้จัดทำได้เลือกพิจารณาข้อมูลที่มีคะแนนความเชื่อมั่นอย่างน้อย 900 หรือมีคะแนนความเชื่อมั่นอย่างน้อย 90% เท่านั้น

โดยข้อมูลโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีน จะเก็บในลักษณะดังรูปต่อไปนี้

protein1	protein2	combined_score
511145.b0002	511145.b1605	915
511145.b0002	511145.b3172	962
511145.b0002	511145.b4132	915
511145.b0002	511145.b2478	986
511145.b0002	511145.b3770	958
511145.b0002	511145.b3078	915
511145.b0002	511145.b3829	917
511145.b0002	511145.b0928	910
511145.b0002	511145.b2014	915
511145.b0002	511145.b2599	944
511145.b0002	511145.b1492	915
511145.b0002	511145.b1296	915
511145.b0002	511145.b4141	915
511145.b0002	511145.b4115	915
511145.b0002	511145.b0268	979
511145.b0002	511145.b3981	950
511145.b0002	511145.b0486	915
511145.b0002	511145.b3370	915

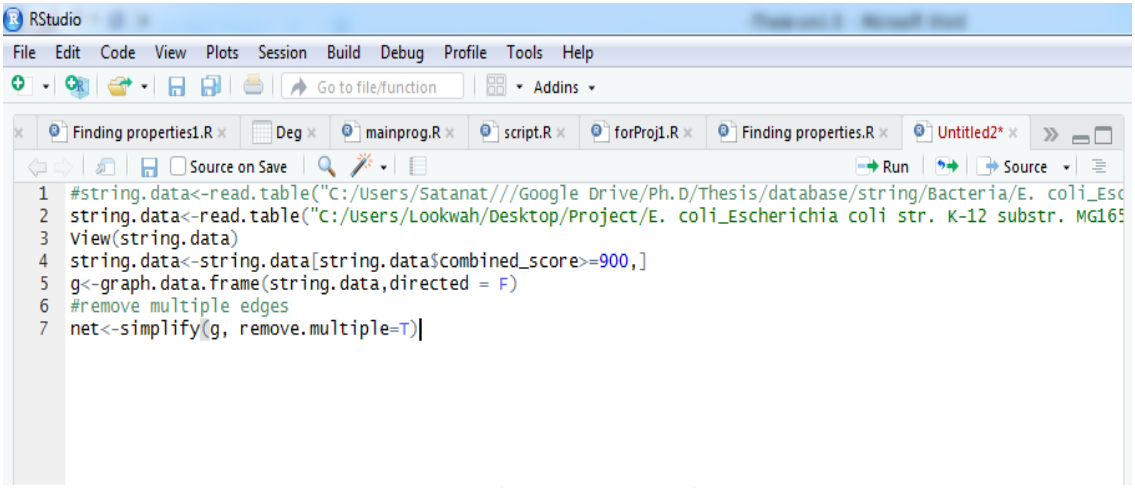
รูปที่ 3.1 แสดงตัวอย่างข้อมูลโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีน

3.2 การเตรียมข้อมูลก่อนการประมวลผล

จากการพิจารณาข้อมูลพบว่า มีทั้งหมด 3 คอลัมน์ ประกอบด้วย โปรตีนชนิดที่ 1 (protein1) โปรตีนชนิดที่ 2 (protein2) และคะแนนความเชื่อมั่น (combined_score) ซึ่งคะแนนความเชื่อมั่นในตารางจะมีค่าตั้งแต่ 0 ถึง 1000 ซึ่งผู้จัดทำได้เลือกพิจารณาเฉพาะความสัมพันธ์ระหว่างโปรตีนที่มีความเชื่อมั่น 90% ขึ้นไปเท่านั้น นั่นก็คือมีคะแนนความเชื่อมั่น (combined_score) ตั้งแต่ 900 คะแนนขึ้นไป

เนื่องจากการระบุยีนที่สำคัญบนโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีน เป็นการคำนวณที่ใช้ข้อมูลค่อนข้างมากและมีความซับซ้อน เพราะฉะนั้นโปรแกรมหนึ่งที่รองรับการคำนวณและจัดการข้อมูลได้คือ RStudio ซึ่งเป็นโปรแกรมที่ใช้ภาษา R ซึ่งทางสถิติใช้กันอย่างแพร่หลาย สะดวกและไม่เสียค่าใช้จ่าย

นำข้อมูลลงไปใน RStudio เขียนคำสั่งโปรแกรม RStudio เพื่อเลือกพิจารณาเฉพาะความสัมพันธ์ระหว่างโปรตีนที่มีความเชื่อมั่น 90% ขึ้นไปเท่านั้น จากรูปภาพที่ 3.2 แสดงตัวอย่างคำสั่งในการเลือกชุดข้อมูลปฏิสัมพันธ์ระหว่างโปรตีนมีความเชื่อมั่น 90% ขึ้นไปเท่านั้น



รูปภาพที่ 3.2 ตัวอย่างคำสั่ง

3.3 การสร้างโมเดลการจำแนกประเภท (Classification Model Creation)

หลังจากนั้นนำชุดข้อมูลโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีนเข้าในโปรแกรม Rstudio เพื่อนำไปทดสอบการสร้างโมเดลการจำแนกประเภทด้วยอัลกอริทึมแบบต่างๆ ได้แก่ ความเป็นจุดศูนย์กลางโดยวัดจากระดับ, การค้นหาเพื่อนบ้านใกล้ที่สุด, การค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน

3.3.1 การระบุยีนที่สำคัญโดยใช้ความเป็นจุดศูนย์กลางโดยวัดจากระดับ (Degree Centrality)

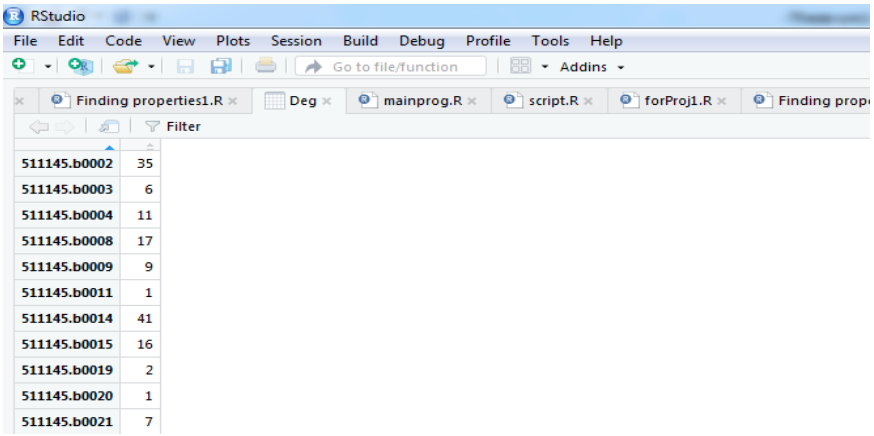
สร้างโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีนของแบคทีเรียด้วยการใช้ โปรแกรม Rstudio จากข้อมูลการเชื่อมโยงกับโปรตีนอื่นของโปรตีนที่สนใจ โดยถ้าโปรตีนใดมีความเชื่อมโยงกันค่าเท่ากับ 1 ถ้าโปรตีนใดไม่มีความเชื่อมโยงกันค่าเท่ากับ 0 และนำค่าที่ได้รวมออกมาเป็นดีกรีรวมของโปรตีนสำหรับโปรตีนตัวอย่างนั้นๆ จากรูปภาพที่ 3.3 แสดงตัวอย่างคำสั่งในการหาดีกรีของโปรตีน

```
#string.data<-read.table("C:/Users/Satanat///Google Drive/Ph.D/Thesis/database/string/Bacteria/E. coli_Es
string.data<-read.table("C:/Users/Lookwah/Desktop/Project/E. coli_Escherichia coli str. K-12 substr. MG16
View(string.data)
string.data<-string.data[string.data$combined_score>=900,]
g<-graph.data.frame(string.data,directed = F)
#remove multiple edges
net<-simplify(g, remove.multiple=T)

#Finding properties of network
#Degree
#Node <- V(net)$name
Deg <- degree(net)
```

รูปภาพที่ 3.3 แสดงตัวอย่างคำสั่งในการหาดีกรีของโปรตีน

รูปภาพที่ 3.4 แสดงตัวอย่างผลค่าดีกรีของโปรตีนแต่ละตัว โดยแสดงเป็น 2 คอลัมน์ ประกอบด้วยชื่อของโปรตีนแต่ละตัว และค่าดีกรีของโปรตีนตัวนั้นๆ



511145.b0002	35
511145.b0003	6
511145.b0004	11
511145.b0008	17
511145.b0009	9
511145.b0011	1
511145.b0014	41
511145.b0015	16
511145.b0019	2
511145.b0020	1
511145.b0021	7

รูปภาพที่ 3.4 แสดงตัวอย่างดีกรีของโปรตีน

หลังจากได้ผลลัพธ์ของดีกรี ขั้นต่อไปเป็นการเขียนคำสั่งเพื่อระบุยีนที่สำคัญโดยการตัดสินใจว่าจะทำนายเป็นคลาสใดแบ่งเป็น 3 วิธีคือ

1) ขึ้นอยู่กับว่าถ้า ดีกรี มากกว่าหรือเท่ากับ ควอไทล์ที่ 1 ของจำนวนดีกรีทั้งหมด จะระบุว่าโปรตีนชนิดนั้นเป็นโปรตีนที่สำคัญ
2) ขึ้นอยู่กับว่าถ้า ดีกรี มากกว่าหรือเท่ากับ ควอไทล์ที่ 2 ของจำนวนดีกรีทั้งหมด จะระบุว่าโปรตีนชนิดนั้นเป็นโปรตีนที่สำคัญ
3) ขึ้นอยู่กับว่าถ้า ดีกรี มากกว่าหรือเท่ากับ ควอไทล์ที่ 3 ของจำนวนดีกรีทั้งหมด จะระบุว่าโปรตีนชนิดนั้นเป็นโปรตีนที่สำคัญ

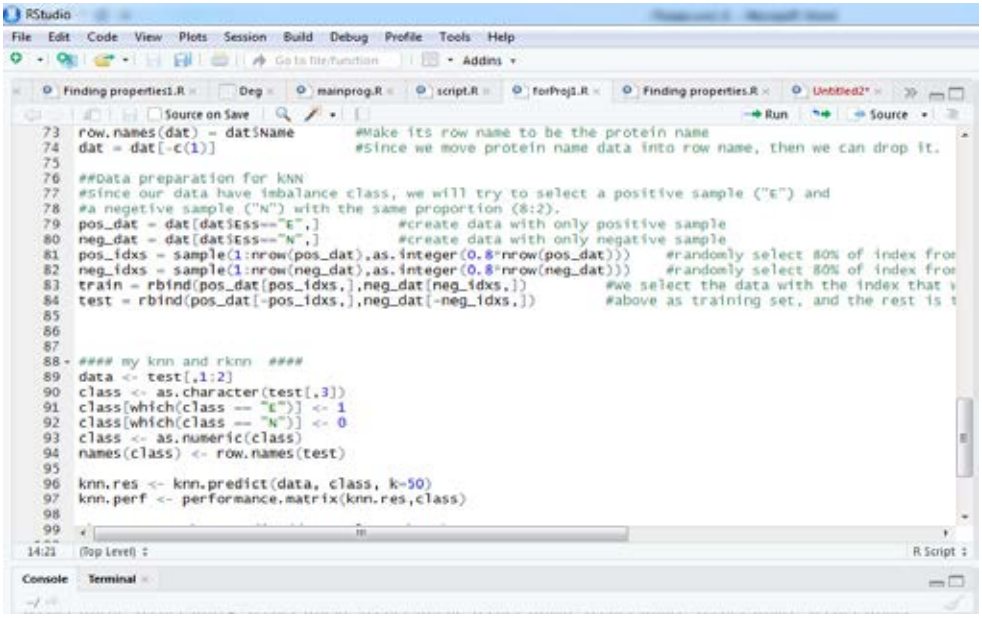
จากนั้นจะนำผลลัพธ์ของการทำนายคลาสของยีนว่าเป็นยีนที่สำคัญหรือเป็นยีนที่ไม่สำคัญที่ได้ไปเปรียบเทียบกับผลการทดลองการระบุยีนที่สำคัญที่ได้จริงจากการทดลองในห้องปฏิบัติการ เพื่อวัดประสิทธิภาพและนำไปแสดงผลต่อไป

3.3.2 การระบุยีนที่สำคัญโดยใช้การค้นหาเพื่อนบ้านใกล้ที่สุด

ขั้นตอนวิธีการค้นหาเพื่อนบ้านที่ใกล้สุดจะทำการจำแนกข้อมูลที่ได้รับมาใหม่โดยอ้างอิงจากข้อมูลเดิมที่มีคุณสมบัติที่ใกล้เคียงที่สุดจำนวน k ตัวโดยจะมีขั้นตอนดังต่อไปนี้

1. กำหนดขนาดของ k
2. คำนวณระยะห่าง (Distance) ระหว่างแต่ละตัวแปร(Attribute) ของข้อมูลที่ต้องการพิจารณากับกลุ่มข้อมูลตัวอย่างโดยผู้จัดทำได้เลือกตัวแปร(Attribute) ที่นำมาคำนวณคือความเป็นจุดศูนย์กลางโดยวัดจากระดับ (Degree Centrality) และความเป็นจุดศูนย์กลางโดยวัดจากการคั่นกลาง (Betweenness Centrality)
3. จัดเรียงลำดับของระยะห่าง และเลือกพิจารณาชุดข้อมูลที่ใกล้จุดที่ต้องการพิจารณาตามจำนวน k ที่กำหนดไว้
4. พิจารณาข้อมูลจำนวน k ชุด และสังเกตว่ากลุ่ม (class) ไหนที่ใกล้จุดที่พิจารณาเป็นจำนวนมากที่สุด
5. กำหนด class ให้กับจุดที่พิจารณา โดยการตัดสินจว่าจะระบุเป็นคลาสใด ขึ้นอยู่กับว่า neighbors เป็น essential มากกว่า 25% ของจำนวน Neighbor ทั้งหมด เนื่องจากข้อมูลจริงจากการทดลองโปรตีนที่เป็นโปรตีนที่สำคัญพบประมาณ 10% ของยีนทั้งหมด ดังนั้นหากเราพบว่าโปรตีนที่พิจารณามีเพื่อนบ้านใกล้เคียงเป็นโปรตีนที่สำคัญมากกว่า 25% ของจำนวน Neighbor ทั้งหมด จึงถือว่ามีโอกาสที่จะเป็นโปรตีนที่สำคัญสูง

ซึ่งผู้จัดทำได้คำนวณโดยการเขียนคำสั่งในโปรแกรม Rstudio จากรูปภาพที่ 3.5 แสดงตัวอย่างคำสั่งในการหาเพื่อนบ้านใกล้ที่สุด



รูปภาพที่ 3.5 แสดงตัวอย่างคำสั่งในการหาเพื่อนบ้านใกล้ที่สุด

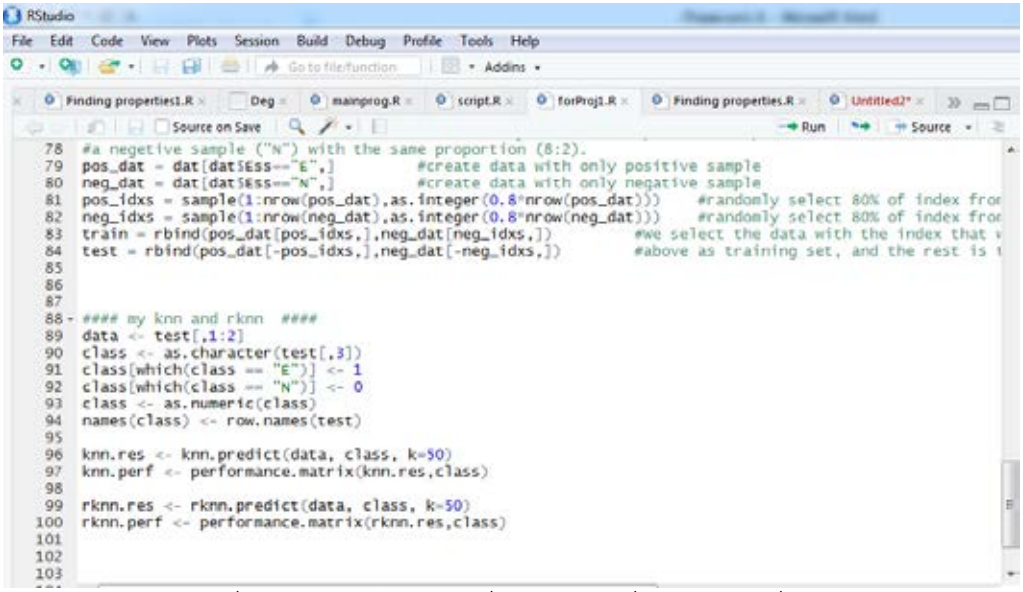
จากนั้นจะนำผลลัพธ์ของการทำนายคลาสของยีนว่าเป็นยีนที่สำคัญหรือเป็นยีนที่ไม่สำคัญที่ได้ไปเปรียบเทียบกับผลการทดลองการระบุยีนที่สำคัญที่ได้จริงจากการทดลองในห้องปฏิบัติการ เพื่อวัดประสิทธิภาพและนำไปแสดงผลต่อไป

3.3.3 การระบุยีนที่สำคัญโดยใช้การค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน

ให้ระยะทางเป็นระยะทางระหว่างโปรตีน 2 ตัวและให้ P เป็นกลุ่มโปรตีนในโครงข่าย และเพื่อนบ้านที่ใกล้ที่สุดของโปรตีน q คือโปรตีนที่มีระยะทางใกล้เคียงที่สุดกับโปรตีน q สามารถหาได้จาก $RkNN(q)$ ซึ่งได้กล่าวไว้แล้วในสมการ 2.5

โดยกำหนด class ให้กับจุดที่พิจารณา q โดยการตัดสินใจว่าจะระบุเป็นคลาสใดขึ้นอยู่กับว่าจุดที่สนใจ q เป็น neighbors ของโปรตีนอื่นซึ่งเป็น essential มากกว่า 25% ของจำนวนโปรตีน ที่ q ไปเป็น neighbor ทั้งหมด เนื่องจากข้อมูลจริงจากการทดลองโปรตีนที่เป็นยีนที่สำคัญพบประมาณ 10% ของโปรตีนทั้งหมด ดังนั้นหากเราพบว่าโปรตีนที่พิจารณาเป็นเพื่อนบ้านใกล้เคียงของโปรตีนที่สำคัญมากกว่า 25% ของจำนวน Neighbor ทั้งหมด จึงถือว่ามีโอกาสที่จะเป็นโปรตีนที่สำคัญสูง

ซึ่งผู้จัดทำได้คำนวณโดยการเขียนคำสั่งในโปรแกรม Rstudio จากรูปภาพที่ 3.6 แสดงตัวอย่างคำสั่งในการหาเพื่อนบ้านใกล้ที่สุดผกผัน



รูปภาพที่ 3.6 แสดงตัวอย่างคำสั่งในการหาเพื่อนบ้านใกล้ที่สุดผกผัน

จากนั้นจะนำผลลัพธ์ของการทำนายคลาสของยีนว่าเป็นยีนที่สำคัญหรือเป็นยีนที่ไม่สำคัญที่ได้ไปเปรียบเทียบกับผลการทดลองการระบุยีนที่สำคัญที่ได้จริงจากการทดลองในห้องปฏิบัติการ เพื่อวัดประสิทธิภาพและนำไปแสดงผลต่อไป

บทที่ 4
ผลการดำเนินงาน

บทที่ 4 จะกล่าวถึงผลการดำเนินงาน ซึ่งมีอยู่ด้วยกัน 4 ส่วนหลัก ๆ คือ ผลที่ได้จากการระบุยีนที่สำคัญโดยใช้ความเป็นจุดศูนย์กลางโดยวัดจากระดับ ผลที่ได้จากการระบุยีนที่สำคัญโดยใช้การค้นหาเพื่อนบ้านใกล้ที่สุด ผลที่ได้จากการระบุยีนที่สำคัญโดยใช้การค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน และผลการระบุยีนที่สำคัญด้วยวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดและการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน ที่ค่า k = 1, 2, ... , 100

1) ผลการระบุยีนที่สำคัญโดยใช้ความเป็นจุดศูนย์กลางโดยวัดจากระดับ

ผลจากการดำเนินงานพบว่าจะได้ผลลัพธ์ดังนี้

1.1) เมื่อใช้ ควอไทล์ที่ 1 เป็นตัววัด

ตารางที่ 4.1 คอนฟิวชันเมทริกซ์ของผลการระบุยีนที่สำคัญโดยใช้ความเป็นจุดศูนย์กลางโดยวัดจากระดับ เมื่อใช้ ควอไทล์ที่ 1 เป็นตัววัด

		Predicted Class	
		Essential	Nonessential
Actual Class	Essential	427	68
	Nonessential	1781	384

ผลการดำเนินงานจะได้ ผลลัพธ์ความถูกต้อง 30.49% ค่าความแม่นยำ 19.34% ค่าความไว 86.26% และค่าความจำเพาะ 17.74%

1.2) เมื่อใช้ ควอไทล์ที่ 2 เป็นตัววัด

ตารางที่ 4.2 คอนฟิวชันเมทริกซ์ของผลการระบุยีนที่สำคัญโดยใช้ความเป็นจุดศูนย์กลางโดยวัดจากระดับ เมื่อใช้ ควอไทล์ที่ 2 เป็นตัววัด

		Predicted Class	
		Essential	Nonessential
Actual Class	Essential	312	183
	Nonessential	1120	1045

ผลการดำเนินงานจะได้ ผลลัพธ์ความถูกต้อง 51.02% ค่าความแม่นยำ 21.79% ค่าความไว 63.03% และค่าความจำเพาะ 48.27%

1.3) เมื่อใช้ ควอไทล์ที่ 3 เป็นตัววัด

ตารางที่ 4.3 คอนฟิวชันเมทริกซ์ของผลการระบุยีนที่สำคัญโดยใช้ความเป็นจุดศูนย์กลางโดยวัดจากระดับเมื่อใช้ ควอไทล์ที่ 3 เป็นตัววัด

		Predicted Class	
		Essential	Nonessential
Actual Class	Essential	175	320
	Nonessential	505	1660

ผลการดำเนินงานจะได้ ผลลัพธ์ความถูกต้อง 68.98% ค่าความแม่นยำ 25.74% ค่าความไว 35.35% และค่าความจำเพาะ 76.67%

2) ผลการระบุยีนที่สำคัญโดยใช้การค้นหาเพื่อนบ้านใกล้ที่สุด

ตารางที่ 4.4 คอนฟิวชันเมทริกซ์ของผลการระบุยีนที่สำคัญโดยใช้การค้นหาเพื่อนบ้านใกล้ที่สุด

		Predicted Class	
		Essential	Nonessential
Actual Class	Essential	216	279
	Nonessential	986	1179

จากการระบุยีนที่สำคัญโดยใช้การค้นหาเพื่อนบ้านใกล้ที่สุด ที่พารามิเตอร์ k = 50 ผลการดำเนินงานจะได้ผลลัพธ์ดังนี้ ผลลัพธ์ความถูกต้อง 52.44% ค่าความแม่นยำ 17.97% ค่าความไว 43.64% และค่าความจำเพาะ 54.46%

3) ผลการระบุยีนที่สำคัญโดยใช้การค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน

ตารางที่ 4.5 คอนฟิวชันเมทริกซ์ของผลการระบุยีนที่สำคัญโดยใช้การค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน

		Predicted Class	
		Essential	Nonessential
Actual Class	Essential	203	292
	Nonessential	460	1705

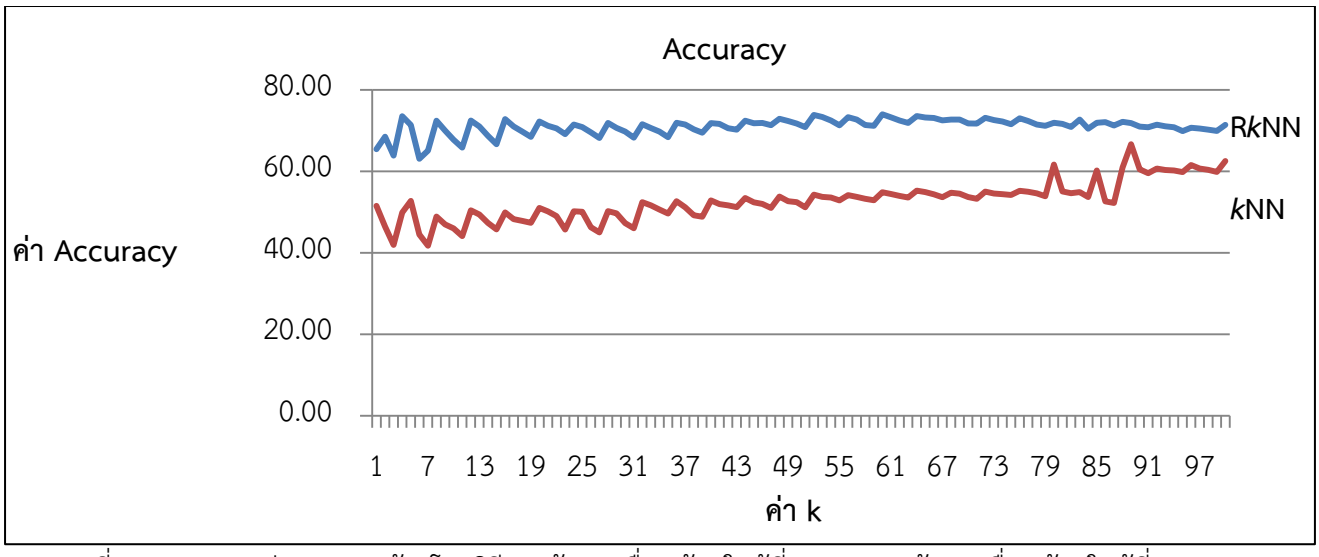
จากการระบุยีนที่สำคัญโดยใช้การค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน ที่พารามิเตอร์ k = 50 ผลการดำเนินงานจะได้ผลลัพธ์ดังนี้ ผลลัพธ์ความถูกต้อง 71.73% ค่าความแม่นยำ 30.62% ค่าความไว 41.01% และค่าความจำเพาะ 78.75%

ตารางที่ 4.6 ผลการระบุยีนที่สำคัญด้วยวิธีต่างๆ

	Degree Centrality		
	ควอไทล์ที่ 1	ควอไทล์ที่ 2	ควอไทล์ที่ 3
Accuracy(%)	30.49	51.02	68.98
precision(%)	19.34	21.79	25.76
sensitivity(%)	86.26	63.03	35.35
specificity(%)	17.74	48.27	76.67

4) ผลการระบุยีนที่สำคัญด้วยวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดและการค้นหาเพื่อนบ้านใกล้ที่สุด ผกผัน ที่ค่า k = 1, 2, ... , 100

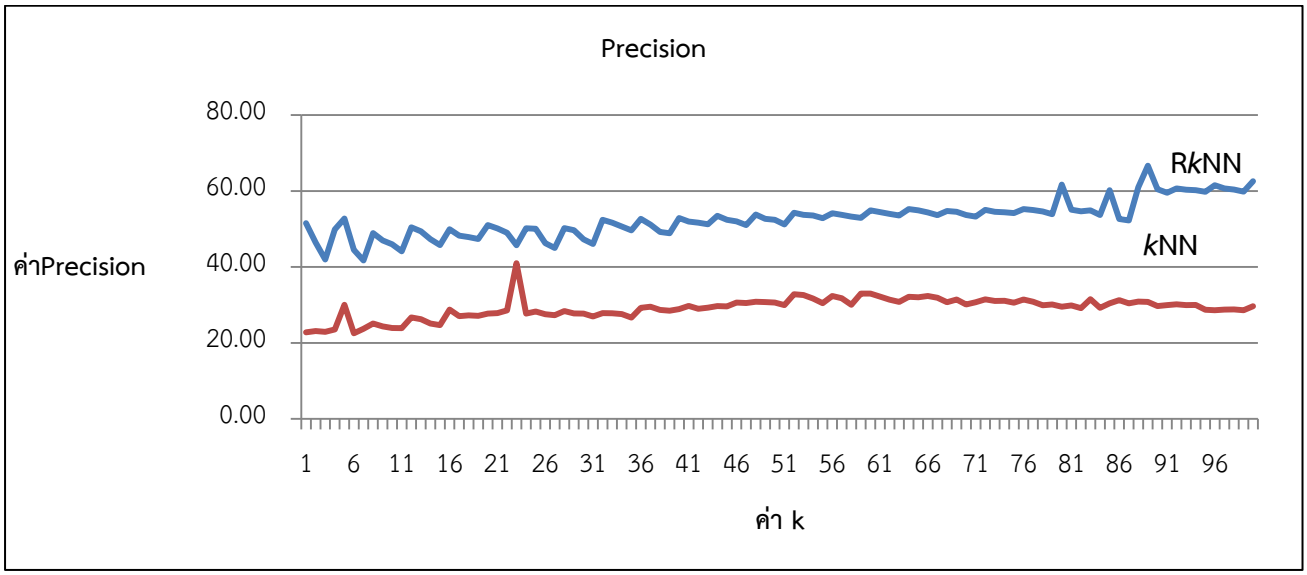
4.1) ค่าความถูกต้องโดยวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดและการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน ที่ค่า k = 1, 2, ... , 100



กราฟที่ 4.1 แสดงผลค่าความถูกต้องโดยวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดและการค้นหาเพื่อนบ้านใกล้ที่สุด ผกผัน ที่ค่า k = 1, 2, ... , 100

ผลการดำเนินงานจะได้ว่าวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน มีค่าความถูกต้องมากกว่าวิธีการค้นหาเพื่อนบ้านใกล้ที่สุด ทุกค่า k = 1, 2, ... , 100

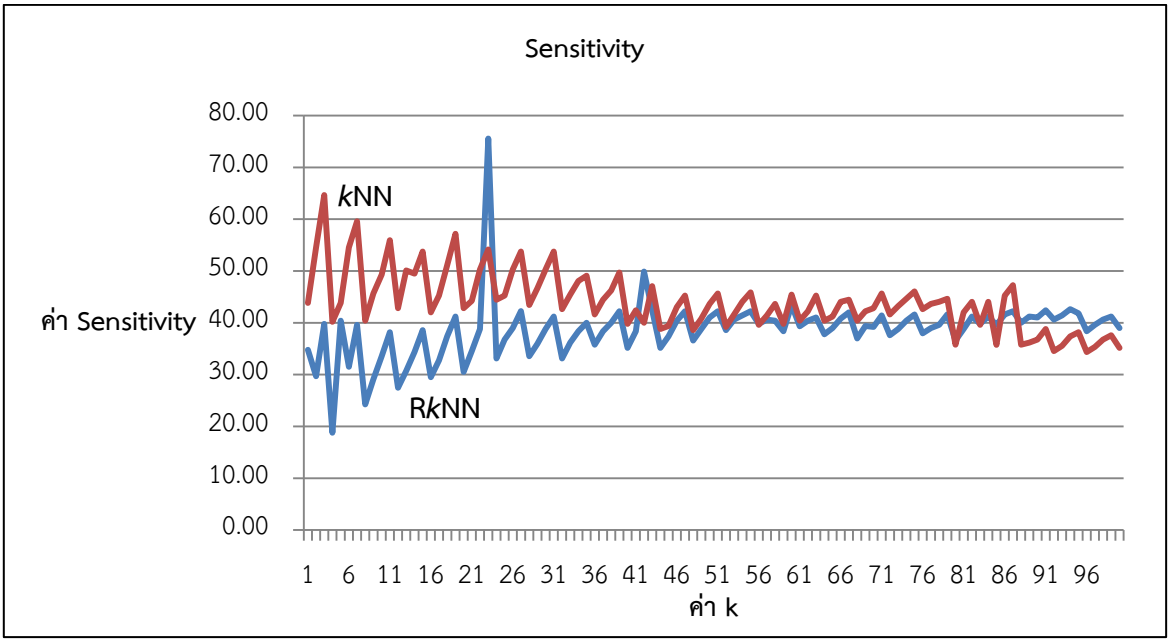
4.2) ค่าความแม่นยำโดยวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดและการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน ที่ค่า k = 1, 2, ... , 100



กราฟที่ 4.2 แสดงผลค่าความแม่นยำโดยวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดและการค้นหาเพื่อนบ้านใกล้ที่สุด ผกผัน ที่ค่า k = 1, 2, ... , 100

ผลการดำเนินงานจะได้ว่าวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน มีค่าความแม่นยำมากกว่าวิธีการค้นหาเพื่อนบ้านใกล้ที่สุด ทุกค่า k = 1, 2, ... , 100

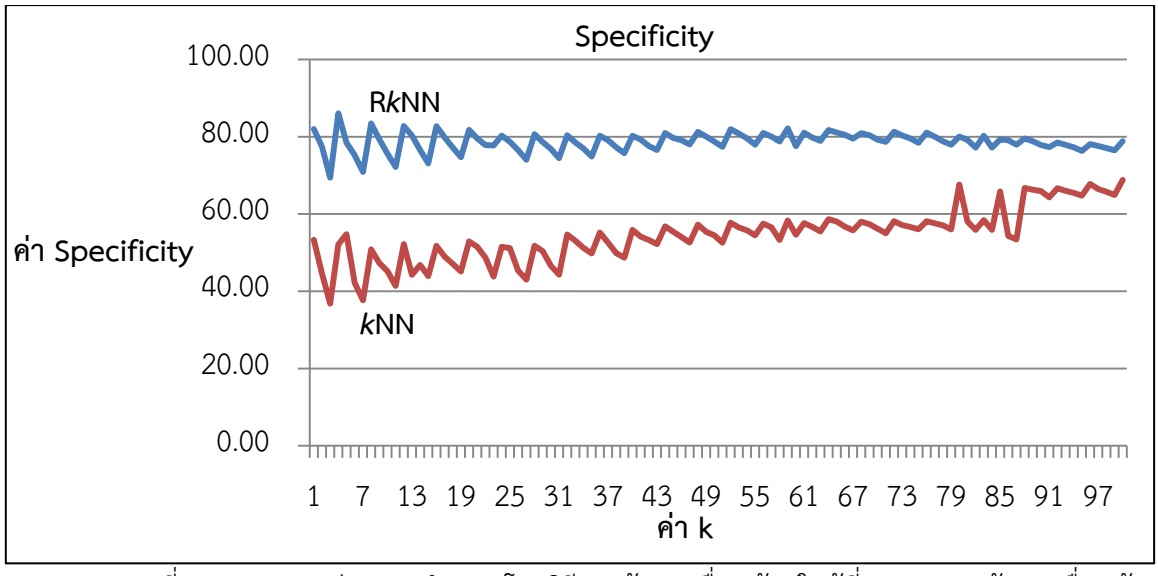
4.3) ค่าความไวโดยวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดและการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน ที่ค่า k = 1, 2, ... , 100



กราฟที่ 4.3 แสดงผลค่าความไวโดยวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดและการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน ที่ค่า k = 1, 2, ... , 100

ผลการดำเนินงานจะได้ว่าวิธีการค้นหาเพื่อนบ้านใกล้ที่สุด มีค่ามีค่าความไวมากกว่าวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน ที่ค่า k = 1, 2, ... , 100 ทั้งหมด 81 วิธีการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน มีค่าความไวมากกว่าวิธีการค้นหาเพื่อนบ้านใกล้ที่สุด ที่ค่า k = 1, 2, ... , 100 ทั้งหมด 19 ค่า

4.4) ค่าความจำเพาะโดยวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดและการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน ที่ค่า k = 1, 2, … , 100



กราฟที่ 4.4 แสดงผลค่าความจำเพาะโดยวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดและการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน ที่ค่า k = 1, 2, … , 100

ผลการดำเนินงานจะได้ว่าวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน มีค่าความจำเพาะมากกว่าวิธีการค้นหาเพื่อนบ้านใกล้ที่สุด ทุกค่า k = 1, 2, … , 100

บทที่ 5
ข้อสรุปและข้อเสนอแนะ

ในบทนี้จะสรุปผลจากบทที่ 4 คือ วิเคราะห์ผลหลังจากการดำเนินคำสั่งของโปรแกรมต่าง ๆ เรียบร้อยแล้วนำมาสู่ข้อสรุปว่าอย่างไร ซึ่งจะแบ่งออกเป็น 2 ส่วนใหญ่ ๆ

5.1 ข้อสรุป

จากการระบุยีนที่สำคัญด้วยชุดข้อมูลทดสอบพบว่า วิธีการระบุยีนที่สำคัญที่มีค่าความถูกต้องมากที่สุด เป็น 74.02% คือวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน ที่พารามิเตอร์ k = 60 วิธีการระบุยีนที่สำคัญที่มีค่าความแม่นยำมากที่สุด เป็น 41.01% คือวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน ที่พารามิเตอร์ k = 23 วิธีการระบุยีนที่สำคัญที่มีค่าความไวมากที่สุด เป็น 86.26% คือการระบุยีนที่สำคัญโดยใช้ความเป็นจุดศูนย์กลางโดยวัดจากระดับเมื่อใช้ ควอไทล์ที่ 1 เป็นตัววัด วิธีการระบุยีนที่สำคัญที่มีค่าความจำเพาะมากที่สุด เป็น 83.46% คือการระบุยีนที่สำคัญโดยใช้การค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน ที่พารามิเตอร์ k = 8 และจากการทดสอบพบว่าวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน มีค่าความถูกต้อง ค่าความแม่นยำ และค่าความจำเพาะ มากกว่าวิธีการค้นหาเพื่อนบ้านใกล้ที่สุด ทุกค่า k = 1, 2, ... , 100 และวิธีการค้นหาเพื่อนบ้านใกล้ที่สุด มีค่าความไวมากกว่าวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน ที่ค่า k = 1, 2, ... , 100 ทั้งหมด 81 ค่า

ดังนั้นสามารถสรุปผลได้ว่า การระบุยีนที่สำคัญโดยใช้การค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน มีประสิทธิภาพในการระบุยีนที่สำคัญมากที่สุด เนื่องจากมีค่าความถูกต้อง ค่าความแม่นยำ และค่าความจำเพาะมากกว่าวิธีการค้นหาเพื่อนบ้านใกล้ที่สุด ทุกค่า k = 1, 2, ... , 100 และวิธีการระบุยีนที่สำคัญโดยใช้ความเป็นจุดศูนย์กลางโดยวัดจากระดับ

5.2 ข้อเสนอแนะจากการทำโครงงานนี้

1. ควรระบุยีนที่สำคัญโดยใช้การค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน ที่พารามิเตอร์ k อื่นๆเพื่อนำมาเปรียบเทียบประสิทธิภาพที่ค่า k ต่างๆ
2. สามารถนำแนวคิดที่ได้ไปพัฒนาต่อในการระบุยีนที่สำคัญของสิ่งมีชีวิตอื่นๆต่อไปได้

เอกสารอ้างอิง

[1] R. Zhang, “DEG: a database of essential genes” [Online] . 2005 . Available from: https://www.ncbi.nlm.nih.gov/m/pubmed/14681410/[2005,January 1].

[2] K. Plaimas, R. Eils, and R. Konig, “Identifying essential genes in bacterial metabolic networks with machine learning methods”, BMC Syst Biol, vol.4, pp. S1-S8, 2010.

[3] A. Cheema, “Reverse nearest neighbors,” (Master’s thesis, Computer Science & Engineering, The University of New South Wales, Sydney Australia, 2007).

[4] K. Plaimas and A. Suratanee. “Reverse Nearest Neighbor Search on a Protein-Protein

Interaction Network to Infer Protein-Disease Associations”, Bioinformatics and Biology Insights, Vol. 11, pp. 1–11, 2017.

[5] ยนี ภู่วรวรรณ. (2003). กราฟ [Online]. Available HTTP: http://web.ku.ac.th/ schoolnet/

snet2/knowlodge_math/graph1.htm [2005, June 9].

[6] C. Haythornthwaite, “Social network analysis: An approach and technique for the study of information exchange”, Library & information science research, 18(4), 323-342, 1996.

[7] K. Flip and S. Muthukrishnan, “Influence sets based on reverse nearest neighbor queries”, SIGMOD Rec. 2000; 29: 201–212.

[8] H. Wang, Z. Xu, H. Fujita and S. Liu. (2016). Towards felicitous decision making: An overview on challenges and trends of Big Data. Information Sciences, 367–368, 747–765.

ภาคผนวก

ภาคผนวก ก
แบบเสนอหัวข้อโครงงาน รายวิชา 2301399 Project Proposal
ปีการศึกษา 2561

ชื่อโครงงาน (ภาษาไทย)	การค้นหาเพื่อนบ้านใกล้ที่สุดผกผันสำหรับการระบุยีนที่สำคัญบนโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีน
ชื่อโครงงาน (ภาษาอังกฤษ)	Reverse nearest neighbors search for identification of bacteria's essential genes in protein-protein interaction network
อาจารย์ที่ปรึกษา	ผศ.ดร. กิติพร พลายมาศ
ผู้ดำเนินการ	น.ส. นัยขวัญ ไชยศรี เลขประจำตัวนิสิต 5833528723 สาขาวิชา คณิตศาสตร์ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

หลักการและเหตุผล

ยีนที่สำคัญ (essential genes) คือยีนที่สำคัญต่อการดำรงชีวิตของสิ่งมีชีวิต ยีนที่สำคัญถือเป็นรากฐานของเซลล์สิ่งมีชีวิต เซลล์ทุกเซลล์จึงมียีนที่สำคัญอยู่จำนวนหนึ่ง การระบุยีนที่สำคัญไม่เพียงแต่จะเป็นความต้องการขั้นพื้นฐานสำหรับการอยู่รอดของสิ่งมีชีวิต แต่ยังสำคัญสำหรับการค้นหายีนของมนุษย์และยาต้านเชื้อแบคทีเรียและเชื้อโรคต่างๆ อีกด้วย วิธีการทดลองในการระบุยีนที่สำคัญมีค่าใช้จ่ายสูง มีหลายขั้นตอนและใช้เวลานาน ด้วยการเก็บข้อมูลของลำดับยีนและข้อมูลการทดลองที่มีปริมาณมาก จึงมีการนำเสนอวิธีการจำนวนมากสำหรับการระบุโปรตีนที่สำคัญซึ่งเป็นประโยชน์สำหรับการคัดกรองเลือกยีนสำหรับศึกษาเพิ่มเติมในห้องทดลอง เป็นการลดค่าใช้จ่ายในการทำการทดลองต่อไป โดยวิธีการที่ทันสมัยที่สุดสำหรับการระบุโปรตีนที่สำคัญโดยอาศัยการเรียนรู้ด้วยเครื่องและคุณสมบัติโทโปโลยีของโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีน ซึ่งชี้ให้เห็นถึงความก้าวหน้าและข้อจำกัดของวิธีการปัจจุบัน ในปัจจุบันมีการสร้างฐานข้อมูลของยีนที่สำคัญ (a database of essential genes: DEG) [1] ทั้งหมดที่มีอยู่ในปัจจุบันไว้ เพราะฉะนั้นการวิเคราะห์หายีนที่สำคัญจึงสามารถช่วยในการตอบคำถามเกี่ยวกับสิ่งที่เป็นหน้าที่พื้นฐานที่จำเป็นต่อการดำรงชีวิตของเซลล์ได้

จากการศึกษาเกี่ยวกับการระบุยีนที่สำคัญในแบคทีเรียเพื่อระบุเป้าหมายยาที่มีประสิทธิภาพในการยับยั้งการเจริญเติบโตของแบคทีเรีย โดยใช้คุณสมบัติลักษณะทางโทโปโลยีของโครงข่ายทางชีววิทยา [2] พบว่าคุณสมบัติลักษณะทางโทโปโลยีหนึ่งที่นำมาใช้คือ ความเป็นศูนย์กลาง(Centrality) ประกอบด้วย ความเป็นจุดศูนย์กลางดีกรีหรือความเป็นจุดศูนย์กลางโดยวัดจากระดับ (Degree Centrality), ความเป็นจุดศูนย์กลางโดยวัดจากความใกล้ชิด(Closeness Centrality), ความเป็นจุดศูนย์กลางโดยวัดจากการคั่นกลาง (Betweenness Centrality) และความเป็นจุดศูนย์กลางโดยวัดจากเวกเตอร์ลักษณะเฉพาะ(Eigenvector Centrality) ของโหนดยีนในโครงข่ายทางชีววิทยาที่ศึกษา ซึ่งจากการศึกษาด้วยความเป็นจุดศูนย์กลางโดยวัด

จากระดับ (Degree Centrality) เป็นการค้นหาว่ายีนใดบ้างที่เป็นจุดศูนย์กลางของการเชื่อมโยง (Hub) ซึ่งถือเป็นตำแหน่งที่มีอิทธิพลสูงสุดในโครงข่าย วัดได้จากจำนวนเส้นเชื่อมโยงทั้งหมดที่โยงเป็นโหนดยีนนั้น โดยพบว่าหากยีนตัวใดที่เป็นจุดศูนย์กลางของการเชื่อมโยง (Hub) เป็นไปได้ว่ายีนนั้นเป็นยีนที่สำคัญ [2] และความเป็นศูนย์กลาง (Centrality) อื่นๆมีแนวโน้มในลักษณะเดียวกันด้วย [2] จะเห็นได้ว่าการใช้คุณลักษณะทางโทโปโลยีของโหนดในโครงข่ายมีประสิทธิภาพเพียงพอที่จะใช้ระบุยีนที่สำคัญ โดยเฉพาะความเป็นจุดศูนย์กลางดีกรี อย่างไรก็ตามโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีนสามารถจำลองได้ด้วยกราฟถ่วงน้ำหนัก ทำให้การพิจารณาดีกรีของโหนดสามารถพิจารณาด้วยดีกรีแบบถ่วงน้ำหนัก และสามารถขยายการประยุกต์ใช้กับการค้นหาเพื่อนบ้านใกล้ที่สุด (*k*-Nearest Neighbors search (kNN)) และการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน (Reverse *k*-Nearest Neighbors search (RkNN)) ได้

วิธีการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน [3] เป็นวิธีการหาโหนดเพื่อนบ้านที่ได้รับอิทธิพลจากโหนดของข้อมูลที่นำมาวิเคราะห์ในโครงข่าย สำหรับโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีนมีการศึกษาและประยุกต์ใช้ในการอนุมานความสัมพันธ์ระหว่างโปรตีนและโรค โดยอาศัยการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน ใช้เพื่อระบุผลกระทบของโปรตีนที่สนใจต่อโปรตีนอื่นในโครงข่าย จากนั้นความสัมพันธ์ระหว่างโปรตีนและโรคจะถูกอนุมานได้โดยใช้วิธีทดสอบข้อมูลทางสถิติ [4]

ในโครงงานนี้ ผู้ดำเนินการจึงสนใจที่จะศึกษาและประยุกต์ใช้การค้นหาเพื่อนบ้านใกล้ที่สุดผกผันเข้ามาช่วยหายีนที่สำคัญในโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีนในแบคทีเรีย โดยจะนำข้อมูลโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีนในแบคทีเรีย และข้อมูลยีนที่สำคัญ มาวิเคราะห์ร่วมกันเพื่ออนุมานหายีนหรือโปรตีนที่สำคัญตัวใหม่ ซึ่งจะเป็นประโยชน์ต่อการพัฒนายาต้านเชื้อแบคทีเรียต่อไปได้ในอนาคต

วัตถุประสงค์

เพื่อค้นหาเพื่อนบ้านใกล้ที่สุดผกผันบนโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีนในการระบุยีนที่สำคัญในแบคทีเรีย

ขอบเขตของโครงงาน

1. สนใจศึกษาการระบุยีนที่สำคัญของแบคทีเรีย Escherichia coli เท่านั้น
2. ข้อมูลยีนที่สำคัญได้มาจาก DEG: a database of essential genes.
3. ข้อมูลโครงข่ายปฏิสัมพันธระหว่างโปรตีน STRING database

วิธีการดำเนินงาน

ก. แผนการศึกษา

1. ศึกษาค้นคว้าทฤษฎี
2. ศึกษางานวิจัยที่เกี่ยวข้อง
3. กำหนดแนวทางในการดำเนินการค้นหาข้อมูลและวิเคราะห์ข้อมูล
4. สร้างโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีนของแบคทีเรีย
5. เขียนโปรแกรมเพื่อระบุยีนที่สำคัญโดยวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน
6. ประเมินประสิทธิภาพกับวิธีการอื่น
7. ปรับปรุงแก้ไขและอภิปรายผล
8. จัดทำเอกสารรายงานฉบับสมบูรณ์

ข. ระยะเวลาที่ศึกษา

ขั้นตอนการดำเนินการ	ปี 2561					ปี 2562			
	ส.ค.	ก.ย.	ต.ค.	พ.ย.	ธ.ค.	ม.ค.	ก.พ.	มี.ค.	เม.ย.
1.ศึกษาค้นคว้าทฤษฎี	■	■	■						
2.ศึกษางานวิจัยที่เกี่ยวข้อง			■	■					
3.กำหนดแนวทางในการดำเนินการค้นหาข้อมูลและวิเคราะห์ข้อมูล				■	■				
4.สร้างโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีนของแบคทีเรีย					■	■	■		
5.เขียนโปรแกรมเพื่อระบุยีนที่สำคัญโดยวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดผกผัน						■	■	■	
6.ประเมินประสิทธิภาพกับวิธีการอื่น							■	■	■
7.ปรับปรุงแก้ไขและอภิปรายผล								■	■
8. จัดทำเอกสารรายงานฉบับสมบูรณ์								■	■

ประโยชน์ที่คาดว่าจะได้รับ

ก. ประโยชน์ด้านความรู้และประสบการณ์ต่อนิสิต

1. ได้ความรู้และความเข้าใจเกี่ยวกับวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดผกผันมากยิ่งขึ้น
2. ได้ความรู้และความเข้าใจเกี่ยวกับการระบุยีนที่สำคัญบนโครงข่ายโปรตีนมากยิ่งขึ้น
3. ได้เพิ่มประสบการณ์การเขียนโปรแกรม R สำหรับการทำงานวิเคราะห์ข้อมูลซึ่งเป็นที่นิยมในปัจจุบัน

ข. ประโยชน์ที่ได้จากโครงงานที่พัฒนาขึ้น

1. ได้วิธีการระบุยีนที่สำคัญบนโครงข่ายปฏิสัมพันธ์ระหว่างโปรตีนโดยการวิเคราะห์ข้อมูลผ่านวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดผกผันได้
2. สามารถนำแนวคิดที่ได้ไปพัฒนาต่อในการระบุยีนที่สำคัญของสิ่งมีชีวิตอื่นๆต่อไปได้

อุปกรณ์และเครื่องมือที่ใช้

ก. ฮาร์ดแวร์

1. คอมพิวเตอร์
2. อุปกรณ์จัดเก็บข้อมูลสำรอง
3. กระดาษ A4

ข. ซอฟต์แวร์

1. โปรแกรม R
2. โปรแกรม Microsoft Word

งบประมาณ

1. ค่าอุปกรณ์จัดเก็บข้อมูลสำรอง 2500 บาท.
2. ค่ากระดาษ A4 400 บาท.
3. ค่าเข้าเล่มโครงงาน 600 บาท
4. ค่าถ่ายเอกสาร 400 บาท
5. Handy Drive 16 GB 500 บาท
6. ค่าปริ้นส์งาน 600 บาท

รวมทั้งสิ้น 5,000 บาท

เอกสารอ้างอิง

[1] R. Zhang, “DEG: a database of essential genes” [Online] . 2005 . Available from: https://www.ncbi.nlm.nih.gov/m/pubmed/14681410/[2005,January 1].

[2] K. Plaimas, R. Eils, and R. Konig, “Identifying essential genes in bacterial metabolic networks with machine learning methods”, BMC Syst Biol, vol.4, pp. S1-S8, 2010.

[3] A. Cheema, “Reverse nearest neighbors,” (Master’s thesis, Computer Science & Engineering, The University of New South Wales, Sydney Australia, 2007).

[4] K. Plaimas and A. Suratanee. “Reverse Nearest Neighbor Search on a Protein-Protein Interaction Network to Infer Protein-Disease Associations”, Bioinformatics and Biology Insights, Vol. 11, pp. 1–11, 2017.

ภาคผนวก ข
ข้อมูลโปรแกรม R

1. ข้อมูลเกี่ยวกับภาษา R และโปรแกรม R Studio

1.1 ภาษา R

R เป็นภาษาโปรแกรมที่ใช้สำหรับการประมวลผลทางด้านสถิติและแสดงผลทางด้านกราฟฟิกต่าง ๆ ตัวโปรแกรมนั้นมีใช้กันอย่างแพร่หลายซึ่งคล้ายกับภาษาอีกภาษาหนึ่งที่ใช้กันกว้างขว้างเช่นกัน นั่นคือภาษา S ซึ่งถูกพัฒนาจากทางห้องวิจัยเบล (Bell Laboratories) โดย จอห์น แชมเบอร์ส (John Chambers) และคณะ ในทางการศึกษาและงานวิจัยงานต่าง ๆ นิยมใช้ภาษา R ข้อดีของภาษานี้คือ เป็นฟรีแวร์ที่สามารถบรรจุลง (Download) คอมพิวเตอร์ได้ทั่วไป และสามารถหาวิธีแก้ไขต่าง ๆ ได้ง่ายหากเจอสถานการณ์ผิดปกติทางโปรแกรม

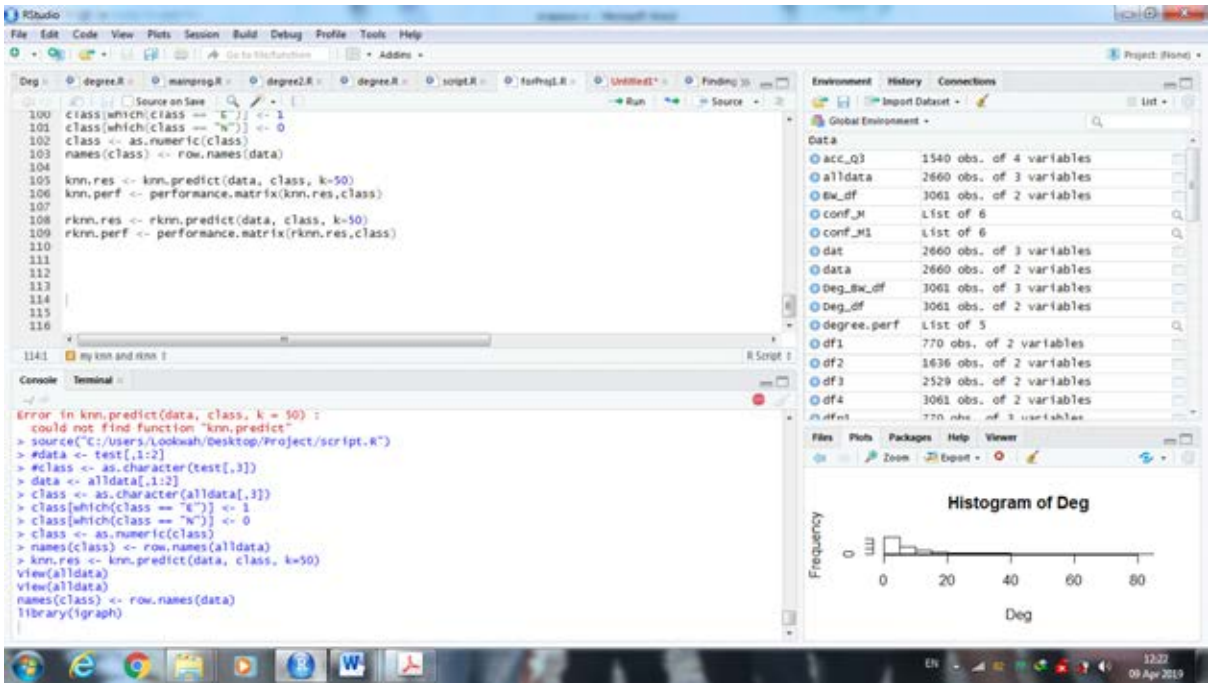
1.2 RStudio

RStudio เป็นโปรแกรมต่อประสานที่ถูกรวบรวมและพัฒนาขึ้นจาก R ซึ่งประกอบไปด้วยหลาย ๆ ส่วนซึ่งจะกล่าวในหัวข้อถัดไป ทางโปรแกรมนั้นมีชุดคำสั่งสำเร็จจรูปที่สนับสนุนการคำนวณทางสถิติค่อนข้างเยอะและรวมไปถึงงานในโครงงานครั้งนี ด้วย จึงทำให้การใช้งานนั้นมีความสะดวกและเป็นที่นิยมมากกว่าแบบภาษา R

1.3 หน้าต่างของ RStudio

หลังจากที่ลงโปรแกรมเป็นที่เรียบร้อยแล้ว พอเข้าหน้าโปรแกรมก็จะพบกับหน้าต่างที่ใช้ปฏิบัติงานซึ่งประกอบด้วย 4 ส่วนหลัก ๆ ได้แก่

- มุมซ้ายล่าง: ส่วนเฝ้าคุม (Console window) เป็นส่วนที่ใช้ในการดำเนินงานคำสั่งต่าง ๆ เมื่อกด Enter ตัวโปรแกรมจะดำเนินการคำสั่งที่ถูกเขียนไว้ทันที โดยสามารถเขียนรหัสหรือคำสั่งต่าง ๆ ตรงตัวพร้อม (prompt) คือ “>”
- มุมบนซ้าย: หน้าต่างแก้ไข (Script window) เป็นส่วนที่ใช้พิมพ์หรือแก้ไขคำสั่งและรหัสต่าง ๆ โดยที่โปรแกรมจะไม่ดำเนินการจนกว่าจะสั่ง นอกจากนี้ยังสามารถเรียกคำสั่งเก่าออกมาตรวจสอบได้ หรือแสดงผลของตารางงานได้
- มุมขวาบน: พื้นที่ทำงาน (Workspace) เป็นส่วนที่จะเก็บประวัติการทำงานและชุดคำสั่งหรือแฟ้มงานเอาไว้สามารถเรียกดูได้ และเก็บชื่อตัวแปร ชนิดของตัวแปร ขนาดและความยาว
- มุมล่างขวา: แฟ้มงาน / จอแสดงผลกราฟหรือ แผนภาพ / คำแนะนำชุดคำสั่งสำเร็จรูป / ตัวช่วยหาข้อมูล



รูปภาพที่ 1: หน้าต่างของ RStudio

1.4 สัญลักษณ์ที่ใช้บ่อยในโปรแกรม

สัญลักษณ์	ความหมาย
a <– b	นำค่า b เก็บลงใน a
a : b	ลำดับจำนวนเต็มตั งแต่ a ถึง b
a$b	ข้อมูล b จาก a
c(a,b)	เวกเตอร์ a,b
a[b]	ดึงข้อมูล a ที่มีเงื่อนไข b
a[,b]	ดึงแถวแนวตั้ง b ของข้อมูล a
#a	หมายเหตุ
&	และ
NA	ไม่มีข้อมูล

ตารางที่ 1: สัญลักษณ์ที่ใช้ในคำสั่ง

1.5 รหัสและคาสั่งต่าง ๆ ที่ใช้ในโครงงาน

```
library()
```

คำอธิบาย

เริ่มต้นใช้งาน

```
read_csv ()
```

คำอธิบาย

อ่านไฟล์คั่นด้วยเครื่องหมายจุลภาค

```
read_csv2 ()
```

คำอธิบาย

อ่านไฟล์ที่คั่นด้วยเครื่องหมายอัฒภาค

```
read_delim ()
```

คำอธิบาย

อ่านในไฟล์ที่มีตัวคั่น

```
matrix(nrow = 1, ncol = 1)
```

คำอธิบาย

สร้าง Matrix โดย nrow และ ncol คือจำนวนของแถวและคอลัมน์ตามลำดับ

```
data.frame()
```

คำอธิบาย

เป็นการจัดเก็บข้อมูล ซึ่งเป็นข้อมูลแบบตาราง

```
install.packages()
```

คำอธิบาย

ติดตั้งและทำการ load package

```
hist()
```

คำอธิบาย

แสดงข้องมูลในรูปแบบฮิสโตแกรม

length()

คำอธิบาย

จะแสดงความยาวของอ็อบเจ็กต์หนึ่ง

row.names()

คำอธิบาย

ให้ชื่อของแถว

rbind()

คำอธิบาย

เชื่อมเมทริกซ์ตามลักษณะของจำนวนแถว

cbind()

คำอธิบาย

เชื่อมเมทริกซ์ตามลักษณะของจำนวนคอลัมน์

qqplot(x, y)

คำอธิบาย

ควอร์ไทล์ของ y ตามควอร์ไทล์ของ x

ภาคผนวก ค
คำสั่งต่าง ๆ ในภาษา R

โหลดข้อมูลและเลือกข้อมูลที่จะนำมาทดสอบ

```
string.data<-read.table("C:/Users/Lookwah/Desktop/Project/E. coli_Escherichia coli
str. K-12 substr. MG1655_511145.protein.links.v10.5 (1) (1).txt", header = T)
View(string.data)
string.data<-string.data[string.data$combined_score>=900,]
g<-graph.data.frame(string.data,directed = F)
```

การระบุยีนที่สำคัญโดยใช้ความเป็นจุดศูนย์กลางโดยวัดจากระดับและสร้างคอนฟิวเมทริกซ์จากผลการทดสอบ

```
Deg <- degree(net)

hist(Deg)

lab_ess = read.table("Lab-Essential-data.txt",header=T)
lab_ess$Name = row.names(lab_ess)

SelQ3 <- Deg[which(Deg >= quantile(Deg)[4])]
names(SelQ3)
df1<-as.data.frame(cbind(names(SelQ3), SelQ3 , "E"))
colnames(df1)<-c("Protein","SelQ3" , "Predict")
head(df1)

nonSelQ3 <- Deg[which(Deg < quantile(Deg)[4])]
names(nonSelQ3)
```

```
dfn1<-as.data.frame(cbind(names(nonSelQ3), nonSelQ3 , "N"))
colnames(dfn1)<-c("Protein","SelQ3" , "Predict")
head(dfn1)

Q3 <- merge(df1,dfn1, all = TRUE)

acc_Q3 <- merge(lab_ess, Q3, by.x="Name",by.y="Protein")
head(acc_Q3)

C3 = confusionMatrix(acc_Q3$Predict,acc_Q3$Ess,mode="everything")

SelQ2 <- Deg[which(Deg >= quantile(Deg)[3])]
names(SelQ2)
df2<-as.data.frame(cbind(names(SelQ2), SelQ2, "E"))
colnames(df2)<-c("Protein","SelQ2","Predict")
head(df2)

nonSelQ2 <- Deg[which(Deg < quantile(Deg)[3])]
names(nonSelQ2)
dfn2<-as.data.frame(cbind(names(nonSelQ2), nonSelQ2 , "N"))
colnames(dfn2)<-c("Protein","SelQ2" , "Predict")
head(dfn2)

Q2 <- merge(df2,dfn2, all = TRUE)

acc_Q2 <- merge(lab_ess, Q2, by.x="Name",by.y="Protein")
head(acc_Q2)

C2 = confusionMatrix(acc_Q2$Predict,acc_Q2$Ess,mode="everything")
```

```
SelQ1 <- Deg[which(Deg >= quantile(Deg)[2])]
names(SelQ1)
df3<-as.data.frame(cbind(names(SelQ1), SelQ1, "E"))
colnames(df3)<-c("Protein","SelQ1", "Predict")
head(df3)

nonSelQ1 <- Deg[which(Deg < quantile(Deg)[2])]
names(nonSelQ1)
dfn3<-as.data.frame(cbind(names(nonSelQ1), nonSelQ1 , "N"))
colnames(dfn3)<-c("Protein","SelQ1" , "Predict")
head(dfn3)

Q1 <- merge(df3,dfn3, all = TRUE)

acc_Q1 <- merge(lab_ess, Q1, by.x="Name",by.y="Protein")
head(acc_Q1)

C1 = confusionMatrix(acc_Q1$Predict,acc_Q1$Ess,mode="everything")
```

การระบุยีนที่สำคัญโดยใช้การค้นหาเพื่อนบ้านใกล้ที่สุดและสร้างคอนฟิวเมทริกซ์จากผลการทดสอบ

```
library(igraph)

string.data<-read.table("C:/Users/Lookwah/Desktop/Project/E. coli_Escherichia coli
str. K-12 substr. MG1655_511145.protein.links.v10.5 (1) (1).txt", header = T)

#View(string.data)
```

```
string.data<-string.data[string.data$combined_score>=900,]
g<-graph.data.frame(string.data,directed = F)
#remove multiple edges
net<-simplify(g, remove.multiple=T)


#Finding properties of network
#Degree
#Node <- V(net)$name
Deg <- degree(net)

#Betweenness
bw <- betweenness(net, directed = F)
names(bw)
df4<- as.data.frame(cbind(names(bw), bw))
colnames(df4)<-c("Protein","bw")
head(df4)

# Import known essential protein data
lab_dat = read.delim("C:/Users/Lookwah/Desktop/Project/inline-supplementary-material-5.txt")

# Create Name of each protein in this dataset to match with its form on string database
# (I guess it must be "511145.bxxxx" so, I combine "511145" and its "b number".)
lab_dat$Name =
data.frame(lapply(lab_dat[9],function(x){paste("511145",x,sep=".")}))
```

```
#Create dataframe that contains only protein name (b number) and it genetic properties
lab_ess = data.frame("Name" = lab_dat$Name,"Ess" = lab_dat$X.19)
names(lab_ess) = c("Name","Ess")    #Change column name to make it easy to understand
lab_ess = lab_ess[-c(1,2,3),]       #Drop row 1,2,3 because its "Name" is blank.

#Crate dataframe of protein name and its degree.
Deg_df = data.frame(Deg)
Deg_df$Name = row.names(Deg_df)

#Crate dataframe of protein name and its bw.
BW_df = data.frame(bw)
BW_df$Name = row.names(BW_df)

#Combine 3 dataframe with same "Name"
Deg_BW_df = merge(BW_df,Deg_df,by.x = "Name", by.y = "Name")
full_dat = merge(Deg_BW_df,lab_ess,by.x = "Name", by.y = "Name")

#Delete row that the genetic properties is not "E" or "N"
dat = full_dat[full_dat$Ess == "E" | full_dat$Ess == "N",]
dat = unique(dat)                   #Also delete duplicate rows.
row.names(dat) = dat$Name           #Make its row name to be the protein name
dat = dat[-c(1)]                    #Since we move protein name data into row name, then we can drop it.

##Data preparation for kNN
#Since our data have imbalance class, we will try to select a positive sample ("E") and
```

```
#a negetive sample ("N") with the same proportion (8:2).
pos_dat = dat[dat$Ess=="E",]          #create data with only positive sample
neg_dat = dat[dat$Ess=="N",]          #create data with only negative sample
pos_idxs = sample(1:nrow(pos_dat),as.integer(0.8*nrow(pos_dat)))        #randomly
select 80% of index from pos_dat
neg_idxs = sample(1:nrow(neg_dat),as.integer(0.8*nrow(neg_dat)))        #randomly
select 80% of index from neg_dat
train = rbind(pos_dat[pos_idxs,],neg_dat[neg_idxs,])          #We select the data with
the index that we randomly select
test = rbind(pos_dat[-pos_idxs,],neg_dat[-neg_idxs,])          #above as training set,
and the rest is test set.

alldata <- rbind(train,test)

#### my knn   ####

source("C:/Users/Lookwah/Desktop/Project/script.R")

#data <- test[,1:2]
#class <- as.character(test[,3])
data <- alldata[,1:2]
class <- as.character(alldata[,3])



class[which(class == "E")] <- 1
class[which(class == "N")] <- 0
class <- as.numeric(class)
names(class) <- row.names(data)
```

```
knn.res <- knn.predict(data, class, k=50)
knn.perf <- performance.matrix(knn.res,class)
```

การระบุยีนที่สำคัญโดยใช้การค้นหาเพื่อนบ้านใกล้ที่สุดผกผันและสร้างคอนฟิวเมทริกซ์จากผลการทดสอบ

```
library(igraph)

string.data<-read.table("C:/Users/Lookwah/Desktop/Project/E. coli_Escherichia coli str. K-12 substr. MG1655_511145.protein.links.v10.5 (1) (1).txt", header = T)

#View(string.data)
string.data<-string.data[string.data$combined_score>=900,]
g<-graph.data.frame(string.data,directed = F)
#remove multiple edges
net<-simplify(g, remove.multiple=T)



#Finding properties of network
#Degree
#Node <- V(net)$name
Deg <- degree(net)

#Betweenness
bw <- betweenness(net, directed = F)
names(bw)
df4<- as.data.frame(cbind(names(bw), bw))
colnames(df4)<-c("Protein","bw")
head(df4)
```

```
# Import known essential protein data
lab_dat = read.delim("C:/Users/Lookwah/Desktop/Project/inline-supplementary-material-5.txt")

# Create Name of each protein in this dataset to match with its form on string database
# (I guess it must be "511145.bxxxx" so, I combine "511145" and its "b number".)
lab_dat$Name = data.frame(lapply(lab_dat[9],function(x){paste("511145",x,sep=".")}))

#Create dataframe that contains only protein name (b number) and it genetic properties
lab_ess = data.frame("Name" = lab_dat$Name,"Ess" = lab_dat$X.19)
names(lab_ess) = c("Name","Ess")  #Change column name to make it easy to understand
lab_ess = lab_ess[-c(1,2,3),]  #Drop row 1,2,3 because its "Name" is blank.

#Crate dataframe of protein name and its degree.
Deg_df = data.frame(Deg)
Deg_df$Name = row.names(Deg_df)

#Crate dataframe of protein name and its bw.
BW_df = data.frame(bw)
BW_df$Name = row.names(BW_df)

#Combine 3 dataframe with same "Name"
Deg_BW_df = merge(BW_df,Deg_df,by.x = "Name", by.y = "Name")
full_dat = merge(Deg_BW_df,lab_ess,by.x = "Name", by.y = "Name")
```

```
#Delete row that the genetic properties is not "E" or "N"
dat = full_dat[full_dat$Ess == "E" | full_dat$Ess == "N",]
dat = unique(dat)                #Also delete duplicate rows.
row.names(dat) = dat$Name        #Make its row name to be the protein name
dat = dat[-c(1)]                 #Since we move protein name data into row name, then we can drop it.

##Data preparation for kNN
#Since our data have imbalance class, we will try to select a positive sample ("E") and
#a negetive sample ("N") with the same proportion (8:2).
pos_dat = dat[dat$Ess=="E",]          #create data with only positive sample
neg_dat = dat[dat$Ess=="N",]          #create data with only negative sample
pos_idxs = sample(1:nrow(pos_dat),as.integer(0.8*nrow(pos_dat)))        #randomly select 80% of index from pos_dat
neg_idxs = sample(1:nrow(neg_dat),as.integer(0.8*nrow(neg_dat)))        #randomly select 80% of index from neg_dat
train = rbind(pos_dat[pos_idxs,],neg_dat[neg_idxs,])          #We select the data with the index that we randomly select
test = rbind(pos_dat[-pos_idxs,],neg_dat[-neg_idxs,])          #above as training set, and the rest is test set.

alldata <- rbind(train,test)

#### my rknn  ####

source("C:/Users/Lookwah/Desktop/Project/script.R")

#data <- test[,1:2]
```

```
#class <- as.character(test[,3])
data <- alldata[,1:2]
class <- as.character(alldata[,3])

class[which(class == "E")] <- 1
class[which(class == "N")] <- 0
class <- as.numeric(class)
names(class) <- row.names(data)

rknn.res <- rknn.predict(data, class, k=50)
rknn.perf <- performance.matrix(rknn.res,class)
```

script.R

```
getNeighbors <- function(data, k)
{
  dmatrix <- dist(data)
  ele <- row.names(data)

  nb <- NULL
  for(i in 1:nrow(data)){
    slist <- names(sort(as.matrix(dmatrix)[,i]))
    slist <- slist[-1]
    if(is.null(slist)){
      nb[[ele[i]]] <- NA
    }else{
      nb[[ele[i]]] <- slist[1:k]
    }
  }
  return(nb)
```

```
}

getReNeighbors <- function(data, k){
  knb <- getNeighbors(data, k)
  ele <- row.names(data)

  rnb <- NULL
  for(e in ele){
    listrn <- NULL
    for(i in 1:length(knb)){
      fl <- which(knb[[i]] == e)
      if(length(fl) > 0){
        listrn <- c(listrn,ele[i])
      }
    }
    #print(listrn)
    if(is.null(listrn)){
      rnb[[e]] <- NA
    }else{
      rnb[[e]] <- listrn
    }
  }
  return(rnb)
}
knn.predict <- function(data, class, k){
  knb <- getNeighbors(data, k)
  ele <- row.names(data)

  cls <- NULL
```

```
  for(i in 1:length(knb)){
    if(is.na(sum(class[knb[[i]]]))){
      score <- 0
    }else{
      score <- sum(class[knb[[i]]])
    }
    #print(i)
    #print(score)
    #if(score >= (length(knb[[i]])-score) ){  # Majority
    if(score/k > 0.25){     # found > 50%  ## Majority (>50%)
      cls[i] <- 1
    }else{
      cls[i] <- 0
    }
  }
  names(cls) <- ele
  return(cls)
}

rknn.predict <- function(data, class, k){
  rnb <- getReNeighbors(data, k)
  ele <- row.names(data)

  cls <- NULL
  for(i in 1:length(rnb)){
    if(is.na(sum(class[rnb[[i]]]))){
      score <- 0
    }else{
      score <- sum(class[rnb[[i]]])
```

```
    }
    #if(score >= (length(rnb[[i]])-score) ){  # Majority
    if(score/k > 0.25){     # found > 25%
      cls[i] <- 1
    }else{
      cls[i] <- 0
    }
  }
  names(cls) <- ele
  return(cls)
}

performance.matrix <- function(predict, class){
  TP <- length(which(class[which(predict == 1)] == 1))
  FN <- length(which(class[which(predict == 0)] == 1))
  FP <- length(which(class[which(predict == 1)] == 0))
  TN <- length(which(class[which(predict == 0)] == 0))

  cmat <- matrix(
    c(TP, FN,
      FP, TN),
    nrow=2, ncol=2, # number of rows and columns
    byrow = TRUE, # fill matrix by rows
    dimnames = list(c("True class 1","True class 0"),
                    c("Pred class 1","Pred class 0")) )

  acc <- (TP+TN)/(TP+TN+FP+FN)
  sens <- TP/(TP+FN)
  spec <- TN/(TN+FP)
```

```
  prec <- TP/(TP+FP)

  perf <- list(cmat = cmat,
               accuracy = acc,
               sensitivity = sens,
               specificity = spec,
               precision = prec)
  return(perf)
}
```

ภาคผนวก ง
ผลการระบุยีนที่สำคัญด้วยวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดและการค้นหาเพื่อนบ้านใกล้ที่สุด ผกผัน ที่ค่า k = 1,2,...,100

ค่า k	Accuracy		Precision		Sensitivity		Specificity	
	rkNN	kNN	rkNN	kNN	rkNN	kNN	rkNN	kNN
1	65.44	51.54	22.79	37.67	34.77	43.84	82.01	53.30
2	68.53	46.43	23.11	18.37	29.70	54.55	77.41	44.57
3	63.87	41.95	22.91	18.95	39.80	64.65	69.38	36.77
4	73.53	49.89	23.54	16.10	18.79	40.20	86.05	52.10
5	71.39	52.74	30.03	18.14	40.40	43.84	78.48	54.78
6	63.11	44.51	22.54	17.75	31.52	54.55	75.24	42.21
7	65.08	41.73	23.73	17.93	39.60	59.60	70.90	37.64
8	72.44	48.91	25.10	15.82	24.24	40.40	83.46	50.85
9	70.00	46.99	24.37	16.53	29.09	45.66	79.35	47.30
10	67.78	45.98	23.92	17.06	33.54	49.29	75.61	45.22
11	65.86	44.10	23.89	17.92	38.18	55.96	72.19	41.39
12	72.48	50.45	26.72	17.00	27.47	42.83	82.77	52.19
13	71.05	49.40	26.25	17.49	30.71	50.12	80.28	44.23
14	68.72	47.33	25.11	17.55	34.34	49.49	76.58	46.83
15	66.65	45.75	24.68	17.97	38.59	53.74	73.07	43.93
16	72.82	49.92	28.77	16.60	29.49	42.02	82.73	51.73
17	71.05	48.27	27.05	16.85	32.73	45.25	79.82	48.96
18	69.77	47.86	27.25	18.10	37.37	51.11	77.18	47.11
19	68.46	47.37	27.13	19.24	41.21	57.17	74.69	45.13
20	72.26	51.02	27.71	17.21	30.51	42.83	81.80	52.89
21	71.17	50.15	27.85	17.26	34.55	44.24	79.54	51.50
22	70.56	49.02	28.57	18.22	38.79	50.30	77.83	48.73
23	69.14	45.71	41.01	18.05	75.57	54.14	77.73	43.79
24	71.50	50.19	27.74	17.32	33.13	44.44	80.28	51.50

25	70.86	50.08	28.26	17.49	36.77	45.25	78.66	51.18
26	69.59	46.28	27.57	17.39	38.99	50.30	76.58	45.36
27	68.20	45.00	27.30	17.73	42.26	53.74	74.04	43.00
28	71.88	50.23	28.38	17.08	33.54	43.43	80.65	51.78
29	70.68	49.70	27.77	17.70	35.96	46.67	78.61	50.39
30	69.77	47.29	27.71	17.72	38.79	50.30	76.86	46.61
31	68.27	46.05	26.95	18.07	41.21	53.74	74.46	44.30
32	71.58	52.44	27.84	17.70	33.13	42.63	80.37	54.69
33	70.64	51.65	27.80	18.13	36.16	45.45	78.52	53.07
34	69.77	50.64	27.58	18.39	38.38	48.08	76.95	51.22
35	68.38	49.66	26.68	18.27	40.00	49.09	74.87	49.79
36	71.95	52.67	29.26	17.52	35.76	41.62	80.23	55.20
37	71.50	51.17	29.55	17.79	38.38	44.45	79.08	52.61
38	70.34	49.21	28.70	17.43	40.00	46.26	77.27	49.88
39	69.51	48.87	28.47	18.13	42.22	49.70	75.75	48.68
40	71.84	52.89	28.90	17.10	35.15	39.80	80.23	55.89
41	71.65	51.95	29.73	17.46	38.38	42.42	79.26	54.13
42	70.60	51.65	28.99	17.92	49.89	40.00	77.60	53.26
43	70.23	51.24	29.29	18.38	42.42	47.07	76.58	52.19
44	72.44	53.46	29.69	17.04	35.15	38.79	80.97	56.81
45	71.80	52.41	29.60	16.80	37.37	39.39	79.68	55.38
46	71.88	51.99	30.63	17.63	40.40	43.03	79.08	54.04
47	71.35	51.04	30.51	17.92	42.22	45.25	78.01	52.61
48	72.93	53.80	30.83	17.11	36.57	38.59	81.25	57.27
49	72.37	52.67	30.77	17.29	38.79	40.81	80.05	55.38
50	71.73	52.44	30.62	17.97	41.01	43.64	78.75	54.46
51	70.86	51.20	29.94	18.01	42.22	45.66	77.41	52.57
52	73.87	54.29	32.82	17.49	38.59	39.19	81.94	57.74
53	73.35	53.72	32.63	17.94	40.61	41.62	80.83	56.49
54	72.48	53.57	31.68	18.54	41.41	44.04	79.58	55.75
55	71.32	52.86	30.47	18.71	42.22	45.86	77.91	54.46

56	73.31	54.17	32.35	17.56	39.80	39.60	80.97	57.51
57	72.71	53.72	31.75	17.88	40.61	41.41	80.05	56.54
58	71.39	53.27	30.03	18.31	40.40	43.64	78.76	53.27
59	71.20	52.93	32.99	17.93	38.38	39.80	82.17	58.34
60	74.02	54.89	32.99	18.64	43.43	45.45	77.55	54.64
61	73.27	54.44	32.18	17.91	39.34	40.40	81.02	57.64
62	72.48	53.98	31.40	18.22	40.40	42.22	79.81	56.67
63	71.88	53.57	30.80	18.86	41.01	45.25	78.94	55.47
64	73.57	55.26	32.13	18.26	37.78	40.40	81.76	58.66
65	73.23	54.92	32.01	18.35	38.99	41.21	81.06	58.06
66	73.12	54.36	32.37	18.87	40.81	44.04	80.51	56.72
67	72.52	53.65	31.90	18.68	42.02	44.44	79.49	55.75
68	72.74	54.74	30.70	18.03	36.97	40.40	80.92	58.01
69	72.74	54.51	31.45	18.45	39.39	42.22	80.37	57.32
70	71.77	53.68	30.12	18.26	39.19	42.82	79.21	56.17
71	71.73	53.27	30.73	18.83	41.41	45.66	78.66	55.01
72	73.16	55.04	31.47	18.51	37.58	41.62	81.29	58.11
73	72.59	54.55	31.07	18.74	38.79	43.23	80.32	57.14
74	72.26	54.40	31.10	19.05	40.40	44.65	79.54	56.63
75	71.58	54.21	30.61	19.34	41.62	46.06	78.43	56.07
76	73.05	55.23	31.44	18.87	37.98	42.63	81.06	58.11
77	72.37	55.00	30.83	19.05	38.99	43.64	80.00	57.60
78	71.50	54.62	29.92	18.90	39.60	44.04	78.80	57.04
79	71.17	53.91	30.12	18.84	41.62	44.65	77.92	56.03
80	71.95	61.69	29.53	20.16	36.57	35.76	80.05	67.62
81	71.65	55.08	29.86	18.64	38.79	42.02	79.17	58.06
82	70.90	54.66	29.15	18.79	41.21	44.04	77.18	55.89
83	72.71	54.92	31.52	17.88	39.80	39.60	80.23	58.43
84	70.49	53.68	29.23	18.58	41.21	44.04	77.18	55.89
85	71.92	60.23	30.43	19.30	39.60	35.75	79.31	65.82
86	72.11	52.63	31.26	18.47	41.62	45.25	79.08	54.32

87	71.28	52.26	30.42	18.83	42.22	47.27	77.92	53.39
88	72.18	60.98	30.89	19.73	40.00	35.76	79.54	66.74
89	71.84	66.68	30.82	19.69	41.21	36.16	78.84	66.28
90	70.98	60.49	29.72	19.78	41.01	36.77	77.83	65.91
91	70.83	59.55	29.96	19.90	42.42	38.79	77.32	64.30
92	71.47	60.68	30.18	19.15	40.61	34.55	78.52	66.65
93	71.09	60.34	29.97	19.30	41.41	35.56	77.88	66.00
94	70.83	60.23	30.01	19.83	42.63	37.38	77.27	65.45
95	69.89	59.81	28.75	19.85	41.82	38.18	76.30	64.75
96	70.71	61.54	28.61	19.59	38.38	34.34	78.11	67.76
97	70.53	60.68	28.78	19.42	39.60	35.35	77.60	66.47
98	70.26	60.38	28.80	19.72	40.61	36.77	77.04	65.77
99	69.92	59.85	28.61	19.68	41.21	37.58	76.49	64.94
100	71.43	62.56	29.65	20.49	38.99	35.15	78.85	68.82

ประวัติผู้เขียน

Miss Naiyakwan Chaisri
นางสาวนัยขวัญ ไชยศรี
วัน เดือน ปีเกิด: 1 เมษายน 2540
สถานที่เกิด: จังหวัดพัทลุง
ชั้นปีที่ 4 คณะวิทยาศาสตร์
ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
สาขาวิทยาการคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย
มือถือ: 086-400-5783
อีเมล: wah_za52@hotmail.com