

การวิเคราะห์บรรณนิทัศน์หลังการจำแนกข้อมูลไมโครอะเรย์

นายรัฐพล หลิน

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาดุษฎีบัณฑิต

สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์

คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

ปีการศึกษา 2560


Annotation Analysis after Microarray Data Classification

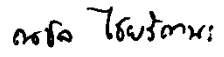
Mr. Rattaphon Lin

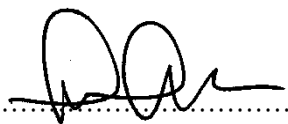
A PROJECT REPORT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE DEGREE OF BACHELOR OF COMPUTER ENGINEERING  
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING  
FACULTY OF ENGINEERING  
KING MONGKUT'S UNIVERSITY OF TECHNOLOGY NORTH BANGKOK  
ACADEMIC YEAR 2017

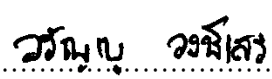
ปริญญานิพนธ์เรื่อง : การวิเคราะห์บรรณนิทัศน์หลังการจำแนกข้อมูลไมโครอะเรย์  
ชื่อ : นายรัฐพล หลิน  
สาขาวิชา : วิศวกรรมคอมพิวเตอร์  
ภาควิชา : วิศวกรรมไฟฟ้าและคอมพิวเตอร์  
คณะ : วิศวกรรมศาสตร์  
อาจารย์ที่ปรึกษา : รองศาสตราจารย์ ดร.ณชด ไชยรัตน์  
ผู้ช่วยศาสตราจารย์ ดร.วรัญญู วงษ์เสรี  
ผู้ช่วยศาสตราจารย์ ดร.ดำรงฤทธิ์ เศรษฐศิริโชค  
ปีการศึกษา : 2560

คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ อนุมัติให้  
ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรบัณฑิต สาขา  
วิชาวิศวกรรมคอมพิวเตอร์

  
..... หัวหน้าภาควิชาวิศวกรรมไฟฟ้า  
(ผู้ช่วยศาสตราจารย์ ดร.ณชด ไชยรัตน์) และคอมพิวเตอร์

  
..... ประธานกรรมการ  
(รองศาสตราจารย์ ดร.ณชด ไชยรัตน์)

  
..... กรรมการ  
(รองศาสตราจารย์ ดร. วรา วราวิทย์)

  
..... กรรมการ  
(ผู้ช่วยศาสตราจารย์ ดร.วรัญญู วงษ์เสรี)

.....

.....


กรรมการ


(ผู้ช่วยศาสตราจารย์ ดร.ดำรงฤทธิ์ เศรษฐศิริโชค)

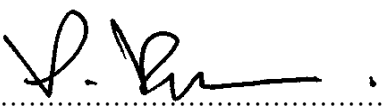
ลิขสิทธิ์ของภาควิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ คณะวิศวกรรมศาสตร์  
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

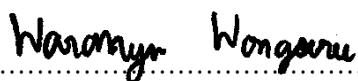
Project Report Title : Annotation Analysis after Microarray Data Classification  
Name : Mr. Rattaphon Lin  
Major Field : Computer Engineering  
Department : Electrical and Computer Engineering  
Faculty : Engineering  
Project Advisors : Assoc. Prof. Dr. Nachol Chaiyaratana  
Asst. Prof. Dr. Waranyu Wongseree  
Asst. Prof. Dr. Damrongrit Setsirichok  
Academic Year : 2017

Accepted by the Faculty of Engineering, King Mongkut's University of Technology  
North Bangkok in Partial Fulfillment of the Requirements for the Degree of Bachelor of Computer  
Engineering

  
.....  
(Asst. Prof. Dr. Nophadon Wiwatcharagoses) Chairperson of Department of Electrical  
and Computer Engineering

  
.....  
(Assoc. Prof. Dr. Nachol Chaiyaratana) Chairperson

  
.....  
(Assoc. Prof. Dr. Vara Varavithya) Member

  
.....  
(Asst. Prof. Dr. Waranyu Wongseree) Member

*Damrongrit Setsirichok.*

Member

(Asst. Prof. Dr. Damrongrit Setsirichok)

Copyright of the Department of Electrical and Computer Engineering, Faculty of Engineering  
King Mongkut's University of Technology North Bangkok

## บทคัดย่อ

ปริญญานิพนธ์นี้นำเสนอผลของการคัดเลือกยีนจากการวิเคราะห์บรรณนิทัศน์ของข้อมูลการแสดงออกของยีนบนไมโครอะเรย์ ตามปกติแล้วการวิเคราะห์บรรณนิทัศน์จะถูกใช้เพื่อระบุฟังก์ชันของยีนที่สำคัญจากชุดข้อมูล โดยฟังก์ชันถูกกำหนดตามทฤษฎีชีววิทยา เนื่องจากยีนบางตัวไม่มีความสำคัญสำหรับการจำแนกตัวอย่าง การเลือกเฉพาะยีนที่มีความสำคัญจะส่งผลต่อผลลัพธ์ในการวิเคราะห์บรรณนิทัศน์ ยีนที่มีความสำคัญจะถูกเลือกโดยใช้เทคนิคประเภทฝังตัวสำหรับการคัดเลือกลักษณะประจำ ปริญญานิพนธ์นี้ใช้ 3 วิธีการทำให้เป็นปกติสำหรับการทดลองโลจิสติก คือ Ridge Regression, Lasso และ Elastic Net บน 3 ชุดข้อมูลคือ Colon Cancer, Leukemia และ Prostate Cancer ผลลัพธ์ที่ได้แสดงให้เห็นว่าจำนวนของยีนที่ถูกเลือกลดลงเมื่อค่าการลงโทษของการทำให้เป็นปกติถูกเปลี่ยนจาก  $l_2$  norm ไปเป็นการรวม  $l_1/l_2$  norm และจากการรวม  $l_1/l_2$  norm ไปเป็น  $l_1$  norm อย่างไรก็ตาม ไม่มีผลแนวโน้มที่ชัดเจนสำหรับจำนวนฟังก์ชันของยีนที่มีความสำคัญ ปริญญานิพนธ์นี้จึงเสนอการใช้เทคนิคการคัดเลือกลักษณะประจำแบบต่าง ๆ สำหรับข้อมูลการแสดงออกของยีนบนไมโครอะเรย์ก่อนการวิเคราะห์บรรณนิทัศน์เนื่องจากผลลัพธ์ที่ได้นั้นยากต่อการทำนาย

## Abstract

This project presents the effects of gene selection on annotation analysis of gene expression microarray data. Usually, annotation analysis is performed to identify significant gene functions, as defined by gene ontology, from the data. Since some genes are not informative for sample classification, selecting only informative genes would affect the outcome of annotation analysis. Informative genes were selected using an embedded approach for attribute selection. Three regularized logistic regression techniques namely ridge regression, lasso and an elastic net were applied to three datasets: colon cancer, leukemia and prostate cancer datasets. The results indicated that the number of selected genes decreased when the regularization penalty changed from an  $l_2$  norm to a combined  $l_1/l_2$  norm and from a combined  $l_1/l_2$  norm to an  $l_1$  norm. However, there was no obvious trend for the resulting number of significant gene functions. This suggests that different attribute selection techniques should be applied to gene expression microarray data prior to annotation analysis since the outcome is difficult to predict.



## กิตติกรรมประกาศ

ปริญญานิพนธ์เล่มนี้ไม่อาจจะเสร็จสมบูรณ์ได้ถ้าปราศจากความช่วยเหลือจาก รองศาสตราจารย์ ดร.ณชล ไชยรัตน์นะ ผู้ช่วยศาสตราจารย์ ดร.วรัญญู วงษ์เสรี และ ผู้ช่วยศาสตราจารย์ ดร.ดำรงฤทธิ เศรษฐศิริโชค ที่คอยช่วยเหลือ ให้คำแนะนำ ตลอดทั้งสนับสนุน ให้ความช่วยเหลือในทุก ๆ ด้าน จนทำให้ปริญญานิพนธ์เล่มนี้เสร็จออกมาครบถ้วนและสมบูรณ์ ต้องขอขอบพระคุณอาจารย์ทุกท่านมา ณ โอกาสนี้

ข้าพเจ้าขอขอบคุณอาจารย์ท่านอื่น ๆ ในภาควิชาวิศวกรรมไฟฟ้าและคอมพิวเตอร์ คณะ วิศวกรรมศาสตร์ ทุกท่านที่คอยสั่งสอน ให้คำแนะนำและให้ความรู้กับข้าพเจ้ามาตลอด 4 ปี จนทำให้ข้าพเจ้ามีความรู้ที่จะสามารถนำไปประกอบอาชีพได้ในอนาคต

ข้าพเจ้าขอขอบคุณมารดาของข้าพเจ้าที่คอยสนับสนุนและให้กำลังใจข้าพเจ้าในตลอดการ เรียนที่ผ่านมา

สุดท้ายนี้ขอขอบคุณเพื่อน ๆ รุ่นพี่ รุ่นน้อง และบุคลากรทุกท่านของภาควิชาวิศวกรรมไฟฟ้า และคอมพิวเตอร์ สาขาวิศวกรรมคอมพิวเตอร์ทุกท่าน ที่คอยให้ความช่วยเหลือในตลอดระยะเวลาที่ ศึกษาอยู่ ณ ที่แห่งนี้

รัฐพล หลิน

## สารบัญ

	หน้า
บทคัดย่อภาษาไทย	๗
บทคัดย่อภาษาอังกฤษ	๗
กิตติกรรมประกาศ	๘
สารบัญตาราง	๙
สารบัญภาพ	๙
บทที่ 1. บทนำ	1
บทที่ 2. ทฤษฎี	2
2.1 การถดถอยโลจิสติก (Logistic Regression)	2
2.2 การคัดเลือกลักษณะประจำ (Feature Selection)	3
2.3 การวิเคราะห์นัยสำคัญของฟังก์ชันและการแสดงออก (Significance Analysis of Function and Expression)	4
2.4 อัตราการค้นพบเท็จ	8
2.5 ขั้นตอนการสร้างแบบจำลอง	9
บทที่ 3. ผลการทดลอง	14
3.1 ผลลัพธ์จากการทำการคัดเลือกลักษณะประจำ	14
3.2 ผลการประเมินประสิทธิภาพแบบจำลอง	15
3.3 ผลลัพธ์จากการทำ SAFE	16
3.4 ผลลัพธ์จากการทำ FDR	19
3.5 ตัวอย่างภววิทยายีนที่พบ	25
3.6 เวลาที่ใช้ในแต่ละการทดลอง	27
บทที่ 4. สรุปผลวิจัย	29
เอกสารอ้างอิง	30
ประวัติผู้แต่ง	32

## สารบัญตาราง

ตารางที่	หน้า
2-1 แหล่งที่มาของชุดข้อมูลและวิธีการประมวลผลก่อน	9
2-2 แสดงวิธีการทำการประมวลผลก่อนด้วยวิธีต่าง ๆ	10
2-3 จำนวนยีนที่เหลื้ก่อนและหลังการประมวลผลก่อน	10
3-1 จำนวนยีนที่สำคัญหลังจากผ่านวิธีการทำให้เป็นปรกติในแต่ละแบบของแต่ละชุดข้อมูล	14
3-2 ค่าเฉลี่ยของผลการประเมินประสิทธิภาพแบบจำลองของชุดข้อมูล Colon Cancer โดยค่าที่อยู่ในวงเล็บคือส่วนเบี่ยงเบนมาตรฐาน	15
3-3 ค่าเฉลี่ยของผลการประเมินประสิทธิภาพแบบจำลองของชุดข้อมูล Leukemia โดยค่าที่อยู่ในวงเล็บคือส่วนเบี่ยงเบนมาตรฐาน	15
3-4 ค่าเฉลี่ยของผลการประเมินประสิทธิภาพแบบจำลองของชุดข้อมูล Prostate Cancer โดยค่าที่อยู่ในวงเล็บคือส่วนเบี่ยงเบนมาตรฐาน	16
3-5 จำนวนประเภทของยีนที่พบในแต่ละวิธีการทำให้เป็นปรกติของแต่ละชุดข้อมูล	16
3-6 ผลลัพธ์ที่ได้จากการทำ FDR ของ Leukemia ด้วยค่า $p$ -value ที่ได้มาจากเทคนิค SAFE ทั้ง 100 ครั้งและแสดงค่าเฉลี่ยของค่า $\pi_0$ ที่พบ	19
3-7 แสดงผลลัพธ์ที่ได้จากการทำ FDR ของ Colon Cancer ด้วยค่า $p$ -value ที่ได้มาจากเทคนิค SAFE ทั้ง 100 ครั้งและแสดงค่าเฉลี่ยของค่า $\pi_0$ ที่พบ	22
3-8 ผลลัพธ์ที่ได้จากการทำ FDR ของ Prostate Cancer ด้วยค่า $p$ -value ที่ได้มาจากเทคนิค SAFE ทั้ง 100 ครั้งและแสดงค่าเฉลี่ยของค่า $\pi_0$ ที่พบ	24
3-9 ตัวอย่างภววิทยายีนที่พบของ Leukemia ด้วยวิธี Lasso ที่ $q$ -value $\leq 0.1$	25
3-10 ตัวอย่างภววิทยายีนที่พบของ Colon Cancer ด้วยวิธี Lasso ที่ $q$ -value $\leq 0.1$	26
3-11 เวลาเฉลี่ยที่ใช้ในการทำบูทสแตรป์แต่ละรอบในหน่วยวินาที โดยค่าที่อยู่ในวงเล็บคือส่วนเบี่ยงเบนมาตรฐาน	27
3-12 เวลาเฉลี่ยในการทำ SAFE แต่ละรอบในหน่วยวินาที โดยค่าที่อยู่ในวงเล็บคือส่วนเบี่ยงเบนมาตรฐาน	28
3-13 เวลาเฉลี่ยในการทำ FDR แต่ละรอบในหน่วยวินาที โดยค่าที่อยู่ในวงเล็บคือส่วนเบี่ยงเบนมาตรฐาน	28

## สารบัญภาพ

ภาพที่	หน้า
2-1 การทำเทคนิค SAFE ซึ่งแสดงลำดับขั้นตอนที่สำคัญต่าง ๆ	4
2-2 ตัวอย่างชุดข้อมูลเข้าสำหรับเทคนิค SAFE	5
2-3 เมทริกซ์ของเวกเตอร์ตอบสนองที่ถูกเรียงสับเปลี่ยน $K$ ครั้ง	6
2-4 เวกเตอร์ผลลัพธ์การคำนวณของสถิติเฉพาะที่ครั้งที่ 1 จนถึงครั้งที่ $K$	6
2-5 เมทริกซ์ประเภทของยีน	7
2-6 เมทริกซ์สถิติครบคลุม	7
2-7 ขั้นตอนการทำงานวิจัย	9
2-8 การตั้งค่าพารามิเตอร์ของการสร้างแบบจำลอง	11
2-9 การตั้งค่าพารามิเตอร์ของเทคนิค SAFE	12
2-10 การตั้งค่าพารามิเตอร์ของ FDR	13
3-1 ตัวอย่างการแจกแจงค่า $p$ -value จากเทคนิค SAFE ของชุดข้อมูล Leukemia	17
3-2 ตัวอย่างการแจกแจงค่า $p$ -value จากเทคนิค SAFE ของชุดข้อมูล Colon Cancer	17
3-3 ตัวอย่างการแจกแจงค่า $p$ -value จากเทคนิค SAFE ของชุดข้อมูล Prostate Cancer	18
3-4 ความถี่ของแต่ละภาววิทยายีนที่พบจากการทำเทคนิค SAFE ทั้งหมด 100 ครั้งของชุดข้อมูล Leukemia ที่ $q\text{-value} \leq 0.1$	20
3-5 ความถี่ของแต่ละภาววิทยายีนที่พบจากการทำเทคนิค SAFE ทั้งหมด 100 ครั้งของชุดข้อมูล Leukemia ที่ $q\text{-value} \leq 0.05$	20
3-6 ตัวอย่างความสัมพันธ์ของการแจกแจง $p$ -value จากเทคนิค SAFE กับค่าเฉลี่ยของ $\pi_0$ ของชุดข้อมูล Leukemia	21
3-7 ความถี่ของแต่ละภาววิทยายีนที่พบจากการทำเทคนิค SAFE ทั้งหมด 100 ครั้งของชุดข้อมูล Colon Cancer	23
3-8 ตัวอย่างความสัมพันธ์ของการแจกแจง $p$ -value จากเทคนิค SAFE กับค่าเฉลี่ยของค่า $\pi_0$ ของชุดข้อมูล Colon Cancer	23

## บทที่ 1

### บทนำ

เป้าหมายของการวิเคราะห์ข้อมูลไมโครอะเรย์ (Microarray Data Analysis) คือการคัดเลือกลักษณะประจำ (Attribute Selection) ทางพันธุกรรมที่จำเป็นสำหรับการจำแนก (Classification) ระหว่างคลาส (Class) ที่สนใจ ในปัจจุบันมีหลายเทคนิคการจำแนกที่เหมาะสมสำหรับการวิเคราะห์ข้อมูลไมโครอะเรย์ อย่างไรก็ตาม การวิเคราะห์ข้อมูลไมโครอะเรย์โดยการจำแนกมักสิ้นสุดที่การระบุลักษณะประจำที่จำเป็นและการประเมินสมรรถนะการจำแนก (Classification Performance) ดังนั้น การวิเคราะห์บรรณนิทัศน์ (Annotation Analysis) หลังการวิเคราะห์ข้อมูลไมโครอะเรย์โดยการจำแนกจะช่วยเพิ่มความสามารถในการอธิบายผลทางพันธุกรรม

ปริญญานิพนธ์นี้นำเสนอวิธีการทำการคัดเลือกลักษณะประจำจากขั้นตอนวิธีการจำแนกของข้อมูลไมโครอะเรย์แล้วจึงทำการวิเคราะห์บรรณนิทัศน์เพื่อหาผลสรุปว่าการทำการจำแนกข้อมูลพร้อมทั้งการคัดเลือกลักษณะประจำก่อนการวิเคราะห์บรรณนิทัศน์มีผลต่อการอธิบายผลทางพันธุกรรมหรือไม่

เนื้อหาปริญญานิพนธ์นี้ประกอบไปด้วย 3 บทคือ บทที่ 2 อธิบายทฤษฎีที่ใช้ในการสร้างแบบจำลองและเทคนิคต่าง ๆ ที่ใช้ในงานวิจัย บทที่ 3 แสดงผลการทดลองของข้อมูลไมโครอะเรย์แต่ละรูปแบบที่ผ่านการทำการจำแนกและการวิเคราะห์บรรณนิทัศน์ บทที่ 4 แสดงผลสรุปของการทดลองของปริญญานิพนธ์นี้

## บทที่ 2

### ทฤษฎี

#### 2.1 การถดถอยโลจิสติก (Logistic Regression)

เป็นการสร้างแบบจำลองเพื่อใช้จำแนกข้อมูลเข้าที่มีผลตอบสนองแบ่งเป็นประเภทหรือเป็นคลาส (Category or Class) โดยพิจารณาจากความน่าจะเป็นของข้อมูลเข้าว่ามีโอกาสที่จะเป็นคลาสใด ในการจำแนกข้อมูลเข้าโดยใช้การถดถอยโลจิสติก ซึ่งข้อมูลเข้าประกอบไปด้วยสองคลาสจะเรียกว่า Binary Logistic Regression (Hastie *et al.*, 2009) และสองคลาสขึ้นไปจะเรียกว่า Multinomial Logistic Regression (Hastie *et al.*, 2009) โดยสมการฟังก์ชันโลจิสติกแบบปกติจะสามารถเขียนได้ในรูป

$$\begin{aligned} f(x) &= \frac{e^x}{1 + e^x} \\ &= \frac{1}{1 + e^{-x}} \end{aligned} \quad (2-1)$$

โดยที่  $x$  คือข้อมูลเข้า

สมการที่ 2-1 สามารถจัดให้อยู่ในรูปของความน่าจะเป็นคลาสเพื่อที่จะใช้ในการสร้างแบบจำลองการถดถอยโลจิสติกได้ดังสมการที่ 2-2

$$p(c_i | x) = \frac{1}{1 + e^{(-\beta^T x)}} \quad (2-2)$$

เมื่อ

$$\beta^T x = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \quad (2-3)$$

โดยที่  $x$  คือเวกเตอร์ข้อมูลเข้า (Input Vector)  $\beta$  คือเวกเตอร์ของสัมประสิทธิ์ (Coefficient Vector)  $p(c_i | x)$  คือความน่าจะเป็นภายหลังของข้อมูลเข้า  $x$  จากคลาส  $c_i$  การประมาณค่าสัมประสิทธิ์ของการถดถอย ( $\beta$ ) สามารถกระทำได้โดยใช้ฟังก์ชันควรจะเป็น (Likelihood Function) ดังสมการที่ 2-4

$$l(\beta) = \prod_{i=1}^{|S|} p(c_i | x_i)^{d_i} (1 - p(c_i | x_i))^{(1-d_i)} \quad (2-4)$$

โดยที่  $d_i$  คือผลลัพธ์ที่ต้องการ (Desired Output) และ  $S$  คือเซตตัวอย่าง

เนื่องจากต้องการหาวิธีเพื่อที่จะให้ได้ค่าฟังก์ชันค่าใช้จ่ายต่ำสุด (Minimum Cost Function) จึงใช้ ฟังก์ชันลอการจะเป็นลบ (Negative Log Likelihood) หรือก็คือฟังก์ชันค่าคลาดเคลื่อนแบบเอนโทรปีไขว้ (Cross Entropy Error Function) สามารถเขียนได้ดังสมการ

$$\begin{aligned}\mathcal{E}_{ce}(\beta) &= -\ln(l(\beta)) \\ &= -\sum_{i=1}^{|S|} (\ln(p(c_1 | x_i)^{d_i}) + \ln(1 - p(c_1 | x_i))^{(1-d_i)}) \\ &= -\sum_{i=1}^{|S|} (d_i \ln p(c_1 | x_i) + (1 - d_i) \ln(1 - p(c_1 | x_i)))\end{aligned}\quad (2-5)$$

และสามารถเขียนให้อยู่ในรูปของฟังก์ชันค่าใช้จ่ายได้ดังสมการ

$$E(\beta) = -\left[\frac{1}{|S|} \sum_{i=1}^{|S|} (d_i \ln p(c_1 | x_i) + (1 - d_i) \ln(1 - p(c_1 | x_i)))\right] \quad (2-6)$$

หลังจากนั้นใช้วิธีการลดตามความชัน (Gradient Descent) เพื่อใช้หาค่าสัมประสิทธิ์ที่เหมาะสมที่สุด

## 2.2 การคัดเลือกลักษณะประจำ (Feature Selection)

การคัดเลือกลักษณะประจำคือขั้นตอนที่ใช้ในการคัดเลือกลักษณะประจำของข้อมูลเข้ามาส่วนหนึ่งจากทั้งหมด ซึ่งจะมีประโยชน์ในการสร้างแบบจำลองเช่น สามารถช่วยลดเวลาในการฝึกสอน ทำให้แบบจำลองง่ายขึ้น และแก้ปัญหการเข้ากับตัวอย่างเกินพอเหมาะ (Overfitting)

วิธีการคัดเลือกลักษณะประจำสามารถทำได้โดยการเพิ่มการลงโทษไปที่ฟังก์ชันค่าใช้จ่าย เพื่อทำการลดลักษณะประจำที่ซ้ำซ้อนหรือไม่เกี่ยวข้องกับแบบจำลองออกไป  
จากสมการที่ 2-6 สามารถเขียนฟังก์ชันค่าใช้จ่ายรวมกับการลงโทษค่าสัมประสิทธิ์ได้ดังนี้

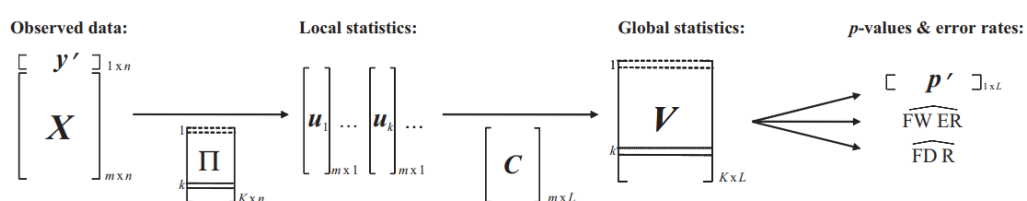
$$\begin{aligned}E(\beta) &= -\left[\frac{1}{|S|} \sum_{i=1}^{|S|} (d_i \ln p(c_1 | x_i) + (1 - d_i) \ln(1 - p(c_1 | x_i)))\right] \\ &\quad + \lambda \left[ \frac{(1 - \alpha) \|\beta\|_2^2}{2} + \alpha \|\beta\|_1 \right]\end{aligned}\quad (2-7)$$

โดย  $\lambda$  คือ พารามิเตอร์ที่ทำให้เป็นปรกติ โดยจะถูกปรับเพื่อหาค่าที่เหมาะสมที่สุดสำหรับแบบจำลอง และ  $\alpha$  คือพารามิเตอร์ที่ใช้ในการเลือกวิธีการลงโทษค่าสัมประสิทธิ์โดย  $\alpha = 0$  คือ Lasso,  $\alpha = 1$  คือ Ridge Regression และ  $0 < \alpha < 1$  คือ Elastic Net

สำหรับวิธีการเลือกค่า  $\lambda$  ให้เหมาะสมนั้นมาจากการสุ่มค่าของ  $\lambda$  ที่ทำให้แบบจำลองมีค่าฟังก์ชันค่าใช้จ่ายหรือค่าความคลาดเคลื่อนน้อยที่สุด

## 2.3 การวิเคราะห์นัยสำคัญของฟังก์ชันและการแสดงออก (Significance Analysis of Function and Expression)

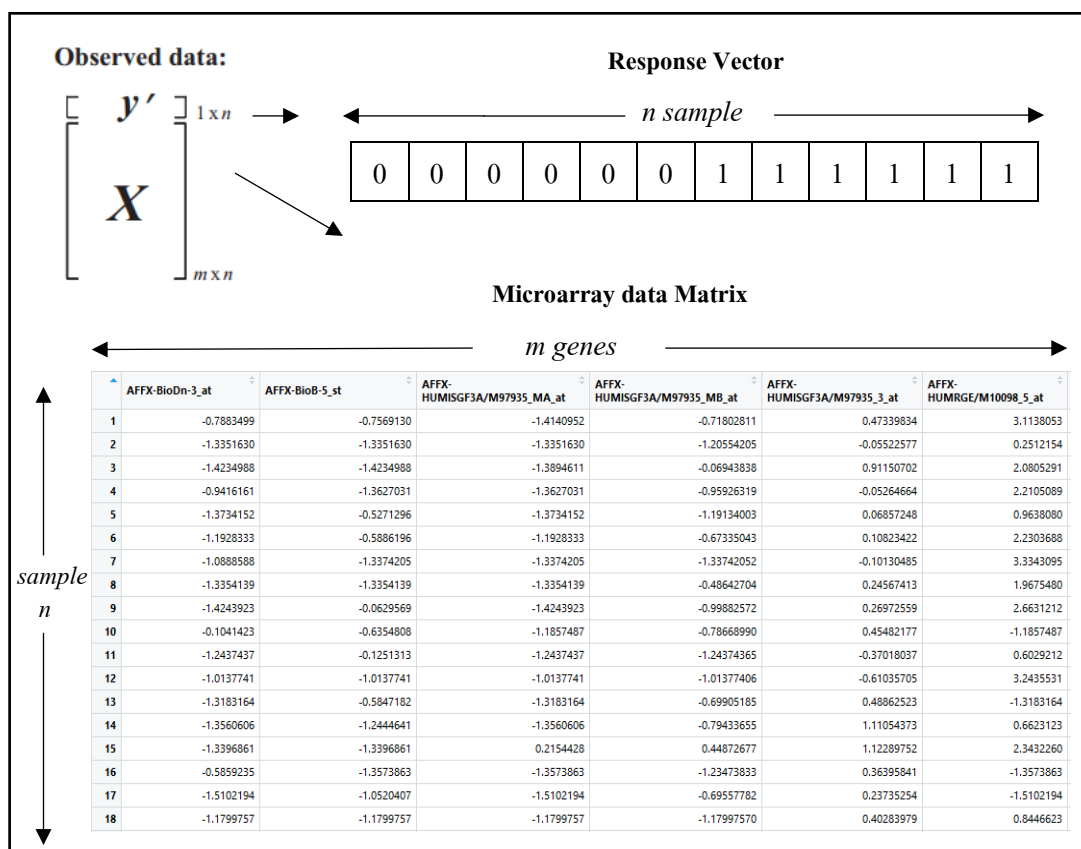
ในส่วนของการวิเคราะห์ในการวิเคราะห์เพื่อหาประเภทของยีน (Gene Category) หรือ ภาววิทยายีน (Gene Ontology) ที่สำคัญจะใช้เทคนิค SAFE (Significance Analysis of Function and Expression) จาก Barry *et al.* (2005) ซึ่งมาจากการคำนวณสถิติเฉพาะที่ (Local Statistic) สถิติครอบคลุม (Global Statistic) และใช้วิธีการประมาณค่าความผิดพลาด (Error Estimation) เพื่อหาประเภทของยีนหรือ ภาววิทยายีนที่สำคัญ โดยภาพรวมของเทคนิค SAFE สามารถแสดงได้ดังภาพที่ 2-1



ภาพที่ 2-1 การทำเทคนิค SAFE ซึ่งแสดงลำดับขั้นตอนที่สำคัญต่าง ๆ (Barry *et al.*, 2004)

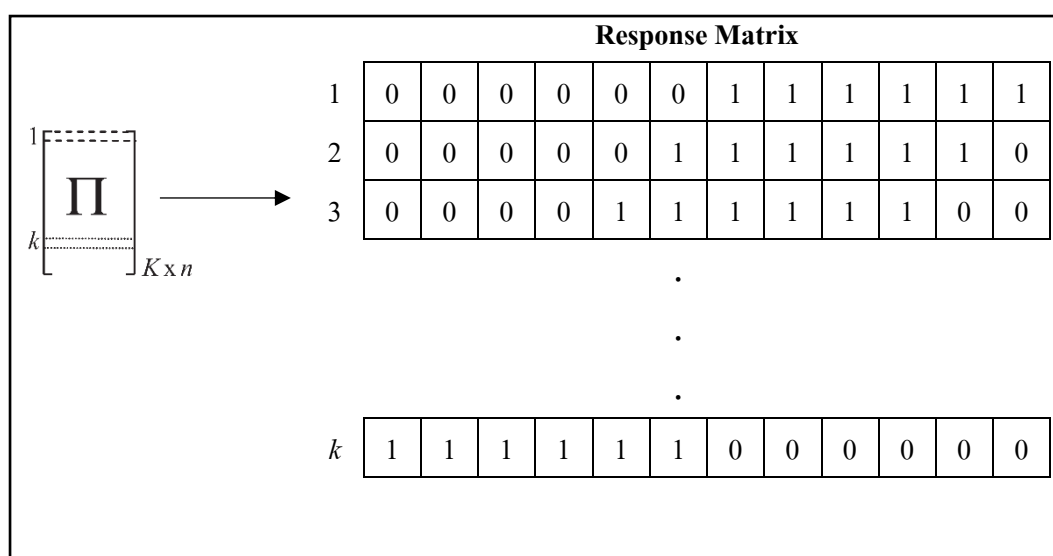
จากภาพที่ 2-1 สามารถอธิบายหลักการทำงานของเทคนิค SAFE ได้ดังนี้ ชุดข้อมูล (Observed data) โดย  $X$  และ  $y'$  คือเมทริกซ์ของข้อมูลเข้าและเวกเตอร์ตอบสนอง โดย  $m$  คือจำนวนยีนและ  $n$  คือขนาดของตัวอย่าง และข้อมูลเข้าต้องถูกทำให้เป็นมาตรฐาน (Standardize) ก่อน โดยตัวอย่างของชุดข้อมูลสามารถแสดงได้ดังภาพที่ 2-2





ภาพที่ 2-2 ตัวอย่างชุดข้อมูลเข้าสำหรับเทคนิค SAFE

$\Pi$  คือเมทริกซ์ของเวกเตอร์ตอบสนองของ  $n$  ตัวอย่าง ที่ถูกเรียงสับเปลี่ยน  $K$  ครั้งเพื่อที่จะใช้ในการคำนวณสถิติครอบคลุมตัวอย่างของการเรียงสับเปลี่ยนแสดงดังภาพที่ 2-3



ภาพที่ 2-3 เมทริกซ์ของเวกเตอร์ตอบสนองที่ถูกเรียงสับเปลี่ยน  $K$  ครั้ง

สถิติเฉพาะที่ (Local Statistic) กำหนดให้  $U(x_i, y)$  คำนวณจากความสัมพันธ์ระหว่างการแสดงออกของยีน  $i$  กับเวกเตอร์ผลตอบสนอง โดยเปรียบเทียบการแจกแจง (Distribution) ของการแสดงออกของยีนในแต่ละคลาสของผลตอบสนอง โดยสามารถเลือกวิธีในการคำนวณสถิติเฉพาะที่ได้ เช่น ใช้  $t$ -test ก็จะได้ค่า  $p$ -value ของแต่ละยีน

$$\begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_k \\ \vdots \end{bmatrix}_{m \times 1}$$

ภาพที่ 2-4 เวกเตอร์ผลลัพธ์การคำนวณของสถิติเฉพาะที่ครั้งที่ 1 จนถึงครั้งที่  $K$

$C$  คือเมทริกซ์ประเภทของยีน โดย  $m$  คือจำนวนของยีนและ  $L$  คือจำนวนประเภทของยีนทั้งหมดที่พบ สมาชิกของเมทริกซ์จะเป็น 1 ก็ต่อเมื่อพบยีน  $m$  อยู่ในประเภทของยีน  $L$  และจะเป็น 0 ถ้ายีน  $m$  ไม่ได้อยู่ในประเภทของยีน  $L$

$$\begin{bmatrix} \mathbf{C} \end{bmatrix}_{m \times L}$$

ภาพที่ 2-5 เมทริกซ์ประเภทของยีน

สถิติครอบคลุม (Global Statistic)  $V$  เปรียบเทียบการแจกแจงระหว่างสถิติเฉพาะที่ที่อยู่ในประเภทของยีนและอยู่นอกประเภทของยีน และทำการคำนวณใหม่ซ้ำโดยใช้เวกเตอร์ตอบสนองที่ถูกเรียงสับเปลี่ยนแทนที่แล้วทำการคิดค่า  $p$ -value จากสมการ

$$p_l = \frac{\sum_{k=1}^K I\{v_{kl} > v_{ll}\}}{K} \quad (2-8)$$

โดยที่  $I$  เป็นฟังก์ชันตัวชี้บอก (Indicator Function)  $v_{ll}$  คือค่าสถิติครอบคลุมที่คิดจากการทำครั้งแรก  $v_{kl}$  คือค่าสถิติครอบคลุมที่คิดจากการทำการเรียงสับเปลี่ยนครั้งที่  $k$

$$\begin{bmatrix} \mathbf{V} \end{bmatrix}_{K \times L}$$

ภาพที่ 2-6 เมทริกซ์สถิติครอบคลุม

## 2.4 อัตราการค้นพบเท็จ (False Discovery Rate; FDR)

อัตราการค้นพบเท็จ เป็นการแก้ไขสำหรับการทดสอบหลายสมมุติฐาน (Storey *et al.*, 2004; Storey & Tibshirani, 2003) โดย FDR มีการคำนวณค่า  $\pi_0$  และ  $q$ -value ในการทดสอบความสำคัญ ซึ่งมีขั้นตอนการคำนวณดังนี้

1. กำหนดให้  $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$  คือค่าของ  $p$ -value ที่ถูกเรียงลำดับ

2. สำหรับช่วงของ  $\lambda \in R$  โดยที่  $R = \{0, 0.05, 0.01, \dots, 0.95\}$  ทำการคำนวณค่า  $\pi_0(\lambda)$  จากสมการ

$$\pi_0(\lambda) = \frac{\#\{p_j > \lambda\}}{m(1 - \lambda)} \quad (2-9)$$

3. สำหรับช่วงของ  $\lambda \in R$  จาก  $B$  ของการทำบูทสแตรป์ (Boostrapping) ทำการคำนวณ  $\pi_0^{*b}(\lambda)$  โดยที่  $b = 1, \dots, B$  โดยใช้ตัวอย่างของ  $p$ -value ในการทำบูทสแตรป์

4. ในแต่ละ  $\lambda \in R$  ทำการคำนวณค่า  $MSE$  ได้จาก

$$MSE(\lambda) = \frac{1}{B} \sum_{b=1}^B \left[ \pi_0^{*b}(\lambda) - \min_{\lambda' \in R} \{\pi_0(\lambda')\} \right]^2 \quad (2-10)$$

5. กำหนดให้

$$\hat{\lambda} = \arg \min_{\lambda \in R} \{MSE(\lambda)\} \quad (2-11)$$

สามารถประมาณค่า  $\pi_0$  ได้จาก

$$\pi_0 = \pi_0(\hat{\lambda}) \quad (2-12)$$

6. ทำการคำนวณ

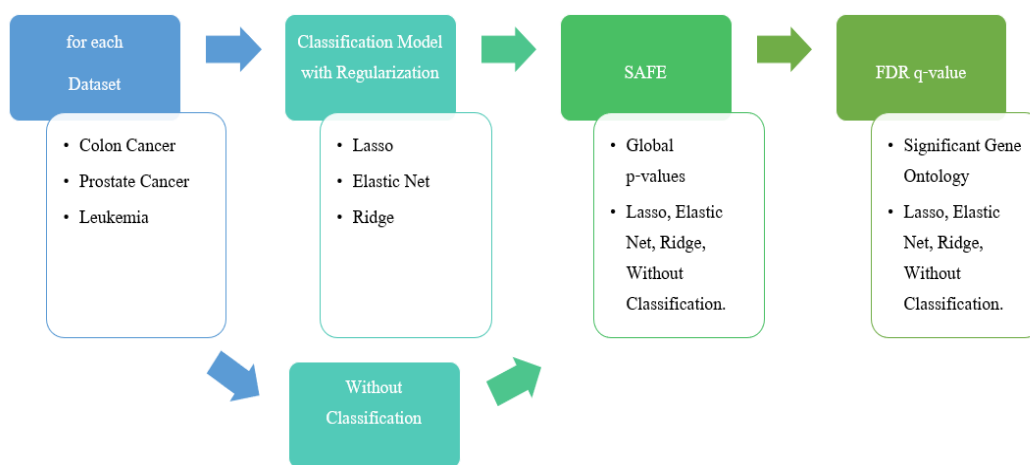
$$\hat{q}(p_{(m)}) = \min_{t \geq p_{(m)}} \frac{\hat{\pi}_0 m \cdot t}{\#\{p_j \leq t\}} = \hat{\pi}_0 \cdot p_{(m)} \quad (2-13)$$

7. สำหรับ  $i = m-1, m-2, \dots, 1$  ทำการคำนวณ

$$\hat{q}(p_{(i)}) = \min_{t \geq p_{(i)}} \frac{\hat{\pi}_0 m \cdot t}{\#\{p_j \leq t\}} = \left( \min \left( \frac{\hat{\pi}_0 m \cdot p_{(i)}}{i}, \hat{q}(p_{(i+1)}) \right) \right) \quad (2-14)$$

## 2.5 ขั้นตอนการสร้างแบบจำลอง

งานวิจัยนี้ถูกทำขึ้นบนภาษา R ทั้งหมด และมีขั้นตอนสองส่วนหลักที่สำคัญคือ การสร้างแบบจำลองเพื่อทำการจำแนกพร้อมวิธีการคัดเลือกลักษณะประจำ และการทำการวิเคราะห์บรรณนิทัศน์ หลังการจำแนกเพื่อหาความสำคัญของภาววิทยายีน (Gene Ontology) โดยหัวข้อ 2.5.1-2.5.2 จะอธิบายวิธีการสร้างแบบจำลอง และหัวข้อ 2.5.3-2.5.4 จะอธิบายวิธีการทำการวิเคราะห์บรรณนิทัศน์ และขั้นตอนการทำวิจัยสามารถสรุปได้ดังภาพที่ 2-7



ภาพที่ 2-7 ขั้นตอนการทำงานวิจัย

### 2.5.1 ขั้นตอนการประมวลผลก่อน (Preprocessing)

งานวิจัยนี้ใช้ชุดข้อมูล 3 ชุดคือ Colon Cancer, Prostate Cancer และ Leukemia โดยทั้ง 3 ชุดข้อมูลประกอบไปด้วย 2 คลาส แหล่งที่มาของชุดข้อมูลและวิธีการประมวลผลก่อนแสดงดังตารางที่ 2-1

ตารางที่ 2-1 แหล่งที่มาของชุดข้อมูลและวิธีการประมวลผลก่อน

Dataset	Dataset Publication	Preprocess Publication
Colon Cancer	Alon <i>et al.</i> (1999)	Dettling (2004)
Prostate Cancer	Singh <i>et al.</i> (2002)	Chang <i>et al.</i> (2007)
Leukemia	Golub <i>et al.</i> (1999)	Dudoit <i>et al.</i> (2002)

หลังจากได้รับชุดข้อมูลครบแล้วจึงทำการประมวลผลก่อนในแต่ละชุดข้อมูล ยกเว้น Colon Cancer เนื่องจากได้ถูกประมวลผลก่อนมาแล้ว ส่วนชุดข้อมูลที่เหลือทำการประมวลผลก่อนโดยวิธีการดังนี้

1. การปรับค่าตามขีดเริ่มเปลี่ยน (Thresholding) ด้วยวิธีการจำกัดค่าที่เป็นไปได้สูงสุดและต่ำสุด (Floor, Ceiling)
2. การกรอง (Filtering) เลือกยีนซึ่งผลหารของค่าน้อยสุดกับค่ามากที่สุด (Max/Min) มากกว่าค่าที่กำหนด และผลต่างระหว่างค่ามากที่สุดและค่าน้อยสุด (Max-Min) มากกว่าค่าที่กำหนด
3. การแปลงลอการิทึมฐาน 10
4. การทำให้เป็นมาตรฐาน (Standardize) ระหว่างยีน

รายละเอียดของการทำ Thresholding และ Filtering ของแต่ละชุดข้อมูลแสดงดังตารางที่ 2-2

ตารางที่ 2-2 แสดงวิธีการทำการประมวลผลก่อนด้วยวิธีต่าง ๆ

Dataset	Floor	Ceiling	Max/Min	Max-Min
Colon Cancer	-	-	-	-
Leukemia	100	16000	5	500
Prostate	10	16000	5	50

หลังจากผ่านการทำประมวลผลก่อนแล้วจำนวนยีนที่เหลือถูกแสดงดังตารางที่ 2-3

ตารางที่ 2-3 จำนวนยีนที่เหลือก่อนและหลังการประมวลผลก่อน

Dataset	Samples	Original Genes	Preprocessed Genes	Category
Colon	62	2000	1872	Tumor/Normal
Leukemia	72	7129	3571	Subtypes of Leukemia
Prostate	102	12600	5966	Tumor/Normal

จากตารางที่ 2-3 ที่จำนวนยีนใน Colon ลดจาก 2000 เหลือ 1872 ทั้ง ๆ ที่ชุดข้อมูล Colon ถูกประมวลผลก่อนมาแล้วเนื่องจากข้อมูล EST ไม่สามารถที่จะระบุออกมาเป็นยีนมนุษย์ได้จึงตัดบางยีนออก

### 2.5.2 องค์ประกอบและวิธีการสร้างแบบจำลองการจำแนก

ในส่วนของการสร้างแบบจำลองจะใช้ library glmnet ซึ่งสามารถสร้างแบบจำลองการจำแนก พร้อมทำ การตรวจสอบความสมเหตุสมผลไขว้ (Cross Validation) และทำการคัดเลือกลักษณะประจำพร้อมกันได้ โดยแบบจำลองที่ใช้คือแบบจำลองการถดถอยโลจิสติก โดยชุดข้อมูลที่ถูกนำมาใช้เป็นข้อมูลเข้าของแบบจำลองคือชุดข้อมูลทั้ง 3 ชุดที่แสดงใน หัวข้อ 2.5.1 ที่ถูกผ่านการประมวลผลก่อนแล้ว โดยการสร้างแบบจำลองจะมีการตั้งค่าพารามิเตอร์ต่าง ๆ ตามภาพที่ 2-8

```
cv.glmnet(data.x, data.y, family = "binomial", alpha = alph , type.measure = "deviance",
nolds = 10 )
```

**ภาพที่ 2-8** การตั้งค่าพารามิเตอร์ของการสร้างแบบจำลอง

จากภาพที่ 2-8 อธิบายพารามิเตอร์ต่าง ๆ ได้ดังนี้

- data.x คือเมทริกซ์ของข้อมูลเข้า และ data.y คือเวกเตอร์ตอบสนองของข้อมูลเข้า
- family = “binomial” คือการสร้างแบบจำลองถดถอยโลจิสติกและเวกเตอร์ตอบสนองมีสองคลาส
- alpha คือการกำหนดวิธีการทำให้เป็นปกติ โดยตั้งค่าเป็น 0, 0.5 และ 1 ซึ่งก็คือ Lasso, Elastic Net และ Ridge Regression ตามลำดับ
- type.measure = “deviance” วิธีการประเมินแบบจำลองโดยใช้ดัชนี Deviance
- nolds = 10 คือการทำการตรวจสอบความสมเหตุสมผลไขว้ (Cross Validation) ด้วย 10 folds

หลังจากผ่านขั้นตอนการจำแนกในแต่ละชุดข้อมูลก็จะสามารถแบ่งเพิ่มเป็นชุดข้อมูลได้อีก 3 แบบคือ ชุดข้อมูลที่ถูกจำแนกพร้อมทำให้เป็นปกติ ด้วย Lasso, Elastic Net และ Ridge Regression ซึ่งในแต่ละวิธีในการทำให้เป็นปกติ ก็จะได้อ่า  $\lambda$  (ที่ถูกอธิบายไว้ในหัวข้อ 2.2) ซึ่ง  $\lambda_{\min}$  คือค่า  $\lambda$  ที่ทำให้แบบจำลองมีค่า Deviance น้อยสุด และ  $\lambda_{1se}$  คือค่า  $\lambda$  ที่มากที่สุดที่มี Deviance น้อยกว่าค่า Deviance น้อยสุดบวก 1 ค่าคลาดเคลื่อนมาตรฐาน (Standard Error)

เนื่องจากชุดข้อมูลมีจำนวนตัวอย่างมีน้อย จึงใช้วิธีบูทสแตรป์ 200 ครั้งกับชุดตัวอย่าง โดยใช้ชุดข้อมูลที่ถูกสุ่มได้มาทำการสร้างแบบจำลองการถดถอยในบูทสแตรป์อีกครั้ง พร้อมใช้ค่า  $\lambda$  ที่ได้จากการสร้างแบบจำลองการถดถอยก่อนการทำบูทสแตรป์ โดยใช้  $\lambda_{1se}$  ที่ได้มาจากการสร้างแบบจำลองที่ได้อธิบายไว้ข้างต้น หลังจากนั้นทำการเก็บค่าสัมประสิทธิ์ (Coefficients) ที่

ได้จากการสร้างแบบจำลองในบูทสแตรป์แต่ละรอบเพื่อมาทำ  $t$ -test โดยตั้งสมมุติฐานหลักเท่ากับ 0 และขอบเขตความเชื่อมั่นเท่ากับ 95% ก็จะได้อินที่ค่าสำคัญสำหรับแต่ละแบบจำลองออกมา

ในส่วนของการประเมินประสิทธิภาพของแบบจำลองจะใช้แบบจำลองที่ได้จากการทำบูทสแตรป์ในแต่ละรอบมาทำการประเมินประสิทธิภาพ โดยข้อมูลที่ถูกสุ่มเพื่อใช้ในการสร้างแบบจำลองจะเป็นข้อมูลฝึกสอน (Training Set) และข้อมูลที่ไม่ถูกสุ่มจะเป็นข้อมูลที่ใช้ทดสอบ หลังจากนั้นทำการสร้างเมทริกซ์ความสับสน (Confusion Matrix) จากผลการทำนายของแบบจำลอง เพื่อคำนวณ ความแม่นยำ (Accuracy) ความไว (Sensitivity) ความจำเพาะ (Specificity) และพื้นที่ใต้เส้นโค้งอาร์โอซี (Area Under the Curve; AUC) โดยผลลัพธ์ต้องถูกถ่วงน้ำหนักดังสมการ

$$e_{Bootstrap} = 0.632e_{Test} + 0.368e_{Train} \quad (2-15)$$

### 2.5.3 การทำการวิเคราะห์หัยสำคัญของฟังก์ชันและการแสดงออก

นำข้อมูลที่ได้จากการสร้างแบบจำลองมาทำการทำการวิเคราะห์หัยสำคัญของฟังก์ชันและการแสดงออก (SAFE) ซึ่งผลลัพธ์ที่ได้ก็คือ  $p$ -value ของแต่ละประเภทของยีน โดยทำการทดลอง 100 ครั้ง และการทำเทคนิค SAFE จะมีการตั้งค่าพารามิเตอร์ต่าง ๆ ได้ตามภาพที่ 2-9

```
safe(data.x, data.y, platform = chip, annotate = c("GO.BP", "GO.MF", "GO.CC"),
error = "none", local = "t.Student", global = "Wilcoxon", Pi.mat = 10000)
```

ภาพที่ 2-9 การตั้งค่าพารามิเตอร์ของเทคนิค SAFE

จากภาพที่ 2-9 อธิบายพารามิเตอร์ต่าง ๆ ได้ดังนี้

- data.x และ data.y คือชุดข้อมูลเข้าและเวกเตอร์ตอบสนองตามลำดับ
- platform คือชิปที่ใช้ในการสร้างเมทริกซ์ประเภทของยีนในแต่ละชุดข้อมูล โดย Colon Cancer, Leukemia และ Prostate Cancer ใช้ hgu133plus2.db, hu6800.db และ ใช้ hgu95av2.db ตามลำดับ
- annotate คือการเลือก GO term แต่ละชนิดที่ต้องการทำการวิเคราะห์บรรณนิทัศน์
- local และ global คือวิธีการทางสถิติที่เลือกใช้ โดยใช้  $t$ -test และ Wilcoxon rank sum ตามลำดับ
- Pi.mat คือจำนวนการทำการเรียงสับเปลี่ยนเท่ากับ 10,000 ครั้ง



- error คือ การแก้ไขสำหรับการทดสอบหลายสมมุติฐานในขั้นตอนนี้ไม่ได้ทำเพราะจะใช้ FDR ในการคำนวณที่หลัง

#### 2.5.4 การใช้ FDR เพื่อวิเคราะห์หาทวิทายีนที่สำคัญ

ใช้  $p$ -value ที่ได้จากการทำเทคนิค SAFE 100 ครั้ง โดยผลลัพธ์ที่ได้จะแสดงในบทที่ 3 โดยการทำให้ FDR จะมีการตั้งค่าพารามิเตอร์ต่าง ๆ ตามภาพที่ 2-10

```
qvalue(pval, fdr.level=0.1, pi0.method="bootstrap")
```

ภาพที่ 2-10 การตั้งค่าพารามิเตอร์ของ FDR

จากภาพที่ 2-10 อธิบายพารามิเตอร์ต่าง ๆ ได้ดังนี้

- pval คือ  $p$ -value ที่ได้จาก SAFE
- fdr.level คือ ระดับที่ใช้ในการควบคุม FDR โดยจะดูว่าค่า  $q$ -value ที่ได้มาน้อยกว่าค่าที่กำหนดไว้หรือไม่ โดยกำหนดให้เท่ากับ 0.05 และ 0.1
- pi0.method คือ วิธีที่ใช้ในการปรับพารามิเตอร์สำหรับประมาณค่า  $\pi_0$  โดยเลือกใช้วิธีบูทสแตรัป

## บทที่ 3

### ผลการทดลอง

การวิเคราะห์ผลที่เกิดขึ้นจากการทำงานวิจัยนี้ต้องใช้ผลการทดลองที่มาจากการทำเทคนิค SAFE และผลลัพธ์ที่ได้จากการทำ FDR ด้วยวิธี  $q$ -value เพื่อที่จะหาสาเหตุที่ทำให้พบภววิทยายีนที่สำคัญ

#### 3.1 ผลลัพธ์จากการทำการคัดเลือกลักษณะประจำ

จากวิธีการในหัวข้อ 2.5.2 จำนวนยีนที่สำคัญหลักจากผ่านวิธีการทำให้เป็นปรกติแสดงได้ดังตารางที่ 3-1

ตารางที่ 3-1 จำนวนยีนที่สำคัญหลังจากผ่านวิธีการทำให้เป็นปรกติในแต่ละแบบของแต่ละชุดข้อมูล

Dataset	Original	Ridge Regression	Elastic Net	Lasso
Colon	1872	1754	122	37
Leukemia	3571	3323	157	74
Prostate	5966	5555	222	73

### 3.2 ผลการประเมินประสิทธิภาพแบบจำลอง

ผลลัพธ์จากการประเมินประสิทธิภาพแบบจำลองที่บทสแตร์ปของชุดข้อมูล Colon Cancer Leukemia และ Prostate Cancer แสดงดังตารางที่ 3-2, 3-3 และ 3-4 ตามลำดับ

**ตารางที่ 3-2** ค่าเฉลี่ยของผลการประเมินประสิทธิภาพแบบจำลองของชุดข้อมูล Colon Cancer โดยค่าที่อยู่ในวงเล็บคือส่วนเบี่ยงเบนมาตรฐาน

Regularization	Accuracy	Sensitivity	Specificity	AUC
Lasso	0.8472 (0.0654)	0.7012 (0.1871)	0.9361 (0.0466)	0.9238 (0.0409)
Elastic Net	0.8578 (0.0695)	0.7289 (0.1922)	0.9380 (0.0431)	0.9264 (0.0394)
Ridge Regression	0.8406 (0.0730)	0.6976 (0.1995)	0.9313 (0.0497)	0.9272 (0.0363)

**ตารางที่ 3-3** ค่าเฉลี่ยของผลการประเมินประสิทธิภาพแบบจำลองของชุดข้อมูล Leukemia โดยค่าที่อยู่ในวงเล็บคือส่วนเบี่ยงเบนมาตรฐาน

Regularization	Accuracy	Sensitivity	Specificity	AUC
Lasso	0.7777 (0.0547)	0.9355 (0.0634)	0.4893 (0.1631)	0.8367 (0.0618)
Elastic	0.7804 (0.0603)	0.9519 (0.0591)	0.4684 (0.1768)	0.8555 (0.0542)
Ridge Regression	0.7971 (0.0614)	0.9661 (0.0512)	0.4934 (0.1734)	0.8706 (0.0599)

**ตารางที่ 3-4** ค่าเฉลี่ยของผลการประเมินประสิทธิภาพแบบจำลองของชุดข้อมูล Prostate Cancer โดยค่าที่อยู่ในวงเล็บคือส่วนเบี่ยงเบนมาตรฐาน

Regularization	Accuracy	Sensitivity	Specificity	AUC
Lasso	0.9355 (0.0225)	0.9380 (0.0415)	0.9366 (0.0392)	0.9747 (0.0152)
Elastic	0.9438 (0.0229)	0.9493 (0.0380)	0.9414 (0.0386)	0.9749 (0.0180)
Ridge Regression	0.9244 (0.0333)	0.9265 (0.0583)	0.9271 (0.0384)	0.9597 (0.0227)

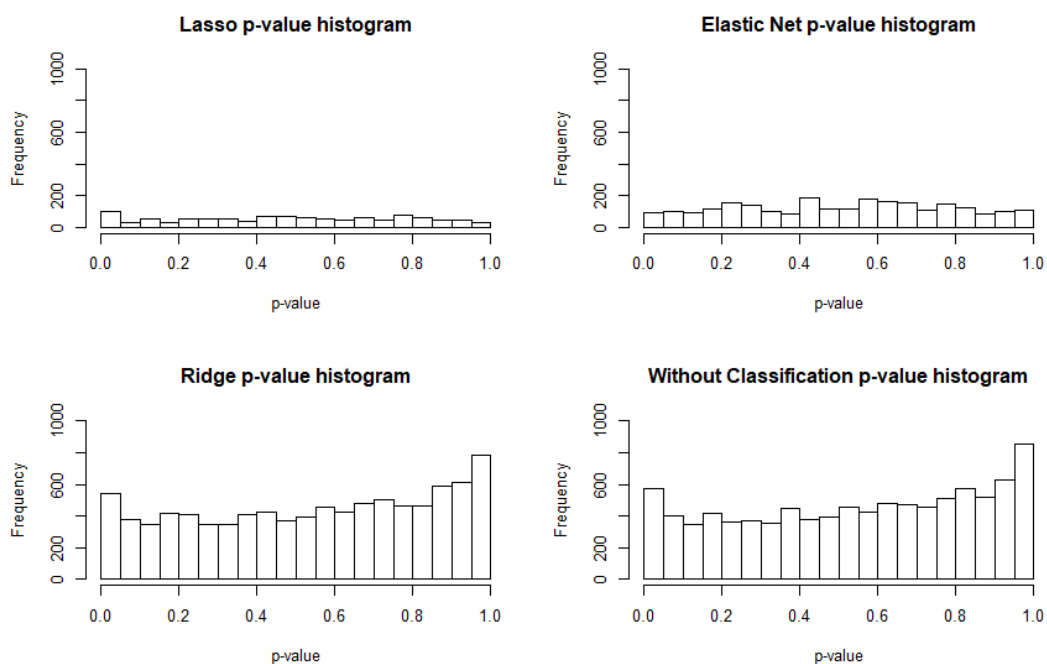
### 3.3 ผลลัพธ์จากการทำ SAFE

การทำเทคนิค SAFE สิ่งที่น่าสนใจคือจำนวนประเภทของยีนที่พบของแต่ละชุดข้อมูลในแต่ละวิธีการทำให้เป็นปกติ ซึ่งจะแสดงให้เห็นความแตกต่างของจำนวนยีนที่พบและจำนวนยีนที่สำคัญดังตารางที่ 3-5 อีกส่วนหนึ่งที่น่าสนใจคือการแจกแจงของค่า  $p$ -value ของสถิติครอบคลุม

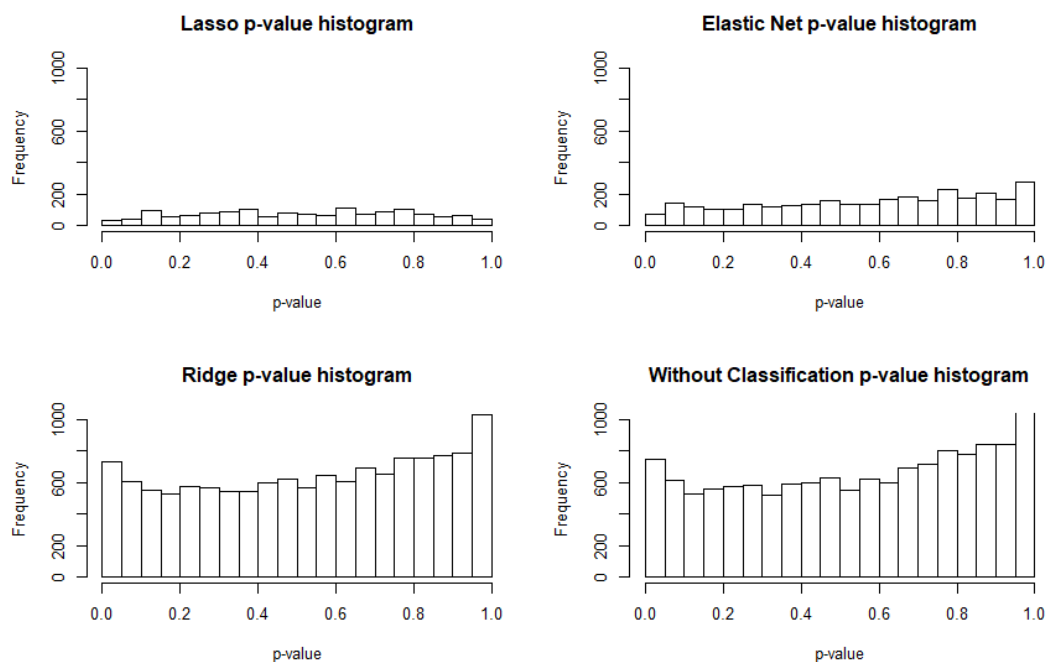
**ตารางที่ 3-5** จำนวนประเภทของยีนที่พบในแต่ละวิธีการทำให้เป็นปกติของแต่ละชุดข้อมูล

Classification Model	Colon Cancer	Leukemia	Prostate Cancer
Lasso	1034	1555	1429
Elastic Net	2434	2466	3045
Ridge Regression	9142	11276	13144
Without Classification Model	9392	11635	13464

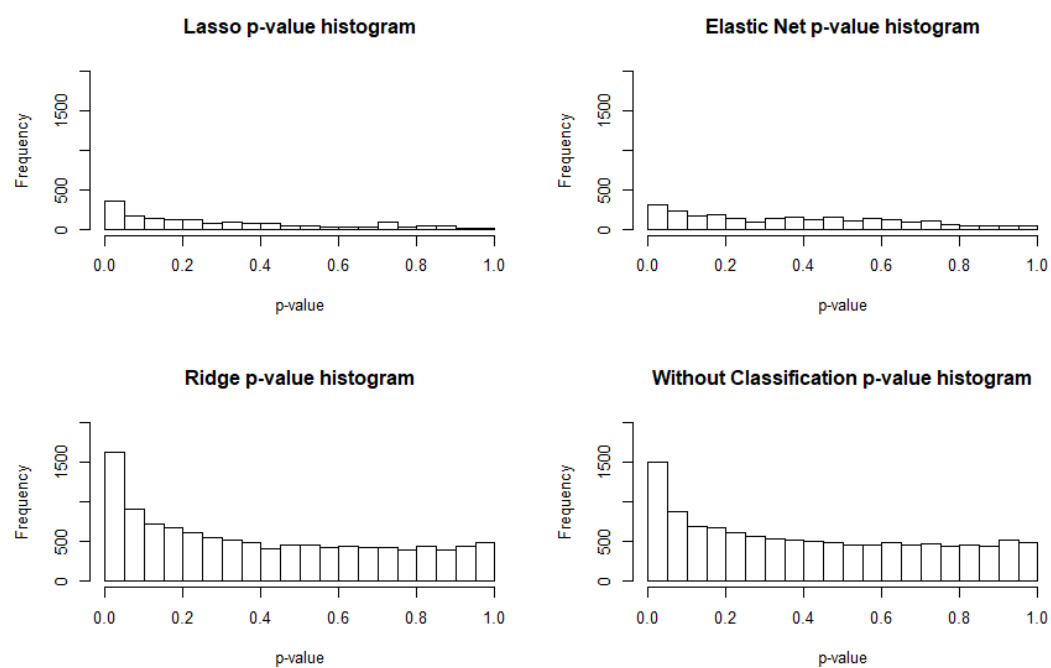
จากการใช้เทคนิค SAFE ในการทดลองทั้งหมด 100 ครั้ง ทำการยกตัวอย่างการแจกแจงค่า  $p$ -value ของสถิติครอบคลุม 1 ครั้งจากการทดลองทั้งหมด 100 ครั้ง เพื่อแสดงตัวอย่างการแจกแจงของค่า  $p$ -value ของสถิติครอบคลุมของแต่ละชุดข้อมูลในแต่ละวิธีการทำให้เป็นปกติโดยค่า  $p$ -value สามารถคำนวณได้จากสมการ 2-7 ซึ่งสามารถแสดงได้ดังภาพที่ 3-1, 3-2 และ 3-3 โดยแสดงข้อมูล Leukemia, Colon Cancer และ Prostate Cancer ตามลำดับ



ภาพที่ 3-1 ตัวอย่างการแจกแจงค่า  $p$ -value จากเทคนิค SAFE ของชุดข้อมูล Leukemia



ภาพที่ 3-2 ตัวอย่างการแจกแจงค่า  $p$ -value จากเทคนิค SAFE ของชุดข้อมูล Colon Cancer



ภาพที่ 3-3 ตัวอย่างการแจกแจงค่า  $p$ -value จากเทคนิค SAFE ของชุดข้อมูล Prostate Cancer

### 3.4 ผลลัพธ์จากการทำ FDR

ผลลัพธ์จากการใช้ FDR จาก global  $p$ -value ที่ได้มาจากการทำเทคนิค SAFE เพื่อหาทวิทายีนที่สำคัญของแต่ละชุดข้อมูลในแต่ละวิธีการทำให้เป็นปรกติจากการทดลองทั้งหมด 100 ครั้งและสนใจจำนวนการทดลองที่ให้ผลตรงกัน 80 ครั้งขึ้นไป โดยค่า  $\pi_0$  และ  $q$ -value คัดจากวิธีการในหัวข้อ 2.4

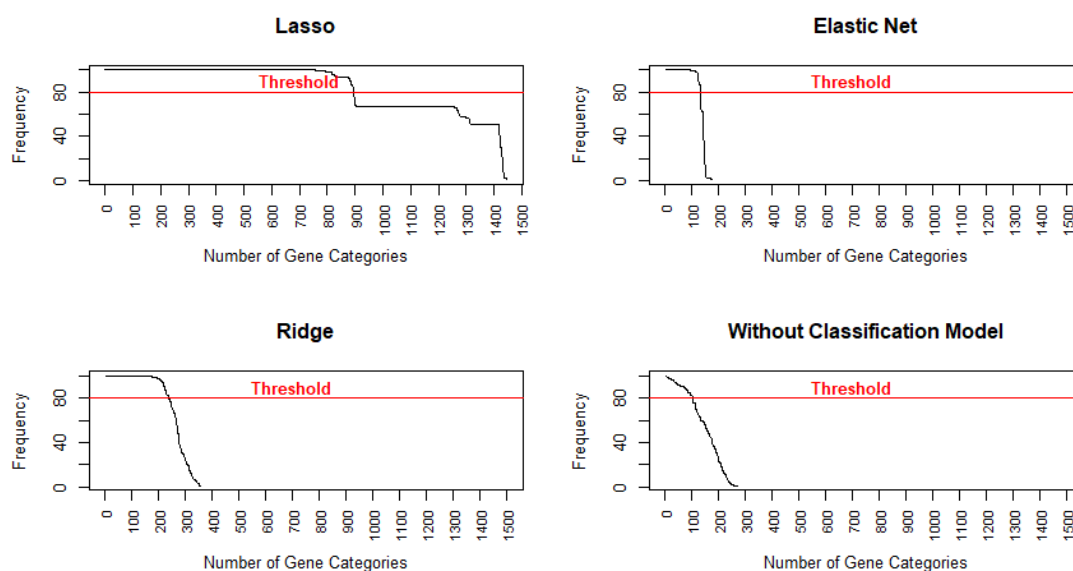
#### 3.4.1 ผลลัพธ์จากชุดข้อมูล Leukemia

ผลลัพธ์จากการใช้วิธี FDR พบจำนวนทวิทายีนที่สำคัญดังตารางที่ 3-6

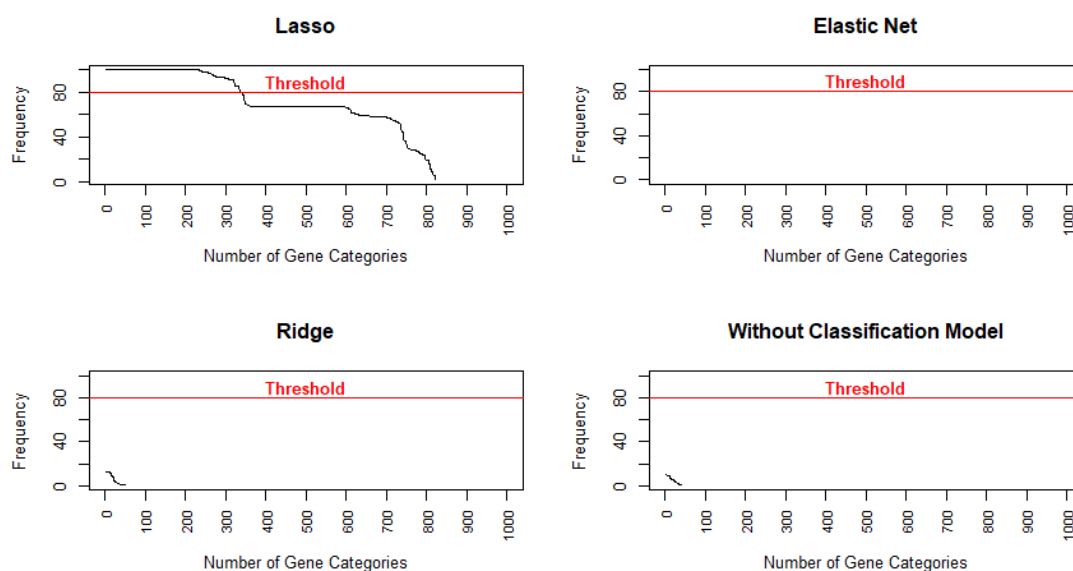
ตารางที่ 3-6 ผลลัพธ์ที่ได้จากการทำ FDR ของ Leukemia ด้วยค่า  $p$ -value ที่ได้มาจากเทคนิค SAFE ทั้ง 100 ครั้งและแสดงค่าเฉลี่ยของค่า  $\pi_0$  ที่พบ

Classification Model	$\pi_0$	Total Number of Gene Category			
		$q\text{-value} \leq 0.1$		$q\text{-value} \leq 0.05$	
		Total	Freq. $\geq 80$	Total	Freq. $\geq 80$
Lasso	0.16214	1447	894	821	339
Elastic Net	0.33001	175	130	0	0
Ridge Regression	0.76809	356	237	39	0
Without Classification Model	0.80686	269	102	49	0

ข้อมูลจากที่ตาราง 3-6 สามารถสร้างกราฟความถี่ของแต่ละทวิทายีนที่พบได้ดังภาพที่ 3-4 และ 3-5 และแสดงตัวอย่างความสัมพันธ์ของการแจกแจง  $p$ -value จากเทคนิค SAFE กับค่าเฉลี่ยของ  $\pi_0$  ได้ดังภาพที่ 3-6

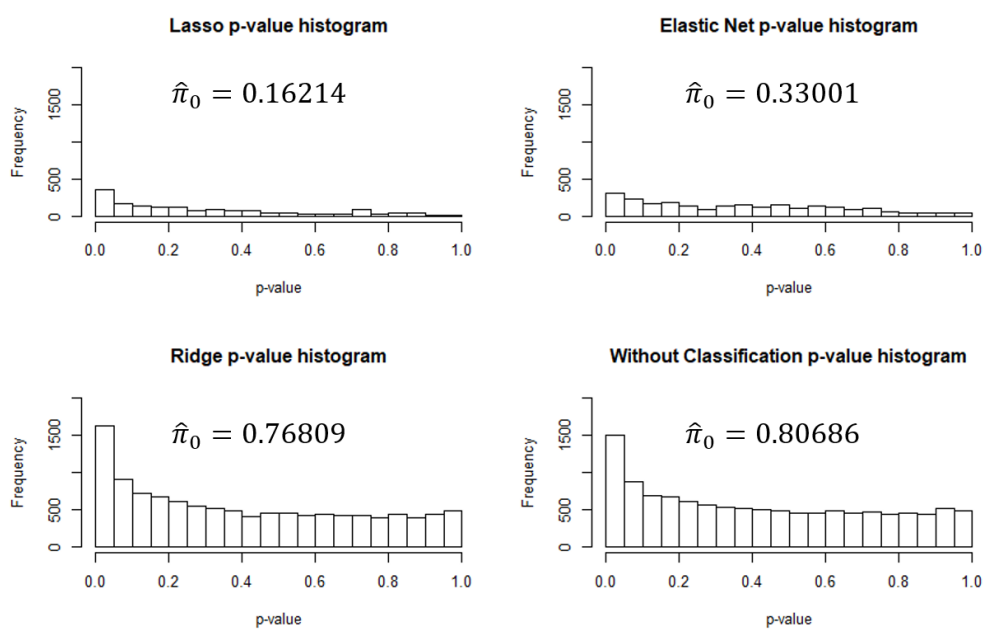


ภาพที่ 3-4 ความถี่ของแต่ละภาววิทยีนที่พบจากการทำเทคนิค SAFE ทั้งหมด 100 ครั้งของชุดข้อมูล Leukemia ที่  $q\text{-value} \leq 0.1$



ภาพที่ 3-5 ความถี่ของแต่ละภาววิทยีนที่พบจากการทำเทคนิค SAFE ทั้งหมด 100 ครั้งของชุดข้อมูล Leukemia ที่  $q\text{-value} \leq 0.05$





ภาพที่ 3-6 ตัวอย่างความสัมพันธ์ของการแจกแจง  $p$ -value จากเทคนิค SAFE กับค่าเฉลี่ยของ  $\pi_0$  ของชุดข้อมูล Leukemia

### 3.4.2 ผลลัพธ์จากชุดข้อมูล Colon Cancer

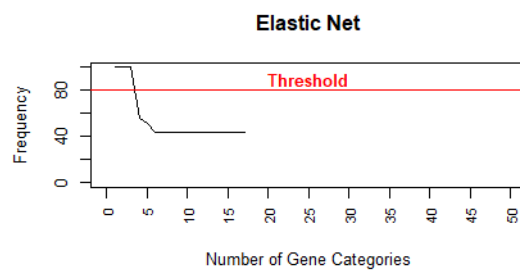
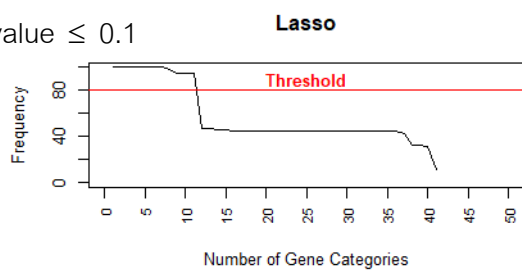
ผลลัพธ์จากการใช้วิธี FDR พบจำนวนภววิทยายีนที่สำคัญดังตารางที่ 3-7

ตารางที่ 3-7 แสดงผลลัพธ์ที่ได้จากการทำ FDR ของ Colon Cancer ด้วยค่า  $p$ -value ที่ได้จากเทคนิค SAFE ทั้ง 100 ครั้งและแสดงค่าเฉลี่ยของค่า  $\pi_0$  ที่พบ

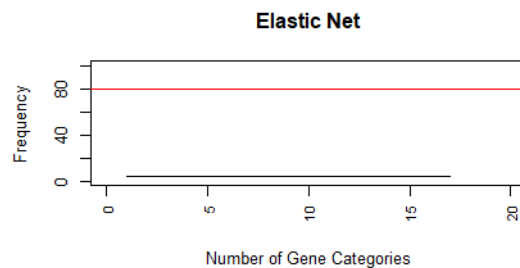
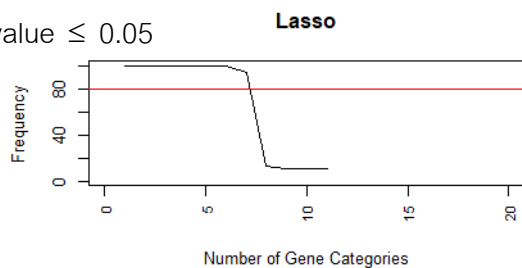
Classification Model	$\pi_0$	Total Number of Gene Category			
		q-value $\leq 0.1$		q-value $\leq 0.05$	
		Total	Freq. $\geq 80$	Total	Freq. $\geq 80$
Lasso	0.74520	41	11	11	7
Elastic Net	0.84373	17	3	17	0
Ridge Regression	1.00000	0	0	0	0
Without Classification Model	1.00000	0	0	0	0

ข้อมูลที่แสดงจากตารางที่ 3-7 สามารถสร้างกราฟความถี่ของแต่ละภววิทยายีนที่พบได้ดังภาพที่ 3-7 และแสดงตัวอย่างความสัมพันธ์ของการแจกแจง  $p$ -value จากเทคนิค SAFE กับค่าเฉลี่ยของ  $\pi_0$  ได้ดังภาพที่ 3-8

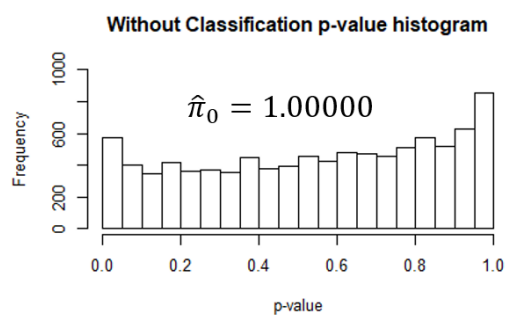
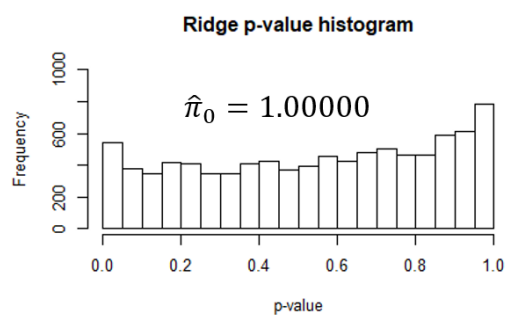
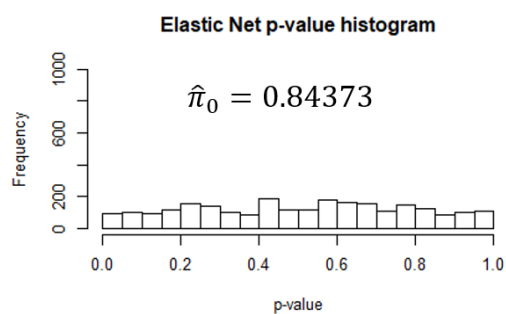
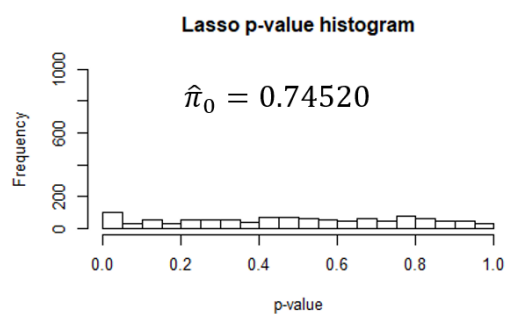
q-value  $\leq 0.1$



q-value  $\leq 0.05$



ภาพที่ 3-7 ความถี่ของแต่ละภาววิทยายีนที่พบจากการทำเทคนิค SAFE ทั้งหมด 100 ครั้งของชุดข้อมูล Colon Cancer



ภาพที่ 3-8 ตัวอย่างความสัมพันธ์ของการแจกแจง p-value จากเทคนิค SAFE กับค่าเฉลี่ยของค่า  $\pi_0$  ของชุดข้อมูล Colon Cancer

### 3.3.3 ผลลัพธ์จากชุดข้อมูล Prostate Cancer

ไม่พบทวิพลาซิมที่มีนัยสำคัญ ค่าเฉลี่ยของค่า  $\pi_0$  สามารถแสดงได้ดังตารางที่ 3-8

ตารางที่ 3-8 ผลลัพธ์ที่ได้จากการทำ FDR ของ Prostate Cancer ด้วยค่า  $p$ -value ที่ได้มาจากเทคนิค SAFE ทั้ง 100 ครั้งและแสดงค่าเฉลี่ยของค่า  $\pi_0$  ที่พบ

Classification Model	$\pi_0$	Total Number of Gene Category			
		$q\text{-value} \leq 0.1$		$q\text{-value} \leq 0.05$	
		Total	Freq. $\geq 80$	Total	Freq. $\geq 80$
Lasso	0.76002	0	0	0	0
Elastic Net	1.00000	0	0	0	0
Ridge Regression	1.00000	0	0	0	0
Without Classification Model	1.00000	0	0	0	0

### 3.5 ตัวอย่างภาววิทยายีนที่พบ

#### 3.5.1 ตัวอย่างภาววิทยายีนที่พบของชุดข้อมูล Leukemia

ผลลัพธ์ของตัวอย่างภาววิทยายีนที่พบในการใช้ FDR ของ  $p$ -value จากเทคนิค SAFE ด้วยข้อมูล Leukemia สามารถแสดงตัวอย่างของภาววิทยายีนที่พบได้ดังตารางที่ 3-9

ตารางที่ 3-9 ตัวอย่างภาววิทยายีนที่พบของ Leukemia ด้วยวิธี Lasso ที่  $q$ -value  $\leq 0.1$

No.	GOID	GO Term	Category	Frequency
1	GO:0003674	negative regulation of transcription from RNA polymerase II promoter	BP	100
2	GO:0005575	MAPK cascade	BP	100
3	GO:0005623	microtubule cytoskeleton organization	BP	100
4	GO:0008150	nuclear chromosome	CC	100
5	GO:0009987	response to reactive oxygen species	BP	100
6	GO:0044464	lytic vacuole	CC	100
7	GO:0050878	chromosome, telomeric region	CC	100
8	GO:0031594	nuclear chromosome, telomeric region	CC	100
9	GO:0043230	nuclear chromatin	CC	100
10	GO:0070062	RNA polymerase II distal enhancer sequence-specific DNA binding	MF	100

### 3.5.2 ตัวอย่างภาววิทยายีนที่พบของชุดข้อมูล Colon Cancer

ผลลัพธ์ของตัวอย่างภาววิทยายีนที่พบในการใช้ FDR ของ  $p$ -value จากเทคนิค SAFE ด้วยชุดข้อมูล Colon Cancer สามารถแสดงตัวอย่างของภาววิทยายีนที่พบได้ดังตารางที่ 3-10

ตารางที่ 3-10 ตัวอย่างภาววิทยายีนที่พบของ Colon Cancer ด้วยวิธี Lasso ที่  $q$ -value  $\leq 0.1$

No.	GOID	GO Term	Category	Frequency
1	GO:0003674	Molecular function	MF	100
2	GO:0005575	Cellular component	CC	100
3	GO:0005623	Cell	CC	100
4	GO:0008150	Biological process	BP	100
5	GO:0009987	cellular process	BP	100
6	GO:0044464	cell part	CC	100
7	GO:0050878	regulation of body fluid levels	BP	100
8	GO:0031594	neuromuscular junction	CC	98
9	GO:0043230	extracellular organelle	CC	95
10	GO:0070062	extracellular exosome	CC	95

### 3.6 เวลาที่ใช้ในแต่ละการทดลอง

ทำการทำการทดลองบน Google Cloud โดย Google Cloud มีข้อกำหนด (Specification) ดังนี้

- CPU 2.3 GHz Intel Xeon E5 v3 (Haswell) 8 vCPUs
- RAM 30 GB memory

#### 3.6.1 เวลาเฉลี่ยที่ใช้ในการทำบัพทสแตร์ป

เวลาเฉลี่ยที่ใช้ในการทำบัพทสแตร์ป ของแต่ละชุดข้อมูลทั้ง 200 ครั้ง เป็นไปตามตารางที่

3-11

**ตารางที่ 3-11** เวลาเฉลี่ยที่ใช้ในการทำบัพทสแตร์ปแต่ละรอบในหน่วยวินาที โดยค่าที่อยู่ในวงเล็บคือ ส่วนเบี่ยงเบนมาตรฐาน

ชุดข้อมูล	Lasso	Elastic Net	Ridge Regression
Colon	0.0980 (0.0786)	0.0958 (0.0346)	0.3442 (0.0956)
Leukemia	0.1294 (0.0994)	0.1090 (0.0431)	0.6057 (0.0952)
Prostate	0.1820 (0.0204)	0.1814 (0.0108)	1.009 (0.2345)

### 3.6.2 เวลาเฉลี่ยที่ใช้ในการทำ SAFE

เวลาเฉลี่ยที่ใช้ในการทำ SAFE ของแต่ละชุดข้อมูลทั้ง 100 ครั้ง เป็นไปตามตารางที่ 3-12

**ตารางที่ 3-12** เวลาเฉลี่ยในการทำ SAFE แต่ละรอบในหน่วยวินาที โดยค่าที่อยู่ในวงเล็บคือส่วนเบี่ยงเบนมาตรฐาน

ชุดข้อมูล	Lasso	Elastic Net	Ridge Regression	Without Classification
Colon	152.8096 (17.7996)	188.7180 (21.4391)	1062.5888 (77.6274)	1110.2798 (91.3757)
Leukemia	40.97038 (3.2008)	43.34656 (3.3340)	482.1301 (5.8017)	487.1236 (4.2220)
Prostate	43.87012 (12.6340)	112.5146 (13.1528)	482.1301 (85.5645)	487.1236 (44.6573)

### 3.6.3 เวลาเฉลี่ยที่ใช้ในการทำ FDR

เวลาเฉลี่ยที่ใช้ในการทำ FDR ของแต่ละชุดข้อมูลทั้ง 100 ครั้ง เป็นไปตามตารางที่ 3-13

**ตารางที่ 3-13** เวลาเฉลี่ยในการทำ FDR แต่ละรอบในหน่วยวินาที โดยค่าที่อยู่ในวงเล็บคือส่วนเบี่ยงเบนมาตรฐาน

ชุดข้อมูล	Lasso	Elastic Net	Ridge Regression	Without Classification
Colon	0.0053 (0.0015)	0.0077 (0.0152)	0.022 (0.0815)	0.014 (0.010)
Leukemia	0.0057 (0.0011)	0.0075 (0.0116)	0.015 (0.0019)	0.016 (0.0089)
Prostate	0.0029 (0.0029)	0.0046 (0.0033)	0.01006 (0.0019)	0.0103 (0.0021)



## บทที่ 4

### สรุปผลการวิจัย

ปัญหานี้เสนอการวิเคราะห์แบบกรณีศึกษาข้อมูลไมโครอะเรย์ หลังผ่านการทำการจำแนกในแต่ละวิธีการทำให้เป็นปรกติรวมทั้งที่ไม่ผ่านการจำแนก

ปัญหานี้ใช้วิธีเทคนิคประเภทฝังตัวในการคัดเลือกลักษณะประจำคือ Lasso, Elastic Net และ Ridge Regression ซึ่งผลลัพธ์ที่พบคือจำนวนของยีนจะลดลงจากเดิมในแต่ละวิธีการทำให้เป็นปรกติโดยที่จำนวนยีนที่พบเรียงจากน้อยไปมากคือ Lasso, Elastic Net และ Ridge Regression นอกจากนี้จำนวนยีนที่พบในแต่ละวิธีการทำให้เป็นปรกติก็มีค่าต่างกัน

จากการทดลองทั้งหมด 100 ครั้งของเทคนิค SAFE พบว่าปัจจัยสำคัญที่ทำให้จำนวนยีนที่พบในแต่ละวิธีการทำให้เป็นปรกติมีความแตกต่างกันคือค่า  $\pi_0$  ที่ได้มาจากการทำ FDR ซึ่งจากผลลัพธ์จะพบว่าค่า  $\pi_0$  มีผลสำคัญต่อจำนวนของยีนที่พบ โดยเมื่อค่า  $\pi_0$  เข้าใกล้ 1 จะยังทำให้จำนวนยีนที่มีนัยสำคัญลดลงจนถึงไม่มีเลย

## เอกสารอ้างอิง

1. Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, 96(12), 6745–6750.
2. Barry, W. T., Nobel, A. B., and Wright, F. A. (2005). Significance analysis of functional categories in gene expression studies: A structured permutation approach. *Bioinformatics*, 21(9), 1943-1949.
3. Chang, K.-H., Kwon, Y. K., & Parker, D. S. (2007). Finding minimal sets of informative genes in microarray data. *Lecture Notes in Bioinformatics*, 4463, 227–236.
4. Dettling, M. (2004). BagBoosting for tumor classification with gene expression data. *Bioinformatics*, 20(18), 3583–3593.
5. Dudoit, S., Fridlyand, J., & Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457), 77–87.
6. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., & Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439), 531–537.
7. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2<sup>nd</sup> ed.). New York, NY: Springer.
8. Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R., & Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2), 203–209.

**เอกสารอ้างอิง (ต่อ)**

9. Storey, J. D., Taylor, J. E., & Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 66(1), 187–205.
10. Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16), 9440–9445.

### ประวัติผู้แต่ง

ปรินิพนธ์เรื่อง : การวิเคราะห์บรรณทัศน์หลังการจำแนกข้อมูลไมโครอะเรย์  
 สาขาวิชา : วิศวกรรมคอมพิวเตอร์  
 ภาควิชา : วิศวกรรมไฟฟ้าและคอมพิวเตอร์  
 คณะ : วิศวกรรมศาสตร์  
 ชื่อ : นายรัฐพล หลิน

### ประวัติ

เกิดเมื่อวันที่ 4 เมษายน พ.ศ. 2538 อยู่บ้านเลข 49/133 หมู่บ้านสัมมากร ถนนนิมิตใหม่ เขต  
 คลองสามวา แขวงสามวาตะวันออก จังหวัดกรุงเทพมหานคร สำเร็จการศึกษาจากโรงเรียนเตรียม  
 อุดมศึกษาน้อมเกล้า จังหวัดกรุงเทพมหานคร สาขาวิทยาศาสตร์-คณิตศาสตร์ ปีการศึกษา 2556 และ  
 สำเร็จการศึกษาในระดับปริญญาตรี สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมไฟฟ้าและ  
 คอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ปี  
 การศึกษา 2560