การออกแบบและพัฒนาเครื่องมือการจัดกลุ่มยีนโดยใช้ข้อมูลการแสดงออกของยีน และข้อมูลของส่วนหน้าของยีน*

A Design and Development of Clustering Tool Using Gene Expression and Upstream Region Data

วัฒน์วิบูลย์ แพรมงคล[1], อนันต์ ทองทา[2], ดวงดาว วิชาดากุล[3]

[1]หลักสูตรชีวสารสนเทศ, คณะเทคโนโลยีสารสนเทศและคณะเทคโนโลยีชีวภาพและเทคโนโลยี มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

[2]คณะเทคโนโลยีชีวภาพและเทคโนโลยี มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

[3]หน่วยชีวสารสนเทศ ศูนย์พันธุวิศกรรมและเทคโนโลยีชีวภาพแห่งชาติ

บทคัดย่อ – เทคโนโลยีไมโครอาเรย์ได้มีบทบาทในงานวิจัยทางด้านเทคโนโลยีชีวภาพในปัจจุบันเป็นอย่างมาก ไม่ว่าจะเป็นในด้านการศึกษาหน้าที่ของยีน (Gene function) การศึกษาโครงข่ายการควบคุมการแสดงออกของยีน (Gene regulatory networks) และการค้นหายาใหม่ เป็นต้น ข้อมูลที่ได้จากไมโครอาเรย์นั้นเป็นข้อมูลของการแสดงออกของยีน (Gene expression data) ทั้งหมดภายในเซลล์ ซึ่งข้อมูลดังกล่าวจะสามารถนำเราไปสู่ความเข้าใจการทำงานระดับของเซลล์ภายในสิ่งแวดล้อมที่นักชีววิทยาต้องการศึกษา โดยการจัดการข้อมูลของการแสดงออกของยีนนั้นโดยทั่วไปแล้วจะทำการจัดกลุ่มของยีนเข้าด้วยกันโดยดูเพียงความคล้ายคลึงของรูปแบบของการแสดงออกของยีนเท่านั้น ซึ่งเครื่องมือที่ใช้ในการจัดกลุ่มข้อมูลของการแสดงออกของยีนในปัจจุบันล้วนแต่ใช้เพียงข้อมูลของการแสดงออกของยีน ซึ่งการใช้ชุดของข้อมูลเพียงรูปแบบของการแสดงออกของยีนนั้นอาจไม่เพียงพอ เนื่องจากข้อมูลการแสดงออกของยีนมักเกิดการรบกวนจากสิ่งแวดล้อม และการแสดงออกของยีนที่มีรูปแบบคล้ายคลึงกันก็เป็นตัวบ่งชี้ที่ไม่ชัดเจนที่จะสรุปได้ว่ายีนเหล่านั้นควรจะอยู่ในกลุ่มเดียวกัน ดังนั้นหากเราทำการเพิ่มข้อมูลอื่นๆ ลงไปร่วมกับข้อมูลการแสดงออกของยีน ก็น่าจะทำให้การจัดกลุ่มของยีนทำได้ถูกต้องมากขึ้น โดยข้อมูลที่นำมาเพิ่มนั้นก็คือข้อมูลของส่วนหน้าของยีน (Upstream regions data) ซึ่งเป็นส่วนควบคุมการแสดงออกของยีน โดยเชื่อว่าหากยีนที่มีส่วนหน้าของยีนที่คล้ายคลึงกัน ก็น่าที่จะถูกแสดงออกในระบบของเมตาบอลิซึมเดียวกัน การจัดกลุ่มแบบนาอีฟเบย์ (Naïve Bayes- classification) ถูกนำมาใช้ในการจัดกลุ่มของ 2 ชุดข้อมูล และนำผลจากการจัดกลุ่มของทั้ง 2 ชุดข้อมูลมาซ้อนทับกัน (Overlapping- algorithm) โดยจะทำให้ผลหลังการทับซ้อนของข้อมูลมีประสิทธิภาพมากขึ้น

คำสำคัญ - การจัดกลุ่มของยีน (Gene clustering), การจัดกลุ่มแบบนาอีฟเบย์ (Naïve Bayes classification), การแสดงออกของยีน (Gene expression), บริเวณส่วนหน้าของยีน (Upstream regions)

*งานวิจัยชิ้นนี้ได้รับทุนสนับสนุนจากศูนย์พันธุวิศกรรมและเทคโนโลยีชีวภาพแห่งชาติ

1. บทนำ

ปัจจุบันงานวิจัยในด้านโพสจีโนมมิก (Post-genomic) กำลังเป็นที่สนใจแก่นักชีววิทยาระดับโมเลกุล โดยเครื่องมือหนึ่งที่มีความสำคัญต่อการศึกษางานวิจัยด้านนี้คือ ไมโครอาเรย์ ดีเอ็นเอ ซึ่งเป็นเครื่องมือที่ทำให้นักวิจัยสามารถศึกษาการแสดงออกของยีนจำนวนหลายพันยีนโดยทำการทดลองเพียงครั้งเดียว จึงทำให้การศึกษางานวิจัยด้านโพสจีโนมมิกทำได้รวดเร็ว และได้ข้อมูลที่ถูกต้องมากขึ้น วิธีการที่จะจัดการกับข้อมูลการแสดงออกของยีนโดยทั่วไปนิยมที่จะจัดกลุ่มของยีนเข้าด้วยกัน โดยการหาความคล้ายคลึงกันของรูปแบบการแสดงออกของยีน (Pattern of gene expression) [1] ซึ่งเครื่องมือที่ใช้ในการจัดกลุ่มของยีนในปัจจุบันมีอยู่มากมาย เช่น Cluster (Eisen), SAM (Stanford), GeneCluster (MIT) และ J-Express (Molmine) [1,2,3,4] แต่เครื่องมือที่ใช้ในการจัดกลุ่มยีนในปัจจุบัน จะทำการพิจารณาข้อมูลการแสดงออกของยีนเพียงข้อมูลเดียว ด้วยการดูรูปแบบการแสดงออกของยีนเป็นหลัก หากยีนใดมีรูปแบบของการแสดงออกของยีนที่คล้ายคลึงกัน ก็จะจัดกลุ่มยีนเหล่านั้นเข้าด้วยกัน อย่างไรก็ตามการพิจารณาข้อมูลการแสดงออกของยีนเพียงอย่างเดียวนั้น อาจไม่เพียงพอที่จะแสดงว่ายีนที่ถูกจัดกลุ่มเข้าด้วยกันมีความสัมพันธ์ไปในทางเดียวกันหรือไม่ เนื่องจากข้อมูลการแสดงออกของยีนจากไมโครอาเรย์มักเกิดสิ่งรบกวนต่อตัวข้อมูลได้มาก และการที่ยีนมีรูปแบบของการแสดงออกของยีนที่เหมือนกันก็ไม่ได้เป็นตัวบ่งชี้ที่เด่นชัดว่ายีนเหล่านั้นควรจะถูกจัดอยู่ในกลุ่มเดียวกัน [5] ด้วยข้อจำกัดของข้อมูลดังกล่าว เราจึงเสนอที่จะเพิ่มข้อมูลทางชีววิทยาอื่นๆ เข้าไปร่วมกับข้อมูลของการแสดงออกของยีนเพื่อช่วยในการจัดกลุ่มยีน โดยหวังว่าการเพิ่มข้อมูลบางอย่างลงไปร่วมนั้นน่าจะสามารถเพิ่มประสิทธิภาพในการจัดกลุ่มของยีนได้ดีขึ้น ชุดข้อมูลทางชีววิทยาที่เราเลือกเข้ามาร่วมด้วยในการจัดกลุ่มยีนนั้นก็คือ ชุดข้อมูลของส่วนหน้าของยีน ด้วยเหตุผลที่ว่า ในกระบวนการทางชีววิทยาระดับเซลล์ส่วนใหญ่นั้นจะถูกควบคุมพฤติกรรมและการแสดงออกในระดับการถอดรหัสพันธุกรรมของดีเอ็นเอ (Transcription) ด้วยการที่ Transcription factor จะเข้าไปจับกับตำแหน่งของดีเอ็นเอ(Motif) ที่จำเพาะกับโปรตีนนั้นๆ ซึ่งอยู่บนตำแหน่งที่ไม่สามารถถอดรหัสพันธุกรรม (Non-coding region) โดยตำแหน่งที่ไม่สามารถถอดรหัสพันธุกรรมจะเป็นได้ทั้งส่วนหน้าของยีน ส่วนหลังของยีน และแฝงอยู่ในส่วนที่สามารถถอดรหัสพันธุกรรมของยีนคนละชนิดซึ่งพบในยูคาริโอต ในงานวิจัยนี้เลือกใช้ส่วนหน้ายีนเนื่องจาก เป็นส่วนที่พบได้ทั้งในโปรคาริโอต และยูคาริโอต ที่สำคัญในส่วนหน้ายีนนี้เป็นส่วนที่มีบทบาทอย่างมากในการควบคุมการแสดงออกของยีน ดังนั้นเราจึงคาดว่ายีนที่ถูกควบคุมการทำงานร่วมกันนั้นน่าจะมีรูปแบบของส่วนหน้าของยีนที่คล้ายคลึงกัน เพราะฉะนั้นชุดของข้อมูลของส่วนหน้ายีนจึงเหมาะสมที่จะถูกเลือกเพื่อนำมาช่วยในการจัดกลุ่มยีน

ผลของการวิจัยเพื่อเพิ่มประสิทธิภาพการจัดกลุ่มยีนโดยใช้ข้อมูลร่วม เราได้พัฒนาโปรแกรมที่มีชื่อว่า "ซู-คลัสเตอร์" (XU-Cluster) ที่สามารถจัดกลุ่มยีนโดยใช้ข้อมูลการแสดงออกของยีน และ ข้อมูลส่วนหน้าของยีน โดยซู-คลัสเตอร์ จะทำการเรียกโปรแกรม AutoClass [6] เพื่อจัดกลุ่มยีนสำหรับแต่ละชุดข้อมูล หลังจากนั้นซู-คลัสเตอร์จะนำกลุ่มยีนที่ผ่านการจัดกลุ่มโดย AutoClass มาหายีนร่วม (ยีนที่พบทั้งในกลุ่มยีนที่ถูกจัดกลุ่มด้วยข้อมูลการแสดงออกของยีนและข้อมูลส่วนหน้าของยีน) ซึ่งกลุ่มยีนใหม่จากขั้นตอนการหายีนร่วมนั้น เราเชื่อว่าจะเป็นกลุ่มยีนที่มีความละเอียด และมีความถูกต้องมากขึ้น นอกจากนี้ซู-คลัสเตอร์ได้จัดเตรียมเครื่องมือที่เกี่ยวข้องกับการแสดงผลเป็นภาพ ซึ่งประกอบด้วย การแสดงผลในรูปของสี และการแสดงผลในรูปของกราฟ เพื่อให้ผู้ใช้มีความสะดวก และง่ายต่อการแปลผลลัพธ์มากยิ่งขึ้น

2. วิธีการศึกษา

2.1 ข้อมูลที่ใช้ในการทดลอง

งานวิจัยนี้ใช้ชุดข้อมูลทั้งหมด 3 ชุด โดยมีรายละเอียดดังต่อไปนี้

2.1.1 ชุดข้อมูลของการแสดงออกของยีนจากไมโครอาเรย์ (Gene expression data)

ชุดข้อมูลนี้นำมาจากฐานข้อมูลไมโครอาเรย์ของสแตนฟอร์ด (genome-www5.stanford.edu) ในงานวิจัยนี้ได้เลือกชุดข้อมูลที่ทำการศึกษาเกี่ยวกับการควบคุมการทำงานของส่วนหน้าของยีนในวัฏจักรเซลล์ และการเจริญของเส้นใยเทียมในยีสต์ (Two yeast forkhead genes regulate the cell cycle and- pseudohyphal growth) [7] โดยชุดข้อมูลนี้เป็นการทดลองแบบอนุกรมเวลา โดยมียีนทั้งหมด 5,171 ยีน และทำการทดลองทั้งหมด 26 การทดลอง (fkh1,2 alpha factor 0 – 210 นาที : 13 การทดลอง, cell cycle-alpha factor, fkh1, fkh2 0 - 180 นาที : 13 การทดลอง) ข้อมูลที่ได้จากฐานข้อมูลจะถูกนำมาจัดให้อยู่ในสภาวะบรรทัดฐานเดียวกัน (Normalization) ด้วย $\log_2$ และทำการกรองยีนที่มีข้อมูลที่ขาดหายไป (Missing data) และยีนที่ไม่มีนัยสำคัญต่อการวิเคราะห์ ด้วยชุดเครื่องมือทางชีวสารสนเทศใน Matlab เวอร์ชั่น 7 [8] โดยข้อมูลหลังจากการกรองแล้วเหลือ 731 ยีน ซึ่งจะถูกจัดเก็บในรูปของไฟล์ข้อความ (Text file)

2.1.2 ชุดข้อมูลของส่วนหน้าของยีน (Upstream sequence data)

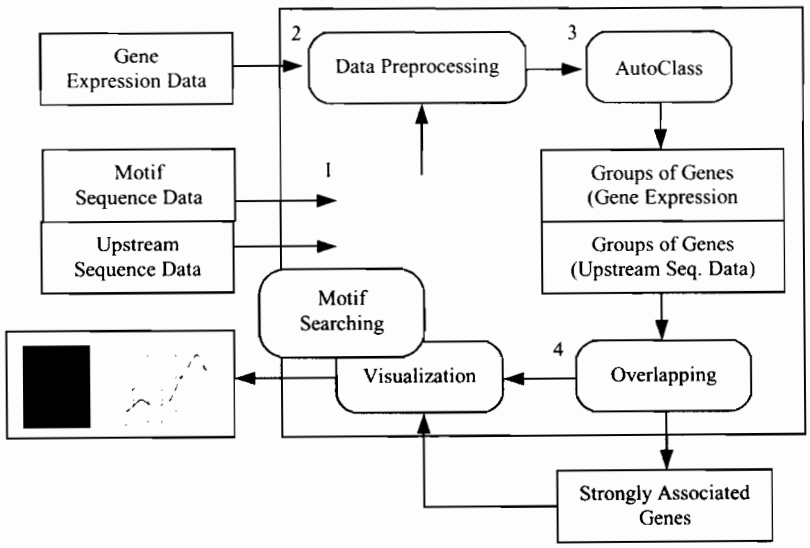
ข้อมูลของส่วนหน้าของยีนนี้ได้มาจาก Regulatory Sequence Analysis Tools [9] ซึ่งเป็นเครื่องมือที่ใช้ในการดึงของข้อมูลของส่วนหน้าของยีน และงานวิจัยนี้ได้เลือกส่วนของหน้ายีนที่ตั้งแต่ –800 ถึง –1 เป็นจำนวน 800 นิวคลีโอไทด์ จากนั้นทำการจัดเก็บข้อมูลในรูปของไฟล์ข้อความ (Text file) ในรูปแบบของ FASTA

2.1.3 ชุดข้อมูลของโมทีฟ (Motif sequence data)

ชุดข้อมูลโมทีฟในงานวิจัยนี้นำมาจาก 2 ฐานข้อมูล คือ ฐานข้อมูลของส่วนส่งเสริมการทำงานของยีน (Promoter region) ในยีสต์ (The Promoter Database of *Saccharomyces cerevisiae* : SCPD[10]) และฐานข้อมูลของโปรตีนที่จับบนส่วนที่ไม่สามารถถอดรหัสพันธุกรรมได้ (TRANSFAC® 6.0 – Public [11]) ในงานวิจัยนี้ผู้วิจัยได้ทำการเลือกชุดของโมทีฟที่เกี่ยวข้องกับการทำงานของวัฏจักรเซลล์ เนื่องจากวัฏจักรเซลล์เป็นกระบวนการทำงานที่สำคัญของสิ่งมีชีวิตจึงน่าจะเป็นที่สนใจของนักวิทยาศาสตร์ นอกจากนี้ชุดของโมทีฟที่เกี่ยวข้องกับวัฏจักรเซลล์มีผู้ทำการทดลองมากมาย เราจึงสามารถเลือกชุดโมทีฟได้มากกว่าเมื่อเปรียบเทียบกับชุดของโมทีฟชุดอื่นๆ สำหรับความยาวของโมทีฟนั้นผู้ทำวิจัยได้เลือกความยาวที่อยู่ระหว่าง 6 – 10 เบส โดยไฟล์ของข้อมูลนี้มีลักษณะเป็นไฟล์ข้อความที่ประกอบด้วยชื่อโมทีฟ รูปแบบของโมทีฟ และตำแหน่งที่พบบนส่วนหน้าของยีน

2.2 โครงสร้างการทำงานของซู-คลัสเตอร์

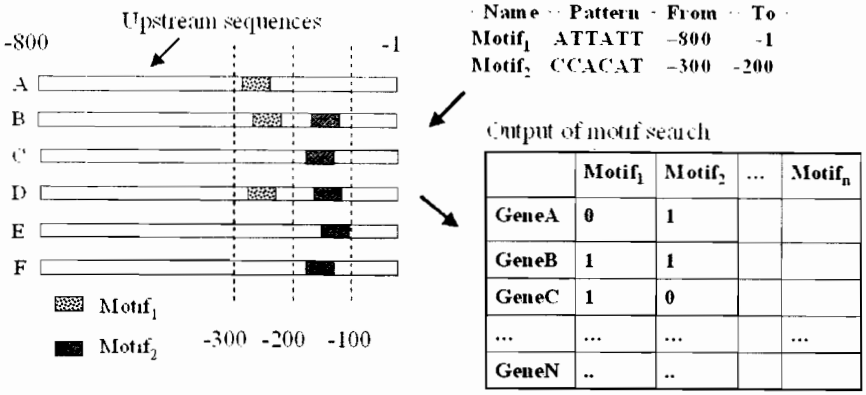
ซู-คลัสเตอร์ประกอบด้วย 5 ส่วนหลัก คือ 1) ส่วนการค้นหาโมทีฟบนส่วนหน้าของยีน 2) ส่วนการเตรียมข้อมูลเพื่อการประมวลผล 3) ส่วนการประมวลผลโดยโปรแกรม AutoClass 4) ส่วนการหายีนร่วม และ 5) ส่วนแสดงผลด้วยภาพ ดังรูปที่ 1



รูปที่ 1. แสดงโครงสร้างการทำงานของซู-คลัสเตอร์

2.2.1 ส่วนการค้นหาโมทีฟบนส่วนหน้าของยีน

ข้อมูลที่ใช้ในขั้นตอนนี้คือ ข้อมูลโมทีฟ และข้อมูลส่วนหน้าของยีน โดยที่ไฟล์ของชุดโมทีฟจะถูกนำไปตรวจหาเพื่อดูว่าพบรูปแบบของโมทีฟเหล่านี้บนส่วนหน้าของยีนหรือไม่ หากพบโปรแกรมจะทำการให้คะแนนเป็น 1 แต่หากไม่พบรูปแบบของโมทีฟนั้นๆ บนส่วนหน้าของยีนโปรแกรมจะให้คะแนนเป็น 0 โดยตำแหน่งของโมทีฟสามารถระบุตำแหน่งที่ชัดเจนหากเราทราบข้อมูลที่แน่นอนของโมทีฟนั้นๆ และหากไม่ทราบตำแหน่งที่ชัดเจนเราสามารถระบุตำแหน่งของข้อมูลอย่างกว้างๆ ได้ ขั้นตอนการค้นหาโมทีฟบนส่วนหน้าของยีนแสดงดังรูปที่ 2



รูปที่ 2. แสดงส่วนการค้นหาโมทีฟบนส่วนหน้าของยีน

ผลลัพธ์จากส่วนค้นหาโมทีฟ และ ชุดของข้อมูลการแสดงออกของยีน จะถูกส่งไปยังส่วนของการเตรียมข้อมูลก่อนการประมวลผล โดยส่วนนี้จะทำการสร้างไฟล์ที่อยู่ในรูปแบบที่ โปรแกรม AutoClass [6] ต้องการ ซึ่งไฟล์ดังกล่าวประกอบด้วย ไฟล์ข้อมูล (Data file), ไฟล์ส่วนหัวของข้อมูล (Header file), โมเดลไฟล์ (Model file), ไฟล์พารามิเตอร์ของการจัดกลุ่ม (Search parameter file) และ ไฟล์พารามิเตอร์ของการสร้างรายงาน (Report parameter file) ซึ่งสามารถดูรายละเอียดของไฟล์ข้อมูลได้ใน

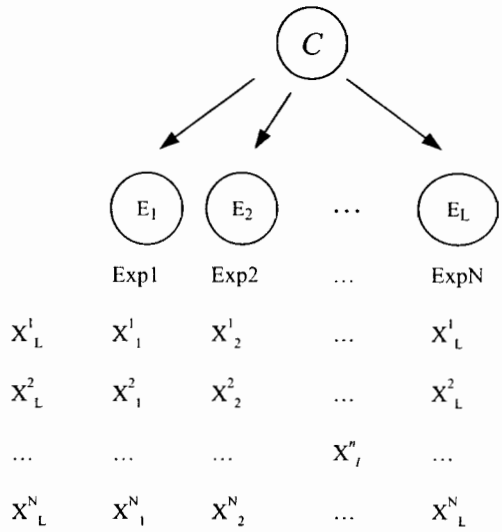
(http://ic.arc.nasa.gov/ic/projects/bayes-group/autoclass/)

2.2.3 ส่วนการประมวลผลโดยโปรแกรม AutoClass

หลังจากได้ไฟล์จากขั้นเตรียมข้อมูล ไฟล์เหล่านั้นจะถูกส่งไปประมวลยัง AutoClass โดย AutoClass จะทำการจัดกลุ่มของชุดข้อมูล และกำหนดจำนวนกลุ่มที่เหมาะสมแก่ชุดข้อมูลทั้งสอง โดยใช้หลักการของการจัดกลุ่มแบบไม่ต้องผ่านการรู้จำ (Unsupervised cluster) อัลกอริทึมที่ถูกใช้ก็คือ การจัดกลุ่มแบบนาอีฟ เบย์

2.2.3.1 โมเดล นาอีฟ เบย์ สำหรับข้อมูลการแสดงออกของยีน (Naïve Bayes Model For Gene Expression Data)

กราฟโมเดลถูกนิยมใช้เป็นโครงสร้างของระบบ Stochastic ซึ่งมีความสามารถในการเรียนรู้ความสัมพันธ์ระหว่างตัวแปรที่มีความซับซ้อนได้เป็นอย่างดี [12] นาอีฟ เบย์ โมเดล เป็นกราฟโมเดลพื้นฐานชนิดหนึ่ง ซึ่งในงานวิจัยนี้จะนำโมเดลชนิดนี้มาประยุกต์ใช้ในการจัดกลุ่มชุดข้อมูลการแสดงออกของยีนในไมโครอาเรย์นั้นถูกแสดงในรูป 3



รูปที่ 3. แสดงโมเดลนาอีฟ เบย์ สำหรับข้อมูลการแสดงออกของยีน [12]

จากรูป C เป็น Hidden classification ในแต่ละ E_l แสดงถึงระดับของการแสดงออกของยีนในการทดลองที่ *l* และ X^n_l แสดงถึงค่าระดับการแสดงออกของยีนที่ *n* ในการทดลองที่ *l* โมเดลนาอีฟ เบย์ ที่เราใช้ในการจัดกลุ่มยีนอยู่ในสมการที่ 1

$$P(X_L^n \mid C_k) = P(C_k) \prod_{l=1}^{L} P(x_l^n \mid C_{k,l}) \qquad (1)$$

ในที่นี้ AutoClass จะหา Posterior probability ที่สูงที่สุด (Maximum of posterior probability) โดยที่ $P(C_k)$ เป็น Prior probability ของกลุ่มที่ k และ $\sum_{k=1}^{K} P(C_k) = 1$ และ $0 \le P(C_k) \le 1$ เราใช้ Conditional probability distribution (CPD) $P(x_l^n \mid C_{k,l})$ โดยสมมติให้มีการ

กระจายของข้อมูลเป็นการกระจายข้อมูลแบบปกติ (Gaussian distribution)

$$P(x_l^n \mid C_{k,l}) = \frac{1}{\sqrt{2\pi\sigma_{k,l}}} \exp - \frac{1}{2} \left(\frac{x_l^n - \mu_{k,l}}{\sigma_{k,l}} \right)^2 \quad (2)$$

2.2.3.2 โมเดล นาอีฟ เบย์ สำหรับข้อมูลส่วนหน้าของยีน (Naïve Bayes Model For Upstream Sequence Data)

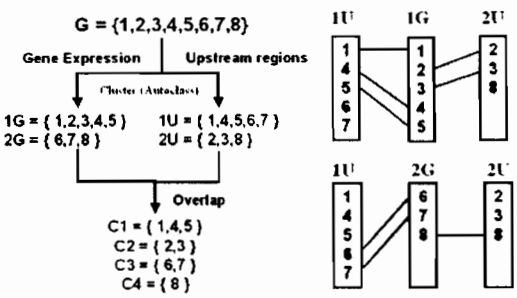
โมเดล นาอีฟ เบย์ ที่ใช้สำหรับส่วนหน้าของยีนนั้น ได้แสดงในสมการที่ 3

$$P(U_M^n \mid C_k) = P(C_k) \prod_{m=1}^{M} P(u_m^n \mid C_{k,m}) \quad (3)$$

โดยที่ u_m^n เป็น Discrete value ที่แสดงถึงการพบหรือไม่พบโมทีฟบนส่วนหน้าของยีน จะสังเกตได้ว่าโมเดลนี้มีลักษณะคล้ายกับโมเดลที่ใช้กับการแสดงออกของยีน แต่จะที่แตกต่างกันคือ การกระจายตัวของข้อมูล โดยการกระจายตัวของข้อมูลส่วนหน้าของยีนเป็นการกระจายตัวแบบไม่ต่อเนื่อง (Discrete distribution) ในขณะที่การกระจายตัวของข้อมูลการแสดงออกของยีนเป็นแบบต่อเนื่อง (Continuous distribution) ดังนั้น $P(u_m^n \mid C_{k,m})$ ที่แสดงถึง CPD ของข้อมูลส่วนหน้าของยีนจะใช้การกระจายแบบ Multinomial distribution

2.2.4 ส่วนการหายีนร่วม (Overlapping)

เมื่อได้กลุ่มของยีนจากการจัดกลุ่มโดย AutoClass โดยใช้ชุดของข้อมูลที่ต่างกันคือ การแสดงออกของยีน และส่วนหน้าของยีน กลุ่มของชุดข้อมูลเหล่านั้นจะถูกนำมาหายีนร่วมระหว่างสองกลุ่มข้อมูล ด้วยการจับคู่ของยีนแล้วดึงส่วนของยีนที่พบทั้งใน 2 กลุ่มข้อมูล ซึ่งเชื่อว่าหลังจากการทำการหายีนร่วมแล้ว ยีนที่ถูกจัดเป็นกลุ่มใหม่น่าที่จะมีความสัมพันธ์มากยิ่งขึ้น แนวความคิดในการหายีนร่วมได้ในรูปที่ 4



รูปที่ 4. แสดงส่วนการหายีนร่วม (Overlapping)

จากรูปให้ G เป็นชุดของยีนทั้งหมดที่เราต้องการจัดกลุ่ม ประกอบด้วย 1, 2, 3, 4, 5, 6, 7 และ 8 หลังจากนั้นยีนชุดนี้จะถูกนำไปจัดกลุ่มด้วยโปรแกรม AutoClass โดยใช้ข้อมูลของการแสดงออกของยีน และ ชุดข้อมูลส่วนหน้าของยีน จากการจัดกลุ่มจะพบว่าชุดข้อมูลการแสดงออกของยีนถูกแบ่งออกเป็น 2 กลุ่มคือ 1G = {1,2,3,4,5} และ 2G = {6,7,8} และชุดข้อมูลส่วนหน้าของยีนถูกแบ่งออกเป็น 2 กลุ่มเช่นกัน คือ 1U = {1,4,5,6, 7} และ 2U = {2,3,8} ขั้นตอนต่อมาเป็นส่วนที่จะทำการหายีนร่วมระหว่างยีนในกลุ่มข้อมูลการแสดงออกของยีน และกลุ่มข้อมูลส่วนหน้าของยีน กลุ่มใหม่ที่เป็นผลลัพท์จากการหายีนร่วมก็คือ C1 = {1,4,5}, C2 = {2,3}, C3 = {6,7} และ C4 = {8}

2.2.5 ส่วนการแสดงผลด้วยภาพ (Visualization)

เพื่อให้ง่ายต่อการตีความหมายของผลลัพธ์ ชู-คลัสเตอร์ได้เตรียมฟังก์ชันของการแสดงผลในรูปของสี และกราฟ

2.2.5.1 การแสดงผลในรูปสี

การศึกษาการแสดงออกของยีนนั้น โดยทั่วไปเป็นการศึกษาถึงการทำงานของยีนว่ายีนใด มีการแสดงออกของยีนเพิ่มขึ้น (Up-regulated) หรือลดลง (Down-regulated) อย่างไร เมื่อเทียบกับการทดลองอ้างอิง โปรแกรมจะกำหนดให้สีเพื่อแสดงการแสดงออกของยีน กล่าวคือ สีแดง แสดงถึงการที่ยีนมีการแสดงออกของยีนมากกว่าตัวอ้างอิง สีเขียว แสดงถึงการที่ยีนมีการแสดงออกของยีนน้อยกว่าตัวอ้างอิง ระดับของสีจะมี 5-ระดับด้วยกัน เพื่อระบุการแสดงออกของยีนได้ถูกต้องและชัดเจนมากขึ้น

2.2.5.2 การแสดงผลในรูปของกราฟ

การแสดงผลโดยกราฟนั้นจะนำข้อมูลของความเข้มข้นของเอ็มอาร์เอ็นเอของแต่ละยีนในแต่ละการทดลองมาพล็อทลงบนกราฟ โดยการแสดงผลในรูปกราฟนั้นทำให้ผู้ใช้งานสามารถดูระดับของการเปลี่ยนแปลงความเข้มข้นของเอ็มอาร์เอ็นเอเพื่อนำมาวิเคราะห์ผลได้ละเอียด และแม่นยำยิ่งขึ้น

3. ผลการทดลอง

จากการกลุ่มจัดยีนโดยใช้การแสดงออกของยีนเป็นเกณฑ์ในการจัดกลุ่มเพียงอย่างเดียว ได้ผลการทดลองดังนี้ กลุ่ม 0 = 382 ยีน, กลุ่ม 1 = 150 ยีน, กลุ่ม 2 =142 ยีน, กลุ่ม 3 = 20 ยีน, กลุ่ม 4 =18 ยีน, กลุ่ม 5 = 10 ยีน และ กลุ่ม 6 = 9 ยีน ส่วนการจัดกลุ่มยีนโดยใช้ส่วนหน้าของยีนเป็นเกณฑ์เพียงอย่างเดียว ได้ผลการทดลองดังนี้ กลุ่ม 0 = 636 ยีน และ

กลุ่ม 1= 45 ยีน สังเกตได้ว่าผลรวมของจำนวนยีนจะไม่เท่ากันในทั้งสองกลุ่ม เนื่องจากยีนบางตัวยังไม่มีข้อมูลส่วนหน้าของยีน ดังนั้นจำนวนยีนที่ใช้ในการจัดกลุ่มโดยใช้ส่วนหน้าของยีนจึงมีจำนวนน้อยกว่า ผลลัพธ์ของการจัดกลุ่มยีนที่ใช้ข้อมูลการแสดงออกของยีน และข้อมูลส่วนหน้าของยีนจะถูกนำมาหายีนร่วมโดยในที่นี้จะกล่าวในรายละเอียดเฉพาะกลุ่มยีนร่วม ที่เกิดมาจากกลุ่มของยีนใดๆ ที่ใช้การแสดงออกของยีนเป็นเกณฑ์ กับกลุ่มที่พบโมทีฟที่เราสนใจบนส่วนหน้าของยีน นั้นคือ กลุ่ม 0-1, กลุ่ม1-1, กลุ่ม 2-1, กลุ่ม 3-1 และ กลุ่ม 6-1 ซึ่งกลุ่มของยีนร่วมที่ได้ออกมานั้นจะถูกนำมาหาหน้าที่ของยีน เพื่อให้ทราบว่ายีนใดบ้างเป็นยีนที่เกี่ยวข้องกับวัฏจักรเซลล์ เป็นการทดสอบประสิทธิภาพของเครื่องมือที่เราทำการพัฒนาขึ้น เว็บไซต์ที่เราใช้ในการตรวจสอบหน้าที่ของยีนคือ www.yeastgenome.com หลังจากนั้นเราก็จะนำกลุ่มยีนร่วมที่ได้ มาเทียบกับวารสารต่างๆ ที่ได้มีรายงานถึงยีนที่ทำหน้าที่เกี่ยวข้องกับวัฏจักรเซลล์ โดยวารสารหลักที่เราใช้เป็นวารสารอ้างอิงคือ Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization [13] วารสารฉบับนี้แสดงถึงกลุ่มยีนที่ทำงานเกี่ยวข้องกับวัฏจักรเซลล์ รายละเอียดของยีนในกลุ่มต่างๆ มีดังนี้ กลุ่ม 0-1 ประกอบด้วยยีนทั้งหมด 16 ยีน โดย 12 ยีนได้ถูกกำหนดหน้าที่ของยีนไว้แล้ว แต่อีก 4 ยีนยังไม่ได้ถูกกำหนดหน้าที่ของยีน (YOR131C, YGL050W, YCL005W และ YGL242C) โดยมี 6 ยีนมีหน้าที่เกี่ยวกับวัฏจักรเซลล์ คือ DOT1, SMC2, BIM1, TOP2, SPC110 และ APC4 กลุ่ม1-1 ประกอบด้วยยีนทั้งหมด 9 ยีน โดยมี 6 ยีนมีหน้าที่เกี่ยวกับวัฏจักรเซลล์ คือ RPS4A, FKS1, PXR1, IPL1, HYS2 และ PRO2 กลุ่ม2-1 ประกอบด้วยยีนทั้งหมด 11 ยีน โดยมี 8 ยีนมีหน้าที่เกี่ยวกับวัฏจักรเซลล์ คือ SWI4, TUB4, SMC1, SMC3, PRE10, JEM1, RFA1 และ RHO1 กลุ่ม3-1 ประกอบด้วยยีนทั้งหมด 6 ยีน โดย 2 ยีนยังไม่ถูกกำหนดหน้าที่ของยีน คือ YLR049C และ YDL156W และ มี 4 ยีน มีหน้าที่เกี่ยวกับวัฏจักรเซลล์ คือ PRI2, RNR1, CRH1 และ POL2 กลุ่ม6-1 ประกอบด้วยยีนทั้งหมด 3 ยีน โดยยีนทั้งหมดในกลุ่มนี้เป็นยีนที่มีหน้าที่เกี่ยวกับวัฏจักรเซลล์ คือ SPT21, KCC4 และ HTA2 ดังแสดงในตารางที่ 1

ตารางที่ 1 แสดงผลลัพธ์ของกลุ่มยีนร่วม

กลุ่มยีน	ยีนทั้งหมด	ยีนที่เกี่ยวข้องกับวัฏจักรเซลล์	C/N
0-1	SPA2, SEC31, DOT1, YOR131C, YGL050W, SMC2, MRPS35, BIM1,YCL005W, SAR1, TOP2,SMT3,SPC110,YGL242C, GID8 และ APC4	DOT1, SMC2, BIM1, TOP2, SPC110 และ APC4	6/16
1-1	RPS4A, FKS1, PXR1, IPL1, HYS2, RFC4 NAM2, PRO2 และ MNL1	RPS4A, FKS1, PXR1, IPL1, HYS2 และ PRO2	6/9
2-1	SWI4, TUB4, UNG1, SMC1, SMC3, PRE10, APJ1, JEM1, RFA1, EUG1 และ RHO1	SWI4, TUB4, SMC1, SMC3, PRE10, JEM1, RFA1 และ RHO1	8/11
3-1	YLR049C, PRI2, RNR1, CRH1, POL2 และYDL156W	PRI2, RNR1, CRH1 และ POL2	4/6
6-1	SPT21, KCC4 และ HTA2	SPT21, KCC4 และ HTA2	3/3

C/N = สัดส่วนจำนวนยีนที่เกี่ยวข้องกับการทำงานของวัฏจักรเซลล์กับจำนวนยีนทั้งหมดในแต่ละกลุ่มยีน

4. บทสรุป

การจัดกลุ่มยีนโดยการใช้ข้อมูลของการแสดงออกของยีนร่วมกับข้อมูลส่วนหน้าของยีนสามารถที่จะเพิ่มประสิทธิภาพของการจัดกลุ่มยีนให้มีความละเอียด และมีความถูกต้องมากยิ่งขึ้น โดยเราจะให้ความสำคัญแก่กลุ่มที่หายีนร่วมระหว่างกลุ่มยีนที่พบโมทีฟ กับกลุ่มยีนใดๆ ของการแสดงออกของยีน กล่าวคือ เราจะให้สนใจเฉพาะกลุ่ม 0-1, กลุ่ม1-1, กลุ่ม2-1, กลุ่ม3-1 และ กลุ่ม6-1 ยีนจากกลุ่มเหล่านี้รวมกันทั้งหมดเท่ากับ 45 ยีน ซึ่งจากผลการทดลองพบว่า 60% ของยีนทั้งหมดเกี่ยวข้องกับการควบคุมยีนในวัฏจักรเซลล์ ได้แก่ SMS2, BIM1, SPC110, TOP2, APC4, FKS1, PRO2, IPL1, RPS44, SWI, SMC1, SMC2, TUB4, JEM1, RFA1, POL2, PRI2, RNR1, CRH1, HTA2, KCC4, SPT2, DOT1, PXR1, HYS2, PRE2 และ RHO1 และประมาณ 27% เป็นยีนที่ไม่เกี่ยวข้องกับการควบคุมการทำงานของวัฏจักรเซลล์ ได้แก่ SPA2, SEC31, MRPS35, SAR1, SMT3, GID8, RFC4, NAM2 ,MNL1 ,UNG1 ,APJ1 และ EUG1 อีกประมาณ 13% เป็นยีนที่ยังไม่ถูกกำหนดหน้าที่ ได้แก่ YOR131C, YGL050W, YCL005W, YGL242C, YLR049C และ YDL156W จากผลการทดลองพบว่า เราได้กลุ่มยีนที่เกี่ยวข้องกับการทำงานในวัฏจักรเซลล์ อยู่ 45 ยีน จากทั้งหมด 800 ยีน [13] ซึ่งจำนวนยีนที่ได้ออกมานั้นยังไม่มากพอ ทั้งนี้อาจเกิดจากการเลือกชนิดของโมทีฟที่ยังไม่มีจำนวนและความจำเพาะมากพอ ในการที่จะดึงเอาข้อมูลในส่วนที่เราต้องการออกมาได้ นอกจากนี้ในขั้นการกรองข้อมูลการแสดงออกของยีน กลุ่มยีนที่เราสนใจอาจถูกกรองออกไป เนื่องจากยีนเหล่านั้นเกิด missing data หรือยีนเหล่านั้นไม่มีนัยสำคัญมากพอที่จะผ่านกระบวนการกรองข้อมูลก็เป็นได้ แต่ทั้งนี้ทั้งนั้น ยีนที่ได้มาทั้ง 45 ยีน เป็นยีนที่เกี่ยวข้องกับโมทีฟที่เราสนใจมากกว่าครึ่งหนึ่ง ซึ่งนับเป็นที่น่าพอใจ เมื่อเทียบกับการใช้ข้อมูลการแสดงออกของยีนเพียงข้อมูลเดียว และพบว่ายังมีกลุ่มยีนที่ยังไม่ได้ทำการกำหนดหน้าที่อีกจำนวนหนึ่งซึ่งกลุ่มยีนเหล่านี้อาจเป็นกลุ่มยีนที่เกี่ยวข้องกับกระบวนการควบคุมการทำงานของวัฏจักรเซลล์ก็เป็นไปได้ ดังนั้นจึงต้องมีการทำการทดลองเพื่อหาหน้าที่ของยีนในกลุ่มนี้ต่อไป ในงานวิจัยนี้เราได้เลือกส่วนหน้ายีนที่มีความยาว 800 เบส เนื่องจากเป็นความยาวที่เหมาะสมสำหรับยีสต์ [14] แต่หากเราทำการเปลี่ยนชนิดของสิ่งมีชีวิตที่จะนำมาทำการทดลอง เราก็ต้องคำนึงถึงความยาวของส่วนหน้าของยีนของสิ่งมีชีวิตนั้นด้วย เช่น หากเราสนใจจะทำการศึกษาการจัดกลุ่มของยีนในพืช เราควรจะเพิ่มขนาดของส่วนหน้าของยีน เป็น 1,000 – 1,500 เบส [15] เนื่องจากตำแหน่งของการควบคุมการทำงานของยีนในพืชนั้น อยู่ไกลจากตำแหน่ง +1 มาก เพราะฉะนั้นเราต้องทำการเลือกส่วนหน้ายีนให้ครอบคลุมกับตำแหน่งของการควบคุมการทำงานของยีน อีกทั้งความยาวของโมทีฟก็นับเป็นส่วนสำคัญในการจัดกลุ่มยีนใหม่ เนื่องจากหากความยาวของโมทีฟมีขนาดที่ยาวจนเกินไป อาจส่งผลให้การค้นหาโมทีฟบนส่วนหน้าของยีน ไม่พบรูปแบบของโมทีฟนั้นๆ อยู่เลย ในทางตรงกันข้ามหากเราเลือกชนิดของโมทีฟที่มีขนาดสั้นจนเกินไป จะทำให้พบรูปแบบของโมทีฟบนส่วนหน้าของยีนมากกว่าความเป็นจริง ดังนั้นในงานวิจัยนี้เราได้ทำการเลือกชุดของโมทีฟที่มีความยาวประมาณ 6 –10 เบส ซึ่งเป็นความยาวที่เหมาะสม [15] กล่าวโดยสรุป การพัฒนาเครื่องมือการจัดกลุ่มยีนโดยการใช้ข้อมูลการแสดงออกของยีน และข้อมูลส่วนหน้าของยีน สามารถช่วยเพิ่มประสิทธิภาพของการจัดกลุ่มยีนให้มีความละเอียด และมีความถูกต้องมากขึ้น

เอกสารอ้างอิง

[1] Eisen MB, Spellman PT, Brown PO and Botstein D. (1998). Cluster Analysis and Display of Genome-Wide Expression Patterns. *Proc Natl Acad Sci U S A* 95, pp. 14863-8.

[2] http://www-stat.stanford.edu/~tibs/SAM/index.html

[3] Reich M, Ohm K, Angelo M, Tamayo P and Mesirov JP. "GeneCluster 2.0: an advanced toolset for bioarray analysis." Bioinformatics. 2004 Jul 22; 20(11) pp.1797-8

[4] http://www.molmine.com/frameset/frm_jexpress2.asp

[5] E. Segal , H. Wang and D. Koller "Discovering molecular pathways from protein interaction and gene expression data" Bioinformatics Vol. 19 Suppl. 1 2003 pp. i264-i272

[6] Peter Cheeseman and John Stutz "Bayesian Classification (AutoClass) Theory and Results" Advances in knowledge discovery and data mining 1996 pp.153 - 180

[7] Zhu G, Spellman PT, Volpe T, Brown PO, Botstein D, Davis TN and Futcher B. "Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth." Nature. 2000 Jul 6;406(6791) pp.90-4.

[8] http://www.mathworks.com/products/bioinfo/

[9] http://rsat.ulb.ac.be/rsat/

[10] http://cgsigma.cshl.org/jian/

[11] http://www.gene-regulation.com/pub/databases.html#transfac

[12] E. J. Moler , D. C. Radisky and I. S. Mian "Integrating naive Bayes models and external knowledge to examine copper and iron homeostasis in *S. cerevisiae*" *Physiol. Genomics* 4: 2000, pp.127-135

[13] Paul T. Spellman, Gavin Sherlock, Michael Q. Zhang, Vishwanath R. Iyer, Kirk Anders, Michael B. Eisen, Patrick O. Brown, David Botstein, and Bruce Futcher, "Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization" Mol Biol Cell. 1998 Dec;9(12) pp. 3273-97.

[14] Van Helden J ,"Metrics for comparing regulatory sequences on the basis of pattern counts."Bioinformatics. 2004 Feb 12;20(3):399-406. Epub 2004 Feb 5.

[15] Chintalapati Janaki and Rajendra R. Joshi, " Motif detection in *Arabidopsis*: Correlation with gene expression data", *In Silico* Biology 4, 0014 (2004);

วัฒน์วิบูลย์ แพรงมงคล จบการศึกษาระดับปริญญาตรี ในสาขาวิชาจุลชีววิทยา และกำลังศึกษาระดับปริญญาโท ในสาขาชีวสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี งานวิจัยที่สนใจคือ การวิเคราะห์ข้อมูลการแสดงออกของยีนที่จากไมโครอาเรย์ดีเอ็นเอ การวิเคราะห์ลำดับสายนิวคลีโอไทด์สายคู่และเดี่ยว และการพัฒนาซอฟท์แวร์ที่เกี่ยวข้องกับการวิเคราะห์ข้อมูลทางชีววิทยาโมเลกุลระดับโอมิก

Anan Tongta M.S., Ph.D. (Chemical Engineering), University of Missouri-Rolla, Rolla, Missouri, USA. B.S. (Chemical Engineering), Chulalongkorn University, Bangkok, Thailand. Research Interests: Modeling of Biological Processes, Adsorption of Biological Molecules

ดวงดาว วิชาดากุล จบการศึกษาปริญญาโทและเอกทางวิทยาศาสตร์คอมพิวเตอร์ จากมหาวิทยาลัยแห่งรัฐอิลลินอยส์ ความรู้และประสบการณ์ที่ผ่านมาเกี่ยวข้องกับการวิจัยและพัฒนาซอฟท์แวร์ระบบมิดเดิลแวร์และเมตาดาต้าเพื่อสนับสนุนการพัฒนา โปรแกรมเชิงประกอบ ในระบบเครือข่ายแบบกระจาย ที่ต้องการการประกันคุณภาพ (Quality of Service) เช่นประสิทธิภาพ ความเสถียร และความปลอดภัย ความสนใจปัจจุบันเกี่ยวข้องกับการวิจัยและพัฒนาทาง ชีวสารสนเทศโดยเฉพาะในสาขาที่เกี่ยวข้องกับการศึกษาการทำงานของกลุ่มยีนเครือข่ายควบคุมการทำงานของกลุ่มยีนในระดับจีโนม ความสัมพันธ์ของเครือข่ายควบคุมการทำงานของกลุ่มยีนระหว่างจีโนม เครือข่ายควบคุมการทำงานของกลุ่มโปรตีน รวมทั้งการประสานการทำงานของระบบในระดับยีนโปรตีน และเมตาโบลิซึ่ม