บทที่ 4

การวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์ ด้วยวิธีวิเคราะห์ปัจจัย

การวิเคราะห์ปัจจัย เป็นวิธีการวิเคราะห์ข้อมูลหลายตัวแปรที่มีการศึกษาค้นคว้าตั้งแต่ศตวรรษที่ 20 โดย คาร์ล เพียร์สัน และชาร์ล สเปียร์แมน ซึ่งต่อมามีนักวิทยาศาสตร์อีกหลายๆ ท่านได้ศึกษาและปรับปรุงวิธีการให้ดียิ่งขึ้น ซึ่งเริ่มแรกนั้นนำไปใช้ในการศึกษาทางด้านจิตวิทยา โดยปัญหานั้นเริ่มจากมีตัวแปรมากมายที่ใช้ในการศึกษาพฤติกรรมของมนุษย์หรือสัตว์ แต่ตัวแปรเหล่านี้มีจำนวนมากเกินกว่าที่จะหาคำอธิบายว่าจริงๆแล้วสิ่งที่มีผลหรือปัจจัยที่มีผลต่อพฤติกรรมที่ศึกษาคืออะไร

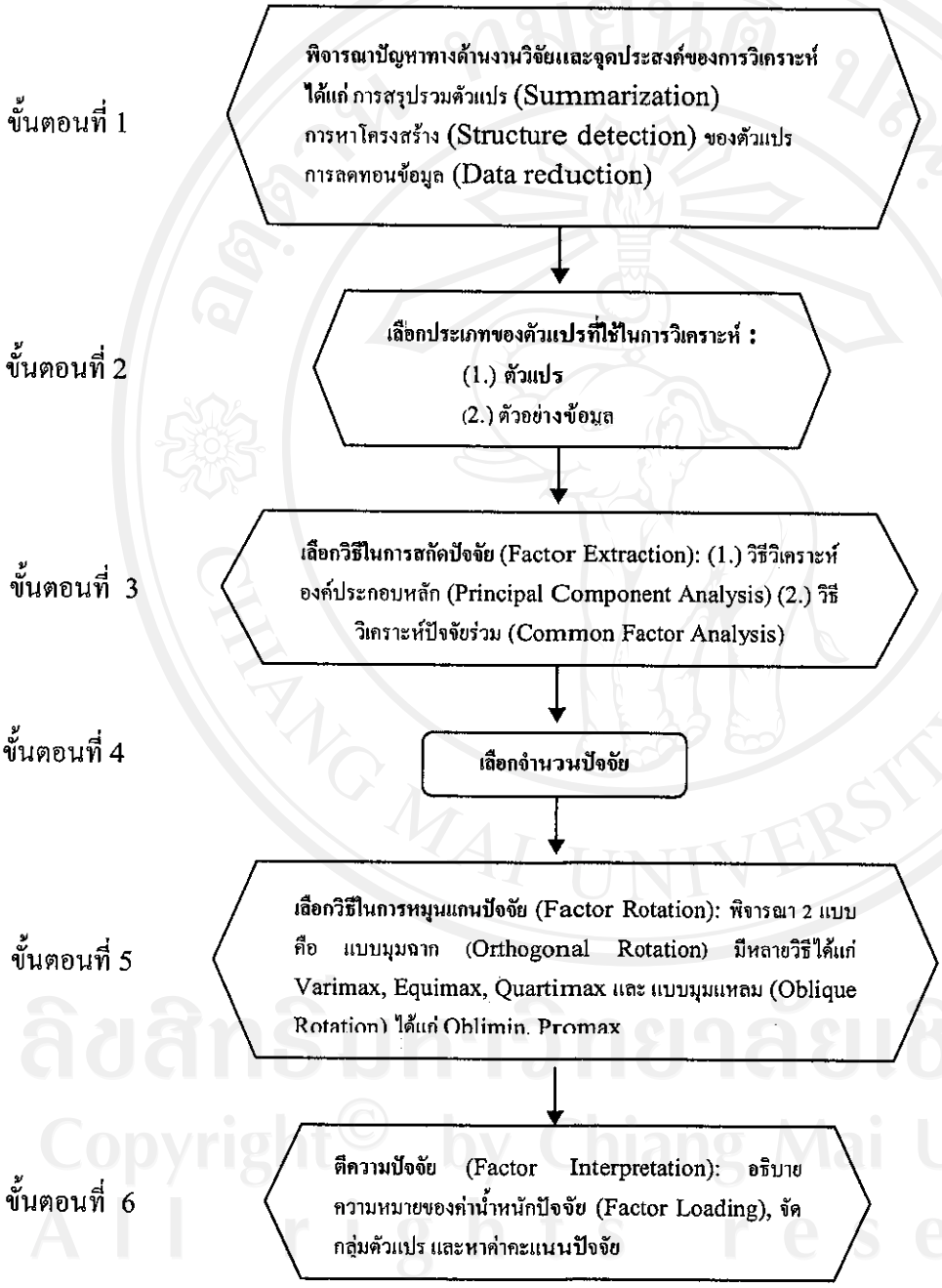
การวิเคราะห์ปัจจัยจึงเป็นกระบวนการวิเคราะห์ความสัมพันธ์ของตัวแปรเพื่อที่จะจัดกลุ่มของตัวแปร (Summarization) ที่มีแบบแผนโครงสร้างความสัมพันธ์ร่วมกันไว้ในลักษณะของตัวแปรแฝง (Latent Variables) หรือ ปัจจัย โดยปัจจัยที่ได้จะถูกนำกลับใช้ในการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปร การใช้ปัจจัยในการปรับเปลี่ยนข้อมูลให้อยู่ในรูปแบบของปัจจัยที่มีมิติของข้อมูลลดน้อยลง (Data Reduction) และการนำปัจจัยไปใช้เป็นข้อมูลตั้งต้นในกระบวนการวิเคราะห์อื่นๆ เช่น เป็นข้อมูลตั้งต้นสำหรับกระบวนการจัดกลุ่มข้อมูล (Clustering Data)

ในข้อมูลดีเอ็นเอไมโครอาร์เรย์ของสิ่งมีชีวิตหนึ่งๆ ของเซลล์หนึ่งๆ หรือในชุดการทดลองหนึ่งๆ นั้น เมื่อให้ยีนเป็นตัวแปรกลุ่มของยีนที่เกี่ยวข้อง ย่อมมีหน้าที่หรือคุณสมบัติต่างๆ ที่สัมพันธ์กันอยู่ภายใน ไม่สามารถที่จะสังเกตเห็น และหากพิจารณาให้เงื่อนไขการทดลองต่างๆเป็นตัวแปร ตัวแปรดังกล่าวย่อมมีทั้งส่วนที่สัมพันธ์กันและไม่สัมพันธ์กันซ่อนอยู่ภายในและไม่สามารถที่จะสังเกตเห็นเช่นเดียวกัน ดังนั้นเมื่อนำวิธีการนี้ไปวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์ ตัวอย่างหนึ่งของผลการวิเคราะห์นี้ก็คือ สามารถที่จะจัดกลุ่มยีนที่มีหน้าที่ความสัมพันธ์กันไว้ด้วยกันได้ และเราสามารถที่จะให้ความหมาย ของกลุ่มยีนดังกล่าวนี้ได้ในลักษณะของปัจจัยต่างๆ เป็นต้น

4.1 หลักการของวิธีวิเคราะห์ปัจจัย

4.1.1 ขั้นตอนการวิเคราะห์ปัจจัยโดยภาพรวม

การวิเคราะห์ปัจจัย มีข้อกำหนด และกระบวนการต่างๆ 6 ขั้นตอน ในรูป 4.1



รูป 4.1 แผนภาพแสดงขั้นตอนการวิเคราะห์ปัจจัย

จากแผนภาพขั้นตอนการวิเคราะห์ปัจจัยอธิบายได้ดังนี้

ขั้นตอนที่ 1

พิจารณาปัญหาทางด้านงานวิจัยนั่นคือการพิจารณาว่างานวิจัยดังกล่าวเป็นงานวิจัยในเชิงสำรวจ (Exploratory) หรือเป็นปัญหาในเชิงการยืนยันผล (Confirmatory) ในงานวิจัยนี้จะอาศัยการวิเคราะห์ปัจจัยในลักษณะของการสำรวจ โดยกำหนดวัตถุประสงค์ของการวิเคราะห์ข้อมูลได้แก่

1.) การสรุปรวมและการหาโครงสร้างของตัวแปร (Summarization and Structure detection)

เป็นการพิจารณาโครงสร้างของตัวแปรเพื่อที่จะหาตัวแปรแฝง (Latent Variable) หรือปัจจัยที่ซ่อนอยู่ภายในกลุ่มของตัวแปรที่มีความสัมพันธ์กัน หรือเป็นการสรุปรวมตัวแปรที่มีความสัมพันธ์กันไว้ในตัวแปรแฝงนั่นเอง ค่าความสัมพันธ์ของตัวแปรและปัจจัยจะเรียกว่า ค่าน้ำหนักปัจจัย (Factor Loading) ซึ่งเป็นค่าที่ใช้ระบุได้ว่า ตัวแปรควรจะอยู่ในปัจจัยใด การวิเคราะห์ปัจจัยสามารถทำได้ทั้งในลักษณะของตัวแปร และตัวอย่างข้อมูล

2.) การลดทอนข้อมูล (Data Reduction)

เป็นการสร้างข้อมูลขึ้นมาใหม่จากชุดข้อมูลเดิม โดยจำนวนตัวแปรของข้อมูลใหม่มีจำนวนน้อยกว่าเดิม นั่นคือตัวแปรของข้อมูลใหม่นี้คือ ปัจจัย ซึ่งข้อมูลที่ได้จากการลดทอนนี้จะเรียกว่า คะแนนปัจจัย (Factor Score) และผลจากการลดทอนข้อมูลนี้จะนำไปใช้ในการวิเคราะห์ข้อมูลหลายตัวแปรในลักษณะอื่นๆ ต่อไปได้

ขั้นตอนที่ 2

กำหนดตัวแปรที่ใช้ในการวิเคราะห์ว่าเป็น ตัวแปร หรือตัวอย่างข้อมูล ทั้งนี้การวิเคราะห์ปัจจัยในลักษณะของตัวอย่างข้อมูลจะเป็นลักษณะที่คล้ายกับ การวิเคราะห์การแบ่งกลุ่ม (Cluster Analysis) แต่วิธีการวิเคราะห์ปัจจัยในลักษณะนี้จะไม่กล่าวถึง เนื่องจากในงานวิจัยไม่ได้สนใจศึกษา

ขั้นตอนที่ 3

การสกัดปัจจัย (Factor Extraction) เป็นขั้นตอนของการหาปัจจัยซึ่งมี 2 วิธีการใหญ่ๆ นั่นคือ วิธีวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis) และวิธีวิเคราะห์ปัจจัยร่วม (Common Factor Analysis) ในการเลือกใช้วิธีการใดนั้น ขึ้นกับจุดประสงค์ของงานวิจัย นั่นคือ วิธีวิเคราะห์องค์ประกอบหลัก จะสนใจที่การลดจำนวนของตัวแปรทั้งหมดให้อยู่ในปัจจัยที่มีจำนวนน้อยเพื่อให้ปัจจัย เป็นตัวเก็บค่าความเป็นข้อมูลเดิมให้ได้มากที่สุด ซึ่งค่าความเป็นข้อมูลเดิมนี้วัดได้จากค่าความแปรปรวน ส่วนวิธีวิเคราะห์ปัจจัยร่วมจะสนใจที่การหาปัจจัยที่เป็นตัวแทนของกลุ่มตัวแปรบางกลุ่มที่มีค่าร่วมกัน(Common) กับปัจจัยนั้นๆเท่านั้น นั่นคือตัวแปรทั้งหมดไม่ได้เป็นตัวก่อให้เกิด

ปัจจัยใดปัจจัยหนึ่ง ด้วยเหตุนี้การวิเคราะห์ปัจจัยในลักษณะนี้เราจะสามารถบอกได้ว่าตัวแปรใดบ้างที่เกี่ยวข้องกับปัจจัยที่เราพิจารณา และด้วยค่ามากน้อยเท่าไร

ขั้นตอนที่ 4

การเลือกจำนวนปัจจัยที่เหมาะสมจะช่วยให้ปัจจัยที่มีทั้งหมดสามารถเป็นตัวแทนของข้อมูลเดิมได้ดี ซึ่งมีวิธีการเลือกอยู่หลายวิธีการ ดังจะได้กล่าวถึงต่อไป

ขั้นตอนที่ 5

ผลจากการสกัดปัจจัยและการเลือกปัจจัย จะทำให้เราได้ปัจจัยที่เป็นตัวแทนของตัวแปรต่างๆ ความสัมพันธ์ระหว่างปัจจัยและตัวแปรแสดงออกมาในลักษณะของค่าร่วมกัน หรือค่าน้ำหนักปัจจัย แต่เนื่องจากสัดส่วนของค่าหรือสเกลของค่าเหล่านี้บางครั้งมีความแตกต่างกันในระดับที่เราไม่สามารถอ่านหรือตีความหมายได้ นั่นคือไม่สามารถบอกได้ว่าตัวแปรตัวไหนบ้างมีความสัมพันธ์กับปัจจัยที่เราพิจารณา ทำให้ปัจจัยที่ได้ไม่มีความหมาย ด้วยเหตุนี้การหมุนแกนปัจจัยจึงถือเป็นขั้นตอนสำคัญที่นำมาใช้ในการ ให้ความหมายของปัจจัยแต่ละปัจจัยเด่นชัดขึ้น ซึ่งมีหลักการคือ การกระจายความแปรปรวนของปัจจัยแต่ละปัจจัยขึ้นใหม่ ให้แยกความแตกต่างของตัวแปรที่มีความสัมพันธ์ กับปัจจัยที่พิจารณา และตัวแปรที่ไม่มีความสัมพันธ์กับปัจจัยดังกล่าวให้เด่นชัด ทั้งนี้ไม่ได้ทำให้อัตราส่วนของค่าความแปรปรวนทั้งหมด ที่อธิบายโดยปัจจัยที่ต่างกัน เปลี่ยนแปลงไป การเลือกวิธีการหมุนแกนปัจจัย มีหลายวิธีได้แก่ การหมุนแกนแบบมุมฉาก (Orthogonal Rotation) ซึ่งการคำนวณมี 3 วิธีคือ วาริแมกซ์ (Varimax) อีคิว์แมกซ์ (Equimax) และ คัวร์ติแมกซ์ (Quartimax) และ การหมุนแกนแบบมุมแหลม (Oblique Rotation) ซึ่งการคำนวณมี 2 วิธีคือ ออบลิมิน (Oblimin) และโปรแมกซ์ (Promax) สำหรับรายละเอียดจะกล่าวถึงต่อไป

ขั้นตอนที่ 6

การตีความปัจจัย (Factor Interpretation) เป็นกระบวนการที่นำพารามิเตอร์ที่ได้จากการวิเคราะห์ปัจจัย มาใช้อธิบายหาความหมายของข้อมูล ซึ่งพารามิเตอร์ที่ได้นั้นก็ได้แก่ ค่าน้ำหนักปัจจัย (Factor Loading) เป็นค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรกับปัจจัยซึ่งใช้ในการอธิบายความสัมพันธ์ระหว่างปัจจัยและตัวแปร ทำให้สามารถจัดกลุ่มตัวแปร (Summarization) ที่สัมพันธ์กันไว้ในปัจจัยเหล่านั้นได้ นอกจากนี้ พารามิเตอร์จากการวิเคราะห์ปัจจัยอีกตัวที่สำคัญก็คือ ค่าคะแนนปัจจัย (Factor Score) ซึ่งเป็นค่าที่ได้จากการแปลงชุดข้อมูลเดิมให้อยู่ในลักษณะของปัจจัยข้อมูล โดยแปลงข้อมูลด้วยกระบวนการโปรเจคชั่น ข้อมูลเดิมจากตัวแปรต่างๆ ลงในปัจจัยที่ได้จาก กระบวนการสกัดและการหมุนแกนปัจจัย ผลก็คือได้ข้อมูลชุดใหม่ที่มีมิติของข้อมูลน้อยลง (Data Reduction)

4.1.2 โมเดลการวิเคราะห์ปัจจัย

กำหนดให้เวกเตอร์ X คือเวกเตอร์ของข้อมูลที่สังเกตเห็น (Observation Data) ซึ่งประกอบไปด้วย p ตัวแปร มีค่าเฉลี่ยของ X ในแต่ละตัวแปรเป็นเวกเตอร์ μ และมีเมตริกซ์ความแปรปรวนร่วม Σ

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_p \end{bmatrix} \tag{31}$$

การวิเคราะห์ปัจจัยจะตั้งสมมุติฐานว่า X คือผลรวมเชิงเส้นของปัจจัยร่วม (Common factors) ใช้สัญลักษณ์ F มีขนาดขนาด m กับน้ำหนักปัจจัย (Loading Factor) สัญลักษณ์ l รวมกับเวกเตอร์ ε ซึ่งเป็นเวกเตอร์ของค่าผิดพลาด (Errors) หรือ ปัจจัยเฉพาะ (Specific Factors) แสดงได้ดังสมการ

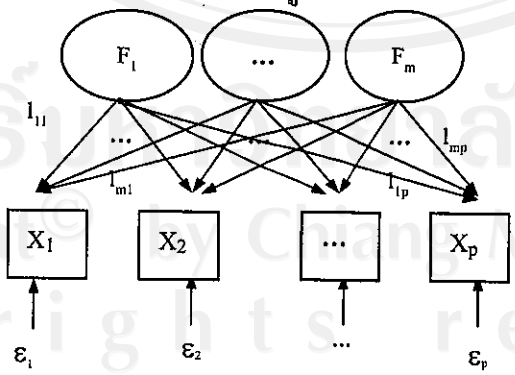
$$\begin{aligned} X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \varepsilon_1 \\ X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \varepsilon_2 \\ &\vdots \\ X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pm}F_m + \varepsilon_p \end{aligned} \tag{32}$$

ถ้าให้ X เป็นเมตริกซ์ขนาด $p \times 1$ จะแสดงสมการ(32) ใหม่ได้เป็น

$$\underset{(p\times 1)}{X - \mu} = \underset{(p\times m)}{L}\ \underset{(m\times 1)}{F} + \underset{(p\times 1)}{\varepsilon} \tag{33}$$

จากสมการ (32) และ (33) l_{ij} คือ ค่าน้ำหนักปัจจัย (Factor loading) ที่ตัวแปร i และปัจจัยที่ j และ L คือ เมตริกซ์ของน้ำหนักปัจจัย ซึ่งค่าน้ำหนักปัจจัยจะเป็นตัวช่วยอธิบายได้ว่าตัวแปรต่างๆ ที่สังเกตเห็นได้นั้น อธิบายได้ โดยปัจจัยที่ซ่อนอยู่ด้วยความแปรปรวนมากน้อยเท่าไร

จากสมการแสดงแผนภาพโมเดลการวิเคราะห์ได้ในรูป 4.2



รูป 4.2 แผนภาพโมเดลการวิเคราะห์ปัจจัย

จากสมการ (33) และ แผนภาพในรูป 4.2 จะเห็นว่า ข้อมูลสังเกตการณ์ในแต่ละตัวแปร เกิดจากผลรวมเชิงเส้นของปัจจัยร่วมที่อยู่ภายในด้วยค่าน้ำหนักที่ต่างๆกันรวมกับค่าความผิดพลาดหรือปัจจัยเฉพาะ

กำหนดให้ $E(F)$ คือเวกเตอร์ค่าเฉลี่ยของ F ส่วน $Cov(F)$ คือเมตริกซ์ความแปรปวนร่วม (Covariance Matrix) ของ F และ ψ คือเมตริกซ์ทแยงมุม (Diagonal Matrix) ซึ่งเป็นเมตริกซ์ความแปรปรวนร่วมของปัจจัยเฉพาะ ความแปรปรวนนี้อาจเรียกว่าความแปรปรวนเฉพาะ (Specific Variance) หรือ ในบางตำราจะเรียกว่า ยูนิคเนส (Uniqueness)

โมเดลการวิเคราะห์ปัจจัยจากสมการ (33) มีข้อสมมุติฐานเบื้องต้นดังนี้

$$E(F) = \underset{(m\times 1)}{0}, Cov(F) = E(FF') = \underset{(m\times m)}{I}$$

$$E(\varepsilon) = \underset{(p\times 1)}{0}, Cov(\varepsilon) = E(\varepsilon\varepsilon') = \underset{(p\times p)}{\psi} = \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix} \tag{34}$$

จากข้อมุติฐานดังกล่าวมีการพิสูจน์และพบว่า

$$\Sigma = Cov(X) = LL' + \psi$$

นั่นคือ

$$\begin{aligned} Var(X_i) &= l_{i1}^2 + \cdots + l_{im}^2 + \psi_i \\ Cov(X_i, X_k) &= l_{i1}l_{k1} + \cdots + l_{im}l_{kmi}; \quad i,k = 1,2,...,p \end{aligned} \tag{35}$$

และ

$$Cov(X,F) = L$$

นั่นคือ

$$Cov(X_i, F_j) = l_{ij} \tag{36}$$

จากสมการ (35) กำหนดให้

$$h_i^2 = l_{i1}^2 + l_{i2}^2 + \cdots + l_{im}^2 \tag{37}$$

ดังนั้น

$$\sigma_{ii} = h_i^2 + \psi_i \quad ,i = 1,2,\cdots, p \tag{38}$$

เรียก h_i^2 ว่าค่าร่วมกัน(communality) ซึ่งเป็นผลรวมของค่าความแปรปรวนของปัจจัยร่วม (common factor) ต่อตัวแปรที่ i ทั้ง m ปัจจัย และ ψ_i คือ ค่าความแปรปรวนเฉพาะ (specific variance) ของตัวแปรตัวที่ i ซึ่งเป็นค่าความแปรปรวนที่อธิบายถึงปัจจัยเฉพาะ หรือค่าความผิดพลาด ดังนั้นในสมการ (38) อธิบายได้ว่า ค่าความแปรปรวนของข้อมูล σ ณ ตัวแปรใดๆ เกิดจากค่าความแปรปรวนของปัจจัยร่วมทุกๆตัวหรือเรียกว่าค่าร่วมกัน รวมกับ ค่าความแปรปรวนของปัจจัยเฉพาะ ที่มีต่อตัวแปรนั้นๆ ซึ่งความสัมพันธ์ดังกล่าวจะถูกนำไปใช้ สำหรับการประมาณค่าพารามิเตอร์ต่างๆ ในโมเดลการวิเคราะห์ ต่อไป

จากโมเดลการวิเคราะห์ปัจจัย พารามิเตอร์ที่สามารถประมาณค่าได้เป็นอันดับแรกคือ ค่าน้ำหนักปัจจัย และปัจจัยเฉพาะ ซึ่งหาได้จากกระบวนการสกัดปัจจัย

4.1.3 การสกัดปัจจัย (Factor Extraction)

วิธีการที่ใช้ในการสกัดปัจจัยมี 2 วิธีใหญ่ๆ คือ วิธีวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis) และ วิธีวิเคราะห์ปัจจัยร่วม (Common Factor Analysis)

1) วิธีวิเคราะห์องค์ประกอบหลัก

หลักการพื้นฐานของการวิเคราะห์องค์ประกอบหลัก จะอาศัยความแปรปรวนร่วมของข้อมูลทั้งหมด (ความแปรปรวนจากปัจจัยร่วม และ ความแปรปรวนจากปัจจัยเฉพาะ) มาใช้ในการประมาณค่าน้ำหนักปัจจัย ซึ่งจากสมการ (31) เมตริกซ์ความแปรปรวนร่วมของข้อมูล Σ จะเป็นเมตริกซ์ข้อมูลตั้งต้นสำหรับใช้ในการประมาณค่า สำหรับวิธีการหาเมตริกซ์เมตริกซ์ความแปรปรวนร่วมของข้อมูลนี้ได้กล่าวถึงแล้วในบทที่ 3

จากเมตริกซ์ความแปรปรวนร่วม Σ มีคู่อันดับของไอเกนแวลยู-ไอเกนเวกเตอร์เป็น (λ_i, e_i) โดยที่ $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_i \geq \cdots \lambda_p \geq 0$

ซึ่งความสัมพันธ์ของเมตริกซ์ความแปรปรวนร่วม และ ไอเกนแวลยู-ไอเกนเวกเตอร์ แสดงได้ดังสมการ (39) และสมการ (40)

$$\Sigma = \lambda_1 e_1 e_1' + \lambda_2 e_2 e_2' + \cdots + \lambda_p e_p e_p' \tag{39}$$

$$\Sigma = \left[\sqrt{\lambda_1} e_1 \vdots \sqrt{\lambda_2} e_2 \vdots \cdots \vdots \sqrt{\lambda_p} e_p\right] \begin{bmatrix} \sqrt{\lambda_1} e_1' \\ \sqrt{\lambda_2} e_2' \\ \vdots \\ \sqrt{\lambda_p} e_p' \end{bmatrix} \tag{40}$$

จากสมการ (39) เมื่อนำมาพิจารณาร่วมกับสมการ (35) โดยให้ จำนวนปัจจัย m เท่ากับจำนวนตัวแปร p จะทำให้ค่าความแปรปรวนเฉพาะ ψ ในสมการ (35) มีค่าเท่ากับ 0 นั่นคือ

$$\underset{(p\times p)}{\sum} = \underset{(p\times p)}{L}\underset{(p\times p)}{L'} + \underset{(p\times p)}{0} = LL' \tag{41}$$

ดังนั้นจากสมการ (35) และ สมการ (40) เมื่อให้ $m < p$ จะได้

$$\begin{aligned}\Sigma &= LL' + \psi \\ &= \left[\sqrt{\lambda_1}e_1 \vdots \sqrt{\lambda_2}e_2 \vdots \cdots \vdots \sqrt{\lambda_m}e_m\right]\begin{bmatrix}\sqrt{\lambda_1}e_1' \\ \sqrt{\lambda_2}e_2' \\ \vdots \\ \sqrt{\lambda_m}e_m'\end{bmatrix} + \begin{bmatrix}\psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p\end{bmatrix}\end{aligned} \tag{42}$$

ซึ่ง $\psi_i = \sigma_{ii} - h_i^2 \quad , i = 1,2,\cdots,p$

ผลจากสมการ(42) ทำให้หาค่าน้ำหนักปัจจัย (factor loading) L ได้ ดังสมการ (43)

$$L = \left[\sqrt{\lambda_1}e_1 \vdots \sqrt{\lambda_2}e_2 \vdots \cdots \vdots \sqrt{\lambda_m}e_m\right] \tag{43}$$

เนื่องจากในกระบวนการวิเคราะห์ข้อมูลนั้น ข้อมูลจะอยู่ในรูปของกลุ่มตัวอย่าง ดังนั้น เมตริกซ์ความแปรปรวนร่วมของประชากร Σ จึงแทนด้วยเมตริกซ์ความแปรปรวนร่วมของกลุ่มตัวอย่างด้วย S และ s_{ii} คือความแปรปรวนของข้อมูลในตัวแปรตัวที่ i ซึ่งค่าความแปรปรวนของแต่ละปัจจัยได้จาก สมการ

$$Var(F_j) = l_{1j}^2 + l_{2j}^2 + \cdots + l_{pj}^2 = \left(\sqrt{\lambda_j}e_j\right)'\left(\sqrt{\lambda_j}e_j\right) = \lambda_j \tag{44}$$

ดังนั้น

$$\begin{bmatrix}\text{สัดส่วนของความแปรปรวนของ} \\ \text{ข้อมูลที่ปัจจัย } j\end{bmatrix} = \frac{\lambda_j}{s_{11}+s_{22}+\cdots+s_{pp}} \quad j = 1,2,\cdots,m \tag{45}$$

สำหรับเมตริกซ์ความแปรปรวนที่ใช้ ถ้าหากใช้ เป็นเมตริกซ์สหสัมพันธ์ R สัดส่วนความแปรปรวนของข้อมูล จะเป็น

$$\begin{bmatrix}\text{สัดส่วนของความแปรปรวนของ}\\ \text{ข้อมูลที่ปัจจัย } j\end{bmatrix} = \frac{\lambda_j}{p} \quad j=1,2,\cdots,m \tag{46}$$

จากสัดส่วนความแปรปรวนนี้ จะใช้เป็นเงื่อนไขในการเลือกจำนวนของปัจจัยร่วมที่เหมาะสม

2) วิธีวิเคราะห์ปัจจัยร่วม

หลักการพื้นฐานคือ การประมาณค่าน้ำหนักปัจจัยโดยวิธีนี้จะใช้ค่าความแปรปรวน จากความแปรปรวนของปัจจัยร่วมเท่านั้น สำหรับความแปรปรวนจากปัจจัยเฉพาะไม่นำมาใช้ในการประมาณค่าน้ำหนักปัจจัย ซึ่งต่างกับวิธีวิเคราะห์องค์ประกอบที่พิจารณาร่วมกันทั้งหมด นั่นคือจากสมการ (35) เมตริกซ์ความแปรปรวนร่วม Σ จะต้องเป็นเมตริกซ์ความแปรปรวนของข้อมูลที่เกิดจากปัจจัยร่วมเท่านั้น จึงจะสามารถนำไปใช้ในการหาค่าน้ำหนักปัจจัยได้ ด้วยเหตุนี้การที่จะประมาณค่าน้ำหนักปัจจัยได้ จะต้องประมาณค่าความแปรปรวนร่วมของปัจจัยร่วมของข้อมูลให้ได้ก่อน ซึ่งวิธีการในการประมาณค่า เมตริกซ์ความแปรปรวนร่วมดังกล่าวนี้ จะอาศัย ทฤษฎีความเป็นไปได้สูงสุด(Maximum Likelihood Method) ในการประมาณค่า ซึ่งจากสมการ (35) นั้น เมตริกซ์ความแปรปรวนร่วม จะมีพารามิเตอร์ที่สัมพันธ์กันอยู่ คือ น้ำหนักปัจจัย ที่อยู่ในรูปของค่าร่วมกัน(communality) และ ปัจจัยเฉพาะ ดังนั้นจากทฤษฎีความเป็นไปได้สูงสุดนี้ เมื่อประมาณค่าเมตริกซ์ความแปรปรวนร่วมจากปัจจัยร่วมได้ ก็จะสามารถประมาณค่าน้ำหนักปัจจัย และปัจจัยเฉพาะได้

จากโมเดลของการวิเคราะห์ปัจจัยในสมการ (33) จะตั้งสมมุติฐานว่า ปัจจัยร่วม และปัจจัยเฉพาะนั้น มีการแจกแจงแบบปกติ เป็นผลให้ข้อมูลสังเกตการณ์ ที่ได้นั้นมีการแจกแจงแบบปกติด้วย และจากคุณสมบัติดังกล่าวนี้เอง จึงสามารถประมาณค่าความแปรปรวนร่วมของปัจจัยร่วม ได้โดยการทำให้ฟังก์ชันความเป็นไปได้ในสมการต่อไปนี้มีค่าสูงที่สุด (Maximization)

$$L(\mu,\Sigma) = (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n}{2}} e^{-\left(\frac{1}{2}\right)tr\left[\Sigma^{-1}\left(\sum_{j=1}^{n}(x_j-\bar{x})\ (x_j-\bar{x})'+n(\bar{x}-\mu)\ (\bar{x}-\mu)'\right)\right]} \tag{47}$$

ในสมการ (47) กำหนดให้ $X_1, X_2, \cdots, X_n$ คือชุดข้อมูลจำนวน n ตัวอย่าง จะเห็นว่าในสมการดังกล่าวเป็นการประมาณค่าหาความแปรปรวนร่วม (Σ) ดังนั้นจากสมการ (35) เมื่อแทนค่าความแปรปรวนร่วมด้วย $LL'+\psi$ ก็จะทำให้เราสามารถประมาณค่าน้ำหนักปัจจัย $\hat{L}$ และ ความ

แปรปรวนเฉพาะ $\hat{\psi}$ ได้ นอกจากนี้ผลจากการประมาณค่า ทั้งค่าน้ำหนักปัจจัยและปัจจัยเฉพาะ จะต้องเป็นไปตามเงื่อนไขดังนี้

$$L'\psi L = \Delta \tag{48}$$

โดยที่ Δ คือเมตริกซ์ทแยง(Diagonal Matrix)

เนื่องจากในวิธีการประมาณค่า มีความสลับซับซ้อนในกระบวนการคำนวณจึงไม่ อธิบายลงในรายละเอียด สำหรับการวิเคราะห์ข้อมูลจะใช้ฟังก์ชันสำเร็จรูป จากโปรแกรมการคำนวณในการวิเคราะห์โดยตรง

หลังจากประมาณค่าน้ำหนักปัจจัยได้แล้ว ค่าร่วมกันของข้อมูล จากการประมาณค่าจะเป็น

$$\hat{h}_i^2 = \hat{l}_{i1}^2 + \hat{l}_{i2}^2 + \cdots + \hat{l}_{im}^2 \quad , i = 1,2,\cdots, p \tag{49}$$

หาสัดส่วนความแปรปรวนของข้อมูลได้จาก

$$\left[\begin{array}{l}\text{สัดส่วนของความแปรปรวนของ} \\ \text{ข้อมูลที่ปัจจัย } j\end{array}\right] = \frac{\hat{l}_{1j}^2 + \hat{l}_{2j}^2 + \cdots + \hat{l}_{pj}^2}{s_{11} + s_{22} + \cdots + s_{pp}} \quad j = 1,2,\cdots, m \tag{50}$$

จากวิธีในการสกัดปัจจัยทั้งสองวิธี การเลือกใช้วิธีใดต้องพิจารณา วัตถุประสงค์ของการวิเคราะห์ และ ลักษณะของข้อมูลตั้งต้นที่จะใช้ในการวิเคราะห์ นั่นคือ วิธีวิเคราะห์องค์ประกอบหลักจะเป็นวิธีที่เหมาะสม เมื่อจุดประสงค์คือ การลดจำนวนของข้อมูลให้อยู่ในรูปของปัจจัยที่มีสัดส่วนของความแปรปรวนของข้อมูลมากที่สุดซึ่งสามารถอธิบายความเป็นข้อมูลเดิมได้ และลักษณะของข้อมูลที่นำเข้านั้น ไม่สนใจที่จะพิจารณาค่าความผิดพลาดหรือปัจจัยเฉพาะของข้อมูล หรือ มั่นใจได้ว่าปัจจัยเฉพาะดังกล่าวนั้นมีความแปรปรวนเพียงเล็กน้อย ไม่มีผลต่อการประมาณค่าน้ำหนักปัจจัย สำหรับในกรณีที่ใช้ วิธีการวิเคราะห์ปัจจัยร่วม จะใช้ได้ดี หากจุดประสงค์ของการวิเคราะห์นั้นคือ การจัดกลุ่มตัวแปร ให้กับปัจจัย ซึ่งต้องแยกแยะปัจจัยร่วมและปัจจัยเฉพาะออกจากกัน ดังนั้น ลักษณะของข้อมูลที่นำเข้านั้นจะเป็นข้อมูลที่มีค่าความผิดพลาด หรือ ความแปรปรวนจากปัจจัยเฉพาะมาก จึงต้องอาศัยวิธีการนี้ในการแยกออกจากกัน เพื่อประมาณค่าน้ำหนักปัจจัยที่เหมาะสม

4.1.4 การเลือกปัจจัย และหาจำนวนปัจจัยที่เหมาะสม

เงื่อนไขในการเลือกจำนวนปัจจัยที่เหมาะสมในเบื้องต้นนั้นคือ จำนวนปัจจัยต้องน้อยกว่าจำนวนตัวแปร และกรณีที่ใช้การสกัดปัจจัยโดยวิธีวิเคราะห์ปัจจัยร่วม จำนวนปัจจัยที่เหมาะสมต้องเป็นไปตามเงื่อนไขดังสมการ (51)

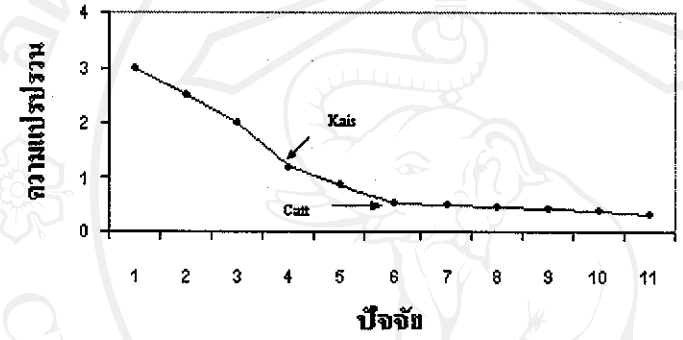
$$\frac{1}{2}\left[(p-m)^2 - p - m\right] > 0 \qquad (51)$$

ทั้งนี้การสกัดปัจจัยโดยวิธีวิเคราะห์องค์ประกอบหลัก เงื่อนไขในสมการ(51) ไม่จำเป็นต้องใช้

สำหรับหลักการพิจารณา เลือกปัจจัยและหาจำนวนปัจจัยที่เหมาะสม เมื่อจำนวนปัจจัยอยู่ภายในเงื่อนไขเบื้องต้นที่กำหนดแล้วมีหลักการดังนี้

1) กำหนดจำนวนปัจจัยโดยผู้วิจัยเอง นั่นคือ การพิจารณาจากลักษณะของข้อมูลตั้งต้น แล้วกำหนดจำนวนปัจจัย มาก่อนที่จะวิเคราะห์

2) เลือกปัจจัย โดยพิจารณาจาก สครีพล็อต (Scree plot) แสดงตัวอย่าง ดังรูป



รูป 4.3 สครีพล็อตอธิบายหลักการเลือกจำนวนปัจจัย

สครีพล็อต เป็นกราฟที่แสดงถึงค่าความแปรปรวนของข้อมูลในแต่ละปัจจัย ซึ่งโดยวิธีวิเคราะห์องค์ประกอบหลัก ค่าความแปรปรวนก็คือ ค่าของไอเกนแวลยู

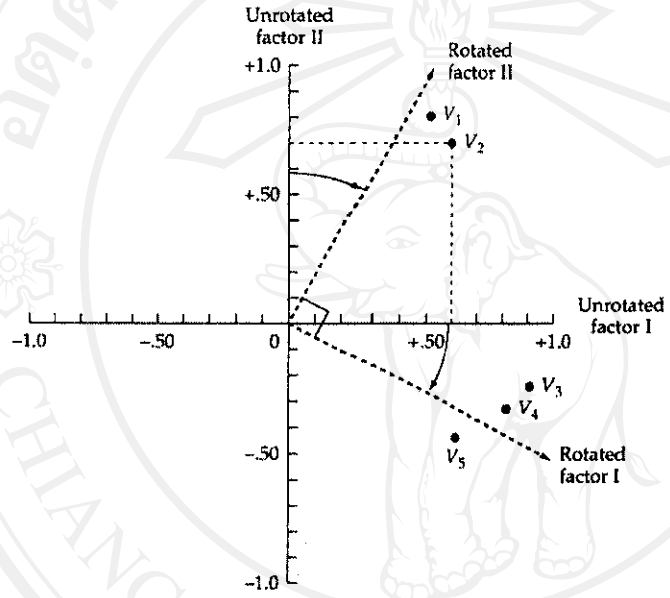
การพิจารณา มี 2 วิธีคือ โดยเลือกเอาปัจจัยที่มีค่าของความแปรปรวน มากกว่าหรือเท่ากับ 1 วิธีนี้นำเสนอโดยไคเซอร์ (Kaiser) จึงเรียกอีกอย่างว่า วิธีไคเซอร์ จากตัวอย่างในรูป 4.3 จะได้ 4 ปัจจัย และ อีกวิธีหนึ่งจะเลือกเอาปัจจัยทางด้านซ้าย ของตำแหน่งที่ความชันของกราฟเริ่มราบเรียบ (smooth) จากตัวอย่าง จะได้ 6 ปัจจัย วิธีนี้นำเสนอโดย คาท์เทลล์ (Cattell)

3) เลือกปัจจัย โดยพิจารณาจากค่าของสัดส่วนความแปรปรวนที่มีค่ามากๆ และหาจำนวนปัจจัยโดยการพิจารณาจากสัดส่วนความแปรปรวนรวมของปัจจัยที่เลือกมา ซึ่งขึ้นกับผู้วิจัยเอง ว่าจะพิจารณาที่ร้อยละเท่าไร

4) เลือกปัจจัยโดยพิจารณาจากค่าของน้ำหนักปัจจัยปัจจัย นั่นคือเมื่อสามารถจัดกลุ่มของตัวแปรในแต่ละปัจจัย และอธิบายความหมายของปัจจัยได้แล้ว ปัจจัยอื่นๆ ที่เหลือ ก็ไม่เลือกอีก

4.1.5 การหมุนแกนปัจจัยและการอธิบายความหมายของปัจจัย(Rotation and Interpretation)

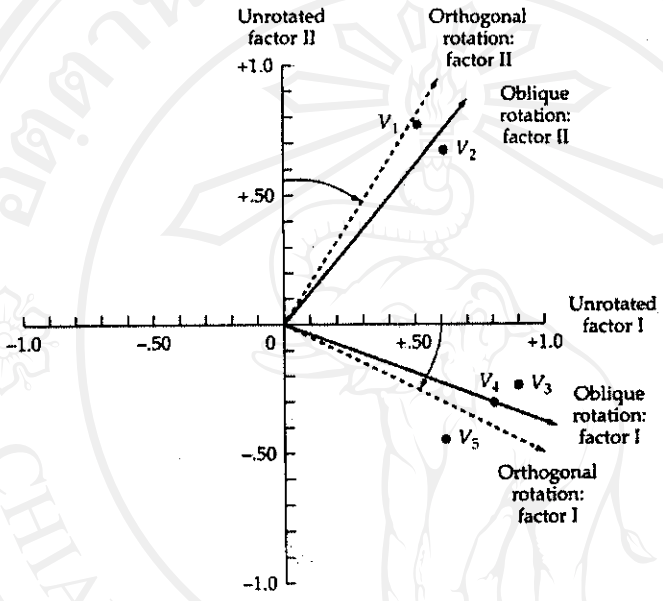
จุดประสงค์ของการหมุนแกนปัจจัยคือ เพื่อทำให้น้ำหนักปัจจัยที่สกัดได้นั้น นำมาตีความหมายได้ง่ายขึ้น ในการหมุนแกนปัจจัย มี 2 วิธีขึ้นกับว่าผู้วิจัย จะกำหนดให้ปัจจัยที่ได้นั้น มีความสัมพันธ์กันหรือไม่ (Correlated) โดยวิธีแรกเรียกว่า การหมุนแกนปัจจัยแบบมุมฉาก (Orthogonal Rotation) วิธีนี้ปัจจัยแต่ละปัจจัยเป็นอิสระต่อกันอย่างชัดเจน (Uncorrelated) และอีกวิธีหนึ่งเรียกว่า การหมุนแกนปัจจัยแบบมุมแหลม (Oblique Rotation) โดยวิธีนี้แต่ละปัจจัยมีความสัมพันธ์กัน แสดงภาพการหมุนแกนปัจจัยได้ดัง รูป 4.4 และรูป 4.5



รูป 4.4 การหมุนแกนปัจจัยแบบมุมฉาก

จากรูป 4.4 แสดงกราฟ การหมุนแกนปัจจัยแบบมุมฉาก แกนหลักแสดงถึง ปัจจัยที่ยังไม่มีการหมุนแกน ซึ่งมีสองปัจจัย มีตัวแปร (Variable : V) ที่เกี่ยวข้อง 5 ตัวแปร ซึ่งเมื่อพิจารณาที่ตัวแปร จะเห็นว่าตัวแปรตัวที่ 1 และ 2 จะมีความสัมพันธ์ใกล้ชิดกันและสัมพันธ์กับปัจจัยที่ 2 มากกว่า ตัวแปรตัวที่ 3, 4 ซึ่งสัมพันธ์กับ ปัจจัยตัวที่ 1 แต่ทั้งนี้แม้ว่าปัจจัยที่พิจารณาขณะที่ยังไม่หมุนแกนนั้นจะพอตีความหมายของข้อมูลได้ แต่ว่าก็มีตัวแปรบางตัวที่ไม่สามารถระบุได้อย่างชัดเจนว่าควรอยู่ในปัจจัยใด เช่นตัวแปรตัวที่ 5 ด้วยเหตุนี้การหมุนแกนปัจจัยจึงนำมาใช้ เพื่อที่จะเปลี่ยนแกนปัจจัยให้เป็นแกนใหม่ที่ทำให้สามารถจัดกลุ่มตัวแปร เข้ากับปัจจัยได้ดีขึ้นได้โดยที่ความแปรปรวนของข้อมูลยังเท่าเดิม ซึ่งจะช่วยในการตีความหมายของข้อมูลได้ง่ายขึ้น จะเห็นว่า ตัวแปรตัว 5 หลังจากหมุนแกนปัจจัย ตัวแปรดังกล่าวจะมีความสัมพันธ์กับปัจจัยตัวที่ 1 มาก

ลักษณะการหมุนแกนแบบมุมฉากนั้นทุกๆ แกนจะตั้งฉากกันเสมอ ทำให้ปัจจัยแต่ละปัจจัยมีคุณสมบัติที่เป็นอิสระต่อกัน ไม่มีความสัมพันธ์กัน ซึ่งต่างจากการหมุนแกนแบบมุมแหลมที่ผลจากการหมุนแกนจะทำให้มุมของแกนใหม่ ระหว่างปัจจัย ไม่ได้อยู่ในลักษณะของมุมฉากเสมอไป นั่นคือ การหมุนแกนปัจจัย จะพิจารณาถึงความสัมพันธ์กันระหว่างปัจจัยด้วย และ แกนของปัจจัยที่มีความสัมพันธ์กันมาก มุมระหว่างปัจจัยดังกล่าวนั้นก็จะเป็นมุมแหลมดังรูป 4.5



รูป 4.5 การหมุนแกนปัจจัยแบบมุมแหลม

ในงานวิจัยฉบับนี้ จะเลือกใช้การหมุนแกนปัจจัยแบบมุมฉากในการวิเคราะห์ข้อมูล ดังนั้นในรายละเอียดของวิธีการ จะไม่สนใจที่การหมุนแกนแบบมุมแหลม

แสดงสมการของการหมุนแกนแบบมุมฉากได้ดังสมการต่อไปนี้

$$L_{new} = L_{old}T \tag{52}$$

จากสมการ T คือ เมตริกซ์ของการย้ายแกน(Transition matrix) L_{old} คือน้ำหนักปัจจัยก่อนการหมุนแกน L_{new} คือน้ำหนักปัจจัยหลังการหมุนแกน

จากสมการจะเห็น T คือพารามิเตอร์ที่ต้องมีการประมาณค่าซึ่งมี 3 วิธีการ คือ วาริแมกซ์ (Varimax), อีคิว์แมกซ์ (Equimax) และ ค้วร์ติแมกซ์ (Quartimax) ซึ่งวิธีที่นิยมกันมากที่สุดคือ วาริแมกซ์ วิธีการนี้พยายามที่จะลดจำนวนตัวแปรที่มีน้ำหนักปัจจัยมากบนแต่ละปัจจัยให้เหลือน้อยที่สุด

ในการวิเคราะห์ข้อมูล ฟังก์ชันที่ใช้สำหรับการหมุนแกนจะเป็นฟังก์ชันสำเร็จรูปที่มีอยู่แล้วในคอมพิวเตอร์ ในส่วนของรายละเอียดของการคำนวณจึงไม่กล่าวถึง

4.1.6 การสร้างคะแนนปัจจัย (Factor Score)

จากโมเดลของการวิเคราะห์ปัจจัย ในสมการ (33) พารามิเตอร์ที่จะต้องประมาณค่า หลังจากทำการสกัดปัจจัยเพื่อให้ได้ค่าน้ำหนักปัจจัยแล้ว นั้นก็คือ คะแนนปัจจัยซึ่งใช้สัญลักษณ์ F

วิธีการที่ได้รับการยอมรับและนำมาใช้ในการหาคะแนนปัจจัย มีอยู่ 2 วิธีคือ วิธีกำลังสองน้อยที่สุดแบบถ่วงน้ำหนัก (The Weighted Least Squares Method) และ วิธีการวิเคราะห์ถดถอย (The Regression Method)

1) วิธีกำลังสองน้อยที่สุดแบบถ่วงน้ำหนัก หรือวิธีการของบาร์ทเลทท์ (Bartlett Method)

จากโมเดลการวิเคราะห์ปัจจัยในสมการ (33) เพื่อจะหาคะแนนปัจจัย จึงเริ่มด้วยการพิจารณาผลรวมกำลังสอง ของปัจจัยเฉพาะหรือค่าผิดพลาด (error) ซึ่งใช้สัญลักษณ์ ε และ ถ่วงน้ำหนักด้วยความแปรปรวนเฉพาะ ψ ดังสมการ (53)

$$\sum_{i=1}^{p}\frac{\varepsilon_i^2}{\psi_i}=\varepsilon'\psi^{-1}\varepsilon=(x-\mu-Lf)'\psi^{-1}(x-\mu-Lf) \tag{53}$$

ด้วยการหาค่าต่ำสุด(Minimization) ของผลรวมกำลังสองในสมการ (53) จะทำให้ประมาณคะแนนปัจจัย f ได้ดังสมการ (54)

$$f=(L'\psi^{-1}L)^{-1}L'\psi^{-1}(x-\mu) \tag{54}$$

กำหนดให้ $\mu=\bar{x}$ โดยที่ $\bar{x}$ เป็นเวกเตอร์ค่าเฉลี่ยของข้อมูลโดยมีขนาดเป็น $p\times 1$ และ x_j เป็นเวกเตอร์ข้อมูลในตัวอย่างที่ j มีขนาด $p\times 1$ และ f_j เป็นเวกเตอร์คะแนนปัจจัยของตัวอย่างที่ j มีขนาด $m\times 1$ โดยที่ $j=1,2,...,n$ จะหาคะแนนปัจจัยของแต่ละข้อมูลตัวอย่างได้ดังสมการ (55)

$$f_j=(L'\psi^{-1}L)^{-1}L'\psi^{-1}(x_j-\bar{x}) \tag{55}$$

จากสมการในการหาคะแนนปัจจัยดังที่กล่าวมาข้างต้น หากในการวิเคราะห์ข้อมูลใช้วิธีการประมาณค่าความเป็นไปได้สูงสุดในการสกัดปัจจัย จะเขียนสมการในการหาคะแนนปัจจัยขึ้นใหม่ได้ดังนี้

$$f_j=\Delta^{-1}L'\psi^{-1}(x_j-\bar{x}) \tag{56}$$

โดยที่ Δ เป็นเมตริกซ์ทแยงจากสมการ(48)

สำหรับการวิเคราะห์ข้อมูลที่ใช้วิธีวิเคราะห์องค์ประกอบหลักในการสกัดปัจจัย จะเขียนสมการในการหาคะแนนปัจจัยได้ดังนี้

$$f_j = (L'L)^{-1} L'(x_j - \bar{x}) \tag{57}$$

ในการวิเคราะห์ข้อมูล บางครั้งเมตริกซ์ความแปรปรวนร่วมที่ใช้ในการหาปัจจัย อยู่ในรูปของเมตริกซ์สหสัมพันธ์ (Correlation Matrix) สูตรการคำนวณเพื่อหาคะแนนปัจจัยจึงต้องปรับใหม่ โดยการแทนที่ $x_j - \bar{x}$ ด้วย z_j ซึ่ง

$$z_j = \frac{x_j - \bar{x}}{\sqrt{\sigma}} \tag{58}$$

จากสมการ (58) σ คือเวกเตอร์ความแปรปรวนของข้อมูลซึ่งมีขนาดเท่ากับเวกเตอร์ค่าเฉลี่ย

2) วิธีการวิเคราะห์ถดถอย

สมการในการหาคะแนนปัจจัยโดยวิธีนี้แสดงได้ดังนี้

$$f_j = L'S^{-1}(x_j - \bar{x}) \tag{59}$$

จากสูตร S คือเมตริกซ์ความแปรปรวนร่วม (Covariance Matrix) ถ้าการวิเคราะห์ใช้เมตริกซ์สหสัมพันธ์ R สมการในการหาคะแนนปัจจัยจะแสดงได้ดังนี้

$$f_j = L'R^{-1}z_j \tag{60}$$

4.2 การวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์

เพื่อที่จะให้มองเห็นประโยชน์และข้อจำกัด ของการวิเคราะห์ปัจจัยกับข้อมูลดีเอ็นเอไมโครอาร์เรย์ จึงนำเสนอ การประยุกต์ใช้วิธีการวิเคราะห์ปัจจัย ในหลายๆ ลักษณะ ดังนี้

4.2.1 การวิเคราะห์ปัจจัยเพื่อวิเคราะห์โครงสร้างของตัวแปรและการนำเสนอข้อมูล ในชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิส เซอริวิสิเอ

- ปัญหา และ วัตถุประสงค์ของการวิเคราะห์

ข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิส เซอริวิสิเอ เป็นข้อมูลที่ได้จากการวัดค่าการแสดงออกของยีนระหว่างเกิดกระบวนการไดอ๊อคซิคชิฟท์ มีช่วงเวลาเป็นตัวแปร สำหรับการวัดผล ทั้งนี้ในบางช่วงเวลา จะให้ค่าการแสดงออกของยีนที่มีลักษณะ คล้ายกับช่วงเวลาอื่นๆ ความสัมพันธ์กันระหว่างตัวแปรบางครั้งเรียกว่าปัจจัย ซึ่งจะต้องค้นหา เพื่ออธิบายโครงสร้างภายในของตัวแปรเหล่านี้ และเพื่อการลดจำนวนตัวแปรที่มีความสัมพันธ์กันให้อยู่ในลักษณะของมิติข้อมูลที่น้อยลง อันจะมีประโยชน์ต่อการนำเสนอข้อมูล

- แหล่งข้อมูลและลักษณะของข้อมูล

ชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิส เซอริวิสิเอ เป็นชุดข้อมูลเดียวกับชุดข้อมูลที่อยู่ใน บทที่ 3 หัวข้อ 3.1.1

- วิธีการวิเคราะห์

1) กำหนดให้ตัวแปรของข้อมูลคือช่วงเวลา (Time) ซึ่งมีจำนวน 7 ช่วงเวลา และตัวอย่างข้อมูลคือ ยีน ซึ่งมีจำนวนทั้งสิ้น 6,153 ยีน กรองข้อมูลโดยกรองเอายีนที่ข้อมูลขาดหาย ออกไป เหลือยีน 6,119 ยีน

2) ใช้สมการ (51) หาจำนวนปัจจัยตั้งต้น

3) สกัดปัจจัยเพื่อหาค่าน้ำหนักปัจจัย โดยใช้วิธีวิเคราะห์ปัจจัยร่วม เนื่องจากจุดประสงค์ของการวิเคราะห์คือ การวิเคราะห์โครงสร้างของตัวแปร

4) หาความแปรปรวนของแต่ละปัจจัย เพื่อใช้ในการเลือกปัจจัย และกำหนดจำนวนปัจจัยที่เหมาะสม

5) หมุนแกนปัจจัยแบบมุมฉากโดยวิธี วาริแมกซ์

6) วิเคราะห์โครงสร้างของตัวแปร จากค่าน้ำหนักปัจจัย ค่าร่วมกัน ค่าความแปรปรวนเฉพาะ และค่าความแปรปรวนของแต่ละปัจจัย

7) หาคะแนนปัจจัยโดยวิธีการของบาร์ทเลทท์ เพื่อดูลักษณะการกระจายตัวของยีน กับ ปัจจัยใน 2 มิติ

- ผลการวิเคราะห์

การพิจารณาจำนวนปัจจัยในขั้นต้นนั้น พบว่า จำนวนปัจจัยต้องมีค่า น้อยกว่า 4 ปัจจัย ดังนั้น เมื่อพิจารณาที่ผลของการสกัดปัจจัย เมื่อกำหนดจำนวนปัจจัย ด้วย 1 ปัจจัย จะได้อัตราส่วนของผลรวมของค่าความแปรแปรปรวน จะมีค่าเท่ากับร้อยละ 34.71 ของความแปรปรวนทั้งหมด และเมื่อกำหนดจำนวนปัจจัย ด้วย 2 ปัจจัย อัตราส่วนของผลรวมของค่าความแปรแปรปรวนจะมีค่าเท่ากับร้อยละ 59 ของความแปรปรวนทั้งหมด และสุดท้ายเมื่อกำหนดจำนวนปัจจัย ด้วย 3 ปัจจัย อัตราส่วนของผลรวมของค่าความแปรแปรปรวนจะมีค่าเท่ากับร้อยละ 69.2 ของความแปรปรวนทั้งหมด จากการพิจารณาค่าความแปรปรวนดังกล่าว จึงพิจารณาจำนวนปัจจัยที่ 3 ปัจจัย ซึ่งมีค่าความแปรปรวนสะสมเป็น 69.2 เปอร์เซ็นต์

หลังจากสกัดปัจจัยและหมุนแกนปัจจัย จะแสดงค่าน้ำหนักปัจจัย ค่าความแปรปรวนของแต่ละปัจจัย ค่าร่วมกัน (Communality) และค่าความแปรปรวนเฉพาะ (Uniqueness) ได้ดัง ตาราง 4.1

ตาราง 4.1 ผลการวิเคราะห์ปัจจัยโดยวิธีวิเคราะห์ปัจจัยร่วมที่หมุนแกนปัจจัยแบบวาริแมกซ์กับชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิสเซอริวิสิเอโดยใช้ช่วงเวลาเป็นตัวแปร

	Factor1	Factor2	Factor3	Communality	Uniqueness
Time 1	0.160	-0.265	**0.648**	0.515	0.485
Time 2	0.153	-0.023	**0.887**	0.811	0.189
Time 3	**0.534**	0.104	0.165	0.323	0.677
Time 4	**0.928**	0.263	0.081	0.936	0.064
Time 5	**0.690**	0.280	0.154	0.578	0.422
Time 6	0.394	**0.778**	-0.225	0.811	0.189
Time 7	0.256	**0.889**	-0.123	0.870	0.130
Var.	1.891	1.624	1.329	4.844	2.156
Proportion Var.	0.27	0.232	0.19	0.692	0.308

ผลจากตาราง 4.1 แสดงให้เห็นว่าปัจจัยที่ 1 มีความสัมพันธ์กับช่วงเวลาที่ 3, 4, 5 มาก เนื่องจากมีค่าน้ำหนักปัจจัยที่สูง ส่วนในปัจจัยที่ 2 จะมีความสัมพันธ์กับช่วงเวลาที่ 6 และ 7 มาก และปัจจัยที่ 3 จะมีความสัมพันธ์กับ ช่วงเวลาที่1 กับ 2 มาก ซึ่งจากค่าร่วมกันที่ได้แสดงให้เห็นว่าช่วงเวลาที่ 2, 4, 6 และ 7 มีความเกี่ยวพันธ์กับปัจจัยทั้งสามปัจจัยนี้มาก ส่วนช่วงเวลาที่ 1, 3 และ 5 เป็นตัวแปรที่มีความเฉพาะกับตัวมันเองมาก โดยเฉพาะ ในช่วงเวลาที่ 3 มีค่าความแปรปรวนเฉพาะถึง 0.677 แสดงให้เห็นว่าตัวแปรตัวนี้ อาจเป็นปัจจัยเฉพาะที่ไม่ได้อยู่ร่วมกับตัวแปรอื่นๆ

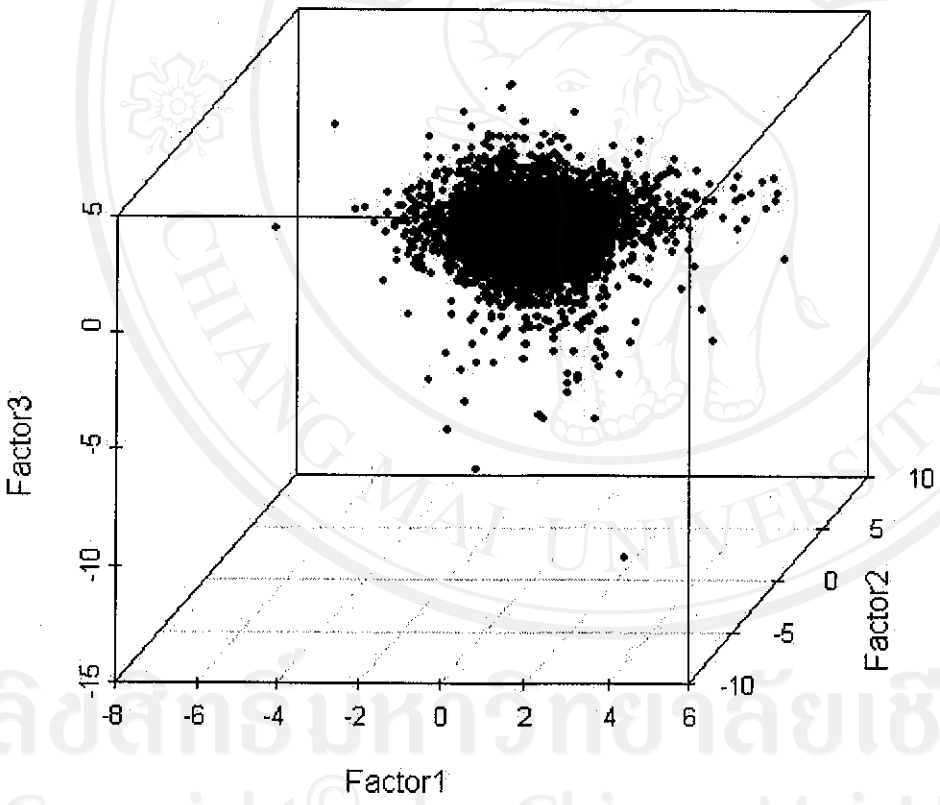
จากผลการวิเคราะห์ปัจจัยนี้เอง เมื่อพิจารณาในกระบวนการไดอ๊อคซิคชิฟท์(Diauxic Shift) จากช่วงเวลาหนึ่งไปยังช่วงเวลาหนึ่งนั้น อาจสรุปได้ว่า

ในช่วงเวลาที่ 1 และ 2 เป็นช่วงเวลาที่มีผลต่อการทำงานของยีนหรือการแสดงออกของยีนที่เหมือนกัน นั่นคือช่วงเวลาทั้งสองช่วงเวลานี้ มีปัจจัยที่ส่งผลต่อการทำงานของยีนเหมือนๆ กันซึ่งจากการทดลองช่วงเวลาดังกล่าว จะเป็นช่วงที่ยีสต์มีสารอาหารซึ่งได้แก่น้ำตาลกลูโคสอุดมสมบูรณ์

ในช่วงเวลาที่ 3, 4, 5 เมื่อพิจารณาที่การทำงานของยีน ช่วงเวลาดังกล่าวเป็นช่วงเวลาที่การทำงานในกระบวนการไดอ๊อคซิคชิฟท์อยู่ในช่วงเวลากลางๆ ซึ่งจะเป็นช่วงที่น้ำตาลกลูโคสเริ่มลดลง

ในช่วงเวลาที่ 6, 7 ปัจจัยดังกล่าวคือช่วงเวลาที่สารอาหารคือน้ำตาลกลูโคสหมด จึงต้องเปลี่ยนไปสังเคราะห์เอทานอล เป็นสารอาหารแทน

จากปัจจัยที่ได้ เมื่อนำมาหาค่าคะแนนปัจจัย จะทำให้ลักษณะการกระจายตัวของกลุ่มยีน แสดงใน 3 มิติ ดังรูป 4.6



รูป 4.6 แผนภาพการกระจายของค่าคะแนนปัจจัยโดยวิธี วิเคราะห์ปัจจัยร่วมที่หมุนแกนปัจจัยแบบวาริแมกซ์ กับชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิส เซอริวิสิเอ โดยใช้ช่วงเวลาเป็นตัวแปร

จากแผนภาพ ในรูป 4.6 จุดแต่จุดในกราฟ แทนค่าคะแนนปัจจัยของยีนแต่ละตัว ทั้ง 6,119 ยีน ผลที่ได้แสดงให้เห็นว่า ข้อมูลการแสดงออกของยีนชุดนี้ สามารถนำเสนอใน 3 มิติได้ด้วยความแปรปรวน 69.2 เปอร์เซ็นต์ ซึ่งจากรูปจะเห็นว่ายีนมีการกระจายตัวกันอย่างไม่ระเบียบ ยีนส่วนใหญ่มีการกระจุกตัวกันอยู่ตรงกลาง จึงไม่สามารถให้ความหมายของข้อมูลเหล่านี้ได้

- สรุปผลการวิเคราะห์

การวิเคราะห์ปัจจัยสามารถที่จะวิเคราะห์โครงสร้างของตัวแปรในข้อมูลดีเอ็นเอไมโครอาร์เรย์ ซึ่งในที่นี้ เป็นช่วงเวลาในกระบวนการไดอ๊อคซิคชิฟท์ของยีสต์ ทำให้เราสามารถที่จะหาความสัมพันธ์ของตัวแปรดังกล่าวออกมาเป็นลักษณะเชิงตัวเลข ซึ่งจะช่วยในการเปรียบเทียบความสัมพันธ์ระหว่างตัวแปรและทำให้สามารถตีความหมายของตัวแปรเหล่านี้ได้ พร้อมกันนั้นจากปัจจัยดังกล่าวยังสามารถที่จะใช้ในการนำเสนอข้อมูลให้อยู่ในรูปที่มนุษย์สามารถที่จะรับรู้ได้ ใน 2 หรือ 3 มิติ

4.2.2 การหาปัจจัยที่มีผลต่อกระบวนการไดอ๊อคซิคชิฟท์ ในชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิสเซอริวิสิเอ

- ปัญหา และวัตถุประสงค์ของการวิเคราะห์

ข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิสเซอริวิสิเอ ได้จากการวัดค่าการแสดงออกของยีนในกระบวนการไดอ๊อคซิคชิฟท์ โดยมียีนที่เกี่ยวข้องกับกระบวนการดังกล่าวมากมาย ยีนแต่ละตัวจะมีคุณสมบัติที่สามารถอธิบายได้ด้วยยีนออนโทโลยี และความสัมพันธ์ของยีนต่างๆสามารถอธิบายด้วยปัจจัย ซึ่งค่าการแสดงออกของยีน คุณสมบัติของยีน รวมทั้งความสัมพันธ์ของยีนเหล่านี้เป็นสิ่งที่ช่วยในการอธิบายกลไกการทำงานในกระบวนการไดอ๊อคซิคชิฟท์ได้ ดังนั้นงานวิจัยจึงทำการวิเคราะห์ความสัมพันธ์ของยีนให้อยู่ในรูปของปัจจัย เพื่อหาปัจจัยที่ส่งผลต่อกระบวนการไดอ๊อคซิคชิฟท์ โดยอาศัยยีนออนโทโลยีในการอธิบายความหมาย ของปัจจัยดังกล่าว

- แหล่งข้อมูลและลักษณะของข้อมูล

ชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิสเซอริวิสิเอ เป็นชุดข้อมูลเดียวกับชุดข้อมูลที่อยู่ใน บทที่ 3 หัวข้อ 3.1.1 และ ยีนออนโทโลยีที่อธิบายคุณสมบัติของยีนในชุดข้อมูลนี้ มีรายละเอียดกล่าวถึงในบทที่ 3 ใน หัวข้อ 3.1.1 การทดลองที่ 2

- วิธีการวิเคราะห์

1) กำหนดให้ตัวแปรของข้อมูลคือยีน (Genes) ซึ่งมีจำนวนทั้งสิ้น 6,153 ยีน กรองข้อมูลขั้นต้นโดยตัดเอายีนที่ข้อมูลขาดหาย ออกไป เหลือเพียง 6,119 ยีน

2) กรองข้อมูลขั้นต่อไป โดยเลือกยีนที่ให้ค่าการแสดงออกแตกต่างกันมากในทุกๆช่วงเวลาไว้ใช้วิเคราะห์ ซึ่งยีนดังกล่าวนี้จะเป็นยีนที่มีความหมายสำหรับการวิเคราะห์ ค่าความแตกต่างของยีนแต่ละตัวในทุกๆ ช่วงเวลานี้จะเรียกว่าค่าความแปรปรวน วิธีกรองข้อมูลนี้ทำได้โดยการ ตัดเอายีนที่มีความแปรปรวนของข้อมูลน้อยๆ ออกไป นั่นคือทำการหาความแปรปรวนของยีนที่มีต่อช่วงเวลาทั้ง 7 ช่วงเวลา เลือกยีนที่มีความแปรปรวนสูงที่สุดเป็นลำดับต้นๆ โดยยีนที่เลือก กำหนดให้มีจำนวน 10 เปอร์เซ็นต์ ของจำนวนยีนทั้งหมด สำหรับนำไปวิเคราะห์ นั่นคือ จากยีน 6,119 ยีน จะเลือกยีนที่มีความแปรปรวนสูงที่สุดจำนวน 612 ยีน ไปวิเคราะห์

3) สกัดปัจจัยเพื่อหาค่าน้ำหนักปัจจัยโดยใช้วิธีวิเคราะห์องค์ประกอบหลักและใช้เมตริกซ์ความแปรปรวนร่วมเป็นเมตริกซ์สหสัมพันธ์

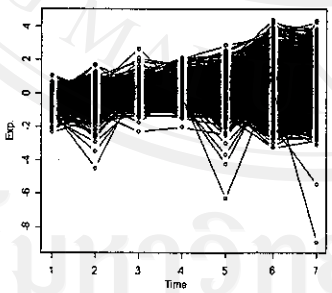
4) เลือกจำนวนปัจจัยที่ 2 ปัจจัย เพื่อที่จะใช้ในการนำเสนอข้อมูลในลักษณะ 2 มิติ

5) หมุนแกนปัจจัยโดยวิธีวาริแมกซ์

6) หาปัจจัยและอธิบายความหมายของปัจจัยที่ได้ โดยการ พิจารณาค่าน้ำหนักปัจจัย เพื่อหายีนที่มี ค่าน้ำหนักปัจจัยมาก 50 ตัวแรกในแต่ละปัจจัยเป็นตัวแทนของปัจจัยนั้นๆ และ หาออนโทโลยีที่อธิบายกลุ่มยีนดังกล่าวจากข้อมูลยีนออนโทโลยีเพื่อให้ความหมายของปัจจัย โดยแยกอธิบายเป็น 3 ออนโทโลยีหลักที่อิสระต่อกันนั่นคือ ฟังก์ชันในระดับโมเลกุล (Molecular Function) กระบวนการทางชีววิทยา (Biological Process) และองค์ประกอบของเซลล์ (Cellular Component)

- ผลการวิเคราะห์

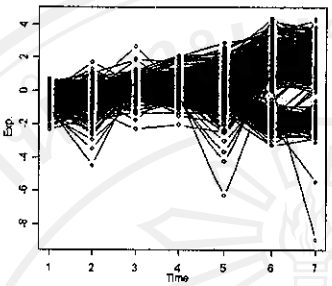
ข้อมูลดีเอ็นเอไมโครอาร์เรย์ทั้ง 6,119 ยีนเมื่อนำมาพล็อตลงในกราฟจะได้ดังรูป 4.7



รูป 4.7 กราฟแสดงค่าการแสดงออกของยีนในกระบวนการไดอ๊อคซิคชิฟของยีสต์
จากข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ ซัคคาโรไมซิส เซอริวิสิเอ

จากรูป 4.7 เส้นกราฟคือยีนแต่ละตัวที่มีค่าการแสดงออก ณ เวลาต่างๆใน 7 ช่วงเวลาของกระบวนไดอ๊อคซิคชิฟท์ จะเห็นว่าในทุกๆ ช่วงเวลายีนบางตัวให้ค่าการแสดงออก ที่เหมือนๆกัน นั่นแสดงว่าความแปรปรวนของข้อมูลในยีนดังกล่าวมีค่าน้อย เนื่องจากยีนที่มีค่าความแปรปรวนน้อยไม่มี

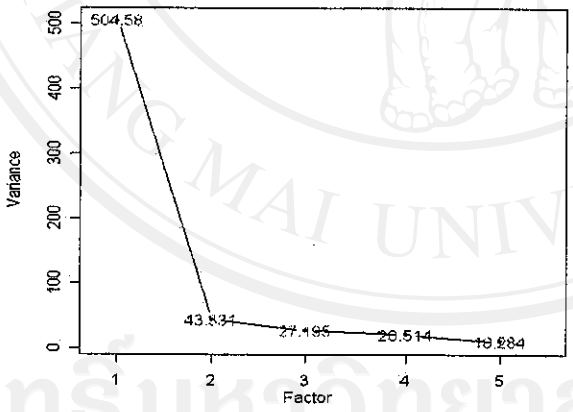
นัยสำคัญต่อผลการวิเคราะห์ การนำยีนที่มีค่าความแปรปรวนน้อยๆ ไปวิเคราะห์จะทำให้ผลการวิเคราะห์ไม่น่าเชื่อถือ ดังนั้นจึงจำเป็นต้องตัดเอายีนที่มีค่าความแปรปรวนน้อยๆ ออกไป และจะแสดงลักษณะการกระจายของข้อมูลในยีนที่เลือกซึ่งเป็นยีนที่มีความแปรปรวนสูงที่สุด ได้ดังกราฟในรูป 4.8



รูป 4.8 กราฟแสดงค่าการแสดงออกของยีนที่มีค่าความแปรปรวนสูงที่สุด ในกระบวนการไดอ๊อคซิค ชิฟท์จากข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ ซัคคาโรไมซิส เซอริวิสิเอ

จากรูป 4.8 ยีนที่เลือกมาจะเป็นยีนที่มีค่าการแสดงออก ในแต่ละช่วงเวลา แตกต่างกัน หรือมีความแปรปรวนมาก ยีนดังกล่าวจึงเหมาะสมกับการนำไปวิเคราะห์

เมื่อนำยีนที่ผ่านกระบวนการกรองข้อมูลมาวิเคราะห์ปัจจัย ภายหลังการสกัดปัจจัย จะแสดงสครีพล็อต อธิบายความแปรปรวนของข้อมูล ที่ 5 ปัจจัยแรก หลังจากสกัดปัจจัยได้ดังนี้



รูป 4.9 สครีพล็อต จากผลการวิเคราะห์ปัจจัยโดยวิธีวิเคราะห์องค์ประกอบหลักกับชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิส เซอริวิสิเอ โดยใช้ยีน (Gene) เป็นตัวแปร

แสดงค่าความแปรปรวนของแต่ละปัจจัย ได้ดังตาราง 4.2

ตาราง 4.2 ความแปรปรวนของปัจจัยจากผลการวิเคราะห์ปัจจัยโดยวิธีวิเคราะห์องค์ประกอบหลักกับชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิส เซอริวิสิเอ โดยใช้ยีนเป็นตัวแปร

ปัจจัย	1	2	3	4	5
ความแปรปรวน	504.58	43.33	27.19	20.51	10.28
อัตราส่วนความแปรปรวน	0.82	0.07	0.04	0.03	0.02
ความแปรปรวนสะสม	504.58	547.91	575.11	595.62	605.90
อัตราส่วนความแปรปรวนสะสม	0.82	0.90	0.94	0.97	0.99

จากรูป 4.9 และตาราง 4.2 จะเห็นว่าปัจจัยที่ 2 มีค่าความแปรปรวนสะสม 90 เปอร์เซ็นต์ และปัจจัยที่ 3 มีความแปรปรวนสะสม 94 เปอร์เซ็นต์ ด้วยค่าความแปรปรวนสะสมที่ 90 เปอร์เซ็นต์ ก็เพียงพอจะเป็นตัวแทนของข้อมูลเดิมได้แล้วจึงเลือกจำนวนปัจจัย ที่ 2 ปัจจัย และภายหลังจากสกัดปัจจัยจะแสดงตัวอย่างของเมตริกซ์ค่าน้ำหนักปัจจัย ก่อนที่จะหมุนแกน ได้ดังตาราง 4.3

ตาราง 4.3 ตัวอย่างเมตริกซ์ค่าน้ำหนักปัจจัยจากผลการวิเคราะห์ปัจจัยโดยวิธีวิเคราะห์องค์ประกอบหลัก แบบไม่หมุนแกนปัจจัยกับชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิส เซอริวิสิเอ โดยใช้ยีนเป็นตัวแปร

Gene	Factor 1	Factor 2	Gene	Factor 1	Factor 2
YJL009W	-0.028	-0.099	YKR097W	0.033	0.084
YJL216C	0.017	0.055	YGR088W	0.043	-0.016
YGL184C	0.031	0.017	YFL014W	0.042	-0.030
YGR225W	0.029	-0.067	YLR377C	0.034	0.082
YGR043C	0.043	-0.016	YKL187C	0.040	0.062
YKL217W	0.043	0.025	YML128C	0.043	-0.025
YDR398W	-0.039	-0.055	YOR215C	0.043	-0.023
YNL194C	0.043	0.003	YDR171W	0.044	-0.007
YGR236C	0.042	0.021	YBL015W	0.043	0.029
YDL204W	0.041	0.005	YLR327C	0.043	-0.032
YAL054C	0.041	0.050	YER150W	0.043	-0.030
YIL136W	0.037	-0.065	YNL200C	0.042	-0.045
YGR243W	0.043	0.004	YER024W	0.038	0.077
YGR248W	0.041	-0.054	YER065C	0.032	0.087
YLR174W	0.038	0.043	YLR312C	0.041	-0.011
YCR021C	0.041	-0.029	YML054C	0.044	0.022
YJR095W	0.040	0.047	YBR132C	0.040	-0.026
YNL036W	0.032	0.063	YJL089W	0.040	0.043
YKL026C	0.044	-0.010	YBR072W	0.043	-0.026
YFR015C	0.040	-0.050	YDR219C	0.031	0.024
YMR206W	0.042	0.027	YBL064C	0.044	-0.009
YBL049W	0.042	0.035	YMR250W	0.039	-0.044
YGL138C	0.032	0.022	YAL034C	0.038	0.057
YMR114C	0.036	-0.043	YMR170C	0.043	-0.020
YLR149C	0.044	0.008	YIR039C	0.038	-0.044

จากเมตริกซ์ของค่าน้ำหนักปัจจัย ทำการหมุนแกนปัจจัยด้วยวิธีวาริแมกซ์ ได้เมตริกซ์ของค่าน้ำหนักปัจจัยใหม่ และเพื่อที่จะตีความหมายของปัจจัย จะเลือกยีน ที่มีค่าน้ำหนักปัจจัยสูงที่สุด 50 ตัวแรก ในทุกๆ ปัจจัยมาใช้ในการอธิบายความหมาย ซึ่งเมตริกซ์ค่าน้ำหนักปัจจัยของยีนดังกล่าว แสดงได้ดังตาราง 4.4 และ ตาราง 4.5

ตาราง 4.4 เมตริกซ์ค่าน้ำหนักปัจจัยของ 50 ยีนที่มีค่าน้ำหนักปัจจัย ณ ปัจจัยที่ 1 สูงที่สุดจากการวิเคราะห์ปัจจัยโดยวิธีวิเคราะห์องค์ประกอบหลักกับชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิสเซอริวิสิเอ โดยใช้ยีนเป็นตัวแปร

ลำดับ	ยีน	ปัจจัย 1	ปัจจัย 2	ลำดับ	ยีน	ปัจจัย 1	ปัจจัย 2
1	YNL061W	-0.103	0.089	26	YDL124W	0.073	-0.020
2	YDR101C	-0.103	0.070	27	YER061C	0.073	-0.027
3	YFR053C	0.096	-0.059	28	YGL062W	-0.072	0.101
4	YOR095C	-0.095	0.066	29	YKL103C	0.072	-0.018
5	YKL191W	-0.090	0.043	30	YIL136W	0.072	-0.020
6	YBR155W	-0.089	0.044	31	YDL021W	0.072	-0.017
7	YIL111W	0.088	-0.039	32	YKL142W	0.071	-0.017
8	YCL042W	0.085	-0.045	33	YPL036W	-0.070	0.032
9	YPR136C	-0.083	0.034	34	YMR090W	0.070	-0.014
10	YMR018W	0.080	-0.029	35	YJR096W	0.070	-0.014
11	YCL040W	0.079	-0.034	36	YCL060C	0.069	-0.027
12	YPL221W	-0.079	0.046	37	YJL109C	-0.068	0.013
13	YBR218C	-0.078	0.101	38	YPL123C	0.068	-0.013
14	YHL022C	0.077	-0.025	39	YGR225W	0.068	-0.027
15	YDL037C	-0.077	0.036	40	YAL061W	0.068	-0.017
16	YNL174W	-0.077	0.027	41	YGR248W	0.068	-0.010
17	YLR252W	0.077	-0.024	42	YLR355C	-0.068	0.010
18	YDR516C	0.077	-0.023	43	YJR021C	0.067	-0.015
19	YPL017C	-0.076	0.103	44	YGL047W	0.067	-0.011
20	YHR215W	-0.076	0.021	45	YDR165W	-0.066	0.009
21	YDR342C	0.076	-0.024	46	YJR073C	0.066	-0.010
22	YJR130C	0.074	-0.022	47	YBR052C	0.066	-0.009
23	YBR183W	0.074	-0.018	48	YGL212W	0.065	-0.009
24	YIL098C	0.073	-0.022	49	YDR130C	0.065	-0.019
25	YHR087W	0.073	-0.019	50	YJL164C	0.065	-0.012

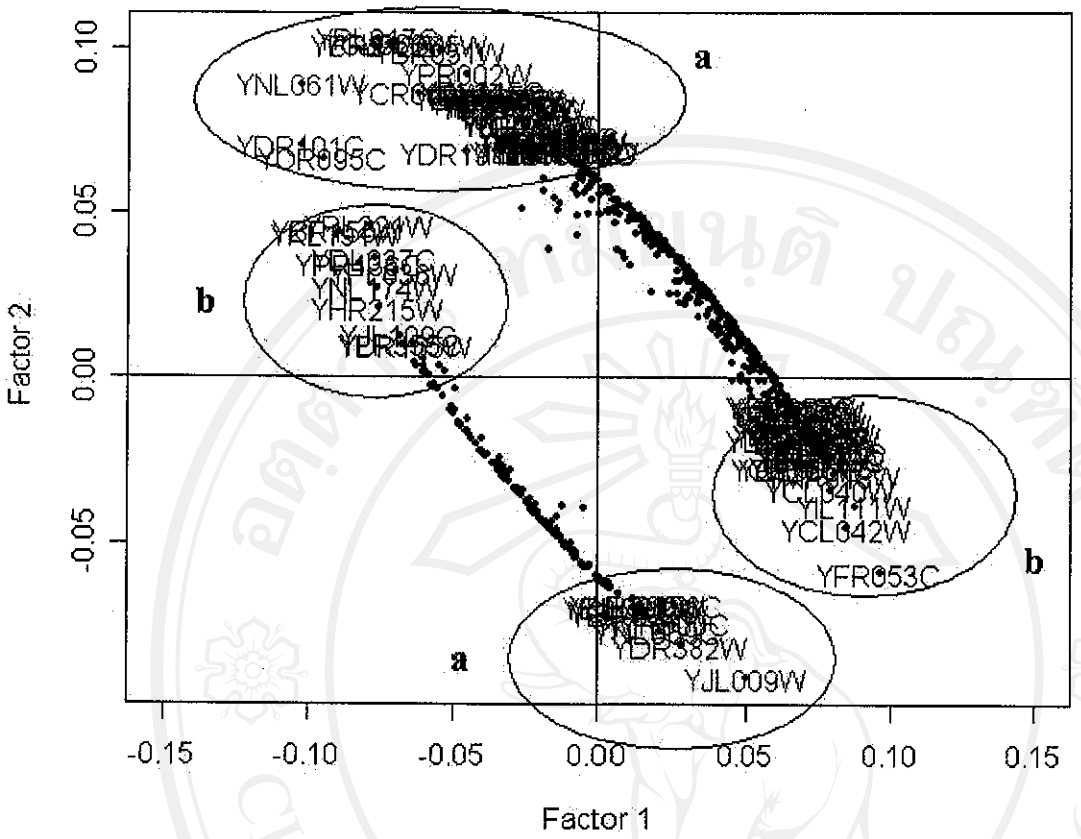
ตาราง 4.4 แสดงยีน และ ค่าน้ำหนักปัจจัยของยีนจำนวน 50 ยีน ที่มีค่าน้ำหนักปัจจัย ในปัจจัยที่ 1 มากที่สุด จะเห็นว่า ค่าน้ำหนักปัจจัยที่มีค่ามากที่สุดในปัจจัยที่ 1 คือ -0.103 ทั้งนี้เครื่องหมายไม่ได้เป็นตัวบอกขนาดของค่าน้ำหนักปัจจัย นั่นคือในการคำนวณเปรียบเทียบค่าดังกล่าวจะใช้ค่าสัมบูรณ์ (Absolute)

ตาราง 4.5 เมตริกซ์ค่าน้ำหนักปัจจัยของ 50 ยีนที่มีค่าน้ำหนักปัจจัย ณ ปัจจัยที่ 2 สูงที่สุด จากการวิเคราะห์ปัจจัยโดยวิธีวิเคราะห์องค์ประกอบหลักกับชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิสเซอริวิสิเอ โดยใช้ยีนเป็นตัวแปร

ลำดับ	ยีน	ปัจจัย 1	ปัจจัย 2	ลำดับ	ยีน	ปัจจัย 1	ปัจจัย 2
1	YPL017C	-0.076	0.103	26	YNL117W	-0.024	0.077
2	YGL062W	-0.072	0.101	27	YIL057C	-0.021	0.076
3	YPL265W	-0.060	0.101	28	YCR005C	-0.020	0.076
4	YBR218C	-0.078	0.101	29	YMR300C	0.022	-0.074
5	YBR051W	-0.055	0.099	30	YDL085W	-0.019	0.073
6	YPR002W	-0.046	0.092	31	YLR142W	-0.019	0.073
7	YJL009W	0.050	-0.091	32	YKL187C	-0.015	0.073
8	YNL061W	-0.103	0.089	33	YIL125W	-0.017	0.072
9	YDL215C	-0.040	0.087	34	YGR110W	-0.015	0.072
10	YCR062W	-0.062	0.086	35	YBR032W	0.014	-0.072
11	YDL215C	-0.039	0.086	36	YLR267W	-0.013	0.071
12	YKL044W	-0.033	0.085	37	YPL262W	-0.014	0.071
13	YER065C	-0.038	0.085	38	YGL102C	0.012	-0.070
14	YBR296C	-0.042	0.085	39	YDR101C	-0.103	0.070
15	YGR067C	-0.032	0.084	40	YIL146C	-0.014	0.070
16	YKR097W	-0.035	0.083	41	YBL027W	0.011	-0.069
17	YDL223C	-0.031	0.083	42	YLR300W	0.014	-0.069
18	YER096W	-0.030	0.082	43	YDR018C	-0.020	0.069
19	YLR377C	-0.033	0.082	44	YMR118C	-0.010	0.069
20	YBR117C	-0.027	0.082	45	YLR058C	0.021	-0.069
21	YER024W	-0.027	0.081	46	YNL195C	-0.013	0.069
22	YDR382W	0.027	-0.081	47	YPR030W	-0.010	0.068
23	YJL088W	-0.028	0.079	48	YDR191W	-0.046	0.068
24	YDR536W	-0.024	0.077	49	YBR116C	-0.010	0.068
25	YNL069C	0.020	-0.077	50	YNL036W	-0.021	0.068

ตาราง 4.5 แสดง ยีน และ ค่าน้ำหนักปัจจัยของยีนจำนวน 50 ยีน ที่มีค่าน้ำหนักปัจจัย ในปัจจัยที่ 2 มากที่สุด จะเห็นว่า ค่าน้ำหนักปัจจัยที่มีค่ามากที่สุดในปัจจัยที่ 2 คือ 0.103 และจากทั้งสองตารางจะสังเกตเห็นว่ายีน YNL061W และ YDR101C มีค่าน้ำหนักปัจจัยสูงในทั้ง 2 ปัจจัย ซึ่งจะทำให้ค่าร่วมกันของทั้ง 2 ยีน ที่มีต่อปัจจัยทั้ง 2 ปัจจัยมีค่าสูงตามไปด้วย นอกจากนี้ เรายังไม่สามารถใช้ยีนทั้ง 2 ตัว ในการแยกความแตกต่างของปัจจัยทั้ง 2 ปัจจัยนี้ได้ นั่นแสดงว่าคุณสมบัติหรือออนโทโลยีของยีนดังกล่าวใช้อธิบายปัจจัยทั้งสองปัจจัยนี้เหมือนๆ กัน

นำเมตริกซ์ของค่าน้ำหนักปัจจัย (Loadings Matrix) หลังจากการหมุนแกนปัจจัย ทั้ง 612 ยีน มาพล็อตในกราฟ 2 มิติ เพื่อสังเกตการกระจายตัวของกลุ่มยีนจากตาราง 4.4 และ 4.5 ซึ่งเป็นยีนที่มีค่าน้ำหนักปัจจัยในแต่ละปัจจัยมากที่สุด แสดงกราฟดังกล่าว ดังรูป 4.10



รูป 4.10 กราฟการกระจายตัวของค่าน้ำหนักปัจจัยใน 2 ปัจจัย จากผลการวิเคราะห์ปัจจัยโดยวิธีวิเคราะห์องค์ประกอบหลักภายหลังหมุนแกนปัจจัยแบบวาริแมกซ์กับชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ ซัคคาโรไมซิส เซอริวิสิเอ โดยใช้ยีนเป็นตัวแปร

จากรูป 4.10 เมื่อนำค่าน้ำหนักปัจจัยของยีนทั้ง 612 ยีนที่ 2 ปัจจัยแรก มาพล็อตลงในกราฟ 2 มิติ โดยให้แต่ละจุดคือคู่อันดับของค่าน้ำหนักปัจจัยของยีน ที่สัมพันธ์กับปัจจัยที่ 1 และปัจจัยที่ 2 เมื่อพล็อตค่าน้ำหนักปัจจัย ของทุกๆยีนลงไปในกราฟ จะเห็นว่ากลุ่มของยีนที่มีค่าน้ำหนักปัจจัยในปัจจัยที่ 1 สูงที่สุดจำนวน 50 ยีนนั้น จะเกาะกลุ่มกันอยู่ติดกับแกนของปัจจัยที่ 1 ส่วนกลุ่มของยีนที่มีค่าน้ำหนักปัจจัยในปัจจัยที่ 2 สูงที่สุดจำนวน 50 ยีน จะเกาะกลุ่มกันอยู่ติดกับแกนของปัจจัยที่ 2 ซึ่งจากกราฟกลุ่มของยีนที่มีแนวโน้มความสัมพันธ์กับปัจจัยที่ 1 มากที่สุดคือกลุ่มยีน b ส่วนกลุ่มของยีนที่มีแนวโน้มความสัมพันธ์กับปัจจัยที่ 2 มากที่สุดคือกลุ่มยีน a นั่นเอง

จากกลุ่มยีนที่ได้ใน ตาราง 4.4 และ ตาราง 4.5 จะเป็นกลุ่มยีน ที่มีค่าน้ำหนักปัจจัยที่สูง ในปัจจัยที่ 1 และ ปัจจัยที่ 2 ตามลำดับ และเนื่องจากค่าน้ำหนักปัจจัยเป็นค่าที่อธิบายความสัมพันธ์ระหว่างตัวแปรและปัจจัย ซึ่งตัวแปรที่มีค่าน้ำหนักปัจจัยที่สูงกับปัจจัยใดปัจจัยหนึ่ง จะหมายถึงตัวแปร

ดังกล่าวสามารถอธิบายได้ด้วยปัจจัยนั้นได้ดี ทั้งนี้หากมองในทางกลับกัน การอธิบายความหมายของปัจจัยก็จะต้องอาศัยคุณสมบัติของตัวแปรที่มีความสัมพันธ์กันสูง ในการให้ความหมายด้วยเช่นเดียวกัน ดังนั้น จากกลุ่มของยีนดังกล่าว จึงอาศัยคุณสมบัติของยีนเหล่านี้ในการให้ความหมายของปัจจัย ซึ่งคุณสมบัติของยีนในความหมายของงานวิจัยนี้ก็คือ ยีนออนโทโลยี

เนื่องจากยีนออนโทโลยีมีการจำแนกการอธิบายคุณสมบัติของยีนออกเป็น 3 ลักษณะหรือ 3 ออนโทโลยีหลักที่เป็นอิสระต่อกัน ดังนั้นผลการวิเคราะห์จึงต้องอธิบายปัจจัย โดยแยกอธิบายออกเป็น 3 ยีนออนโทโลยีหลักด้วยเช่นเดียวกัน ซึ่งผลการวิเคราะห์แสดงได้ดังตาราง 4.6 และ ตาราง 4.7 ดังนี้

ตาราง 4.6 ยีนออนโทโลยี ของ 50 ยีนที่มีค่าน้ำหนักปัจจัย ณ ปัจจัยที่ 1 สูงที่สุดจากการวิเคราะห์ปัจจัยโดยวิธีวิเคราะห์องค์ประกอบหลักกับชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิสเซอริวิสิเอ โดยใช้ยีนเป็นตัวแปร

Molecular Function	**Biological Process**	**Cellular Component**
transferase activity	RNA metabolism	nucleolus
isomerase activity	ribosome biogenesis and assembly	cytoplasm
chaperone activity	carbohydrate metabolism	nucleus
oxidoreductase activity	protein modification	mitochondrial membrane
ligase activity	cellular respiration	bud
hydrolase activity	DNA metabolism	chromosome
transporter activity	Transport	vacuole
peptidase activity	lipid metabolism	plasma membrane
structural molecule activity	protein catabolism	endoplasmic reticulum
RNA binding	protein biosynthesis	mitochondrion
	morphogenesis	ribosome
	organelle organization and biogenesis	
	vesicle-mediated transport	
	pseudohyphal growth	

จากตาราง 4.6 แสดงให้เห็นถึงยีนออนโทโลยีที่เป็นความหมายของปัจจัยที่ 1

ซึ่งเมื่อพิจารณาที่ฟังก์ชันการทำงานของยีน (Molecular Function) ความหมายของปัจจัยที่ 1 จะหมายถึง transferase activity, isomerase activity และ chaperone activity เป็นต้น

เมื่อพิจารณาที่ฟังก์ชันกระบวนการทางชีววิทยา (Biological Process) ความหมายของปัจจัยที่ 1 จะหมายถึง RNA metabolism, ribosome biogenesis and assembly และ carbohydrate metabolism เป็นต้น

และเมื่อพิจารณาที่ส่วนประกอบของเซลล์ (Cellular Component) ความหมายของปัจจัยที่ 1 จะหมายถึง nucleolus, cytoplasm และ mitochondrial membrane เป็นต้น

ตาราง 4.7 ยีนออนโทโลยี ของ 50 ยีนที่มีค่าน้ำหนักปัจจัย ณ ปัจจัยที่ 2 สูงที่สุดจากการวิเคราะห์ปัจจัย โดยวิธีวิเคราะห์องค์ประกอบหลักกับชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิสเซอริวิสิเอ โดยใช้ยีนเป็นตัวแปร

Molecular Function	Biological Process	Cellular Component
transferase activity	carbohydrate metabolism	cytoplasm
ligase activity	Transport	plasma membrane
transporter activity	RNA metabolism	mitochondrion
oxidoreductase activity	ribosome biogenesis and assembly	nucleolus
lyase activity	Conjugation	site of polarized growth
enzyme regulator activity	Morphogenesis	nucleus
hydrolase activity	Sporulation	ribosome
structural molecule activity	vitamin metabolism	membrane
RNA binding	protein biosynthesis	peroxisome
	amino acid and derivative metabolism	cell wall
	energy pathways	
	cell wall organization and biogenesis	
	lipid metabolism	

จากตาราง 4.7 แสดงให้เห็นถึงยีนออนโทโลยีที่เป็นความหมายของปัจจัยที่ 2

ซึ่งเมื่อพิจารณาที่ฟังก์ชันการทำงานของยีน (Molecular Function) ความหมายของปัจจัยที่ 2 จะหมายถึง transferase activity, ligase activity และ transporter activity เป็นต้น

เมื่อพิจารณาที่ฟังก์ชันกระบวนการทางชีววิทยา (Biological Process) ความหมายของปัจจัยที่ 2 จะหมายถึง carbohydrate metabolism, RNA metabolism และ ribosome biogenesis and assembly เป็นต้น

และเมื่อพิจารณาที่ส่วนประกอบของเซลล์ (Cellular Component) ความหมายของปัจจัยที่ 2 จะหมายถึง cytoplasm, plasma membrane และ mitochondrion เป็นต้น

- สรุปผลการวิเคราะห์

ปัจจัยที่มีผลต่อกระบวนไดอ๊อคซิคชิฟท์ ซึ่งได้จากกระบวนการวิเคราะห์ปัจจัย สามารถอธิบายได้ด้วยยีนออนโทโลยี ของกลุ่มยีนที่มีค่าความสัมพันธ์กับปัจจัยนั้นๆสูง แต่ปัญหาก็คือด้วยข้อมูลยีนออนโทโลยีและเทคนิควิธีการนี้ การอธิบายความหมายของปัจจัยดังกล่าวยังมีขอบเขตที่กว้าง ไม่สามารถระบุไปอย่างชัดเจนว่าปัจจัยแต่ละปัจจัยนั้นคืออะไร ซึ่งจะสังเกตได้จากผลการทดลองที่พบว่ายีนออนโทโลยีบางตัวสามารถอธิบายปัจจัยได้ทั้งสองปัจจัย ทั้งนี้อาจเป็นผลมาจากการที่ปัจจัยทั้งสองปัจจัยนี้ มียีนที่มีค่าน้ำหนักปัจจัยที่สูง เป็นยีนตัวเดียวกัน หรืออาจเป็นเพราะกลุ่มยีน ที่เป็นตัวแทนของแต่ละปัจจัย มียีนออนโทโลยีเหมือนกัน ดังนั้นข้อสรุปของปัจจัยดังกล่าวจึงยังต้องมีการศึกษาเพื่อให้ได้ผลสรุปเชิงชีววิทยาที่ดีที่สุดต่อไป

4.2.3 การวิเคราะห์ปัจจัยเพื่อ จัดกลุ่มยีนในชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิส เซอริวิสิเอ

- ปัญหาและวัตถุประสงค์ของการวิเคราะห์

การจัดกลุ่มข้อมูล โดยเฉพาะข้อมูลทางด้านดีเอ็นเอไมโครอาร์เรย์ ปัญหาอย่างหนึ่งที่มักพบก็คือ จำนวนตัวแปรมีมากเกินไป บางครั้งมากกว่าตัวอย่างข้อมูลด้วยซ้ำไป ผลก็คือ ทำให้เกิดปัญหาในการประมาณค่าพารามิเตอร์ในโมเดลของการวิเคราะห์ เช่น ปัญหาโอเวอร์ฟิต (Overfitting) และปัญหาทางด้านการประมวลผลที่ต้องใช้เวลามาก (Overloading) เป็นต้น นอกจากนี้ ตัวแปรที่ใช้บางตัวไม่จำเป็นสำหรับการจัดกลุ่ม การนำตัวแปรดังกล่าวมาใช้ จะทำให้ผลการวิเคราะห์ไม่น่าเชื่อถือ แต่ทั้งนี้ในกรณีของข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิส เซอริวิสิเอ ที่นำมาวิเคราะห์ มีตัวแปรเพียง 7 ตัวแปร เมื่อนำมาวิเคราะห์การจัดกลุ่มปัญหาดังที่กล่าวมาอาจจะไม่เกิดขึ้น ซึ่งก็มีงานวิจัยมากมายที่วิเคราะห์การจัดกลุ่มโดยใช้ตัวแปรดังกล่าว แต่เพื่อที่จะสร้างแบบจำลองสำหรับแก้ปัญหาการวิเคราะห์ข้อมูลที่จำนวนตัวแปรมีจำนวนมาก และ มากกว่าตัวอย่างข้อมูล ซึ่งอาจก่อให้เกิดปัญหาดังที่กล่าวมา ในการวิเคราะห์ข้อมูลชุดนี้จึง อาศัยวิธีการวิเคราะห์ปัจจัยมาประยุกต์สำหรับการจัดกลุ่มข้อมูล ซึ่งในที่นี้คือ ยีน โดยการสร้างตัวแปรจำนวนน้อยๆ ขึ้นมาใหม่ในลักษณะของ ปัจจัยที่เป็นตัวแทนของยีนที่จะจัดกลุ่ม ทั้งนี้มีสมมุติฐานว่า ยีนที่มีรูปแบบความสัมพันธ์กับปัจจัยทุกๆปัจจัยคล้ายๆ กัน ย่อมจะอยู่ในกลุ่มเดียวกัน ดังนั้น ทฤษฎีการจัดกลุ่มซึ่งมีการนำเสนอในงานวิจัยต่างๆ ดังเช่น การจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) จึงนำมาใช้ในการจัดกลุ่มยีนโดยใช้ปัจจัยเป็นตัวแปร

- แหล่งข้อมูลและลักษณะของข้อมูล

ชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิส เซอริวิสิเอ เป็นชุดข้อมูลเดียวกับชุดข้อมูลที่อยู่ใน บทที่ 3 หัวข้อ 3.1.1 ทั้งนี้ข้อมูลดีเอ็นไมโครอาร์เรย์ที่เลือกมาใช้นั้น เลือกมาเพียง 35 ยีน ที่มีการจัดกลุ่มแล้วตามงานวิจัยที่เป็นแหล่งของข้อมูลนี้ ซึ่งแสดงชุดยีนดังกล่าวได้ในตาราง 3.10 บทที่ 3 เพื่อเปรียบเทียบผล

- วิธีการวิเคราะห์

1) จากข้อมูลที่ประกอบด้วยยีน (Genes) ซึ่งมีจำนวนทั้งสิ้น 6,153 ยีน และกรองข้อมูลโดยกรองเอายีนที่ข้อมูลขาดหาย ออกไป เหลือเพียง 6,119 ยีน ดังในหัวข้อที่ผ่านมา จะทำการเลือกยีนเฉพาะที่ได้ทำการจัดกลุ่มแล้วดังตาราง 3.10 ในบทที่ 3 มาใช้วิเคราะห์

2) กำหนดให้ ยีนเป็นตัวแปร ซึ่งมีทั้งสิ้น 35 ยีน

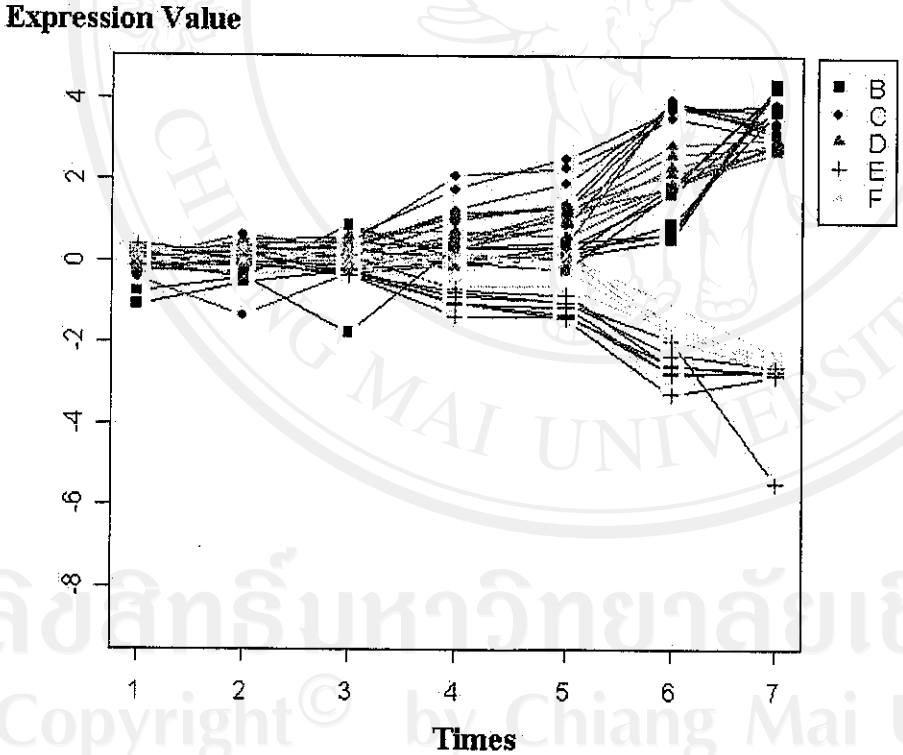
3) สกัดปัจจัยเพื่อหาเมตริกซ์ของค่าน้ำหนักปัจจัย โดยใช้วิธีวิเคราะห์องค์ประกอบหลักและใช้เมตริกซ์ความแปรปรวนร่วมเป็นเมตริกซ์สหสัมพันธ์

4) เลือกจำนวนปัจจัยโดยพิจารณาที่ค่าความแปรปรวนมากกว่า 1 ตามวิธีการเลือกจำนวนปัจจัย

5) นำเมตริกซ์ของค่าน้ำหนักปัจจัย มาใช้เป็นข้อมูลตั้งต้นสำหรับการจัดกลุ่มยีน โดยกำหนดให้ปัจจัยเป็นตัวแปร และยีนเป็นตัวอย่างข้อมูล ซึ่งทฤษฎีที่ใช้ร่วมกับการจัดกลุ่มยีน คือทฤษฎีการจัดกลุ่มข้อมูลแบบลำดับชั้น ทั้งนี้ในการวิเคราะห์ข้อมูลจะใช้ฟังก์ชัน ในการจัดกลุ่มข้อมูลแบบลำดับชั้นที่มีมากับโปรแกรมคอมพิวเตอร์วิเคราะห์โดยตรง ดังนั้นในส่วนของรายละเอียดเชิงทฤษฎีจึงไม่กล่าวถึง

- ผลการวิเคราะห์

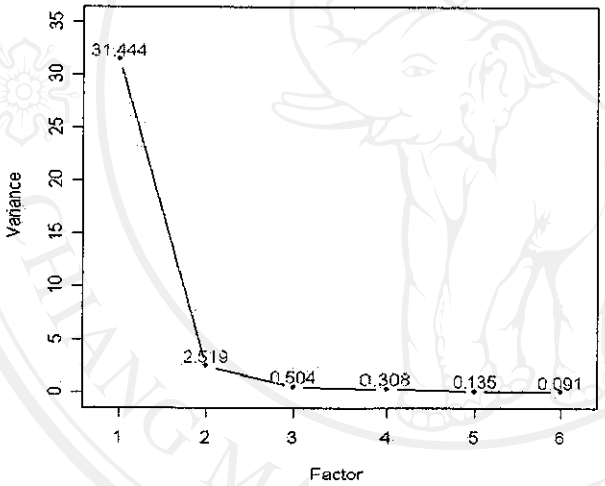
จากยีนที่แสดงใน ตาราง 3.10 ซึ่งมีจำนวน 35 ยีน เมื่อนำค่าการแสดงออกมาพล็อตลงในกราฟเพื่อแสดงลักษณะการแสดงออกของยีนใน 7 ช่วงเวลา ของกระบวนการไดอ๊อคซิคชิฟ จะแสดงได้ดังรูป 4.11



รูป 4.11 กราฟแสดงค่าการแสดงออกของยีนในกระบวนการไดอ๊อคซิคชิฟของยีสต์ จากข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ ซัคคาโรไมซิสเซอริวิสิเอ จำนวน 35 ยีน

จากรูป 4.11 ในแกนนอนจะแสดงเวลาในกระบวนการไดอ๊อคซิคชิฟท์ ส่วนในแกนตั้งแสดงค่าการแสดงออกของยีน กราฟแต่ละเส้นแสดงถึงลักษณะการแสดงออกของยีนแต่ละตัวซึ่งจะเห็นว่ายีนที่อยู่ในกลุ่มเดียวกันจะมีรูปแบบของค่าการแสดงออกทั้ง 7 ช่วงเวลาคล้ายๆ กัน แต่เนื่องจาก ถ้าการวิเคราะห์การจัดกลุ่มยีนกำหนดให้เวลาทั้ง 7 ช่วงเวลาเป็นตัวแปร จำนวนตัวแปรจะมีจำนวนมาก ซึ่งทำให้เกิดปัญหาดังที่กล่าวมาก่อนหน้า ในการทดลองนี้จึงทำการวิเคราะห์ปัจจัยก่อน เพื่อสร้างตัวแปรขึ้นใหม่สำหรับการจัดกลุ่มยีนแทนช่วงเวลาทั้ง 7 ช่วงเวลา โดยตัวแปรใหม่นี้คือปัจจัย นั่นคือในขั้นตอนของการวิเคราะห์ปัจจัย จะกำหนดให้ยีนเป็นตัวแปร ส่วนในขั้นตอนของการวิเคราะห์การจัดกลุ่มจะกำหนดให้ปัจจัยเป็นตัวแปร

ผลจากการวิเคราะห์ปัจจัย จะแสดงค่าความแปรปรวนของข้อมูลแต่ละปัจจัยได้ดัง สครีพล็อตในรูป 4.12



รูป 4.12 สครีพล็อตจากวิธีการวิเคราะห์ปัจจัยกับข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิส เซอริวิสิเอ โดยใช้ยีนเป็นตัวแปร จำนวน 35 ยีน

จากรูป 4.12 ค่าตัวเลขในแกนนอนของกราฟหมายถึง ปัจจัย ส่วนค่าตัวเลขในแกนดิ่งหมายถึงค่าความแปรปรวนของข้อมูล ซึ่งด้วยวิธีการวิเคราะห์ปัจจัยที่ใช้เมตริกซ์สหสัมพันธ์ในการวิเคราะห์ จะทำให้ค่าความแปรปรวนของยีนแต่ละตัวนั้น มีค่าเท่ากับ 1 ผลก็คือค่าความแปรปรวนของข้อมูลทั้งหมดจะเท่ากับจำนวนยีน นั่นคือ 35 จากกราฟค่าความแปรปรวนของข้อมูลในแต่ละปัจจัยจะแสดงออกมาในลักษณะของกราฟที่ลาดชันลงมาจากปัจจัยที่ 1 ไปหาปัจจัย ตัวท้ายๆ ซึ่งจะเห็นว่า ปัจจัยที่ 1 ความแปรปรวนของข้อมูลจะมีค่าเท่ากับ 31.44 หรือคิดเป็นร้อยละ 90 ของความแปรปรวนทั้งหมด และที่ปัจจัยที่ 2 ค่าความแปรปรวนเป็น 2.519 หรือ คิดเป็นร้อยละ 7.2 ของค่าความแปรปรวนทั้งหมด ดังนั้น

ความแปรปรวนรวมของทั้งสองปัจจัยจะมีค่าเท่ากับร้อยละ 97 ของค่าความแปรปรวนทั้งหมดซึ่งถือว่ามากพอที่ปัจจัยทั้งสองตัวนี้จะเป็นตัวแทนของข้อมูลได้ทั้งหมด และด้วยหลักการเลือกจำนวนปัจจัยที่เหมาะสมในหัวข้อ 4.1.4 ทำให้เราเลือกปัจจัยที่มีค่าความแปรปรวนมากกว่า 1 ได้ที่ 2 ปัจจัยแรก

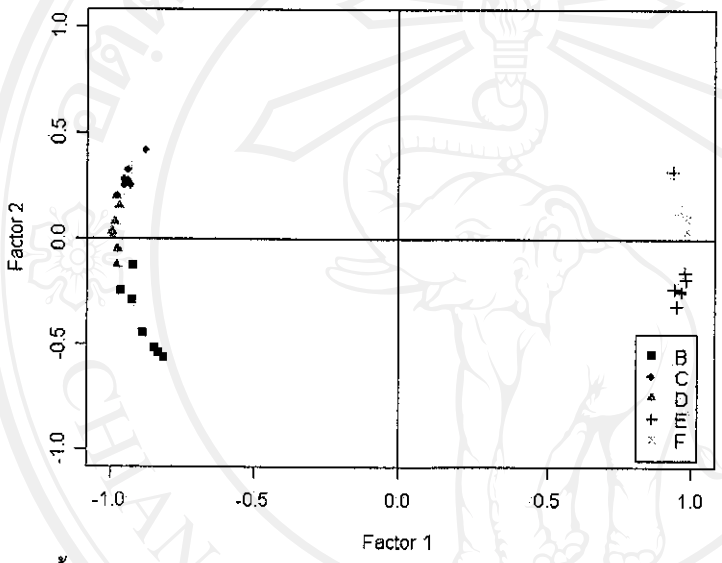
เมตริกซ์ของค่าน้ำหนักปัจจัย ค่าความแปรปรวนของแต่ละตัวแปร ค่าร่วมกัน และ ค่าความแปรปรวนเฉพาะ จากการสกัดปัจจัย ที่ 2 ปัจจัยแรกจะแสดงได้ในตาราง 4.8

ตาราง 4.8 ค่าพารามิเตอร์ต่างๆ จากผลการวิเคราะห์ปัจจัยในชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิสเซอริวิสิเอ โดยใช้ยีนเป็นตัวแปร จำนวน 35 ยีน

Gene	Factor 1	Factor 2	Communality	Uniqueness
YNL117W	-0.888	-0.448	0.989	0.011
YLR174W	-0.923	-0.123	0.866	0.134
YER065C	-0.816	-0.562	0.983	0.017
YAL054C	-0.963	-0.248	0.989	0.011
YJR095W	-0.925	-0.292	0.940	0.060
YLR377C	-0.848	-0.519	0.988	0.012
YKR097W	-0.833	-0.543	0.989	0.011
YLR258W	-0.879	0.417	0.947	0.053
YGR088W	-0.953	0.279	0.987	0.013
YDR171W	-0.976	0.199	0.992	0.008
YBR072W	-0.940	0.321	0.988	0.012
YFL014W	-0.938	0.273	0.955	0.045
YKL026C	-0.951	0.251	0.967	0.033
YGR043C	-0.934	0.247	0.933	0.067
YOR065W	-0.977	-0.053	0.957	0.043
YNL052W	-0.991	0.017	0.983	0.017
YHR051W	-0.994	0.030	0.989	0.011
YGL191W	-0.970	0.152	0.964	0.036
YEL024W	-0.986	0.078	0.977	0.023
YDR529C	-0.980	0.195	0.998	0.002
YBL045C	-0.980	-0.124	0.975	0.025
YPL012W	0.948	-0.316	0.998	0.002
YNL141W	0.964	-0.248	0.990	0.010
YMR290C	0.964	-0.246	0.989	0.011
YLR180W	0.980	-0.191	0.996	0.004
YIL053W	0.978	-0.157	0.980	0.020
YGR160W	0.941	-0.233	0.939	0.061
YDR398W	0.936	0.323	0.980	0.020
YNL069C	0.933	0.303	0.963	0.037
YPL220W	0.961	0.134	0.942	0.058
YLR340W	0.991	0.009	0.983	0.017
YGL076C	0.986	0.040	0.973	0.027
YHL015W	0.975	0.110	0.963	0.037
YDR418W	0.985	0.097	0.979	0.021
YLL045C	0.952	0.149	0.929	0.071
Var.	31.444	2.519	**33.963**	**1.037**
Proportion Var.	0.898	0.072	**0.970**	**0.030**

ผลจากตาราง 4.8 จะได้เมตริกซ์ของค่าน้ำหนักปัจจัย ซึ่งจะสังเกตว่า ค่าความแปรปรวนของข้อมูลของทั้งสองปัจจัยนี้มีค่าถึง 97 เปอร์เซ็นต์ จึงทำให้ค่าร่วมกัน (Communality) ของยีนแต่ละตัวกับปัจจัยทั้งสองปัจจัยนี้มีค่าสูง และค่าความแปรปรวนเฉพาะ (Uniqueness) มีค่าต่ำ นั่นก็แสดงว่าปัจจัยทั้งสองนี้ เป็นปัจจัยร่วมของยีนทุกๆ ตัว และเป็นตัวแทนของยีนเหล่านี้ได้ดี

จากตารางดังกล่าว เมื่อนำค่าน้ำหนักปัจจัยมาพล๊อตลงในกราฟ 2 มิติโดยกำหนดให้แกนนอนเป็นค่าน้ำหนักปัจจัยของยีน ในปัจจัยที่ 1 และแนวดิ่งเป็นของปัจจัยที่ 2 รวมทั้งกำหนดกลุ่มให้กับยีนตามกลุ่มของข้อมูลที่อยู่ในตาราง 3.10 จะแสดงได้ดังรูป 4.12



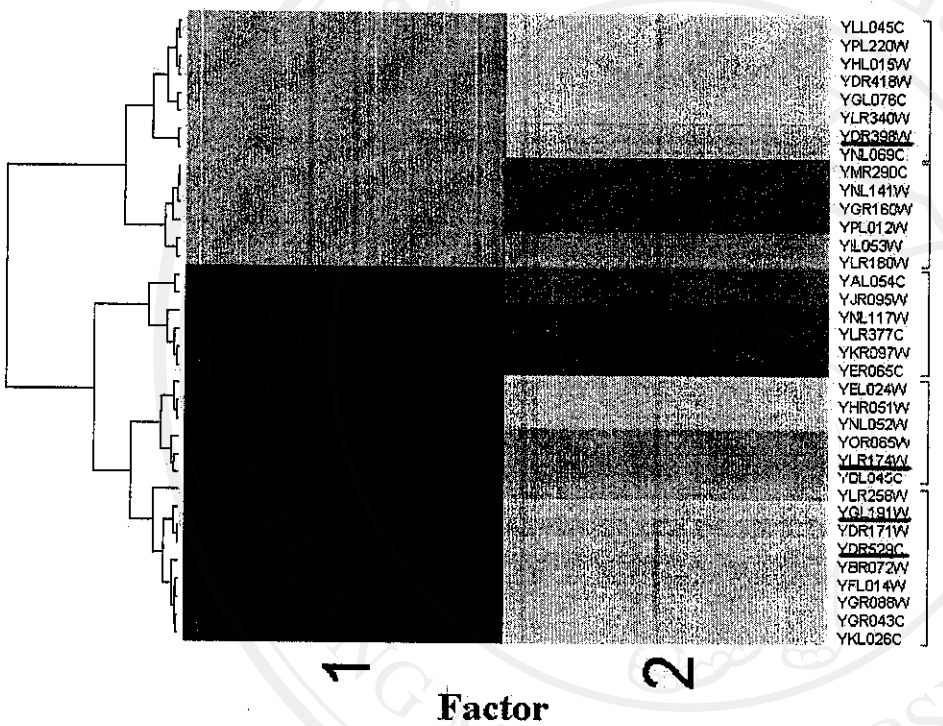
รูป 4.13 กราฟของค่าน้ำหนักปัจจัย จากผลการวิเคราะห์ปัจจัยโดยวิธีวิเคราะห์ปัจจัยกับชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิส เซอริวิสิเอ โดยใช้ยีนเป็นตัวแปรจำนวน 35 ยีน กำหนดกลุ่มยีนโดยอาศัยกลุ่มยีนจากผลงานวิจัยที่เป็นแหล่งของข้อมูล

จากรูป 4.13 กราฟที่ได้ แสดงให้เห็นว่า ยีนที่อยู่ในกลุ่มเดียวกัน จะมีโครงสร้างของข้อมูล ที่เหมือนกัน นั่นคือค่าน้ำหนักปัจจัยทั้งสองปัจจัย มีลักษณะใกล้เคียงกัน ซึ่งแสดงให้เห็นความแตกต่างของกลุ่มยีนดังกล่าวได้ แต่ทั้งนี้จะสังเกตว่า มียีนบางตัวที่ไม่ได้แยกออกมาจากกลุ่มอื่น ซึ่งอาจมี สาเหตุมากจากกลุ่มยีนที่กำหนดไว้ในตอนต้นนั้นยังจัดกลุ่มได้ไม่ดีที่สุด หรือ เนื่องจากค่าน้ำหนักปัจจัยที่ได้นั้นเป็นค่าที่เป็นตัวแทนของข้อมูลเพียง 97 เปอร์เซ็นต์ ของข้อมูลเดิม จึงอาจให้ค่าที่ผิดพลาดในบางยีนได้ นอกจากนี้อาจจะมีสาเหตุ อื่นๆ ที่ผู้วิจัยไม่สามารถคาดการณ์ได้

จากผลของการวิเคราะห์ปัจจัยดังกล่าว ข้อสังเกตที่พบแม้ว่าจะไม่สามารถบอกได้ว่ามีสาเหตุมาจากอะไรแน่ชัด แต่ด้วยข้อสังเกตหลักที่สำคัญ นั่นก็คือยีนส่วนใหญ่ที่อยู่ในกลุ่มเดียวกัน จะมี

โครงสร้างของปัจจัยทั้งสองปัจจัยที่เหมือนกันซึ่งเห็นได้จาก กลุ่มยีนที่กำหนดมานั้นมีการจับกลุ่มอยู่ด้วยกัน จึงทำให้ ข้อสมมุติฐานที่ว่าปัจจัย สามารถใช้เป็นตัวแปรตั้งต้นสำหรับการจัดกลุ่มยีน มีความเป็นไปได้ยิ่งขึ้น แต่ทั้งนี้ด้วยสาเหตุดังที่กล่าวมา ผลที่ได้จากการจัดกลุ่มโดยใช้ปัจจัย อาจจะให้ผลที่ต่างจากกลุ่มยีนที่มีการกำหนดมาในข้างต้น

ในการจัดกลุ่มยีนโดยอาศัยวิธีการจัดกลุ่มข้อมูลแบบลำดับชั้นและกำหนดให้ปัจจัยเป็นตัวแปรตั้งต้น ทั้งนี้ค่าที่ใช้วัดผลคือค่าน้ำหนักปัจจัย จะให้ผลการจัดกลุ่มได้ดังรูป 4.14



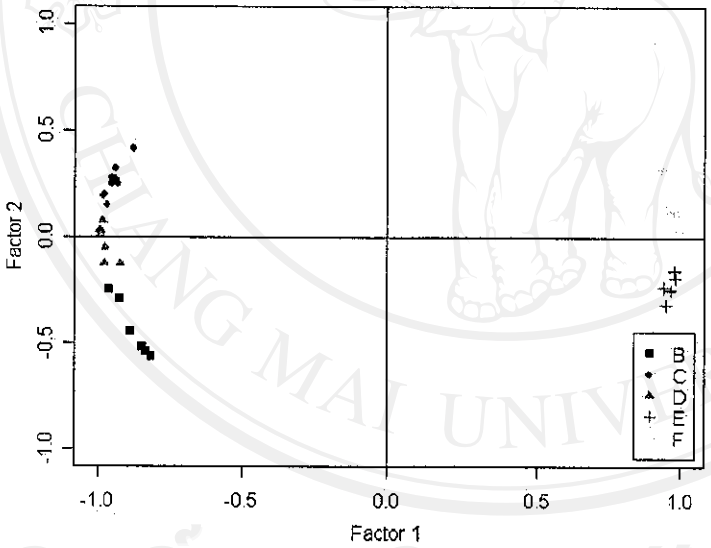
รูป 4.14 ผลการจัดกลุ่มยีนในชุดข้อมูล ดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิส เซอริวิสิเอ โดยวิธีการจัดกลุ่มยีนแบบลำดับชั้น (Hierarchical Clustering)

จากรูป 4.14 ค่าน้ำหนักปัจจัยแต่ละปัจจัยของยีน จะแสดงออกมาลักษณะของแถบสี โดยแถบสีที่มีค่าน้ำหนักปัจจัย มากกว่า 0 และมีค่ามากแถบสีจะเป็นสีเขียว และที่มีค่าน้ำหนักปัจจัยใกล้ๆ 0 จะมีสีเหลือง สำหรับแถบสีของยีนที่มีค่าน้ำหนักปัจจัยมากๆ แต่มีทิศทางเป็นลบ จะมีแถบสีแดง และผลจากการจัดกลุ่มยีน แบบลำดับชั้น สามารถที่จะจัดกลุ่มยีนได้หลายๆ กลุ่ม ทั้งนี้เราจะเลือกออกมา 5 กลุ่ม ยีนให้เหมือนกับกลุ่มยีนที่กำหนดมาในข้างต้น ซึ่งจากแถบสีจะสังเกตเห็นว่ายีนที่อยู่ในกลุ่มเดียวกัน จะมีลักษณะของแถบสีทั้งสองปัจจัยคล้ายๆกัน นั่นคือมีรูปแบบของค่าน้ำหนักปัจจัยทั้งสองปัจจัยเหมือนกัน กลุ่มยีนที่ได้จากการจัดกลุ่มดังกล่าว จะแสดงได้ดังตาราง 4.9

ตาราง 4.9 ผลการจัดกลุ่มยีนในชุดข้อมูล ดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิสเซอริวิสิเอ โดยวิธีการจัดกลุ่มยีนแบบลำดับชั้น (Hierarchical Clustering)

กลุ่ม	แหล่งข้อมูล	วิธีวิเคราะห์ปัจจัย ร่วมกับ วิธีวิเคราะห์แบบลำดับชั้น
1	YNL117W, YLR174W, YER065C, YAL054C, YJR095W, YLR377C, YKR097W	YJR095W, YAL054C,YNL117W, YLR377C, YER065C, YKR097W, YOR065W, YLR174W
2	YLR258W, YGR088W, YDR171W,YBR072W, YFL014W, YRL026C, YGR043C	YLR258W, **YGL191W**, YDR171W, **YDR529C**, YBR072W, YGR043C, YRL026C, YFL014W, YGR088W
3	YOR065W, YNL052W, YHR051W,YGL191W, YEL024W, YDR529C, YBL045C	YOR065W, **YLR174W**, YBL045C, YEL024W, YNL052W, YHR051W
4	YPL012W, YNL141W, YMR290C, YLR180W, YIL053W, YGR160W, YDR398W	YMR290C, YNL141W, YGR160W, YPL012W, YIL053W, YLR180W
5	YNL069C, YPL220W, YLR340W, YGL076C, YHL015W, YDR418W, YLL045C	YDR418W, YHL015W, YPL220W, YLL045C, YLR340W, YGL076C, **YDR398W**

จากรูป 4.14 และ ตาราง 4.9 จะสังเกตเห็นว่า มียีน 4 ตัว ได้แก่ YGL191W, YDR529C, YLR174W และ YDR398W อยู่ในกลุ่มยีนที่แตกต่างจากกลุ่มยีนที่กำหนดมาในข้างต้น และ เมื่อนำค่าน้ำหนักปัจจัยของยีนทั้ง 35 ยีน มาพล็อตลงในกราฟ 2 มิติเช่นเดียวกับในรูป 4.13 โดย กำหนดกลุ่มยีนเป็นกลุ่มยีนที่เราวิเคราะห์ได้ใหม่ จะให้ผลแสดงได้ดังรูป 4.15



รูป 4.15 กราฟของค่าน้ำหนักปัจจัย จากผลการวิเคราะห์ปัจจัยโดยวิธีวิเคราะห์ปัจจัยกับชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิสเซอริวิสิเอ โดยใช้ยีนเป็นตัวแปรจำนวน 35 ยีน กำหนดกลุ่มยีนโดยวิธีการจัดกลุ่มแบบลำดับชั้น

จากรูป 4.15 จะสังเกตเห็นว่า กลุ่มยีนที่อยู่ในกลุ่มเดียวกัน จะจัดกลุ่มอยู่ด้วยกัน

- สรุปผลการวิเคราะห์

ผลจากการวิเคราะห์ปัจจัยจะได้ปัจจัยที่สามารถอธิบายกลุ่มข้อมูลที่กำหนดได้ ทั้งยังใช้เป็นตัวแปร ในการวิเคราะห์การจัดกลุ่มแบบอื่นๆ เช่น วิธีการจัดกลุ่มแบบลำดับชั้น แต่ปัญหาที่พบคือ ผลการจัดกลุ่มยีน ไม่สามารถบอกได้ว่าดีหรือไม่อย่างไร เนื่องจากผลที่ได้ แตกต่างจากกลุ่มยีนที่กำหนดมาในตอนต้น และกลุ่มยีนที่กำหนดมาในขั้นต้นนั้นก็ไม่สามารถบอกได้ว่า จัดกลุ่มได้ดีเพียงใด แต่ทั้งนี้ความแตกต่างของผลการวิเคราะห์กับกลุ่มยีนที่กำหนดมาก่อนหน้านั้น แตกต่างกันเพียงเล็กน้อย จึงเป็นไปได้ว่า ส่วนที่เหมือนกันนั้น น่าจะเป็นกลุ่มยีนที่มีการจัดกลุ่มได้ถูกต้อง แต่ส่วนที่ต่างกันนี้ ต้องปรับปรุงวิธีการวิเคราะห์เพื่อหาสาเหตุต่อไป

4.2.4 การวิเคราะห์ปัจจัยในชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์วัณโรค เพื่อจัดกลุ่มยาที่มีผลต่อโรควัณโรค (Drug Clustering) การหาเป้าหมายของยา (Drug Target Detection) และการหาพาทเวย์ที่กลุ่มยามีผลกระทบ (Pathway Detection)

- ปัญหาและวัตถุประสงค์ของการวิเคราะห์

จากการศึกษาในงานวิจัยหนึ่งที่วิเคราะห์ชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของโรคมะเร็งซึ่งเป็นชุดข้อมูลที่วัดค่าการแสดงออกของยีน ในเซลล์มะเร็งเพาะเลี้ยง (Cell Lines) ชนิดต่างๆ (Lozano, 2005) พบว่าลักษณะของข้อมูลนั้น จะประกอบได้ด้วยตัวแปรคือ ยีนของเซลล์มะเร็ง และตัวอย่างข้อมูลคือเซลล์มะเร็งเพาะเลี้ยง โดยจำนวนยีนมีมากกว่าตัวอย่างข้อมูล ซึ่งงานวิจัยวิจัยดังกล่าวต้องการที่จะจัดกลุ่มเซลล์มะเร็งที่มีความ สัมพันธ์กันเข้าด้วยกัน พร้อมกันนั้น ยังต้องการที่จะระบุยีนที่เกี่ยวข้องกับกลุ่มเซลล์นั้นๆ (Relevant Gene Extraction) เพื่อที่จะหาว่ากลุ่มเซลล์ดังกล่าวมีพาทเวย์ ของการทำงานภายในเซลล์ เป็นอย่างไรบ้าง ซึ่งจากการศึกษาในระยะเริ่มต้น เทคนิควิธีการที่ใช้วิเคราะห์ข้อมูลในลักษณะนี้ได้แก่ การเลือกยีนแบบไม่ควบคุม (Unsupervised Gene Selection) การจัดกลุ่มข้อมูลที่มีความสัมพันธ์กัน (Interrelated Clustering) และการจัดกลุ่มไบคลัสเตอร์ (Biclustering) แต่ปัญหาที่พบจากการวิเคราะห์ข้อมูลด้วยวิธีการเหล่านี้ได้แก่ ปัญหาโอเวอร์ฟิต (Overfitting Problem) ปัญหายีนมีจำนวนมากเกินไป (Irrelevant or Redundant Genes Problem) ซึ่งยีนบางตัวที่ไม่จำเป็น (Noise) ต่อการจัดกลุ่มข้อมูล ทำให้การจัดกลุ่มข้อมูลโดยโมเดลการวิเคราะห์ต่างๆ เช่น โมเดลการวิเคราะห์แบบลำดับชั้น ไม่มีความน่าเชื่อถือ และสุดท้ายเป็นปัญหาในเรื่องที่ไม่สามารถจัดกลุ่มยีนให้เข้ากับกลุ่มเซลล์มะเร็งต่างๆ อันเนื่องมาจาก ยีนบางตัวสามารถที่จะมีความสัมพันธ์กับของกลุ่มเซลล์มะเร็งได้มากกว่า 2 กลุ่ม (Non-Overlapping Gene Cluster) ด้วยปัญหาที่พบทำให้งานวิจัยดังกล่าว ได้ปรับวิธีการวิเคราะห์ข้อมูลขึ้นมาใหม่สำหรับแก้ปัญหาดังกล่าว ซึ่งวิธีการที่นำมาใช้นี้ เป็นวิธีการเช่นเดียวกับที่นำเสนอมาแล้วในการทดลอง 4.3.3 นั่นคือ การใช้ วิธีการวิเคราะห์ปัจจัยร่วมกับ

วิธีการจัดกลุ่มข้อมูลแบบลำดับชั้นสำหรับการจัดกลุ่มตัวอย่างข้อมูล ซึ่งในที่นี้คือกลุ่มเซลล์มะเร็งเพาะเลี้ยง นอกจากนี้ยังอาศัยวิธีการทางสถิติที่เรียกว่า ที-เทส (t-test method)ในการจัดกลุ่มยีนที่มีความสัมพันธ์กับกลุ่มของเซลล์มะเร็ง และใช้ยีนออนโทโลยีในการหาพาทเวย์ที่เกี่ยวข้อง

ด้วยวิธีการในงานวิจัยที่กล่าวมาขั้นต้น จะเห็นได้ว่า เป็นการประยุกต์เทคนิควิธีการวิเคราะห์ปัจจัยกับการจัดกลุ่มข้อมูลที่ให้ผลในระดับที่น่าเชื่อถือ และได้รับการยอมรับอย่างกว้างขวางในหมู่นักวิจัย งานวิจัยที่จะทำต่อไปนี้ จึงเห็นสมควรที่จะประยุกต์วิธีการดังกล่าวกับข้อมูลอื่นซึ่งคาดว่าน่าจะให้ผลการวิเคราะห์ที่ดี ไม่แตกต่างกัน และข้อมูล ที่นำมาใช้คือ ชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์วัณโรค (*Mycobacterium tuberculosis*) วัตถุประสงค์ของงานวิจัยที่จะทำนั้น ก็เพื่อที่จะจัดกลุ่มยา และ การหาเป้าหมายของยา ซึ่งส่งผลไปถึงการหาพาทเวย์ของกลุ่มยีนที่กลุ่มยาเหล่านี้มีผลกระทบด้วย ทั้งนี้จากการศึกษาข้อมูลชุดนี้พบว่าจำนวนยีนมีมากกว่าจำนวนของยาซึ่งถือเป็นตัวอย่างข้อมูลที่จัดกลุ่ม ดังนั้นปัญหาที่เกิดขึ้นจึงไม่ต่างกับปัญหาจากงานวิจัยที่วิเคราะห์ชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของโรคมะเร็งดังที่กล่าวมา

- แหล่งข้อมูลและลักษณะของข้อมูล

ข้อมูลที่นำมาใช้เป็นชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์วัณโรค (Boshoff, 2004) จากผลงานวิจัยเรื่อง "การตอบสนองในระดับทรานสคริปชั่นของเชื้อวัณโรคต่อยา (The Transcriptional Response of *Mycobacterium tuberculosis* to inhibitors of Metabolism)" ทั้งนี้ลักษณะข้อมูลไมโครอาร์เรย์ประกอบไปด้วยยีนจำนวน 4,320 ยีน และยาจำนวน 436 ตัวยา เช่น [1ug/mL No.121940: DMSO, 12h (mAdb expid=44709)], [1ug/mL No.111891: DMSO, 12h (mAdb expid=44710)] และ [24ug/mL clotrimazole: DMSO, 6h (mAdb expid=44713)] เป็นต้น สำหรับข้อมูลดาวน์โหลดได้จากเวบไซต์ http://www.ncbi.nlm.nih.gov/geo/ ที่หมายเลขจีอีโอ (GEO Accession Number) GSE1642 และ GSE1694 นอกจากนี้ ข้อมูลที่สำคัญอีกส่วนหนึ่งสำหรับนำไปใช้ในการหาพาทเวย์ ก็คือ ข้อมูลยีนออนโทโลยีของยีนที่เกี่ยวข้องวัณโรค จะสามารถดาวน์โหลดได้จากเวบไซต์ www.geneontology.org

- วิธีการวิเคราะห์

1) สร้างเมตริกซ์ของข้อมูลดีเอ็นเอไมโครอาร์เรย์โดยกำหนดให้ ชุดของยาเป็นตัวแปร และ ยีนเป็นตัวอย่างข้อมูล

2) กรองข้อมูล โดยตัดชุดยาที่ข้อมูลขาดหายออกไป ทั้งนี้เนื่องจากชุดข้อมูลที่นำมาวิเคราะห์นี้มีข้อมูลที่ขาดหายเกือบทุกๆ ชุดยา การตัดออกทั้งหมดจะไม่สามารถนำข้อมูลดังกล่าวมาวิเคราะห์ต่อไปได้ ดังนั้นวิธีการตัดก็คือ จะเลือกตัดเอาชุดยาที่ข้อมูลมีการขาดหายจำนวนมากออกไป เพียงบางตัว แล้ว

ชุดยาที่เหลือก็นำไปพิจารณาว่ามียีนตัวใดบ้างที่ข้อมูลขาดหาย จากนั้นจึงตัดเอายีน เหล่านั้นทิ้งไป ซึ่งทั้งนี้จะพิจารณาตัดชุดยาและยีน จากเปอร์เซ็นต์ของข้อมูลที่ขาดหายไปในชุดยาแต่ละตัว แล้วพิจารณาสัดส่วนของชุดยาและยีนที่เหลืออยู่ ซึ่งเหมาะสม

3) นอร์มอลไลซ์เซชันข้อมูลโดยวิธีการสเกล นอร์มอลไลซ์เซชัน

4) ทำการวิเคราะห์ปัจจัยโดยใช้เมตริกซ์สหสัมพันธ์ สกัดปัจจัยโดยวิธีวิเคราะห์องค์ประกอบหลัก เลือกจำนวนปัจจัย ที่มีความแปรปรวน มากกว่า 1

5) ใช้เมตริกซ์ของค่าน้ำหนักปัจจัย มาวิเคราะห์การจัดกลุ่มโดยวิธีวิเคราะห์การจัดกลุ่มแบบลำดับชั้น ซึ่งจากเมตริกซ์ของค่าน้ำหนักปัจจัยที่ได้ จะทำให้ชุดข้อมูลเข้า ในการจัดกลุ่ม มีตัวแปรและตัวอย่างข้อมูลคือ ชุดยา ชนิดต่างๆ นอกจากนี้ เนื่องจากวิธีการจัดกลุ่มเป็นการวิเคราะห์แบบลำดับชั้น การกำหนดกลุ่มยาจึงอาศัยการพิจารณาโดยผู้วิจัยเอง

6) กลุ่มของยาที่ได้ในแต่ละกลุ่มนั้น จะนำมาหายีนที่ได้รับผลจากยาดังกล่าว (Drug Targets Detection) โดยการใช้ หลักการทางสถิติในเรื่องความแตกต่างของค่าเฉลี่ยของข้อมูลในแต่ละกลุ่ม ทั้งนี้มีสมมุติฐานว่า ค่าการแสดงออกของยีนแต่ละตัวที่มีความสัมพันธ์อยู่ในกลุ่มยากลุ่มใดกลุ่มหนึ่ง จะต้องมีค่าเฉลี่ยของข้อมูลที่แตกต่างกับ กลุ่มยากลุ่มอื่นๆ ที่ระดับนัยสำคัญที่เชื่อถือได้ นั่นคือ วิธีการวิเคราะห์ จะพิจารณา ความแตกต่างของค่าเฉลี่ยของค่าการแสดงออกของยีนแต่ละตัวเมื่อกำหนดกลุ่มของยาขึ้นมากลุ่มหนึ่ง เทียบกับ ค่าความแตกต่างของค่าเฉลี่ยของค่าการแสดงออกของยีนดังกล่าว ในชุดยาอื่นๆ ที่ไม่ได้อยู่ในกลุ่มที่ระบุในตอนต้น ซึ่งวิธีทดสอบความแตกต่างของค่าเฉลี่ยของยีนแต่ละตัวนั้นจะเรียกว่า ที-เทส (t-test) โดยค่าทางสถิติที่ได้จากการทดสอบและเป็นตัวตัดสินก็คือ ค่า พีแวลยู (p-value) ทั้งนี้โดยหลักการพื้นฐานจะเลือกยีนที่มีค่า พีแวลยู น้อยกว่า 0.05 ในการวิเคราะห์จะต้องพิจารณาทุกๆ ยีนที่เกี่ยวข้องและ ทุกๆ กลุ่มยา ซึ่งผลที่ได้จะทำให้มียีนบางตัวอาจได้รับผลจากยาได้ในหลายๆ กลุ่มยา อย่างไรก็ตามในรายละเอียดของเทคนิควิธีการนี้เป็นวิธีการพื้นฐานที่รับทราบกันโดยทั่วไป ดังนั้นวิทยานิพนธ์ฉบับนี้จะไม่กล่าวถึง แต่จะอาศัย ฟังก์ชันสำเร็จรูป จากโปรแกรมคอมพิวเตอร์มาใช้วิเคราะห์

7)นำกลุ่มยีนที่ได้ไปพิจารณาหาพาทย์เวย์ที่กลุ่มของยาซึ่งเกี่ยวข้องส่งผลกระทบ(Pathway Detection)โดยชุดข้อมูลยีนออนโทโลยีซึ่งในการวิเคราะห์จะแยกพิจารณาออกเป็น 3 ออนโทโลยีหลัก

- ผลการวิเคราะห์

ชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์วัณโรค แสดงตัวอย่างของข้อมูลได้ดังตาราง 4.10

ตาราง 4.10 เมตริกซ์แสดงตัวอย่างข้อมูลดีเอ็นเอไมโครอาร์เรย์วัณโรค (*Mycobacterium tuberculosis*)

	Drugs ID						
Genes	**GSM28217**	**GSM28218**	**GSM28219**	**GSM28220**	**GSM28221**	**GSM28222**	**GSM28223**
Rv0001	0.644006	0.434344	0.196201	0.168632	-0.258616	-0.142778	-1.488008
Rv0002	-0.922116	-0.831289	-0.873130	-1.381421	-1.006963	-1.589476	-0.204861
Rv0003	NA	NA	NA	2.262851	NA	1.945534	NA
Rv0004	1.371151	NA	1.673556	NA	NA	1.553278	1.859460
Rv0005	NA	NA	NA	NA	NA	NA	NA
Rv0006	0.231554	0.187419	NA	0.581418	0.657052	NA	0.173467

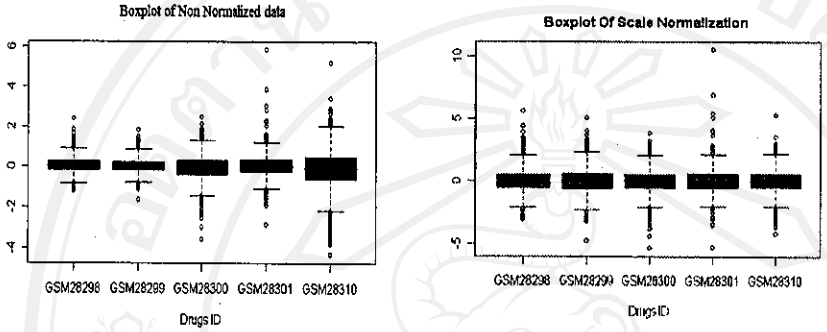
จากตาราง 4.10 ในแต่ละแถวแสดงยีน ส่วนคอลัมน์แสดง หมายเลขของชุดยา (Drug's ID) จากค่าการแสดงออกของยีนในตาราง จะเห็นว่ามีข้อมูลบางชุดยีน และบางตัวยา ที่ค่าการแสดงออกของข้อมูลขาดหายไป (Missing Value) ซึ่งแสดงได้จากเครื่องหมาย "NA" และเพื่อที่จะกรองข้อมูล ส่วนที่ขาดหายนี้ออกไป จึงทำการหาความถี่ของยีนที่ขาดหายไป ในชุดยาแต่ละตัว ซึ่งแสดงออกมาเป็นร้อยละของจำนวนยีนทั้งหมด และร้อยละของความถี่ดังกล่าวจะถูกนำมาใช้ในการพิจาณาเพื่อกรองเอาชุดยาซึ่งมีร้อยละของความถี่น้อยกว่าหรือเท่ากับร้อยละของความถี่ที่เราสนใจออกไป นอกจากนี้จากชุดยาที่เหลืออยู่ จะทำการกรองยีนที่มีข้อมูลขาดหายออกไปทั้งหมด แสดงร้อยละของความถี่ของข้อมูลที่ขาดหายไปในแต่ละชุดยา จำนวนชุดยาที่เหลือหลังจากการกรองขั้นแรก และจำนวนชุดยีนที่เหลือหลังจากการกรองในขั้นที่สองได้ดังตาราง 4.11

ตาราง 4.11 ความถี่ของข้อมูลที่เหลืออยู่หลังจากการกรองข้อมูลที่ขาดหายออกไป

เปอร์เซ็นต์ความถี่ของข้อมูลที่ขาดหายไปใน แต่ละชุดยา	จำนวนชุดยาที่เหลือหลังจากการกรอง	จำนวนชุดยีนที่เหลือหลังจากการกรอง
1%	4	3,827
2%	49	2,940
3%	87	2,183
4%	125	1,524
5%	159	1,143
6%	180	864
7%	205	681
8%	228	548
9%	257	413
10%	269	338

จากตาราง 4.11 จะกรอง เอาชุดยา ซึ่งมีความถี่ของยีนที่ข้อมูลขาดหายไม่มากกว่า 5 เปอร์เซ็นต์ จะได้จำนวนชุดยาที่เหลืออยู่สำหรับนำไปวิเคราะห์ที่ 159 ชุดยา และ จำนวนยีนที่เหลืออยู่ที่ 1,143 ยีน

ดังนั้นชุดข้อมูลที่ได้จากการกรองที่ 159 ชุดยา จะกำหนดให้เป็นตัวแปร ส่วนยีนซึ่งมีจำนวน 1,143 ยีน จะกำหนดให้เป็นตัวอย่างข้อมูล หลังจากนอร์มอลไลซ์เซชันข้อมูลโดยวิธีการสเกล นอร์มอลไลซ์เซชัน ผลของการนอร์มอลไลซ์เซชัน จะทำให้ค่าการแสดงออกของยีน ในยาแต่ละตัว มีค่าเฉลี่ยเท่ากับ 0 และมีความแปรปรวนเท่ากับ 1 ซึ่งแสดงได้ในลักษณะของ กราฟหรือเรียกว่า บ๊อกพล็อต (Box Plot) ดังรูป 4.16



รูป 4.16 บ๊อกพล็อตของตัวอย่างชุดยา 5 ชนิด ก่อนและหลังนอร์มอลไลซ์เซชันข้อมูล (Non Normalization Data and Scale Normalization Data)

จากรูป 4.16 แกนตั้งแสดงค่าเฉลี่ยของข้อมูล และขนาดของแท่งกราฟแสดงให้เห็นถึงส่วนเบี่ยงเบนมาตรฐาน นอกจากนี้เส้นสองเส้นที่อยู่รอบกล่องแสดงให้เห็นถึงค่าการแสดงออกของยีน ผลจากกระบวนการนอร์มอลไลซ์จะเซชั่นจะช่วยปรับบรรทัดฐานของข้อมูลในยาแต่ละชุด ให้มีมาตรฐาน ซึ่งจะช่วยลดปัญหาในเรื่องของ หน่วยวัด และข้อผิดพลาดจากกระบวนการจัดเก็บข้อมูลให้ลดน้อยลง

ข้อมูลที่ผ่านการนอร์มอลไลซ์เซชัน เมื่อนำมาสร้างเป็นเมตริกซ์สหสัมพันธ์จะแสดงได้ดังตัวอย่างในตาราง 4.12

ตาราง 4.12 เมตริกซ์สหสัมพันธ์ของตัวอย่างชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์วัณโรค

Drug ID	GSM28298	GSM28299	GSM28310	GSM27862	GSM27994	GSM28013
GSM28298	1	**0.684987**	0.353959	0.096200	0.253671	0.275987
GSM28299	**0.684987**	1	0.244801	0.213711	0.194772	0.202031
GSM28310	0.353959	0.244801	1	-0.261890	0.378929	0.343558
GSM27862	0.096200	0.213711	-0.261890	1	0.028134	0.010555
GSM27994	0.253671	0.194772	0.378929	0.028134	1	**0.929920**
GSM28013	0.275987	0.202031	0.343558	0.010555	**0.929920**	1

จากตาราง 4.12 ค่าสัมประสิทธิ์สหสัมพันธ์ แสดงให้เห็นถึงความสัมพันธ์ของชุดยาแต่ละตัวกับชุดยาตัวอื่นๆ ซึ่งหากมีค่ามากแสดงว่า ชุดยาทั้ง 2 ตัวนั้นมีความสัมพันธ์กัน และน่าจะจัดกลุ่มอยู่ด้วยกันได้ ซึ่งจะเห็นว่า ชุดยา GSM28298 และ ชุดยา GSM28299 มีความสัมพันธ์กันสูง ชุดยาทั้ง 2

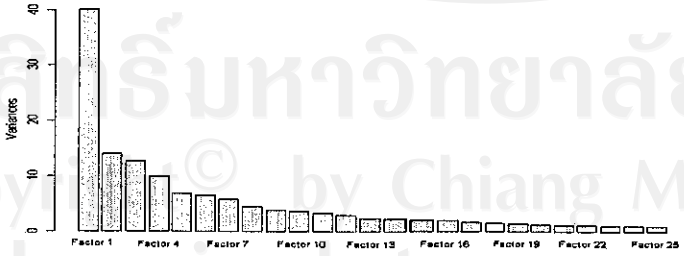
ชุดนี้จึงน่าจะอยู่ในกลุ่มเดียวกัน และชุดยา GSM27994 กับ GSM28013 ก็น่าจะอยู่ในกลุ่มเดียวกันเช่นกัน

ผลจากการวิเคราะห์ปัจจัย แสดงให้เห็นถึง ส่วนเบี่ยงเบนมาตรฐาน (S.D) ค่าความแปรปรวน (Var) ความแปรปรวนสะสม (Cum.Var) สัดส่วนของความแปรปรวน (Prop. of Var) และสัดส่วนของความแปรปรวนสะสม (Cum Prop. of Var) ของแต่ละปัจจัยได้ดังตาราง 4.13

ตาราง 4.13 ข้อสรุปจากการวิเคราะห์ปัจจัยในชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์วัณโรค

Factor	S.D	Var	Cum.Var	Prop. of Var	Cum. Prop. of Var
Factor 1	6.3328	40.1048	40.1048	0.2522	0.2522
Factor 2	3.7436	14.0143	54.1192	0.0881	0.3404
Factor 3	3.5644	12.7051	66.8242	0.0799	0.4203
Factor 4	3.1624	10.0009	76.8251	0.0629	0.4832
Factor 5	2.6073	6.7980	83.6231	0.0428	0.5259
Factor 6	2.5472	6.4880	90.1111	0.0408	0.5667
Factor 7	2.4144	5.8292	95.9403	0.0367	0.6034
Factor 8	2.1151	4.4737	100.4140	0.0281	0.6315
Factor 9	1.9528	3.8133	104.2273	0.0240	0.6555
Factor 10	1.9000	3.6100	107.8373	0.0227	0.6782
Factor 11	1.8093	3.2736	111.1110	0.0206	0.6988
Factor 12	1.6933	2.8672	113.9781	0.0180	0.7168
Factor 13	1.5556	2.4198	116.3980	0.0152	0.7321
Factor 14	1.5160	2.2984	118.6964	0.0145	0.7465
Factor 15	1.4233	2.0256	120.7220	0.0127	0.7593
Factor 16	1.3828	1.9122	122.6343	0.0120	0.7713
Factor 17	1.2809	1.6408	124.2751	0.0103	0.7816
Factor 18	1.2473	1.5558	125.8308	0.0098	0.7914
Factor 19	1.1970	1.4329	127.2638	0.0090	0.8004
Factor 20	1.1227	1.2605	128.5243	0.0079	0.8083
Factor 21	1.0862	1.1799	129.7041	0.0074	0.8157
Factor 22	1.0743	1.1542	130.8583	0.0073	0.8230
Factor 23	1.0355	1.0723	131.9306	0.0067	0.8298
Factor 24	1.0023	1.0045	132.9351	0.0063	0.8361
Factor 25	0.9652	0.9317	133.8668	0.0058	0.8419

จากตารางนำมาพล็อตใน สครีพล็อตแสดงได้ดังรูป 4.17



รูป 4.17 สครีพล็อตจากการวิเคราะห์ปัจจัยในชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์วัณโรค

จากตารางและ สครีพล็อต จะเลือกปัจจัยที่มีค่าความแปรปรวนมากกว่า 1 มาใช้ในการวิเคราะห์ต่อไป ซึ่งจะได้จำนวน 24 ปัจจัย และมีค่าความแปรปรวนสะสมเป็น 83.61 เปอร์เซ็นต์

เมตริกซ์ของค่าน้ำหนักปัจจัยจำนวน 24 ปัจจัย จะแสดงตัวอย่างได้ดังตาราง 4.14

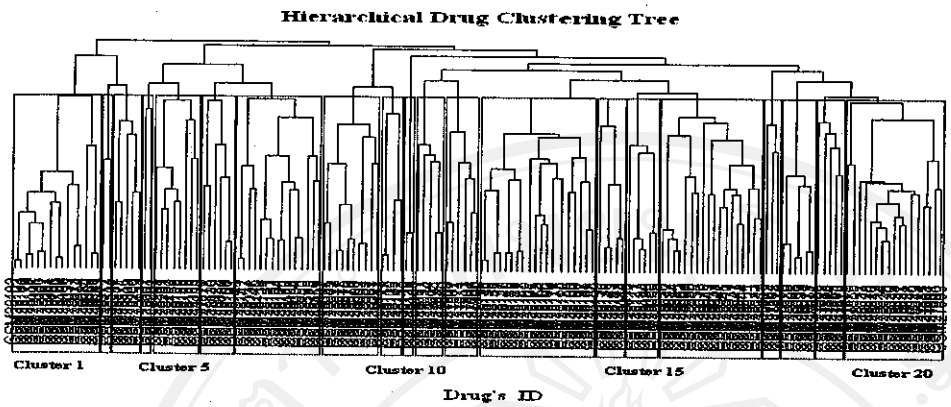
ตาราง 4.14 เมตริกซ์ของค่าน้ำหนักปัจจัยจากการวิเคราะห์ปัจจัยในชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์วัณโรค

	Factor1	Factor 2	Factor3	...	Factor24
GSM28298	-0.327567	0.030084	-0.059268	...	0.040230
GSM28299	-0.229162	0.061318	-0.088495	...	-0.166916
GSM28300	-0.229566	-0.335290	-0.120677	...	-0.045027
GSM28013	-0.189935	-0.163244	-0.050001	...	0.024060
GSM27994	-0.112073	-0.225375	-0.084350	...	-0.011679
GSM28104	-0.042683	-0.781241	-0.022499	...	-0.080183
GSM28105	-0.047431	-0.765431	-0.027159	...	-0.093487
...	...	...	...	...	...

จากตาราง 4.13 จะแสดงค่าน้ำหนักปัจจัยของชุดยาแต่ละตัวในแต่ละปัจจัย ซึ่งการที่ยามีค่าน้ำหนักปัจจัย สูงกับปัจจัยใดมากๆ แสดงว่า ยาตัวดังกล่าวมีความสัมพันธ์กับปัจจัยนั้นๆ มาก แต่หากมีค่าน้อย ก็แสดงว่าตัวยาดังกล่าวมีความสัมพันธ์กับปัจจัยนั้นๆ น้อย และเมื่อพิจารณาทุกๆ ปัจจัย การที่ยาตัวใดตัวหนึ่งจะสัมพันธ์กับยาตัวอื่นๆ หรืออยู่ในกลุ่มเดียวกันนั้น จะต้องมีรูปแบบของค่าน้ำหนักปัจจัยในทุกๆ ปัจจัยเหมือนกัน ซึ่งจะเห็นว่าเมื่อพิจารณาโดยสายตา ชุดยา GSM28298 กับ ชุดยา GSM28299 มีค่าน้ำหนักปัจจัยในทุกๆ ปัจจัยใกล้เคียงกัน ชุดยาทั้งสองชุดนี้จึงน่าจะอยู่ในกลุ่มเดียวกัน และเช่นเดียวกับใน ชุดยา GSM27994 และ GSM28013

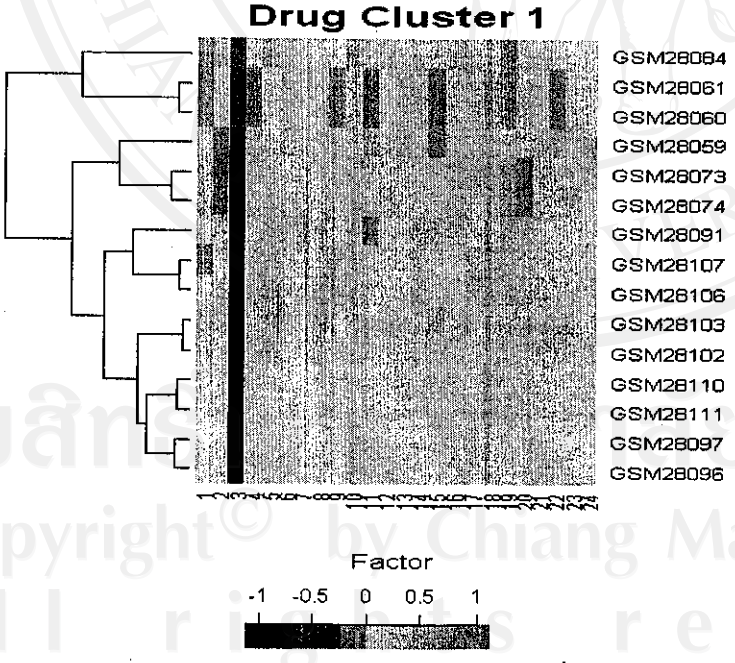
จากค่าน้ำหนักปัจจัยที่ได้ จะเห็นยาที่มีค่าน้ำหนักปัจจัยในทุกๆ ปัจจัยคล้ายคลึงกัน น่าจะอยู่ในกลุ่มเดียวกัน วิธีการจัดกลุ่มที่นำมาใช้ เพื่อพิจารณาความคล้ายกันของชุดข้อมูลยาในทุกๆ ปัจจัย นี้จะอาศัย วิธีการวิเคราะห์จัดกลุ่มแบบลำดับชั้น ซึ่งจะกำหนดให้ปัจจัยเป็นตัวแปร และ ชุดยาเป็นตัวอย่างข้อมูล

ผลของการวิเคราะห์การจัดกลุ่มแบบลำดับชั้นจะแสดงได้ดังรูป 4.18

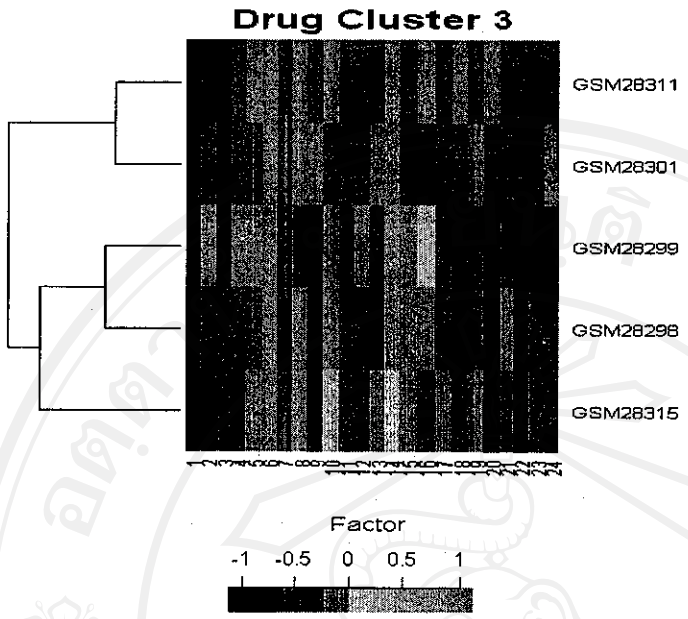


รูป 4.18 ผลของการจัดกลุ่มยาโดยวิธีการจัดกลุ่มแบบลำดับชั้นในชุดข้อมูลดีเอ็นไมโครอาร์เรย์วัณโรค

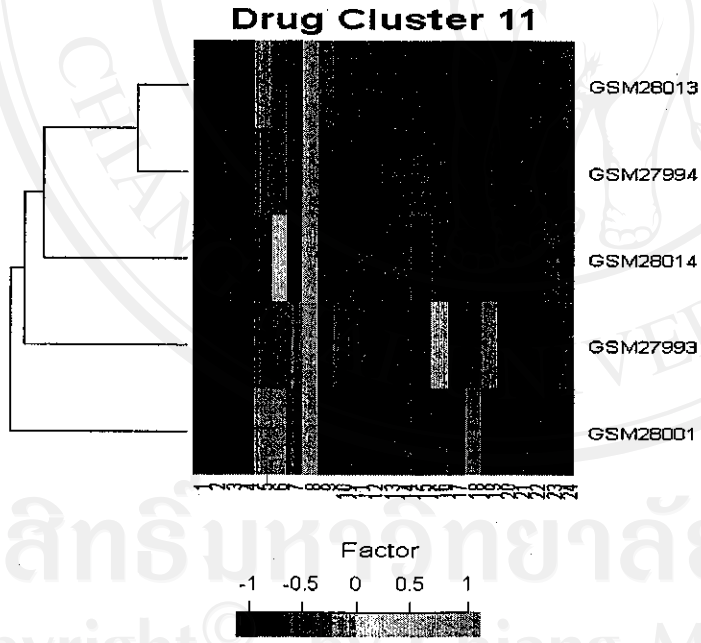
จากรูป 4.18 เป็นผลของการวิเคราะห์การจัดกลุ่มชุดยาซึ่งเมื่อพิจารณาจากกราฟต้นไม้ (Tree) ในรูป จะแสดงให้เห็นถึงการจัดกลุ่มยาเป็นลักษณะของลำดับชั้น ซึ่งความยาวของกิ่งที่เกี่ยวเนื่องกัน จะแสดงให้เห็นถึงความคล้ายคลึงกันของข้อมูล จากรูปเมื่อพิจารณาโดยสายตา ณ ระดับชั้นเดียวกัน จะจัดกลุ่มยาออกมาได้ ทั้งสิ้น 20 กลุ่มยา และเมื่อนำค่าน้ำหนักปัจจัยมาแสดงร่วมในกราฟดังกล่าว โดยแยกพิจารณาเป็นบางกลุ่ม จะแสดง ให้เห็นถึงความสัมพันธ์ของค่าน้ำหนักปัจจัยและการจัดกลุ่มข้อมูลได้ดังรูป 4.19 - 4.21



รูป 4.19 ผลของการจัดกลุ่มยาในกลุ่มที่ 1



รูป 4.20 ผลของการจัดกลุ่มยาในกลุ่มที่ 3



รูป 4.21 ผลของการจัดกลุ่มยาในกลุ่มที่ 11

จากรูป 4.19- 4.21 แสดงตัวอย่างของการจัดกลุ่มยา ใน 3 กลุ่มยาโดยอาศัยรูปแบบของค่าน้ำหนักปัจจัยที่เหมือนกัน ทั้งนี้เนื่องจากความเข้มของแถบสีจะแสดงถึงค่าน้ำหนักปัจจัยที่ยาแต่ละตัวมี

ต่อปัจจัยแต่ละปัจจัย จะเห็นว่ายาที่อยู่ในกลุ่มเดียวกันจะมีรูปแบบของแถบสีในทุกๆ ปัจจัยคล้ายกันนั่น แสดงว่ามีรูปแบบของค่าน้ำหนักปัจจัยในทุกๆ ปัจจัยเหมือนกัน และเมื่อพิจารณาที่ ชุดยา GSM28298 กับ ชุดยา GSM28299 จะเห็นว่าอยู่ในกลุ่ม 3 และ ชุดยา GSM27994 และ GSM28013 จะอยู่ในกลุ่มที่ 11

ผลของการจัดกลุ่มยา จะทำให้ได้กลุ่มยาทั้งสิ้น 20 กลุ่มยา ซึ่งแต่ละกลุ่มนั้น มีจำนวนชุดยาที่แตกต่างกันออกไป และเพื่อที่จะหาว่ากลุ่มของยาแต่ละกลุ่ม มีผลต่อยีนใดบ้าง โดยใช้วิธีการวิเคราะห์ดังที่กล่าวมาจะทำให้สามารถหายีนหรือเป้าหมายของยา (Drug Target) ที่ชุดยาในแต่ละกลุ่มส่งผลกระทบมากที่สุดได้ พร้อมกันนั้น โดยอาศัยยีนออนโทโลยีจะทำให้หาพาทเวย์ที่เกี่ยวเนื่องกับกลุ่มของยีนดังกล่าวนี้ได้ ซึ่งจะช่วยในการอธิบายความหมายได้ว่า ยาในแต่ละกลุ่มนั้นมีหน้าที่ หรือเป้าหมายอย่างไรต่อเชื้อวัณโรค ผลของการวิเคราะห์แสดงได้ดังตาราง 4.15 4.16 และ 4.17

ตาราง 4.15 ผลของการจัดกลุ่มยา การหายีนเป้าหมาย และ พาทเวย์ที่เกี่ยวข้อง ในยากลุ่มที่ 1

Drug ID	Treatments	Genes Target	
		Number	Examples
GSM28059	10ug/mL Amikacin:EtOH, 6h (mAdb expid=35691)	797 genes	Rv0013,Rv0019c,Rv0020c, Rv0040c,Rv0054,Rv0057, Rv0059,Rv0063,Rv0066c, Rv0069c,Rv0073,Rv0074, Rv0088,Rv0112,Rv0113, Rv0157,Rv0169,Rv0175, Rv0176,Rv0184,Rv0199, Rv0200,Rv0202c,Rv0203, Rv0220,Rv0222,Rv0237, Rv0241c,Rv0243,etc.
GSM28060	5ug/mL Streptomycin: EtOH, 6h (mAdb expid=35727)		
GSM28061	5ug/mL Streptomycin: EtOH, 6h (mAdb expid=35728)		
GSM28073	5ug/mL Amikacin: EtOH, 6h (mAdb expid=36042)		
GSM28074	5ug/mL Amikacin: EtOH, 6h (mAdb expid=36043)		
GSM28084	2ug/mL Streptomycin: EtOH, 6h (mAdb expid=36057)		
GSM28091	5ug/mL Tetracycline: EtOH, 6h (mAdb expid=36193)		
GSM28096	10ug/mL capreomycin: EtOH, 6h (mAdb expid=38031)		
GSM28097	10ug/mL Capreomycin: EtOH, 6h (mAdb expid=38032)		
GSM28102	50ug/mL Roxithromycin: EtOH, 6h (mAdb expid=38065)		
GSM28103	50ug/mL Roxithromycin: EtOH, 6h (mAdb expid=38066)		
GSM28106	5ug/mL Capreomycin: EtOH, 6h (mAdb expid=38069)		
GSM28107	5ug/mL Capreomycin: EtOH, 6h (mAdb expid=38070)		
GSM28110	30ug/mL Roxithromycin: EtOH, 6h (mAdb expid=38078)		
GSM28111	30ug/mL Roxithromycin: EtOH, 6h (mAdb expid=38079)		

Examples of Pathways		
Molecular Function	Biological Process	Cellular Component
DNA binding, single-stranded DNA binding, L-serine ammonia-lyase activity, iron ion binding, lyase activity, metal ion binding, nucleotide binding, nucleic acid binding, helicase activity, protein binding, ATP binding, magnesium ion binding, catalytic activity, , etc.	DNA replication, DNA repair, response to DNA damage stimulus, gluconeogenesis, protein targeting, transport, intracellular protein transport, protein transport, protein import, cation transport, metabolism, metal ion transport, protein biosynthesis, protein modification, protein metabolism, etc.	integral to membrane, membrane, cytoplasm, signal recognition particle (sensu Eukaryota), cell wall, intracellular, ribosome, ribonucleoprotein complex, large ribosomal subunit, nucleus, small ribosomal subunit, vacuole, etc

ตาราง 4.16 ผลของการจัดกลุ่มยา การหายีนเป้าหมาย และ พาทเวย์ที่เกี่ยวข้อง ในยากลุ่มที่ 3

Drug ID	Treatments	Genes Target	
		Number	Examples
GSM28298 GSM28299 GSM28301 GSM28311 GSM28315	2mM b-mercaptoethanol: DMSO, 6h (mAdb expid=49262) 2mM DTNB: DMSO (mAdb expid=49263) 50uM Nigericin: DMSO, 6h (mAdb expid=49265) 50uM Nigericin: DMSO, 6h (mAdb expid=49333) 0.1mM GSNO/10ug/mL menadione: DMSO, 6h (mAdb expid=49337)	353 genes	Rv0015c,Rv0016c,Rv0019c, Rv0031,Rv0038,Rv0039c, Rv0044c,Rv0046c,Rv0054, Rv0074,Rv0088,Rv0093c, Rv0096,Rv0103c,Rv0113,Rv0145, Rv0153c,Rv0158,Rv0183,Rv0198c, Rv0200,Rv0215c,Rv0217c,Rv0224c, Rv0261c,Rv0313,Rv0317c,Rv0332, Rv0335c,Rv0337c,Rv0345,Rv0373c, Rv0386,Rv0396,Rv0398c,Rv0413, etc.

Examples of Pathways		
Molecular Function	**Biological Process**	**Cellular Component**
nucleotide binding, protein kinase activity, protein serine/threonine kinase activity, ATP binding, kinase activity, transferase activity, inositol-3-phosphate synthase activity, DNA binding, single-stranded DNA binding, magnesium ion binding, catalytic activity, copper-exporting ATPase activity, ATPase activity, etc.	protein amino acid phosphorylation, myo-inositol biosynthesis, phospholipid biosynthesis, DNA replication, DNA repair, response to DNA damage stimulus, cation transport, metabolism, metal ion transport, carbohydrate metabolism, lipopolysaccharide core region biosynthesis, biosynthesis, glyoxylate cycle, etc.	integral to membrane, membrane, cytoplasm, signal recognition particle (sensu Eukaryota), cell wall, intracellular, ribosome, small ribosomal subunit, ribonucleoprotein complex, large ribosomal subunit, vacuole, molybdopterin synthase complex, etc.

ตาราง 4.17 ผลของการจัดกลุ่มยา การหายีนเป้าหมาย และ พาทเวย์ที่เกี่ยวข้อง ในยากลุ่มที่ 11

Drug ID	Treatments	Genes Target	
		Number	Examples
GSM27993 GSM27994 GSM28001 GSM28013 GSM28014	pH4.8:pH6.8 (2h) (mAdb expid=24193) pH4.8:pH6.8 (2h) (mAdb expid=24194) pH4.8:pH6.8 (7h) (mAdb expid=24206) pH5.6:pH6.8 (4h) (mAdb expid=24224) pH5.6:pH6.8 (7h) (mAdb expid=24226)	393 genes	Rv0013,Rv0019c,Rv0020c,Rv0040c, Rv0054,Rv0057,Rv0059,Rv0063, Rv0066c,Rv0069c,Rv0073,Rv0074, Rv0088,Rv0112,Rv0113,Rv0157, Rv0169,Rv0175,Rv0176,Rv0184, Rv0199,Rv0200,Rv0202c,Rv0203, Rv0220,Rv0222,Rv0237,Rv0241c, Rv0243,Rv0263c,Rv0264c,Rv0268c, Rv0285,Rv0287,Rv0320,Rv0357c, Rv0366c, etc

Examples of Pathways		
Molecular Function	**Biological Process**	**Cellular Component**
DNA binding, single-stranded DNA binding, L-serine ammonia-lyase activity, iron ion binding, lyase activity, metal ion binding, nucleotide binding, nucleic acid binding, helicase activity, protein binding, ATP binding, sugar binding, superoxide dismutase activity ,electron transporter activity, etc.	DNA replication, DNA repair, response to DNA damage stimulus, gluconeogenesis, protein targeting, transport, intracellular protein transport, protein transport, protein import, carbohydrate metabolism, chemotaxis, acetyl-CoAbiosynthesis from acetate, superoxide metabolism, etc.	membrane, cytoplasm, cell wall, integral to membrane, intracellular, ribosome, ribonucleoprotein complex, large ribosomal subunit, small ribosomal subunit, proton-transporting two-sector ATPase complex, proton-transporting ATP synthase complex, etc.

จากตาราง 4.15 ตาราง 4.16 และ ตาราง 4.17 แสดงตัวอย่างส่วนหนึ่งของ ผลการวิเคราะห์ข้อมูล สำหรับการจัดกลุ่มยา การหายีนเป้าหมาย และ พาทเวย์ที่เกี่ยวข้อง ใน 3 กลุ่มชุดยา ทั้งนี้ในตารางจะแสดงถึงหมายเลขของชุดยา (Drug ID) ส่วนประกอบของชุดยา (Treatments) จำนวนยีนเป้าหมาย (Gene Target) ในแต่ละกลุ่มยา ตัวอย่างกลุ่มยีนเป้าหมาย และตัวอย่างพาทเวย์ที่เกี่ยวข้องซึ่งจะแบ่งออกเป็น 3 ยีนออนโทโลยีหลัก ได้แก่ หน้าที่ในระดับโมเลกุล (Molecular Function) กระบวนการทางชีววิทยา (Biological Process) และองค์ประกอบของเซลล์ (Cellular Component)

ผลจากการวิเคราะห์ทั้ง 20 กลุ่มเนื่องจากจำนวนชุดยา จำนวนยีนเป้าหมาย และ จำนวนของพาทย์เวย์ที่เกี่ยวข้อง ในแต่ละกลุ่มยา นั้นมีจำนวนมาก จึงไม่สามารถนำเสนอได้ทั้งหมด ทั้งนี้จะสรุปผลของการวิเคราะห์ในลักษณะของค่าจำนวนที่ได้ ทั้ง 20 กลุ่มยาดังตาราง 4.18

ตาราง 4.18 ข้อสรุปของการจัดกลุ่มยา การหายีนเป้าหมายและพาทเวย์ที่เกี่ยวข้องจำนวน 20 กลุ่ม

Group	Number of Drugs	Number of Related Genes	Number of Genes which Annotated by Gene Ontology	Number of Corresponding Pathways		
				Molecular Function	Biological Process	Cellular Component
1	15	797	189	189	151	23
2	2	237	65	98	81	17
3	5	353	84	135	96	16
4	2	156	42	75	55	12
5	8	476	115	142	121	18
6	6	466	121	141	102	22
7	15	662	160	175	143	23
8	10	609	136	184	122	21
9	4	379	73	115	87	15
10	2	437	110	147	117	18
11	5	393	87	116	96	16
12	6	370	87	136	97	13
13	19	683	145	165	121	23
14	5	506	101	134	95	20
15	6	542	134	150	128	20
16	18	637	134	181	133	18
17	3	245	80	100	87	16
18	6	486	117	131	109	19
19	5	338	337	93	76	16
20	17	690	687	174	141	21

จากตาราง 4.18 แสดงให้เห็นถึงกลุ่มยา (Group) จำนวนของยาที่อยู่ในแต่ละกลุ่ม (Number of Drugs) จำนวนของยีนเป้าหมายที่เกี่ยวข้อง (Number of Related Gene) จำนวนของยีนเป้าหมายที่พบว่าสามารถอธิบายโดยยีนออนโทโลยี (Number of Genes which Annotated by Gene Ontology) และจำนวนของพาทเวย์ที่เกี่ยวข้อง (Number of Corresponding Pathways) ซึ่งแบ่งออกเป็น 3 ออนโทโลยีหลัก

- สรุปผลการวิเคราะห์

จากข้อมูลดีเอ็นเอไมโครอาร์เรย์วัณโรค ในตอนต้น ซึ่งมีจำนวน 4,320 ยีน ใน 436 ชุดยา ผลจากการกรองข้อมูล ทำให้ได้ยีนที่นำมาใช้วิเคราะห์จำนวน 1,143 ยีน และยาจำนวน 159 ชุดยา ทั้งนี้เมื่อนำชุดข้อมูลดังกล่าวไปวิเคราะห์ปัจจัย จะได้ปัจจัยที่เป็นตัวแทนของยาทั้ง 159 ชุดยาจำนวน 24 ปัจจัย ที่ความแปรปรวน 83.61 เปอร์เซ็นต์ จากนั้นเมื่อนำค่าน้ำหนักปัจจัยทั้ง 24 ปัจจัยไปวิเคราะห์เพื่อจัดกลุ่มยาโดยวิธีวิเคราะห์การจัดกลุ่มแบบลำดับชั้น จะได้กลุ่มยาจำนวนทั้งสิ้น 20 กลุ่มยา และจากกลุ่มยาดังกล่าวเมื่อนำไปวิเคราะห์ต่อ จะสามารถหาเป้าหมายของยาในแต่ละกลุ่มได้ในลักษณะของยีนและพาทเวย์ โดยอาศัยวิธี ที-เทส และข้อมูลจากยีนออนโทโลยี ทั้งนี้ทั้งนั้น ผลจากการวิเคราะห์ที่นำเสนอยังไม่ได้หาข้อสรุปโดยข้อมูลเชิงชีววิทยาแต่อย่างใด

4.3 วิจารณ์และสรุปผล

เนื่องด้วยจุดประสงค์ของการวิเคราะห์ปัจจัย มีหลายลักษณะ และเปิดกว้างสำหรับการวิเคราะห์ข้อมูลในลักษณะต่างๆ ดังนั้น การประยุกต์ใช้การวิเคราะห์ปัจจัยในข้อมูลต่างๆ จึงมีหลากหลายรูปแบบขึ้นกับจุดประสงค์ของการวิเคราะห์ข้อมูลนั้นๆ ซึ่งก็รวมถึงข้อมูลดีเอ็นเอไมโครอาร์เรย์ ดังนั้นงานวิจัยที่ในบทนี้ จึงได้ชี้ให้เห็นถึงแนวทางการประยุกต์การวิเคราะห์ปัจจัยกับข้อมูลดีเอ็นเอไมโครอาร์เรย์ในบางลักษณะ ดังต่อนี้

1) ใช้การวิเคราะห์ปัจจัยสำหรับวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์ยีสต์ซัคคาโรไมซิสเซอริวิสิเอ เพื่อวิเคราะห์โครงสร้างของตัวแปรซึ่งได้แก่ช่วงเวลาของการไดอ๊อคซิคชิฟท์ โดยกำหนดให้ตัวอย่างของข้อมูลคือยีน ทั้งนี้วิธีการวิเคราะห์จะใช้วิธีวิเคราะห์ปัจจัยร่วมในการสกัดปัจจัยเนื่องจากวิธีนี้เหมาะสำหรับการวิเคราะห์โครงสร้างของตัวแปร โดยมีเงื่อนไขว่าจำนวนตัวแปรต้องน้อยกว่าจำนวนของตัวอย่างข้อมูล และหมุนแกนปัจจัยโดยวิธีวาริแมกซ์ จากนั้นวิเคราะห์โครงสร้างของข้อมูลจากเมตริกซ์ของค่าน้ำหนักปัจจัย นอกจากนี้ ยังใช้วิธีการวิเคราะห์ปัจจัยในการลดจำนวนตัวแปรเพื่อสังเกตการกระจายตัวของยีนในมิติที่น้อยลง ซึ่งข้อมูลชุดใหม่นี้เรียกว่า คะแนนปัจจัย วิเคราะห์โดยใช้วิธีการของบาร์ทเลทท์

2) ใช้การวิเคราะห์ปัจจัยสำหรับวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์ยีสต์ซัคคาโรไมซิสเซอริวิสิเอ เพื่อวิเคราะห์โครงสร้างของตัวแปร ซึ่งในที่นี้จะกำหนดให้ยีนเป็นตัวแปร โดยยีนที่นำมาใช้วิเคราะห์จะคัดเลือกเอายีนที่มีค่าความแปรปรวนสูง ซึ่งคาดว่าเป็นยีนที่มีนัยสำคัญ สำหรับวิธีการสกัดปัจจัยจะใช้ วิธีการวิเคราะห์องค์ประกอบหลัก เนื่องจากสามารถวิเคราะห์ได้ในกรณีที่จำนวนตัวแปรมีมากกว่าจำนวนตัวอย่าง และจากข้อ 1. ถึงแม้ว่าวิธีการที่เหมาะสมจะเป็น วิธีวิเคราะห์ปัจจัยร่วม แต่ด้วยข้อจำกัดของวิธีการที่ไม่สามารถทำได้ในกรณีที่จำนวนตัวแปรมีมากกว่ากลุ่มตัวอย่าง งานวิจัยนี้จึงเลี่ยง

ใช้วิธีการวิเคราะห์องค์ประกอบหลัก ผลของการสกัดปัจจัย จะเลือกจำนวนปัจจัยที่ 2 ปัจจัยเป็นตัวแทนของยีน ขั้นตอนต่อไปของวิธีการวิเคราะห์คือการหมุนแกนปัจจัยจะใช้วิธีการหมุนแกนปัจจัยแบบวาริแมกซ์ และจากเมตริกซ์ของค่าน้ำหนักปัจจัยที่ได้ภายหลังการหมุนแกน จะนำไปใช้ในการอธิบายความหมายของปัจจัย โดยการคัดเลือกยีนที่มีค่าน้ำหนักปัจจัยสูงที่สุดในแต่ละปัจจัยเป็นตัวอธิบายความหมาย ซึ่งความหมายของปัจจัยเหล่านี้จะอยู่ในรูปของยีนออนโทยีที่เกี่ยวข้องกับยีนดังกล่าว

3) ใช้การวิเคราะห์ปัจจัยสำหรับวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์ยีสต์ซัคคาโรไมซิสเซอริวิสิเอ เพื่อช่วยในการจัดกลุ่มยีน โดยกำหนดให้ยีนเป็นตัวแปร ทั้งนี้วิธีการวิเคราะห์ เริ่มจากการสกัดปัจจัยโดยวิธีวิเคราะห์องค์ประกอบหลักและใช้เมตริกซ์สหสัมพันธ์ในการวิเคราะห์ซึ่งจะช่วยในการเลือกจำนวนปัจจัยได้ง่ายขึ้น โดยเลือกจำนวนปัจจัยที่มีค่าความแปรปรวนมากกว่า 1 ผลของการสกัดปัจจัยทำให้ได้เมตริกซ์ของค่าน้ำหนักปัจจัย ที่อธิบายความสัมพันธ์ระหว่างยีนและปัจจัย ดังนั้นจากเมตริกซ์ดังกล่าว จะนำไปใช้เป็นข้อมูลตั้งต้นในการวิเคราะห์การจัดกลุ่มข้อมูล ด้วยวิธีวิเคราะห์แบบลำดับชั้น โดยกำหนดให้ ปัจจัยเป็นตัวแปร และยีน เป็นตัวอย่างข้อมูล ผลของการวิเคราะห์นำไปเปรียบเทียบกับกลุ่มข้อมูลของยีนที่มีการจัดกลุ่มแล้วจากผลงานวิจัยที่ได้รับการยอมรับ โดยวิธีการนี้จะเห็นว่าไม่มีการหมุนแกนปัจจัย เนื่องจากผลของการหมุนแกนปัจจัย ไม่ได้ทำให้แบบแผนของยีนที่มีกับปัจจัยแต่ละตัวเปลี่ยนไป ผลของการจัดกลุ่มข้อมูลจากเมตริกซ์ของค่าน้ำหนักปัจจัยทั้งก่อนหมุนแกนปัจจัยและภายหลักหมุนแกนปัจจัยจึงให้ผลไม่ต่างกัน

4)ใช้การวิเคราะห์ปัจจัยสำหรับวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์วัณโรค เพื่อช่วยในการจัดกลุ่มยา และการหาเป้าหมายของยา โดยกำหนดให้ยาเป็นตัวแปรและยีนเป็นตัวอย่างข้อมูล สำหรับวิธีการวิเคราะห์ จะเลือก ยีนและยาที่ข้อมูลครบสมบูรณ์มาใช้ในการวิเคราะห์ ข้อมูลที่ได้นำไปสกัดปัจจัยโดยวิธีวิเคราะห์องค์ประกอบหลัก และเลือกปัจจัยที่มีค่าความแปรปรวนมากกว่า 1 ไปใช้เป็นข้อมูลตั้งต้นในการจัดกลุ่มข้อมูลด้วยวิธีวิเคราะห์การจัดกลุ่มแบบลำดับชั้น โดยให้ปัจจัยเป็นตัวแปรและยีนเป็นตัวอย่างข้อมูล จากกลุ่มข้อมูลที่ได้เพื่อหาว่ากลุ่มของยาดังกล่าวนั้นส่งผลต่อยีนเป้าหมายตัวใดบ้าง การวิเคราะห์ในขั้นต่อไป จึงใช้หลักการทางสถิติในเรื่องความแตกต่างของค่าเฉลี่ยของข้อมูลในแต่ละกลุ่ม ทั้งนี้มีสมมุติฐานว่า ค่าการแสดงออกของยีนแต่ละตัวที่มีความสัมพันธ์อยู่ในกลุ่มยากลุ่มใดกลุ่มหนึ่ง จะต้องมีค่าเฉลี่ยของข้อมูลที่แตกต่างกับ กลุ่มยากลุ่มอื่นๆ ที่ระดับนัยสำคัญที่เชื่อถือได้ นั่นคือวิธีการวิเคราะห์ จะพิจารณา ความแตกต่างของค่าเฉลี่ยของค่าการแสดงออกของยีนแต่ละตัวเมื่อกำหนดกลุ่มของยาขึ้นมากลุ่มหนึ่ง เทียบกับ ค่าความแตกต่างของค่าเฉลี่ยของค่าการแสดงออกของยีนดังกล่าวในชุดยาอื่นๆ ที่ไม่ได้อยู่ในกลุ่มที่ระบุในตอนต้น ซึ่งวิธีทดสอบความแตกต่างของค่าเฉลี่ยของยีนแต่ละตัวนั้นจะเรียกว่า ที-เทส (t-test) โดยค่าทางสถิติที่ได้จากการทดสอบและเป็นตัวตัดสินก็คือ ค่า พีแวลยู

(p-value) ทั้งนี้โดยหลักการพื้นฐานจะเลือกยีนที่มีค่า พีแวลยู น้อยกว่า 0.05 ในการวิเคราะห์จะต้องพิจารณาทุกๆ ยีนที่เกี่ยวข้องและ ทุกๆ กลุ่มยา ซึ่งผลที่ได้จะทำให้มียีนบางตัวอาจได้รับผลจากยาได้ในหลายๆ กลุ่มยา และเพื่อที่จะหาว่ากลุ่มของยาแต่ละกลุ่มนั้นมีหน้าที่อย่างไรต่อกลุ่มยีน หรือ ส่งผลต่อตำแหน่งใดในสิ่งมีชีวิต การพิจารณาพาทย์เวย์ที่กลุ่มยาส่งผลจึงเป็นการวิเคราะห์ข้อมูลในขั้นตอนสุดท้ายที่งานวิจัยนี้ทำ โดยวิเคราะห์จากยีนออนโทโลยีที่เกี่ยวข้องกับกลุ่มยีนที่ได้จากการวิเคราะห์ในขั้นตอนก่อนหน้า

ผลการวิเคราะห์ ในลักษณะต่างๆ สรุปได้ดังนี้

1) เมื่อกำหนดให้ช่วงเวลาของกระบวนการไดอ๊อคซิคชิฟท์ของยีสต์เป็นตัวแปร การวิเคราะห์ปัจจัยในข้อมูลดีเอ็นเอไมโครอาร์เรย์นี้ จะช่วยอธิบายความสัมพันธ์ของช่วงเวลาต่างๆ ในกระบวนการไดอ๊อคซิคชิฟท์ได้ พร้อมกันนั้นยังสามารถนำเสนอข้อมูลที่อยู่ในหลายๆ ช่วงเวลา ให้อยู่ในมิติของข้อมูลที่มนุษย์สามารถรับรู้ได้ใน 2 หรือ 3 มิติ

2) เมื่อกำหนดให้ยีนเป็นเป็นตัวแปร การวิเคราะห์ปัจจัยในข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์จะสามารถหาปัจจัยที่มีผลต่อกระบวนการไดอ๊อคซิคชิฟท์และอธิบายความหมายของปัจจัยดังกล่าวได้ด้วยยีนออนโทโลยี

3) เมื่อกำหนดให้ยีนเป็นตัวแปร การวิเคราะห์ปัจจัยในข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ จะสามารถใช้ปัจจัยเป็นข้อมูลตั้งต้น ในการวิเคราะห์การจัดกลุ่มยีนโดยวิธีการวิเคราะห์แบบลำดับชั้นได้ ซึ่งผลที่ได้จากการวิเคราะห์เมื่อเทียบกับผลของการจัดกลุ่มยีนที่นำเสนอในงานวิจัยอื่นๆ พบว่าให้ผลการจัดกลุ่มไม่แตกต่างกันมากนัก แต่ข้อดีของการใช้ปัจจัยเป็นข้อมูลตั้งต้นในการจัดกลุ่มข้อมูล ก็คือวิธีการนี้จะช่วยแก้ปัญหาการวิเคราะห์การจัดกลุ่มข้อมูลดีเอ็นเอไมโครอาร์เรย์ในกรณี ที่ชุดข้อมูลมีจำนวนตัวแปรมากเกินไปได้

4) จากข้อมูลดีเอ็นเอไมโครอาร์เรย์วัณโรคที่ผ่านขั้นตอนของการกรองข้อมูลเรียบร้อยแล้ว จะประกอบไปด้วยค่าการแสดงออกของยีน จำนวน 1,143 ยีน ในชุดยาต่างๆ จำนวน 159 ชุดยา เมื่อกำหนดให้ ยา เป็นตัวแปร ผลของการวิเคราะห์ปัจจัยจะได้ปัจจัยที่เป็นตัวแทนของยาทั้งหมด จำนวน 24 ปัจจัย ที่ความแปรปรวน 83.61 เปอร์เซ็นต์ และ จากปัจจัยดังกล่าว เมื่อนำไปวิเคราะห์การจัดกลุ่มแบบลำดับชั้น เพื่อจัดกลุ่มยา จะได้กลุ่มยาจำนวนทั้งสิ้น 20 กลุ่มยา และจากกลุ่มยาดังกล่าวเมื่อนำไปวิเคราะห์ต่อ โดยวิธี ที-เทส (t-test) และ ยีนออนโทโลยี จะสามารถหาเป้าหมายของยาในแต่ละกลุ่มได้ในลักษณะของยีนและพาทเวย์

ผลจากการวิเคราะห์ทั้ง 4 ลักษณะแม้ว่าจะให้ผล ตามจุดประสงค์ที่วางไว้ แต่ผลการวิเคราะห์ที่ได้ยังไม่ถือว่าดีที่สุด และจำเป็นที่จะต้องมีการศึกษา และ หาข้อสรุปทางชีววิทยาต่อไป