บทที่ 3
การวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์ ด้วยวิธีวิเคราะห์องค์ประกอบหลัก

ชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์โดยทั่วไปนั้น มีลักษณะเป็นข้อมูลหลายตัวแปร ปัญหาพื้นฐานของการวิเคราะห์ข้อมูลในลักษณะนี้คือ ตัวแปรมีจำนวนมาก และมากเกินกว่าที่จะนำมาใช้วิเคราะห์หรืออธิบายความหมายของข้อมูลดังกล่าวได้ วิธีการวิเคราะห์องค์ประกอบหลัก เป็นวิธีการหนึ่งในเทคนิคการวิเคราะห์หลายตัวแปรที่นำมาใช้ลดจำนวนตัวแปรหรือมิติของข้อมูล ตลอดจนอธิบายความหมายข้อมูล (Data Interpretation) แนวคิดพื้นฐานของวิธีการ ได้อาศัยการพิจารณาถึงโครงสร้างภายในของตัวแปร ในการสร้างตัวแปรจำนวนน้อยๆ ขึ้นมาใหม่ โดยตัวแปรใหม่ จะต้องอธิบายความผันแปรของชุดตัวแปรขนาดใหญ่เดิมให้ได้มากที่สุด ตัวแปรที่ได้ใหม่นี้จะเป็นผลรวมเชิงเส้น (Linear Combination) ของตัวแปรดั้งเดิม และจะเรียกชุดตัวแปรใหม่ว่า องค์ประกอบหลัก

ในข้อมูลดีเอ็นเอไมโครอาร์เรย์ ลักษณะของข้อมูลจะอยู่ในรูปของเมตริกซ์ข้อมูล ที่ประกอบไปด้วยยีน และการทดลองต่างๆ (Experimentation) หรือประกอบไปด้วยยีนและตัวอย่างข้อมูล (Samples) ซึ่ง อาจจะเป็นเนื้อเยื่อ อวัยวะ หรือสัตว์ทดลองต่างๆ ทั้งนี้ทั้งนั้น ข้อมูลดังกล่าว ไม่ว่าจะให้อะไรเป็นตัวแปร จะเห็นว่า มีจำนวนมาก เช่น ยีน ในมนุษย์ที่เกี่ยวกับโรคมะเร็งชนิดลิวคีเมีย (Leukemia Cancer) อาจจะมีมากกว่า 100 ยีน หรือ ในการทดลองในกระบวนการไดอ๊อคซิคชิฟท์ (Diauxic Shift) เพื่อจะดูค่าการแสดงออกของยีนในยีสต์ ซึ่งเป็นกระบวนการเกี่ยวกับการหมัก (Fermentation) น้ำตาลกลูโคส (Glucose) ไปเป็นเอทานอล (Ethanol) ซึ่งอาศัยช่วงเวลา เป็นตัวแปรในการวัดผลถึง 7 ช่วงเวลา เป็นต้น การที่จะวิเคราะห์ข้อมูลเหล่านี้นั้น มีความจำเป็นอย่างยิ่งที่จะต้องลดตัวแปรหรือมิติ ให้น้อยลง เพื่อให้สามารถอธิบายความหมายของข้อมูลเหล่านี้ได้ง่ายขึ้น นอกจากนี้ตัวแปรที่ได้ใหม่ สามารถที่จะนำไปใช้เป็นตัวแปรตั้งต้นในการวิเคราะห์ทางสถิติ ในลักษณะอื่นๆ ได้อีก เช่น การวิเคราะห์การถดถอย การจำแนกประเภทข้อมูล การจัดกลุ่มข้อมูลเป็นต้น

3.1 หลักการของวิธีวิเคราะห์องค์ประกอบหลัก

3.1.1 เมตริกซ์ข้อมูล (Data Matrix)

การวิเคราะห์ข้อมูลใดๆก็ตาม จุดเริ่มต้นนั้นมักจะเริ่มจากลักษณะของข้อมูลดั้งต้น ซึ่งอยู่ในรูปของตารางข้อมูลโดยที่แต่ละแถวแสดงถึง กรณี (Case) วัตถุ (Object) หรือตัวอย่างข้อมูล (Sample) ที่ศึกษา ส่วนในแต่ละคอลัมน์ของข้อมูลนั้น จะเรียกว่า ตัวแปร (Variables) เช่น คุณสมบัติ (Feature) เงื่อนไขการทดลอง ที่ใช้ในการศึกษาตัวอย่างข้อมูล เป็นต้น ตารางข้อมูลเหล่านี้สามารถจัดให้อยู่ในรูปของเมตริกซ์ข้อมูล หรือ อาร์เรย์ (Array) ได้ดังนี้

X คือ เมตริกซ์ข้อมูลที่มี n แถว p คอลัมน์ และ $p \geq 1$ ดังนั้น

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix} \quad (5)$$

3.1.2 เมตริกซ์ความแปรปรวนร่วม (Covariance Matrix) และเมตริกซ์สหสัมพันธ์ (Correlation Matrix)

กำหนดให้

S คือเมตริกซ์ ความแปรปรวนร่วมของกลุ่มตัวอย่าง(Covariance Matrix)

R คือเมตริกซ์สหสัมพันธ์ของกลุ่มตัวอย่าง (Correlation Matrix)

$\bar{x}_k$ คือค่าเฉลี่ยของข้อมูล(Sample mean) ในตัวแปรที่ k ซึ่ง

$$\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk} \quad ; \quad k = 1,2,\cdots,p \quad (6)$$

s_k^2 คือค่าความแปรปรวนของข้อมูล(Sample variance) ในตัวแปรตัวที่ k ซึ่ง

$$s_k^2 = s_{kk} = \frac{1}{n}\sum_{j=1}^{n}\left(x_{jk} - \bar{x}_k \right)^2 \quad ; \quad k = 1,2,\cdots,p \quad (7)$$

s_{ik} คือค่าความแปรปรวนร่วมของข้อมูล(Sample covariance) ระหว่างตัวแปรตัวที่ i และ k ซึ่ง

$$s_{ik}=\frac{1}{n}\sum_{j=1}^{n}\left(x_{ji}-\bar{x}_i\right)\left(x_{jk}-\bar{x}_k\right)\quad;\quad i,k=1,2,\cdots,p \tag{8}$$

r_{ik} คือค่าสัมประสิทธิ์สหสัมพันธ์ของข้อมูล (Sample Correlation Coefficient) ระหว่าง ตัวแปรตัวที่ i และ k ซึ่ง

$$r_{ik}=\frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}}=\frac{\sum_{j=1}^{n}\left(x_{ji}-\bar{x}_i\right)\left(x_{jk}-\bar{x}_k\right)}{\sqrt{\sum_{j=1}^{n}\left(x_{ji}-\bar{x}_i\right)^2}\sqrt{\sum_{j=1}^{n}\left(x_{jk}-\bar{x}_k\right)^2}}\quad;\quad i,k=1,2,\cdots,p \tag{9}$$

ดังนั้น

$$S=\left(\frac{1}{n-1}\right)\left(\mathrm{X}-\bar{\mathrm{X}}\right)'\cdot\left(\mathrm{X}-\bar{\mathrm{X}}\right)=\begin{bmatrix}s_{11}&s_{12}&\cdots&s_{1p}\\s_{21}&s_{22}&\cdots&s_{2p}\\\vdots&\vdots&\ddots&\vdots\\s_{p1}&s_{p2}&\cdots&s_{pp}\end{bmatrix} \tag{10}$$

และ

$$R=\left(\frac{1}{n-1}\right)\left(\frac{\mathrm{X}-\bar{\mathrm{X}}}{SD\left(\mathrm{X}\right)}\right)'\cdot\left(\frac{\mathrm{X}-\bar{\mathrm{X}}}{SD\left(\mathrm{X}\right)}\right)=\begin{bmatrix}1&r_{12}&\cdots&r_{1p}\\r_{21}&1&\cdots&r_{2p}\\\vdots&\vdots&\ddots&\vdots\\r_{p1}&r_{p2}&\cdots&1\end{bmatrix} \tag{11}$$

จาก เมตริกซ์ ความแปรปรวนร่วม S และ เมตริกซ์สหสัมพันธ์ R ในกลุ่มตัวอย่าง เพื่อที่จะอธิบายหลักการของการวิเคราะห์องค์ประกอบหลักได้อย่างสมเหตุสมผล จึงพิจารณาข้อมูลที่ระดับประชากร (Population) เราจะให้ Σ แทน เมตริกซ์ความแปรปรวนร่วม และ ρ แทน เมตริกซ์สหสัมพันธ์ ของประชากร ซึ่งจะได้

$$\Sigma=E\left(\mathrm{X}-\mu\right)\left(\mathrm{X}-\mu\right)'=\begin{bmatrix}\sigma_{11}&\sigma_{12}&\cdots&\sigma_{1p}\\\sigma_{21}&\sigma_{22}&\cdots&\sigma_{2p}\\\vdots&\vdots&\ddots&\vdots\\\sigma_{p1}&\sigma_{p2}&\cdots&\sigma_{pp}\end{bmatrix} \tag{12}$$

โดยที่ μ คือเมตริกซ์ของค่าเฉลี่ยของประชากรในแต่ละตัวแปร และ σ_{ik} คือค่าความแปรปรวนร่วมของประชากรระหว่าง ตัวแปรที่ i และ k ซึ่งค่า σ_{ii} ในแนวทแยงของเมตริกซ์ (Diagonal) จะเป็นค่าความแปรปรวนภายในตัวแปรตัวที่ i

สำหรับค่าสัมประสิทธิ์สหสัมพันธ์ของประชากร นั้น จะหาได้จาก

$$\rho_{ik} = \frac{\sigma_{ik}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{kk}}} \quad ; \quad i,k = 1,2,\cdots,p \tag{13}$$

จะได้

$$\rho = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix} \tag{14}$$

จากเมตริกซ์สหสัมพันธ์ ค่าความแปรปรวนของประชากรในแต่ละตัวแปร จะเท่ากับ 1 และ เมื่อให้ V เป็นเมตริกซ์ทแยง (Diagonal Matrix) ซึ่ง

$$V = diag(\Sigma) = \begin{bmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{pp} \end{bmatrix} \tag{15}$$

จะได้

$$\rho = \left(V^{1/2}\right)^{-1} \Sigma \left(V^{1/2}\right)^{-1} \tag{16}$$

3.1.3 ไอเกนแวลยูและไอเกนเวกเตอร์ (Eigenvalue and Eigenvector)

เมื่อ ให้ A เป็นเมตริกซ์ขนาด $p \times p$ จะเรียก สเกลาร์ λ ที่ทำให้ $Av = \lambda v$ มีคำตอบไม่เป็นศูนย์ ว่า เป็น ไอเกนแวลยูของ A และเรียกเวกเตอร์ v ที่ไม่ใช่เวกเตอร์ศูนย์ซึ่งสอดคล้องกับ λ ว่าเป็นไอเกนเวกเตอร์ของ A

สามารถหา ไอเกนแวลยู λ ได้จากการแก้สมการ

$$|A - \lambda I| = \underline{0} \tag{17}$$

ไอเกนเวลยูแต่ละค่า ที่ได้นำไปแทนในสมการ $Av = \lambda v$ จะสามารถหาไอเกนเวกเตอร์ที่สอดคล้องกับไอเกนแวลยู ดังกล่าวได้

3.1.4 โมเดลการวิเคราะห์องค์ประกอบหลัก

องค์ประกอบหลักของข้อมูล (Principal Component) ได้จากผลรวมเชิงเส้นของตัวแปร โดยมีเงื่อนไขว่าค่าความแปรปรวนของข้อมูลในองค์ประกอบหลักต้องมีค่ามากๆจึงจะทำให้องค์ประกอบดังกล่าวยังคงให้ความหมายของข้อมูลดั้งเดิมได้ดี และองค์ประกอบหลักแต่ละองค์ประกอบต้องเป็นอิสระต่อกัน แสดงผลรวมเชิงเส้นดังกล่าวในสมการต่อไปนี้

$$\begin{aligned} Y_1 &= a_1'\mathrm{X} = a_{11}\mathrm{X}_1 + a_{12}\mathrm{X}_2 + \cdots + a_{1p}\mathrm{X}_p \\ Y_2 &= a_2'\mathrm{X} = a_{21}\mathrm{X}_1 + a_{22}\mathrm{X}_2 + \cdots + a_{2p}\mathrm{X}_p \\ &\vdots \\ Y_p &= a_p'\mathrm{X} = a_{p1}\mathrm{X}_1 + a_{p2}\mathrm{X}_2 + \cdots + a_{pp}\mathrm{X}_p \end{aligned} \qquad (18)$$

จากสมการองค์ประกอบหลักที่ได้นั้นคือ $Y_1, Y_2, \cdots, Y_p$ ซึ่งเกิดจากผลรวมเชิงเส้นของตัวแปร $X_1, X_2, \cdots, X_p$ โดยมีสัมประสิทธิ์ที่ใช้คือ a_{ik}, $i,k = 1,2,3,\cdots,p$

3.1.5 วิธีการหาองค์ประกอบหลัก

กำหนดให้

$Var(Y_i)$ คือความแปรปรวนขององค์ประกอบหลัก Y_i

$Cov(Y_i, Y_k)$ คือความแปรปรวนร่วมขององค์ประกอบหลัก Y_i และ Y_k

จากสมการ (18) จะได้

$$Var(Y_i) = a_i'\Sigma a_i \qquad i = 1,2,\cdots,p \qquad (19)$$

$$Cov(Y_i, Y_k) = a_i'\Sigma a_k \qquad i,k = 1,2,\cdots,p \mid \qquad (20)$$

ถ้า Σ, ρ คือ เมตริกซ์ความแปรปรวนร่วม และ เมตริกซ์สหสัมพันธ์ของเมตริกซ์ข้อมูล X มีคู่ของไอเกนแวลยู-ไอเกนเวกเตอร์ $(\lambda_1, e_1), (\lambda_2, e_2), \ldots, (\lambda_p, e_p)$ โดยที่ $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$

จากสมการ (19), (20) เมื่อให้ $a_i = e_i$ พบว่า

$$Var(Y_i) = e_i'\Sigma e_i = \lambda_i \qquad i = 1,2,\cdots,p \qquad (21)$$

$$Cov(Y_i, Y_k) = e_i'\Sigma e_k = 0 \qquad i,k = 1,2,\cdots,p \mid i \neq k \qquad (22)$$

จากสมการ (21) สรุปได้ว่าความแปรปรวนขององค์ประกอบหลักที่ i คือค่าไอเกนแวลยูตัวที่ i ของ Σ และ สมการ (22) ค่าความแปรปรวนร่วมขององค์ประกอบหลักตัวที่ i และ k มีค่าเป็น 0 แสดงว่าแต่ละองค์ประกอบหลักไม่มีสหสัมพันธ์กัน นั่นคือ องค์ประกอบหลักทุกตัวเป็นอิสระต่อกัน (Uncorrelated)

นอกจากนี้ยังมีการพิสูจน์และแสดงให้เห็นว่า

$$\sum_{i=1}^{p} Var(X_i) = \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} Var(Y_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p \qquad (23)$$

สมการ (23) ผลรวมของความแปรปรวนของทุกๆ ตัวแปร ในเมตริกซ์ข้อมูล X และ ผลรวมของความแปรปรวนของทุกๆ องค์ประกอบหลัก นั้นให้ค่าเท่ากัน และเท่ากับผลรวมของไอเกนแวลยูของ Σ จากหลักการ นี้เอง จึงสรุปได้ว่า องค์ประกอบหลักทั้งหมดนั้น สามารถเป็นตัวแทนของข้อมูลดั้งเดิมได้โดยที่ความเป็นข้อมูลนั้นไม่เสียไปเนื่องจากค่าความแปรปรวนของข้อมูลยังมีค่าเท่าเดิม

นอกจากนี้ เมื่อพิจารณาสัดส่วนของความแปรปรวน ของข้อมูลในแต่ละองค์ประกอบนั้น จะสามารถแสดงได้ดังนี้

$$\begin{bmatrix} \text{สัดส่วนของความแปรปรวนของ} \\ \text{ประชากรที่องค์ประกอบหลักที่ } k \end{bmatrix} = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \quad k = 1,2,3,\cdots,p \qquad (24)$$

จากสมการ(24) ถ้าหาก องค์ประกอบหลักตัวที่ k มี สัดส่วนความแปรปรวนมาก แสดงว่าองค์ประกอบหลักตัวดังกล่าวสามารถอธิบายข้อมูลได้ดีและถือเป็นตัวแทนของข้อมูลได้ และเนื่องจาก $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ แสดงว่า ความแปรปรวนขององค์ประกอบหลักที่สอดคล้องกับ λ ตัวแรกๆ จะมีค่าที่สูง ด้วยเหตุนี้เราจึงใช้หลักการดังกล่าวในการพิจารณาหาจำนวนขององค์ประกอบที่เหมาะสมสำหรับการลดมิติของข้อมูล

ผลจากการวิเคราะห์องค์ประกอบหลักนี้ สัมประสิทธิ์ e_{ik} จะเป็นพารามิเตอร์ที่บอกถึงความสำคัญของตัวแปรตัวที่ k ต่อองค์ประกอบหลักตัวที่ i ซึ่งจะช่วยในการอธิบายความหมายขององค์ประกอบหลักได้ แต่ทว่าสัมประสิทธิ์ดังกล่าวนี้ไม่ได้เป็นสัมประสิทธิ์สหสัมพันธ์ระหว่างองค์ประกอบหลัก และ ตัวแปร

สัมประสิทธิ์สหสัมพันธ์ (ρ) ระหว่างองค์ประกอบหลักตัวที่ i และตัวแปรตัวที่ k แสดงได้ดังสมการ

$$\rho_{Y_i X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \qquad (25)$$

ค่าสัมประสิทธิ์สหสัมพันธ์ และสัมประสิทธิ์ขององค์ประกอบหลัก ทั้งสองค่านี้เป็นค่าที่ช่วยในการให้ความหมายขององค์ประกอบหลักซึ่งหากมีค่ามาก แสดงว่า องค์ประกอบหลักและตัวแปรนั้นๆมีความเกี่ยวพันธ์กันอยู่

ในกระบวนการวิเคราะห์ข้อมูลจริงๆ นั้น การอ้างอิงถึงข้อมูลต่างๆ จะพิจารณาในระดับของกลุ่มตัวอย่างดังสมการในตอนต้น ทั้งนี้กระบวนการวิเคราะห์องค์ประกอบหลัก ต่างๆ นั้น ก็จะใช้หลักการพื้นฐานเช่นเดียวกับที่พิจารณาในระดับประชากร ซึ่งสรุปได้ดังนี้

1.) หาเมตริกซ์ความแปรปรวนร่วม S หรือ เมตริกซ์สหสัมพันธ์ของกลุ่มตัวอย่าง R

2.) หาไอเกนแวลยู และไอเกนเวกเตอร์ของเมตริกซ์ ในข้อ 1

3.) เลือกจำนวนองค์ประกอบหลักที่เหมาะสมเพื่อลดมิติข้อมูลซึ่งจะอธิบายต่อในหัวข้อ 3.1.6

4.) จากโมเดลของการวิเคราะห์องค์ประกอบหลักในสมการ (18) จะหาคะแนนองค์ประกอบหลัก ซึ่งเป็นเมตริกซ์ของข้อมูลใหม่ Y ที่มีมิติของข้อมูลเท่ากับจำนวนองค์ประกอบที่เลือก จะต้องแยกพิจารณาออกเป็น 2 สมการ ตามเมตริกซ์ความแปรปรวนที่ใช้นั่นคือ หากการวิเคราะห์ข้อมูลเป็นการวิเคราะห์โดยใช้เมตริกซ์ความแปรปรวนร่วม S จะหาองค์ประกอบหลักได้ดังสมการ

$$Y = \left(\mathrm{X} - 1\bar{x}\right)E \qquad (26)$$

โดยที่ E เป็นเมตริกซ์ของสัมประสิทธิ์ หรือไอเกนเวกเตอร์ที่ได้จากไอเกนแวลยูตัวที่มีค่าสูง ซึ่งได้จากกระบวนการเลือกองค์ประกอบ

สำหรับกรณีของเมตริกซ์สหสัมพันธ์

$$Y = ZE \qquad (27)$$

ซึ่ง

$$Z_{jk} = \frac{\mathrm{X}_{jk} - \overline{\mathrm{X}}_k}{\sqrt{s_k^2}} \quad j = 1,2,\cdots,n \mid k = 1,2,\cdots,p \qquad (28)$$

5.) ในการเลือกใช้เมตริกซ์ความแปรปรวนร่วมหรือ เมตริกซ์สหสัมพันธ์นั้น ขึ้นอยู่กับว่าข้อมูลที่นำมาวิเคราะห์นั้น มีความเป็นมาตรฐานหรือไม่ เช่น หน่วยการวัด สเกลการวัด เท่ากันหรือไม่ ซึ่งสามารถทำให้เป็นมาตรฐานได้โดย กระบวนการนอร์มอลไลซ์เซชัน สำหรับข้อมูลที่ไม่ได้ผ่านกระบวนการดังกล่าวและข้อมูลไม่มีความเป็นมาตรฐาน การเลือกใช้ เมตริกซ์สหสัมพันธ์ในการวิเคราะห์องค์ประกอบหลัก จะถือว่ามีความน่าเชื่อถือ เพราะในการหาเมตริกซ์สหสัมพันธ์นั้นจะมีการปรับข้อมูลให้เป็นมาตรฐานไปในตัว ส่วนข้อมูลที่มีความเป็นมาตรฐานอยู่แล้วสามารถเลือกใช้ ได้ทั้ง 2 วิธี

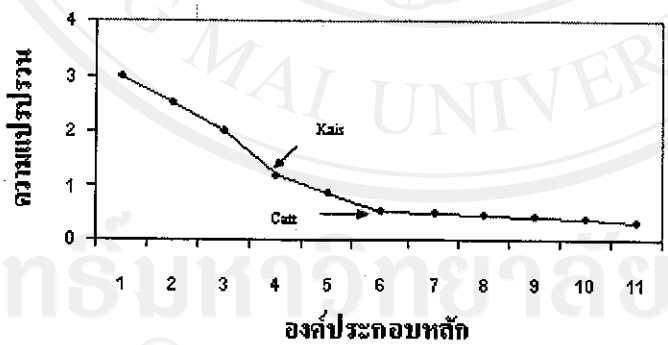
3.1.5 การเลือกองค์ประกอบหลักและจำนวนองค์ประกอบหลักที่เหมาะสม

ในการพิจารณาเลือกองค์ประกอบหลักนั้น ไม่มีหลักที่แน่นอนตายตัวทั้งนี้ ขึ้นกับวิจารณญาณของผู้วิเคราะห์เองว่าจะใช้หลักการใดในการพิจารณา ปัจจุบันมีการนำเสนอวิธีการเลือกองค์ประกอบหลักอยู่หลายวิธีได้แก่

1.) ผู้วิเคราะห์กำหนดจำนวนองค์ประกอบหลักขึ้นเอง และเลือกเอาองค์ประกอบหลักที่ได้จากค่าไอเกนแวลยูในระดับแรกๆ

2.) พิจารณาที่ค่าความแปรปรวนสะสมที่ต้องการ โดยสามารถพิจารณาได้จากเปอร์เซ็นต์ของค่าความแปรปรวนสะสมทั้งหมด ทั้งนี้ผู้วิเคราะห์กำหนดขึ้นเองว่า จะเลือกองค์ประกอบหลักตัวที่มีค่าความแปรปรวนสะสมเท่าไร ซึ่งค่าความแปรปรวนสะสมคือผลรวมของค่าความแปรปรวนขององค์ประกอบหลักตัวแรกๆ จนถึงองค์ประกอบหลักตัวที่ต้องการ

3.) พิจารณาจาก สครีพล็อต (Scree Plot) แสดงตัวอย่าง ดังรูป 3.1



รูป 3.1 ตัวอย่างสครีพล็อต

สครีพล็อต เป็นกราฟที่แสดงถึงค่าความแปรปรวนของข้อมูลในแต่ละองค์ประกอบหลัก ซึ่งเป็นค่าเดียวกับค่าไอเกนแวลยู

การเลือกองค์ประกอบหลักจาก สครีพล็อต มี 2 วิธี

3.1) เลือกพิจารณาองค์ประกอบหลักที่มีค่าของความแปรปรวนมากกว่าหรือเท่ากับ1 นำเสนอโดยไคเซอร์ (Kaiser) จึงเรียกอีกอย่างว่า วิธีไคเซอร์ ซึ่งจากตัวอย่างในรูปที่ 1 จะได้ 4 ปัจจัย

3.2) เลือกพิจารณาเอาองค์ประกอบหลักทางด้านซ้ายของแกน ณ ตำแหน่งที่ความชัน ของกราฟเริ่มราบเรียบ (Smooth) จากตัวอย่าง จะได้ 6 ปัจจัย วิธีนี้นำเสนอโดย คาท์เทลล์ (Cattell)

3.2 การวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์

ประยุกต์การวิเคราะห์องค์ประกอบหลัก กับชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ 2 ชุดข้อมูล ได้แก่ ชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ ซัคคาโรไมซิสเซอริวิสิเอ (*Saccharomyces cerevisiae*) และชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของ มะเร็งชนิด ลิวคีเมีย

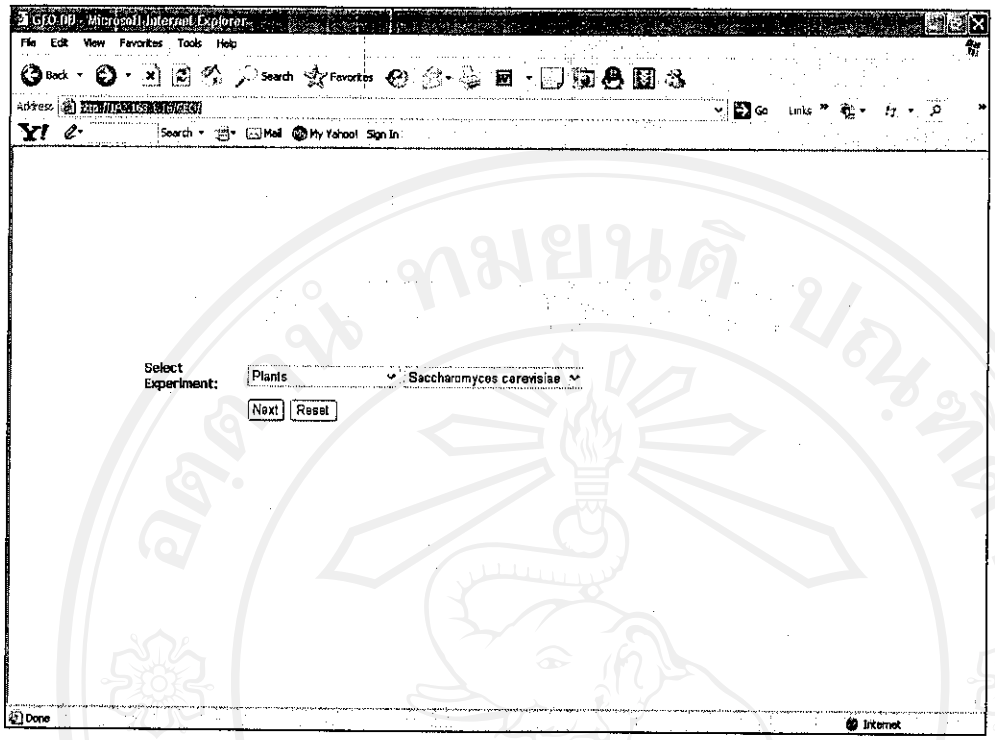
แหล่งที่มาของข้อมูล วิธีการวิเคราะห์ และผลการวิเคราะห์ จะแยกออกเป็น 2 กรณีศึกษาดังนี้

3.2.1 ชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิสเซอริวิสิเอ (*Saccharomyces cerevisiae*)

อาศัยข้อมูลจากผลงานวิจัย ซึ่งเผยแพร่ให้แก่สาธารณชนใน หัวข้อ "การสำรวจกระบวนการเมตาบอลิซึมและการควบคุมยีนในระดับการแสดงออกของยีน (Exploring the Metabolic and Genetic Control of Gene Expression on Genomic Scale)" ซึ่งวิจัยโดย โจเซฟ แอลล์ เดริซี (Joseph L. Derisi) และนักวิจัยร่วมท่านอื่นๆ ซึ่งตีพิมพ์ในวารสารไซน์ (SCIENCE) ตั้งแต่เมื่อปี ค.ศ. 1997 และข้อมูลดังกล่าวดาวน์โหลดได้จากเวบไซต์ http://www.ncbi.nlm.nih.gov/geo/

(DeRisi et al., 1997)

ข้อมูลที่ดาวน์โหลดถูกนำมาเก็บและประมวลผลในฐานข้อมูลเพื่อให้ง่ายสำหรับการวิเคราะห์ โดยฐานข้อมูลนี้พัฒนาขึ้นโดยกลุ่มวิจัยชีวสารสนเทศศาสตร์ (Bioinformatic) คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่ ในชื่อฐานข้อมูลยีนเอ็กเพรสชันโอมนิบัส (Gene Expression Omnibus Database: GEODB) ซึ่งเก็บอยู่ในโปรแกรม ทรานสคริปโตมซีเอ็มยู (Transcriptome CMU) ในเวบไซต์ http://bioinfo.science.cmu.ac.th แสดงโปรแกรมและการดึงข้อมูลจากฐานข้อมูลได้ดังต่อไปนี้

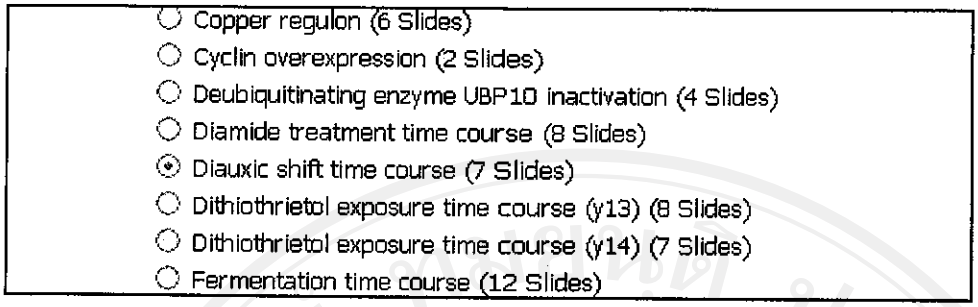


รูป 3.2 หน้าแรกโปรแกรมทรานสคริปโตมซีเอ็มยู (Transcriptome CMU)

ในรูป 3.2 แสดงหน้าแรกของโปรแกรมทรานสคริปโตมซีเอ็มยูที่มีฐานข้อมูลสำหรับใช้ในงานวิจัย ทั้งนี้ฐานข้อมูลดังกล่าวปรับปรุงมาจากฐานข้อมูลที่มีการเผยแพร่อยู่แล้วในอินเทอร์เน็ต ทั้งนี้ในการวิเคราะห์ข้อมูลชุด จะดาวน์โหลดชุดข้อมูลมาจากฐานข้อมูลดังกล่าวเอง ซึ่งดาวน์โหลดได้โดยใช้โปรแกรมทรานสคริปโตมซีเอ็มยู สำหรับคิวรี (Query) ข้อมูลที่ต้องการ และจะแสดงการเลือกชุดข้อมูลที่จะนำมาใช้วิเคราะห์ได้ดัง รูป 3.3

รูป 3.3 ดรอปดาวน์เมนูสำหรับคิวรี่ข้อมูลยีสต์

จาก รูป 3.3 เนื่องจากข้อมูลที่เกี่ยวข้องกับงานวิจัย คือข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ ซัคคาโรไมซิสเซอริวิสิเอ ดังนั้นเราจึงใช้ ดรอปดาวน์เมนู ทำการเลือกชุดข้อมูลที่ต้องการ



รูป 3.4 การเลือกชุดการทดลองที่เกี่ยวข้องกับยีสต์ ซัคคาโรไมซิสเซอริวิสิเอ

ใน รูป 3.4 ทำการเลือกการทดลองในกระบวนการไดอ๊อคซิคชิฟท์ ผลการเลือกจะได้ไมโครอาร์เรย์ที่เกี่ยวข้องเป็นจำนวน 7 สไลด์ ในแต่ละสไลด์จะเก็บข้อมูลการแสดงออกของยีน ทั้งหมดของยีสต์ สำหรับ 1 การทดลอง ข้อมูลที่ได้แสดงตัวอย่างได้ดัง ตาราง 3.1

ตาราง 3.1 ตัวอย่างชุดข้อมูลไมโครอาร์เรย์ของซัคคาโรไมซิสเซอริวิสิเอซึ่งได้จากการคิวรี่จากฐานข้อมูลจีอีโอดีบี (GEODB) โดยโปรแกรมทรานสคริปโตมซีเอ็มยู

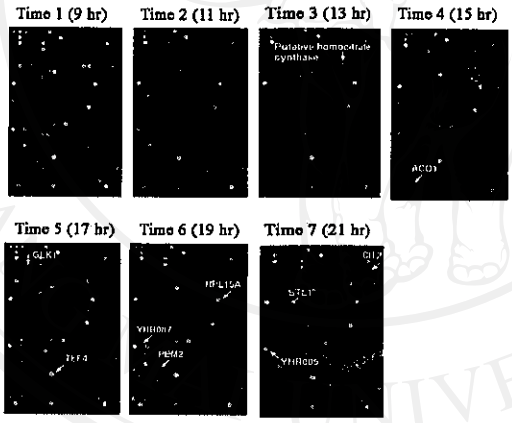
Genes ID	Time Course1 (9 hr)	Time Course2 (11 hr)	Time Course3 (13 hr)	Time Course4 (15 hr)	Time Course5 (17 hr)	Time Course6 (19 hr)	Time Course7 (21 hr)
YKL186C	0.05	0.069	-0.088	0.072	-0.034	0.935	0.698
YKL188C	0.12	-0.456	-0.443	0.079	0.005	0.997	1.446
YKL202W	-0.082	-0.231	-0.449	-0.266	-0.389	-0.727	-0.49
YKL204W	0.085	0.248	-0.176	0.049	-0.172	-0.327	-0.169
YKL206C	-0.285	-0.252	-0.346	0.144	-0.092	-0.056	-0.659
YKL208W	-0.031	-0.23	-0.095	0.261	0.428	0.502	-0.016
YKL210W	0.179	0.149	-0.128	0.2	0.105	-0.023	NA
YKL212W	0.274	0.312	0.093	0.125	-0.065	-0.47	-0.7
YKR001C	0.25	0.154	-0.221	0.171	0.082	-0.24	-1.056
YKR003W	0.008	-0.002	-0.212	-0.114	0.023	-0.432	-0.278
YKR005C	-0.241	-0.577	-0.405	-0.146	-0.229	0.302	0.195
YKR007W	-0.126	-0.266	-0.405	0.09	0.164	0.916	0.735
YKR009C	-0.039	0.114	0.217	0.414	0.39	0.813	1.21

- ลักษณะของข้อมูล

ยีสต์ (*Yeast*) เป็นจุลินทรีย์ที่รู้จักกันดีในแง่ที่ยีสต์มีคุณสมบัติในการเปลี่ยนน้ำตาลให้เป็นคาร์บอนไดออกไซด์และแอลกอฮอล์ได้ โดยเฉพาะ ยีสต์สายพันธุ์ซัคคาโรไมซิสเซอริวิสิเอ ซึ่งพบอยู่ในขนมปัง และเป็นสายพันธ์ที่นำมาใช้ในการผลิต แอลกอฮอล์ นอกจากนี้ในทางการแพทย์ยังใช้ในการผลิตยาเช่นยาเฮปาไตติส บี วัคซีน(hepatitis B vaccine) การสืบพันธ์ของยีสต์นั้นอาศัยกระบวนการแตกหน่อโดยในแต่ละเซลของยีสต์แตกหน่อ (Budding) ได้ถึง

24 เซลล์ ทั้งนี้ข้อมูลดีเอ็นไมโครอาร์เรย์ที่ใช้มาจากการทดลองกับยีนในส่วนของหน่อที่แตกออกไปนี่เอง(Budding Yeast)

ในกระบวนการเมตาโบลิซึมของยีสต์ ซึ่งเป็นกระบวนการที่เกี่ยวข้องกับการเผาผลาญที่เกิดขึ้นภายในเซลนั้น ในยีสต์จะเรียกว่ากระบวนการหมัก (Fermentation) ซึ่งเป็นกระบวนการที่เปลี่ยนสารตัวกลางเช่นแป้ง เป็นน้ำตาลกลูโคส (glucose) และเปลี่ยนน้ำตาลเป็น คาร์บอนไดออกไซด์และเอทานอลแอลกอฮอล์ ในขั้นตอนแรกจะเป็นกระบวนการหายใจที่ไม่ใช้ออกซิเจน (Anaerobic) ส่วนในขั้นตอนต่อมาใช้ออกซิเจน (Aerobic) กระบวนการที่เปลี่ยน (Shift) จากขั้นตอนหนึ่งไปยังขั้นตอนหนึ่งนี้เรียกว่า กระบวนการไดอ๊อคซิคชิฟท์ ซึ่งในการทดลองเพื่อที่จะวัดค่าการแสดงออกของยีน ผู้วิจัยได้สร้างไมโครอาร์เรย์ขึ้นมา โดยไมโครอาร์เรย์แต่ละแผ่นจะเก็บค่าการแสดงออกของยีนไว้ในสปอต(Spot) โดยเก็บไว้ในลักษณะของสีที่แสดงออกมาในยีนแต่ละตัว ซึ่งค่าสีที่ได้นี้เกิดจากกระบวนการไฮบริดจ์ไดเซชั่น (hybridization) โดยไมโครอาร์เรย์ 1 แผ่นจะวัดผล ณ เวลาใดเวลาหนึ่งในกระบวนการไดอ๊อคซิคชิฟท์ จะแสดงแผ่นไมโครอาร์เรย์จากการทดลองนี้ใน รูป 3.5



รูป 3.5 ภาพไมโครอาร์เรย์ของยีสต์ใน 7 ช่วงเวลาของกระบวนการไดอ๊อคซิคชิฟท์
(แหล่งที่มา: DeRisi et al., 1997)

จากรูป ในแต่ละสปอตที่เป็นจุดสีนั้น คือระดับการแสดงออกของยีน ที่แสดงออกมาในแต่ละยีน ซึ่งในกระบวนการไฮบริดจ์ไดเซชั่น จะกำหนดให้ สารเรืองแสงชนิดซีวาย 5 (Cy3) เป็นสีที่ติดฉลากให้กับยีนที่เป็นตัวอ้างอิงหรือยีนในสภาพปกติ ซึ่งจะเป็นยีน ณ เวลาเริ่มต้นของการทดลอง และติดฉลากสารเรืองแสงชนิดซีวาย 5 (Cy5) ให้กับยีนชุดทดลอง ซึ่งเป็นยีนที่อยู่ในช่วงของการเปลี่ยนแปลง ณ เวลาต่างๆ ซึ่งหลังจากผ่านกระบวนการดังกล่าว

แล้ว หากค่าสีที่แสดงออกมาในสปอตมีลักษณะเช่นเดียวกับ สารเรืองแสงชนิดซีวาย 3 จะแสดงว่า ยีนตัวนั้น ให้ค่าการแสดงออกที่สูงในทิศทางที่เพิ่มขึ้นจากสภาพปกติ แต่ถ้ามีลักษณะเช่นเดียวกับสารเรืองแสงชนิดซีวาย 5 แสดงว่ายีนนั้นให้ค่าการแสดงออกที่สูงในทิศทางที่ลดลงจากสภาพปกติ และหากมีค่าสี ที่มีลักษณะของสารเรืองแสงทั้งสองชนิดเท่าๆ กัน แสดงว่ายีนไม่มีการเปลี่ยนแปลงค่าการแสดงออก ซึ่งจากระดับความเข้มข้นของสีดังกล่าวนี้เอง จะเห็นว่า เมื่อเวลาผ่านไปตามกระบวนการไดอ๊อคซิคชิฟท์ ยีนแต่ละยีนจะให้ค่าการแสดงออกที่แตกต่างกันไป ซึ่งความเปลี่ยนแปลงดังกล่าวนี้จะช่วยให้เราสามารถนำมาวิเคราะห์หาความรู้จากข้อมูลเหล่านี้ได้

ในการวิเคราะห์เพื่อที่จะนำข้อมูลซึ่งแสดงออกมาเป็นลักษณะของสีมาใช้คำนวณในเชิงตัวเลขได้ นักวิจัยจึงได้แปลงข้อมูลให้เป็นตัวเลข โดยค่าตัวเลขในตาราง 3.1 คำนวณได้จากสมการ (29)

$$data_{ij} = \log(\frac{I_{ij,Cy3}}{I_{i,j,Cy5}}); i = 1,2,3,...,n, j = 1,2,3,...,p \qquad (29)$$

จากสมการข้อมูลแต่ละตัวเกิดจากค่าล็อกการิทึมของอัตราส่วนระหว่างค่าความเข้มข้นของสีที่มีลักษณะของสารเรืองแสงชนิดซีวาย 3 ต่อสารเรืองแสงชนิดซีวาย 5 ในยีนตัวที่ i ณ เวลาที่ j โดยที่จำนวนยีนทั้งหมดที่ได้จะมีจำนวน 6,153 ยีน แต่ทั้งนี้ ยีนเหล่านี้บางตัวยังไม่มีชื่อดังนั้น จากตาราง 3.1 สิ่งที่เป็นตัวกำกับยีนนั้นคือหมายเลขรหัสของยีน (Gene ID) หรืออีกนัยหนึ่งจะเรียกว่า โอเพนรีดดิ่งเฟรม (ORF) ซึ่งจะเป็นส่วนของระดับสายพันธุกรรม (Genome Sequence) ที่ตัดออกมาจากยีนทั้งหมดของเซล ทั้งนี้ค่าการแสดงออกที่ได้ในแต่ละยีนนั้นบางครั้งค่าที่ได้ เป็นค่าที่ขาดหาย (Missing value) ซึ่งอาจเป็นผลมาจากการที่ไม่สามารถวัดระดับความเข้มข้นของสีได้ การคำนวณผิดพลาด หรือ อาจเกิดจากปัญหาการเก็บข้อมูลในฐานข้อมูล ค่าที่ขาดขายไปนั้น ในตาราง 3.1 จะแสดงออกมาในสัญลักษณ์ “NA” ซึ่งในการวิเคราะห์ข้อมูล ค่าเหล่านี้จะต้องตัดทิ้งไป

(DeRisi et al., 1997)

- วิธีการวิเคราะห์

1) วิเคราะห์ข้อมูลโดยใช้ฟังก์ชั่นการวิเคราะห์และพัฒนาขึ้นเอง จากโปรแกรม ภาษา R เวอร์ชัน 2.3.1ในการคำนวณ

2) ข้อมูลที่ได้จากฐานข้อมูลนำมาเข้าสู่ กระบวนการเตรียมข้อมูล เพื่อที่จะกรองข้อมูลที่ใช้ไม่ได้ ออกไป นั่นคือตัดเอายีนที่ค่าของข้อมูลขาดหายทิ้งไป ซึ่งจากข้อมูลเดิมที่มียีนทั้งสิ้น 6,153 ยีน หลังจากกรองข้อมูลออกไปเหลือเพียง 6,119 ยีน

3) เพื่อที่จะทำให้ข้อมูลมีความเป็นมาตรฐานโดยไม่ต้องผ่านกระบวนการนอร์มอลไลซ์เซชัน ในการวิเคราะห์องค์ประกอบหลักในที่นี้ จะสร้างเมตริกซ์สหสัมพันธ์แทนเมตริกซ์ความแปรปรวนร่วม ในการวิเคราะห์

4) แบ่งการวิเคราะห์ข้อมูลออกเป็น 3 การทดลอง นั่นคือ

4.1) วิเคราะห์องค์ประกอบหลัก โดยใช้ยีนทั้งหมดที่ผ่านกระบวนการกรองข้อมูลแล้ว ทั้งสิ้น 6,119 ยีน

4.2) วิเคราะห์องค์ประกอบหลักโดยใช้ ยีนออนโทโลยี ในการกำหนดกลุ่มยีน เนื่องจากยีนออนโทโลยี จะแบ่งการพิจารณา ยีนออกเป็น 3 ลักษณะ หรือ 3 ออนโทโลยีหลัก นั่นคือองค์ประกอบของเซลล์ (Cellular Component) กระบวนการทางชีวภาพ (Biological Process) และหน้าที่ในระดับโมเลกุล (Molecular Function) ในแต่ละลักษณะนั้น จะแบ่งออนโทโลยี หรือคุณสมบัติของยีนลงไปอีก เป็นหลายๆ กลุ่มย่อย ซึ่งคุณสมบัติของยีนดังกล่าวนี้เอง ที่เราจะใช้ในการกำหนดกลุ่มให้กับยีนแต่ละตัว ทั้งนี้ยีนแต่ละตัวนั้นอาจจะ มีคุณสมบัติอยู่ในหลายๆ กลุ่ม และยีนบางตัวอาจจะไม่ได้ถูกพิจารณาให้ อยู่ในออนโทโลยีใด ออนโทโลยีหนึ่ง ใน 3 ออนโทโลยีหลัก ก็เป็นได้

4.3) วิเคราะห์องค์ประกอบหลัก โดยอาศัย กลุ่มของยีนที่จัดกลุ่มแล้ว และมีอ้างอิงอยู่แล้วในงานวิจัยที่เรานำข้อมูลมาวิเคราะห์

- ผลการวิเคราะห์

  - <u>การทดลองที่ 1</u>

จากชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ทั้งสิ้น 7 ช่วงเวลา ในยีน 6,153 ยีน แสดงดัง ตาราง 3.1 หลังจากผ่านกระบวนการกรองข้อมูลโดยตัดเอายีนที่ข้อมูลขาดหายทิ้งไป ผลที่ได้คือ ชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ชุดใหม่มีจำนวนยีนทั้งสิ้น 6,119 ยีน ซึ่งมีค่าความแปรปรวนและค่ากลางของข้อมูลแสดงได้ดัง ตาราง 3.2

ตาราง 3.2 ค่ากลางและความแปรปรวนของข้อมูลดีเอ็นเอไมโครอาร์เรย์ในซัคคาโรไมซิส เซอริวิสิเอ

	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6	Time 7
Min	-2.336	-4.479	-2.302	-2.043	-6.310	-3.278	-8.950
1st Quartile	-0.168	-0.169	-0.187	-0.208	-0.301	-0.287	-0.278
Median	-0.001	0.026	0.029	-0.007	-0.032	0.233	0.205
Mean	-0.034	0.002	0.043	0.002	-0.018	0.234	0.205
3rd Quartile	0.146	0.207	0.269	0.202	0.247	0.753	0.678
Max	1.083	1.720	2.672	2.131	2.888	4.372	4.354
Var	0.081	0.136	0.120	0.132	0.258	0.813	0.853

จากตาราง 3.2 จะได้ความแปรปรวนสะสมของทุกๆช่วงเวลาเป็น 2.393 และจะหาเมตริกซ์ความแปรปรวนร่วมและเมตริกซ์สหสัมพันธ์ จากสูตรการคำนวณ (10) และ (11) โดยกำหนดให้ ตัวแปร X เก็บข้อมูลไมโครอาร์เรย์โดยแต่ละแถวคือยีนแต่ละตัว จะได้ 6,119 แถว และคอลัมน์คือตัวแปรซึ่งในที่นี้คือช่วงเวลาโดยมีทั้งหมด 7 คอลัมน์ ได้ตารางเมตริกซ์ความแปรแปรปรวนร่วมและเมตริกซ์สหสัมพันธ์ดังนี้

ตาราง 3.3 เมตริกซ์ความแปรปรวนร่วมของข้อมูลดีเอ็นเอไมโครอาร์เรย์ในซัคคาโรไมซิสเซอริวิสิเอ

	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6	Time 7
Time 1	0.081	0.063	0.016	0.013	0.022	-0.074	-0.072
Time 2	0.063	0.136	0.029	0.028	0.044	-0.053	-0.031
Time 3	0.016	0.029	0.120	0.068	0.069	0.086	0.064
Time 4	0.013	0.028	0.068	0.132	0.134	0.180	0.155
Time 5	0.022	0.044	0.069	0.134	0.258	0.211	0.189
Time 6	-0.074	-0.053	0.086	0.180	0.211	0.813	0.683
Time 7	-0.072	-0.031	0.064	0.155	0.189	0.683	0.853

ตาราง 3.4 เมตริกซ์สหสัมพันธ์ของข้อมูลดีเอ็นเอไมโครอาร์เรย์ในซัคคาโรไมซิสเซอริวิสิเอ

	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6	Time 7
Time 1	1.000	0.604	0.164	0.129	0.151	-0.289	-0.274
Time 2	0.604	1.000	0.230	0.208	0.234	-0.158	-0.090
Time 3	0.164	0.230	1.000	0.537	0.393	0.275	0.199
Time 4	0.129	0.208	0.537	1.000	0.727	0.551	0.462
Time 5	0.151	0.234	0.393	0.727	1.000	0.462	0.404
Time 6	-0.289	-0.158	0.275	0.551	0.462	1.000	0.820
Time 7	-0.274	-0.090	0.199	0.462	0.404	0.820	1.000

จากตารางเมตริกซ์ความแปรปรวนร่วมและเมตริกซ์สหสัมพันธ์ แสดงให้เห็นว่า ค่าในแนวทแยงคือค่าความแปรปรวนของข้อมูลในแต่ละช่วงเวลา และ พบว่าในช่วงเวลาที่ 1 กับช่วงเวลาที่ 2 มีความสัมพันธ์กัน ด้วยค่าความแปรปรวนร่วม 0.063 และมีค่าสัมประสิทธิ์

สหสัมพันธ์ 0.604 ซึ่งค่าเหล่านี้ถ้ามีค่ามากแสดงว่าช่วงเวลาทั้ง 2 ช่วงนั้นมีความสัมพันธ์กันมาก นั่นคือช่วงเวลาทั้งสองนี้เป็นตัวแปรที่ส่งผลต่อค่าการแสดงออกของยีน เหมือนกัน นอกจากนั้น จากค่าในแนวเส้นทแยงของเมตริกซ์จะเป็นค่าความแปรปรวนของข้อมูลในแต่ละช่วงเวลา ซึ่งจากเมตริกซ์ความแปรปรวนร่วมของข้อมูลจะได้ความแปรปรวนสะสมของทุกๆ ตัวแปร เป็น 2.393 และจากเมตริกซ์สหสัมพันธ์ค่าความแปรปรวนของแต่ละช่วงเวลาเป็น 1 ดังนั้นค่าความแปรปรวนสะสมทั้งหมดจึงเท่ากับ 7 ซึ่งเท่ากับจำนวนตัวแปรทั้งหมด

เมื่อนำเมตริกซ์ความแปรปรวนร่วม และเมตริกซ์สหสัมพันธ์ไปวิเคราะห์องค์ประกอบหลัก จะแสดงค่าความแปรปรวนขององค์ประกอบหลัก (Principal Component: PC) แต่ละตัวได้ดังตาราง 3.5

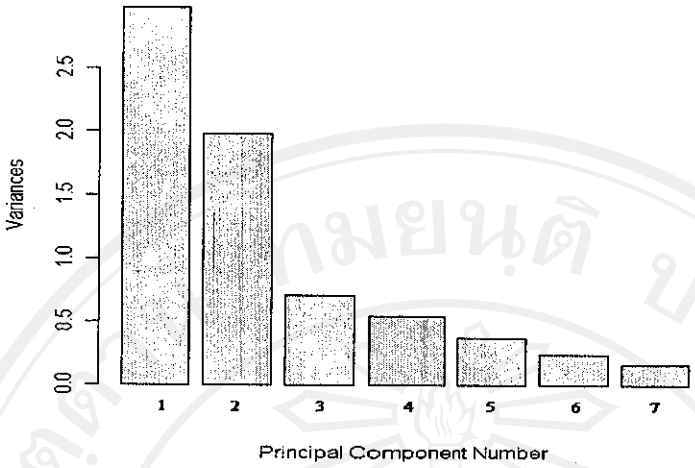
ตาราง 3.5 ค่าความแปรปรวนขององค์ประกอบหลัก จากการวิเคราะห์กับข้อมูลดีเอ็นเอไมโครอาร์เรย์ของซัคคาโรไมซิสเซอริวิสิเอ

Data Matrix	Variance							Cumulative of Variance
	1'st PC	2'nd PC	3'rd PC	4'th PC	5'th PC	6'th PC	7'th PC	
Covariance Matrix	1.639	0.324	0.165	0.105	0.086	0.038	0.035	2.393
Correlation Matrix	2.984	1.979	0.712	0.545	0.375	0.242	0.162	7.000

จาก ตาราง 3.5 ค่าความแปรปรวนสะสม (Cumulative of Variance) ของทุกๆ องค์ประกอบหลักซึ่งได้จากเมตริกซ์ความแปรปรวนร่วม (Covariance Matrix) จะเท่ากับความแปรปรวนสะสมของทุกๆตัวแปรเมื่อพิจารณาที่เมตริกซ์ความแปรปรวนร่วมซึ่งเท่ากับ 2.393 และ ค่าความแปรปรวนสะสมของทุกๆองค์ประกอบหลัก ซึ่งได้จากเมตริกซ์สหสัมพันธ์ (Correlation Matrix) จะเท่ากับความแปรปรวนสะสมของทุกๆตัวแปรเมื่อพิจารณาจากเมตริกซ์สหสัมพันธ์ซึ่งเท่ากับจำนวนตัวแปรทั้งหมด นั่นคือ 7

เมื่อพิจารณาที่ 2 องค์ประกอบแรก โดยใช้เมตริกซ์ความแปรปรวนร่วม จะได้ร้อยละของความแปรปรวนสะสม มีค่าเป็น 82.05 และ จากเมตริกซ์สหสัมพันธ์มีค่าเป็นร้อยละ 70.9

จะสร้าง สครีพล็อตเพื่อใช้ในการพิจารณาจำนวนองค์ประกอบที่เหมาะสม จากวิธีการวิเคราะห์องค์ประกอบหลักโดยใช้เมตริกซ์สหสัมพันธ์ แสดงได้ดัง รูป 3.6



รูป 3.6 สครีพล็อตจากการวิเคราะห์องค์ประกอบหลักโดยใช้เมตริกซ์สหสัมพันธ์ของข้อมูลดีเอ็น เอไมโครอาร์เรย์ในซัคคาโรไมซิสเซอริวิสิเอ

จากรูป 3.6 จะใช้หลักการของไคเซอร์ (Kaiser) ในการเลือกจำนวนองค์ประกอบหลักที่มีค่าความแปรปรวนมากกว่า 1 ซึ่งจะได้ 2 องค์ประกอบหลัก

เมตริกซ์ของสัมประสิทธิ์ หรือไอเกนเวกเตอร์ ที่ใช้สำหรับการสร้างเมตริกซ์ของข้อมูลใหม่จากโมเดลการวิเคราะห์องค์ประกอบหลักใน สมการ(18) แสดงได้ดัง ตาราง 3.6

ตาราง 3.6 เมตริกซ์ของไอเกนเวกเตอร์จากการวิเคราะห์องค์ประกอบหลักโดยใช้เมตริกซ์สหสัมพันธ์ของข้อมูลดีเอ็น เอไมโครอาร์เรย์ในซัคคาโรไมซิสเซอริวิสิเอ

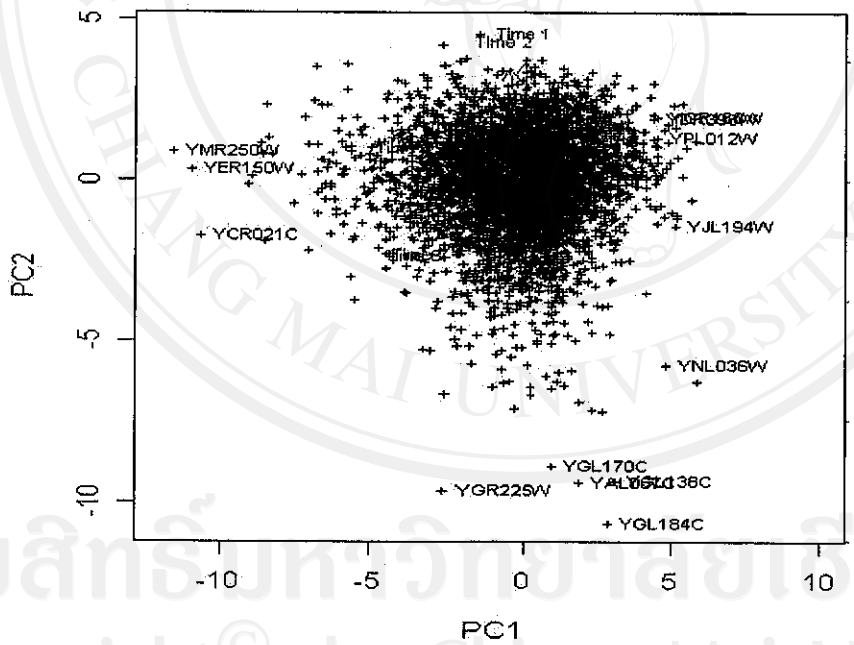
	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Time 1	0.005	0.615	-0.236	0.033	-0.746	-0.092	-0.011
Time 2	-0.089	0.580	-0.368	0.372	0.598	0.093	0.126
Time 3	-0.345	0.234	0.780	0.403	-0.009	-0.232	-0.024
Time 4	-0.502	0.139	0.115	-0.305	-0.030	0.767	-0.183
Time 5	-0.460	0.162	-0.086	-0.636	0.198	-0.557	0.028
Time 6	-0.468	-0.306	-0.200	0.225	-0.198	0.038	0.746
Time 7	-0.435	-0.301	-0.374	0.387	-0.085	-0.174	-0.627

จะใช้เวกเตอร์ ของสัมประสิทธิ์ที่องค์ประกอบหลักที่ 1 (PC1) และองค์ประกอบหลักที่ 2 (PC2) ในการสร้างชุดข้อมูลใหม่ที่มีมิติ 2 มิติ โดยมีองค์ประกอบหลักเป็นตัวแปร และแสดงผลการลดมิติ ได้ดัง ตาราง 3.7

ตาราง 3.7 ตัวอย่างของข้อมูลที่ผ่านกระบวนการวิเคราะห์องค์ประกอบหลักโดยใช้เมตริกสหสัมพันธ์ของข้อมูลดีเอ็นเอไมโครอาร์เรย์ในซัคคาโรไมซิสเซอริวิสิเอ

Gene ID	PC1	PC2
YGL138C	4.728	-9.297
YGL184C	4.446	-10.604
YGR225W	-0.868	-9.581
YAL067C	3.480	-9.348
YGL170C	2.555	-8.813
YCR021C	-9.033	-1.632
YER150W	-9.288	0.382
YMR250W	-9.826	0.925
YPL012W	6.220	1.370
YGR160W	6.247	1.986
YNL036W	6.498	-5.670
YDR398W	6.155	1.946
YJL194W	6.749	-1.343

จากตัวอย่างของข้อมูลที่ผ่านกระบวนการวิเคราะห์องค์ประกอบหลัก ข้อมูลใหม่ที่ได้จะมี 2 มิติซึ่งแสดงในกราฟ 2 มิติได้ ดัง รูป 3.7



รูป 3.7 กราฟ 2 มิติของข้อมูลดีเอ็นเอไมโครอาร์เรย์ของซัคคาโรไมซิสเซอริวิสิเอ จากการลดมิติโดยใช้กระบวนการวิเคราะห์องค์ประกอบหลัก

จากกราฟใน รูป 3.7 แสดงให้เห็นว่าผลจากการวิเคราะห์องค์ประกอบหลักทำให้เราสามารถสังเกต เห็นข้อมูลที่มีหลายๆ มิติได้ใน 2 มิติ โดยจุดแต่ละจุดแทนยีนแต่ละตัวโดยค่า

ในแนวนอนคือค่าจากองค์ประกอบหลักที่ 1 และค่าในแกนตั้งคือค่าจากองค์ประกอบหลักที่ 2 และเส้นลูกศรแทนแกนที่เป็นตัวแปรเดิมซึ่งในที่นี้คือช่วงเวลา จะเห็นว่าในช่วงเวลาที่ 1 กับช่วงเวลาที่ 2 มีลูกศรไปในทิศทางเดียวกันและมีความใกล้ชิดกัน สอดคล้องค่าประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรทั้ง 2 ที่มีค่ามาก ความสัมพันธ์ระหว่างตัวแปรนี้เอง ทำให้มีการนำเทคนิคการวิเคราะห์องค์ประกอบหลักไปใช้กับการวิเคราะห์ปัจจัย

ผลที่ได้จากการวิเคราะห์องค์ประกอบหลักทำให้เราเห็นแนวโน้ม หรือลักษณะของความผันแปรของยีนได้ง่ายขึ้น ดังนั้นหากมีการกำหนดกลุ่มให้กับยีนแล้ว การวิเคราะห์องค์ประกอบหลัก ก็อาจจะช่วยให้เราสามารถ สังเกตเห็นความแตกต่างของกลุ่มยีนดังกล่าวได้ จึงนำไปสู่การทดลองต่อไปซึ่งจะใช้ออนโทโลยีในการกำหนดกลุ่มให้กับยีน จากนั้นจะใช้กระบวนการวิเคราะห์องค์ประกอบหลักในการลดมิติ ซึ่งจะมีประโยชน์ในการสังเกตความแตกต่างของกลุ่มยีนดังกล่าวนี้ได้

o การทดลองที่ 2

การทดลองนี้จะแยกการวิเคราะห์ข้อมูลออกเป็น 3 กลุ่มที่อิสระต่อกัน นั่นคือ ออนโทโลยีองค์ประกอบของเซลล์ (Cellular Component) ออนโทโลยีกระบวนการทางชีวภาพ (Biological Process) และออนโทโลยีหน้าที่ในระดับโมเลกุล (Molecular Function)

ข้อมูลยีนออนโทโลยีของยีสต์ซัคคาโรไมซิสเซอริวิสิเอ ดาวน์โหลดได้จากเวบไซต์ http://www.yeastgenome.org/ ในไฟล์ชื่อ go_slim_mapping.tab และ แสดงยีนออนโทโลยีย่อยที่อยู่ในยีนออนโทโลยีหลักทั้ง 3 ดัง ตาราง 3.8

ตาราง 3.8 ออนโทโลยีย่อยของกลุ่มยีนของยีสต์ซัคคาโรไมซิสเซอริวิสิเอ

Cellular Component Ontology	**Biological Process Ontology**	**Molecular Function Ontology**
1. ribosome (253 genes)	1. protein biosynthesis (658 genes)	1. oxidoreductase activity (214 genes)
2. plasma membrane (200 genes)	2. transport (481 genes)	2. DNA binding (186 genes)
3. cytoplasmic vesicle (59 genes)	3. protein catabolism (98 genes)	3. ligase activity (96 genes)
4. nucleus (1,558 genes)	4. cytoskeleton organization and biogenesis (85 genes)	4. signal transducer activity (68 genes)
5. endoplasmic reticulum (339 genes)	5. protein modification. (273 genes)	5. transcription regulator activity (275 genes)
6. cytoplasm (1,805 genes)	6. cytokinesis (60 genes)	6. translation regulator activity (54 genes)
7. membrane (257 genes)	7. RNA metabolism (377 genes)	7. chaperone activity (75 genes)

ตาราง 3.8 (ต่อ) ออนโทโลยีย่อยของกลุ่มยีนของยีสต์ซัคคาโรไมซิสเซอริวิสิเอ

Cellular Component Ontology	Biological Process Ontology	Molecular Function Ontology
8. bud (83 genes)	8. DNA metabolism (272 genes)	8. lyase activity (77 genes)
9. mitochondrion (524 genes)	9. organelle organization and biogenesis (237 genes)	9. hydrolase activity (347 genes)
10. nucleolus (252 genes)	10. ribosome biogenesis and assembly (140 genes)	10. motor activity (17 genes)
11. site of polarized growth (7 genes)	11. pseudohyphal growth (49 genes)	11. protein kinase activity (147 genes)
12. chromosome (103 genes)	12. response to stress (185 genes)	12. peptidase activity (99 genes)
13. vacuole (137 genes)	13. cell cycle (201 genes)	13. transporter activity (376 genes)
14. peroxisome (49 genes)	14. meiosis (77 genes)	14. structural molecule activity (317 genes)
15. cell wall (68 genes)	15. nuclear organization and biogenesis (113 genes)	15. transferase activity (355 genes)
16. mitochondrial membrane (156 genes)	16. lipid metabolism (143 genes)	16. nucleotidyltransferase activity (72 genes)
17. Golgi apparatus (105 genes)	17. transcription (268 genes)	17. protein phosphatase activity (50 genes)
18. cell cortex (58 genes)	18. electron transport (9 genes)	18. enzyme regulator activity (120 genes)
19. membrane fraction (53 genes)	19. cell homeostasis (45 genes)	19. helicase activity (62 genes)
20. extracellular (16 genes)	20. amino acid and derivative metabolism (114 genes)	20. RNA binding (271 genes)
21. cytoskeleton (48 genes)	21. energy pathways (27 genes)	21. protein binding (223 genes)
22. endomembrane system (78 genes)	22. cell wall organization and biogenesis (127 genes)	22. isomerase activity (41 genes)
23. microtubule organizing center (43 genes)	23. conjugation (59 genes)	23. molecular_function unknown (2,371 genes)
24. cellular_component unknown (1,015 genes)	24. signal transduction (74 genes)	
	25. sporulation (48 genes)	
	26. carbohydrate metabolism (124 genes)	
	27. cellular respiration (55 genes)	
	28. vesicle-mediated transport (201 genes)	
	29. coenzymes and prosthetic group metabolism. (88 genes)	
	30. morphogenesis (24 genes)	
	31. budding (25 genes)	
	32. vitamin metabolism (44 genes)	
	33. membrane organization and biogenesis (14 genes)	
	34. biological process unknown (1,799 genes)	

จากตาราง 3.8 แสดงออนโทโลยีย่อย ในออนโทโลยีหลัก 3 ออนโทโลยี พร้อมทั้งแสดงจำนวนยีนที่อธิบายได้ด้วยออนโทโลยีดังกล่าว นำมาใช้วิเคราะห์องค์ประกอบหลักกับข้อมูลดีเอ็นเอไมโครอาร์เรย์ ได้ผลการทดลองเป็น 3 ชุด ตามออนโทโลยีหลัก ดังนี้

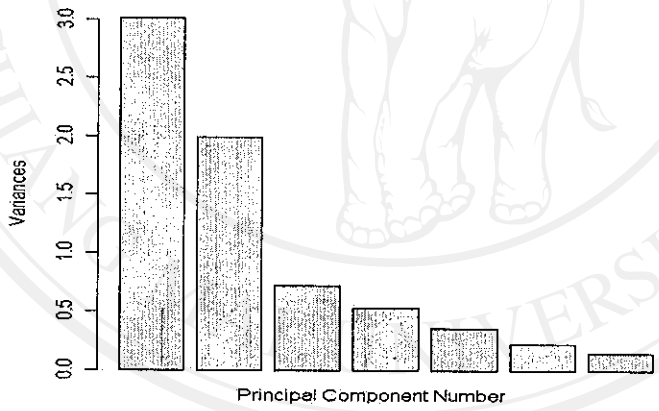
1) ผลการทดลอง ในออนโทโลยีองค์ประกอบของเซลล์ (Cellular Component)

แสดงค่าความแปรปรวนขององค์ประกอบหลักโดยใช้เมตริกซ์สหสัมพันธ์ในการวิเคราะห์ดัง ตาราง 3.9

ตาราง 3.9 ค่าความแปรปรวนและความแปรปรวนสะสมขององค์ประกอบหลักในกลุ่มยีนของซัคคาโรไมซิสเซอริวิสิเอที่อยู่ในออนโทโลยีองค์ประกอบของเซลล์

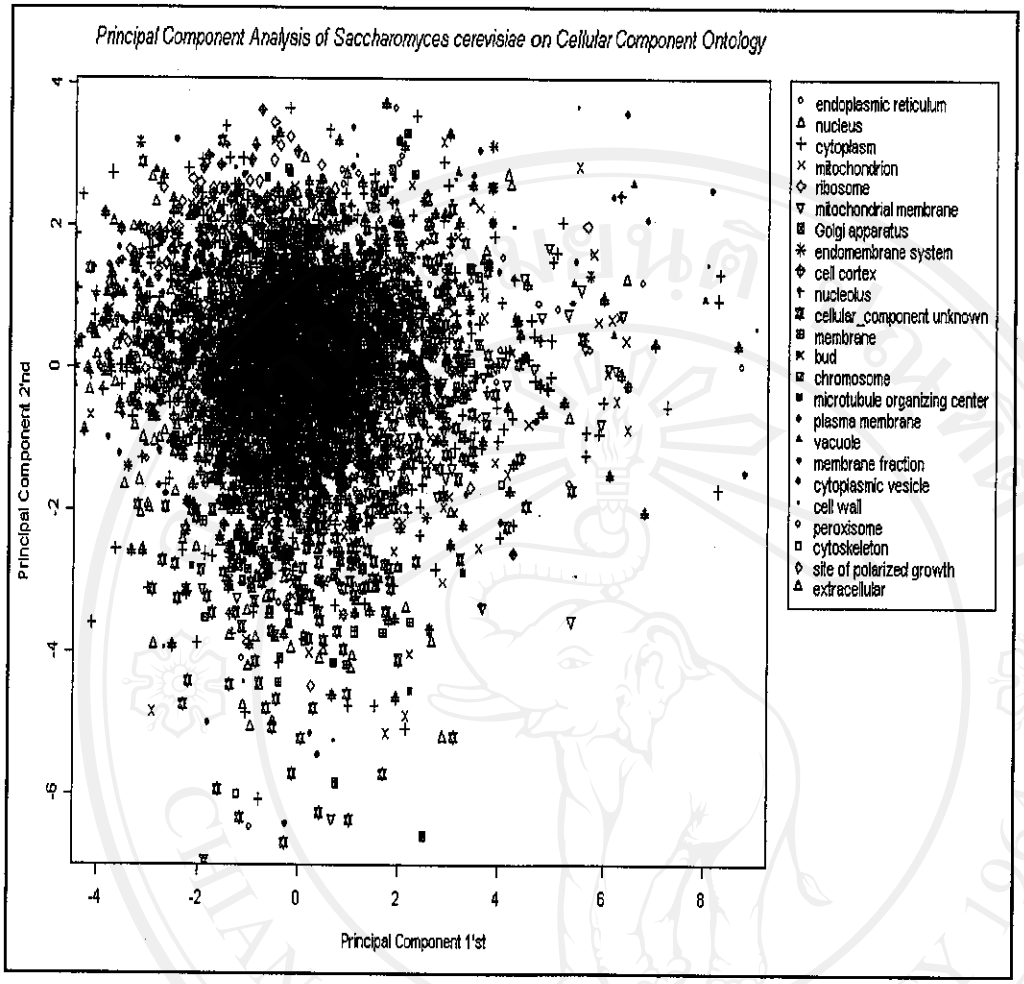
Data Matrix	1'st PC	2'nd PC	3'rd PC	4'th PC	5'th PC	6'th PC	7'th PC
Variance	3.013	1.989	0.723	0.531	0.363	0.230	0.151
Cumulative of Variance	3.013	5.002	5.725	6.256	6.619	6.849	7.000

จากตาราง 3.9 เมื่อพิจารณาที่ 2 องค์ประกอบหลักแรกจะได้ค่าความแปรปรวนสะสมเป็น 71.457 % และเมื่อนำค่าความแปรปรวนมาสร้างเป็นสครีพล็อตจะได้ดังรูป 3.8



รูป 3.8 สครีพล็อตจากการวิเคราะห์องค์ประกอบหลักในกลุ่มยีนของซัคคาโรไมซิสเซอริวิสิเอที่อยู่ในออนโทโลยีองค์ประกอบของเซลล์

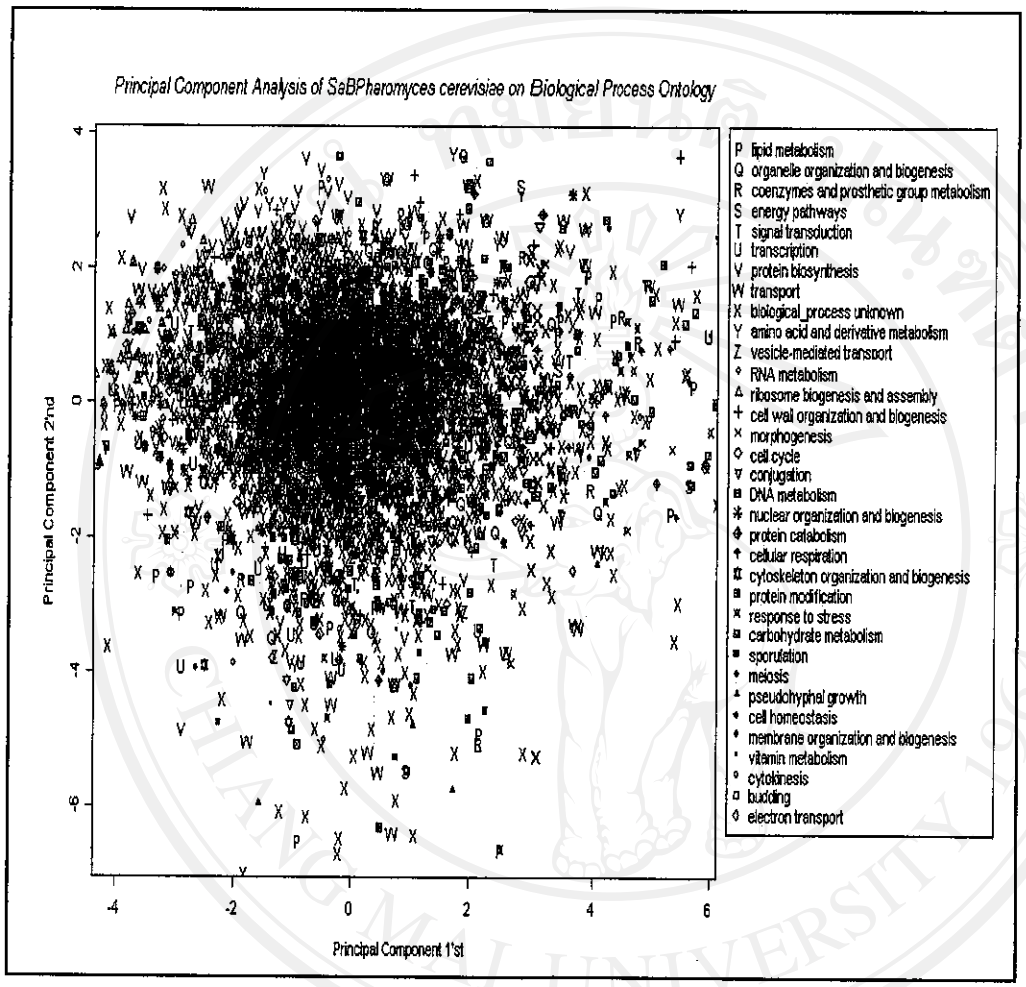
จากสครีพล็อตจะทำการเลือกจำนวนองค์ประกอบหลักที่มีค่าความแปรปรวนมากกว่า 1 จะได้ 2 องค์ประกอบหลัก นำไปหาค่าคะแนนองค์ประกอบหลัก เพื่อให้ได้ข้อมูลชุดใหม่ จะได้ภาพของการลดทอนมิติของข้อมูลในรูป 3.9



รูป 3.9 กราฟ 2 มิติของข้อมูลดีเอ็นเอไมโครอาร์เรย์ของซัคคาโรไมซิสเซอริวิสิเอจาก การลดมิติโดยกระบวนการวิเคราะห์องค์ประกอบหลัก และแยกข้อแตกต่างของข้อมูลโดยกลุ่มของออนโทโลยีใน ออนโทโลยีองค์ประกอบของเซลล์

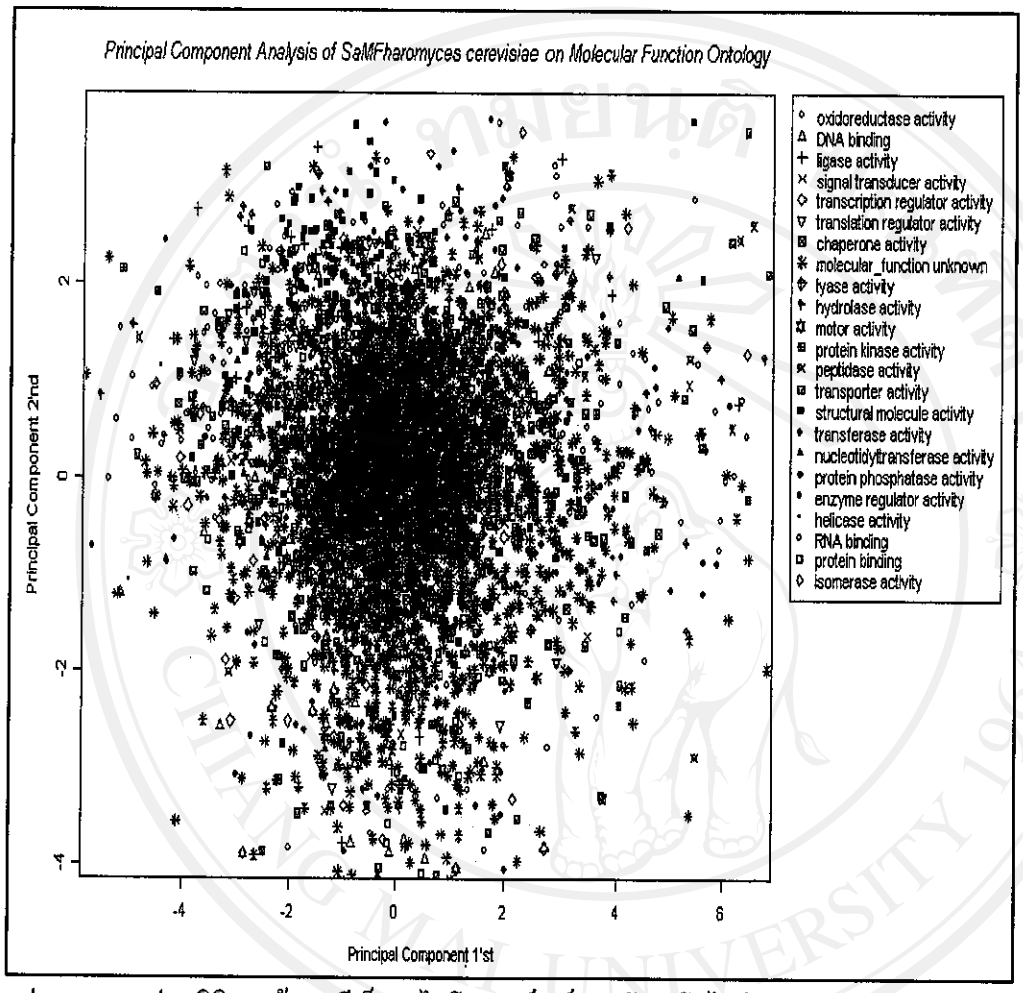
รูป 3.9 จุดแต่ละจุดแทนค่าของยีนแต่ละตัวใน 2 องค์ประกอบหลัก ซึ่งสัญลักษณ์ที่ต่างๆ กันในแต่ละจุดนั้น หมายถึงกลุ่มหรือออนโทโลยีที่แตกต่างกันด้วย จากรูปจะพบว่าการกระจายตัวของข้อมูลในแต่ละกลุ่มนั้นไม่แตกต่างกันเลย จึงน่าจะมีเหตุผลอยู่สองประการที่ทำให้เกิดผลดังกล่าวขึ้นคือ หนึ่ง ออนโทโลยีไม่ได้เป็นตัวกำหนดกลุ่มยีนที่ดี และสององค์ประกอบหลักไม่ได้ช่วยในการจำแนกกลุ่มข้อมูลได้ จากผลการทดลองที่ได้นี้เมื่อนำไปวิเคราะห์กับออนโทโลยีหลักที่เหลืออยู่พบว่าได้ผลไม่แตกต่างกันดังกราฟใน รูป 3.10-3.11

2) ผลการทดลองในออนโทโลยีกระบวนการทางชีวภาพ (Biological Process) แสดงแผนภาพ 2 มิติของข้อมูลได้ในรูป 3.10



รูป 3.10 กราฟ 2 มิติของข้อมูลดีเอ็นเอไมโครอาร์เรย์ของซัคคาโรไมซิสเซอริวิสิเอจาก การลดมิติโดยกระบวนการวิเคราะห์องค์ประกอบหลัก และแยกข้อแตกต่างของข้อมูลโดยกลุ่มของออนโทโลยีใน ออนโทโลยีกระบวนการทางชีวภาพ

3) ผลการทดลองในออนโทโลยีหน้าที่ในระดับโมเลกุล (Molecular Function) แสดงแผนภาพ 2 มิติของข้อมูลได้ดังรูป 3.11

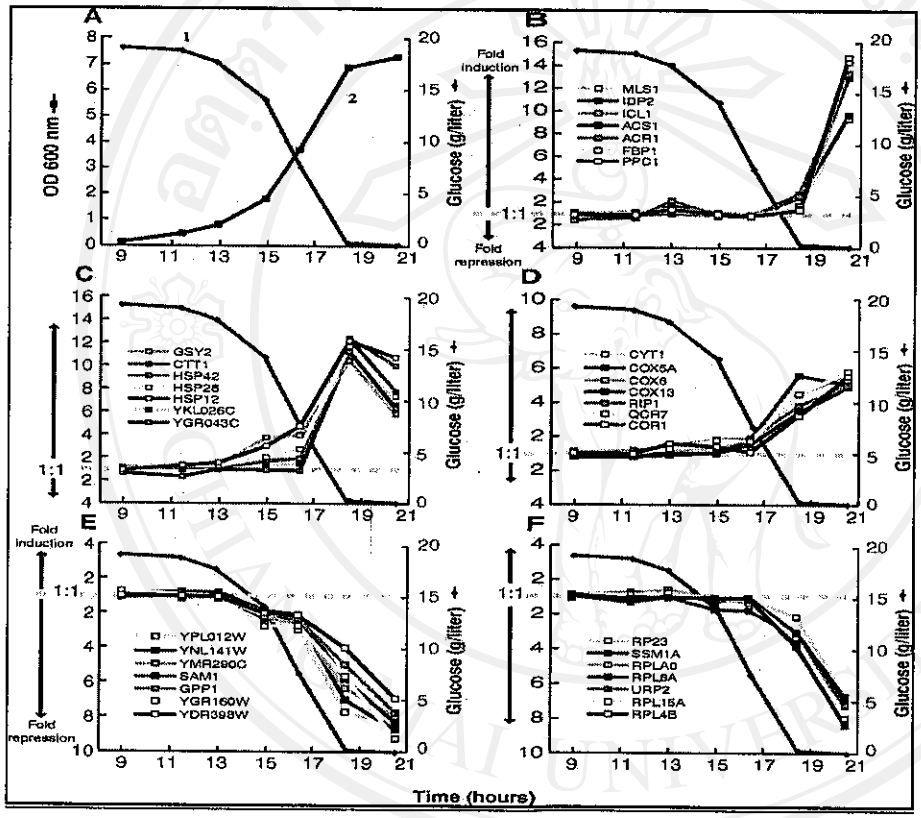


รูป 3.11 กราฟ 2 มิติของข้อมูลดีเอ็นเอไมโครอาร์เรย์ของซัคคาโรไมซิสเซอริวิสิเอจากการลดมิติโดยกระบวนการวิเคราะห์องค์ประกอบหลัก และแยกข้อแตกต่างของข้อมูลโดยกลุ่มของออนโทโลยีในออนโทโลยีหน้าที่ในระดับโมเลกุล

จากผลการวิเคราะห์ข้อมูลจะพบว่า ผลการวิเคราะห์องค์ประกอบที่ได้นั้น ไม่ได้ช่วยแยกความแตกต่างของยีนที่อยู่ออนโทโลยีที่แตกต่างกันได้ จึงทำให้ข้อมูลทั้งหมดนั้นอยู่กระจัดกระจายไปทั่วไม่แบ่งออกเป็นกลุ่มของออนโทโลยีที่ชัดเจน ด้วยเหตุนี้ในการวิเคราะห์ต่อไปจึงใช้กลุ่มยีนที่มีการค้นพบจากผลงานวิจัยมาแล้ว มาใช้ในการกำหนดกลุ่มข้อมูล เพื่อที่จะดูความแตกต่างของกลุ่มยีน จากการวิเคราะห์องค์ประกอบหลัก

o การทดลองที่ 3

จากผลงานวิจัย ที่ปรากฏในแหล่งของข้อมูลนี้ ได้อาศัยรูปแบบการแสดงออกของยีนที่คล้ายๆ กันมาใช้ในการจัดกลุ่มยีน โดยยีนที่นำมาจัดกลุ่มนั้น เป็นเพียงยีนจำนวนหนึ่งจาก 6,000 กว่ายีน ที่มีลักษณะของการแสดงออกที่คล้ายกันมากในกลุ่มยีนเดียวกัน และมีลักษณะที่แตกต่างกับกลุ่มยีนอื่นๆอย่างชัดเจน ซึ่งจากผลงานวิจัย ได้แสดงรูปแบบการแสดงออกของยีน ดังกล่าวดังแผนภาพ ใน รูป 3.12



รูป 3.12 รูปแบบการแสดงออกของยีนที่แตกต่างกันในแต่ละกลุ่ม ซึ่ง รูป 3.12 Aใน เส้นกราฟเส้นที่ 1 จะแสดงปริมาณน้ำตาลที่ลดลองเมื่อเวลาผ่านไป และเส้นที่ 2 แสดงค่าระดับการแสดงออกของยีน จากลักษณะการแสดงออกของยีนที่แตกต่างกันจะแบ่งออกเป็น 5 กลุ่มดังรูป 3.12 B - 3.12 F ซึ่งในแต่ละกลุ่มยีนนั้นจะมียีนที่เกี่ยวข้อง 7 ยีน รวมทั้งสิ้น 35 ยีน

(แหล่งที่มา: DeRisi et al., 1997)

กลุ่มของยีนในแต่ละกลุ่มแสดงได้ดัง ตาราง 3.10

ตาราง 3.10 กลุ่มยีนจากรูป 3.12

Group	Gene Name	Gene ID	Group	Gene Name	Gene ID
B	MLS1	YNL117W	E	YPL012W	YPL012W
	IDP2	YLR174W		YNL141W	YNL141W
	ICL1	YER065C		YMR290C	YMR290C
	ACS1	YAL054C		SAM1	YLR180W
	ACR1	YJR095W		GPP1	YIL053W
	FBP1	YLR377C		YGR160W	YGR160W
	PPC1	YKR097W		YDR398W	YDR398W
C	GSY2	YLR258W	F	RP23	YNL069C
	CTT1	YGR088W		SSM1A	YPL220W
	HSP42	YDR171W		RPLA0	YLR340W
	HSP26	YBR072W		RPL6A	YGL076C
	HSP12	YFL014W		URP2	YHL015W
	YKL026C	YRL026C		RPL15A	YDR418W
	YGR043C	YGR043C		RPL4B	YLL045C
D	CYT1	YOR065W			
	COX5A	YNL052W			
	COX6	YHR051W			
	COX13	YGL191W			
	RIP1	YEL024W			
	QCR7	YDR529C			
	COR1	YBL045C			

จะแสดงค่าของข้อมูลการแสดงออกได้ใน ตาราง 3.11

ตาราง 3.11 ค่าระดับการแสดงออกของยีนในชุดข้อมูลไมโครอาร์เรย์ของยีสต์ ซัคคาโรไม ซิสเซอริวิสิเอจากยีน 35 ตัวที่มีการจัดกลุ่มแล้ว

	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6	Time 7
YNL117W	-0.227	-0.274	0.343	-0.104	0.111	0.92	3.696
YLR174W	-0.092	-0.386	-1.758	0.347	0.328	1.654	3.768
YER065C	-0.01	0.4	0.465	0.309	0.446	0.68	4.354
YAL054C	-1.07	-0.514	-0.22	-0.012	-0.215	1.741	4.239
YJR095W	-0.755	-0.41	0.89	0.073	0.154	1.834	4.24
YLR377C	-0.247	0.057	0.197	-0.045	0.266	0.667	4.28
YKR097W	-0.124	-0.13	0.224	0.146	0.074	0.553	4.283
YLR258W	0.133	0.457	0.402	2.097	2.275	3.533	3.045
YGR088W	0.218	0.197	0.077	0.982	1.393	3.937	3.389
YDR171W	0.178	0.277	0.333	1.174	1.295	3.789	3.677
YBR072W	0.044	0.624	0.305	1.246	1.911	3.802	3.313
YFL014W	-0.078	0.5	0.561	1.766	2.508	3.716	3.849
YRL026C	-0.131	-0.015	-0.07	0.483	0.878	3.843	3.191
YGR043C	-0.4	-1.345	-0.282	-0.071	0.379	3.962	3.122

ตาราง 3.11(ต่อ) ค่าระดับการแสดงออกของยีนในชุดข้อมูลไมโครอาร์เรย์ของยีสต์ ซัค คาโรไมซิสเซอริวิสิเอจากยีน 35 ตัวที่มีการจัดกลุ่มแล้ว

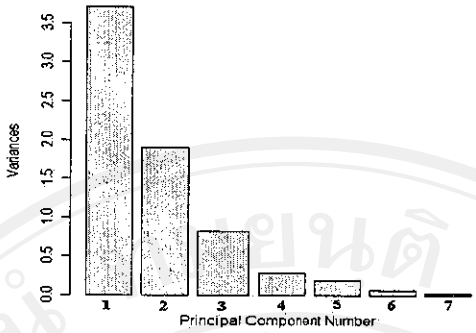
	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6	Time 7
YOR065W	0.211	0.087	0.239	0.464	1.239	1.803	2.846
YNL052W	-0.19	-0.074	-0.176	0.282	0.902	1.867	2.686
YHR051W	0.274	0.223	0.117	0.607	1.008	2.302	2.934
YGL191W	0.104	0.063	-0.031	0.245	1.196	2.618	2.781
YEL024W	0.175	0.37	0.49	1.098	1.313	2.142	2.753
YDR529C	0.21	0.266	0.305	1.094	1.38	2.851	2.996
YBL045C	0.09	0.029	0.609	0.745	0.602	1.876	2.907
YPL012W	0.146	-0.111	-0.235	-1.014	-1.405	-3.278	-2.823
YNL141W	-0.11	0.097	-0.19	-1.057	-1.132	-2.774	-2.682
YMR290C	0.191	-0.086	-0.33	-1.002	-1.319	-2.735	-2.712
YLR180W	0.248	0.358	0.149	-0.676	-0.841	-2.588	-2.706
YIL053W	0.34	0.023	-0.337	-0.743	-1.022	-2.307	-2.573
YGR160W	0.391	0.225	-0.345	-1.341	-1.342	-2.551	-2.773
YDR398W	0.071	-0.028	-0.224	-0.855	-0.999	-1.933	-5.481
YNL069C	-0.048	-0.005	0.216	0.005	0.134	-1.037	-2.297
YPL220W	0.082	0.039	0.304	-0.21	0.198	-1.55	-2.396
YLR340W	0.266	0.395	0.399	0.112	-0.219	-1.745	-2.387
YGL076C	0.147	-0.384	-0.093	-0.597	-0.616	-1.65	-2.376
YHL015W	0.223	0.312	0.011	0.147	0.038	-1.722	-2.511
YDR418W	0.275	0.243	0.218	0.051	0.017	-1.678	-2.514
YLL045C	0.243	-0.047	-0.171	0.099	0.17	-1.89	-2.692

จากข้อมูลนำมาวิเคราะห์องค์ประกอบหลักโดยสร้างเมตริกซ์สหสัมพันธ์ แสดงค่าความแปรปรวนและค่าความแปรปรวนสะสมขององค์ประกอบหลักได้ดัง ตาราง 3.12

ตาราง 3.12 ค่าความแปรปรวนและความแปรปรวนสะสมขององค์ประกอบหลักในกลุ่มยีนของซัคคาโรไมซิสเซอริวิสิเอซึ่งมีการจัดกลุ่มแล้ว

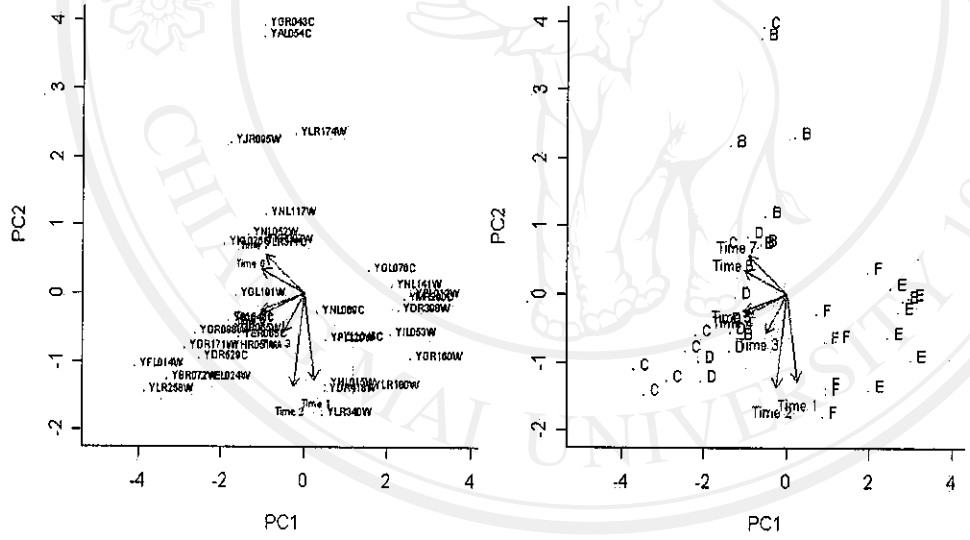
Data Matrix	1'st PC	2'nd PC	3'rd PC	4'th PC	5'th PC	6'th PC	7'th PC
Variance	3.708	1.898	0.819	0.288	0.192	0.063	0.032
Cumulative of Variance	3.708	5.606	6.425	6.713	6.905	6.968	7.000
% of Cumulative of Variance	52.97	80.09	91.79	95.90	98.64	99.54	100

จากค่าความแปรปรวนที่ 2 องค์ประกอบแรกให้ค่าความแปรปรวนสะสมที่ 80.09 % และจะสร้างเป็น สครีพล็อตได้ดัง รูป 3.13



รูป 3.13 สครีพล็อตจากการวิเคราะห์องค์ประกอบหลักในกลุ่มยีนของซัคคาโรไมซิส เซอริวิสิเอที่มีการจัดกลุ่มแล้ว

จากสครีพล็อตเลือกที่ 2 องค์ประกอบหลัก เป็นองค์ประกอบหลักของข้อมูลในการลดมิติของข้อมูล แสดงได้ในกราฟ 2 มิติได้ดังรูป 3.14



รูป 3.14 กราฟ 2 มิติของข้อมูลดีเอ็นเอไมโครอาร์เรย์ของซัคคาโรไมซิสเซอริวิสิเอ จากการลดมิติโดยกระบวนการวิเคราะห์องค์ประกอบหลัก และแยกข้อแตกต่างของกลุ่มยีนใน 5 กลุ่มยีน

จากรูปผลจากการวิเคราะห์องค์ประกอบหลักในยีนชุดนี้ สามารถที่จะช่วยให้มองเห็นความแตกต่างของกลุ่มยีน ในลักษณะของภาพ 2 มิติได้ นั่นจึงสรุปได้ว่า กระบวนการวิเคราะห์

องค์ประกอบหลักช่วยในการสังเกตข้อแตกต่างของข้อมูลที่อยู่ในต่างกลุ่มได้ แต่ไม่สามารถแยกข้อข้อแตกต่างของยีนที่อยู่ต่างออนโทโลยีได้

3.2.2 ชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของ มะเร็งชนิด ลิวคีเมีย (Leukemia)

ข้อมูลดีเอ็นเอไมโครอาร์เรย์ของมะเร็งชนิดลิวคีเมียหรือมะเร็งเม็ดเลือดขาว เป็นข้อมูลจากผลงานวิจัยเรื่องการจัดกลุ่มของมะเร็งในระดับโมเลกุล โดยอาศัยค่าการแสดงออกของยีน (Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression) ซึ่งวิจัยโดย ที อาร์ กอลับ (T.R. Golub) และนักวิจัยร่วมท่านอื่นๆ ซึ่งตีพิมพ์ในวารสารไซน์แมกกาซีน (SCIENCE) เมื่อปี ค.ศ. 1999 และข้อมูลดังกล่าวดาวน์โหลดได้จากอินเทอร์เน็ต: http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi

(T.R. Golub, 1999)

- ลักษณะของข้อมูล

มะเร็ง เป็นโรคที่เกิดจากเซลล์ของร่างกายชนิดหนึ่งเพิ่มจำนวนมากขึ้นผิดปกติ และกระจายไปตำแหน่งอื่น ๆ ของร่างกาย มะเร็งลิวคีเมียหรือมะเร็งเม็ดเลือดขาวเป็นโรคมะเร็งที่เซลล์เม็ดเลือดชนิดใดชนิดหนึ่ง เกิดการเพิ่มจำนวนอย่างไม่หยุดยั้ง จึงทำให้เกิดความผิดปกติของปริมาณและการทำงานของเซลล์เม็ดเลือดชนิดอื่น ๆ ในร่างกาย

ในร่างกายคนปกติมีเลือดประมาณ 70 ซีซีต่อน้ำหนักตัว 1 กิโลกรัม ประมาณ 55 ถึง 60 เปอร์เซ็นต์ ของเลือด เป็นส่วนที่เรียกว่า พลาสมาหรือ น้ำเลือด ที่เหลือเป็นส่วนของเม็ดเลือดในส่วนที่เป็นเม็ดเลือดยังแบ่งออกเป็นชนิดใหญ่ๆได้3ชนิดคือ

1) เม็ดเลือดแดงมีหน้าที่นำออกซิเจนไปเลี้ยงส่วนต่างๆของร่างกาย
2) เม็ดเลือดขาวทำหน้าที่คล้ายตำรวจคอยต่อสู้กับเชื้อโรคซึ่งเปรียบได้กับผู้ร้าย
3) เกร็ดเลือดมีหน้าที่เกี่ยวกับการห้ามเลือดในเวลาที่ร่างกายมีเลือดออก

ในภาวะปกติ เม็ดเลือดทั้ง 3 ชนิด สร้างจากเซลล์ในไขกระดูก (Bone Marrow) ซึ่งเป็นเนื้อเยื่อที่อยู่ในแกนกลางของกระดูกทั่วไป เม็ดเลือดแต่ละชนิดจะมีอายุขัยในร่างกายคนแตกต่างกันไป เช่นเม็ดเลือดแดงมีอายุ 120 วัน เม็ดเลือดขาวมีอายุ 2-3 สัปดาห์ เป็นต้น เมื่อเจริญเติบโตเต็มที่แล้วก็จะถูกปล่อยออกมาจากไขกระดูก เข้าเส้นเลือดออกไป ทำหน้าที่ต่างๆกัน แล้วแต่ชนิดของเม็ดเลือดนั้นๆ เมื่อครบอายุของเม็ดเลือดนั้นๆ แล้วก็จะถูกทำลายไป ซึ่งส่วนใหญ่การทำลายนี้จะเกิดขึ้นที่ม้าม แล้วก็จะมีเม็ดเลือดที่เกิดมาใหม่มาทำหน้าที่ทดแทนเป็นเช่นนี้ไปเรื่อยๆ ในคนปกติการสร้างและการทำลายเม็ดเลือดนี้จะอยู่ภายใต้กลไกการควบคุม

ของร่างกายเพื่อให้เกิดความสมดุลกัน เมื่อใดก็ตามที่กลไกการควบคุมดังกล่าวเสียไป ทำให้การสร้างและการทำงานไม่ได้สมดุลกันก็จะเกิดเป็นพยาธิสภาพขึ้น ความผิดปกติเกิดขึ้นกับการแบ่งตัวเหล่านี้ เช่นได้รับสารรังสี สารเคมี หรือไวรัส บางชนิด การแบ่งตัวจะผิดปกติไป หากร่างกายไม่สามารถควบคุมการแบ่งตัวที่ผิดปกติเหล่านี้ไว้ได้ จำนวนเม็ดเลือดที่เกิดผิดปกติก็จะมากขึ้นเรื่อยๆและเกิดภาวะมะเร็งขึ้น

ลิวคีเมียหมายถึงภาวะที่ร่างกายมีการสร้างเม็ดเลือดชนิดใดชนิดหนึ่งมากกว่าปกติหลายเท่าจนเป็นอันตรายต่อเม็ดเลือดข้างเคียง และอวัยวะอื่นๆ ในความหมายเช่นนี้ ลิวคีเมียก็เปรียบได้กับมะเร็งของเม็ดเลือดนั่นเอง ส่วนมะเร็งเม็ดเลือดขาว เป็นมะเร็งชนิดหนึ่งของระบบเลือดที่เกิดจากการที่เซลล์เม็ดเลือดขาวในไขกระดูกเจริญเติบโตผิดปกติทำให้มีการสร้างเม็ดเลือดขาวออกมามากในกระแสเลือดเป็นผลให้ระบบการทำงานของเม็ดเลือดเสียไป และพบว่าเซลล์ที่สร้างเม็ดเลือดแดงก็เป็นมะเร็งได้ แต่เนื่องจากเซลล์ที่ตรวจพบนั้นยังไม่มีการสร้างฮีโมโกลบิน(hemoglobin) จึงรวมเรียกว่าเป็นมะเร็งเม็ดเลือดขาวชนิดหนึ่งเช่นกัน นอกจากนี้พบว่าเซลล์ที่สร้างเกล็ดเลือดก็เป็นมะเร็งได้เช่นกัน ปัจจุบันคำว่ามะเร็งเม็ดเลือดขาว หรือมะเร็งชนิดชนิดลิวคีเมีย(leukemia) จึงใช้รวมเรียก มะเร็งของเซลล์เม็ดเลือดทุกชนิด

ส่วนสาเหตุของการเกิดมะเร็งชนิดนี้ยังไม่ทราบแน่ชัดแต่พบว่าปัจจัยทางพันธุกรรม การติดเชื้อไวรัสบางชนิดและการได้รับสารเคมีบางอย่าง เช่น ยาฆ่าแมลง และสารกัมมันตภาพรังสีสามารถทำให้เกิดมะเร็งเม็ดเลือดได้ ลิวคีเมีย แบ่งออกได้เป็นชนิดใหญ่ๆ 2 ชนิด คือ ชนิดเฉียบพลัน และชนิดเรื้อรัง โดยทั่วไปแล้วมะเร็งเม็ดเลือดขาวชนิดเฉียบพลันจะมีอาการรุนแรงกว่าชนิดที่เกิดช้า ๆ หรือเรื้อรัง

ลิวคีเมียชนิดเฉียบพลัน(Acute leukemia) ลักษณะคือร่างกายมีการสร้างเม็ดเลือดขาวมากกว่าปกติอย่างรวดเร็ว การสร้างเม็ดเลือดแดงและเกล็ดเลือดในไขกระดูก ถูกแย่งที่ไปเกือบหมด ทำให้ เม็ดเลือดชนิดปกติมีจำนวนลดต่ำลงอย่างรวดเร็วตามไปด้วย เป็นผลให้ผู้ป่วยมีอาการของโรคมะเร็งในระยะเวลาอันสั้น

ลิวคีเมียชนิดเรื้อรัง (Chronic leukemia) มีลักษณะที่เซลล์ในไขกระดูกที่เป็นต้นกำเนิดของเซลล์เม็ดเลือดทั้ง 3 ชนิด มีการเปลี่ยนแปลงโครงสร้างไป ทำให้โครงสร้างของเม็ดเลือดทั้ง 3 ชนิดเปลี่ยนแปลงไปด้วย ส่วนใหญ่จะแสดงออกทางเม็ดเลือดขาวมากที่สุดคือมีการสร้างเม็ดเลือดขาวมากกว่าปกตินอกจากนี้เม็ดเลือดขาวที่สร้างขึ้นมานี้ยังมีอายุยืนยาวกว่าปกติมาก จึงพอที่จะสามารถทำหน้าที่ต่อสู้กับเชื้อโรคได้พอสมควร เพราะฉะนั้นในลิวคีเมียชนิดนี้จึงไม่ค่อยมีการติดเชื้อบ่อยอย่างเช่นในลิวคีเมียชนิดเฉียบพลัน นั่นคือเซลล์ มะเร็งที่เกิดขึ้นจะ

ยังพอทำหน้าที่แทนเซลล์ปกติได้บ้างและรบกวนการสร้างเซลล์ปกติไม่มากนักเป็นผลให้ผู้ป่วยจะมีอาการค่อยเป็นค่อยไป

นอกจากนี้ ในแต่ละชนิดของมะเร็งลิวคีเมียดังที่กล่าวมาสามารถแบ่งย่อยตามลักษณะของเซลล์เม็ดเลือดขาวที่เป็นต้นกำเนิดของเซลล์มะเร็ง ได้แก่ ไมลอยด์ลูคีเมีย(Myeloid Leukemia) และลิมโฟบาสติกลูคีเมีย (Lymphoblastic Leukemia) ผลจากการแบ่งย่อยประเภทของมะเร็งลิวคีเมีย ลงไปนี้ทำให้สามารถแบ่งชนิดของโรคมะเร็งลิวคีเมียได้เป็น 4 ชนิด คือลิวคีเมียชนิดเฉียบพลันประเภทไมลอยด์ (Acute Myeloid Leukemia) ลิวคีเมียชนิดเฉียบพลันประเภทลิมโฟบาสติก (Acute Lymphoblastic Leukemia) ลิวคีเมียชนิดเรื้อรังประเภทไมลอยด์ (Chronic Myeloid Leukemia) และ ลิวคีเมียชนิดเรื้อรังประเภทลิมโฟบาสติก (Chronic Lymphoblastic Leukemia)

(แสงสุรีย์ จุฑา, 2523)

จากผลงานวิจัยซึ่งเราได้อาศัยข้อมูลสำหรับวิเคราะห์ สนใจศึกษาที่สองชนิดแรกของมะเร็งนั่นคือ ลิวคีเมียชนิดเฉียบพลันประเภทไมลอยด์ (Acute Myeloid Leukemia) และลิวคีเมียชนิดเฉียบพลันประเภทลิมโฟบาสติก (Acute Lymphoblastic Leukemia) ซึ่งจะแทนด้วยอักษรย่อเป็น เอเอ็มแอล (AML) และ เอแอลแอล (ALL) ตามลำดับ

ในงานวิจัยดังกล่าวนี้ มีจุดประสงค์ที่จะศึกษาในเรื่องของการแบ่งชนิดของเซลล์มะเร็งโดยอาศัยค่าการแสดงออกของยีนต่างๆ ที่อยู่ในเซลล์มะเร็ง ซึ่งดังที่เคยกล่าวมา ค่าการแสดงออกของยีนนั้นได้จากการใช้เทคโนโลยีทางดีเอ็นเอไมโครอาร์เรย์ในการวัดค่า ทั้งนี้ค่าการแสดงออกดังกล่าวเกิดขึ้นมาจากกระบวนการไฮบริดจ์ไดเซชัน (hybridization) ระหว่างนิวคลีโอไทด์ของซีดีเอนเอ (cDNA) หรือนิวคลีโอไทด์สายสั้นๆ (oligonucleotide) ที่อยู่ในเลือดหรือไขกระดูก ในเนื้อเยื่อที่เป็นมะเร็งกับไม่เป็นมะเร็ง นิวคลีโอไทด์เหล่านี้บางส่วนเป็นยีน บางส่วนไม่ใช่ยีนและจะเรียกส่วนนี้ว่า อีเอสที (ESTs : Expressed Sequence Tag) ค่าการแสดงออกเหล่านี้เก็บไว้ในแผ่นไมโครอาร์เรย์ที่ประกอบไปด้วย นิวคลีโอไทด์ ที่เป็นทั้งยีนและส่วนที่ไม่ใช่ยีน ทั้งสิ้น 7,129 พร็อบ(prob) ซึ่งจะแยกออกเป็นยีนทั้งสิ้น 6,187 ยีน และส่วนที่เหลือ คือ อีเอสที ในการทดลองดังกล่าวได้ทดลองกับเนื้อเยื่อที่เป็นมะเร็งลิวคีเมียทั้งสิ้น 72 ตัวอย่างโดยแบ่งเป็นชุดข้อมูลสำหรับเรียนรู้(training data) จำนวน 38 ตัวอย่าง และข้อมูลสำหรับทดสอบ(testing data) 34 ตัวอย่าง ทั้งนี้ในจำนวนข้อมูลตัวอย่างที่ใช้ศึกษานั้นประกอบไปด้วย กลุ่มตัวอย่างของเนื้อเยื่อที่เป็นมะเร็งลิวคีเมียประเภทเอแอลแอล จำนวน 27

ตัวอย่าง และประเภทเอเอ็มแอล 11 ตัวอย่าง สำหรับข้อมูลที่ใช้ทดสอบนั้น ประกอบไปด้วยมะเร็งลิวคีเมียประเภทเอแอลแอล จำนวน 20 ตัวอย่าง และประเภทเอเอ็มแอล 14 ตัวอย่าง (T.R. Golub, 1999)

แสดงตัวอย่างของข้อมูลได้ดังตาราง 3.13

ตาราง 3.13 ชุดข้อมูลการแสดงออกของยีนจากดีเอ็นเอไมโครอาร์เรย์ของมะเร็งชนิดลิวคีเมีย จำนวน 72 ตัวอย่าง

Gene Description	Gene Accession Number	Sample 1	Sample 2	...	Sample 71	Sample 72
AFFX-BioB-5_at (endogenous control)	AFFX-BioB-5_at	-214	-139		-48	-176
AFFX-BioB-M_at (endogenous control)	AFFX-BioB-M_at	-153	-73		-531	-284
AFFX-BioB-3_at (endogenous control)	AFFX-BioB-3_at	-58	-1		-124	-81
AFFX-BioC-5_at (endogenous control)	AFFX-BioC-5_at	88	283		431	9
AFFX-BioC-3_at (endogenous control)	AFFX-BioC-3_at	-295	-264		-496	-294
AFFX-BioDn-5_at (endogenous control)	AFFX-BioDn-5_at	-558	-400		-696	-493
AFFX-BioDn-3_at (endogenous control)	AFFX-BioDn-3_at	199	-330		-1038	-393
AFFX-CreX-5_at (endogenous control)	AFFX-CreX-5_at	-176	-168	⋮	-441	-141
AFFX-CreX-3_at (endogenous control)	AFFX-CreX-3_at	252	101		235	166
AFFX-BioB-5_st (endogenous control)	AFFX-BioB-5_st	206	74		157	-37
AFFX-BioB-M_st (endogenous control)	AFFX-BioB-M_st	-41	19		-162	-39
AFFX-BioB-3_st (endogenous control)	AFFX-BioB-3_st	-831	-743		-1453	-859
AFFX-BioC-5_st (endogenous control)	AFFX-BioC-5_st	-653	-239		-1013	-572
AFFX-BioC-3_st (endogenous control)	AFFX-BioC-3_st	-462	-83		-317	-150
AFFX-BioDn-5_st (endogenous control)	AFFX-BioDn-5_st	75	182		173	414
Type of Leukemia Cancer		ALL	ALL	...	AML	AML

จาก ตาราง 3.13 แสดงตัวอย่างของข้อมูลการแสดงออกของยีน โดยแต่ละแถวแสดงถึง รหัสของนิวคลีโอไทด์สายสั้นๆ ซึ่งบางตัวเป็นยีน บางตัวเป็น อีเอสที (ESTs) แต่เพื่อให้ง่ายต่อการวิเคราะห์ต่อไปจะเรียกทั้งสองส่วนนี้เป็นยีนเช่นเดียวกันหมด และจากตารางในแต่

ละคอลัมน์เป็นตัวอย่างข้อมูล ซึ่งมีจำนวน 72 ตัวอย่างแบ่งเป็นกลุ่ม เอแอลแอลจำนวน 47 ตัวอย่างเป็น กลุ่มเอเอ็มแอล 25 ตัวอย่าง

- วิธีการวิเคราะห์

1) วิเคราะห์ข้อมูลโดยใช้ฟังก์ชั่นการวิเคราะห์และพัฒนาขึ้นเอง จากโปรแกรมภาษา R เวอร์ชัน 2.3.1ในการคำนวณ

2) เตรียมข้อมูล ในขั้นตอนของการเตรียมข้อมูลนี้ จะใช้วิธีการเดียวกับผลงานวิจัยที่เป็นแหล่งของข้อมูลชุดนี้ ซึ่งมีวิธีการได้แก่การปรับข้อมูลที่มีค่ามากหรือน้อยเกินไปให้มีค่าอยู่ในระดับที่เหมาะสม การกรองข้อมูลที่ไม่จำเป็นออกไป และการปรับข้อมูลให้อยู่ในลักษณะของค่าล๊อกการิทึม ทั้งนี้ชุดข้อมูลที่จะใช้ในการวิเคราะห์จะใช้ชุดข้อมูลเรียนรู้ (training) เพียง 38 ตัวอย่าง เนื่องจากจะเป็นส่วนของข้อมูลที่จะนำไปใช้ในการเรียนรู้เพื่อที่จะจำแนกประเภทของข้อมูลในบทต่อไป สำหรับขั้นตอนของการปรับข้อมูล จะแทนที่ยีนที่มีค่าการแสดงออกของข้อมูล น้อยกว่า 100 ด้วยค่า 100 และยีนที่มีค่ามากกว่า 16,000 ด้วยค่า 16,000 ในขั้นต่อไปจะเป็นขั้นตอนของการกรองข้อมูล ในขั้นนี้จะกรองเอายีนที่ ค่าการแสดงออกของยีนดังกล่าวไม่มีนัยสำคัญออกไป ด้วยเงื่อนไขดังสมการ

$$\text{Max}(X_i) / \text{Min}(X_i) > 5 \text{ และ } \text{Max}(X_i) - \text{Min}(X_i) > 500 \quad \text{โดยที่ } i = 1,\ldots,n \text{ ยีน} \tag{30}$$

ในสมการ (30) จะกรองเอาข้อมูลที่ อัตราส่วนระหว่างค่าการแสดงออกของยีนสูงสุดกับค่าการแสดงออกต่ำสุดในยีนแต่ละตัวมีค่ามากกว่า 5 และผลต่างระหว่างค่าการแสดงออกของยีนที่มีค่าสูงสุดกับค่าต่ำสุด มากกว่า 500 เก็บไว้ ส่วนข้อมูลที่ไม่ตรงตามเงื่อนไขตัดทิ้ง ผลก็คือจะทำให้ค่าการแสดงออกของยีนในทุกๆ ตัวอย่าง (Sample) มีค่าที่แตกต่างกันในระดับนัยสำคัญที่เหมาะสม ผลสุดท้ายที่ได้จากการกรองข้อมูลด้วยกระบวนการดังกล่าว จะได้ยีนที่นำมาใช้ในการวิเคราะห์จริงๆ เพียง 3,051 ยีนเท่านั้น ขั้นตอนสุดท้ายของการเตรียมข้อมูล จะทำการหาค่าล๊อกการิทึมของข้อมูล ผลก็คือจะให้ข้อมูลตัวเลขที่มีค่ามากๆ อยู่ในลักษณะของค่าล๊อกการิทึมที่มีค่าตัวเลขน้อยลง

(T.R. Golub, 1999)

แสดงข้อมูลตัวอย่างได้ดัง ตาราง 3.14

ตาราง 3.14 ชุดข้อมูลการแสดงออกของยีนจากดีเอ็นเอไมโครอาร์เรย์ของมะเร็งชนิดลิวคีเมียโดยผ่านกระบวนการเตรียมข้อมูลแล้ว จำนวน 38 ตัวอย่าง

Gene ID	Sample 1	Sample 2	Sample 3	...	Sample 37	Sample 38
AFFX-HUMISGF3A/M97935_MA_at	2.000	2.000	2.017		3.110	2.316
AFFX-HUMISGF3A/M97935_MB_at	2.332	2.064	2.678		2.908	2.415
AFFX-HUMISGF3A/M97935_3_at	2.901	2.636	3.168		3.139	2.857
AFFX-HUMRGE/M10098_5_at	4.163	2.789	3.754		3.339	2.913
AFFX-HUMRGE/M10098_M_at	3.988	2.061	3.515	⋮	3.194	2.710
AFFX-HUMRGE/M10098_3_at	3.931	3.181	3.564		2.940	3.287
AFFX-HUMGAPDH/M33197_5_at	4.178	4.204	4.204		4.204	4.204
AFFX-HUMGAPDH/M33197_M_at	4.046	4.133	4.204		4.177	4.177
AFFX-HSAC07/X00351_5_at	4.204	4.204	4.204		4.204	4.166
Type of Leukemia Cancer	ALL	ALL	ALL	...	AML	AML

3) สร้างเมตริกซ์สหสัมพันธ์ของข้อมูล โดยใช้ยีนเป็นตัวแปร นั่นคือถ้าให้ X เป็นเมตริกซ์ของข้อมูลดังในตาราง 3.14 จะสร้างเมตริกซ์สหสัมพันธ์ตามสมการ (7) โดยที่ ให้ชุดข้อมูลที่ใช้แทนในสมการดังกล่าวด้วย A ซึ่ง A หาได้จากการทรานสโพส (transpose) เมตริกซ์ X

4) หาองค์ประกอบหลัก ความแปรปรวนขององค์ประกอบหลัก และค่าคะแนนขององค์ประกอบหลัก จากกระบวนการวิเคราะห์องค์ประกอบหลัก เพื่อแสดงให้เห็นความแตกต่างของข้อมูลระหว่างกลุ่มตัวอย่างที่มีชนิดของมะเร็งต่างๆกัน

- ผลการวิเคราะห์

1) แสดงค่าความแปรปรวนและความแปรปรวนสะสม ใน 30 องค์ประกอบหลักแรก ได้ดังตารางต่อไปนี้

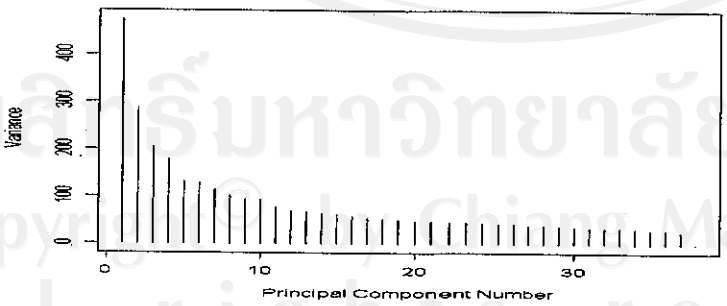
ตาราง 3.15 ค่าความแปรปรวนและความแปรปรวนสะสมขององค์ประกอบหลักจากการวิเคราะห์องค์ประกอบหลักโดยใช้ข้อมูลการแสดงออกของยีนจากดีเอ็นเอไมโครอาร์เรย์ของมะเร็งชนิด ลิวคีเมีย

Prin Comp No.	Variance	Cum. of Var	% of Cum. of Var	Prin Comp No.	Variance	Cum. of Var	% of Cum. of Var
1	475.06	475.06	15.57	16	57.93	2207.77	72.36
2	286.72	761.78	24.97	17	54.99	2262.76	74.16
3	204.00	965.78	31.65	18	52.90	2315.66	75.90
4	179.09	1144.87	37.52	19	50.47	2366.14	77.55
5	132.01	1276.89	41.85	20	48.57	2414.70	79.14

ตาราง 3.15(ต่อ) ค่าความแปรปรวนและความแปรปรวนสะสมขององค์ประกอบหลักจากการวิเคราะห์องค์ประกอบหลัก โดยใช้ข้อมูลการแสดงออกของยีนจากดีเอ็นเอไมโครอาร์เรย์ของมะเร็งชนิด ลิวคีเมีย

Prin Comp No.	Variance	Cum. of Var	% of Cum. of Var	Prin Comp No.	Variance	Cum. of Var	% of Cum. of Var
6	129.16	1406.04	46.08	21	47.73	2462.44	80.71
7	114.06	1520.10	49.82	22	45.98	2508.41	82.22
8	101.62	1621.72	53.15	23	44.35	2552.76	83.67
9	95.15	1716.88	56.27	24	42.99	2595.75	85.08
10	92.96	1809.83	59.32	25	41.79	2637.54	86.45
11	75.74	1885.57	61.80	26	41.18	2678.72	87.80
12	69.27	1954.84	64.07	27	39.30	2718.02	89.09
13	67.99	2022.83	66.30	28	38.37	2756.39	90.34
14	64.15	2086.98	68.40	29	37.48	2793.87	91.57
15	62.86	2149.85	70.46	30	36.79	2830.66	92.78

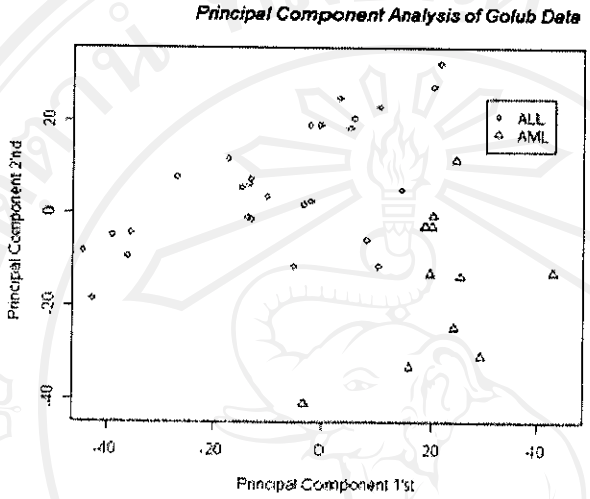
จากตารางจะเห็นว่า 10 องค์ประกอบหลักแรก มีค่าความแปรปรวนสะสมประมาณร้อยละ 50 ซึ่งยังนับว่าน้อย และเป็นตัวแทนของข้อมูลไม่ดีนัก และจากผลการวิเคราะห์พบว่าที่องค์ประกอบหลักที่ 28 จะได้ความแปรปรวนสะสมร่วมกับองค์ประกอบหลักก่อนหน้า เป็น 90 เปอร์เซ็นต์ ดังนั้น เมื่อเราต้องการข้อมูลชุดใหม่ ซึ่งมีจำนวนตัวแปรน้อยกว่าข้อมูลเดิม และมีค่าความแปรปรวนทั้งหมดเป็น 90 เปอร์เซ็นต์ ของข้อมูลเดิม ข้อมูลชุดใหม่นั้นจะมีตัวแปรใหม่เป็นองค์ประกอบหลักที่ 28 องค์ประกอบหลักแรก ซึ่งองค์ประกอบหลักเหล่านี้ สามารถนำไปใช้เป็นตัวแปรสำหรับการวิเคราะห์ข้อมูลเพื่อจำแนกกลุ่มของข้อมูลได้ และเมื่อนำไปสร้าง สครีพล็อต จะแสดงแนวโน้มของค่าความแปรปรวนของข้อมูลใน 38 องค์ประกอบหลักแรกได้ดังรูป 3.15



รูป 3.15 สครีพล็อตจากการวิเคราะห์องค์ประกอบข้อมูลดีเอ็นเอไมโครอาร์เรย์ของชุดข้อมูลมะเร็งลิวคีเมีย

จากรูป แสดงให้เห็นว่าแนวโน้มของความแปรปรวนของข้อมูลที่ได้จากองค์ประกอบหลักแรกๆ จะมีค่าสูง ดังนั้นองค์ประกอบหลักในลำดับต้นๆ จึงเป็นตัวแทนของข้อมูลได้ดี และนำไปใช้ในการหาคะแนนองค์ประกอบหลักต่อไปได้

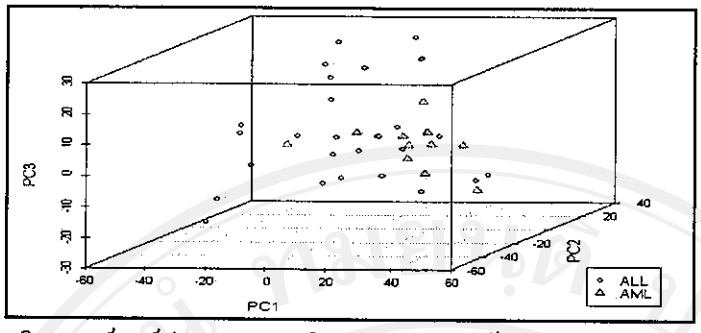
2) เมตริกซ์ของคะแนนองค์ประกอบที่ได้จากการโปรเจคชั่น ลงใน 2 องค์ประกอบหลักแรก พล๊อตลงในกราฟ 2 มิติดังรูป



รูป 3.16 ผลการวิเคราะห์องค์ประกอบหลักใน 2 มิติโดยชุดข้อมูลลิวคีเมียไมโครอาร์เรย์

จากรูปแสดงผลของการนำค่าคะแนนองค์ประกอบหลักจาก 2 องค์ประกอบหลักแรกมาวาดลงในกราฟ 2 มิติ จะช่วยให้เห็นแนวโน้ม หรือข้อแตกต่างของข้อมูลได้แม้ว่าที่ 2 องค์ประกอบหลักแรกจะมีค่าความแปรปรวนไม่มากนักก็ตามแต่ก็มากพอที่จะช่วยแยกให้มองเห็นข้อแตกต่างของค่าการแสดงออกในกลุ่มตัวอย่าง ทั้งสองกลุ่มตัวอย่างได้ ซึ่งจากรูปจะเห็นว่ากลุ่มตัวอย่างที่เป็นมะเร็งลิวคีเมียชนิดเอแอลแอล (ALL) จะเกาะกลุ่มกันอยู่ในพื้นที่ส่วนบนซ้ายของกราฟ และ มะเร็งลิวคีเมียชนิดเอเอ็มแอล (AML) จะเกาะกลุ่มกันอยู่ในพื้นที่ส่วนร่างขวาของกราฟ

3) เมตริกซ์ของคะแนนองค์ประกอบที่ได้จากการโปรเจคชั่น ลงใน 3 องค์ประกอบหลักแรก พล๊อตลงในกราฟ 3 มิติดังรูป



รูป 3.17 ผลการวิเคราะห์องค์ประกอบหลักใน 3 มิติโดยชุดข้อมูลลิวคีเมียไมโครอาร์เรย์

จากรูป 3.17 ลักษณะการกระจายตัวของข้อมูลที่ 3 มิติ พบว่าตัวอย่างข้อมูล มีการแยกกลุ่มกันอย่างชัดเจน ตามกลุ่มที่กำหนด

3.3 วิจารณ์และสรุปผล

แนวทางการประยุกต์วิธีการวิเคราะห์องค์ประกอบหลักกับข้อมูลดีเอ็นเอไมโครอาร์เรย์ในงานวิจัยนี้ สรุปได้ดังต่อไปนี้

1) ใช้การวิเคราะห์องค์ประกอบหลักในการวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิสเซอริวิสิเอ สำหรับลดจำนวนตัวแปรของข้อมูล เพื่อสังเกตลักษณะการกระจายตัวของยีน ในข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิสเซอริวิสิเอ ใน 2 มิติ โดยใช้ยีนทั้งหมดของชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ที่เกี่ยวข้องกับการวิเคราะห์

2)ใช้การวิเคราะห์องค์ประกอบหลักในการวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิสเซอริวิสิเอ สำหรับลดจำนวนตัวแปรของข้อมูล เพื่อสังเกตลักษณะการกระจายตัวของยีนที่อยู่ต่างยีนออนโทโลยี โดยใช้ยีนที่มีการกำหนดยีนออนโทโลยี ทั้งนี้แยกวิเคราะห์ข้อมูลออกเป็น 3 ยีนออนโทโลยีหลักที่อิสระต่อกัน จากนั้นกำหนดยีนออนโทโลยีย่อยให้กับยีนแต่ละตัว แล้วสังเกตลักษณะการกระจายตัวของยีนที่อยู่ต่างยีนออนโทโลยีย่อย ในภาพ 2 มิติ

3)ใช้การวิเคราะห์องค์ประกอบหลักในการวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิสเซอริวิสิเอ สำหรับลดจำนวนตัวแปรของข้อมูล เพื่อสังเกตลักษณะการกระจายตัวของยีนในกลุ่มที่ต่างกัน โดยใช้ยีนที่มีการกำหนดกลุ่มแล้วจากผลงานวิจัย ซึ่งจะทำให้สังเกตลักษณะการกระจายตัวของยีนที่อยู่ต่างกลุ่มกันในภาพ 2 มิติ

4)ใช้การวิเคราะห์องค์ประกอบหลักในการวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์มะเร็งลิวคีเมีย สำหรับลดจำนวนยีน เพื่อสังเกตลักษณะการกระจายตัวของกลุ่มตัวอย่างข้อมูล ในกลุ่มที่แตกต่างกันใน 2 หรือ 3 มิติ

จากกรณีศึกษา การวิเคราะห์องค์ประกอบหลักในการวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ซัคคาโรไมซิสเซอริวิสิเอ ซึ่งแบ่งเป็น 3 การทดลอง สรุปได้ดังนี้

การทดลองที่ 1 พบว่า การวิเคราะห์องค์ประกอบหลักทำให้สามารถลดขนาดของตัวแปรซึ่งเป็นช่วงเวลาของการไดอ๊อคซิคชิฟท์ ทั้ง 7 ช่วงเวลา ในลักษณะองค์ประกอบหลักจำนวน 2 ประกอบหลัก ที่ความแปรปรวน 70.9 เปอร์เซ็นต์ ซึ่งผลดังกล่าวจะทำให้สามารถสังเกตเห็นแนวโน้มการกระจายตัวของยีนต่างๆ ในลักษณะสองมิติ

การทดลองที่ 2 พบว่า ผลจากการวิเคราะห์องค์ประกอบหลักไม่สามารถที่จะใช้สังเกตข้อแตกต่างของยีน ที่อยู่ต่างยีนออนโทโลยีได้ ซึ่งผู้วิจัยได้ตั้งสมมุติฐานถึงสาเหตุ จากหลายๆ ด้าน เช่น ยีนออนโทโลยีไม่สามารถที่จะอธิบายได้โดยอาศัยค่าการแสดงออกของยีน ยีนออนโทโลยีไม่ได้เป็นตัวกำหนดกลุ่มข้อมูลที่ดี ยีนออนโทโลยีไม่สามารถใช้การวิเคราะห์องค์ประกอบหลักในการสังเกตความแตกต่างของข้อมูล ยีนออนโทโลยีมีจำนวนมากเกินไปรวมทั้งยีนแต่ละตัวมีการคาบเกี่ยวกันในหลายๆออนโทโลยี ทำให้การสังเกตข้อแตกต่างของข้อมูลในสองหรือสามมิติ ไม่สามารถทำได้ และผลจากการวิเคราะห์องค์ประกอบหลักไม่สามารถที่จะสังเกตข้อแตกต่างของกลุ่มข้อมูลได้ เป็นต้น ทั้งนี้การสรุปว่าเป็นเพราะสาเหตุใดนั้น สมควรที่จะต้องมีการศึกษาต่อไป

การทดลองที่ 3 เมื่อกำหนดกลุ่มยีน ที่มีรูปแบบของค่าการแสดงออกทั้ง 7 ตัวแปรแตกต่างกันอย่างชัดเจน เป็นข้อมูลที่นำมาใช้ในการวิเคราะห์องค์ประกอบหลัก พบว่าลักษณะการกระจายตัวของข้อมูลใน 2 องค์ประกอบหลักที่ความแปรปรวน 80.09 เปอร์เซ็นต์ มีการแบ่งแยกกลุ่มของข้อมูล ได้อย่างชัดเจน นั่นแสดงว่ากระบวนการวิเคราะห์องค์ประกอบหลักช่วยในการสังเกตข้อแตกต่างของข้อมูลที่อยู่ในต่างกลุ่มได้

จากกรณีศึกษา การวิเคราะห์องค์ประกอบหลักในการวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์ของมะเร็ง ลิวคีเมีย ในผู้ป่วยจำนวน 38 คน ซึ่งแบ่งผู้ป่วยออกเป็นสองกลุ่มตามประเภทของมะเร็ง โดยกำหนดให้ยีนที่ผ่านกระบวนการกรองข้อมูลจำนวน 3,051 ยีนเป็นตัวแปร เมื่อนำเสนอข้อมูลใน 2 มิติจาก 2 องค์ประกอบหลักแรก ทำให้แยกข้อแตกต่างของผู้ป่วยที่เป็นมะเร็งลิวคีเมียทั้งสองประเภทนั้นได้อย่างชัดเจน ซึ่งเมื่อพิจารณาที่ 3 องค์ประกอบหลัก ก็ให้ผลไม่แตกต่างกัน

นอกจากนี้เมื่อพิจารณาที่จำนวนองค์ประกอบหลักที่มีค่าความแปรปรวนมากๆ โดยจำนวนองค์ประกอบหลักดังกล่าวมีจำนวนน้อยกว่ายีนที่กำหนดให้เป็นตัวแปร จะเห็นว่าองค์ประกอบหลักดังกล่าว น่าจะนำไปใช้เป็นข้อมูลตั้งต้นสำหรับการวิเคราะห์อื่นๆ ที่มีความต้องการลดจำนวนตัวแปร เช่น การวิเคราะห์การจำแนกประเภท การวิเคราะห์การถดถอยได้