บทที่ 1
บทนำ

1.1 หลักการ ทฤษฎี เหตุผล และ สมมุติฐาน

เทคโนโลยีดีเอ็นเอไมโครอาร์เรย์ (DNA Mocroarrays) ในปัจจุบัน ทำให้เราสามารถตรวจวัดค่าระดับการแสดงออก (Transcriptional Expression Level) ของยีนในสิ่งมีชีวิตภายใต้สภาวะต่างๆ ได้อย่างรวดเร็ว ทำให้นักวิทยาศาสตร์สามารถสร้างความเข้าใจถึงกระบวนการต่างๆ ที่เกิดขึ้นภายในเซลล์ทั้งระบบ ซึ่งจะนำไปใช้สร้างองค์ความรู้ใหม่ในการพัฒนาเทคโนโลยีทางด้านชีวภาพให้ก้าวหน้ายิ่งขึ้น

เป้าหมายสูงสุดของการสร้างข้อมูลดีเอ็นเอไมโครอาร์เรย์ คือการวิเคราะห์หาความหมายของข้อมูลที่ได้จากการทดลองเพื่อหาข้อสรุปเชิงชีววิทยา ทั้งนี้การวิเคราะห์โดยทั่วไปนั้นอาศัยหลักการวิเคราะห์ข้อมูลทั้งทางด้านคณิตศาสตร์และสถิติ โดยใช้เทคโนโลยีทางคอมพิวเตอร์ในการประมวลผล ซึ่งผลการวิเคราะห์ที่ได้จะนำมาเปรียบเทียบกับข้อมูลทางชีววิทยาเพื่อหาข้อสรุป

ข้อมูลดีเอ็นเอไมโครอาร์เรย์เป็นข้อมูลสังเกตการณ์ที่แสดงให้เห็นปริมาณการแสดงออกของยีนในระดับ ทรานสคริปชัน (Transcription) ซึ่งค่าการแสดงออกดังกล่าวได้จากการวัดโดยอาศัยการทดลองทางวิทยาศาสตร์แล้วเก็บไว้ในรูปของไมโครอาร์เรย์ การนำข้อมูลนี้มาใช้ต้องผ่านกระบวนการเตรียมข้อมูล(Data Preprocessing) เพื่อที่จะสามารถนำมาใช้ได้อย่างมีประสิทธิผล เช่น การแปลงข้อมูลจากข้อมูลไมโครอาร์เรย์ที่เป็นภาพให้อยู่ในลักษณะของตัวเลข กระบวนการนอร์มอลไลซ์เซชัน (Normalization) เป็นการลดผลกระทบต่อการวิเคราะห์อันเนื่องมาจากหน่วยวัดหรือบริบทของการทดลอง ซึ่งจะช่วยให้ปรับบรรทัดฐาน(Scaling)ของค่าตัวเลขให้มีขอบเขตที่เป็นมาตรฐาน นอกจากนี้กระบวนการรวบรวมข้อมูลที่เกี่ยวข้องไว้ด้วยกันก็ถือเป็นขั้นตอนหนึ่งของการจัดเตรียมข้อมูล ซึ่งจะช่วยให้เราสามารถนำข้อมูลมาใช้ได้อย่างครบถ้วน เช่น การนำข้อมูลที่ระบุกลุ่มของยีนมาใช้ควบคู่กับข้อมูลการแสดงออกของยีนเป็นต้น ท้ายที่สุดแล้วหลังจากที่ข้อมูลผ่านกระบวนการของการจัดเตรียมข้อมูลก็จะนำไปสู่ขั้นตอนของการวิเคราะห์ต่อไปได้ ในส่วนของการวิเคราะห์นี้นักวิทยาศาสตร์มีการศึกษา และวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์ด้วยวิธีการที่หลากหลายตามจุดประสงค์ต่างๆ เช่น เพื่อแบ่งกลุ่มยีน เพื่อจำแนกกลุ่มของยีน เพื่อหาหน้าที่ความสัมพันธ์ระหว่างยีน เป็นต้น

ในงานวิจัยนี้ ได้อาศัยเทคนิคการวิเคราะห์ข้อมูลทางสถิติที่มีชื่อว่า กระบวนการวิเคราะห์หลายตัวแปร (Multivariate Analysis) มาใช้ในการวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์ โดยเทคนิคดังกล่าวนี้เป็นวิธีการที่เหมาะกับการนำไปวิเคราะห์ข้อมูลที่มีหลายตัวแปร มีเป้าหมายเพื่อการอนุมานความสัมพันธ์ระหว่างตัวแปร ดังนั้นกระบวนการดังกล่าวจึงเหมาะกับการนำมาใช้วิเคราะห์ข้อมูลดีเอ็น

เอไมโครอาร์เรย์ ซึ่งเป็นข้อมูลที่มีตัวแปรมากกว่า 1 ตัวแปร และในที่นี้ตัวแปรที่จะใช้ในการวิเคราะห์คือยีนและ เงื่อนไขการทดลองต่างๆ(Experiment Condition) สำหรับทฤษฎีที่ใช้ในการวิเคราะห์หลายตัวแปรมีหลายทฤษฎี โดยมีจุดประสงค์ของการวิเคราะห์ที่ต่างๆ กัน ดังนี้

(1) การวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis) เป็นวิธีการที่ใช้สำหรับการลดจำนวนตัวแปรที่มีหลายตัวแปรให้มีจำนวนน้อยลง (Data Reduction) ประโยชน์คือทำให้นำเสนอ (Visualization) และตีความหมาย (Interpretation) ของข้อมูลได้ง่ายขึ้น โดยอาศัยหลักการที่ว่าตัวแปรใหม่จากการลดมิติของข้อมูลเป็นตัวแปรที่ให้ความหมายหรือความแปรปรวน (Variance) ของข้อมูลเดิมได้มากที่สุด ในการวิเคราะห์กับข้อมูลดีเอ็นเอไมโครอาร์เรย์จะนำไปใช้ในการลดขนาดของตัวแปรซึ่งในที่นี้อาจจะเป็นยีนหรือเงื่อนไขการทดลองต่างๆ

(2) การวิเคราะห์ปัจจัย (Factor Analysis) เป็นวิธีการที่มีจุดประสงค์เพื่อหาโครงสร้างความสัมพันธ์ระหว่างตัวแปร(Structure Detection) นั่นคือ การหาว่าภายใต้ตัวแปรที่สังเกตการณ์ (Observation Variables) มีตัวแปรที่ซ่อนอยู่ภายใน(Latency Variable) หรือปัจจัยแฝง(Abstract Factor) อะไรบ้าง ที่เป็นตัวให้ความหมายหรือสื่อถึงตัวแปรที่เราสังเกตเห็นได้ ประโยชน์ของการหาโครงสร้างภายในของตัวแปรนี้ก็คือ จะช่วยให้เราสามารถที่จะจัดกลุ่มของตัวแปร ที่มีความสัมพันธ์กันทั้งในด้านหน้าที่ หรือ คุณสมบัติ ที่เป็นไปทางเดียวกันไว้ด้วยกัน นอกจากนี้ ผลของการวิเคราะห์ จะช่วยให้เราระบุตัวแปรที่ไม่มีความสัมพันธ์กับตัวแปรหรือข้อมูลอื่นได้ ซึ่งทำให้เราสามารถเลือกตัวแปร หรือเลือกคุณสมบัติ (Feature Selection) เพียงส่วนที่มีผลต่อข้อมูลที่เราจะวิเคราะห์เท่านั้น สำหรับในข้อมูลดีเอ็นเอไมโครอาร์เรย์ ของสิ่งมีชีวิตหนึ่งๆ ของเซลล์หนึ่งๆ หรือในชุดการทดลองหนึ่งๆ นั้น เมื่อให้ยีนเป็นตัวแปรกลุ่มของยีนที่เกี่ยวข้อง ย่อมมีหน้าที่หรือคุณสมบัติต่างๆ ที่สัมพันธ์กันอยู่ภายใน ไม่สามารถที่จะสังเกตเห็น นอกจากนี้เมื่อพิจารณา ให้เงื่อนไขการทดลองต่างๆเป็นตัวแปร ตัวแปรดังกล่าวย่อมมีทั้งส่วนที่สัมพันธ์กันและไม่สัมพันธ์กัน ซึ่งซ่อนอยู่ภายในและไม่สามารถสังเกตเห็นเช่นเดียวกัน ดังนั้นเมื่อนำวิธีการนี้ไปวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์ ตัวอย่างหนึ่งของผลการวิเคราะห์นี้ก็คือ สามารถที่จะจัดกลุ่มยีนที่มีหน้าที่ความสัมพันธ์กันไว้ด้วยกัน และให้ความหมายของกลุ่มยีนดังกล่าวนี้ได้

(3) การวิเคราะห์การจำแนกประเภท (Discriminant Analysis) เป็นวิธีการวิเคราะห์ข้อมูลที่อาศัย ความสัมพันธ์ของข้อมูลในกลุ่มตัวอย่าง (Training Data)ที่มีการแบ่งกลุ่มข้อมูลเรียบร้อยแล้วนำมาสร้างเป็นฟังก์ชัน (Function) หรือโมเดล (Model) เพื่อใช้ในการจำแนกกลุ่มของข้อมูล (Data Classification) โดยในหลักการจะอาศัย ความแตกต่างระหว่างค่าเฉลี่ยของตัวแปร ทั้งภายในกลุ่มและระหว่างกลุ่มของข้อมูล เพื่อนำมาใช้ประมาณค่าพารามิเตอร์ สำหรับสร้างเป็นโมเดลการจำแนกประเภท

ข้อมูลที่ดีที่สุด ซึ่งหลังจากได้โมเดลดังกล่าวแล้วนั้น ในขั้นตอนการวิเคราะห์ต่อไปจะอาศัยโมเดลดังกล่าวในการจำแนกประเภทของข้อมูลที่เป็นตัวทดสอบ (Testing Data) หรือชุดของข้อมูลที่ต้องการทำนายกลุ่มได้ นอกจากนี้ในขั้นตอนของการวิเคราะห์ มีขั้นตอนหนึ่งที่จะต้องทำ นั่นก็คือ ขั้นตอนของการเลือกตัวแปรสำหรับการจำแนกประเภท(Feature Selection) เพราะว่าตัวแปรเดิมนั้นมีจำนวนมาก แต่ตัวแปรบางตัวอาจจะไม่เหมาะสมกับการจำแนกประเภท ในกระบวนการวิเคราะห์จะมีวิธีการเลือกตัวแปรที่เหมาะสมสำหรับการนำไปจำแนกประเภทของข้อมูล ตัวอย่างการวิเคราะห์ เช่น ใช้ในการจำแนกประเภทของผู้ป่วยที่เป็นโรคต่างๆ โดยอาศัยข้อมูลดีเอ็นเอไมโครอาร์เรย์ หรือ การนำไปใช้ในการจำแนกกลุ่มของยีน ตามหน้าที่ความสัมพันธ์ต่างๆ

(4) การวิเคราะห์การถดถอยแบบโลจิสติก(Logistic Regression Analysis) เป็นรูปแบบหนึ่งของการวิเคราะห์การถดถอย ที่ใช้ตัวแปรทำนายค่าในการสร้างโมเดล เพื่อที่จะทำนายค่าของข้อมูล ซึ่งค่าของข้อมูลที่ได้จะเป็นค่าความน่าจะเป็นที่ชุดข้อมูลที่เป็นตัวทดสอบนั้นจะอยู่ในกลุ่มของข้อมูลที่ต้องการ ทั้งนี้ ประโยชน์ของวิธีการดังกล่าวกับการวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์ก็คือ สามารถที่จะทำนายได้ว่า ชุดข้อมูลที่เลือกมานั้น น่าจะอยู่ในกลุ่มที่กำหนดให้ ด้วยความน่าจะเป็นเท่าไหร่ ซึ่งสามารถนำไปประยุกต์ใช้กับการจำแนกประเภทของข้อมูลได้

1.2 วัตถุประสงค์ของการวิจัย

1.2.1 เพื่อนำเสนอแนวทางการประยุกต์ เทคนิคการวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์โดยใช้เทคนิคการวิเคราะห์หลายตัวแปร

1.2.2 เพื่อวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์ โดยอาศัยเทคนิคที่นำเสนอ ผลที่ได้จะนำไปใช้ในการหาข้อสรุปเชิงชีววิทยาต่อไป

1.3 ประโยชน์ที่จะได้รับจากการศึกษาเชิงทฤษฎี และ เชิงประยุกต์

1.3.1 ในเชิงทฤษฎี จะทำให้เข้าใจถึงเทคนิคการวิเคราะห์หลายตัวแปร ทั้งแนวคิด และขอบเขตของเทคนิคการวิเคราะห์ รวมทั้งแนวทางการประยุกต์เทคนิคการวิเคราะห์หลายตัวแปรกับการวิเคราะห์กับข้อมูลดีเอ็นเอไมโครอาร์เรย์

1.3.2 ในเชิงประยุกต์ จะได้นำเทคนิคการวิเคราะห์หลายตัวแปรไปวิเคราะห์กับข้อมูลดีเอ็นเอไมโครอาร์เรย์ซึ่งผลที่ได้รับจะเป็นประโยชน์ต่อการศึกษาค้นคว้าของนักวิจัยทางด้านเทคโนโลยีชีวภาพ

1.4 ขอบเขตการวิจัย

ศึกษาและประยุกต์ เทคนิคการวิเคราะห์หลายตัวแปรต่างๆสำหรับ การวิเคราะห์ข้อมูลดีเอ็นเอไมโครอาร์เรย์ ได้แก่ การวิเคราะห์องค์ประกอบหลัก การวิเคราะห์ปัจจัย การวิเคราะห์การจำแนกประเภท และการวิเคราะห์การถดถอยแบบโลจิสติก โดยกำหนดเป้าหมายของการศึกษาและวิเคราะห์ข้อมูลเป็น 2 เป้าหมาย ได้แก่

1.4.1 การหาโมเดลความสัมพันธ์หรือโครงสร้างความสัมพันธ์ระหว่างตัวแปรซึ่งเป็น ยีน และเงื่อนไขการทดลอง เพื่อลดจำนวนตัวแปรและเลือกตัวแปร สำหรับใช้ในการนำเสนอข้อมูล การจำแนกกลุ่มข้อมูล และการจัดกลุ่มข้อมูล โดยอาศัยทฤษฎีการวิเคราะห์องค์ประกอบหลักและการวิเคราะห์ปัจจัย ทั้งนี้จะใช้ชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของยีสต์ที่มีชื่อทางวิทยาศาสตร์ว่า ซัคคาโรไมซิสเซอริวิสิเอ (*Saccharomyces cerevisiae*) ข้อมูลดีเอ็นเอไมโครอาร์เรย์ของผู้ป่วยที่เป็นโรคมะเร็ง และชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของแบคทีเรียซึ่งทำให้เกิดเชื้อวัณโรค หรือ ไมโครแบคทีเรียม ทิวเบอร์คูโลซิส (*Mycobacterium tuberculosis*) เป็นกรณีศึกษา

1.4.2 การจำแนกกลุ่มของข้อมูล โดยอาศัยทฤษฎีการวิเคราะห์การจำแนกประเภท และ ทฤษฎีการวิเคราะห์การถดถอยแบบโลจิสติกทั้งนี้จะใช้ชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ของผู้ป่วยที่เป็นโรคมะเร็ง เป็นกรณีศึกษา

1.5 วิธีการวิจัย

1.5.1 ศึกษาเทคนิคการวิเคราะห์หลายตัวแปรต่างๆ ในเชิงทฤษฎี และการนำไปใช้งาน

1.5.2 ศึกษาค้นคว้าเกี่ยวกับข้อมูลดีเอ็นเอไมโครอาร์เรย์ เพื่อที่จะเลือกชุดของดีเอ็นเอไมโครอาร์เรย์มาวิเคราะห์ให้เหมาะสมกับเทคนิคการวิเคราะห์หลายตัวแปรในแต่ละวิธี

1.5.3 พัฒนาโปรแกรมการคำนวณ ในส่วนที่เกี่ยวข้องกับเทคนิคการวิเคราะห์หลายตัวแปรโดยอาศัยฟังก์ชันที่มีอยู่แล้ว และเขียนขึ้นเพิ่มเติมในส่วนที่ขาดหาย นำโปรแกรมดังกล่าวไปใช้ในการวิเคราะห์ ชุดของข้อมูลดีเอ็นเอไมโครอาร์เรย์ที่ได้เลือกมา ซึ่งในงานวิจัยนี้ ข้อมูลดีเอ็นเอไมโครอาร์เรย์ที่นำมาวิเคราะห์เป็นข้อมูลไมโครอาร์เรย์ที่มีการเก็บรวบรวมจากงานวิจัยซึ่งเผยแพร่ผ่านอินเทอร์เน็ต โดยข้อมูลที่นำมาวิเคราะห์จะอยู่ในรูปของไฟล์ข้อมูลที่ผ่านกระบวนกาจัดเตรียมข้อมูลเบื้องต้นมาแล้วในบางส่วน ทำให้เราสามารถ นำข้อมูลเหล่านี้ มาวิเคราะห์ได้อย่างง่ายดาย

1.5.4 สรุปผลการวิเคราะห์ชุดข้อมูลดีเอ็นเอไมโครอาร์เรย์ในแต่ละชุดข้อมูล สำหรับในแต่ละเทคนิควิธีการ

1.5.5 จัดทำเอกสารสรุปผลงานวิจัยทั้งหมด

1.6 สถานที่ ที่ใช้ในการดำเนินการวิจัยและรวบรวมข้อมูล

1.6.1 สถานที่

(1) ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

(2) สำนักหอสมุด มหาวิทยาลัยเชียงใหม่

(3) ห้องสมุดคณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

(4) ห้องปฏิบัติการวิจัย ชีวสารสนเทศศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

1.6.2 อุปกรณ์ที่ใช้ในการวิจัย

(1) เครื่องคอมพิวเตอร์ส่วนบุคคล (Personal Computer)

(2) โปรแกรมระบบปฏิบัติการไมโครซอฟท์วินโดว์เอ็กซ์พี โปรเฟสชันนั่ล (Windows XP professional)

(3) โปรแกรม อินเทอร์เน็ต เอ็กซ์พลอเรอร์ (Internet Explorer)

(4) โปรแกรมภาษา อาร์ (R) เวอร์ชัน 2.3.1