บทที่ 2
หลักการและทฤษฎีที่เกี่ยวข้อง

สำหรับเนื้อหาในส่วนของหลักการและทฤษฎีที่เกี่ยวข้องนี้ เป็นการอธิบายถึงหลักการทางชีววิทยาและวิธีการต่าง ๆ ที่ถูกนำมาใช้ในการวิเคราะห์ข้อมูล เนื่องจากงานวิจัยในครั้งนี้เป็นการศึกษาข้อมูลทางชีววิทยาที่เป็นข้อมูลการแสดงออกของยีนที่เรียกว่าข้อมูลดีเอ็นเอไมโครอาร์เรย์ โดยอาศัยคอมพิวเตอร์เข้ามาช่วยในการวิเคราะห์ข้อมูลดังกล่าว ซึ่งเป็นศาสตร์ที่เรียกว่าชีวสารสนเทศ (Bioinformatics) ดังนั้นในส่วนแรกของบทนี้จะเป็นการอธิบายถึงความหมายของชีวสารสนเทศศาสตร์ ข้อมูลรหัสพันธุกรรมของดีเอ็นเอ การถอดรหัสพันธุกรรมของดีเอ็นเอ ไมโครอาร์เรย์ดีเอ็นเอ ซึ่งเป็นข้อมูลที่ถูกนำมาใช้ในการศึกษาวิเคราะห์ครั้งนี้ และเครือข่ายการควบคุมกันระหว่างยีน จากนั้นในส่วนถัดไปจะเป็นการอธิบายถึงวิธีการต่างๆ ที่ถูกนำมาใช้ในการวิเคราะห์ข้อมูลในงายวิจัยนี้ อันประกอบไปด้วย การอนุมานด้วยเบย์เซียน วิธีการมาร์คอฟเชนมอนติคาร์โล ซึ่งแบ่งออกเป็นวิธีการต่างๆ หลายวิธีการด้วยกัน ได้แก่ อัลกอริทึมเมโทรโพลิส - แฮสทติงส์ (Metropolis – Hastings Algorithm) วิธีการซิงเกิล - คอมโพเน็นท์ เมโทรโพลิส - แฮสติงส์ (Single – component Metropolis - Hastings) วิธีการกิบบ์ แซมพลิง (Gibb Sampling) และวิธีการของรีเวิร์สซิเบิลจัมพ์ มาร์คอฟเชนมอนติคาร์โล (Reversible Jump MCMC) วิธีการสร้างแบบจำลองและการปรับแบบจำลองให้เหมาะสมด้วยวิธีการกิบบ์ แซมพลิง และในสว่นสุดท้ายขะเป็นการอธิบายถึงโปรแกรมวินบักส์ (WinBUGS program) ซึ่งเป็นโปรแกรมสำเร็จรูปที่ถูกนำมาใช้ในการวิเคราะห์ข้อมูลการแสดงออกของยีนในการทดลองส่วนแรก ซึ่งสามารถอธิบายได้ดังต่อไปนี้คือ

2.1 ชีวสารสนเทศศาสตร์

ชีวสารสนเทศศาสตร์ (Bioinformatics) เป็นการประยุกต์เทคโนโลยีเพื่อการจัดการและการวิเคราะห์ข้อมูลทางชีววิทยา และนำคอมพิวเตอร์มาใช้เพื่อรวบรวม จัดเก็บ วิเคราะห์ และเชื่อมโยงข้อมูลทางชีววิทยาเหล่านั้น หรืออาจกล่าวได้ว่าชีวสารสนเทศเป็นแนวทางการศึกษาค้นคว้าแบบสหวิทยาการ (Interdisciplinary) ระหว่างสาขาชีววิทยากับสาขาการคำนวณ เป้าหมายสูงสุดของชีวสารสนเทศคือ การค้นหาความหมายที่ซ่อนอยู่ในสารสนเทศทางชีววิทยาซึ่งมีปริมาณมาก แล้วนำมาใช้ให้เกิดประโยชน์สูงสุด และให้เกิดความเข้าใจอย่างลึกซึ้งต่อชีววิทยาในระดับพื้นฐานของสิ่งมีชีวิต มีการคาดการณ์ว่าความรู้ใหม่ที่ได้จะส่งผลกระทบอย่างใหญ่หลวงในหลายๆ วงการ เช่น การรักษาโรค เกษตรกรรม สิ่งแวดล้อม พลังงาน และเทคโนโลยีชีวภาพ

การก่อกำเนิดลำดับเบส (Sequence generation) ซึ่งเป็นหน่วยพื้นฐานของยีน การจัดเก็บลำดับย่อย (Subsequent storage) ของลำดับเบส รวมทั้งการตีความ (Interpretation) และการวิเคราะห์ (Analysis) ลำดับเบสเหล่านั้น ทั้งหมดล้วนเป็นงานที่ขึ้นอยู่กับคอมพิวเตอร์โดยตรง อย่างไรก็ตาม ชีววิทยาในระดับโมเลกุลของสิ่งมีชีวิตหนึ่งๆ เป็นเรื่องที่ซับซ้อนมาก จากการวิจัยที่ผ่านมาในระดับที่แตกต่างกัน ทั้งในระดับของกลุ่มยีน (Genome) กลุ่มโปรตีน (Proteome) กลุ่มการถ่ายสำเนา (Transcriptome) และกลุ่มการเผาผลาญพลังงาน (Metabolome) ทำให้ปริมาณข้อมูลจีโนมิกส์มีการเพิ่มขึ้นอย่างรวดเร็ว สิ่งท้าทายอย่างมากในวงการชีวสารสนเทศทุกวันนี้คือการจัดเก็บข้อมูลที่มีปริมาณมากให้มีประสิทธิภาพสูงสุด ดังนั้นจึงเป็นงานสำคัญที่จะต้องจัดให้มีการเข้าถึงข้อมูลได้โดยง่ายและสามารถเชื่อถือได้ ข้อมูลทั้งหมดที่เก็บรวบรวมมาได้ จะต้องถูกนำมาวิเคราะห์เพื่อหาความหมายของข้อมูลก่อนนำไปใช้งานต่อไป ดังนั้นเครื่องมือคอมพิวเตอร์ที่ฉลาดเฉลียวจึงต้องถูกพัฒนาขึ้นมาเพื่อให้สามารถสกัดเอาสารสนเทศทางชีววิทยาที่มีความหมายออกมาจากข้อมูลที่ถูกเก็บรวบรวมมา

มีการคาดการณ์เกี่ยวกับเป้าหมายในระยะยาวของการผนวกเข้าด้วยกัน ของสารสนเทศเพื่อศึกษาเกี่ยวกับกระบวนการดังกล่าวข้างต้น ว่าจะทำให้เราสามารถทำความเข้าใจชีววิทยาของสิ่งมีชีวิตต่างๆ ได้อย่างสมบูรณ์

(Nilges and Ling, 2002)

2.2 ข้อมูลรหัสพันธุกรรมของดีเอ็นเอ (DNA Data)

เนื่องจากงานวิจัยที่ทำมีความเกี่ยวข้องกับงานทางด้านชีววิทยาอย่างมากโดยเฉพาะเรื่องดี-เอ็นเอ จึงจำเป็นจะต้องอธิบายให้เข้าใจเกี่ยวกับเรื่องของดีเอ็นเอและเทคนิคการได้มาของข้อมูลพอสังเขป

สิ่งมีชีวิตทุกสิ่งในโลกล้วนดำรงพันธุ์อยู่ได้โดยการถ่ายทอดลักษณะทางพันธุกรรม (genetic character) จากรุ่นหนึ่งไปสู่อีกรุ่นหนึ่งต่อๆ กันไป ตัวอย่างเช่น มนุษย์เราถ่ายทอดลักษณะทางกายภาพ (สีผม สีผิว สีตา ฯลฯ) จากพ่อแม่ไปสู่ลูก เป็นต้น ลักษณะทางพันธุกรรมที่สืบทอดต่อกันไปนี้แตกต่างกันไปตามเผ่าพันธุ์ของสิ่งมีชีวิต และถูกควบคุมโดยสิ่งที่เรียกว่ายีน (gene) ที่มีอยู่มากมายในเซลล์ของสิ่งมีชีวิต ยีนเหล่านี้เป็นตัวควบคุมลักษณะทางพันธุกรรมให้แสดงออกมาในสิ่งมีชีวิตแต่ละรุ่น ทำให้การแสดงออกของลักษณะทางพันธุกรรมที่ได้รับการถ่ายทอดมามีเพียงบางลักษณะ ซึ่งมีจำนวนน้อยมากเมื่อเทียบกับจำนวนของลักษณะทางพันธุกรรมที่ได้รับการถ่ายทอดมาทั้งหมด ลักษณะทางพันธุกรรมทั้งหมดทั้งที่แสดงออกมาและไม่แสดงออกมาจะถูกถ่ายทอดไปสู่รุ่นต่อไป การถ่ายทอดลักษณะทางพันธุกรรมของสิ่งมีชีวิตนี้จำเป็นต้องอาศัยข้อมูลทางพันธุกรรมของสิ่งมีชีวิตหรือที่เรียกว่า สารพันธุกรรม โดยสำหรับสิ่งมีชีวิตส่วน

ใหญ่แล้ว จะมีดีเอ็นเอ (DNA) เป็นสารพันธุกรรม หรืออาจกล่าวได้ว่า ดีเอ็นเอเป็นที่เก็บข่าวสารทางพันธุกรรมของสิ่งมีชีวิต ข้อมูลทางพันธุกรรมถูกจัดเก็บอยู่ในรูปแบบที่เป็นรหัสต่างๆ จัดเรียงอยู่บนดีเอ็นเอ รหัสเหล่านี้เรียกว่า รหัสพันธุกรรม (genetic code) การถ่ายทอดลักษณะทางพันธุกรรมจึงเป็นการถ่ายทอดรหัสพันธุกรรม และเนื่องจากยีนเป็นชุดของรหัสพันธุกรรมที่จัดเรียงอยู่บนดีเอ็นเอ การถ่ายทอดลักษณะทางพันธุกรรมจึงเป็นการถ่ายทอดยีนทั้งหมดด้วย

ความสามารถในการควบคุมการแสดงออกของลักษณะทางพันธุกรรมในสิ่งมีชีวิตแต่ละรุ่นของยีน เกี่ยวเนื่องมาจากการที่ยีนมีความเกี่ยวข้องกับการสังเคราะห์โปรตีนของเซลล์ ซึ่งถือว่ามีความเกี่ยวข้องกับการแสดงลักษณะและการดำรงชีวิตของสิ่งมีชีวิต ดังนั้นจึงกล่าวได้ว่าดีเอ็นเอเป็นตัวควบคุมลักษณะทางพันธุกรรมของสิ่งมีชีวิต โดยควบคุมการสังเคราะห์โปรตีนนั่นเอง

การที่ดีเอ็นสามารถควบคุมลักษณะทางพันธุกรรมและการสังเคราะห์โปรตีนของสิ่งมีชีวิตนี่เอง จึงมีผู้พยายามศึกษาและทำความเข้าใจความหมายของรหัสพันธุกรรมที่ดีเอ็นเอเก็บไว้เพื่อเข้าใจลักษณะทางพันธุกรรมของสิ่งมีชีวิต และกลไกการตอบสนองต่อความต้องการหรือการกระตุ้นจากสิ่งแวดล้อมของเซลล์สิ่งมีชีวิต ในปัจจุบันจึงมีการคิดค้นเครื่องมือเพื่อช่วยในการศึกษาและวิเคราะห์ข่าวสารทางพันธุกรรมเหล่านี้ หนึ่งในเครื่องมือที่ช่วยในการศึกษาข่าวสารทางพันธุกรรม คือ ไมโครอาร์เรย์ดีเอ็นเอ (DNA microarray) ซึ่งได้มาจากกระบวนการถอดรหัสพันธุกรรมของเซลล์สิ่งมีชีวิตในการทดลองเชิงชีววิทยา ข้อมูลที่ได้จากไมโครอาร์เรย์จะถูกจัดเก็บอยู่ในรูปแบบของตารางที่เก็บข้อมูลเชิงตัวเลขเรียกว่า ประวัติการถอดรหัสพันธุกรรม (transcription profile) ซึ่งสามารถนำข้อมูลนี้ไปวิเคราะห์โดยใช้วิธีการเชิงคำนวณต่อไปได้

(ชโลธร เหลี่ยมวิรัช, 2546)

2.3 การถอดรหัสพันธุกรรมของดีเอ็นเอ

สิ่งมีชีวิตส่วนใหญ่มีดีเอ็นเอ (DNA: Deoxyribonucleic acid) เป็นสารพันธุกรรม ซึ่งบรรจุข่าวสารทางพันธุกรรมของสิ่งมีชีวิตในรูปแบบที่เป็นรหัสพันธุกรรม (genetic code) ดีเอ็นเอใช้รหัสพันธุกรรมที่เก็บไว้นี้ในการควบคุมลักษณะทางพันธุกรรมและการสังเคราะห์โปรตีนของสิ่งมีชีวิต สิ่งมีชีวิตถ่ายทอดลักษณะทางพันธุกรรมจากรุ่นหนึ่งไปสู่อีกรุ่นหนึ่งได้โดยการถ่ายทอดดีเอ็นเอนั่นเอง

ดีเอ็นเอเป็นสารอินทรีย์ที่มีคุณสมบัติเป็นกรดและพบในนิวเคลียสของเซลล์สิ่งมีชีวิต ดีเอ็นเอประกอบขึ้นจากการเชื่อมกันของ สายพอลินิวคลีโอไทด์(polynucleotide) หรือสายพันธุกรรม 2 สาย ซึ่งทำให้เกิดโครงสร้างที่มีลักษณะเป็นเกลียว สายพันธุกรรมแต่ละสายประกอบด้วยนิวคลีโอไทด์ 4 ชนิด ได้แก่ อะดีนีน (Adenine-A), ไทอะมีน (Thymine-T),

กัวนีน (Guanine-G), และไซโตซีน (Cytosine-C) นิวคลีโอไทด์เหล่านี้มีธาตุไนโตรเจนเป็นองค์ประกอบหลักและมีคุณสมบัติเป็นเบส จึงมีการเรียกนิวคลีโอไทด์เหล่านี้สั้นๆว่า เบส (base) นิวคลีโอไทด์ทั้ง 4 ชนิดนี้แตกต่างกันตามส่วนประกอบที่เป็นเบส และมีชื่อเรียกสั้นๆว่า เบส A เบส G เบส C และเบส T ตามลำดับ นิวคลีโอไทด์เหล่านี้เชื่อมต่อๆ กันไปเป็นสายกลายเป็นสายพันธุกรรม ลำดับการเชื่อมต่อของนิวคลีโอไทด์หรือที่เรียกว่าลำดับเบส (base sequence) ที่แตกต่างกันทำให้สายพันธุกรรมแต่ละสายมีความแตกต่างกันด้วย แต่สำหรับสายพันธุกรรม 2 สายที่เชื่อมต่อกันกลายเป็นดีเอ็นเอนั้น แต่ละสายมีลำดับเบสที่เบสแต่ละตำแหน่งสามารถจับคู่เข้ากับเบสในลำดับเบสของอีกสายหนึ่งในตำแหน่งที่ตรงกันได้ โดยเบส A จับคู่กับเบส T และ เบส G จับคู่กับเบส C เสมอ เราเรียกลำดับเบสที่สามารถเข้าคู่กับลำดับเบสอีกลำดับหนึ่งได้ว่า ลำดับเบสประกอบ (complementary base sequence) ตัวอย่างเช่น ลำดับเบส G-T-C-C-T-A มีลำดับเบสประกอบเป็น C-A-G-G-A-T เป็นต้น ลำดับเบสที่ต่างกันเมื่อใช้เบสในจำนวนที่แตกต่างกันบนสายพันธุกรรมทำให้เกิดรหัสที่แตกต่างกันอย่างมากมาย ตัวอย่างเช่น ถ้าพิจารณาให้เบส 2 โมเลกุลเรียงต่อกันเป็นรหัสแล้วจะได้จำนวนรหัสที่เกิดจากการจัดเรียงลำดับของเบสเพียง 2 โมเลกุลนี้มากถึง 16 แบบ ($4^2 = 16$) เป็นต้น ในความเป็นจริงแล้ว สายพันธุกรรมประกอบด้วยการจัดเรียงลำดับของเบสหลายโมเลกุลซึ่งอาจมีมากถึงระดับหมื่นโมเลกุล จึงก่อให้เกิดความหลากหลายทางพันธุกรรมของสิ่งมีชีวิตมากมาย

(ชโลธร เหลี่ยมวิรัช, 2546)

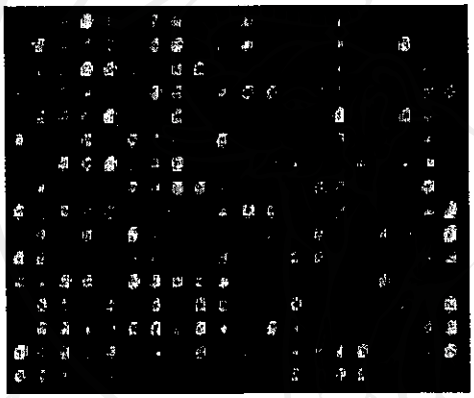
2.4 ไมโครอาร์เรย์ดีเอ็นเอ

โดยปกติแล้ว ทุกๆ เซลล์ของร่างกายจะมีกลุ่มของโครโมโซมและกลุ่มยีนที่สอดคล้องกันบรรจุอยู่ภายในอย่างสมบูรณ์ เมื่อยีนเหล่านี้ทำงาน แม้อาจเป็นเพียงแค่ส่วนย่อยๆของยีนก็ตาม แต่ก็เป็นส่วนย่อยๆที่มีความเฉพาะเจาะจง และร่วมกันแสดงคุณสมบัติที่เป็นเอกลักษณ์ของเซลล์แต่ละรูปแบบ การแสดงออกของยีน (Gene expression) เป็นคำที่ใช้อธิบายถึงการถอดรหัสสารสนเทศที่บรรจุอยู่ภายในดีเอ็นเอ ไปเป็นโมเลกุลเอ็มอาร์เอ็นเอ (mRNA) ซึ่งทำหน้าที่ส่งสาร และจะถูกแปลไปเป็นโปรตีนที่กระทำหน้าที่ที่สำคัญทั้งหมดของเซลล์ นักวิทยาศาสตร์ศึกษาคุณลักษณะและปริมาณของเอ็มอาร์เอ็นเอที่ผลิตโดยเซลล์หนึ่งๆ เพื่อศึกษาเกี่ยวกับการแสดงออกของยีน ซึ่งจะช่วยให้มีความเข้าใจเกี่ยวกับเซลล์ได้ลึกซึ้งมากขึ้นว่ามีการตอบสนองต่อความต้องการที่เปลี่ยนแปลงต่างๆ ของตัวมันเองได้อย่างไร การแสดงออกของยีนมีความซับซ้อนสูงมาก และมีกระบวนการที่มีกฎเกณฑ์เข้มงวดเกี่ยวกับการตอบสนองของเซลล์ทั้งต่อสิ่งแวดล้อมที่มากระตุ้นและต่อความ

ต้องการที่เปลี่ยนแปลงของตัวมันเอง กลไกนี้เปรียบเสมือนเป็นทั้งเครื่องมือในการเปิดและปิด เพื่อควบคุมให้ยีนมีระดับการแสดงออกเพิ่มขึ้นหรือลดลงตามสภาวะที่เป็นจริง

ไมโครอาร์เรย์ดีเอ็นเอแสดงดังรูป 2.1 ซึ่งช่วยในการศึกษาวิเคราะห์ข้อมูลการแสดงออกของยีนได้ ซึ่งเป็นผลจากการพัฒนาของเทคโนโลยีไมโครอาร์เรย์ดีเอ็นเอแสดงดังรูป 2.2 ที่ช่วยให้เราสามารถวิเคราะห์ลำดับของดีเอ็นเอจำนวนมากได้ในเวลาเดียวกันเพื่อการวิจัยด้านจีโนมิกส์และการวินิจฉัยโรค ในตอนเริ่มต้น การประยุกต์ใช้ไมโครอาร์เรย์เป็นไปเพื่อการศึกษาเรื่องการแสดงออกของยีนที่แตกต่างกัน (Differential gene expression) และการจัดทำแผนผังยีน (Gene mapping) และไมโครอาร์เรย์ถูกใช้ครั้งแรกในปี 1997 เพื่อศึกษาการแสดงออกของยีนโดยรวม

(DeRisi et al., 1997)



รูป 2.1 ไมโครอาร์เรย์ดีเอ็นเอ

(แหล่งที่มา: Yukhananov and Loguinov, 2003)

การวิเคราะห์การแสดงออกของยีนในระยะเริ่มต้น นักวิจัยจะทำการค้นคว้ายีนที่สัมพันธ์กันได้ในจำนวนน้อยๆเท่านั้นในการทดลองแต่ละครั้ง แต่การเกิดขึ้นอย่างรวดเร็วของเครื่องมือใหม่ๆ ทำให้นักวิทยาศาสตร์สามารถวิเคราะห์การแสดงออกของยีนจำนวนมากได้ในการทดลองเพียงครั้งเดียวอย่างรวดเร็วและมีประสิทธิภาพ นั่นแสดงให้เห็นว่าเครื่องมือที่มีศักยภาพสำหรับนักวิจัยได้เกิดขึ้น ส่วนหนึ่งเป็นเพราะความก้าวหน้าในด้านเทคโนโลยี นักวิทยาศาสตร์ใช้เทคโนโลยีไมโครอาร์เรย์ทั้งเพื่อพยายามทำความเข้าใจแนวคิดพื้นฐานของการเจริญเติบโตและการพัฒนา และเพื่อค้นหาสาเหตุสำคัญทางพันธุกรรมที่ก่อให้เกิดโรคในมนุษย์หลายๆโรค

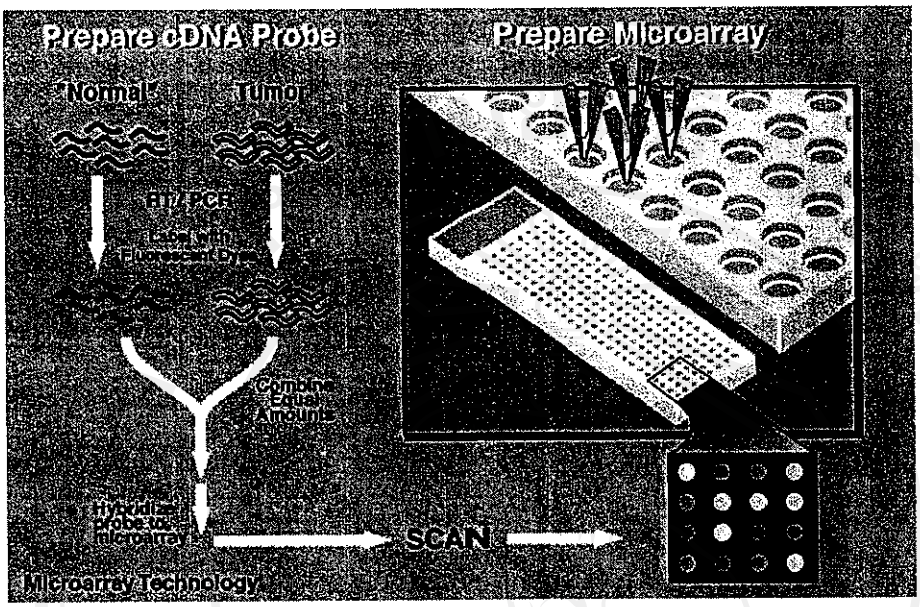
การพัฒนาร่วมกันทั้งในด้านของความรู้และด้านเทคโนโลยี นับว่ามีส่วนอย่างมากในการช่วยอำนวยความสะดวกต่อการศึกษาการแสดงออกของยีนและการค้นพบบทบาทหน้าที่ของยีนที่เฉพาะเจาะจงในการพัฒนาของโรค ผลที่ได้รับจากโครงการศึกษาจีโนมมนุษย์ (Human genome

project) ทำให้เกิดสารสนเทศที่เกี่ยวกับลำดับดีเอ็นเอของจีโนมมนุษย์เพิ่มขึ้นในปริมาณมาก และจากลำดับดีเอ็นเอที่เพิ่มขึ้นเหล่านั้น ทำให้นักวิจัยค้นพบยีนใหม่ๆ เพิ่มขึ้นด้วย แต่ความท้าทายที่นักวิทยาศาสตร์กำลังเผชิญอยู่ตอนนี้คือการค้นหาวิธีการที่จะรวบรวมและจัดทำบัญชีรายชื่อของสารสนเทศปริมาณมหาศาลเหล่านี้ให้อยู่ในรูปแบบที่สามารถนำไปใช้งานได้ ผลกระทบที่แต่ละวงการจะได้รับจากโครงการจีโนมมนุษย์จะเห็นชัดเจนขึ้นก็ต่อเมื่อเราสามารถระบุหน้าที่ของยีนใหม่ได้แล้วเท่านั้น

การพัฒนาทางด้านเทคโนโลยี อาจทำให้การระบุเอกลักษณ์และการจัดแบ่งประเภทสารสนเทศของลำดับดีเอ็นเอเหล่านี้ รวมทั้งการกำหนดหน้าที่ของการทำงานให้แก่ยีนใหม่เหล่านี้ทำได้สะดวกมากขึ้น เนื่องจากไมโครอาร์เรย์หนึ่งๆ ทำงานได้โดยอาศัยความสามารถของโมเลกุลเอ็มอาร์เอ็นเอที่กำหนดให้ ในการเข้าไปจับคู่ (Binding) หรือเข้าคู่กัน (Hybridize) กับแผ่นดีเอ็นเอที่เฉพาะเจาะจง และเนื่องจากการใช้เพียงอาร์เรย์เดียวที่สามารถบรรจุตัวอย่างดีเอ็นเอได้เป็นจำนวนมาก จึงทำให้นักวิทยาศาสตร์สามารถกำหนดระดับการแสดงออกของยีนเป็นร้อยหรือเป็นพันยีนภายในเซลล์ หนึ่งๆได้จากการทดลองเพียงครั้งเดียว โดยการตรวจวัดปริมาณของเอ็มอาร์เอ็นเอที่เข้าไปติดในแต่ละตำแหน่งของอาร์เรย์ โดยอาศัยคอมพิวเตอร์เป็นตัวช่วย ปริมาณของเอ็มอาร์เอ็นเอดังกล่าวจะถูกตรวจวัดด้วยความแม่นยำ แล้วสร้างข้อมูลโดยรวมของการแสดงออกของยีนภายในเซลล์ ดังนั้นจึงกล่าวได้ว่าไมโครอาร์เรย์เป็นเครื่องมือสำหรับการวิเคราะห์การแสดงออกของยีนที่ประกอบด้วยเยื่อบุขนาดเล็กหรือเป็นแผ่นสไลด์ที่ทำมาจากแก้วแล้วบรรจุตัวอย่างของหลายๆยีนเอาไว้โดยจัดวางในรูปแบบที่สม่ำเสมอ

เมื่อสารสนเทศถูกเก็บสะสมไว้มากขึ้น นักวิทยาศาสตร์จึงสามารถใช้ไมโครอาร์เรย์เพื่อตั้งคำถามและทำการทดลองที่มีความซับซ้อนมากขึ้นได้ ด้วยการพัฒนาใหม่ๆ นักวิจัยจะสามารถอนุมานหน้าที่ที่เป็นไปได้ของยีนใหม่ได้ โดยพิจารณาจากความคล้ายกันในรูปแบบการแสดงออกกับยีนที่รู้จักแล้ว ท้ายที่สุดการศึกษานี้จะทำให้เรามั่นใจได้ถึงการขยายขนาดของกลุ่มยีนที่มีอยู่ เปิดเผยรูปแบบใหม่ๆของการแสดงออกร่วมกันของยีนผ่านกลุ่มยีน และเปิดเผยกลุ่มใหม่ทั้งหมดของยีน ยิ่งไปกว่านั้น เนื่องจากผลผลิตของยีนหนึ่ง โดยปกติแล้วจะมีผลกระทบต่อยีนอื่นอีกหลายตัว ความเข้าใจของเราเกี่ยวกับว่ายีนเหล่านี้ทำงานร่วมกันได้อย่างไรกำลังจะชัดเจนขึ้นโดยผ่านการวิเคราะห์ดังกล่าว และความรู้ที่ชัดเจนของความสัมพันธ์ภายในเหล่านี้ก็จะถูกรวบรวมไว้ด้วยกัน นอกจากนี้การใช้ไมโครอาร์เรย์อาจทำให้การระบุความเฉพาะเจาะจงของยีนที่เกี่ยวข้องกับการเจริญเติบโตของเชื้อโรคทำได้เร็วมากขึ้น โดยการทำให้นักวิทยาศาสตร์สามารถตรวจสอบยีนจำนวนมากๆได้ นอกจากนี้ เทคโนโลยีนี้ยังช่วยให้สามารถทำการทดสอบเกี่ยวกับการร่วมกันแสดงออกของยีนและหน้าที่ของยีนเหล่านั้นในระดับเซลล์ได้ด้วย และทำให้ทราบว่าผลผลิตของยีนจำนวน

มากเท่าใดที่ทำงานร่วมกันเพื่อก่อให้เกิดการตอบสนองทางกายภาพและทางเคมีต่อความต้องการต่างๆของเซลล์



รูป 2.2 เทคโนโลยีไมโครอาร์เรย์

(แหล่งที่มา http://www.genome.gov/10000533, 2006[online])

ไมโครอาร์เรย์ดีเอ็นเอมีขนาดเล็ก มีลักษณะเป็นแผ่นสี่เหลี่ยมแข็งที่มีการวางลำดับของยีนจำนวนมากที่แตกต่างกันอย่างมีแบบแผนในตำแหน่งที่แน่นอน แผ่นที่ใช้วางมักจะเป็นแผ่นสไลด์แก้วแบบที่ใช้กับกล้องจุลทรรศน์ มีขนาดประมาณสองนิ้วก้อยวางชิดกัน หรืออาจใช้แผ่นซิลิคอนชิป (Silicon chips) หรือใช้เยื่อไนลอน (Nylon membranes) ก็ได้ ดีเอ็นเอที่ต้องการจะถูกพิมพ์หรือสังเคราะห์ไปวางเป็นจุดเล็กๆบนแผ่นดังกล่าว เรียกแต่ละจุดนี้ว่าสปอต (Spot) นอกจากนี้แต่ละสปอตอาจจะใช้เป็นดีเอ็นเอ ซีดีเอ็นเอหรือโอลิโกนิวคลีโอไทด์ (Oligonucleotides) ก็ได้โดยที่โอลิโกนิวคลีโอไทด์คือสายดีเอ็นเอเส้นสั้นๆที่มีนิวคลีโอไทด์ยาวประมาณ 5 ถึง 50 นิวคลีโอไทด์ และเรียกดีเอ็นเอบนแผ่นสไลด์นี้ว่าดีเอ็นเอเป้าหมาย (Target DNA)

นักวิทยาศาสตร์สามารถสกัดสารสนเทศเกี่ยวกับสภาวะของโรคออกมาจากแผ่นแก้ว หรือแผ่นซิลิคอนชิปขนาดเล็กที่บรรจุลำดับของยีนจำนวนมากและมีความแตกต่างกันได้ เพราะอาศัยการตรวจสอบการเข้าคู่กัน (Hybridization probing) ซึ่งเป็นเทคนิคหนึ่งที่ใช้การติดฉลากเรืองแสงให้แก่โมเลกุลกรดนิวคลีอิคที่เป็นโพรบที่เคลื่อนที่ (Mobile probes) เพื่อพิสูจน์โมเลกุลประกอบ (Complementary molecules) ซึ่งเกิดจากการที่ลำดับย่อยสองลำดับที่สอดคล้องกันมาจับคู่กัน แต่ละสายดีเอ็นเอย่อยประกอบด้วยนิวคลีโอไทด์ที่แตกต่างกันสี่ตัวคือ อะดีนีน, ไทอะมีน,

กัวนีน, และไซโตซีน ที่เชื่อมต่อกัน โดยที่อะดีนีนจะจับกับไทอะมีน ในขณะที่กัวนีนจับกับไซโตซีน ดังนั้นลำดับที่เป็นคู่ของ G-T-C-C-T-A คือ C-A-G-G-A-T เมื่อลำดับที่เข้าคู่กันได้ของดีเอ็นเอเป้าหมาย (Target DNA) ที่ถูกติดอยู่บนแผ่นแก้วกับโพรบดีเอ็นเอที่เคลื่อนที่ได้ (Mobile probe DNA) มาพบกัน มันจะล็อคตัวเองเข้าไว้ด้วยกันเรียกว่ามันเกิดการเข้าคู่กัน (Hybridize)

ยกตัวอย่างเช่นเรากำลังพิจารณากลุ่มเซลล์สองกลุ่มที่แตกต่างกัน โดยที่กลุ่มที่หนึ่งเป็นเซลล์ปกติที่มีสุขภาพดี และกลุ่มที่สองเป็นเซลล์ที่ติดโรค ทั้งสองเซลล์ต่างก็มีกลุ่มของยีนแบบเดียวกันสี่ตัวคือ ยีน A, B, C และ D นักวิทยาศาสตร์สนใจที่จะค้นหาข้อมูลโดยรวมเกี่ยวกับการแสดงออกของยีนสี่ตัวนี้ในทั้งสองเซลล์ เพื่อจะทำการทดลองนี้ นักวิทยาศาสตร์ได้แยกเอ็มอาร์เอ็นเอ ออกจากแต่ละเซลล์ และใช้ เอ็มอาร์เอ็นเอนี้เป็นแม่แบบในการสร้างซีดีเอ็นเอที่มีป้ายเรืองแสงติดไว้ โดยใช้ป้ายที่มีสีแตกต่างกัน (ในที่นี้มีสีแดงและสีเขียว) เพื่อให้สามารถจำแนกความแตกต่างจากผลการทดลองที่ตามมาได้ ตัวอย่างทั้งสองกลุ่มที่ถูกติดป้ายไว้จะผสมกับยีนที่ถูกผนึกติดอยู่กับไมโครอาร์เรย์และเกิดเป็นโมเลกุลใหม่ที่สัมพันธ์กับการแสดงออกของยีนในแต่ละเซลล์

หลังจากขั้นตอนการจับคู่กันเสร็จสมบูรณ์แล้ว นักวิจัยจะอ่านไมโครอาร์เรย์โดยใช้เครื่องอ่านหรือเครื่องสแกน (Scanner) ที่ประกอบด้วยแสงเลเซอร์ กล้องจุลทรรศน์ และกล้องถ่ายภาพ ป้ายเรืองแสงจะถูกกระตุ้นด้วยเลเซอร์ ส่วนกล้องจุลทรรศน์และกล้องถ่ายภาพจะทำงานร่วมกันในการสร้างภาพดิจิตอล (Digital image) ของอาร์เรย์ จากนั้นข้อมูลเหล่านี้จะถูกเก็บไว้ในคอมพิวเตอร์ และใช้โปรแกรมพิเศษเพื่อคำนวณสัดส่วนสีแดงต่อสีเขียว หรือใช้วิธีลบข้อมูลพื้นหลัง (Background data) ออกจากแต่ละสปอตของไมโครอาร์เรย์โดยการวิเคราะห์ภาพดิจิตอลของอาร์เรย์ หลังจากนั้นโปรแกรมจะทำการสร้างตารางเพื่อเก็บค่าความเข้มของสีของแต่ละอาร์เรย์เพื่อนำไปใช้ในการวิเคราะห์หาความหมายในเชิงชีววิทยาต่อไป นี่เป็นเพียงตัวอย่างเบื้องต้นที่ใช้อธิบายความสำคัญของการออกแบบการทดลอง บางการทดลองไมโครอาร์เรย์สามารถบรรจุสปอตเป้าหมาย (Target spots) ได้ถึงสามหมื่นสปอต ดังนั้น ข้อมูลที่ได้จากอาร์เรย์เพียงอาร์เรย์เดียวจึงสามารถเพิ่มปริมาณขึ้นได้อย่างรวดเร็ว และเราสามารถสรุปขั้นตอนการทดลองไมโครอาร์เรย์ดีเอ็นเอได้ดังต่อไปนี้

(1) เตรียมไมโครอาร์เรย์ที่มีสปอตของดีเอ็นเอหรือลำดับยีนที่ต้องการถูกผนึกไว้ เรียกดีเอ็นเอหรือลำดับยีนที่อยู่บนไมโครอาร์เรย์เหล่านี้ว่า ดีเอ็นเอเป้าหมาย ซึ่งดีเอ็นเอเป้าหมายนี้อาจเป็นดีเอ็นเอที่เราต้องการทราบลักษณะเฉพาะ (Identity) ของมันหรือเป็นดีเอ็นเอที่เราสามารถหาได้ง่าย ซึ่งสามารถนำมาใช้เป็นตัวทดลองเพื่อเปรียบเทียบการแสดงออกของยีน

(2) เตรียม cDNA ซึ่งอาจได้มาจาก mRNA ที่มีลำดับยีนที่ทราบอยู่ก่อนแล้ว หรือ mRNA ที่เราต้องการทราบยีนที่ถูกแสดงออกมา

(3) นำ cDNA เข้าสู่กระบวนการติดป้ายเรืองแสง เรียก cDNA ที่ติดป้ายเรืองแสงว่า โพรบเคลื่อนที่ (Mobile probe)

(4) นำ cDNA ที่ได้จากขั้นตอนที่ (3) ไปเพาะตัวบนไมโครอาร์เรย์ที่เตรียมไว้ในขั้นตอนที่ (1)

(5) ตรวจสอบการผสมของโพรบเคลื่อนที่ กับดีเอ็นเอเป้าหมายโดยใช้เลเซอร์ยิงกระตุ้นให้มีการเรืองแสง แล้วสร้างเป็นรูปภาพดิจิตอลของไมโครอาร์เรย์เก็บไว้ในเครื่องคอมพิวเตอร์

(6) วิเคราะห์ข้อมูลโดยใช้วิธีการเชิงคำนวณ (Computational method)

(ชโลธร เหลี่ยมวิรัช, 2546)

2.5 เครือข่ายการควบคุมกันระหว่างยีน (Gene Regulatory Network)

กระบวนการทำงานภายในเซลล์เป็นการทำงานร่วมกันของยีนต่างๆ รวมทั้งยังมีปัจจัยอื่นที่เข้ามามีส่วนเกี่ยวข้อง เช่น โปรตีน เอนไซม์ เป็นต้น ซึ่งการทำงานประสานร่วมมือกันระหว่างสิ่งต่าง ๆ ภายในเซลล์นี้เป็นกระบวนการที่มีความซับซ้อนมาก สำหรับกระบวนการทำงานระหว่างยีนนั้น การแสดงออกของยีนหนึ่งอาจถูกควบคุมโดยการแสดงออกของยีนอื่นๆ หรือปัจจัยอื่นอีกทีหนึ่ง หรือยีนดังกล่าวนี้อาจทำหน้าที่เสมือนเป็นตัวกลางเพื่อไปควบคุมการแสดงออกของยีนอื่นๆ ต่อไป ซึ่งสามารถมองการทำงานร่วมกันของยีนต่าง ๆ นี้ได้เป็นลักษณะความสัมพันธ์แบบเครือข่ายการทำงานระหว่างยีน

ภายในเครือข่ายยีนที่มีความสัมพันธ์กันอย่างซับซ้อน ยีนสามารถถูกมองเป็นโหนดต่าง ๆ ภายในเครือข่าย โดยที่เครือข่ายจะมีข้อมูลนำเข้า (Input) เป็นโปรตีน ตัวอย่างเช่น โปรตีนทรานสคริปท์ชันแฟกเตอร์ (Transcriptional factor: TF) เป็นต้น และได้ผลลัพธ์หรือข้อมูลนำออก (Output) เป็นระดับการแสดงออกของยีน (Gene expression level) นอกจากนี้ภายในตัวของโหนดเองสามารถมองเป็นฟังก์ชัน (functions) ได้ โดยฟังก์ชันนี้จะได้มาจากการรวมกันของฟังก์ชันพื้นฐาน (Basic function) ของข้อมูลนำเข้าของเครือข่าย ฟังก์ชันต่างๆ นี้จะถูกตีความในระหว่างกระบวนการดำเนินการเกี่ยวกับสารสนเทศ (Information processing) ภายในเซลล์ซึ่งเป็นตัวกำหนดพฤติกรรมของเซลล์ โดยที่ตัวขับเคลื่อนหลักภายในเซลล์คือระดับของโปรตีนบางตัวที่ส่งผลกระทบต่อการทำงานภายในเซลล์ ในขณะนี้พบว่าความเข้าใจเกี่ยวกับเครือข่ายของยีนนั้นยัง

อยู่ในระดับเบื้องต้นเท่านั้น และถือว่าจะเป็นก้าวต่อไปในทางด้านชีววิทยาที่จะต้องพยายามทำการหาข้อสรุปเกี่ยวกับฟังก์ชันของแต่ละโหนดยีน เพื่อเป็นตัวช่วยในการจำลองพฤติกรรมของเซลล์

แบบจำลองทางคณิตศาสตร์ (Mathematical models) ได้ถูกนำมาประยุกต์ใช้กับเครือข่ายการควบคุมกันระหว่างยีนเพื่อทำการ ทำนายเครือข่ายของยีน เทคนิคในการสร้างแบบจำลอง (modeling technique) หลากหลายเทคนิคได้ถูกนำเข้ามาใช้ อาทิเช่น เครือข่ายแบบบูลีน เครือข่ายเพทรี (Petri nets) เครือข่ายเบย์เซียน (Bayesian networks) แบบจำลองเกาส์เซียนเชิงกราฟ (Graphical Gaussian model) สโตคาสติก โพรเซส แคลคูลิ (Stochastic Process Calculi) และสมการระบบอนุพันธ์ (Differential equations)

(Wikipedia, 2006[online])

2.6 การอนุมานด้วยเบย์เซียน (Bayesian Inference)

เราจะพบว่าการใช้งานวิธีการมาร์คอฟเชนมอนติคาร์โล (Markov Chain Monte Carlo) มักจะใช้งานควบคู่ไปกับวิธีการอนุมานด้วยเบย์เซียน (Bayesian Inference) ดังนั้นในส่วนนี้จึงขออธิบายถึงวิธีการอนุมานด้วยเบย์เซียนก่อน แล้วจึงกล่าวถึงวิธีการมาร์คอฟเชนมอนติคาร์โลในหัวข้อถัดไป ดังต่อไปนี้

การอนุมานด้วยเบย์เซียนเป็นวิธีการในการอนุมานหาการแจกแจงร่วมของตัวแปรต่าง ๆ แทนด้วยสัญลักษณ์ $P(H,X)$ กำหนดให้ X เป็นตัวอย่างข้อมูลหนึ่งที่เราสนใจ และ H เป็นสมมติฐานเกี่ยวกับตัวอย่างข้อมูล X แล้ว $P(H \mid X)$ จะหมายถึงความน่าจะเป็นที่สมมติฐาน H จะเป็นจริงเมื่อกำหนดตัวอย่างข้อมูล X มาให้ เราสามารถนำทฤษฎีของเบย์ (Baye's Theory) เข้ามาช่วยคำนวณหาการแจกแจงร่วมของตัวแปรต่าง ๆ $P(H,X)$ โดยการคำนวณหาความน่าจะเป็นโพสทีเรีย (Posterior Probability) หรือการแจกแจงของสมมติฐาน H ที่มีเงื่อนไขภายใต้ข้อมูลที่เราสนใจ X แสดงสมการตามทฤษฎีของเบย์ได้ดังสมการที่ (1)

$$P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)} \qquad (1)$$

จากการแจกแจงโพสทีเรียในสมการ (1) จะประกอบด้วยเทอมของความน่าจะเป็นสี่ค่าด้วยกัน เพื่อความเข้าใจจะกำหนดให้ข้อมูลที่เราสนใจที่นำมาพิจารณาเป็นเซตของยีนชนิดต่างๆ ที่มีการแสดงออกเพื่อผลิตเอนไซม์ได้ต่างชนิดกัน ถ้าเรามีข้อมูล X หนึ่งที่เป็นข้อมูลการแสดงออกของยีน A และ H เป็นสมมติฐานที่ว่า A เป็นยีนที่ผลิตเอนไซม์ a โดย กำหนดให้แต่ละตัวอย่างข้อมูล $X = (x_1, x_2, ..., x_n)$ ถูกแทนด้วยเวกเตอร์ขนาด n โดยที่ n เป็นจำนวนของคุณลักษณะต่างๆ ของ

ตัวอย่างข้อมูล X หรือคือจำนวนแอททริบิวต์ (Attribute) ของข้อมูลนั่นเอง ซึ่งข้อมูลในแอทริบิวต์ต่างๆ จะแทนด้วย $A_1, A_2, ..., A_n$ ตามลำดับ แล้วค่าความน่าจะเป็นทั้งสี่ค่าจะอธิบายได้ดังนี้

$P(H \mid X)$ เป็นการแจกแจงโพสทีเรีย (Posterior Distribution) ของ H ที่มีเงื่อนไขคือ X นั่นคือ $P(H \mid X)$ จะบอกถึงระดับความเชื่อมั่นที่จะเชื่อได้ว่า A คือยีนที่ผลิตเอนไซม์ a เมื่อเราทราบว่า A มีการแสดงออกตามข้อมูล X

$P(X \mid H)$ คือความน่าจะเป็นอย่างมีเงื่อนไข (Conditional probability) ของ X โดยที่เงื่อนไขคือ H นั่นคือความน่าจะเป็นที่ยีน A จะมีการแสดงออกในลักษณะ X เมื่อเราทราบว่า A ผลิตเอนไซม์ a

$P(X)$ เป็นการกระจายที่เป็นไปได้ (Probabilistic distribution) ของ X โดยที่จากตัวอย่างของเรา จะได้ว่า $P(X)$ คือความน่าจะเป็นที่ยีน A จะมีการแสดงออกในลักษณะ X

$P(H)$ เป็นการแจกแจงไพรเออร์ (Prior distribution) ของ H จากตัวอย่างข้างต้น $P(H)$ หมายถึงความน่าจะเป็นที่ยีน A จะผลิตเอนไซม์ a โดยที่เราไม่ต้องพิจารณาว่ายีน A จะมีการแสดงออกในลักษณะใด

โดยทั่วไป $P(H)$, $P(X)$ และ $P(X \mid H)$ สามารถถูกประมาณได้จากข้อมูลที่มีอยู่ ส่วน $P(H \mid X)$ เราสามารถคำนวณได้โดยใช้ทฤษฎีของเบย์เข้ามาช่วยในการคำนวณจาก $P(H)$, $P(X)$ และ $P(X \mid H)$ ดังสมการที่ (1)

(ออมพิไล มโนรัตน์, 2548)

2.7 มาร์คอฟเชนมอนติคาร์โล (Markov Chain Monte Carlo)

วิธีการมาร์คอฟเชนมอนติคาร์โล หรือ MCMC เกิดจากการรวมกันระหว่างวิธีการมอนติคาร์โล (Monte Carlo) และมาร์คอฟเชน (Markov Chain) โดยทั่วไปแล้ววิธีการ MCMC นี้มักจะทำงานร่วมกับวิธีการเบย์เซียนและในบางครั้งจะอาศัยวิธีการทางสถิติที่เกี่ยวกับความถี่ (Frequentist) เพื่อทำการอนุมาน (Inference) หาค่าของปัจจัยของแบบจำลอง (Model parameters) หรือเพื่อทำการทำนาย (Prediction) โดยวิธีการเบย์เซียนจะทำการอนุมานหาการแจกแจงโพสทีเรีย (Posterior distribution) ของปัจจัยของแบบจำลองเมื่อทราบข้อมูล ส่วนวิธีการทางสถิติที่เกี่ยวกับความถี่นั้นจะเป็นการหาการแจกแจง (distribution) ของข้อมูลที่เราสนใจ (Observation) เมื่อทราบค่าของปัจจัย (Parameters) จากนั้นด้วยวิธีการของมอนติคาร์โลจะทำการสร้างข้อมูลตัวอย่าง (draw samples) ขึ้นมาจากการแจกแจงโพสทีเรียที่ได้ จากนั้นจึงทำการประมาณหาค่าคาดหวัง (Expectation) จากข้อมูลที่ถูกสร้างขึ้นมา วิธีการ MCMC นี้จะสร้างข้อมูลโดยทำการสร้างห่วงโซ่ข้อมูล (Markov chain) ด้วยระยะเวลาที่เพียงพอ ซึ่งข้อมูลที่ถูกสร้าง

ขึ้นในแต่ละครั้งจะขึ้นอยู่กับข้อมูลที่สร้างขึ้นมาก่อนหน้าตามหลักการของวิธีการมาร์คอฟเชน วิธีการที่ใช้ในการสร้างลำดับข้อมูลที่มีลักษณะเป็นห่วงโซ่ (chain) ของวิธีการ MCMC นี้มีด้วยกันหลายวิธี เช่น วิธีการกิบบ์แซมพลิง (Gibb sampling) และวิธีการเมโทรโพลิส – แฮสติงส์ (Metropolis – Hastings) เป็นต้น

จากที่ได้กล่าวมาแล้วข้างต้นว่า MCMC เกิดจากการรวมกันของวิธีการมอนติคาร์โลและมาร์คอฟเชน ดังนั้นในส่วนถัดไปจะเป็นการอธิบายถึงการทำงานของทั้งสองวิธีการนี้ก่อน แล้วจึงเข้าสู่การอธิบายถึงการทำงานของวิธีการ MCMC ซึ่งแบ่งออกเป็นหลายวิธีการด้วยกัน

2.7.1 วิธีการมาร์คอฟเชน

วิธีการมาร์คอฟเชนเป็นการสร้างลำดับของตัวแปรสุ่ม $\{X_0, X_1, X_2, ...\}$ ขึ้นมา โดยตัวแปรที่ได้จากการสุ่มจะขึ้นอยู่กับตัวแปรที่สุ่มมาได้ก่อนหน้าเพียงหนึ่งขั้น นั่นคือ ที่แต่ละเวลา $t \geq 0$ ตัวแปรสุ่มในสถานะถัดไป คือ X_{t+1} จะถูกสร้างมาจากการแจกแจง $P(X_{t+1} | X_t)$ ซึ่งขึ้นอยู่กับค่าของตัวแปรในสถานะ X_t ที่ถูกสร้างขึ้นมาก่อนหน้าเท่านั้น ย่อมหมายความว่าเมื่อมีลำดับห่วงโซ่ของตัวแปรสุ่ม X_t ตัวแปรสุ่มในสถานะถัดไปของการสุ่ม X_{t+1} จะไม่ขึ้นอยู่กับตัวแปรสุ่มที่อยู่ด้านหน้าของลำดับตัวแปรสุ่ม $\{X_0, X_1, ..., X_{t-1}\}$ ซึ่งลำดับของตัวแปรที่ได้จากการสุ่มซึ่งมีลักษณะเป็นแบบห่วงโซ่นี้ จะเรียกว่า "มาร์คอฟเชน" (Markov Chian) และ $P(.|.)$ จะเรียกว่าเป็นทรานสิชันเคอร์-เนล (transition kernel) ของลำดับของห่วงโซ่ที่ไม่ขึ้นอยู่กับเวลา t

อาจจะมีคำถามที่ว่า ตัวแปรสุ่มในสถานะเริ่มต้น X_0 จะส่งผลกระทบต่อสถานะ X_t อย่างไร ซึ่งเป็นคำถามที่เกี่ยวข้องกับการแจกแจงของ X_t ภายใต้เงื่อนไขของ X_0 แทนด้วย $P^{(t)}(X_t | X_0)$ ทำการตรวจสอบโดยในระหว่างการสร้างลำดับของตัวแปรสุ่มจะไม่ให้เกิดตัวแปรสุ่มในช่วง $\{X_1, X_2, ..., X_{t+1}\}$ ดังนั้น X_t จะขึ้นอยู่กับ X_0 โดยตรง ซึ่งนำไปสู่สภาวะของความสม่ำเสมอ โดยห่วงโซ่ที่เกิดขึ้นจะลืมสถานะเริ่มต้นของมันและในท้ายที่สุด $P^{(t)}(.|X_t)$ จะลู่เข้าสู่ค่าที่ไม่เปลี่ยนแปลงหรือการแจกแจงที่คงที่ (Stationary distribution) แทนด้วยสัญลักษณ์ $\phi(.)$ โดยไม่ขึ้นอยู่กับเวลา t หรือ X_0 ดังนั้นหลังจากการสร้างห่วงโซ่ของตัวแปรสุ่มจนได้ความยาวของช่วงเบิร์น – อิน (burn- in) ที่เพียงพอแล้ว โดยข้อมูลในช่วงเบิร์น - อินนี้เป็นช่วงที่ข้อมูลที่ถูกสร้างขึ้นมายังมีการเปลี่ยนแปลงอยู่ กำหนดให้เป็นช่วงที่มีความยาว m ครั้ง นั่นคือ $\{X_t; t = 0, ..., m\}$ จากนั้นเราสามารถใช้ลำดับของตัวแปรสุ่มที่สร้างมาได้ในการประมาณค่าหาค่าคาดหวังที่ต้องการทราบ $E[f(X)]$ โดยจะตัดส่วนของตัวแปรสุ่มที่อยู่ในช่วงเบิร์น – อิน ออกไม่นำมาพิจารณาร่วมด้วย

2.7.2 วิธีการมอนติคาร์โล

มอนติคาร์โลเป็นวิธีการที่ต้องการประมาณหาค่าคาดหวัง โดยการสร้างข้อมูลตัวอย่าง $\{X_t, t=1,...,n\}$ ขึ้นมาจากการแจกแจงหนึ่ง ๆ ซึ่งในที่นี้ให้เป็นการแจกแจง $\pi(.)$ แล้วจึงทำการประมาณค่าเพื่อหาค่าคาดหวังจากข้อมูลที่สร้างขึ้นมา โดยแสดงค่าคาดหวังได้ดังสมการที่ (2)

$$\mathrm{E}[f(X)] \approx \frac{1}{n}\sum_{t=1}^{n} f(X_t) \qquad (2)$$

เนื่องจากข้อมูลตัวอย่างที่สร้างขึ้นมานั้นจะสร้างมาจากข้อมูลที่เราสนใจที่มีอยู่ ดังนั้นข้อมูลตัวอย่างที่ได้จากการสร้างนี้จึงถือว่าเป็นตัวแทนของข้อมูลของประชากรของข้อมูลที่เราสนใจ ดังนั้นในการหาค่าเฉลี่ยของประชากรของ $f(X)$ ก็จะถูกประมาณค่ามาจากค่าเฉลี่ยของข้อมูลที่สร้างขึ้นมา เมื่อข้อมูลที่สร้างขึ้นมา $\{X_t\}$ มีความเป็นอิสระต่อกัน และสร้างออกมาเป็นจำนวนมากพอก็จะทำให้มั่นใจได้ว่าการประมาณค่าคาดหวังที่ต้องการทราบนี้มีความถูกต้อง ซึ่งความถูกต้องของการประมาณค่าคาดหวังนี้จะแปรผันตามขนาดของข้อมูลที่ถูกสร้างขึ้นมา

(Gilks el al., 1996)

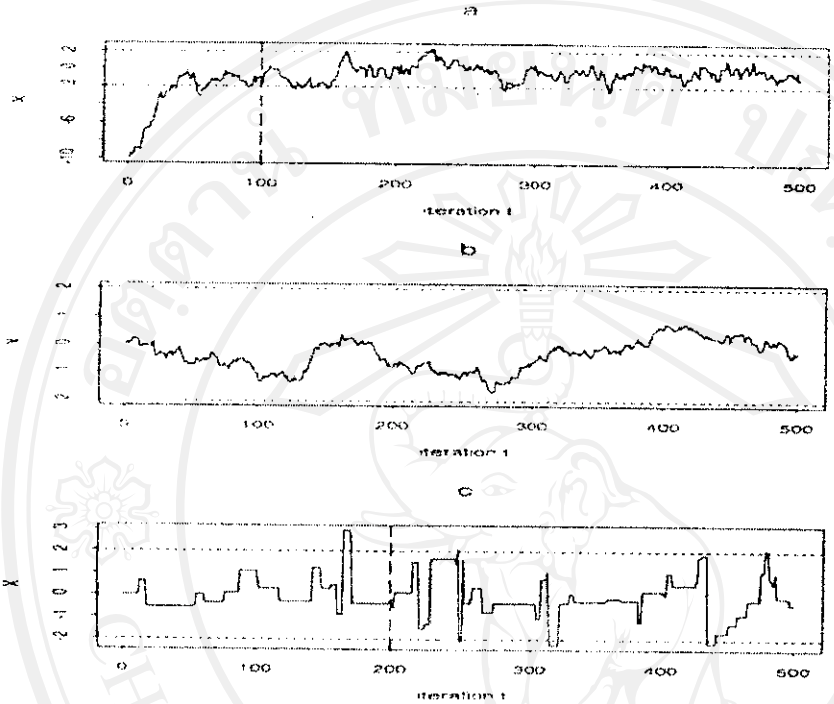
2.8 อัลกอริทึมเมโทรโพลิส – แฮสติงส์ (Metropolis – Hastings Algorithm)

อัลกอริทึมเมโทรโพลิส – แฮสติงส์เป็นวิธีการหนึ่งของมาร์คอฟเชนมอนติคาร์โล ที่ใช้สร้างห่วงโซ่ลำดับของตัวแปรสุ่มขึ้นมาซึ่งมีลักษณะเป็นแบบมาร์คอฟเชน โดยที่แต่ละเวลา t สถานะถัดไปของลำดับตัวแปรสุ่ม X_{t+1} จะถูกเลือกมาจากแคนดิเดทพอยท์ (Candidate point) Y ซึ่งเป็นค่าที่สุ่มได้จากการแจกแจงที่นำเสนอ (proposal distribution) $q(.|X_t)$ ที่ขึ้นอยู่กับค่าของสถานะ X_t ตัวอย่างเช่น ให้การแจกแจงที่นำเสนอ $q(.|X)$ เป็นการแจกแจงปกติแบบหลายตัวแปร (Multivariate normal distribution) ที่มีค่าเฉลี่ย (mean) เท่ากับ X และกำหนดค่าความแปรปรวนร่วม (covariance) ที่คงที่ แสดงดังรูป 2.3 ซึ่งเป็นการสร้างห่วงโซ่ของข้อมูลชุดเดียวกันโดยมีการแจกแจงที่นำเสนอสำหรับการแจกแจงในแต่ละครั้งที่ที่แตกต่างกันไป เป็นต้น

ในการพิจารณาว่าจะยอมรับค่าแคนดิเดทพอยท์ Y ที่ทำการสุ่มได้ ให้เข้าเป็นสมาชิกของห่วงโซ่ลำดับของตัวแปรสุ่มหรือไม่ จะพิจารณาจากค่าความน่าจะเป็นที่ยอมรับได้ (Acceptance probability) $\alpha(X,Y)$ ซึ่งคำนวณได้จากสมการที่ (3)

$$\alpha(X,Y) = \min\left(1, \frac{\pi(Y)q(X|Y)}{\pi(X)q(Y|X)}\right) \qquad (3)$$

โดยที่ การแจกแจงที่นำเสนอ $q(.|.)$ สามารถอยู่ในรูปแบบใดก็ได้ และการแจกแจงที่คงที่ของห่วงโซ่ข้อมูลก็จะเป็นการแจกแจง $\pi(.)$ ซึ่งเราคาดหวังว่าจะเป็นการแจกแจงของประชากรที่สร้างขึ้นมาได้



รูป 2.3 ห่วงโซ่ข้อมูลที่สร้างขึ้นจากทำงานของอัลกอริทึมเมโทรโพลิส – แฮสติงส์ จำนวน 500 รอบ โดยมีการแจกแจงที่คงที่เป็น N(0,1) และการแจกแจงที่นำเสนอ (ก) $q(.|X) = \mathrm{N}(X, 0.5)$; (ข) $q(.|X) = \mathrm{N}(X, 0.1)$; และ (ค) $q(.|X) = \mathrm{N}(X, 10.0)$ โดยข้อมูลที่อยู่ในช่วงเบิร์น – อิน จะอยู่ทางด้านซ้ายของเส้นประในแนวตั้ง

(แหล่งที่มา Gilks el al., 1996)

แสดงตัวอย่างของห่วงโซ่ข้อมูลที่ถูกสร้างขึ้นมาด้วยวิธีการของอัลกอริทึมเมโทรโพลิส – แฮสติงส์ได้ดังรูป 2.3 ซึ่งข้อมูลที่เราสนใจ คือ ชุดข้อมูล X ซึ่งมีการแจกแจงที่คงที่หรือการแจกแจงของชุดข้อมูลนี้เป็นการแจกแจงปกติที่มีค่าเฉลี่ย (mean) เป็นค่าเฉลี่ยของข้อมูล X และมีส่วนเบี่ยงเบนมาตรฐาน (standard deviation) เป็น 1 แทนด้วยสัญลักษณ์ N(0,1) โดยกำหนดการแจกที่นำเสนอที่แตกต่างกันสำหรับการสร้างข้อมูลในแต่ละครั้ง โดย (ก) เป็นการแจกแจงแบบปกติที่มีค่าเฉลี่ยเป็น X และส่วนเบี่ยงเบนมาตรฐานเป็น 0.5 แทนด้วยสัญลักษณ์ $\mathrm{N}(X, 0.5)$ (ข) เป็นการแจกแจงแบบปกติที่มีค่าเฉลี่ยเป็น X และส่วนเบี่ยงเบนมาตรฐานเป็น 0.1 แทนด้วยสัญลักษณ์ $\mathrm{N}(X, 0.1)$ และ (ค) เป็นการแจกแจงแบบปกติที่มีค่าเฉลี่ยเป็น X และส่วนเบี่ยงเบนมาตรฐานเป็น

10 แทนด้วยสัญลักษณ์ N(X,10.0) พบว่าถึงแม้จะเริ่มต้นสร้างข้อมูลจากการแจกแจงที่นำนำเสนอที่แตกต่างกัน ท้ายที่สุดแล้วข้อมูลที่ถูกสร้างขึ้นมาจากการแจกแจงที่นำเสนอทั้งสามกรณีนี้จะค่อย ๆ เปลี่ยนแปลงจนลู่เข้าสู่การแจกแจงปกติ N(0,1) ซึ่งเป็นการแจกแจงของข้อมูลชุดนี้

ในการพิจารณายอมรับข้อมูลที่ถูกสร้างขึ้นมานั้น ถ้ายอมรับแคนดิเดทพอยท์ Y จะทำให้สถานะถัดไปของลำดับของตัวแปรสุ่ม $X_{t+1} = Y$ แต่ถ้าไม่ยอมรับแคนดิเดทพอยท์ Y ลำดับถัดไปของห่วงโซ่ข้อมูล ก็จะไม่เปลี่ยนแปลง นั่นคือ $X_{t+1} = X_t$ ซึ่งสามารถแสดงอัลกอลิทึมเมโทรโพลิส – แฮสติงส์ ได้ดังต่อไปนี้

Initialize X_0 ; set $t = 0$
Repeat {
Sample a point from $q(.|X_t)$: Y
Sample a Uniform(0,1) random variable : U
If $U \le \alpha(X_t, Y)$ then
Set $X_{t+1} = Y$
Otherwise set $X_{t+1} = X_t$
Increment t }

(Gilks el al., 1996)

2.9 ซิงเกิล – คอมโพเน็นท์ เมโทรโพลิส – แฮสติงส์ (Single – component Metropolis – Hastings)

แทนที่จะทำการปรับปรุง (Updating) ห่วงโซ่ข้อมูลทั้งหมด n บล็อก (blocks) วิธีการนี้จะทำการแบ่งข้อมูล X ออกเป็นส่วน ๆ ที่แตกต่างกัน ทั้งหมด h ส่วน คือ $\{X_{.1}, X_{.2}, ..., X_{.h}\}$ แล้วจึงทำการปรับปรุงข้อมูลทีละส่วน โดยกำหนดให้ $X_{.-i}$ ประกอบด้วยส่วนต่าง ๆ ทั้งหมดของ X ที่ถูกแบ่ง ยกเว้นส่วนที่ $X_{.i}$ นั่นคือ $X_{.-i} = \{X_{.1}, ..., X_{.i-1}, X_{.i+1}, ..., X_{.h}\}$

ในแต่ละรอบของการทำงานแบบวนซ้ำของวิธีการซิงเกิล – คอมโพเน็นท์ เมโทรโพลิส – แฮสติงส์ จะทำการปรังปรุงเพื่อสร้างข้อมูลใหม่ทั้งหมด h ขั้นตอน ดังนี้ ให้ $X_{t,i}$ แสดงถึงสถานะของข้อมูลในส่วนที่ $X_{.i}$ ที่จุดสุดท้ายของการทำซ้ำรอบที่ t สำหรับการปรับปรุงในขั้นตอนที่ i ของการทำซ้ำรอบที่ $t + 1$ ข้อมูลในส่วน $X_{.i}$ จะถูกปรับปรุงด้วยวิธีการของเมโทรโพลิส – แฮสติงส์ ซึ่งแคนดิเดทพอยท์ $Y_{.i}$ ซึ่งถูกสร้างมาจากการแจกแจงที่นำเสนอ $q_i(Y_{.i} | X_{t,i}, X_{t,-i})$ โดยที่ $X_{t,-i}$ หมายถึงค่าของส่วนที่ $X_{.-i}$ หลังจากทำการปรับปรุงค่าในขั้นตอนที่ 1 ถึง i-1 ของการทำซ้ำในรอบที่ $t + 1$ เรียบร้อยแล้ว นั่นคือ

$$X_{t,-i} = \{X_{t+1,1}, ..., X_{t+1,i-1}, X_{t,i+1}, ..., X_{t,h}\}$$

โดยที่ข้อมูลในส่วนที่ 1, 2,..., i – 1 ต้องถูกปรับปรุงเรียบร้อยแล้ว ดังนั้นการแจกแจงที่นำเสนอ $q_i(.|.,.)$ ในลำดับที่ i^{th} ก็จะสร้างแคนดิเดทพอยท์ Y_i สำหรับข้อมูลในส่วนที่ i^{th} เท่านั้น และอาจจะขึ้นอยู่กับค่า ณ ขณะนั้น (current values) ของข้อมูลในส่วนนั้น ๆ ของข้อมูลทั้งหมด X ซึ่งค่าแคนดิเดทพอยท์นี้จะถูกพิจารณายอมรับด้วยความน่าจะเป็นที่ยอมรับได้ของวิธีการซิงเกิล – คอมโพเน็นท์ เมโทรโพลิส – แฮสติงส์ $\alpha(X_{-i}, X_i, Y_i)$ ดังสมการที่ (4)

$$\alpha(X_{-i}, X_i, Y_i) = \min\left(1, \frac{\pi(Y_i|X_{-i})q_i(X_i|Y_i, X_{-i})}{\pi(X_i|X_{-i})q_i(Y_i|X_i, X_{-i})}\right) \quad (4)$$

ถ้าแคนดิเดทพอยท์ Y_i ถูกยอมรับ จะกำหนดค่าให้กับสถานะถัดไปของข้อมูลส่วนที่ i เป็นค่าแคนดิเดทพอยท์ที่สร้างมาได้ นั่นคือ $X_{t+1,i} = Y_i$ แต่ถ้า Y_i ถูกปฏิเสธ จะกำหนดค่าให้ $X_{t+1,i} = X_{t,i}$ โดยส่วนอื่นของข้อมูลที่เหลือจะยังไม่ถูกทำการปรับปรุงที่ขั้นตอนที่ i นี้

ในที่นี้ $\pi(X_i|X_{-i})$ เป็นการแจกแจงที่มีเงื่อนไข (Full conditional distribution) สำหรับ X_i ภายใต้การแจกแจง $\pi(.)$ นั่นคือ เป็นการแจกแจงของข้อมูลส่วนที่ i ของ X ซึ่งอยู่ภายใต้เงื่อนไขของส่วนอื่น ๆ ที่เหลือ โดยข้อมูล X มีการแจกแจงเป็น $\pi(.)$ แสดงการแจกแจงที่มีเงื่อนไขดังสมการที่ (5)

$$\pi(X_i|X_{-i}) = \frac{\pi(X)}{\int \pi(X)dX_i} \quad (5)$$

จากการทำงานของวิธีการซิงเกิล – คอมโพเน็นท์ เมโทรโพลิส – แฮสติงส์ด้วยความน่าจะเป็นที่ยอมรับได้ในสมการที่ (4) ที่ทำการสร้างข้อมูลมาจากการแจกแจง $\pi(.)$ ซึ่งอันที่จริงแล้ว $\pi(.)$ ก็ได้รับอิทธิพลมาจากการแจกแจงที่มีเงื่อนไขของตัวมันเอง

(Gilks el al., 1996)

จากการทำงานของวิธีการซิงเกิล – คอมโพเน็นท์ เมโทรโพลิส – แฮสติงส์ ที่ได้กล่าวไว้ข้างต้น สามารถแสดงอัลกอริทึมของการทำงาน ได้ดังต่อไปนี้

- *Initialization*
  Select randomly or deterministically $X^{(0)} = (X_1^{(0)}, ..., X_h^{(0)})$
- *Iteration t* $(t \geq 1)$
  For $i = 1: h$
  Sample $Y_i \sim q_i(. \mid X_{1:i-1}^{(t)}, X_{i:h}^{(t-1)})$
  Sample $U \sim$ Uniform(0,1)

If $U \leq \alpha(X_{-i}, X_i, Y_i)$ then

Set $X_i^{(t)} = Y_i$

Otherwise $X_i^{(t)} = X_i^{(t-1)}$

End For

(Doucet and Wang, 2005)

2.10 กิบบ์ แซมพลิง (Gibbs Sampling)

วิธีการนี้ถือว่าเป็นกรณีพิเศษกรณีหนึ่งของวิธีการซิงเกิล – คอมโพเน็นท์ เมโทร-โพลิส – แฮสติงส์ ซึ่งในปัจจุบันนี้พบว่าเป็นวิธีที่ได้รับความนิยมนำไปประยุกต์ทางสถิติเป็นส่วนใหญ่ สำหรับวิธีการกิบบ์ แซมพลิงนี้ มีการแจกแจงที่ใช้นำเสนอเพื่อใช้ในการปรับปรุงข้อมูลในส่วนที่ i ของข้อมูล X แสดงได้ดังสมการที่ (6)

$$q_i(Y_i \mid X_i, X_{-i}) = \pi(Y_i \mid X_{-i}) \tag{6}$$

โดยที่ $\pi(Y_i \mid X_{-i})$ คือการแจกแจงที่มีเงื่อนไขในสมการที่ (5) เมื่อทำการหาค่าของความน่าจะเป็นที่ยอมรับได้โดยแทนค่าการแจกแจงที่นำเสนอในสมการที่ (6) นี้ ลงไปในสมการความน่าจะเป็นที่ยอมรับได้สมการที่ (4) พบว่าจะมีค่าเท่ากับ 1 นั่นคือ แคนดิเดทพอยท์ที่สร้างมาจากวิธีการกิบบ์ แซมพลิง จะถูกยอมรับเสมอ ดังนั้นด้วยวิธีการของกิบบ์ แซมพลิง จะทำให้ได้ห่วงโซ่ข้อมูลที่ประกอบไปด้วยข้อมูลที่สร้างมาจากการแจกแจงที่มีเงื่อนไข

(Gilks el al., 1996)

ข้อดีของวิธีการกิบบ์ แซมพลิง คือ ไม่ต้องทำการเลือกการแจกแจงที่นำเสนอ ทุกอย่างที่ใช้ในการคำนวณจะถูกกำหนดโดยการแจกแจงที่เป็นเป้าหมาย (Target distribution) หรือการแจกแจงที่มีเงื่อนไข π แต่อย่างไรก็ตาม ก็เป็นไปไม่ได้ที่จะสามารถทำการสร้างข้อมูลขึ้นมาจากการแจกแจงที่มีเงื่อนไขเสมอไป ถึงแม้ว่าจะเป็นการสร้างมาจากการแจกแจงที่มีเงื่อนไขเสมอแต่ก็อาจจะไม่ใช้ผลลัพธ์ที่ดีที่สุด

แสดงอัลกอริทึมของกิบบ์ แซมพลิง ได้ดังต่อนี้

- *Initialization*

  Select randomly or deterministically $X^{(0)} = (X_1^{(0)}, ..., X_h^{(0)})$

- *Iteration t* $(t \geq 1)$

  For $i = 1: h$

  Sample $Y_i^{(t)} \sim \pi(. \mid X_{1:i-1}^{(t)}, X_{i:h}^{(t-1)})$

  End For

(Doucet and Wang, 2005)

2.11 การสร้างแบบจำลอง

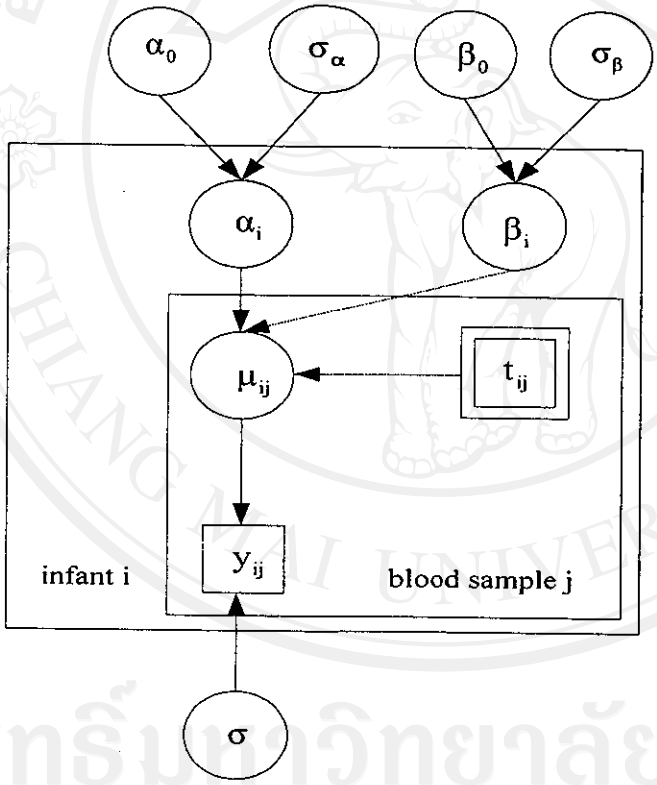
ในการสร้างแบบจำลอง (Modeling) เพื่อทำการอนุมานด้วยวิธีการกิบบ์ แซมพลิง มีองค์ประกอบหลักของการสร้างแบบจำลองความน่าจะเป็น (Full probability model) อยู่ด้วยกันทั้งหมด 3 องค์ประกอบ คือ องค์ประกอบแรกจะต้องทำการกำหนดลักษณะของแบบจำลองในเชิงปริมาณ คือ ขนาดหรือจำนวนโหนดในแบบจำลอง และในเชิงคุณภาพ คือ คุณลักษณะของโครงสร้างที่เป็นอิสระแบบมีเงื่อนไข (Conditional independence structure) ซึ่งเป็นแบบจำลองเชิงกราฟ (Graphical model) แสดงตัวอย่างได้ดังรูป 2.4 องค์ประกอบที่สอง ทำการกำหนดรูปแบบพารามิเตอร์ (parametric form) ของความสัมพันธ์ระหว่างโหนดต่าง ๆ โดยตรงซึ่งเป็นส่วนที่จะนำไปคำนวณเทอมไลค์ลิฮูดของแบบจำลอง และองค์ประกอบที่สาม เป็นการกำหนดการแจกแจงไพรเออร์ให้กับพารามิเตอร์ต่าง ๆ

จากขั้นตอนการสร้างแบบจำลองข้างต้น เพื่อให้เกิดความเข้าใจที่ง่ายขึ้นจึงได้แสดงตัวอย่างการสร้างแบบจำลองของข้อมูลเพื่อทำการอนุมาน ได้ดังแบบจำลองเชิงกราฟในรูป 2.4 ซึ่งเป็นแบบจำลองที่มาจากการเก็บรวบรวมข้อมูลเลือดของทารกและทำการวัดปริมาณของสารแอนติบอดี้ที่เรียกการวัดในครั้งนี้ว่า anti-HBs titre ที่ระยะเวลาต่าง ๆ ภายหลังจากที่ทารกได้รับวัคซีนป้องกันโรคไวรัสตับอักเสบบี (Hepatitis B) ครั้งสุดท้าย ซึ่งพบว่าความสัมพันธ์ระหว่างค่าลอการิทึมของปริมาณสารแอนติบอดี้ (log titre) และค่าลอการิทึมของเวลา (log time) เป็นความสัมพันธ์แบบเส้นตรง (Linear relationship) คือ $y = \alpha_i - 1 * \log t$ โดยที่ y คือ log anti-HBs titre และ α_i คือค่าคงที่หลังจากทารกคนที่ i^{th} ได้รับวัคซีนครั้งสุดท้าย

ในการกำหนดรูปแบบของพารามิเตอร์ ซึ่งเป็นการกำหนดการแจกแจงไพรเออร์หรือสมการความสัมพันธ์ให้กับพารามิเตอร์ต่างๆ ในแบบจำลอง เพื่อเป็นการแสดงความสัมพันธ์ระหว่างโหนดพ่อแม่กับโหนดลูก (Parent – Child) จากตัวอย่างได้กำหนดการแจกแจงไพรเออร์หรือสมการความสัมพันธ์ให้กับพารามิเตอร์ต่างๆ ซึ่งได้มาจากการวิเคราะห์ข้อมูลเบื้องต้นจากข้อมูลที่เราสนใจ ดังต่อไปนี้

$$
\begin{aligned}
y &\sim \mathrm{N}(\mu, \sigma^2) \\
\mu &= \alpha + \beta(\log t - \log 730) \\
\alpha &\sim \mathrm{N}(\alpha_0, \sigma_\alpha^2) \\
\beta &\sim \mathrm{N}(\beta_0, \sigma_\beta^2) \\
\alpha_0, \beta_0 &\sim \mathrm{N}(0, 10000) \\
\sigma^2, \sigma_\alpha^2, \sigma_\beta^2 &\sim Ga(0.01, 0.01)
\end{aligned}
$$

จากแบบจำลองเชิงกราฟในรูป 2.4 พารามิเตอร์ต่าง ๆ จะถูกแทนด้วยแต่ละโหนดในแบบจำลอง ซึ่งรูปแบบของโหนดที่ต่างกันจะบ่งบอกถึงความหมายที่แตกต่างกัน คือ โหนดที่แทนด้วยสี่เหลี่ยมสองชั้นแสดงถึงพารามิเตอร์ที่ถูกกำหนดค่าไว้ตายตัวในขั้นตอนการออกแบบ เช่น พารามิเตอร์เวลา t_{ij} โหนดที่เป็นสี่เหลี่ยมชั้นเดียวจะแสดงถึงข้อมูลที่เราสนใจ (observed data) และโหนดที่แทนด้วยวงกลมจะหมายถึงพารามิเตอร์ที่ไม่ทราบค่า (unknown parameters) และสำหรับเส้นเชื่อมแบบมีทิศทาง (directed links) นั้นจะแสดงถึงความสัมพันธ์ของพารามิเตอร์ต่างๆ ที่ขึ้นอยู่กับกันโดยตรง โดยเส้นเชื่อมที่เป็นเส้นทึบจะแสดงถึงการขึ้นอยู่กับกันที่ว่าด้วยความน่าจะเป็น (probabilistic dependencies) ในขณะที่เส้นเชื่อมที่เป็นเส้นประจะแสดงถึงความสัมพันธ์ระหว่างกันในลักษณะเป็นสมการความสัมพันธ์ (functional or deterministic relationship)



รูป 2.4 แบบจำลองเชิงกราฟสำหรับข้อมูล anti-HBs titre

จากรูป 2.4 กำหนดให้ y_{ij} แทนชุดข้อมูลที่เราสนใจ ในที่นี้ก็คือปริมาณของสารแอนติบอดี้ของตัวอย่างเลือดที่ j^{th} ในทารกคนที่ i^{th} ซึ่งมีสมมติฐานว่ามีลักษณะการแจกแจงเป็นแบบปกติ (Normal distribution) ที่ขึ้นอยู่กับพารามิเตอร์ค่าเฉลี่ย (mean) μ_{ij} และปัจจัยส่วนเบี่ยงเบนมาตรฐาน σ โดยที่ μ_{ij} เกิดจากสมการความสัมพันธ์แบบเส้นตรง (Linear regression model) ระหว่างค่าที่คาดหวังซึ่งเป็นค่าลอการิทึมของปริมาณสารแอนติบอดี้ (log titre) μ_{ij} กับค่าลอก-กา

ริทึมของเวลา (log time) t_{ij} สำหรับแต่ละตัวอย่างเลือดที่ j^{th} ในทารกคนที่ i^{th} ซึ่ง t_{ij} นี้เป็นพารามิเตอร์ที่ถูกกำหนดค่าไว้ตายตัวในการออกแบบ สำหรับพารามิเตอร์อินเตอร์เซพท์ (intercept) ของทารกคนที่ i แทนด้วย α_i และพารามิเตอร์แกรเดียนท์ (gradient) β_i ของทารกคนที่ i ซึ่งเป็นพารามิเตอร์ที่มาจากสมการความสัมพันธ์แบบเส้นตรงที่กำหนดให้แก่พารามิเตอร์ μ_{ij} โดยค่าของพารามิเตอร์ α_i และ β_i จะขึ้นอยู่กับค่าของซุปเปอร์พารามิเตอร์ (Super - parameters) อีกทีหนึ่ง นั่นคือ ค่าของ α_i จะได้มาจากการแจกแจงของตัวเองซึ่งเป็นการแจกแจงแบบปกติที่ขึ้นอยู่กับพารามิเตอร์ค่าเฉลี่ย α_0 และปัจจัยส่วนเบี่ยงเบนมาตรฐาน σ_α เช่นเดียวกันกับพารามิเตอร์ β_i ที่มีการแจกเป็นแบบปกติที่ขึ้นอยู่กับปัจจัยค่าเฉลี่ย β_0 และปัจจัยส่วนเบี่ยงเบนมาตรฐาน σ_β

ในแบบจำลองแบบขั้นลำดับดังเช่นตัวอย่างนี้ จะต้องหลีกเลี่ยงการกำหนดการแจกแจงไพรเออร์ที่ไม่เหมาะสมสำหรับพารามิเตอร์ต่าง ๆ เนื่องมาจากว่าการแจกแจงไพรเออร์ที่ไม่เหมาะสมอาจเป็นสาเหตุทำให้ได้การแจกแจงโพสทีเรียที่ไม่เหมาะสมจากการอนุมาน

หลังจากทำการกำหนดการแจกแจงเบื้องต้นให้กับปัจจัยต่าง ๆ และทำการสร้างแบบจำลองเชิงกราฟขึ้นมาเรียบร้อยแล้ว ต่อไปจึงเข้าสู่การคำนวณเพื่อหาความน่าจะเป็นของแบบจำลอง (Full probability model) นั่นคือ จากแบบจำลองที่ได้ออกแบบไว้ เรามีสมมติฐานว่าการแจกแจงร่วมของทุกพารามิเตอร์แบบสุ่มจะถูกกำหนดด้วยของการแจกแจงแบบมีเงื่อนไขของแต่ละโหนดเมื่อทราบโหนดพ่อแม่ของมัน (Conditional distribution) แสดงการแจกแจงของความน่าจะเป็น (Probability distribution) โดยที่ v คือโหนดในแบบจำลอง และ V คือ เซตของทุกโหนดที่อยู่ในแบบจำลอง ได้ดังสมการที่ (7)

$$P(V) = \prod_{v \in V} P(v \mid parents(v)) \tag{7}$$

(Gilks et al., 1996)

2.12 การปรับแบบจำลองให้เหมาะสมด้วยวิธีการกิบบ์ แซมพลิง (Fitting a model using Gibbs sampling)

โดยทั่วไปในการสร้างข้อมูลด้วยวิธีการของกิบบ์ แซมพลิง จะวนทำงานแบบทำซ้ำเพื่อสร้างข้อมูลทำให้เกิดความเหมาะสมกับแบบจำลองที่ได้ออกแบบไว้มากที่สุด ซึ่งวิธีการทำงานจะประกอบด้วย 4 ขั้นตอนหลัก ดังต่อไปนี้

(1) กำหนดค่าเริ่มต้น (Initialization) โดยกำหนดค่าเริ่มต้นและการแจกแจงไพรเออร์ให้กับทุกโหนดที่ไม่ทราบค่า นั่นคือ เป็นการกำหนดค่าเริ่มต้นให้กับพารามิเตอร์ต่าง ๆ ในแบบจำลอง

(2) สร้างข้อมูลจากการแจกแจงที่มีเงื่อนไข (Sampling from full conditional distributions) ด้วยวิธีการของกิบบ์ แซมพลิง จะทำงานโดยการวนทำซ้ำเพื่อสร้างข้อมูลจากการแจกแจงที่มีเงื่อนไขของโหนดที่ไม่ทราบค่าในแบบจำลองเชิงกราฟ ซึ่งการแจกแจงที่มีเงื่อนไขสำหรับโหนดหนึ่ง ๆ จะเป็นการแจกแจงของโหนดนั้นเมื่อทราบค่าของโหนดอื่น ๆ ที่เหลือที่อยู่ในแบบจำลอง ทำให้เราสามารถแบ่งโครงสร้างของการแจกแจงร่วมในสมการที่ (7) ได้เป็นการแจกแจงที่มีเงื่อนไขดังสมการที่ (8) ซึ่งในท้ายที่สุดแล้วก็จะกลายเป็นการแจกแจงโพสทีเรียที่เราต้องการทราบ ดังนี้

$$\begin{aligned} P(v|V_{-v}) &\propto P(v,V_{-v}) \\ &\propto \text{terms in } P(V) \text{ containing } v \\ &= P(v|parents[v]) * \prod_{w \in children[v]} P(w|parents[w]) \end{aligned} \quad (8)$$

จากสมการที่ (8) ให้ v แทนโหนดต่าง ๆ ในแบบจำลองเชิงกราฟ V_{-v} แทนโหนดอื่นๆ ที่เหลือยกเว้นโหนด v และ w แทนโหนดที่เป็นโหนดลูกของโหนด v ซึ่งในการแจกแจงที่มีเงื่อนไขสำหรับโหนด v จะประกอบไปด้วย 2 ส่วน คือ ส่วนไพรเออร์ $P(v|parents[v])$ ซึ่งพิจารณาจากโหนด v และโหนดพ่อแม่ของโหนด v และส่วนไลค์ลิฮูด (Likelihood) ที่ได้มาจากแต่ละโหนดลูกของโหนด v ดังนั้นการแจกแจงที่มีเงื่อนไขสำหรับแต่ละโหนดก็จะขึ้นอยู่กับค่าของโหนดพ่อแม่ (parents) โหนดลูก (children) และโหนดพ่อแม่ร่วม (co - parents) เท่านั้น โดยที่ โหนดพ่อแม่ร่วม คือโหนดพ่อแม่ของโหนดลูกของโหนด v โหนดอื่น ๆ ที่ไม่ใช่โหนด v

(3) ติดตามผลลัพธ์ (Monitoring the output) โดยทำการตรวจสอบค่าที่สร้างขึ้นมาจากวิธีการของกิบบ์ แซมพลิงว่ามีลักษณะของการ mixing เป็นอย่างไร พิจารณาตัดสินใจความยาวของช่วงเบิร์น – อิน และพิจารณาว่าค่าที่สร้างขึ้นมามีการลู่เข้าการแจกแจงแบบใด

(4) อนุมานจากผลลัพธ์ที่ได้ (Inference from the output) โดยการอนุมานเพื่อให้ทราบลักษณะข้อมูลของประชากรที่เราไม่ทราบค่า ซึ่งจะถูกนำมาทำการหาค่าทางสถิติที่เราสนใจ

(Gilks et al., 1996)

2.13 รีเวิร์สซิเบิลจัมพ์ มาร์คอฟเชนมอนติคาร์โล (Reversible Jump MCMC)

พิจารณากรณีที่เราต้องการสร้างข้อมูลแบบสุ่มจากการแจกแจง π ที่กำหนดอยู่บนเซตข้อมูลที่เกิดจากการยูเนียนกันของเซตข้อมูลย่อย กล่าวคือ $X=\cup_{k=0}^{\infty}\{k\}\times X_k$ ซึ่งสามารถเขียนการแจกแจงได้เป็น $\pi(k,x_k)$ โดยที่ $x_k \in X_k$ ในกรณีนี้ ตัวแปรแบบสุ่ม (random variables) ที่เกี่ยวข้องสามารถเป็นค่าในเซตข้อมูลย่อยที่มีขนาด (dimensions) ต่างกันได้ กรณีเช่นนี้จะเกิดขึ้นเมื่อเราต้องการแก้ปัญหาของการอนุมาน โดยปัญหาคือเราไม่ทราบจำนวนของตัวที่ไม่ทราบค่า (unknown) ตัวอย่างเช่น ปัญหาของการวิเคราะห์สเปกตรัม (Spectral analysis problem) ในกรณีที่เราไม่ทราบจำนวนของซินูซอยด์ (sinusoids) k และพารามิเตอร์ต่าง ๆ ที่ไม่ทราบค่าซึ่งถูกกำหนดอยู่บนเซตข้อมูล โดยขนาดของเซตข้อมูลนั้นขึ้นอยู่กับตัวแปร k

สำหรับปัญหาดังกล่าว จะไม่สามารถเป็นไปได้เลยที่จะประยุกต์ใช้อัลกอริทึมเมโทรโพลิส– แฮสติงส์ สำหรับการคำนวณที่มีการกระโดดจาก X_k ไปสู่ X_l เมื่อ $\dim(X_k) \neq \dim(X_l)$ ทางออกของการแก้ปัญหานี้ได้ถูกนำเสนอโดย Green (1995) ซึ่งได้เสนอวิธีการที่เรียกว่า "รีเวิร์สซิเบิลจัมพ์ เอ็มซีเอ็มซี" (Reversible Jump MCMC) โดยเป็นวิธีที่อยู่บนพื้นฐานของข้อจำกัดของการย้อนกลับของการเคลื่อนที่ระหว่างเซต $\{X_k\}$ ที่แตกต่างกัน สำหรับการกระโดดจาก X_1 ไปยัง X_2 ซึ่งจะทำการสุ่มตัวแปร $u_1 : q_{1\to 2}(.)$ และกำหนดให้

$$(x_2,u_2)=\varphi_{1\to 2}(x_1,u_1) \tag{9}$$

โดยที่ $\varphi_{1\to 2}$ คือฟังก์ชัน (Deterministic mapping) แบบหนึ่งต่อหนึ่ง (one-to-one) นั่นคือเวกเตอร์ (x_1,u_1) และ (x_2,u_2) มีขนาดเท่ากัน หมายเหตุ ถ้า $\dim(X_1) > \dim(X_2)$ โดยทั่วไปจะไม่มีการกำหนดเป็นตัวแปร u_1 ในแนวทางเดียวกัน สำหรับการกระโดดจาก X_2 ไปยัง X_1 ก็จะทำการสุ่มตัวแปร $u_2 : q_{2\to 1}(.)$ และกำหนดให้

$$(x_1,u_1)=\varphi_{2\to 1}(x_2,u_{21}) \tag{10}$$

โดยที่ $\varphi_{2\to 1}(\varphi_{1\to 2}(x_1,u_1))=(x_1,u_1)$

ด้วยวิธีการนี้ ค่าความน่าจะเป็นที่ยอมรับได้สำหรับการเคลื่อนที่จาก X_1 ไปยัง X_2 สามารถแสดงได้ดังสมการที่ (11)

$$\min\left(1,\frac{\pi(2,x_2)}{\pi(1,x_1)}\frac{p_{2\to 1}q_{2\to 1}(u_2)}{p_{1\to 2}q_{1\to 2}(u_1)}\left|\frac{\partial\varphi_{1\to 2}(x_1,u_1)}{\partial(x_1,u_1)}\right|\right) \tag{11}$$

เมื่อ $\partial\varphi_{1\to 2}(x_1,u_1)/\partial(x_1,u_1)$ คือ ดีเทอร์มิเน็นท์ ของจาโคเบียน (Determinant of the Jacobian) ของการเปลี่ยนแปลงจาก $p_{i\to j}$ ซึ่งเป็นความน่าจะเป็นของการเลือกเพื่อพยายามกระโดดจาก X_i ไปยัง X_l และ $q_{1\to 2}(.)$ คือความหนาแน่น (density) ของ u_1

จากขั้นตอนดังกล่าวข้างต้น สามารถสรุปเป็นอัลกอริทึมของรีเวิร์สซิเบิลจัมพ์ ได้ดังต่อไปนี้

- *Initialization*

  Select randomly or deterministically $(k^{(0)},x^{(0)})$

- *Iteration t* $(t\geq 1)$

  Assume we have $x^{(t-1)}=(k^{(t-1)},x_{k^{(t-1)}}^{(t-1)})$

  Propose a move from $X_{k^{(t-1)}}$ to X_l with probability $p_{k^{(t-1)}\to l}$

  Sample $u_{k^{(t-1)}}\sim q_{k^{(t-1)}\to l}(.)$

  Set $(x_l^*,u_l)=\varphi_{k^{(t-1)}\to l}(x_{k^{(t-1)}}^{(t-1)},u_{k^{(t-1)}})$

  With probability

$$\min\left(1,\frac{\pi(l,x_l^*)}{\pi(k^{(t-1)},x_{k^{(t-1)}}^{(t-1)})}\frac{p_{l\to k^{(t-1)}}q_{l\to k^{(t-1)}}(u_l)}{p_{k^{(t-1)}\to l}q_{k^{(t-1)}\to l}(u_{k^{(t-1)}})}\left|\frac{\partial\varphi_{k^{(t-1)}\to l}(x_{k^{(t-1)}}^{(t-1)},u_{k^{(t-1)}})}{\partial(x_{k^{(t-1)}}^{(t-1)},u_{k^{(t-1)}})}\right|\right)$$

  set $x^{(t)}=(l,x_l^*)$; otherwise $x^{(t)}=x^{(t-1)}$

(Doucet and Wang, 2005)

2.14 โปรแกรมวินบักส์ (WinBUGS Program)

โปรแกรมวินบักส์ หรือโปรแกรมการอนุมานด้วยเบย์เซียนโดยอาศัยวิธีการกิบบ์แซมพลิงสำหรับวินโดวส์ (Bayesian inference Using Gibbs Sampling for Windows) เป็นซอฟแวร์ (Software) สำหรับการวิเคราะห์แบบจำลองทางสถิติที่มีความซับซ้อนด้วยวิธีการเบย์เซียนโดยอาศัยวิธีการกิบบ์แซมพลิงซึ่งเป็นวิธีการหนึ่งของมาร์คอฟเชนมอนติคาร์โล โปรแกรมนี้ได้ถูกพัฒนาขึ้นตั้งแต่ปี 1989 โดย MRC Biostatistics Unit ที่ถูกพัฒนาขึ้นด้วยภาษาบักส์ (Bugs language) การทำงานของโปรแกรมวินบักส์เป็นการทำงานเชิงโต้ตอบกับวินโดวส์ (Interactive windows) โดยต้องมีการออกแบบแบบจำลองแสดงความสัมพันธ์ระหว่างโหนดหรือพารามิเตอร์ต่างๆ ที่อยู่ในรูปแบบของแบบจำลองเชิงกราฟ หรือเป็นการบรรยายแบบจำลองด้วยข้อความ (Text – based discription) ของโปรแกรมภาษาบักส์ จากนั้นต้องมีการกำหนดค่าเริ่มต้นและการแจกแจงไพรเออร์ให้กับโหนดต่างๆ แล้วจึงทำการรันโปรแกรมวินบักส์ด้วยจำนวนรอบการทดลองซ้ำที่เพียงพอ เพื่อทำการประมาณค่าของพารามิเตอร์ต่าง ๆ ที่เราสนใจ ซึ่งวิธีการทำงานของโปรแกรม

วินบักส์นี้มีขั้นตอนการทำงานดังที่ได้อธิบายไว้ในหัวข้อที่ 2.12 การปรับแบบจำลองให้เหมาะสมด้วยวิธีการกิบบ์ แซมพลิง

(MRC Biostatistics Unit, 2005[online])

วิธีการที่ใช้ในการสร้างลำดับข้อมูลที่มีลักษณะเป็นห่วงโซ่ของวิธีการมาร์คอฟเชนมอนติ-คาร์โล ที่จะถูกนำมาใช้เป็นวิธีการในการวิเคราะห์ข้อมูลการแสดงออกของยีนในงานวิจัยนี้ มีอยู่ด้วยกัน 2 วิธีการ คือ วิธีการกิบบ์แซมพลิง (Gibb sampling) และวิธีการเมโทรโพลิส-แฮสติงส์ (Metropolis – Hastings) เพื่อทำการสร้างแบบจำลองของข้อมูลไมโครอาร์เรย์ ซึ่งจะได้กล่าวถึงต่อไปในบทที่ 3 และ 4 โดยการทดลองในบทที่ 3 นั้นจะเป็นการใช้โปรแกรมวินบักส์เป็นเครื่องมือช่วยในการวิเคราะห์ข้อมูล ส่วนการทดลองในบทที่ 4 ผู้เขียนได้ทำการทดลองโดยการเขียนโปรแกรมด้วยภาษาอาร์เอง