

บทที่ 2

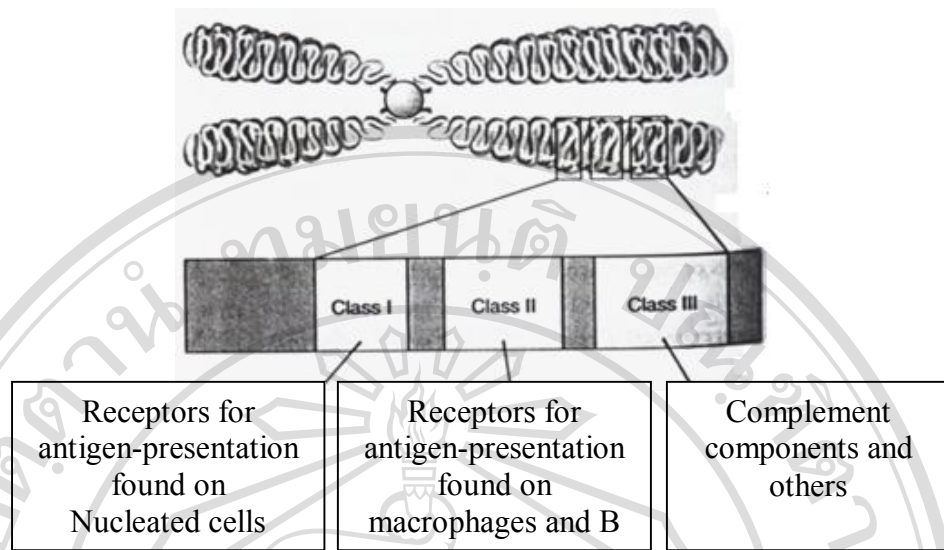
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในขั้นตอนของการศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้องกับการจำแนกชนิดข้อมูลสายลำดับ โดยวิธีซัพพอร์จเวกเตอร์แมชชีน ได้แบ่งเนื้อหาออกเป็น 6 ส่วน โดยมีรายละเอียดดังต่อไปนี้

- เอ็มเอชซี (Major Histocompatibility Complex)
- ข้อมูลรหัสพันธุกรรมของดีเอ็นเอ (DNA Data)
- การถอดรหัสพันธุกรรมของดีเอ็นเอ (DNA Transcription)
- โคดอนยูสเอจ (Codon Usage)
- ได-โคดอนยูสเอจ (Di-codon Usage)
- การเรียนรู้ด้วยเครื่อง (Machine Learning)

2.1 เอ็มเอชซี (MHC)

เอ็มเอชซี คือบริเวณที่เป็นที่ตั้งของกลุ่มยีนบนโครโมโซมที่ 6 ที่ถูกค้นพบครั้งแรกจากการทดลองในหนู ว่าเป็นมีความสำคัญมาก ในการกำหนดการตอบรับหรือการปฏิเสธเนื้อเยื่อ (tissue rejection) โดยพบว่า เกิดจากคุณสมบัติของเอ็มเอชซี แอนติเจน ซึ่งเป็นโมเลกุลสำคัญ ที่สร้างจากยีนบนเอ็มเอชซีโมเลกุลเหล่านี้ โดยจะแสดงออกบนพื้นผิวเซลล์ (cell surface molecule) และมีรูปแบบแตกต่างกันมาก (Highly polymorphic) ซึ่งในหนูหรือมนุษย์ที่มีเอ็มเอชซีที่ต่างกัน จะไม่สามารถเปลี่ยนถ่ายเนื้อเยื่อแก่กันได้ โดยเราเรียกเอ็มเอชซี ในหนูว่า เอช-2 (H-2) ส่วนเอ็มเอชซี ในมนุษย์นั้นเรียกว่า ฮิวแมน ลิวโคไซต์ แอนติเจน (เอชแอลเอ) เนื่องจากโมเลกุลจำพวกนี้ถูกค้นพบครั้งแรกบนผิวของเม็ดเลือดขาวในมนุษย์ (Margulies et al., 1999) โดยการศึกษาต่อมาพบว่า นอกจากคุณสมบัติที่เกี่ยวกับการปฏิเสธเนื้อเยื่อ แล้วเอ็มเอชซียังมีความสำคัญมาก ในการควบคุมปฏิกิริยาตอบสนองของภูมิคุ้มกัน โดยได้ชื่อว่าเป็น ตัวกระตุ้นระบบภูมิคุ้มกัน (Immune response gene: Ir) ซึ่งภายในโมเลกุลของเอชแอลเอจะประกอบด้วยโมเลกุล 3 กลุ่ม (Class) คือเอชแอลเอกลุ่มที่ I, II และ III ดังแสดงในรูป 2.1



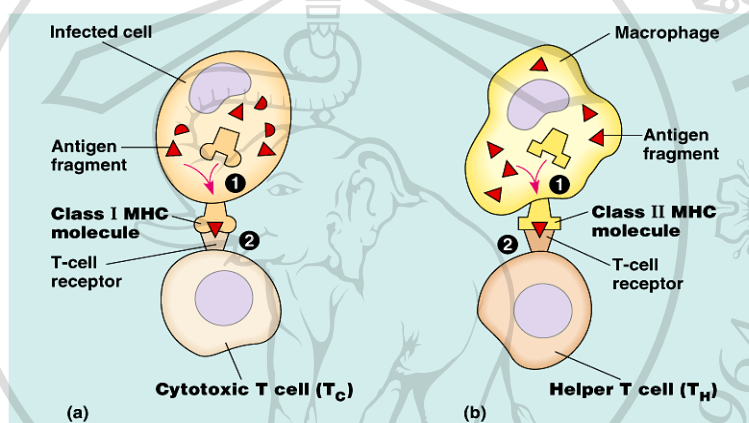
รูป 2.1 เอชแอลเอ

เอชแอลเอกลุ่มที่ I และเอชแอลเอกลุ่มที่ II คือกลุ่มของรหัสโปรตีนที่แสดงแอนติเจนของมนุษย์บนผิวเซลล์ต่างๆ ในร่างกาย เพื่อให้ที-เซลล์ (T-cells) สามารถแยกความแตกต่างและจดจำแอนติเจนแปลกปลอมด้วยการแสดงออกของโปรตีนเอชแอลเอบนผิวเซลล์ปกติในร่างกาย ซึ่งจะทำให้ที-เซลล์ สามารถตรวจพบสิ่งแปลกปลอมและทำการกระตุ้นระบบภูมิคุ้มกันต่อไป ส่วนเอชแอลเอกลุ่มที่ III นั้นคือกลุ่มของรหัสโปรตีนหลายชนิดที่ ซึ่งส่วนใหญ่เกิดขึ้นในปฏิกิริยาการเกิดภาวะอักเสบ (Inflammation response) รวมทั้งโปรตีนคอมพลีเมนต์ต่างๆ ซึ่งเป็นระบบที่ประกอบด้วยการทำงานอย่างต่อเนื่องของโปรตีนหลายชนิด เพื่อช่วยแอนติบอดีในการทำลายแบคทีเรียและรวมทั้งทูเมอร์ นิโครซิส แฟรกเตอร์ (ทีเอ็นเอฟ) คือ ตัวกระตุ้นให้มีการสร้างเซลล์เม็ดเลือดขาวออกมามากขึ้น เพื่อให้เกิดการตอบสนองต่อระบบภูมิคุ้มกันแบบจำเพาะ หรือถ้าจุลชีพนั้นถูกทำลายหมดก็จะกระตุ้นให้เกิดการซ่อมแซมเนื้อเยื่อที่ถูกทำลายต่อไป

ส่วนความแตกต่างของยีนเอชแอลเอกลุ่มที่ I และ II ในมนุษย์แต่ละคนนั้นจะมีความแตกต่างหรือคล้ายคลึงกัน แต่จะไม่เหมือนกันทั้งหมด (Polymorphism) ยกเว้นในกรณีที่เป็นฝาแฝดเกิดจากไข่ใบเดียวกัน (Identical twins) จะมีความเข้ากัน (Matching) ของยีน ซึ่งเอชแอลเอมีความสำคัญมากในการปลูกถ่ายเนื้อเยื่อหรืออวัยวะ (Transplantation) ซึ่งหากมีความเข้ากันของยีนหรือมีความคล้ายคลึงกันมากจะทำให้มีโอกาสเกิดการสัด (Rejection) เนื้อเยื่อหรืออวัยวะได้น้อยลง

2.1.1 กระบวนการนำเสนอแอนติเจน (Antigen Processing and Presentation)

โดยปกติแล้วทุกเซลล์ในร่างกายสามารถแสดงถึงเอชแอลเอกลุ่มที่ I บนผิวเซลล์ได้ แต่ในเซลล์ภูมิคุ้มกันเช่น เซลล์ที่ทำลายจุลชีพหรือสิ่งแปลกปลอม (Macrophages) และบี-เซลล์ (B-cell) นั้นเป็นเซลล์ที่มีลักษณะพิเศษที่สามารถแสดงถึงเอชแอลเอกลุ่มที่ II ได้ ส่วนในการกระตุ้นที-เซลล์ จะมีเซลล์ที่เรียกว่าไซโตโตซิก (Cytotoxic) ซึ่งมีความสามารถในการจดจำแอนติเจนด้วยการจับกับเอชแอลเอกลุ่มที่ I ในขณะที่เซลล์ผู้ช่วย (Helper) นั้นจะมีความสามารถในการจดจำแอนติเจนด้วยการจับกับเอชแอลเอกลุ่มที่ II ซึ่งมีกระบวนการนำเสนอแอนติเจนที่แตกต่างกัน ดังแสดงในรูป 2.2



รูป 2.2 โครงสร้างของเอชแอลเอกลุ่มที่ I และเอชแอลเอกลุ่มที่ II ในการจับกับแอนติเจน พร้อมกับการนำเสนอเซลล์ไซโตโตซิกและเซลล์ผู้ช่วยในการกระตุ้นที-เซลล์

(ที่มา: http://www.sc.chula.ac.th/courseware/GenBio103/Immune-system16_7_3.ppt)

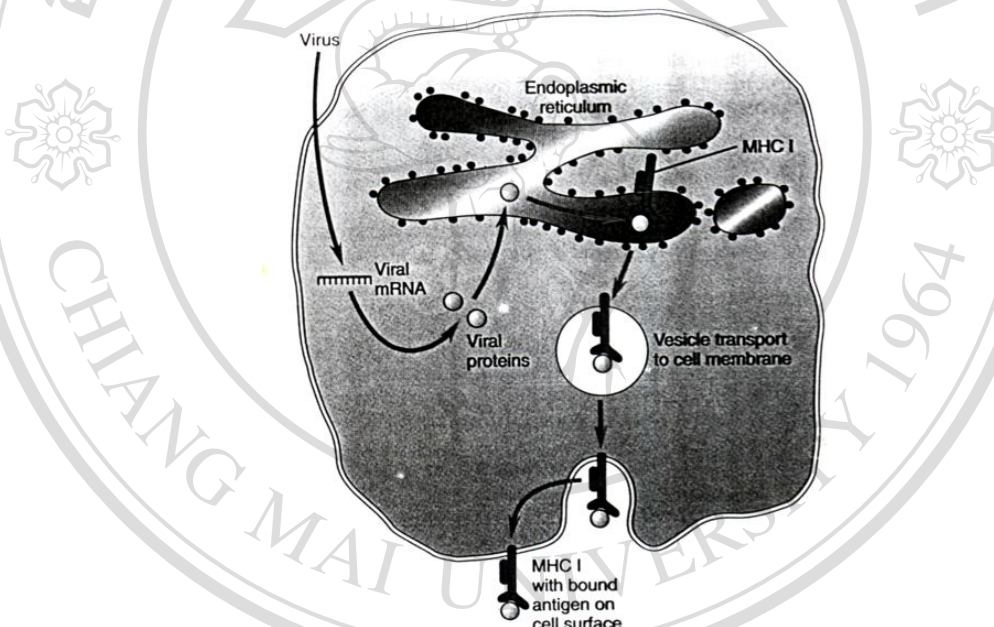
2.1.2 เอชแอลเอกลุ่มที่ I

เอชแอลเอกลุ่มที่ I พบได้ในทุกเซลล์ปกติของร่างกาย เมื่อได้สัมผัสกับแอนติเจนแปลกปลอมภายในเซลล์ร่างกาย (Intracellular sources) จะสามารถย่อยแอนติเจนและชิ้นส่วนที่จับแอนติเจนแปลกปลอมจากภายในเซลล์ออกมาให้แก่ไซโตโตซิก ที-เซลล์ (Cytotoxic T-cell) ซึ่งลักษณะการนำเสนอเช่นนี้ส่วนมากจะพบในการติดเชื้อไวรัส เนื่องจากรูปแบบการติดเชื้อภายในเซลล์นั้น จะทำให้สะดวกต่อการนำเสนอแอนติเจนอีกด้วย และนอกจากนี้โปรตีนในเซลล์ที่ผิดปกติจากเซลล์ร่างกายที่ผ่าเหล่า (Mutant cell) นั้นก็สามารถถูกนำเสนอในลักษณะเดียวกันนี้ได้ เพื่อให้เกิดการกระตุ้นภูมิคุ้มกันมาทำลายเซลล์นั้นต่อไป

ขั้นตอนในการนำเสนอแอนติเจนหลังจากที่เซลล์ได้มีการย่อยจนเป็นชิ้นเดียวกับแอนติเจนแปลกปลอมแล้ว ชิ้นส่วนของแอนติเจนแปลกปลอมนี้จะถูกส่งไปยังเอนโดพลาสมิก เรติคูลัม

(Endoplasmic reticulum) เพื่อจับกับโปรตีนเอชแอลเอกลุ่มที่ I ที่เพิ่งสร้างเสร็จจนกลายเป็น เอชแอลเอกลุ่มที่ I ต่อจากนั้นจึงจะเดินทางไปที่ยื่อหุ้มเซลล์และถูกยื่นออกไปบนผิวเซลล์เพื่อนำเสนอให้แก่ไซโตโตซิก ที-เซลล์มาจับ ทำให้เกิดการกระตุ้นไซโตโตซิก ที-เซลล์ให้หลั่งเอนไซม์และทำการโพลี-ฟอร์มมิง โปรตีน (Pore-forming proteins) ไปยังเซลล์เป้าหมายซึ่งก็คือเซลล์ที่นำเสนอแอนติเจน ซึ่งเป็นผลให้เซลล์นั้นเกิดการแตกตัวออกมา ดังแสดงในรูป 2.3

โดยไซโตโตซิก ที-เซลล์นั้นจะเป็นตัวที่สามารถจดจำหรือทำลายแอนติเจนแปลกปลอมได้จากการนำเสนอที่ผิวเซลล์ร่วมกับโมเลกุลของเอชแอลเอกลุ่มที่ I เท่านั้น โดยอาจเรียกไซโตโตซิก ที-เซลล์ว่าเป็นตัวร่วมยับยั้งของเอชแอลเอกลุ่มที่ I (HLA Class I restricted)



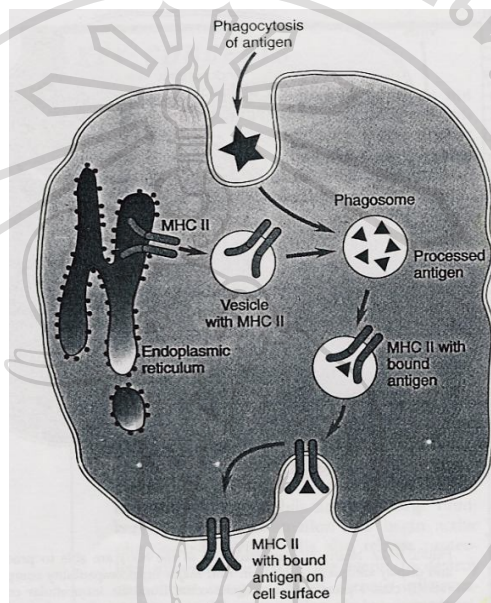
รูป 2.3 ขั้นตอนการนำเสนอแอนติเจนด้วยโมเลกุลของเอชแอลเอกลุ่มที่ I

(ที่มา: <http://cyberclass.msu.ac.th/>)

2.1.3 เอชแอลเอกลุ่มที่ II

เนื่องจากยีนรหัสโมเลกุลของเอชแอลเอกลุ่มที่ II พบเฉพาะในเซลล์เช่นโมโนไซต์ (Monocytes), มาโครฟาจ (Macrophages) และบี-เซลล์ (B-cells) ซึ่งเซลล์เหล่านี้จะสามารถนำเสนอแอนติเจนแปลกปลอมจากนอกเซลล์ได้ (Extracellular sources) โดยเริ่มต้นจากการกลืนกินแอนติเจนแปลกปลอมเข้าไปในเซลล์ก่อนที่จะถูกย่อยเป็นชิ้นภายในถุง (Phagosomes) โดยทั่วไปโมเลกุลของเอชแอลเอกลุ่มที่ II ซึ่งถูกสร้างจากเอนโดพลาสมิก เรติคูลัม จะเคลื่อนมา

จับกับชิ้นส่วนของแอนติเจนแปลกปลอมจนกลายเป็นแอนติเจนของเอชแอลเอกลุ่มที่ II จากนั้นจะเดินทางไปยังเยื่อหุ้มเซลล์เพื่อยื่นออกไปบนผิวเซลล์และนำเสนอให้แก่เซลล์ผู้ช่วยที-เซลล์ (Helper T-cells) ต่อไป และเซลล์ผู้ช่วยที-เซลล์เท่านั้น ที่สามารถจดจำหรือแยกแอนติเจนแปลกปลอมได้ ด้วยการนำเสนอจากโมเลกุลเอชแอลเอกลุ่มที่ II ซึ่งอาจเรียก ผู้ช่วยที-เซลล์ ว่าเป็นตัวร่วมยับยั้งของเอชแอลเอกลุ่มที่ II (HLA Class II restricted) ดังแสดงในรูป 2.4



รูป 2.4 ขั้นตอนการนำเสนอด้วยโมเลกุลโปรตีนเอชแอลเอกลุ่มที่ II

(ที่มา: <http://cyberclass.msu.ac.th/>)

ส่วนเซลล์มาโครฟาเกส มีความสามารถที่จะนำเสนอแอนติเจนแปลกปลอมที่มีลักษณะแตกต่างกันเป็นจำนวนมาก จึงมีบทบาทในการตอบสนองแบบไม่จำเพาะ (Non-specific)

แต่ส่วนในบี-เซลล์นั้นจะจับกับแอนติเจนแปลกปลอมอย่างจำเพาะที่ตัวรับบี-เซลล์ (B-cell Receptors) ก่อนที่จะส่งเข้าไปย่อยภายในเซลล์ บี-เซลล์แต่ละเซลล์ จะมีตัวรับบี-เซลล์เพียงชนิดเดียวซึ่งจะสามารถนำเสนอแอนติเจนอย่างจำเพาะเพียงชนิดเดียว จึงมีบทบาทในการตอบสนองแบบจำเพาะ (Specific)

(Abbas et al., 1997)

2.2 ข้อมูลรหัสพันธุกรรมของดีเอ็นเอ

เนื่องจากงานวิจัยที่ทำความข้องเกี่ยวกับงานทางด้านชีววิทยาอย่างมากโดยเฉพาะเรื่องดีเอ็นเอ จึงจำเป็นต้องอธิบายให้เข้าใจเกี่ยวกับเรื่องของดีเอ็นเอและเทคนิคการได้มาของข้อมูลพอสังเขป

สิ่งมีชีวิตทุกสิ่งในโลกดำรงพันธุ์อยู่ได้โดยการถ่ายทอดลักษณะทางพันธุกรรม (Genetic character) จากรุ่นหนึ่งไปสู่อีกรุ่นหนึ่งต่อไป ตัวอย่างเช่น มนุษย์เราถ่ายทอดลักษณะทางกายภาพ เช่น สีผม สีผิว สีตา ฯลฯ จากพ่อแม่ไปสู่ลูก เป็นต้น ลักษณะทางพันธุกรรมที่สืบทอดต่อกันไปนี้แตกต่างกันไปตามเผ่าพันธุ์ของสิ่งมีชีวิต ยีนเหล่านี้เป็นตัวควบคุมลักษณะทางพันธุกรรมให้แสดงออกมาในสิ่งมีชีวิตแต่ละรุ่น ทำให้การแสดงออกของลักษณะทางพันธุกรรมที่ได้รับการถ่ายทอดมาเพียงบางลักษณะ ซึ่งมีจำนวนน้อยมากเมื่อเทียบกับจำนวนของลักษณะทางพันธุกรรมที่ได้รับการถ่ายทอดมาทั้งหมด ลักษณะทางพันธุกรรมทั้งหมดที่แสดงออกและไม่แสดงออกมา จะถูกถ่ายทอดไปสู่รุ่นต่อไป การถ่ายทอดลักษณะทางพันธุกรรมของสิ่งมีชีวิตนี้จำเป็นต้องอาศัยข่าวสารทางพันธุกรรมของสิ่งมีชีวิตหรือที่เรียกว่า สารพันธุกรรม โดยสำหรับสิ่งมีชีวิตส่วนใหญ่แล้วจะมีดีเอ็นเอ (DNA) เป็นสารพันธุกรรม หรืออาจกล่าวได้ว่า ดีเอ็นเอเป็นที่เก็บข้อมูลทางพันธุกรรมของสิ่งมีชีวิต ข้อมูลทางพันธุกรรมถูกจัดเก็บอยู่ในรูปแบบที่เป็นรหัสที่แตกต่างกัน โดยจัดเรียงอยู่บนดีเอ็นเอ ซึ่งรหัสเหล่านี้เรียกว่า รหัสพันธุกรรม (Genetic code) ทำให้การถ่ายทอดลักษณะทางพันธุกรรมจึงเป็นลักษณะเฉพาะของรหัสพันธุกรรมแต่ละตัว และเนื่องจากยีนเป็นชุดของรหัสพันธุกรรมที่จัดเรียงอยู่บนดีเอ็นเอ การถ่ายทอดลักษณะทางพันธุกรรมจึงเป็นการถ่ายทอดยีนทั้งหมดด้วย

ความสามารถในการควบคุมการแสดงออกของลักษณะทางพันธุกรรมในสิ่งมีชีวิตแต่ละรุ่นของยีน เกี่ยวเนื่องมาจากการที่ยีนมีความเกี่ยวข้องกับการสังเคราะห์โปรตีนของเซลล์ ซึ่งถือว่ามีความเกี่ยวข้องกับการแสดงลักษณะและการดำรงชีวิตของสิ่งมีชีวิต โดยควบคุมการสังเคราะห์โปรตีนนั่นเอง

การที่ดีเอ็นเอสามารถควบคุมลักษณะทางพันธุกรรมและการสังเคราะห์โปรตีนของสิ่งมีชีวิตนี้เอง จึงมีผู้พยายามศึกษาและทำความเข้าใจความหมายของรหัสพันธุกรรมที่ดีเอ็นเอเก็บไว้ เพื่อเข้าใจลักษณะทางพันธุกรรมของสิ่งมีชีวิต และกลไกการตอบสนองต่อความต้องการหรือการกระตุ้นจากสิ่งแวดล้อมของเซลล์สิ่งมีชีวิต ในปัจจุบันจึงมีการคิดค้นเครื่องมือเพื่อช่วยในการศึกษาและวิเคราะห์ข้อมูลทางพันธุกรรมเหล่านี้

2.3 การถอดรหัสพันธุกรรมของดีเอ็นเอ

สิ่งมีชีวิตส่วนใหญ่มีดีเอ็นเอ (DNA: Deoxyribonucleic acid) เป็นสารพันธุกรรม ซึ่งบรรจุข่าวสารทางพันธุกรรมของสิ่งมีชีวิตในรูปแบบที่เป็นรหัสพันธุกรรม โดยดีเอ็นเอใช้รหัสพันธุกรรมที่เก็บไว้ใน การควบคุมลักษณะทางพันธุกรรมและการสังเคราะห์โปรตีนของสิ่งมีชีวิต เนื่องจากสิ่งมีชีวิตมีการถ่ายทอดลักษณะทางพันธุกรรมจากหนึ่งไปสู่อีกรุ่นหนึ่งได้โดยการถ่ายทอดดีเอ็นเอนั่นเอง

ดีเอ็นเอเป็นสารอินทรีย์ที่มีคุณสมบัติเป็นกรดและพบในนิวเคลียสของเซลล์สิ่งมีชีวิต ดีเอ็นเอ จะประกอบขึ้นจากการเชื่อมกันของสายพอลินิวคลีโอไทด์ (Polynucleotide) หรือสายพันธุกรรม 2 สาย ซึ่งทำให้เกิดโครงสร้างที่มีลักษณะเป็นเกลียว ซึ่งสายพันธุกรรมแต่ละสายประกอบด้วยนิวคลีโอไทด์ 4 ชนิด ได้แก่ อะดีนีน (adenine) กัวนีน (guanine) ไซโทซีน (cytosine) และไทมีน (thymine) เนื่องจากนิวคลีโอไทด์เหล่านี้มีธาตุไนโตรเจนเป็นองค์ประกอบหลักและมีคุณสมบัติเป็นเบส จึงมีการเรียกนิวคลีโอไทด์เหล่านี้สั้นๆว่า เบส (base) นิวคลีโอไทด์ทั้ง 4 ชนิดนี้แตกต่างกันตามส่วนประกอบที่เป็นเบส และมีชื่อเรียกสั้นๆว่า เบส A เบส G เบส C และเบส T ตามลำดับ ซึ่งนิวคลีโอไทด์เหล่านี้จะเชื่อมต่อกันไปเป็นสายกลายเป็นสายพันธุกรรม ลำดับการเชื่อมต่อของนิวคลีโอไทด์หรือที่เรียกว่าลำดับเบส (Base sequence) นั้นจะแตกต่างกันทำให้สายพันธุกรรมแต่ละสายมีความแตกต่างกัน แต่สำหรับสายพันธุกรรม 2 สายเชื่อมต่อกันกลายเป็นดีเอ็นเอแต่ละสาย โดยลำดับเบสแต่ละตำแหน่งสามารถจับคู่เข้ากับเบสในลำดับเบสของอีกสายหนึ่งในตำแหน่งที่ตรงกันได้ โดยเบส A จับคู่กับเบส T และเบส G จับคู่กับเบส C เสมอ เราจึงเรียกลำดับเบสที่สามารถจับคู่กับลำดับเบสที่สามารถจับคู่กับลำดับเบสอีกลำดับหนึ่งได้ว่า ลำดับเบสประกอบ (Complementary base sequence) ตัวอย่างเช่น ลำดับเบส G-T-C-C-T-A มีลำดับเบสประกอบเป็น C-A-G-G-A-T เป็นต้น ลำดับเบสที่ต่างกันเมื่อใช้เบสในจำนวนที่แตกต่างกันบนสายพันธุกรรมจะทำให้เกิดรหัสที่แตกต่างกันอย่างมาก ตัวอย่างเช่น ถ้าพิจารณาให้เบส 2 โมเลกุลเรียงต่อกันแบบ $(4^2 = 16)$ เป็นต้น ในความจริงแล้ว สายพันธุกรรมประกอบด้วย การจัดเรียงลำดับของเบสหลายโมเลกุลซึ่งอาจมีมากถึงระดับหมื่น โมเลกุล จึงก่อให้เกิดความหลากหลายทางพันธุกรรมของสิ่งมีชีวิตมากมาย

(ชโลธร, 2546)

2.4 โคดอนยูสเอส

รหัสพันธุกรรมหรือโคดอนหมายถึงลำดับของนิวคลีโอไทด์บนดีเอ็นเอ เพื่อทำการสื่อสารในการถอดข้อความของรหัสพันธุกรรมที่ได้จากดีเอ็นเอ ให้ออกมาเป็นลำดับของกรดอะมิโนในรูปแบบของโมเลกุลโปรตีน จากการค้นคว้าพบว่ามียุสเอสรหัสพันธุกรรมอยู่ 64 ตัวต่อกรดอะมิโนทั้งหมด 20 ชนิด โดยมีการจัดเรียงของนิวคลีโอไทด์เป็น 3 ตัว ซึ่งจะประกอบกันเป็นรหัสพันธุกรรมสำหรับกรดอะมิโนหนึ่งตัว กรดอะมิโนตัวหนึ่งอาจมีรหัสพันธุกรรมมากกว่าหนึ่งได้ ดังนั้นรหัสพันธุกรรมบนอาร์เอ็นเอสื่อสารนี้เรียกว่า ทริเพลตโคดอน (Triplet code) หรือโคดอน (Codon) ดังตาราง 2.1

ตาราง 2.1 ตารางรหัสพันธุกรรม

(ที่มา: http://www.mun.ca/biology/scarr/MGA2_03-20.html)

First letter	Second letter				Third letter
	U	C	A	G	
U	UUU } Phe	UCU } Ser	UAU } Tyr	UGU } Cys	U
	UUC }	UCC }	UAC }	UGC }	C
	UUA } Leu	UCA }	UAA Stop	UGA Stop	A
	UUG }	UCG }	UAG Stop	UGG Trp	G
C	CUU } Leu	CCU } Pro	CAU } His	CGU } Arg	U
	CUC }	CCC }	CAC }	CGC }	C
	CUA }	CCA }	CAA } Gln	CGA }	A
	CUG }	CCG }	CAG }	CGG }	G
A	AUU } Ile	ACU } Thr	AAU } Asn	AGU } Ser	U
	AUC }	ACC }	AAC }	AGC }	C
	AUA }	ACA }	AAA } Lys	AGA } Arg	A
	AUG Met	ACG }	AAG }	AGG }	G
G	GUU } Val	GCU } Ala	GAU } Asp	GGU } Gly	U
	GUC }	GCC }	GAC }	GGC }	C
	GUA }	GCA }	GAA } Glu	GGA }	A
	GUG }	GCG }	GAG }	GGG }	G

ในรหัสพันธุกรรม ที่มีการจัดเรียงของรหัสดีเอ็นเอในแต่ละตำแหน่งจะมีเบสอยู่ 3 โมเลกุล เราจึงเรียกแต่ละรหัสว่าโคดอน ซึ่งถ้าดูในระดับของอาร์เอ็นเอ จะสามารถเข้าใจถึงหน้าที่ของโคดอนแต่ละตัว ได้คือ ในการสังเคราะห์โปรตีนทรานสเฟออาร์เอ็นเอ (tRNA) มีหน้าที่นำกรดอะมิโนมายังแมสเซนเจอร์อาร์เอ็นเอ โดยทรานสเฟออาร์เอ็นเอ จะเข้าจับกับแมสเซนเจอร์อาร์เอ็นเอ ซึ่งทรานสเฟออาร์เอ็นเอแต่ละโมเลกุลจะมีลำดับเบส 3 โมเลกุลที่เข้าคู่กับโคดอนได้ เรียกเบส 3 โมเลกุล โดยทรานสเฟออาร์เอ็นเอ นี้เรียกว่า แอนติโคดอน (Anticodon)

โดยการแปลลำดับของเบสในแมสเซนเจอร์อาร์เอ็นเอ (mRNA) จะต้องเริ่มจากปลาย 5' ไปยังปลาย 3' โดยที่ลำดับของเบสสามตัวติดต่อกันจะกำหนดเป็นรหัสสำหรับกรดอะมิโนหนึ่งตัว สำหรับรหัสเริ่มต้นของกรดอะมิโนตัวแรกคือ AUG ซึ่งเป็นพวกโปรคาริโอท โดย

กรดอะมิโนนั้นคือ เอ็น-ฟอร์มิล เมไทโอนีนซึ่งถูกพามาด้วยทรานสเฟอรัอาร์เอ็นเอ ที่ต่างกับกับ ทรานสเฟอรัอาร์เอ็นเอที่พากรดอะมิโนเมไทโอนีนภายในห่วงโซ่พอลิเพปไทด์ สำหรับพวกยูคาริโอตนั้น กรดอะมิโนตัวแรกคือเมไทโอนีนจะไม่มีการเติมหมู่ฟอร์มิล ที่เมไทโอนีน ยกเว้นการสังเคราะห์โปรตีนที่เกิดขึ้นในไมโทคอนเดรีย (Mitochondria) และคลอโรพลาสต์ (Chloroplasts) ซึ่งจะมี เอ็น-ฟอร์มิล เมไทโอนีนเป็นกรดอะมิโนตัวแรก AUG ซึ่งจะเป็นโคดอน เดียวเท่านั้นที่จำเพาะต่อกรดเมไทโอนีนจึงทำหน้าที่เป็นทั้งโคดอนเริ่มต้นและโคดอนสำหรับ เมไทโอนีนที่อยู่ภายในห่วงโซ่พอลิเพปไทด์ ส่วนการแปลรหัสจะหยุดเมื่อพบโคดอน ที่มีความหมายให้หยุดหรือรหัสหยุด เนื่องจากไม่มีทรานสเฟอรัอาร์เอ็นเอ ที่มีแอนติโคดอนที่ เฉพาะต่อโคดอนนั้น จะมีการกำหนดโคดอนที่เป็นรหัสหยุดคือ UAA

(Nirenberg and Khorana, 1961)

2.4.1 การคำนวณหาความเหมือนเชิงเปรียบเทียบโดยใช้ความถี่ในรูปแบบโคดอน

จากการพัฒนาวิธีการในการจำแนกยีน โดยใช้โคดอนเป็นคุณสมบัติหลักและ ใช้ซอฟต์แวร์เวกเตอร์แมชชีนในการหาตำแหน่งของกลุ่มยีน เพื่อแก้ไขปัญหาการจำแนกชนิดของ ข้อมูล (Ma et al., 2009) โดยจะทำการเปลี่ยนลำดับเบสของข้อมูลเข้าทั้งหมดให้อยู่ในรูปของ เวกเตอร์ที่มี 59 มิติของคุณลักษณะ ซึ่งในแต่ละองค์ประกอบนั้นจะสอดคล้องกับการคำนวณหา ความเหมือนเชิงเปรียบเทียบโดยใช้ความถี่ในรูปแบบโคดอน (อาเอสซียู) ซึ่งสามารถคำนวณได้ ดังสมการ

$$r_k = n_k \times \frac{obs_k}{tot_k} \quad (2.1)$$

โดย	r_k	คือค่าความเหมือนของโคดอนที่ตำแหน่ง k ใดๆ
	obs_k	คือจำนวนของการพบโคดอนที่เหมือนกันในตำแหน่ง k ใดๆ
	tot_k	คือจำนวนทั้งหมดของการพบโคดอนที่เหมือนกันในตำแหน่ง k ใดๆ
	n_k	คือจำนวนของโคดอนที่เหมือนกันที่ตำแหน่ง k ใดๆ

ซึ่งจากวิธีการข้างต้นจะเห็นว่าคุณลักษณะทั้งหมดของโคดอนจะมีเท่ากับ 64 มิติ ของคุณลักษณะดังตาราง 2.1 แต่จะมีอยู่ 5 โคดอนที่เป็นโคดอนที่มีลักษณะและมีหน้าที่เฉพาะเจาะจง ซึ่งจะไม่นำมาพิจารณา คือ UGG, AUG, UGA, UAA, UAG ซึ่งจะเหลือโคดอนที่นำมาพิจารณา เท่ากับ 59 มิติ จากคุณลักษณะทั้งหมดที่ใช้เป็นคุณลักษณะในการจำแนกชนิดข้อมูลต่อไป

โดยวิธีการนี้จะใช้ในการแปลงคุณลักษณะของวิธีการทันสมัยต่างๆ เพื่อใช้ในการปรับให้อยู่บนมาตรฐานเดียวกัน ซึ่งเป็นส่วนหนึ่งของวิธีการเปรียบเทียบประสิทธิภาพในการจำแนกชนิดข้อมูล เป็นต้น

2.5 ไลโคดอนยูสเอส

ไลโคดอนเป็นรหัสพันธุกรรมที่พัฒนาหรือต่อ ยอดการวิเคราะห์ข้อมูลสายลำดับจากไลโคดอน โดยจะพิจารณาข้อมูลตามคุณลักษณะของข้อมูล ซึ่งคุณลักษณะทางพันธุกรรมต่างๆ จะใช้ในการแปลงข้อมูลเข้าให้อยู่ในรูปของคุณลักษณะที่กำหนดสำหรับการจำแนกชนิดกลุ่มข้อมูล โดยจะทำการแจกแจงทุกกรณี ที่เป็นไปได้ทั้งหมด ตามลักษณะการพิจารณา ดังนี้

- ไลโคดอนจะมีคุณลักษณะลักษณะทางพันธุกรรมที่เป็นไปได้ทั้งหมดเท่ากับ 59 รูปแบบ
- ไลโคดอนจะมีคุณลักษณะทางพันธุกรรมที่เป็นไปได้ทั้งหมดเท่ากับ 4096 รูปแบบ

โดยในงานวิจัยลักษณะนี้จะแสดงให้เห็นถึงการใช้งานของไลโคดอน ซึ่งเป็นปัจจัยในการวิเคราะห์คุณลักษณะของข้อมูลสายลำดับนั้นๆ และจะเห็นได้ว่าผลลัพธ์ที่ได้ดีกว่าการพิจารณาโดยไลโคดอนเพียงอย่างเดียว ดังนั้นการใช้ไลโคดอนจึงเป็นตัวชี้วัดที่ดีในการแสดงออกของยีนและการศึกษาในระดับโมเลกุลต่อไป

(Nguyen et al, 2009)

2.6 การเรียนรู้ด้วยเครื่อง

การเรียนรู้ของเครื่องเป็นสาขาหนึ่งของปัญญาประดิษฐ์ ที่เกี่ยวข้องกับการพัฒนาเทคนิควิธีเพื่อให้คอมพิวเตอร์สามารถเรียนรู้ โดยเน้นที่วิธีการเพื่อสร้างโปรแกรมคอมพิวเตอร์จากการวิเคราะห์ชุดข้อมูล โดยการเรียนรู้ด้วยเครื่องจะเกี่ยวข้องอย่างมากในงานทางสถิติศาสตร์และคอมพิวเตอร์ เนื่องจากทั้งสองสาขานี้ จะศึกษาถึงการวิเคราะห์ข้อมูลเช่นเดียวกัน โดยการเรียนรู้ของเครื่องจะแบ่งได้ตามลักษณะผลลัพธ์ โดยทั่วไปได้แก่

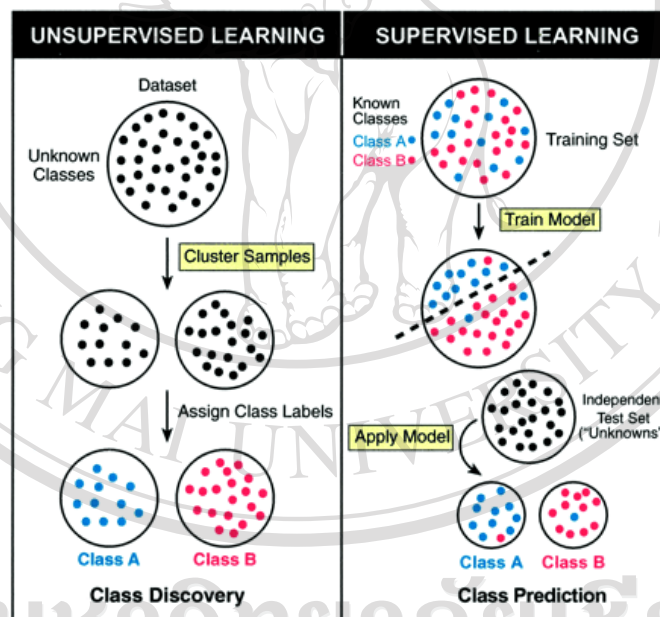
2.6.1 การเรียนแบบมีผู้สอน (Supervised Learning)

การเรียนรู้แบบมีผู้สอนเป็นเทคนิคหนึ่งของการเรียนรู้ของเครื่องซึ่งสร้างฟังก์ชันจากข้อมูลฝึกสอน (Training data) ซึ่งข้อมูลฝึกสอนนี้ จะประกอบด้วยวัตถุเข้าที่อยู่ในรูปของเวกเตอร์ และผลลัพธ์ที่ต้องการ โดยผลจากการเรียนรู้จะเป็นฟังก์ชันที่อาจจะให้ค่าต่อเนื่อง ซึ่งจะอยู่ในรูปแบบของการถดถอย (Regression) หรือใช้ทำนายประเภทของวัตถุ ซึ่งจะเรียกว่าการแบ่งประเภท (classification) โดยหน้าที่หลักของเครื่องเรียนรู้แบบมีผู้สอน จะทำนายค่าของ

ฟังก์ชันจากข้อมูลเข้าที่ถูกต้องโดยใช้ตัวอย่างการฝึกสอนจำนวนน้อย (Training examples) โดยเครื่องเรียนรู้จะต้องทำให้มีลักษณะทั่วไป (Generalize) จากข้อมูลที่มีอยู่ไปยังกรณีที่ไม่เคยพบอย่างมีเหตุผล

2.6.2 การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning)

การเรียนรู้แบบไม่มีผู้สอนเป็นเทคนิคหนึ่งของการเรียนรู้ด้วยเครื่อง โดยการสร้างโมเดลที่เหมาะสมกับข้อมูล โดยการเรียนรู้แบบนี้จะแตกต่างจากการเรียนรู้แบบมีผู้สอน คือจะไม่มีการระบุผลลัพธ์ที่ต้องการหรือประเภทไว้ก่อน ซึ่งการเรียนรู้แบบนี้จะพิจารณาวัตถุเป็นเซตของตัวแปรสุ่ม แล้วจึงสร้างโมเดลความหนาแน่นร่วมของชุดข้อมูล และยังสามารถนำไปใช้ร่วมกับการอนุมานแบบเบย์ เพื่อหาความน่าจะเป็นแบบมีเงื่อนไขของตัวแปรสุ่ม และจัดอยู่ในรูปแบบของการแบ่งกลุ่มข้อมูล



รูป 2.5 การเรียนรู้แบบมีผู้สอนและการเรียนรู้แบบไม่มีผู้สอน

(ที่มา: <http://jco.ascopubs.org/content/20/7/1932.full>)