แบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

DNA Sequence Analysis Model and Pattern Recognition

ธเนศ ปานรัตน์

Tanate Panrat

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์
มหาวิทยาลัยสงขลานครินทร์

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science
Prince of Songkla University
2551

ชื่อวิทยานิพนธ์ แบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม
ผู้เขียน นายธเนศ ปานรัตน์
สาขาวิชา วิทยาการคอมพิวเตอร์

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

.....
(ดร.วิภาดา เวทย์ประสิทธิ์)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม

.....
(ดร.ลัดดา ปรีชาวีรกุล)

คณะกรรมการสอบ

.....ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.กฤษณะ ชินสาร)

.....กรรมการ
(รองศาสตราจารย์ ดร.อมรรัตน์ พงศ์ดารา)

.....กรรมการ
(ดร.ลัดดา ปรีชาวีรกุล)

.....กรรมการ
(ดร.วิภาดา เวทย์ประสิทธิ์)

บัณฑิตวิทยาลัย มหาวิทยาลัยสงขลานครินทร์ อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษา ตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์

.....
(รองศาสตราจารย์ ดร.เกริกชัย ทองหนู)
คณบดีบัณฑิตวิทยาลัย

ชื่อวิทยานิพนธ์ แบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม
ผู้เขียน นายธเนศ ปานรัตน์
สาขาวิชา วิทยาการคอมพิวเตอร์
ปีการศึกษา 2550

บทคัดย่อ

การศึกษารูปแบบและการมีปฏิสัมพันธ์ระหว่างโปรตีนกับโปรตีน (Protein-Protein Interaction) บนสายพันธุกรรม ช่วยให้ทราบลักษณะเด่นของโรคแต่ละชนิด ซึ่งปรากฏบนสายพันธุกรรม งานวิจัยนี้ได้นำเสนอต้นแบบการวิเคราะห์และรู้จำลำดับพันธุกรรมโดยใช้เทคนิคการสกัดลักษณะเฉพาะบนสายโปรตีน (Feature Extraction of Protein Sequence Analysis: FEPSA) ด้วยข้อมูลแบบ 2-Grams ที่มีการปรับค่าน้ำหนัก (Weight Balance Frequency: WBL) ร่วมกับการเลือกลักษณะเฉพาะแบบทั้งชุดข้อมูล (Global Sequences Analyze) เพื่อการหาลักษณะเด่นของสายพันธุกรรมสำหรับการจัดกลุ่มโรค เทคนิคที่นำเสนอประกอบด้วย 3 ขั้นตอนหลัก คือ 1) การเตรียมข้อมูลแบบปรับค่าน้ำหนัก 2) การสกัดลักษณะเฉพาะบนสายพันธุกรรม และ 3) การจำแนกสายพันธุกรรม ทดสอบโดยใช้ข้อมูลสายพันธุกรรมของผู้ป่วยโรคธาลัสซีเมีย เบาหวาน และดาวน์ซินโดรม พบว่าเทคนิคการเรียนรู้ของโครงข่ายประสาทเทียมแบบ Multilayer Perceptron (MLP) ให้ผลการจำแนกข้อมูลที่มีค่าความถูกต้องสูง สำหรับเวลาที่ใช้ในการจำแนกข้อมูลพบว่าขั้นตอนวิธี RandomTree ใช้เวลาน้อยที่สุด

Thesis Title	DNA Sequence Analysis Model and Pattern Recognition
Author	Mr. Tanate Panrat
Major Program	Computer Science
Academic Year	2007

ABSTRACT

The study of pattern and protein-protein interaction on genetic sequences will be beneficial to understand the dominant feature of each disease that occurs on those genetic sequences. This study proposes the prototype of DNA Sequence Analysis Model and Pattern Recognition by the technique of Feature Extraction of Protein Sequence Analysis (FEPSA) using 2-Grams data type with adjusting Weight Balance Frequency (WBL) and global sequences analyze. Finding dominant feature of genetic sequence for disease composes of 3 main steps; step 1) preparing of weight balance frequency data, step 2) extracting dominant feature on genetic sequence, and step 3) using dominant feature to classify genetic sequence. The data used for this experiment were patients' protein sequences of Thalassemia, Diabetes, and Downsyndrome. The study found that the technique of Multilayer Perceptron Neural Networks gave the highest accuracy while Random Tree used the least time.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยดี ด้วยการสนับสนุนจากบุคคลดังนี้

ดร.วิภาดา เวทย์ประสิทธิ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่กรุณาให้คำปรึกษา แนะนำและช่วยเหลือในการแก้ปัญหาต่างๆ พร้อมทั้งตรวจทานและแก้ไขวิทยานิพนธ์ให้แก่ผู้วิจัย

ดร.ลัดดา ปรีชาวีรกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ที่กรุณาให้ข้อเสนอแนะต่าง ๆ รวมทั้งตรวจทานและแก้ไขวิทยานิพนธ์ให้แก่ผู้วิจัย

ดร.ศิริรัตน์ วณิชโยบล ที่กรุณาให้ข้อเสนอแนะต่าง ๆ ในการทำวิทยานิพนธ์ และคำแนะนำสำหรับการดูแลสุขภาพในระหว่างการทำวิทยานิพนธ์แก่ผู้วิจัย

ดร.ประวัติ เวทย์ประสิทธิ์ ที่ให้คำปรึกษา ข้อแนะนำและปรับปรุงไวยกรณ์ เนื้อหา ผลงานวิจัยในส่วนที่เป็นภาษาอังกฤษแก่ผู้วิจัยให้มีความสมบูรณ์

อาจารย์ภาควิชาวิทยาการคอมพิวเตอร์ทุกท่าน ที่ให้ความรู้ซึ่งสามารถนำมาใช้ในการทำวิทยานิพนธ์ได้เป็นอย่างดียิ่ง ผศ.ดร.กฤษณะ ชินสาร ที่กรุณาช่วยตรวจทานและแก้ไขวิทยานิพนธ์ รศ.ดร.อมรรัตน์ พงศ์ดารา ที่ช่วยเหลือและให้คำแนะนำในการทำวิทยานิพนธ์แก่ผู้วิจัย

เจ้าหน้าที่ภาควิชาวิทยาการคอมพิวเตอร์ และเจ้าหน้าที่บัณฑิตวิทยาลัยทุกท่าน ที่ให้ความช่วยเหลือและอำนวยความสะดวกในเรื่องของเอกสาร และการเบิกจ่ายวัสดุต่างๆ ที่ใช้ในงานวิจัย

มูลนิธิเพื่อการศึกษาคอมพิวเตอร์และการสื่อสาร และภาควิชาวิทยาการ-คอมพิวเตอร์ ที่สนับสนุนทุนการศึกษาแก่ผู้วิจัย

นายพีรบูรณ์ พรหมจันทร์ นางสาวพชรกมล อำนวย นางสาวศิวพร ประมวล และนายธีราวุฒิ ภู่สันติสัมพันธ์ ที่เป็นกำลังใจและช่วยเหลือในการทำวิทยานิพนธ์

นางสาวมานิดา นุ้ยเล็ก นางสาวกลิ่นสุคนธ์ นิ่มกาญจนา ที่ช่วยตรวจทานเอกสารวิทยานิพนธ์ เพื่อนๆ และพี่ๆ ภาควิชาวิทยาการคอมพิวเตอร์ ที่ให้การช่วยเหลือในการทำวิทยานิพนธ์ด้วยดีเสมอมา

คุณพ่อคุณแม่และพี่ๆ ครอบครัวปานรัตน์ที่แสนอบอุ่น ที่ให้กำลังใจและให้การสนับสนุนในการทำวิทยานิพนธ์ของผู้วิจัยมาโดยตลอด

ผู้วิจัยขอขอบคุณทุกท่านเป็นอย่างสูงมา ณ โอกาสนี้

ธเนศ ปานรัตน์

สารบัญ

หน้า

สารบัญ..... (6)

รายการตาราง..... (9)

รายการภาพประกอบ..... (11)

บทที่

1 บทนำ

1.1 การตรวจเอกสาร

1.1.1 ชีวสารสนเทศศาสตร์..... 2

1.1.2 การเตรียมข้อมูล..... 4

1.1.3 การสกัดลักษณะเฉพาะ..... 5

1.1.4 การประยุกต์ใช้งานโครงข่ายประสาทเทียม..... 7

1.1.5 การประยุกต์ใช้งานต้นไม้การตัดสินใจ..... 8

1.2 วัตถุประสงค์ของวิทยานิพนธ์..... 9

1.3 ขอบเขตของวิทยานิพนธ์..... 9

1.4 ขั้นตอนและระยะเวลาการดำเนินงาน

1.4.1 ขั้นตอนการดำเนินงาน..... 9

1.4.2 ระยะเวลาดำเนินการวิจัย..... 9

1.4.3 แผนการดำเนินการวิจัย..... 10

1.5 สถานที่และเครื่องมือที่ใช้

1.5.1 สถานที่ทำการวิจัย..... 10

1.5.2 เครื่องมือที่ใช้..... 11

1.6 ประโยชน์ที่คาดว่าจะได้รับ..... 11

2 ทฤษฎีที่เกี่ยวข้องกับการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

2.1 จีโนม

2.1.1 ดีเอ็นเอและอาร์เอ็นเอ..... 13

2.1.2 กรดอะมิโนและโปรตีน..... 14

2.2 ฟาสท์-เอ..... 17

2.3 การทำความสะอาดข้อมูล..... 20

2.4 เอ็นแกรม..... 22

สารบัญ (ต่อ)

หน้า

2.5 โครงข่ายประสาทเทียม
2.5.1 พื้นฐานของโครงข่ายประสาทเทียม
2.5.1.1 ฟังก์ชันผลรวม..... 26
2.5.1.2 ฟังก์ชันกระตุ้น..... 27
2.5.2 การคำนวณภายในโครงข่ายประสาท..... 29
2.5.3 นิวรอนเพอร์เซฟตรอน..... 30
2.5.4 ประเภทของโครงข่ายประสาท..... 31
2.6 ต้นไม้การตัดสินใจ
2.6.1 การแทนข้อมูลในต้นไม้การตัดสินใจ..... 33
2.6.2 กฎจากต้นไม้การตัดสินใจ..... 34
2.6.3 ขั้นตอนวิธีของต้นไม้การตัดสินใจ……..... 35
2.7 การทดสอบประสิทธิภาพแบบ k-Fold Cross Validation..... 36
2.8 การวิเคราะห์ประสิทธิภาพ
2.8.1 ค่าความแม่นยำ..... 37
2.8.2 ค่าระลึกได้..... 38
2.8.3 ค่าถ่วงดุล..... 38
3 แบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม
3.1 แบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม
3.1.1 ขั้นตอนหลักที่ 1 การเตรียมข้อมูลค่าความถี่แบบปรับน้ำหนัก..... 41
3.1.2 ขั้นตอนหลักที่ 2 การสกัดลักษณะเฉพาะบนสายพันธุกรรม..... 47
3.1.3 ขั้นตอนหลักที่ 3 การจำแนกสายพันธุกรรมด้วยลักษณะเฉพาะ..... 58
4 โปรแกรมระบบวิเคราะห์รูปแบบลำดับพันธุกรรม
4.1 แผนภาพกระแสข้อมูลโปรแกรมวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม.. 60
4.2 ผังงานโปรแกรมวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม..... 65
4.3 การพัฒนาโปรแกรมวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม..... 70
5 ผลการทดลองและบทวิจารณ์
5.1 การออกแบบวิธีการทดลอง..... 74
5.2 การเตรียมเซตของคู่ลำดับกรดอะมิโนที่เป็นได้ด้วยเทคนิคเอ็นแกรม..... 76

สารบัญ (ต่อ)

หน้า

5.3 ข้อมูลผู้ป่วยโรคธาลัสซีเมีย

5.3.1 ขั้นตอนที่ 1 การเตรียมข้อมูลผู้ป่วยธาลัสซีเมียแบบปรับน้ำหนัก..... 77

5.3.2 ขั้นตอนที่ 2 การสกัดลักษณะเฉพาะบนข้อมูลผู้ป่วยธาลัสซีเมีย..... 85

5.3.3 ขั้นตอนที่ 3 จำแนกข้อมูลผู้ป่วยธาลัสซีเมียด้วยลักษณะเฉพาะ..... 87

5.3.4 เปรียบเทียบผลการทดลองข้อมูลผู้ป่วยโรคธาลัสซีเมีย..... 92

5.4 ข้อมูลผู้ป่วยโรคเบาหวาน

5.4.1 ขั้นตอนที่ 1 การเตรียมข้อมูลผู้ป่วยเบาหวานแบบปรับน้ำหนัก..... 99

5.4.2 ขั้นตอนที่ 2 การสกัดลักษณะเฉพาะบนข้อมูลผู้ป่วยเบาหวาน..... 107

5.4.3 ขั้นตอนที่ 3 จำแนกข้อมูลผู้ป่วยเบาหวานด้วยลักษณะเฉพาะ..... 109

5.3.4 เปรียบเทียบผลการทดลองข้อมูลผู้ป่วยโรคเบาหวาน..... 113

5.5 ข้อมูลผู้ป่วยโรคดาวน์ซินโดรม

5.5.1 ขั้นตอนที่ 1 การเตรียมข้อมูลผู้ป่วยโรคดาวน์ซินโดรมแบบปรับ น้ำหนัก..... 121

5.5.2 ขั้นตอนที่ 2 การสกัดลักษณะเฉพาะบนข้อมูลผู้ป่วยโรคดาวน์ซินโดรม 129

5.5.3 ขั้นตอนที่ 3 จำแนกข้อมูลผู้ป่วยโรคดาวน์ซินโดรมด้วยลักษณะเฉพาะ 130

5.5.4 เปรียบเทียบผลการทดลองข้อมูลผู้ป่วยโรคดาวน์ซินโดรม..... 135

5.6 วิเคราะห์ผลการทดลอง..... 143

6 บทสรุปและข้อเสนอแนะ

6.1 สรุปผลการวิจัย..... 145

6.2 ปัญหาและอุปสรรค..... 147

6.3 ข้อเสนอแนะ..... 147

บรรณานุกรม..... 148

ภาคผนวก

ก คู่มือผู้ใช้ระบบการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม..... 157

ข คู่มือการติดตั้งระบบการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม..... 170

ค ผลงานวิจัยที่ได้รับการตีพิมพ์ JCSSE 2007..... 176

ง ผลงานวิจัยที่ได้รับการตีพิมพ์ IEEE (RIVF 2008)..... 185

ประวัติผู้เขียน..... 190

รายการตาราง

ตาราง		หน้า
1.1	ระยะเวลาการดำเนินการวิจัย.....	10
2.1	ค่าของคอนฟิวชันเมตริกซ์ (Confusion Matrix) แบบ 2 กลุ่ม.....	37
3.1	ค่าความถี่คู่ลำดับกรดอะมิโนนับบนสายพันธุกรรม.....	44
3.2	ข้อมูลค่าความถี่คู่ลำดับกรดอะมิโนแบบมีการปรับค่าน้ำหนักบนสายพันธุกรรม..	46
3.3	ค่าสถานะของคู่ลำดับกรดอะมิโน เมื่อค่าน้ำหนักเฉลี่ยของชุดข้อมูลเท่ากับ 2.73	52
3.4	ผลรวมคู่ลำดับกรดอะมิโนที่มีค่าสถานะสถานะเท่ากับ 1	53
3.5	ค่าร้อยละของคู่ลำดับกรดอะมิโน (FPV).....	55
3.6	คู่ลำดับกรดอะมิโนที่สกัดได้ด้วยค่าร้อยละการยอมรับได้ (λ) เท่ากับ 60%.....	57
5.1	รายละเอียดการออกแบบการทดลองตามขั้นตอนวิธีของแบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรมด้วยขั้นตอนวิธี FEPSA.....	75
5.2	ค่าความถี่คู่ลำดับอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยธาลัสซีเมีย.....	78
5.3	ค่าสถิติของข้อมูลผู้ป่วยธาลัสซีเมียแบบ 2-Grams ที่ไม่ปรับค่าน้ำหนัก (NBL)...	84
5.4	ค่าสถิติของข้อมูลผู้ป่วยธาลัสซีเมียแบบ 2-Grams ที่ปรับค่าน้ำหนัก (WBL).....	85
5.5	จำนวนลักษณะเฉพาะที่สกัดได้จากข้อมูลผู้ป่วยธาลัสซีเมียจากการทดลองแบบ C และ D เมื่อกำหนดค่าร้อยละการยอมรับได้ (λ) ที่แตกต่างกัน.....	86
5.6	ความถูกต้องการจำแนกข้อมูลผู้ป่วยธาลัสซีเมียของการทดลองแบบ A และ B...	92
5.7	ความถูกต้องการจำแนกข้อมูลผู้ป่วยธาลัสซีเมียของการทดลองแบบ C และ D เมื่อกำหนดค่าร้อยละการยอมรับได้ (λ) ที่แตกต่างกัน.....	93
5.8	เปรียบเทียบค่าความถูกต้องการจำแนกข้อมูลผู้ป่วยธาลัสซีเมีย.....	96
5.9	เวลาที่ใช้ในการจำแนกข้อมูลผู้ป่วยธาลัสซีเมียของการทดลองแบบ A และ B.....	97
5.10	เวลาที่ใช้ในการจำแนกข้อมูลผู้ป่วยธาลัสซีเมียของการทดลองแบบ C และ D เมื่อกำหนดค่าร้อยละการยอมรับได้ (λ) ที่แตกต่างกัน……..	97
5.11	เปรียบเทียบเวลาที่ใช้จำแนกข้อมูลผู้ป่วยธาลัสซีเมีย.....	98
5.12	ค่าความถี่คู่ลำดับอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยเบาหวาน.....	100
5.13	ค่าสถิติของข้อมูลผู้ป่วยเบาหวานแบบ 2-Grams ที่ไม่ปรับค่าน้ำหนัก (NBL)…..	107
5.14	ค่าสถิติของข้อมูลผู้ป่วยเบาหวานแบบ 2-Grams ที่ปรับค่าน้ำหนัก (WBL)……..	107
5.15	จำนวนลักษณะเฉพาะที่สกัดได้จากข้อมูลผู้ป่วยเบาหวาน ของการทดลองแบบ C และ D เมื่อกำหนดค่าร้อยละการยอมรับได้ (λ) ที่แตกต่างกัน.....	108

รายการตาราง (ต่อ)

ตาราง		หน้า
5.16	ความถูกต้องการจำแนกข้อมูลผู้ป่วยเบาหวานของการทดลองแบบ A และ B.....	113
5.17	ความถูกต้องการจำแนกข้อมูลผู้ป่วยเบาหวานของการทดลองแบบ C และ D เมื่อกำหนดค่าร้อยละการยอมรับได้ (λ) ที่แตกต่างกัน.....	115
5.18	เปรียบเทียบค่าความถูกต้องการจำแนกข้อมูลผู้ป่วยเบาหวาน.....	118
5.19	เวลาที่ใช้ในการจำแนกข้อมูลผู้ป่วยเบาหวานของการทดลองแบบ A และ B.....	118
5.20	เวลาที่ใช้ในการจำแนกข้อมูลผู้ป่วยเบาหวานของการทดลองแบบ C และ D เมื่อกำหนดค่าร้อยละการยอมรับได้ (λ) ที่แตกต่างกัน.....	119
5.21	เปรียบเทียบเวลาที่ใช้จำแนกข้อมูลผู้ป่วยเบาหวาน.....	119
5.22	ค่าความถี่คู่ลำดับอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยดาวน์ซินโดรม.....	122
5.23	ค่าสถิติของข้อมูลผู้ป่วยดาวน์ซินโดรมแบบ 2-Grams ที่ไม่ปรับค่าน้ำหนัก (NBL)	128
5.24	ค่าสถิติของข้อมูลผู้ป่วยดาวน์ซินโดรมแบบ 2-Grams ที่ปรับค่าน้ำหนัก (WBL)..	129
5.25	จำนวนลักษณะเฉพาะที่สกัดได้จากข้อมูลผู้ป่วยดาวน์ซินโดรม ของการทดลองแบบ C และ D เมื่อกำหนดค่าร้อยละการยอมรับได้ (λ) ที่แตกต่างกัน.....	129
5.26	ความถูกต้องการจำแนกข้อมูลผู้ป่วยดาวน์ซินโดรมของการทดลองแบบ A และ B	135
5.27	ความถูกต้องการจำแนกข้อมูลผู้ป่วยดาวน์ซินโดรมของการทดลองแบบ Cและ D เมื่อกำหนดค่าร้อยละการยอมรับได้ (λ) ที่แตกต่างกัน.....	136
5.28	เปรียบเทียบค่าความถูกต้องการจำแนกข้อมูลผู้ป่วยดาวน์ซินโดรม.....	140
5.29	เวลาที่ใช้ในการจำแนกข้อมูลผู้ป่วยดาวน์ซินโดรมของการทดลองแบบ A และ B	140
5.30	เวลาที่ใช้ในการจำแนกข้อมูลผู้ป่วยดาวน์ซินโดรมของการทดลองแบบ C และ D เมื่อกำหนดค่าร้อยละการยอมรับได้ที่แตกต่างกัน.....	141
5.31	เปรียบเทียบเวลาที่ใช้จำแนกข้อมูลผู้ป่วยดาวน์ซินโดรม.....	141
5.32	ความถูกต้องในการจำแนกข้อมูลสายพันธุกรรมของการทดลองแบบ D.....	143
5.33	เวลาที่ใช้ในการจำแนกข้อมูลสายพันธุกรรมของการทดลองแบบ D.....	143

รายการภาพประกอบ

ภาพประกอบ		หน้า
2.1	โครงสร้างสารพันธุกรรม (DNA) ของมนุษย์	12
2.2	รายชื่อกรดอะมิโนมาตรฐาน 20 ชนิด	14
2.3	รหัสพันธุกรรม (Genetic Code)	15
2.4	กระบวนการถอดรหัสพันธุกรรมของกรดอะมิโนและโปรตีน	16
2.5	ลักษณะโครงสร้างการจัดเก็บข้อมูลแบบ FASTA	17
2.6	ตัวอย่างรายละเอียดของสายพันธุกรรม	18
2.7	กราฟแนวโน้มของค่าเฉลี่ย	20
2.8	เครือข่ายประสาทในสมองมนุษย์	24
2.9	พื้นฐานของโครงข่ายประสาทเทียม	25
2.10	องค์ประกอบของนิวรอน (Neuron)	26
2.11	ลักษณะของฟังก์ชันสเตป (Step Function)	27
2.12	ลักษณะของฟังก์ชันซิกมอยด์ (Sigmoid Function)	28
2.13	ลักษณะของฟังก์ชันเชิงเส้น (Linear Function)	28
2.14	ลักษณะของฟังก์ชันเกาเซียน (Gaussian Function)	29
2.15	การออกแบบเพอร์เซฟตรอน	30
2.16	โครงข่ายประสาทเทียมแบบไปข้างหน้า (Feedforward Neural Network)	31
2.17	โครงข่ายประสาทเทียมแบบย้อนกลับ (Recurrent Neural Network)	32
2.18	แสดงตัวอย่างการแทนข้อมูลในต้นไม้การตัดสินใจ	33
2.19	การสร้างกฎจากต้นไม้การตัดสินใจ	34
2.20	การแบ่งข้อมูลสำหรับทดสอบแบบ k-Fold Cross Validation	36
3.1	ขั้นตอนการทำงานแบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม	40
3.2	รายละเอียดขั้นตอนการเตรียมข้อมูลค่าความถี่แบบปรับน้ำหนัก	41
3.3	คำนวณเซตของคู่ลำดับที่เป็นได้ (Feature Vector) แบบ 2-Grams	42
3.4	ขั้นตอนวิธีการคำนวณความถี่คู่ลำดับกระอะมิโนบนสายพันธุกรรม	43
3.5	การหาค่าความถี่ด้วยเทคนิค n-Grams เมื่อกำหนดค่าของ n = 2 ตัวอักษร	44
3.6	ขั้นตอนวิธีการปรับน้ำหนักความถี่บนสายพันธุกรรม	45
3.7	รายละเอียดขั้นตอนการสกัดลักษณะเฉพาะบนสายพันธุกรรม	47

รายการภาพประกอบ (ต่อ)

ภาพประกอบ		หน้า
3.8	ขั้นตอนวิธีการคำนวณค่าน้ำหนักเฉลี่ยของชุดข้อมูล.....	48
3.9	ขั้นตอนวิธีการกำหนดค่าสถานะของคู่ลำดับกรดอะมิโน.....	50
3.10	ขั้นตอนวิธีการคำนวณผลรวมค่าสถานะของคู่ลำดับกรดอะมิโน.....	53
3.11	ขั้นตอนวิธีการคำนวณค่าร้อยละของคู่ลำดับกรดอะมิโน.....	54
3.12	ขั้นตอนวิเคราะห์ข้อมูลระดับชุดข้อมูล.....	56
3.13	รายละเอียดขั้นตอนการจำแนกสายพันธุกรรม.....	58
3.14	ขั้นตอนวิธีการจำแนกข้อมูล (Classification) ด้วยโปรแกรม WEKA.....	59
4.1	ภาพกระแสข้อมูลของระบบวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม……....	60
4.2	ภาพกระแสข้อมูลระดับที่ 1 ระบบวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม.....	61
4.3	ภาพกระแสข้อมูลระดับที่ 2 การเตรียมข้อมูลค่าความถี่แบบปรับน้ำหนัก.....	62
4.4	ภาพกระแสข้อมูลระดับที่ 2 การสกัดลักษณะเฉพาะบนสายพันธุกรรม.....	63
4.5	ภาพกระแสข้อมูลระดับที่ 2 การจำแนกสายพันธุกรรมด้วยลักษณะเฉพาะ.....	64
4.6	ผังงานโปรแกรมหลักระบบวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม.....	65
4.7	ผังงานโปรแกรมระบบวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม.....	66
4.8	ผังงานโปรแกรมย่อยการเตรียมข้อมูลค่าความถี่แบบมีการปรับน้ำหนัก.....	67
4.9	ผังงานโปรแกรมย่อยการสกัดลักษณะเฉพาะบนสายพันธุกรรม.....	68
4.10	ผังงานโปรแกรมย่อยการจำแนกสายพันธุกรรมด้วยลักษณะเฉพาะ.....	69
4.11	เมนูหลักของโปรแกรมวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม.....	70
4.12	หน้าต่างโปรแกรมการเตรียมข้อมูลลำดับพันธุกรรม.....	71
4.13	หน้าต่างโปรแกรมการสกัดลักษณะเฉพาะบนสายพันธุกรรม.....	72
4.14	หน้าต่างโปรแกรมการจำแนกสายพันธุกรรมด้วยลักษณะเฉพาะ.....	73
5.1	ภาพรวมการทดลองตามขั้นตอนวิธี FEPSA.....	75
5.2	ลักษณะของข้อมูลผู้ป่วยธาลัสซีเมียด้วยค่าความนิยม.....	77
5.3	จำนวนลักษณะเฉพาะที่สกัดได้จากการทดลองแบบ C และ D บนชุดข้อมูลผู้ป่วยธาลัสซีเมีย เมื่อกำหนดค่าร้อยละการยอมรับได้ที่แตกต่างกัน.....	86
5.4	โครงข่ายประสาทเทียมแบบ MLP สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยธาลัสซีเมีย เมื่อกำหนดค่า $\lambda \geq 10\%$.....	87

รายการภาพประกอบ (ต่อ)

ภาพประกอบ		หน้า
5.5	โครงข่ายประสาทเทียมแบบ MLP สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยธาลัสซีเมีย เมื่อกำหนดค่า $\lambda \geq 15\%$	88
5.6	โครงข่ายประสาทเทียมแบบ MLP สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยธาลัสซีเมีย เมื่อกำหนดค่า $\lambda \geq 20\%$	88
5.7	ต้นไม้การตัดสินใจแบบ ADTree สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยธาลัสซีเมีย เมื่อกำหนดค่า $\lambda \geq 10\%$	89
5.8	ต้นไม้การตัดสินใจแบบ ADTree สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยธาลัสซีเมีย เมื่อกำหนดค่า $\lambda \geq 15\%$	89
5.9	ต้นไม้การตัดสินใจแบบ ADTree สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยธาลัสซีเมีย เมื่อกำหนดค่า $\lambda \geq 20\%$	90
5.10	ต้นไม้การตัดสินใจแบบ J48 สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยธาลัสซีเมีย เมื่อกำหนดค่า $\lambda \geq 10\%$	90
5.11	ต้นไม้การตัดสินใจแบบ J48 สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยธาลัสซีเมีย เมื่อกำหนดค่า $\lambda \geq 15\%$	91
5.12	ต้นไม้การตัดสินใจแบบ J48 สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยธาลัสซีเมีย เมื่อกำหนดค่า $\lambda \geq 20\%$	91
5.13	ความถูกต้องการจำแนกข้อมูลผู้ป่วยธาลัสซีเมียของการทดลองแบบ A และ B	92
5.14	ความถูกต้องการจำแนกข้อมูลผู้ป่วยธาลัสซีเมียของการทดลองแบบ C และ D เมื่อกำหนดค่า $\lambda \geq 5\%$	94
5.15	ความถูกต้องการจำแนกข้อมูลผู้ป่วยธาลัสซีเมียของการทดลองแบบ C และ D เมื่อกำหนดค่า $\lambda \geq 10\%$	94
5.16	ความถูกต้องการจำแนกข้อมูลผู้ป่วยธาลัสซีเมียของการทดลองแบบ C และ D เมื่อกำหนดค่า $\lambda \geq 15\%$	95
5.17	ความถูกต้องการจำแนกข้อมูลผู้ป่วยธาลัสซีเมียของการทดลองแบบ C และ D เมื่อกำหนดค่า $\lambda \geq 20\%$	95
5.18	เปรียบเทียบค่าความถูกต้องการจำแนกข้อมูลผู้ป่วยธาลัสซีเมีย	96
5.19	เปรียบเทียบเวลาที่ใช้บนชุดข้อมูลผู้ป่วยธาลัสซีเมีย	98
5.20	ลักษณะของข้อมูลผู้ป่วยเบาหวานด้วยค่าความนิยม	99

รายการภาพประกอบ (ต่อ)

ภาพประกอบ		หน้า
5.21	จำนวนลักษณะเฉพาะที่สกัดได้จากการทดลองแบบ C และ D บนชุดข้อมูลผู้ป่วยเบาหวาน เมื่อกำหนดค่าร้อยละการยอมรับได้ที่แตกต่างกัน.....	108
5.22	โครงข่ายประสาทเทียมแบบ MLP สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยเบาหวาน เมื่อกำหนดค่า $\lambda \geq 10\%$.....	109
5.23	โครงข่ายประสาทเทียมแบบ MLP สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยเบาหวาน เมื่อกำหนดค่า $\lambda \geq 15\%$.....	109
5.24	โครงข่ายประสาทเทียมแบบ MLP สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยเบาหวาน เมื่อกำหนดค่า $\lambda \geq 20\%$.....	110
5.25	ต้นไม้การตัดสินใจแบบ ADTree สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยเบาหวาน เมื่อกำหนดค่า $\lambda \geq 10\%$.....	110
5.26	ต้นไม้การตัดสินใจแบบ ADTree สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยเบาหวาน เมื่อกำหนดค่า $\lambda \geq 15\%$.....	111
5.27	ต้นไม้การตัดสินใจแบบ ADTree สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยเบาหวาน เมื่อกำหนดค่า $\lambda \geq 20\%$.....	111
5.28	ต้นไม้การตัดสินใจแบบ J48 สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยเบาหวาน เมื่อกำหนดค่า $\lambda \geq 10\%$.....	112
5.29	ต้นไม้การตัดสินใจแบบ J48 สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยเบาหวาน เมื่อกำหนดค่า $\lambda \geq 15\%$.....	112
5.30	ต้นไม้การตัดสินใจแบบ J48 สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยเบาหวาน เมื่อกำหนดค่า $\lambda \geq 20\%$.....	113
5.31	ความถูกต้องการจำแนกข้อมูลผู้ป่วยเบาหวานของการทดลองแบบ A และ B....	114
5.32	ความถูกต้องการจำแนกข้อมูลผู้ป่วยเบาหวานของการทดลองแบบ C และ D เมื่อกำหนดค่า $\lambda \geq 5\%$.....	115
5.33	ความถูกต้องการจำแนกข้อมูลผู้ป่วยเบาหวานของการทดลองแบบ C และ D เมื่อกำหนดค่า $\lambda \geq 10\%$.....	116
5.34	ความถูกต้องการจำแนกข้อมูลผู้ป่วยเบาหวานของการทดลองแบบ C และ D เมื่อกำหนดค่า $\lambda \geq 15\%$.....	116

รายการภาพประกอบ (ต่อ)

ภาพประกอบ		หน้า
5.35	ความถูกต้องการจำแนกข้อมูลผู้ป่วยธาลัสซีเมียของการทดลองแบบ C และ D เมื่อกำหนดค่า $\lambda \geq 20\%$	117
5.36	เปรียบเทียบค่าความถูกต้องการจำแนกข้อมูลผู้ป่วยเบาหวาน	118
5.37	เปรียบเทียบเวลาที่ใช้บนชุดข้อมูลผู้ป่วยเบาหวาน	120
5.38	ลักษณะของข้อมูลผู้ป่วยดาวน์ซินโดรมด้วยค่าความนิยม	121
5.39	จำนวนลักษณะเฉพาะที่สกัดได้จากการทดลองแบบ C และ D บนชุดข้อมูลผู้ป่วยดาวน์ซินโดรม เมื่อกำหนดค่าร้อยละการยอมรับได้ที่แตกต่างกัน	130
5.40	โครงข่ายประสาทเทียมแบบ MLP สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยดาวน์ซินโดรม เมื่อกำหนดค่า $\lambda \geq 10\%$	131
5.41	โครงข่ายประสาทเทียมแบบ MLP สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยดาวน์ซินโดรม เมื่อกำหนดค่า $\lambda \geq 15\%$	131
5.42	โครงข่ายประสาทเทียมแบบ MLP สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยดาวน์ซินโดรม เมื่อกำหนดค่า $\lambda \geq 20\%$	132
5.43	ต้นไม้การตัดสินใจแบบ ADTree สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยดาวน์ซินโดรม เมื่อกำหนดค่า $\lambda \geq 10\%$	132
5.44	ต้นไม้การตัดสินใจแบบ ADTree สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยดาวน์ซินโดรม เมื่อกำหนดค่า $\lambda \geq 15\%$	133
5.45	ต้นไม้การตัดสินใจแบบ ADTree สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยดาวน์ซินโดรม เมื่อกำหนดค่า $\lambda \geq 20\%$	133
5.46	โครงข่ายประสาทเทียมแบบ J48 สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยดาวน์ซินโดรม เมื่อกำหนดค่า $\lambda \geq 10\%$	134
5.47	ต้นไม้การตัดสินใจแบบ J48 สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยดาวน์ซินโดรม เมื่อกำหนดค่า $\lambda \geq 15\%$	134
5.48	ต้นไม้การตัดสินใจแบบ J48 สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยดาวน์ซินโดรม เมื่อกำหนดค่า $\lambda \geq 20\%$	135

รายการภาพประกอบ (ต่อ)

ภาพประกอบ		หน้า
5.49	ความถูกต้องการจำแนกข้อมูลผู้ป่วยดาวน์ซินโดรมของการทดลองแบบ A และB	136
5.50	ความถูกต้องการจำแนกข้อมูลผู้ป่วยดาวน์ซินโดรมของการทดลองแบบ C และ D เมื่อกำหนดค่า $\lambda \geq 5\%$	137
5.51	ความถูกต้องการจำแนกข้อมูลผู้ป่วยดาวน์ซินโดรมของการทดลองแบบ C และ D เมื่อกำหนดค่า $\lambda \geq 10\%$	138
5.52	ความถูกต้องการจำแนกข้อมูลผู้ป่วยดาวน์ซินโดรมของการทดลองแบบ C และ D เมื่อกำหนดค่า $\lambda \geq 15\%$	138
5.53	ความถูกต้องการจำแนกข้อมูลผู้ป่วยดาวน์ซินโดรมของการทดลองแบบ C และ D เมื่อกำหนดค่า $\lambda \geq 20\%$	139
5.54	เปรียบเทียบค่าความถูกต้องการจำแนกข้อมูลผู้ป่วยดาวน์ซินโดรม	140
5.55	เปรียบเทียบเวลาที่ใช้บนชุดข้อมูลผู้ป่วยดาวน์ซินโดรม	142
5.56	เปรียบเทียบค่าความถูกต้องการจำแนกข้อมูลสายพันธุกรรมของการทดลองแบบ D เมื่อกำหนดค่า $\lambda \geq 10\%$	144
5.57	เปรียบเทียบเวลาที่ใช้ในการจำแนกข้อมูลสายพันธุกรรมของการทดลองแบบ D เมื่อกำหนดค่า $\lambda \geq 10\%$	144

บทที่ 1

บทนำ

ชีวสารสนเทศศาสตร์ (Bioinformatics) เป็นการประยุกต์ใช้เทคนิคการเก็บข้อมูลที่เป็นระบบ การประมวลผลและการประเมินผลทางด้านคอมพิวเตอร์ สำหรับหาคำตอบทางด้าน วิทยาศาสตร์ชีวภาพ ผลจากการศึกษาในระดับห้องปฏิบัติการ (Wet Lab) ส่งผลให้ทราบถึงข้อมูลจีโนม (Genome) ของสิ่งมีชีวิต เช่น การสร้างบัญชีรายชื่อยีนและโปรตีนของมนุษย์ (Wiemann *et al.*, 2001) การศึกษาลักษณะจีโนมของกุ้งซึ่งส่งผลต่อการอยู่รอดในสภาพแวดล้อมของกุ้ง (อมรรัตน์ พงศ์ดารา, 2550) เป็นต้น ข้อมูลที่ได้ส่วนใหญ่จะอยู่ในรูปของสายพันธุกรรม ซึ่งเป็นการเรียงลำดับของนิวคลีโอไทด์ (Nucleotide Sequence) และสามารถแปลรหัสพันธุกรรม (Translation) ให้อยู่ในรูปของกรดอะมิโน (Amino Acid) ได้ (สุนทร โสตถิพันธุ์ และคณะ. 2548) สังเกตุได้ว่าข้อมูลที่ได้จากห้องปฏิบัติการอาจมีรูปแบบการจัดเก็บข้อมูล ความยาว มาตรฐานของสายพันธุกรรมซึ่งมีความแตกต่างกัน และมีความหลากหลายทั้งด้านข้อมูลและมีจำนวนมากทั้งลักษณะของข้อมูลที่ใช้ (วสันต์ จันทราทิตย์ และคณะ, 2544)

ปัญหาหลักอย่างหนึ่งที่พบในการศึกษาเกี่ยวกับจีโนมในระดับโมเลกุล คือ ใช้เวลาในห้องปฏิบัติการนาน อีกทั้งมีขั้นตอนในการศึกษาที่ซับซ้อนและต้องการความชำนาญของผู้เชี่ยวชาญเฉพาะด้านเพื่อให้ได้ข้อมูลจีโนมหรือผลการศึกษาที่ต้องการ การนำเทคโนโลยีทางด้านคอมพิวเตอร์เข้ามาช่วยในขั้นตอนการดำเนินงานในห้องปฏิบัติการ โดยการแปลงข้อมูลให้อยู่ในรูปแบบที่สามารถนำมาคำนวณ หรือประมวลผลด้วยคอมพิวเตอร์ได้ สามารถช่วยให้ได้ผลลัพธ์จากห้องปฏิบัติการที่มีความแม่นยำมากขึ้น และช่วยลดเวลาการดำเนินการลงได้ (Cohen, 2004)

การประยุกต์ใช้ขั้นตอนวิธี (Algorithm) ทางด้านคอมพิวเตอร์สำหรับวิเคราะห์และแก้ปัญหาเกี่ยวกับการศึกษาความสัมพันธ์ของข้อมูลในระดับโมเลกุล การหาลักษณะของตำแหน่งซึ่งบ่งบอกถึงลักษณะเด่นของสายพันธุกรรม หรือการแสดงออกของโรคซึ่งปรากฏบนสายพันธุกรรมแต่ละสาย โดยข้อมูลเหล่านี้สามารถนำไปวิเคราะห์หาลักษณะเด่นของกลุ่มสายพันธุกรรมแต่ละกลุ่ม สอดคล้องกับการเกิดโรคแต่ละชนิดได้ เช่น การศึกษาและการทำนายรูปแบบของการเกิดปฏิสัมพันธ์ระหว่างโปรตีนกับโปรตีน (Protein-Protein Interaction) โดยใช้เทคนิคแอ็นแกรม (n-Grams) แปลงข้อมูลสายพันธุกรรม และจำแนกเพื่อเปรียบเทียบความคล้ายคลึงกันของสายพันธุกรรม (Nanni, 2005)

เทคนิค Support Vector Machine (SVM) เป็นเทคนิคการจัดกลุ่มข้อมูลที่นิยมนำมาใช้ในการวิเคราะห์ลำดับโปรตีนบนสายพันธุกรรม เช่น มีการประยุกต์ใช้ SVM เพื่อสร้างตัวจำแนกกลุ่มของข้อมูล (Classifier) สำหรับการเรียนรู้ลักษณะของโปรตีนกลุ่มต่าง ๆ ที่มีความคล้ายคลึงกันบนสายพันธุกรรม (Natalia and Cathy, 2006) และพบว่าเทคนิค SVM ยังเหมาะสำหรับปัญหาการวิเคราะห์ข้อมูลซึ่งมีจำนวนมากได้ดี เช่น การศึกษารูปแบบของลำดับสายพันธุกรรมที่มีสายพันธุกรรมจำนวนมากด้วยขั้นตอนวิธี CSVM (Clustering Support Vector Machine) (Zhong *et al.*, 2007)

วิทยานิพนธ์นี้ได้นำเสนอแบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม สำหรับสกัดหาลักษณะเฉพาะ (Feature) ซึ่งแสดงคุณสมบัติของข้อมูลสายพันธุกรรมด้วยข้อมูลแบบเอ็นแกรม (n-Grams) ในข้อมูลผู้ป่วยโรคธาลัสซีเมีย (Thalassemia Disease) เบาหวาน (Diabetes Disease) และดาวน์ซินโดรม (Downsyndrome Disease) ซึ่งรวบรวมได้จาก The National Center for Biotechnology Information (National Center for Biotechnology Information, 2007) และใช้ขั้นตอนวิธีแบบโครงข่ายประสาทเทียม (Neural Networks) และต้นไม้การตัดสินใจ (Decision Trees) สำหรับจำแนกข้อมูล (Classification) สายพันธุกรรมด้วยลักษณะเฉพาะ

1.1 การตรวจสอบเอกสาร

1.1.1 ชีวสารสนเทศศาสตร์ (Bioinformatics)

1.1.1.1 จีโนม (Genome) คือ หน่วยพันธุกรรมหรือยีนทั้งหมดที่ควบคุมการแสดงลักษณะทางพันธุกรรมของสิ่งมีชีวิต ข้อมูลทางพันธุกรรมจะถูกบรรจุอยู่ภายในโครโมโซม (Chromosomes) โดยโครโมโซมจะประกอบอยู่ภายในนิวเคลียส (Nucleus) ของเซลล์ (Cell) ข้อมูลทางพันธุกรรมของสิ่งมีชีวิต เกิดจากการผสมผสานจากลักษณะพันธุกรรมของสิ่งมีชีวิตรุ่นบรรพบุรุษ เช่น ปู่และย่า ตาและยาย หรือพ่อและแม่ และถ่ายทอดคุณลักษณะของสิ่งมีชีวิตนั้นไปยังรุ่นถัดไป (วิสุทธิ์ ใบไม้, 2536; มนตรี จุฬาวัฒนทล และคณะ, 2543) และมีคำศัพท์ที่เกี่ยวข้อง คือ

1) ดีเอ็นเอ (DNA)

ดีเอ็นเอ คือ ชื่อย่อของสารพันธุกรรมที่มีชื่อวิทยาศาสตร์ว่า “กรดดีออกซีไรโบนิวคลีอิก” (Deoxyribonucleic Acid: DNA) โดยภายในลำดับของสายดีเอ็นเอจะประกอบด้วยเบสพื้นฐาน 4 ชนิด ได้แก่ เบสอะดีนีน (Adenine: A) เบสไทมีน (Thymine: T) เบสไซโตซีน (Cytosine: C) และเบสกัวนีน (Guanine: G) ตามลำดับ

2) กรดอะมิโน (Amino Acid) และโปรตีน (Proteins)

กรดอะมิโน คือ โมเลกุลซึ่งเกิดจากการจับตัวกันของเบสพื้นฐานครั้งละ 3 ชนิด หรือเรียกว่า “โคดอน” (Codon) และสามารถผลิตเป็นกรดอะมิโนมาตรฐาน (Standard Amino Acids) ทั้งสิ้น 20 ชนิด คือ อะลานีน (Alanine) ซิสเทอีน (Cysteine) กรดแอสปาร์ติก (Aspartic Acid) กรดกลูตามิก (Glutamic Acid) ฟีนิลอะลานีน (Phenylalanine) ไกลซีน (Glycine) ฮีสติดีน (Histidine) ไอโซลูซีน (Isoleucine) ไลซีน (Lysine) ลูซีน (Leucine) เมตไทโอนีน (Methionine) แอสปาราจีน (Asparagine) โปรลีน (Proline) กลูตามีน (Glutamine) อาร์จินีน (Arginine) เซอรีน (Serine) ทรีโอนีน (Threonine) วาลีน (Valine) ทริปโตแฟน (Tryptophan) และไทโรซีน (Tyrosine) ตามลำดับ

โปรตีน คือ โมเลกุลซึ่งมีขนาดใหญ่เกิดขึ้นจากการเชื่อมต่อกันของกรดอะมิโนด้วยพันธะเปปไทด์

1.1.1.2 เทคนิคและเครื่องมือสำหรับการคำนวณทางด้านชีวสารสนเทศ

เทคนิคและเครื่องมือที่นิยมทางด้านชีวสารสนเทศศาสตร์ สำหรับประยุกต์กับงานในห้องปฏิบัติการ ได้แก่

1.1.1.2.1 เทคนิคการเทียบเรียงข้อมูลสายพันธุกรรม (Sequence Alignment) สำหรับหาความเหมือน (Identity) และความคล้ายคลึง (Similarity) ของสายพันธุกรรม มี 2 ลักษณะ คือ

1) การเทียบเรียงแบบสองสายพันธุกรรม (Pair-Wise Sequence Alignment) ขั้นตอนวิธีที่นิยม คือ Needleman-Wunsch (Needleman and Wunsch, 1970), Smith-Waterman (Smith and Waterman, 1981), Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1990) และ Position-Specific Iterated Basic Local Alignment Search Tool (PSI-BLAST) (Friedman, 2004; Kesmir, 2007)

2) การเทียบเรียงแบบหลายสายพันธุกรรม (Multiple Sequence Alignment) ขั้นตอนวิธีที่นิยม คือ Progressive Alignment (Thompson *et al.*, 1994)

1.1.1.2.2 เครื่องมือเทียบเรียงข้อมูลสายพันธุกรรม ที่นิยมใช้ได้แก่ โปรแกรม ClustalV และ ClustalW สำหรับทำงานบนระบบปฏิบัติการ DOS หรือ โปรแกรม ClustalX สำหรับทำงานบนระบบปฏิบัติการ Windows (Altschul *et al.*, 1990) และโปรแกรม Phylip สำหรับสร้างต้นไม้สายสัมพันธ์ทางวิวัฒนาการของสิ่งมีชีวิต (Phylogenetic Trees) (Claverie and Notredame, 2003; Golding and Morton, 2006)

1.1.2 การเตรียมข้อมูล (Data Preprocessing)

การเตรียมข้อมูลนับเป็นสิ่งคำคัญอย่างยิ่งต่อการประมวลผลข้อมูล (Data Processing) โดยเฉพาะการศึกษาทางด้านชีวสารสนเทศจำเป็นต้องมีการเตรียมข้อมูล เพื่อเป็นการแปลงรูปแบบของข้อมูลทางด้านชีวภาพให้อยู่ในรูปแบบที่เหมาะสมและสามารถนำมาประมวลผลด้วยคอมพิวเตอร์ได้ (Mitra and Acharaya, 2003) เนื่องจากในบางครั้งข้อมูลสายพันธุกรรมอาจมีค่ามากหรือน้อยกว่าที่ควรจะเป็น ทำให้ข้อมูลที่นำมาประมวลผลไม่สม่ำเสมอ (Inconsistent) ข้อมูลอาจมีลักษณะที่แตกต่างกัน (Heterogeneous) หรือไม่ได้อยู่ในรูปแบบเดียวกัน ส่งผลให้การประมวลผลข้อมูลมีความผิดพลาดหรือไม่มีความน่าเชื่อถือ เทคนิคที่นำมาใช้สำหรับการเตรียมข้อมูลมี 2 ขั้นตอน คือ ขั้นตอนการทำความสะอาดข้อมูล (Data Cleaning) และขั้นตอนการเปลี่ยนรูปแบบข้อมูล (Data Transformation) (Kimball and Margy, 2002; Roiger and Greatz, 2003) อธิบายได้ดังนี้

1.1.2.1 ขั้นตอนการทำความสะอาดข้อมูล (Data Cleaning)

สายพันธุกรรมที่ใช้ในการประมวลผลข้อมูล อาจมีความยาวสายที่แตกต่างกันของข้อมูล การสุ่มเลือกตัวอย่าง (Sampling) เพื่อเลือกสายพันธุกรรมซึ่งมีความยาวที่ใกล้เคียงกันจึงเป็นสิ่งที่จำเป็น การหาตัวแทนของกลุ่มข้อมูลด้วยการวัดค่าแนวโน้มเข้าสู่ส่วนกลางเป็นหลักการทางด้านสถิติ สำหรับการหาศูนย์กลางหรือกึ่งกลางของข้อมูล โดยค่าสถิติที่นิยมใช้ ได้แก่ ค่าเฉลี่ยเลขคณิต (Mean Value) ค่าเฉลี่ยถ่วงน้ำหนัก (Weight Mean) มัธยฐาน (Median) ฐานนิยม (Mode) และเทคนิคการวัดการกระจายของข้อมูลสำหรับศึกษาลักษณะการกระจายของข้อมูล ได้แก่ พิสัย (Range) พิสัยระหว่างควอร์ไทล์ (Interquartile Range) ค่าเบี่ยงเบนเฉลี่ย (Mean Deviation หรือ Average Deviation) ความแปรปรวน (Variance) และค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) (ชัชวาลย์ เรืองประพันธ์, 2541; กัลยา วานิชย์บัญชา, 2546)

1.1.2.2 ขั้นตอนการเปลี่ยนรูปแบบข้อมูล (Data Transformation)

เนื่องจากข้อมูลด้านชีวสารสนเทศส่วนใหญ่อยู่ในรูปของสายอักขระ (Text Sequence) เช่น สายลำดับนิวคลีโอไทด์ (Nucleotide Sequence หรือ DNA Sequence) สายลำดับกรดอะมิโน (Amino Acid Sequence) หรือสายโปรตีน (Protein Sequence) เป็นต้น บางครั้งอาจไม่สามารถนำมาใช้ในการวิเคราะห์ (Data Analysis) หรือการประมวลข้อมูลด้วยขั้นตอนวิธีทางคอมพิวเตอร์บางขั้นตอนวิธีได้ เนื่องจากจำเป็นต้องใช้ข้อมูลซึ่งเป็นค่าสถิติ (Statically Value)

เทคนิคการแปลงข้อมูลแบบเอ็นแกรม (n-Grams) เป็นเทคนิคที่ได้รับความนิยม เป็นเทคนิคที่เปลี่ยนรูปแบบจากข้อมูลสายอักขระให้อยู่ในรูปของค่าสถิติด้วยทฤษฎีความน่าจะเป็น (Probability Theorem) ด้วยสมมติฐานของมาร์คอฟ (Markov Assumption) ที่ว่า “การปรากฏของตัวอักษรตัวหนึ่งขึ้นกับตัวอักษรก่อนหน้าเพียง n-1 ตัว”

การประยุกต์ใช้งานเทคนิคเอ็นแกรมกับงานทางด้านชีวสารสนเทศ เช่น การประยุกต์ใช้ร่วมกับเทคนิคการเทียบอักขระ (String Matching) บนสายพันธุกรรม (Kim and Shawe-taylor, 1994) การค้นหาเครื่องหมายที่บ่งชี้ลักษณะของจีโนม (Genome Signatures) บนสายโปรตีนจากรูปแบบการพับตัวของโปรตีน (Protein Folding) (Ganapathiraju *et al.*, 2002) และใช้สำหรับการรู้จำรูปแบบและการจำแนกสายพันธุกรรมที่มีส่วนของ Binding Sites, Immune Epitopes, Primers ซึ่งมีความเกี่ยวข้องกันบนฐานข้อมูลขนาดใหญ่ (Wren *et al.*, 2005) ประยุกต์ใช้ข้อมูลแบบเอ็นแกรมสำหรับการจำแนกสายพันธุกรรมแบบลำดับชั้น (Andrija *et al.*, 2005) การเทียบเรียงลำดับข้อมูลชีวภาพโดยการพิจราณาความถี่ส่วนย่อยของลำดับโดยใช้เทคนิคเอ็นแกรม (คมสัน จันมา และพันธุปิติ เปี่ยมสง่า, 2005) การทำนายลักษณะโครงสร้างของสายพันธุกรรมด้วยค่าความถี่ของข้อมูลแบบเอ็นแกรม (Nanni, 2005; Mhamdi *et al.*, 2006a) การทำนายรูปแบบของสายพันธุกรรม (Jiampojamarn *et al.*, 2005; Mhamdi *et al.*, 2006b) การเปรียบเทียบรูปแบบ (Pattern Matching) ของสายพันธุกรรมที่มีความคล้ายคลึงกันด้วยเทคนิคเอ็นแกรมร่วมกับเทคนิคแอลจิบรา (Algebraic Signatures) สำหรับการค้นคืนสายพันธุกรรมบนฐานข้อมูล (Litwin and Mokadem, 2007) นอกจากนี้ยังนำเทคนิคเอ็นแกรมมาประยุกต์ใช้สำหรับค้นหารูปแบบและวิเคราะห์ลักษณะของยีนบนฐานข้อมูลจีโนม (Poddar *et al.*, 2007)

จากประยุกต์ใช้ขั้นตอนการเตรียมข้อมูลข้างต้น นอกจากปัญหาของการมีข้อมูลจำนวนมากแล้วปัญหาที่พบอีกประการ คือ การมีจำนวนคุณสมบัติของข้อมูล (Attributes) ที่มาก และข้อมูลมีจำนวนมิติของข้อมูลที่ซับซ้อน ซึ่งส่งผลต่อประสิทธิภาพของการประมวลผล เนื่องจากต้องใช้พื้นที่หน่วยความจำหลัก (Main Memory) จำนวนมาก และใช้เวลาในการประมวลผลนาน

1.1.3 การสกัดลักษณะเฉพาะ (Feature Extraction)

การสกัดลักษณะเฉพาะของข้อมูล เป็นการค้นหารูปแบบของข้อมูล (Characteristics) ซึ่งเป็นลักษณะเด่นของข้อมูลนั้น ทำให้ง่ายต่อการทำความเข้าใจ และง่ายต่อการจำแนกข้อมูลจำนวนมากได้ นอกจากนี้การสกัดลักษณะเฉพาะด้วยหลักการลดขนาดและมิติของข้อมูล (Dimensional Reduction) ช่วยให้การประมวลผลสามารถทำงานได้รวดเร็ว และมีความแม่นยำที่สูงขึ้น (Doda *et al.*, 2001; Burges, 2004) เทคนิคการสกัดลักษณะเฉพาะที่นิยม

ใช้ เช่น ความน่าจะเป็น (Probabilistic) สำหรับการศึกษารูปแบบและการเลือกลักษณะเฉพาะของข้อมูล (Feature Selection) (Devroye *et al.*, 1996; Kohavi and John, 1997) เทคนิคการวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis: PCA) เทคนิคการวิเคราะห์แบบต่อเนื่องของฟิชเชอร์ (Fisher's Linear Discriminant Analysis) เทคนิคการกระจายเมทริกซ์ด้วยวิธีการแยกค่าแบบเดี่ยว (Singular Value Decomposition: SVD) เทคนิคการวิเคราะห์ความเป็นอิสระขององค์ประกอบ (Independent Component Analysis: ICA) (Cios *et al.*, 2007) และเทคนิคการวิเคราะห์รูปแบบข้อมูลทางด้านชีวสารสนเทศแบบ Polynomial Time (Nilsson *et al.*, 2007) เป็นต้น

การประยุกต์ใช้หลักการสกัดลักษณะเฉพาะกับงานทางด้านชีวสารสนเทศ เช่น การหาลักษณะเฉพาะเพื่อหาค่าความสอดคล้องของลำดับเบสสำหรับการวิเคราะห์และค้นคืนสายพันธุกรรม (Pereira *et al.*, 2000) การหาลักษณะเฉพาะจากสายพันธุกรรมด้วยวิธีการวิเคราะห์แบบ Multifractal Analysis (Zhang and Kinsner, 2001) การสกัดลักษณะเด่นเพื่อระบุตำแหน่งของการสังเคราะห์ยีน (Genes Coding) บนสายโปรโมเตอร์ (Promoter Sequences) (Fukui *et al.*, 2003) การจำแนกตระกูล (Superfamily) ของสายลำดับกรดอะมิโนด้วยเทคนิคการสกัดลักษณะเฉพาะ (Bandyopadhyay, 2003; Bandyopadhyay, 2005) การศึกษาชนิดคุณสมบัติและหน้าที่ของโปรตีนที่อยู่ภายในเซลล์ (Akay, 2004) การหาลักษณะเฉพาะของลำดับเบสโดยใช้เทคนิค Non-Negative Least Squares (NNLS) (Andrade-Cetto and Manolakos, 2005) การสกัดหารูปแบบการแสดงออกของยีนร่วมกับโครงข่ายประสาทเทียม (Tan and Pan, 2005) การหาลักษณะเด่นสำหรับทำนายตำแหน่งการเกิด HIV-1 protease Cleavage Site ในสายพันธุกรรมผู้ป่วยเอสด์ (Nanni, 2006a; Nanni and Alessandra, 2006) การศึกษารูปแบบการพับตัวของโปรตีน (Protein Fold Recognition) (Nanni, 2006b) การระบุตำแหน่ง Splice-Junction บนสายพันธุกรรมด้วยหลักการจำแนกแบบลำดับขั้น (Hierarchical Multiclassifier) (Alessandra and Nanni, 2006) การหารูปแบบความสัมพันธ์บนสายพันธุกรรมสำหรับการสร้างข่ายงานพันธุศาสตร์ (Phylogenetic Network) (M.A.H. Zahid *et al.*, 2005a; M.A.H. Zahid *et al.*, 2005b; M.A.H. Zahid *et al.*, 2006) และนอกจากนี้ยังมีการพัฒนาเว็บเซอร์วิส (Web Service) สำหรับการวิเคราะห์หาลักษณะเฉพาะของสายพันธุกรรมด้วยขั้นตอนวิธีแบบเบย์เซียน (Nikolajewa *et al.*, 2007) เป็นต้น

จะสังเกตุได้ว่า การประยุกต์ใช้งานเทคนิคการสกัดลักษณะข้างต้นสามารถลดจำนวนคุณสมบัติและความซับซ้อนของมิติของข้อมูลสำหรับใช้งานลงได้ ทั้งนี้ยังมีข้อจำกัดคือมีขั้นตอนการคำนวณที่ซับซ้อน และยุ่งยากในการออกแบบ

1.1.4 การประยุกต์ใช้งานโครงข่ายประสาทเทียม

โครงข่ายประสาทเทียม (Artificial Neural Networks: ANN) เป็นการประมวลผลสารสนเทศ (Information) ซึ่งมีความคล้ายคลึงกับการทำงานของเครือข่ายประสาทในสมองมนุษย์ โดยลักษณะเด่นของการทำงานโครงข่ายประสาทเทียม คือ สามารถเรียนรู้ได้ทั้งแบบมีผู้สอน (Supervised Learning) และแบบไม่มีผู้สอน (Unsupervised Learning) เหมาะสำหรับการแก้ไขปัญหาซึ่งมีความซับซ้อน และสามารถใช้ได้กับข้อมูลแบบต่อเนื่องและไม่ต่อเนื่องได้ (Rich and Knight, 1991; Hagan *et al.*, 1996; Eliasmith and Anderson, 2003)

ตัวอย่างการใช้งานโครงข่ายประสาทเทียมสำหรับงานทางด้านชีวสารสนเทศด้วยโครงข่ายประสาทเทียม เช่น การประมวลผลด้วยโครงข่ายประสาทเทียม (Neural Networks) สำหรับเรียนรู้รูปแบบของโปรตีนในสายพันธุกรรมสำหรับการจำแนกกลุ่มของโปรตีน (Protein Classification) บนฐานข้อมูลโปรตีน (Wu *et al*, 1991; Wu *et al.*, 1992a; Wu *et al.*, 1992b) การออกแบบโครงข่ายประสาทเทียมสำหรับการระบุกลุ่มของโปรตีนขนาดใหญ่ (Wu *et al.*, 1997; Wu and McLarty, 2000) การจำแนกลำดับโปรตีนบนฐานข้อมูลชีวภาพด้วยโครงข่ายประสาทเทียม (Wang *et al.*, 2000) การวิเคราะห์รูปแบบการแสดงออกของยีนโดยใช้โครงข่ายประสาทเทียมแบบลำดับขั้น (Hierarchical Neural Networks) (Herrero *et al.*, 2001) การใช้ขั้นตอนวิธี Expectation Maximization (EM) ร่วมกับเทคนิคโครงข่ายประสาทเทียมเพื่อวิเคราะห์ลำดับโปรโมเตอร์ (Promoter Sequences) (Ma *et al.*, 2001) การวิเคราะห์สายโปรตีนด้วยโครงข่ายประสาทเทียมแบบเรเดียลเบสิสฟังก์ชัน (Wang *et al.*, 2002) และนำหลักการเรียนรู้ของโครงข่ายประสาทเทียมแบบเรเดียลเบสิสฟังก์ชันมาสร้างโมเดล Bio-Basis Function Neural Network (BBFNN) สำหรับการวิเคราะห์สายโปรตีน (Berry *et al.*, 2003) การศึกษาลักษณะของยีนสำหรับทำนายโอกาสการเป็นโรคมะเร็งลำไส้ใหญ่ (Colon Cancer) จากข้อมูลดีเอ็นเอไมโครแอเรย์โดยใช้เทคนิคโครงข่ายประสาทเทียมแบบวิวัฒนาการ (Evolutionary Neural Networks) (Kim and Cho, 2004) และใช้โครงข่ายประสาทเทียมสร้างโมเดล Mixed Environment-Specific Substitution Mapping (MESSM) สำหรับการสร้างแผนที่เพื่อการทำนายรูปแบบการพับของโปรตีน (Protein Fold) บนสายลำดับกรดอะมิโน (Jiang *et al.*, 2005) การวิเคราะห์โรคมะเร็งต่อมลูกหมากจากข้อมูลดีเอ็นเอไมโครแอเรย์ (Peterson *et al.,* 2005) การใช้โครงข่ายประสาทเทียมแบบจัดกลุ่มเอง (Self-Organizing Neural Networks) สำหรับค้นหาความสัมพันธ์ของ DNA-Binding Motif บนสายพันธุกรรม (Mahony *et al.,* 2006) การทำนายกระบวนการเกิดปฏิสัมพันธ์ระหว่างโปรตีนกับโปรตีน (Protein-Protein Interactions Method) ด้วยโครงข่ายประสาทเทียมแบบย้อนกลับ (Back Propagation Neural Networks) (Ma *et al.*, 2007)

เทคนิคการเรียนรู้ของโครงข่ายประสาทเทียม นับว่าเป็นเทคนิคที่ได้รับความนิยมและเหมาะสมกับการแก้ไขปัญหาซึ่งมีความซับซ้อนของข้อมูลได้ดี และมีความทนทานต่อข้อมูลซึ่งมีสิ่งรบกวน (Noise) ได้ดี และในการแก้ปัญหาไม่จำเป็นต้องเข้าใจว่าปัญหาที่กำลังดำเนินการมีขั้นตอนอย่างไร อย่างไรก็ตามข้อจำกัดของเทคนิคโครงข่ายประสาทเทียม คือ ใช้เวลาในการประมวลผลเป็นเวลานาน

1.1.5 การประยุกต์ใช้งานต้นไม้การตัดสินใจ

ต้นไม้การตัดสินใจเป็นขั้นตอนวิธีการเรียนรู้ที่มีการอนุมานความรู้ (Inference Engine) ที่ได้ในรูปของกฎ “ถ้า...แล้ว...” ซึ่งเข้าใจได้ง่าย และมีความเหมาะสมสำหรับการแก้ปัญหาซึ่งมีข้อมูล 2 กลุ่ม และมีข้อมูลที่ไม่มีความต่อเนื่อง ในการแก้ปัญหาทำโดยเลือกคุณสมบัติของข้อมูล (Attributes) ที่มีคุณภาพดี สำหรับสร้างเป็นโหนดราก (Root Node) และเลือกคุณสมบัติของข้อมูลที่มีคุณภาพรองลงมา สำหรับสร้างโหนดลูกของโหนดรากในชั้นต่อๆ ไป และสิ้นสุดที่โหนดใบ (Leaf Node) ซึ่งที่โหนดใบจะเป็นจุดสิ้นสุดของการแบ่งแยกข้อมูลและใช้สำหรับแสดงคำตอบของปัญหา (Rich and Knight, 1991; Duda *et al.*, 2000; Mitra and Acharaya, 2003; Roiger and Greatz, 2003)

ในงานทางด้านชีวสารสนเทศนั้นมีการนำต้นไม้การตัดสินใจมาประยุกต์ใช้สำหรับการทำนายตำแหน่งเริ่มต้นการแปลรหัสโปรตีน (Protein Coding Regions) บนสายพันธุกรรมของมนุษย์ (Salzberg, 1995) การสร้างระบบตัดสินใจสำหรับค้นหายีนบนสายพันธุกรรม (Salzberg *et al.*, 1998) การแก้ปัญหาทางด้านจีโนมด้วยต้นไม้การตัดสินใจ (Browne *et al.*, 2003) การทำนายคู่ลำดับโปรตีนแบบ Co-Complex (Co-Complexed Protein Pairs) จากการมีปฏิสัมพันธ์กันระหว่างโปรตีนกับโปรตีนเพื่อศึกษาชนิด คุณสมบัติ และหน้าที่ของโปรตีนที่มีอยู่ในเซลล์ (Zhang *et al.*, 2004) เทคนิคการรวบรวมข้อมูลแบบอัตโนมัติด้วยต้นไม้การตัดสินใจสำหรับการแบ่งกลุ่มข้อมูลสายพันธุกรรมในระดับ Family, Superfamily และ Fold (Camoglu *et al.*, 2005) การศึกษาลักษณะของ Single Nucleotide Polymorphisms (SNPs) ซึ่งบ่งชี้ลักษณะการก่อโรคในผู้ป่วยมะเร็งหลอดอาหาร (Esophageal Cancer) (Xie *et al.*, 2005) การทำนายตำแหน่งของการเกิดโรคบนสายโปรตีนด้วยต้นไม้การตัดสินใจร่วมกับเทคนิคการลดรูปแบบของกรดอะมิโน (Han *et al.*, 2006) การประยุกต์ใช้ต้นไม้การตัดสินใจสำหรับศึกษาโครงสร้างของ DNA-Binding ที่มีความสัมพันธ์ต่อการจำแนกกลุ่มข้อมูล จากการเกิดปฏิกริยา Basic Helix–Loop–Helix ของโปรตีน (Atchley and Zhao, 2007) และใช้ต้นไม้การตัดสินใจแบบอัตโนมัติสำหรับทำนายรูปแบบการมีปฏิสัมพันธ์ระหว่างโปรตีนกับโปรตีน (Darnell *et al.*, 2007) การประยุกต์ใช้ต้นไม้การตัดสินใจแบบ Co-evolving สำหรับหาตำแหน่งเริ่มต้นการแปลรหัสโปรตีน (Protein Coding Regions) บนสายพันธุกรรมของมนุษย์ (Aitkenhead, 2008) เป็นต้น

จะเห็นได้ว่า ลักษณะเด่นของต้นไม้การตัดสินใจ คือ ใช้เวลาในการประมวลผลน้อยกว่าการใช้โครงข่ายประสาทเทียม และความรู้ที่ได้จากต้นไม้การตัดสินใจจะเกิดจากการอนุมานความรู้ด้วยค่าคุณสมบัติของข้อมูล (Attributes) แต่ละตัว ข้อจำกัดของการใช้ต้นไม้การตัดสินใจ คือ หากมีจำนวนคุณสมบัติของข้อมูลน้อยเกินไปประสิทธิภาพของการทำงานจะลดลง และหากมีจำนวนมากต้นไม้การตัดสินใจที่ได้จะมีขนาดใหญ่และมีจำนวนกฎ (Rules) สำหรับใช้งานจำนวนมาก

1.2 วัตถุประสงค์ของวิทยานิพนธ์

1.2.1 ออกแบบแบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

1.2.2 พัฒนาโปรแกรมจากแบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

1.3 ขอบเขตของวิทยานิพนธ์

1.3.1 พัฒนาขั้นตอนวิธีและสร้างแบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

1.3.2 พัฒนาโปรแกรมวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

1.3.3 ข้อมูลที่นำมาใช้สำหรับการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม เป็นข้อมูลสายโปรตีนของผู้ป่วยธาลัสซีเมีย เบาหวาน และดาวน์ซินโดรม รวบรวมได้จาก National Center for Biotechnology Information (NCBI)

1.4 ขั้นตอนและระยะเวลาการดำเนินงาน

1.4.1 ขั้นตอนการดำเนินการ

1.4.1.1 ศึกษางานวิจัยและเอกสารที่เกี่ยวข้อง

1.4.1.2 ศึกษาเทคนิคการวิเคราะห์สายพันธุกรรมและเทคนิคอื่นๆ ที่ใช้ในงานวิจัย

1.4.1.3 ศึกษาเครื่องมือและซอฟต์แวร์ที่ใช้สำหรับงานวิจัย

1.4.1.4 วิเคราะห์และออกแบบโปรแกรมวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

1.4.1.5 พัฒนาโปรแกรมวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

1.4.1.6 ทดสอบและติดตั้งโปรแกรมวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

1.4.1.7 จัดทำเอกสารประกอบโปรแกรมวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

1.4.1.8 เขียนผลงานวิจัยและนำเสนอผลงานวิจัย

1.4.1.9 จัดทำเอกสารวิทยานิพนธ์และสอบวิทยานิพนธ์

1.4.2 ระยะเวลาดำเนินการวิจัย

พฤศจิกายน 2549 – เมษายน 2551

1.4.3 แผนการดำเนินการวิจัย

ตารางที่ 1.1 ระยะเวลาการดำเนินการวิจัย

กิจกรรม / ขั้นตอนการดำเนินงาน	ช่วงเวลาดำเนินการ																	
	2549		2550												2551			
	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4
1. ศึกษางานวิจัยและเอกสารที่เกี่ยวข้อง	■	■																
2. ศึกษาเทคนิคการวิเคราะห์สายพันธุกรรมและเทคนิคอื่นๆ ที่ใช้ในงานวิจัย	■	■	■															
3. ศึกษาเครื่องมือและซอฟต์แวร์ที่ใช้สำหรับงานวิจัย		■	■	■														
4. วิเคราะห์และออกแบบโปรแกรมวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม				■	■	■	■											
5. พัฒนาโปรแกรมวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม						■	■	■	■	■	■	■						
6. ทดสอบและติดตั้งโปรแกรมวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม									■	■	■	■	■	■				
7. จัดทำเอกสารประกอบโปรแกรมวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม													■	■	■			
8. เขียนผลงานวิจัยและนำเสนอผลงานวิจัย							■	■	■	■	■	■	■	■	■	■		
9. จัดทำเอกสารวิทยานิพนธ์และสอบวิทยานิพนธ์													■	■	■	■	■	■

1.5 สถานที่และเครื่องมือที่ใช้

1.5.1 สถานที่ทำการวิจัย

ห้องปฏิบัติการวิจัยปัญญาประดิษฐ์ CS207 อาคารภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่

1.5.2 เครื่องมือที่ใช้

1.5.2.1 ด้านเครื่องมือและอุปกรณ์

1.5.2.1.1 ฮาร์ดแวร์

เครื่องคอมพิวเตอร์ส่วนบุคคล หน่วยความจำ 1 กิกะไบต์ ฮาร์ดดิสก์ 80 กิกะไบต์ หน่วยประมวลผลกลางรุ่น Intel Pentium 4 ความเร็วในการประมวลผล 3.0 กิกะเฮิร์ต สำหรับพัฒนาโปรแกรมและทดสอบแบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม จำนวน 1 เครื่อง

1.5.2.1.2 ซอฟต์แวร์

1) Microsoft Windows XP เป็นระบบปฏิบัติการ

2) MATLAB 2007a สำหรับพัฒนาโปรแกรมจากแบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรมและทดสอบแบบจำลอง

3) Adobe Photo Shop CS สำหรับตกแตกหน้าจอโปรแกม

4) WEKA เวอร์ชัน 3.4.10 สำหรับเปรียบเทียบผลการทดลอง

5) Microsoft Office 2007 สำหรับจัดทำเอกสารวิทยานิพนธ์

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1.6.1 ได้แบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

1.6.2 ได้โปรแกรมวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

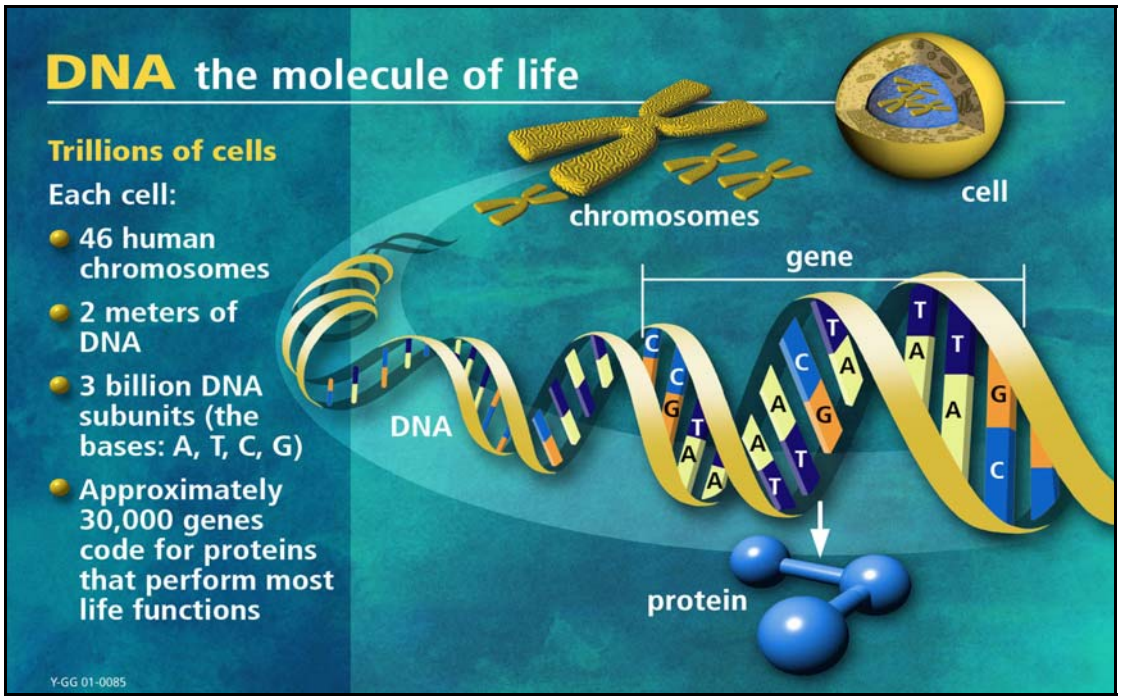
บทที่ 2

ทฤษฎีที่เกี่ยวข้องกับการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

ทฤษฎีที่นำมาใช้ในการสร้างแบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม ประกอบด้วย จีโนม ฟาส์ท-เอ เอ็นแกรม โครงข่ายประสาทเทียม ต้นไม้การตัดสินใจ การทดสอบประสิทธิภาพแบบ k-Fold Cross Validation และการวิเคราะห์ประสิทธิภาพ ซึ่งแสดงรายละเอียดได้ดังนี้

2.1 จีโนม

จีโนม (Genome) หมายถึง หน่วยพันธุกรรมหรือยีนทั้งหมดที่ควบคุมการแสดงลักษณะทางพันธุกรรมของสิ่งมีชีวิต ข้อมูลทางพันธุกรรมถูกบรรจุอยู่ภายในโครโมโซม (Chromosomes) โดยโครโมโซมจะประกอบอยู่ภายในนิวเคลียส (Nucleus) ของเซลล์ (Cell) ซึ่งเซลล์นับเป็นหน่วยย่อยที่เล็กที่สุดภายในระบบของสิ่งมีชีวิต โดยเมื่อเซลล์หลายๆ เซลล์มารวมตัวกัน จะมีทำหน้าที่หรือลักษณะเฉพาะอย่าง เช่น เป็นเนื้อเยื่อ ระบบกล้ามเนื้อ เป็นต้น สามารถแสดงโครงสร้างของหน่วยพันธุกรรมได้ดังภาพประกอบ 2.1



ภาพประกอบ 2.1 โครงสร้างสารพันธุกรรม (DNA) ของมนุษย์

(ที่มา: Human Genome Program in U.S. Department of Energy, 1992)

การศึกษาข้อมูลทางด้านชีวภาพในระดับโมเลกุลนั้น จะศึกษาในองค์ประกอบส่วนต่างๆ ดังต่อไปนี้

2.1.1 ดีเอ็นเอ (DNA) และอาร์เอ็นเอ (RNA)

ดีเอ็นเอ (DNA) เป็นชื่อย่อของสารพันธุกรรม ที่มีชื่อวิทยาศาสตร์ว่า “กรดดีออกซีไรโบนิวคลีอิก” (Deoxyribonucleic Acid: DNA) ซึ่งเป็นกรดนิวคลีอิกที่พบในใจกลางของเซลล์ทุกชนิด เช่น มนุษย์ สัตว์ พืช เชื้อรา แบคทีเรีย ไวรัส เป็นต้น ผู้ค้นพบ ดีเอ็นเอ คือ ฟรีดิช มีสเชอร์ ในปี พ.ศ. 2412 (ค.ศ. 1869) แต่ไม่ทราบว่ามีโครงสร้างอย่างไร จนกระทั่งในปี พ.ศ. 2496 (ค.ศ. 1953) เจมส์ ดี. วัตสัน และ ฟรานซิส คริก เป็นผู้ค้นพบโครงสร้างของดีเอ็นเอ และนับเป็นจุดเริ่มต้นของยุคเทคโนโลยีทางดีเอ็นเอ ดีเอ็นเอจะบรรจุข้อมูลทางพันธุกรรมของสิ่งมีชีวิตชนิดนั้นไว้ ซึ่งมีลักษณะที่เกิดจากการผสมผสานมาจากลักษณะพันธุกรรมของสิ่งมีชีวิตรุ่นบรรพบุรุษ เช่น ปู่และย่า ตาและยาย หรือพ่อและแม่ และถ่ายทอดลักษณะทางพันธุกรรมซึ่งเกิดการผสมผสานไปยังสิ่งมีชีวิตรุ่นถัดไป คือ ลูกและหลาน

จากภาพประกอบ 2.1 ลักษณะโครงสร้างของดีเอ็นเอมีรูปร่างเป็นสายเกลียวคู่ (Double Helix) แต่ละข้างเป็นการเรียงตัวของลำดับนิวคลีโอไทด์ (Nucleotide) โดยนิวคลีโอไทด์เป็นโมเลกุลซึ่งประกอบด้วยน้ำตาล ฟอสเฟต และเบส (หรือด่าง) ภายในลำดับนิวคลีโอไทด์มีเบสพื้นฐาน 4 ชนิด ได้แก่ อะดีนีน (A) ไทมีน (T) ไซโตซีน (C) และกัวนีน (G) ตามลำดับ นิวคลีโอไทด์ทั้งสองข้างจะถูกเชื่อมด้วยเบส โดยที่เบส A จะเชื่อมกับเบส T หรือ {A-T, T-A} และเบส C จะเชื่อมกับ เบส G หรือ {C-G, G-C} เท่านั้น (ในกรณีของดีเอ็นเอ)

ภายในระบบการทำงานของระบบเซลล์ภายในสิ่งมีชีวิต จะมีการถอดรหัสพันธุกรรม เพื่อนำข้อมูลจากดีเอ็นเอไปสังเคราะห์ ผลลัพธ์จากการสังเคราะห์ คือ โปรตีนชนิดต่างๆ ซึ่งเกิดจากการเรียงตัวของโมเลกุลกรดอะมิโนชนิดต่างๆ (จะกล่าวในหัวข้อถัดไป) โดยเมื่อการถอดรหัสเริ่มขึ้น เอนไซม์ (โปรตีนซึ่งทำหน้าที่เป็นตัวเร่งปฏิกิริยา) จะเปิดสายโมเลกุลเกลียวคู่ให้แยกจากกันตามจุดที่ต้องการ และใช้เอนไซม์อีกชนิดหนึ่งสร้างสาย “เมสเซนเจอร์อาร์เอ็นเอ” (Messenger RNA: mRNA) เข้าไปประกบคู่กับสายดีเอ็นเอข้างหนึ่งโดยการเชื่อมเบสนิวโตรจีเนียสเข้าด้วยกัน แต่เบสใน RNA มีเพียง A, C และ G โดยจะไม่มี T (ไทมีน) แต่มี U หรือยูราซิล (Urasil) แทน (วิสุทธ์ ใบไม้, 2536; มนตรี จุฬาวัฒนทล และคณะ, 2543; อุไรวรรณ วิจารณกุลม, 2545; สุนทร โสตถิพันธุ์ และคณะ. 2548)

2.1.2 กรดอะมิโน (Amino Acid) และโปรตีน (Proteins)

2.1.2.1 กรดอะมิโน (Amino Acid)

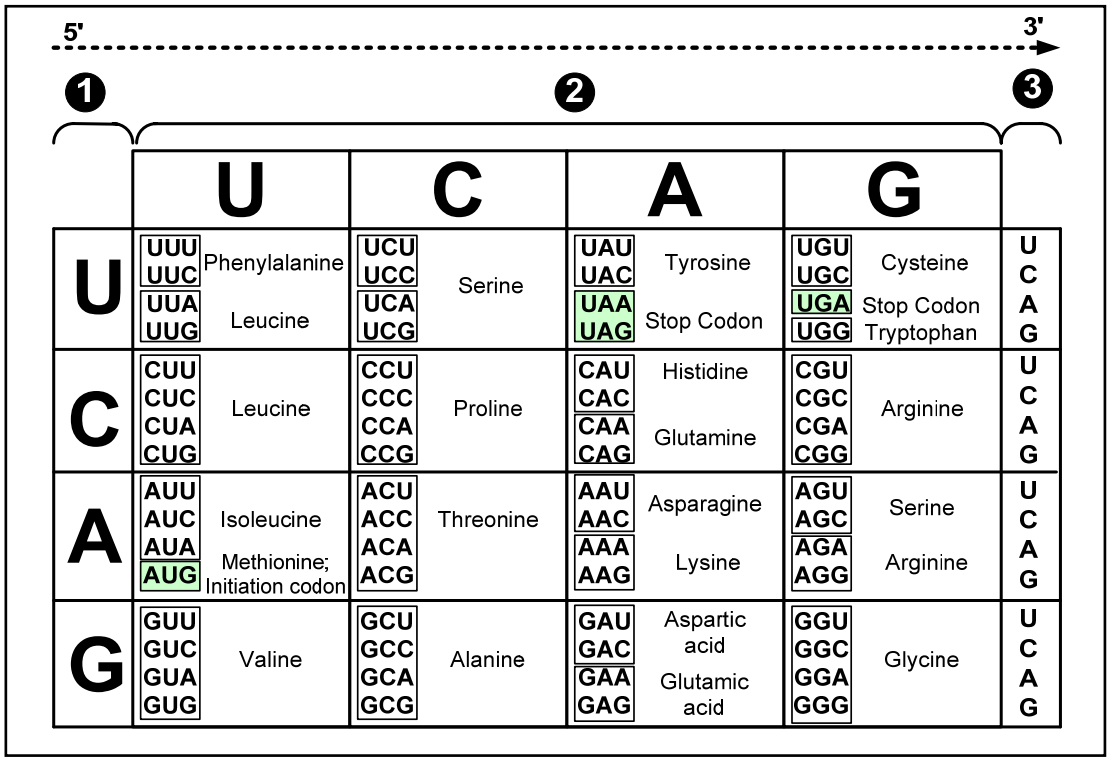
กรดอะมิโน คือ โมเลกุลที่เกิดจากการจับตัวกันของเบสทั้ง 4 ชนิด (A, T, G และ C) ซึ่งมีหมู่ที่เป็นทั้งหมู่ฟังก์ชันอะมิโนและคาร์บอกซิลเป็นส่วนประกอบติดอยู่กับคาร์บอนอะตอมเดียวกัน เรียกว่า “แอลฟา-คาร์บอน” กรดอะมิโนเป็นองค์ประกอบสำคัญของการสังเคราะห์โปรตีนภายในสิ่งมีชีวิต กรดอะมิโนมักหมายความถึง กรดอะมิโนแบบแอลฟา (Alpha Amino Acid) โดยกรดอะมิโนมาตรฐาน (Standard Amino Acid) มีทั้งสิ้น 20 ชนิด และสามารถแสดงรายชื่อได้ดังภาพประกอบ 2.2

ที่	ชื่อเต็ม	ชื่อย่อ	
1	อะลานีน (Alanine)	Ala	A
2	ซิสเทอีน (Cysteine)	Cys	C
3	กรดแอสปาร์ติก (Aspartic acid)	Asp	D
4	กรดกลูตามิก (Glutamic acid)	Glu	E
5	ฟีนิลอะลานีน (Phenylalanine)	Phe	F
6	ไกลซีน (Glycine)	Gly	G
7	ฮีสติดีน (Histidine)	His	H
8	ไอโซลูซีน (Isoleucine)	Ile	I
9	ไลซีน (Lysine)	Lys	K
10	ลูซีน (Leucine)	Leu	L

ที่	ชื่อเต็ม	ชื่อย่อ	
11	เมตไทโอนีน (Methionine)	Met	M
12	แอสปาราจีน (Asparagine)	Asn	N
13	โปรลีน (Proline)	Pro	P
14	กลูตามีน (Glutamine)	Gln	Q
15	อาร์จินีน (Arginine)	Arg	R
16	เซอรีน (Serine)	Ser	S
17	ทรีโอนีน (Threonine)	Thr	T
18	วาลีน (Valine)	Val	V
19	ทริปโตแฟน (Tryptophan)	Trp	W
20	ไทโรซีน (Tyrosine)	Tyr	Y

ภาพประกอบ 2.2 รายชื่อกรดอะมิโนมาตรฐาน 20 ชนิด
(ดัดแปลงจาก: สุนทร โสตถิพันธุ์ และคณะ. 2548)

จากรายชื่อกรดอะมิโนมาตรฐานทั้ง 20 ชนิด ในภาพประกอบ 2.2 และการแปลรหัสพันธุกรรม (Genetic Code) ซึ่งเป็นการเรียงตัวของเบสทั้ง 4 ชนิด เกิดการจับตัวของเบสครั้งละ 3 ชนิดหรือเรียกว่า “โคดอน” (Codon) ในภาพประกอบ 2.3 แสดงรายละเอียดของการเรียงตัวและรูปแบบการแปลรหัสทางด้านพันธุกรรมในระดับนิวคลีโอไทด์



ภาพประกอบ 2.3 รหัสพันธุกรรม (Genetic Code)
(ดัดแปลงจาก: สุนทร โสตถิพันธุ์ และคณะ. 2548)

จากภาพประกอบ 2.3 จะเห็นว่า ในการแปลรหัสพันธุกรรมเบสไทมีน (T) จะถูกแทนด้วยสัญลักษณ์ U หรือยูราซิล (Urasil) และการอ่านรหัสพันธุกรรมจะอ่านเรียงจากซ้าย (ตำแหน่งที่ 1) ไปทางขวา (สิ้นสุด ณ ตำแหน่งที่ 3) เช่น ถ้าสัญลักษณ์ตัวแรก คือ U (Urasil) สัญลักษณ์ตัวที่สอง คือ U และตัวสุดท้าย คือ U จะได้รหัสพันธุกรรม คือ UUU ซึ่งเป็นรหัสพันธุกรรมของกรดอะมิโนชื่อฟีนิลอะลานีน (Phenylalanine: F)

2.1.2.2 โปรตีน (Proteins)

โปรตีน คือ โมเลกุลซึ่งมีขนาดใหญ่เกิดขึ้นจากการเชื่อมต่อกันของกรดอะมิโนตั้งแต่ 2 หน่วยขึ้นไปด้วยพันธะเปปไทด์ (Peptide Bond) และสามารถจำแนกประเภทตามลักษณะการเชื่อมต่อกันของกรดอะมิโนในลักษณะต่างๆ ดังนี้

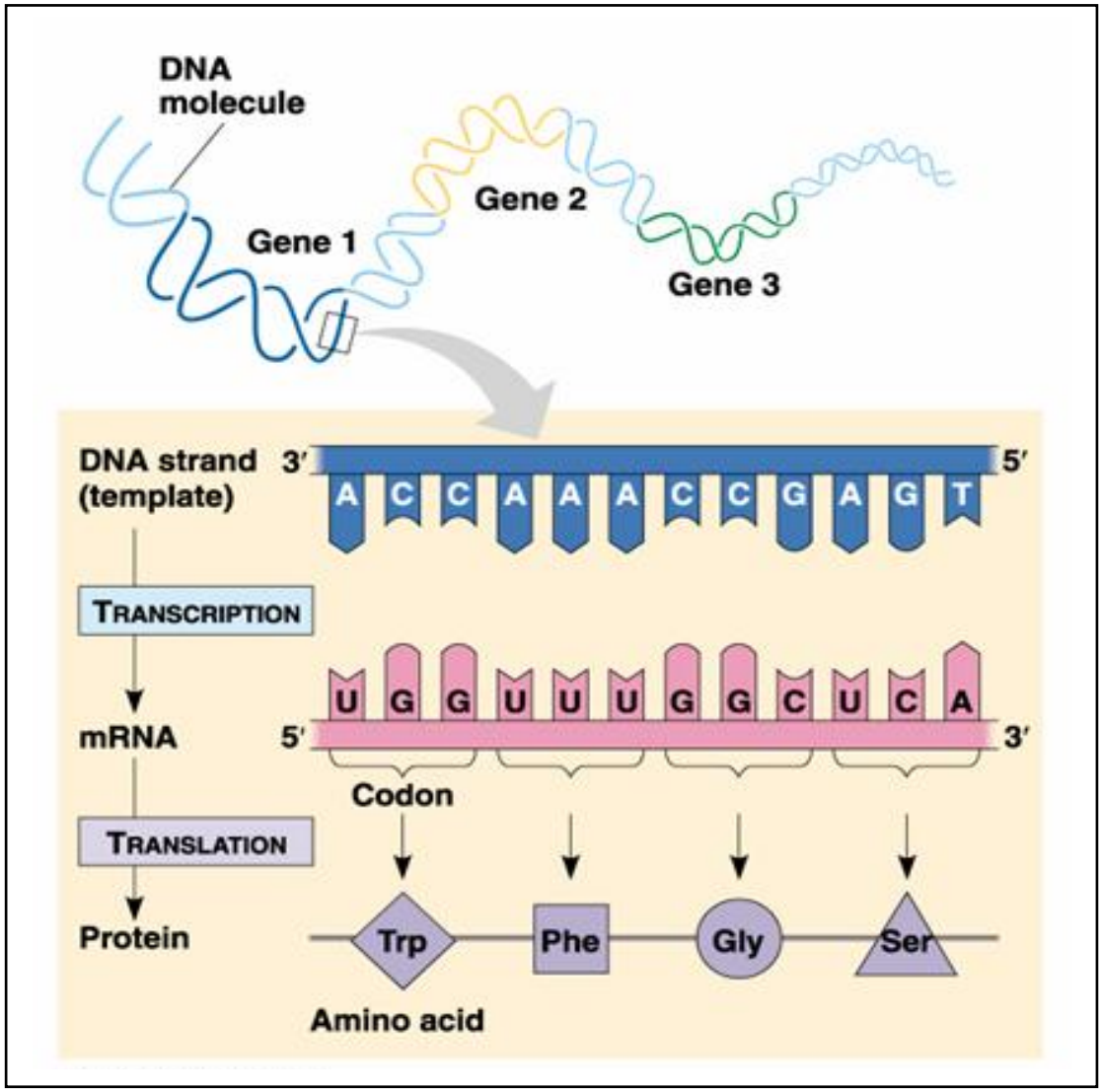
1) ไดเปปไทด์ (Dipeptide) เกิดจากการเชื่อมต่อระหว่างกรดอะมิโน 2 ตัวมาเชื่อมกัน โดยมีหมู่อะมิโน (Amino: $–NH_2$) ซึ่งมีคุณสมบัติเป็นเบสของกรดอะมิโนตัวหนึ่งไปเกิดพันธะกับหมู่คาร์บอกซิล (Carboxyl: –COOH) ซึ่งมีคุณสมบัติเป็นกรดของกรดอะมิโนอีกตัวหนึ่ง

2) ไตรเปปไทด์ (Tripeptide) เกิดจากการเชื่อมต่อระหว่างกรดอะมิโนด้วยพันธะเปปไทด์ 2 พันธะ

3) โพลีเปปไตด์ (Polypeptide) เกิดจากการเชื่อมต่อระหว่างกรดอะมิโนด้วยพันธะเปปไตด์เป็นโซ่ยาว (Long Chain) แต่มีจำนวนน้อยกว่า 50 หน่วย

4) โปรตีน (Proteins) เป็นโพลีเปปไตด์โซ่ยาวที่มีจำนวนของหน่วยกรดอะมิโนที่เชื่อมต่อด้วยพันธะเปปไตด์มากกว่า 50 หน่วยขึ้นไป

จากที่กล่าวมาในข้างต้น สามารถสรุปกระบวนการของการถอดรหัสพันธุกรรม (Transcription) และการแปลรหัสพันธุกรรม (Translation) ของกรดอะมิโนและโปรตีน ได้ดังภาพประกอบ 2.4



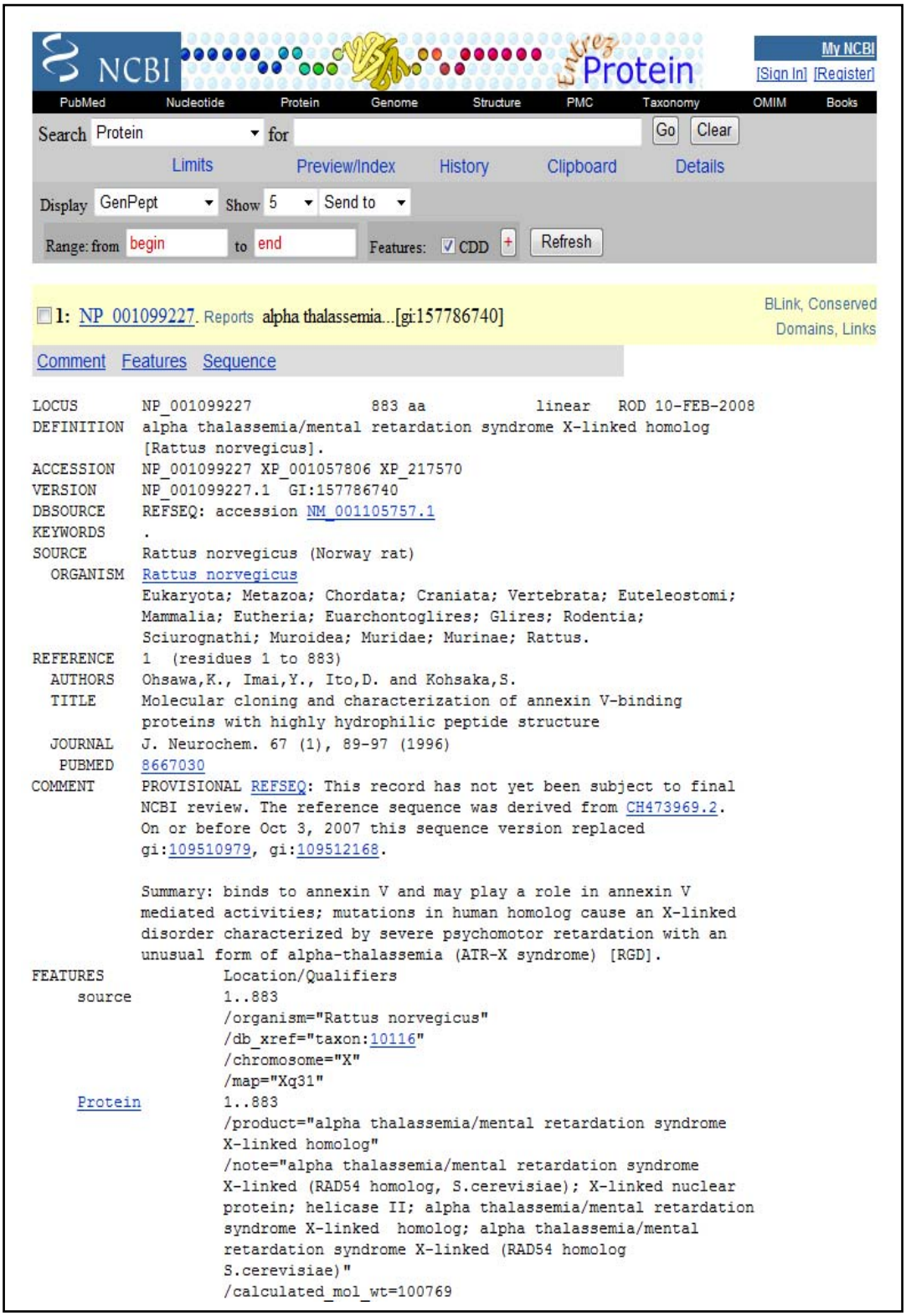
ภาพประกอบ 2.4 กระบวนการถอดรหัสพันธุกรรมของกรดอะมิโนและโปรตีน
(ที่มา: Addison Wesley Longman, 1999)

2.2 ฟาสท์-เอ

ฟาสท์-เอ (FASTA) ย่อมาจากคำว่า FAST-All โดยคำว่า “All” หมายถึง การรวมทั้งในการจัดเรียง FAST-P (โปรตีน) และ FAST-N (นิวคลีโอไทด์) โดย FASTA Format จะจัดเก็บข้อมูลแบบไฟล์ข้อความ (TEXT File) นามสกุลของ FASTA Format คือ *.fasta ซึ่งเก็บข้อมูลที่เป็นลำดับนิวคลีโอไทด์ (Nucleotide) หรือ ลำดับโปรตีน (Proteins Sequences) โครงสร้างการเก็บข้อมูลจะประกอบด้วย 2 ส่วนหลัก คือส่วนหัวของสายพันธุกรรม (Header of Sequences) โดยจะขึ้นต้นด้วยสัญลักษณ์ “>” ตามด้วยชื่อสายพันธุกรรม (Sequence Name) และส่วนของข้อมูลหรือลักษณะของข้อมูลสายพันธุกรรม (Sequence Detail) จะอยู่ในบรรทัดใหม่ (Claveric and Notredame, 2003; Golding and Morton, 2006) สามารถแสดงได้ดังภาพประกอบ 2.5

ภาพประกอบ 2.5 ลักษณะโครงสร้างการจัดเก็บข้อมูลแบบ FASTA

จากภาพประกอบ 2.5 แสดงตัวอย่างการเก็บข้อมูลแบบ FASTA ของสายพันธุกรรม gi|157786740|ref|NP_001099227.1|[157786740] ซึ่งประกอบด้วย ส่วนหัวของสายพันธุกรรมจะระบุหมายเลขสายพันธุกรรม (Sequence ID) คือ “gi|157786740” รหัสอ้างอิง (Accession Number) คือ “NP_001099227.1” และคำอธิบายสายพันธุกรรม (Description) คือ “alpha thalassemia/mental retardation syndrome X-linked homolog [Rattus norvegicus]” และสามารถดาวน์โหลดรายละเอียดของข้อมูลได้จากฐานข้อมูล NCBI ดังภาพประกอบที่ 2.6



ภาพประกอบ 2.6 ตัวอย่างรายละเอียดของสายพันธุกรรม

(ดัดแปลงจาก: National Center for Biotechnology Information, 2007)

```
     Region          15..282
                     /region_name="SNF2_N"
                     /note="SNF2 family N-terminal domain. This domain is found
                     in proteins involved in a variety of processes including
                     transcription regulation (e.g., SNF2, STH1, brahma, MOT1),
                     DNA repair (e.g., ERCC6, RAD16, RAD5), DNA recombination
                     (e.g; pfam00176"
                     /db_xref="CDD:84584"
     Region          <15..143
                     /region_name="DEXHc"
                     /note="DEXH-box helicases. A diverse family of proteins
                     involved in ATP-dependent DNA or RNA unwinding, needed in
                     a variety of cellular processes. The name derives from the
                     sequence of the Walker  B motif (motif II). This domain
                     contains the ATP-binding region; cd00269"
                     /db_xref="CDD:29341"
     Site            112..115
                     /site_type="other"
                     /note="Mg++ binding site"
                     /db_xref="CDD:29341"
     Site            140..142
                     /site_type="other"
                     /note="motif III"
                     /db_xref="CDD:29341"
     Region          408..554
                     /region_name="HELICc"
                     /note="Helicase superfamily c-terminal domain; associated
                     with DEXDc-, DEAD-, and DEAH-box proteins, yeast
                     initiation factor 4A, Ski2p, and Hepatitis C virus NS3
                     helicases; this domain is found in a wide variety of
                     helicases and helicase related proteins; may no; cd00079"
                     /db_xref="CDD:28960"
     Site            order(433..436,476..477,508..510)
                     /site_type="other"
                     /note="nucleotide binding region"
                     /db_xref="CDD:28960"
     Site            order(516,537,541,544)
                     /site_type="other"
                     /note="ATP-binding site"
                     /db_xref="CDD:28960"
     CDS             1..883
                     /gene="Atrx"
                     /coded_by="NM_001105757.1:4813..7464"
                     /db_xref="GeneID:246284"
                     /db_xref="RGD:619795"
ORIGIN
        1 myscplhgpw qdpaalvvcp lntalnwmne fekwqeglnd aeklevsela tvkrpqersy
       61 mlqrwqedgg vmiigyemyr nlaqgrnvks rklkeifnka lvdpgpdfvv cdeghilkne
      121 asavskamns iksrrriilt gtplqnnlie yhcmvnfike nllgsikefr nrfinpiqng
      181 qcadstmvdv rvmkkrahil yemlagcvqr kdytaltkfl ppkheyvlav rmtaiqckly
      241 qyyldhltgv gnstdggrgk agaklfqdfq mlsriwthpw clqldyiske nkgyfdedsm
      301 defiasdsde tsmslssdek kktkgkkgrk dssssgsgsd ndvevikvwn srsrgggegn
      361 veetgnnpsv slkldesktt stsnpsspap dwykdfvtda daevlehsgk mvllfeilrm
      421 aeeigdkvlv fsqslisldl iedflelasr ektedkdkpl iykgegkwlr nidyyrldgs
      481 tnaqsrkkwa eefndetnvr grlfiistka gslginlvaa nrviifdasw npsydiqsif
      541 rvyrfgqtkp vyvyrflaqg tmedkiydrq vtkqslsfrv vdqqqverhf tmneltelyt
      601 fepdllddpn sekkkkrdtp mlpkdtilae llqihkehiv gyhehdslld hkeeeeltee
      661 erkaawaeye aekkgltmrf niptgtnlpp vtftsqtpyi pfnlgalsam snqqledlin
      721 qgrekvveat nsvtsvriqp lediistvwk enmnlseaqv qalalsrqas qeldvkrrea
      781 iyndvltkqq mliscvqril mnrrlqqqyt qqqqqqltyq qatlshlmmp kppnlimtps
      841 nyqqidmrgm yqsvaggmqp pplqrapppm rsknpgpspg ksm
```

ภาพประกอบ 2.6 ตัวอย่างรายละเอียดของสายพันธุกรรม (ต่อ)

(ดัดแปลงจาก: National Center for Biotechnology Information, 2007)

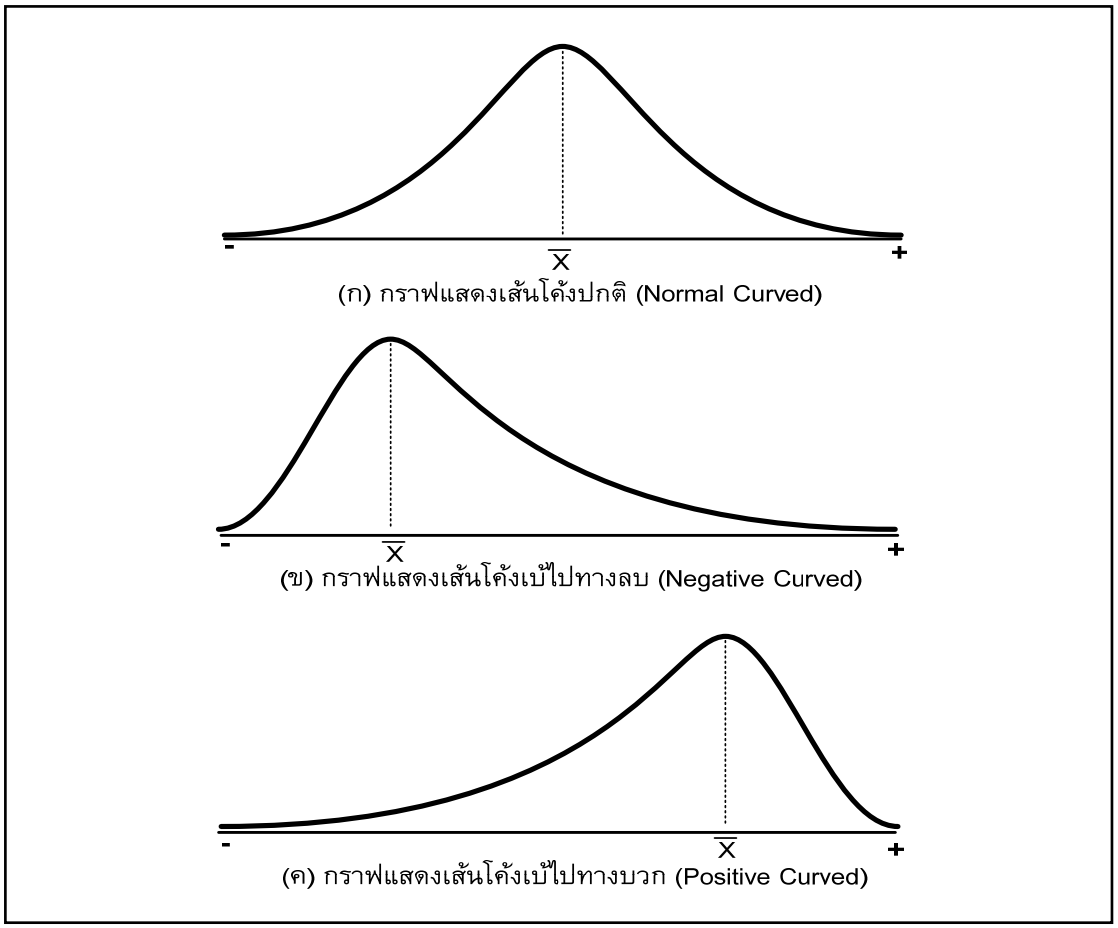
2.3 การทำความสะอาดข้อมูล

การประมวลผลข้อมูลทางด้านชีวสารสนเทศ อาจมีสายพันธุกรรมที่ใช้มีความยาวสายอยู่ในช่วงตั้งแต่ 2 ตัวอักษรขึ้นไปจนถึงล้านตัวอักษร ทำให้เกิดความแตกต่างกันของข้อมูลใช้นั้น การเลือกตัวอย่าง (Sampling) ที่มีความยาวใกล้เคียงกันนั้น จำเป็นต้องหาตัวแทนของกลุ่มข้อมูล ค่าทางสถิติที่นิยมใช้ คือ ค่าเฉลี่ยเลขคณิต (Mean Value) สำหรับหาศูนย์กลางของข้อมูล กำหนดให้ $\overline{X}$ คือ ค่าเฉลี่ยของข้อมูลที่ใช้ สามารถคำนวณดังสมการ (2.1)

$$\overline{X} = \frac{\sum fx}{N} \tag{2.1}$$

เมื่อ ตัวแปร fx คือ ค่าข้อมูล และตัวแปร N คือ จำนวนของข้อมูลทั้งหมดที่ใช้

จากสมการ (2.1) สามารถอธิบายลักษณะของกราฟดังภาพประกอบ 2.7



ภาพประกอบ 2.7 กราฟแนวโน้มของค่าเฉลี่ย

จากภาพประกอบ 2.7 สามารถอธิบายกราฟได้ ดังนี้

1) กราฟโค้งปกติ (Normal Curved)

ภาพประกอบ 2.7 (ก) ลักษณะของกราฟที่ได้ตำแหน่งของค่าเฉลี่ยอยู่กึ่งกลางของกลุ่มข้อมูล ซึ่งมีความหมายว่า ค่าข้อมูลที่ใช้มีความแตกต่างกันน้อยหรือมีค่าที่ใกล้เคียงกัน

2) กราฟเบ้ไปทางลบ (Negative Curved)

ภาพประกอบ 2.7 (ข) ตำแหน่งของค่าเฉลี่ยอยู่เอียงไปทางลบหรือมีความหมายว่า ข้อมูลที่ใช้โดยรวมมีค่าค่อนข้างต่ำ

3) กราฟเบ้ไปทางบวก (Positive Curved)

ภาพประกอบ 2.7(ค) ตำแหน่งของค่าเฉลี่ยอยู่เอียงไปทางบวกหรือมีความหมายว่า ข้อมูลที่ใช้โดยรวมมีค่าค่อนข้างสูง

ในการสุ่มเลือกกลุ่มตัวอย่าง อาจมีการกำหนดเงื่อนไขสำหรับเลือก เช่น
เงื่อนไขที่ 1 เลือกตัวอย่างที่มีค่าของข้อมูลน้อยหรือเท่ากับค่าเฉลี่ย
เงื่อนไขที่ 2 เลือกตัวอย่างที่มีค่าของข้อมูลมากกว่าหรือเท่ากับค่าเฉลี่ย
เป็นต้น

ทั้งนี้อาจมีการปรับเปลี่ยนเงื่อนไข หรืออาจมีเงื่อนไขในการเลือกกลุ่มตัวอย่างได้มากกว่า 1 เงื่อนไข เพื่อให้ได้ข้อมูลที่มีคุณลักษณะที่เหมาะสมและสามารถเทียบเคียงกันได้ ทั้งนี้สิ่งที่ต้องคำนึงถึงคือ ข้อมูลกลุ่มตัวอย่างที่ได้จะต้องมีความสอดคล้องกับปัญหาที่ต้องการแก้ไข ซึ่งจะช่วยให้ผลลัพธ์หรือคำตอบที่ได้มีความแม่นยำมากยิ่งขึ้น (ชัชวาลย์ เรืองประพันธ์, 2541; กัลยา วานิชย์บัญชา, 2546)

2.4 เอ็นแกรม

การคำนวณหากลุ่มย่อยด้วยเทคนิคเอ็นแกรม (n-Grams Technique) เป็นเทคนิคที่ได้รับความนิยมเนื่องจากเป็นวิธีที่เรียบง่าย มีประสิทธิภาพสูง นิยมนำมาใช้ในการแก้ไขทางด้านการประมวลผลข้อความ (Text Processing) ซึ่งนำหลักการประมาณค่าความน่าจะเป็น (Probability) บางครั้งเรียกแบบจำลองมาร์คอฟ (Markov Model)

หลักการคำนวณกลุ่มย่อยที่เป็นไปได้ของข้อมูลด้วยเทคนิค n-Grams คือ การแบ่งข้อมูลที่ต้องการศึกษาออกเป็นส่วนย่อยๆ ที่มีขนาดเท่ากับ n ตัว โดยที่ n คือ ความกว้างของข้อมูลที่ต้องการแบ่ง ซึ่งสามารถเลือกใช้ขนาดของแกรมตั้งแต่ n=2 ตัว เรียก “ไบ-แกรม” (Bi-Grams), n=3 ตัว เรียก “ไตร-แกรม” (Tri-Grams), และ n=4 ตัว เรียก “ควอดิ-แกรม” (Quadi-Grams) เป็นต้น เมื่อกำหนดให้ P แทน ค่าความน่าจะเป็น สามารถเขียนรูปแบบของเอ็นแกรมได้ ดังนี้

กรณี 2-Grams คำนวณดังสมการ (2.2)

$$P(c_1c_2c_3\ldots c_n) = P(c_1)\,P(c_2|c_1)\,P(c_3|c_2)\ldots P(c_n|c_{n-1}) \quad (2.2)$$

กรณี 3-Grams คำนวณดังสมการ (2.3)

$$P(c_1c_2c_3\ldots c_n) = P(c_1)\,P(c_2|c_1)\,P(c_3|c_1c_2)\ldots P(c_n|c_{n-2}c_{n-1}) \quad (2.3)$$

กรณี 4-Grams คำนวณดังสมการ (2.4)

$$P(c_1c_2c_3\ldots c_n) = P(c_1)\,P(c_2|c_1)\,P(c_3|c_1c_2)\,P(c_4|c_1c_2c_3)\ldots P(c_n|c_{n-3}c_{n-2}c_{n-1}) \quad (2.4)$$

เมื่อ ตัวแปร c แทน อักขระหรือตัวอักษร และ $(c_1c_2c_3\ldots c_n)$ แทน ชุดอักขระที่ประกอบด้วยอักขระตั้งแต่ 3 ตัวขึ้นไป จนถึง n ตัว

การประมาณค่าความน่าจะเป็นของกลุ่มข้อมูลย่อย เกิดจากความน่าจะเป็นของชุดอักขระที่รวมกันเป็นกลุ่มข้อมูลย่อย $w_1w_2w_3\ldots.w_n$ โดยใช้สมมติฐานของมาร์คอฟ (Markov Assumption) ที่ว่า “การปรากฏของตัวอักษรตัวหนึ่งขึ้นกับตัวอักษรก่อนหน้าเพียง n-1 ตัว”

กำหนดให้ w แทน ค่าความน่าจะเป็นที่จะเกิดกลุ่มข้อมูลย่อย โดยสามารถคำนวณค่า w ได้ดังสมการ (2.5)

ค่าความน่าจะเป็นที่จะเกิดกลุ่มข้อมูลย่อย $w = P(w|h)$ (2.5)

เมื่อ P(w|h) คือ ค่าความน่าจะเป็นของกลุ่มข้อมูลย่อย w บนสายอักขระ ตัวแปร h แทน ข้อมูลทั้งหมดหรือสายอักขระทั้งหมดที่ใช้ สามารถเขียนสมการแทนการประมาณค่าความน่าจะเป็นของเอ็นแกรมแต่ละขนาดได้ ดังนี้

กรณี 2-Grams คำนวณดังสมการ (2.6)

$$P(w_1w_2w_3\ldots w_n) = P(w_1)P(w_2|w_1)P(w_3|w_2)\ldots P(w_n|w_{n-1}) \quad (2.6)$$

กรณี 3-Grams คำนวณดังสมการ (2.7)

$$P(w_1w_2w_3\ldots w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1w_2)\ldots P(w_n|w_{n-2}w_{n-1}) \quad (2.7)$$

กรณี 4-Grams คำนวณดังสมการ (2.8)

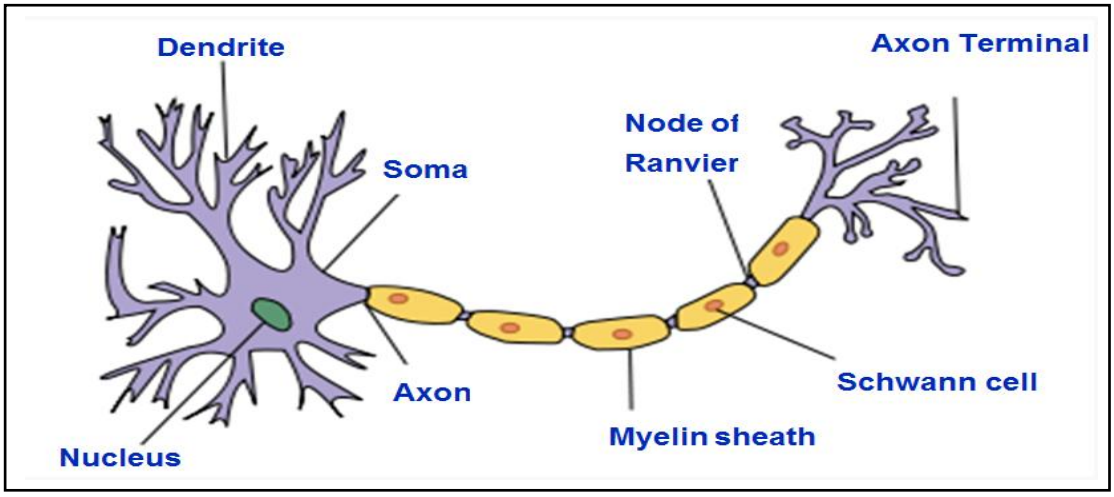
$$P(w_1w_2w_3\ldots w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1w_2)P(w_4|w_1w_2w_3)\ldots P(w_n|w_{n-3}w_{n-2}w_{n-1}) \quad (2.8)$$

สำหรับข้อมูลสายอักขระซึ่งมีขนาด N ตัว จะได้ว่า $w_1\ldots w_n$ หรือ w_1^n และเขียนรูปทั่วไปของการประมาณค่าความน่าจะเป็นของข้อมูล คือ $P(w_1^n)$ คำนวณดังสมการ (2.9)

$$\begin{aligned} P(w_1{}^n) &= P(w_1)P(w_2|w_1)P(w_n|w_1{}^{n-1}) \\ &= \prod_{k-1}^{n} P(w_n|w_1{}^{n-1}) \end{aligned} \quad (2.9)$$

2.5 โครงข่ายประสาทเทียม

โครงข่ายประสาทเทียม (Artificial Neural Networks) คือ โมเดลทางคณิตศาสตร์ สำหรับประมวลผลสารสนเทศด้วยการคำนวณแบบคอนเนคชันนิสต์ (Connectionist) เพื่อจำลองการทำงานของเครือข่ายประสาทในสมองมนุษย์ ด้วยวัตถุประสงค์ที่จะสร้างเครื่องมือซึ่งมีความสามารถในการเรียนรู้การจดจำรูปแบบ (Pattern Recognition) และการอุปมานความรู้ (Knowledge Deduction) เช่นเดียวกับความสามารถที่มีในสมองมนุษย์ แนวคิดเริ่มต้นของเทคนิคนี้ได้มาจากการศึกษาข่ายงานไฟฟ้าชีวภาพ (Bioelectric Networks) ในสมองซึ่งประกอบด้วย เซลล์ประสาทหรือนิวรอน (Neurons) และจุดประสานประสาท (Synapses) แต่ละเซลล์ประสาทประกอบด้วยปลายในการรับกระแสประสาท เรียกว่า "เดนไดรท์" (Dendrite) ซึ่งเป็นข้อมูลเข้า (Input) และปลายในการส่งกระแสประสาทเรียกว่า "แอคซอน" (Axon) ซึ่งเป็นเหมือนข้อมูลผลลัพธ์ (Output) ของเซลล์ซึ่งเซลล์เหล่านี้ทำงานด้วยปฏิกิริยาไฟฟ้าเคมี เมื่อมีการกระตุ้นด้วยสิ่งเร้าภายนอกหรือกระตุ้นด้วยเซลล์ด้วยกัน กระแสประสาทจะวิ่งผ่านเดนไดรท์เข้าสู่นิวเคลียสซึ่งจะเป็นตัวตัดสินว่าต้องกระตุ้นเซลล์อื่น ๆ ต่อหรือไม่ ถ้ากระแสประสาทแรงพอนิวเคลียสก็จะกระตุ้นเซลล์อื่น ๆ ต่อไป ผ่านทางแอคซอน จากรูปแบบการทำงานของเครือข่ายประสาท และเกิดการเชื่อมต่อระหว่างเซลล์ประสาทจนเป็นเครือข่ายที่สามารถทำงานร่วมกัน แสดงลักษณะของเครือข่ายประสาทในสมองมนุษย์ได้ดังภาพประกอบ 2.8 (Rich and Knight, 1991; Hagan *et al.*, 1996)



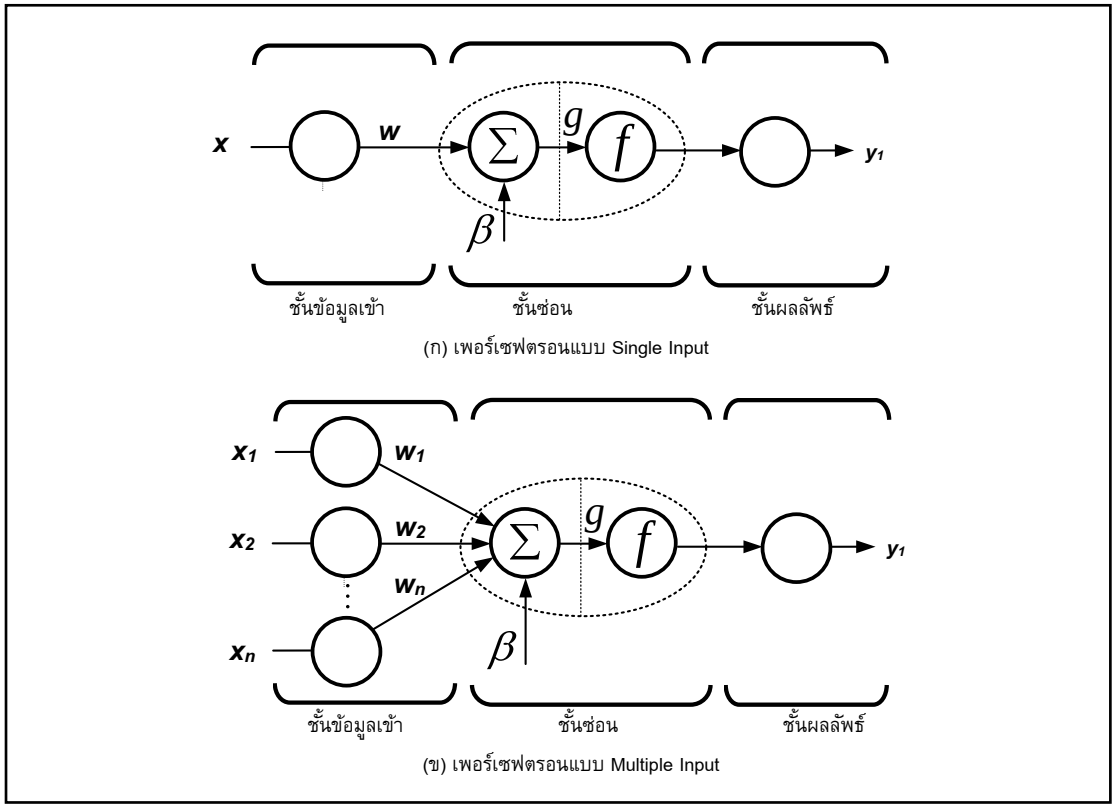
ภาพประกอบ 2.8 เครือข่ายประสาทในสมองมนุษย์
(ที่มา: Wikipedia, 2007)

เทคนิคโครงข่ายประสาทเทียมเป็นเทคนิคที่นิยมใช้ในสาขาปัญญาประดิษฐ์ (Artificial Intelligent) ซึ่งมีความสามารถในการเรียนรู้ที่คล้ายคลึงกับระบบสมองของมนุษย์ดังกล่าวข้างต้น ขั้นตอนของการนำเครือข่ายประสาทมาใช้สำหรับการพยากรณ์ก็มีลักษณะ

เช่นเดียวกับวิธีการพยากรณ์อื่น ๆ โดยอาศัยข้อมูลนำเข้าเพื่อสร้างแบบจำลองในการพยากรณ์ข้อมูลในอนาคตและมีความสามารถในการรวมการวิเคราะห์พื้นฐานและเทคนิคเพื่อสร้างแบบจำลอง โดยที่โครงข่ายประสาทจะพยายามลดจำนวนของการทำนายที่ผิดพลาดให้ต่ำที่สุด และเป็นเทคนิคที่นิยมใช้ในการแก้ปัญหาด้านปัญญาประดิษฐ์ (Artificial Intelligence) (Mitchell, 1997)

2.5.1 พื้นฐานของโครงข่ายประสาทเทียม (Basic of Neural Network)

พื้นฐานของโครงข่ายประสาทเทียม ประกอบด้วย 3 ส่วนหลัก คือ ชั้นข้อมูลเข้า (Input Layer) ชั้นซ่อน (Hidden Layer) และชั้นผลลัพธ์ (Output Layer) แสดงได้ดังภาพประกอบ 2.9 (Hagan *et al.*, 1996)

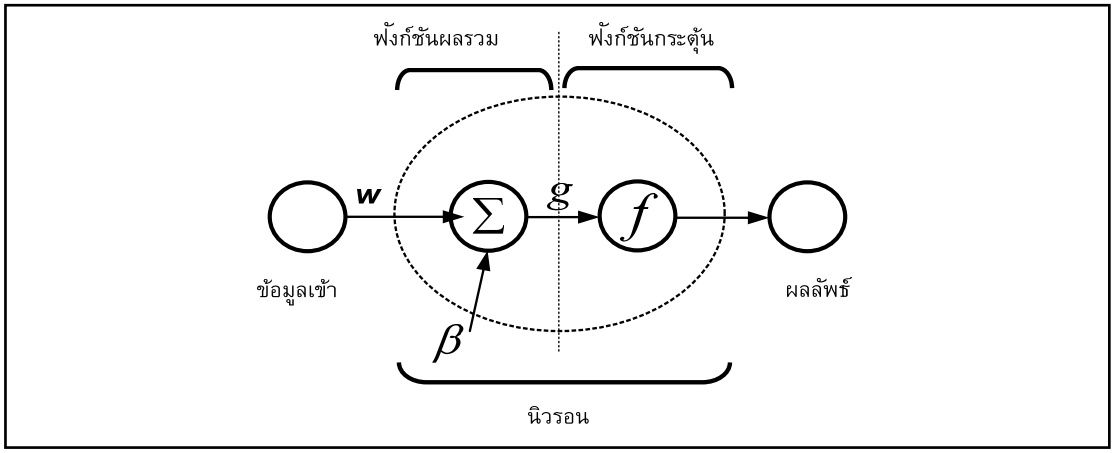


ภาพประกอบ 2.9 พื้นฐานของโครงข่ายประสาทเทียม
(ดัดแปลงจาก: Hagan *et al.*, 1996)

จากภาพประกอบ 2.9 จะเห็นว่าในชั้นซ่อน จะบรรจุนิวรอนตั้งแต่หนึ่งตัวขึ้นไป โดยนิวรอนจะทำหน้าที่ในการประมวลผลลัพธ์และส่งผลลัพธ์ที่ออกทางชั้นผลลัพธ์ ทั้งนี้นิวรอน 1 ตัว สามารถอ่านข้อมูลเข้าได้มากกว่า 1 ค่าข้อมูล โดยหากมีข้อมูลเข้าเพียงแค่ 1 ค่า

เรียกว่า “เพอร์เซฟตรอนแบบ Single Input” และกรณีที่มีค่าข้อมูลเข้ามากกว่าตั้งแต่ 2 ค่าขึ้นไป เรียกว่า “เพอร์เซฟตรอนแบบ Multiple Input”

องค์ประกอบซึ่งถูกบรรจุอยู่ในนิวรอนแต่ละตัว ประกอบด้วย 2 ฟังก์ชันย่อย คือ ฟังก์ชันผลรวม (Summation Function) และ ฟังก์ชันกระตุ้น (Activation Function) และมีรายละเอียดการทำงานของนิวรอน ดังภาพประกอบ 2.10 (Hagan *et al.*, 1996)



ภาพประกอบ 2.10 องค์ประกอบของนิวรอน (Neuron)
(ดัดแปลงจาก: Hagan *et al.*, 1996)

2.5.1.1 ฟังก์ชันผลรวม

ฟังก์ชันผลรวม (Summation Function) ทำหน้าที่ในการคำนวณผลรวมของข้อมูลที่ได้จากชั้นข้อมูลนำเข้า (Input Layer) ซึ่งคำนวณได้ดังสมการ (2.10) (Hagan *et al.*, 1996)

กำหนดให้ ตัวแปร x คือ ค่าข้อมูลนำเข้าตัวที่ i
ตัวแปร w คือ ค่าน้ำหนักของข้อมูลนำเข้าตัวที่ i
ตัวแปร g คือ ผลลัพธ์จากฟังก์ชันผลรวม
ตัวแปร z คือ จำนวนนิวรอนของข้อมูลเข้า
ตัวแปร β คือ ค่าความเอนเอียง (Bias)
ตัวแปร i มีค่าตั้งแต่ 1 จนถึง z

$$g = \sum_{i=1}^{z} x_i \times w_i + \beta \qquad (2.10)$$

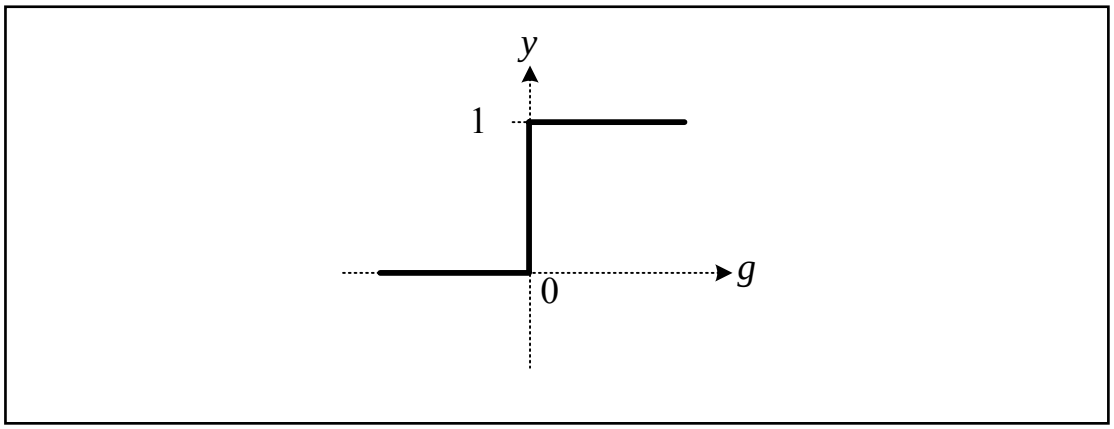
2.5.1.2 ฟังก์ชันกระตุ้น

ฟังก์ชันกระตุ้น (Activation Function) ทำหน้าที่ปรับค่าของข้อมูลที่ได้จากฟังก์ชันผลรวมให้อยู่ในช่วงที่ต้องการ ฟังก์ชันกระตุ้นที่นิยม ได้แก่ ฟังก์ชันสเตป (Step Function) ฟังก์ชันซิกมอยด์ (Sigmoid Function) ฟังก์ชันลิเนียร์ (Linear Function) และฟังก์ชันเกาเซียน (Gaussian Function) โดยกำหนดให้ y คือ ผลลัพธ์ที่ได้จากฟังก์ชันกระตุ้น (Hagan *et al.*, 1996) แสดงรายละเอียดดังนี้

2.4.1.2.1 ฟังก์ชันสเตป (Step Function)

ผลลัพธ์ที่ได้จากฟังก์ชันจะมีค่าเป็น 0 และ 1 ซึ่งคำนวณได้ดังสมการ (2.11) แสดงได้ดังภาพประกอบ 2.11

$$y = \begin{cases} 1, \text{If } g \geq 0 \\ 0, \text{Otherwise} \end{cases} \tag{2.11}$$

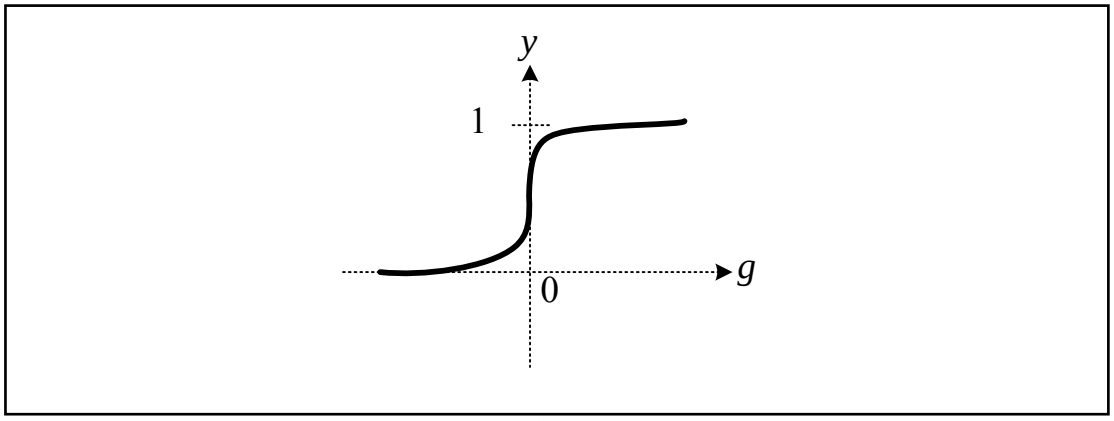


ภาพประกอบ 2.11 ลักษณะของฟังก์ชันสเตป (Step Function)
(ดัดแปลงจาก: Hagan *et al.*, 1996)

2.5.1.2.2 ฟังก์ชันซิกมอยด์ (Sigmoid Function)

ผลลัพธ์ที่ได้จากฟังก์ชันจะมีค่าอยู่ในช่วงระหว่าง 0 ถึง 1 ซึ่งคำนวณได้ดังสมการ (2.12) แสดงได้ดังภาพประกอบ 2.12

$$y = \frac{1}{1+e^{-g}} \tag{2.12}$$

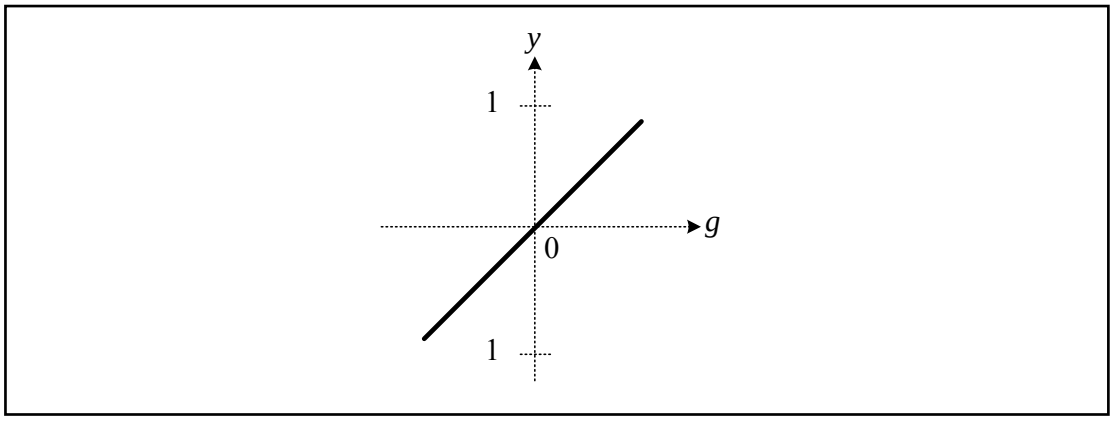


ภาพประกอบ 2.12 ลักษณะของฟังก์ชันซิกมอยด์ (Sigmoid Function)
(ดัดแปลงจาก: Hagan *et al.*, 1996)

2.5.1.2.3 ฟังก์ชันเชิงเส้น (Linear Function)

ผลลัพธ์ที่ได้จากฟังก์ชันจะมีค่าที่การลู่เข้าในลักษณะเชิงเส้น คำนวณได้ดังสมการ (2.13) แสดงได้ดังภาพประกอบ 2.14

$$y = g \tag{2.13}$$



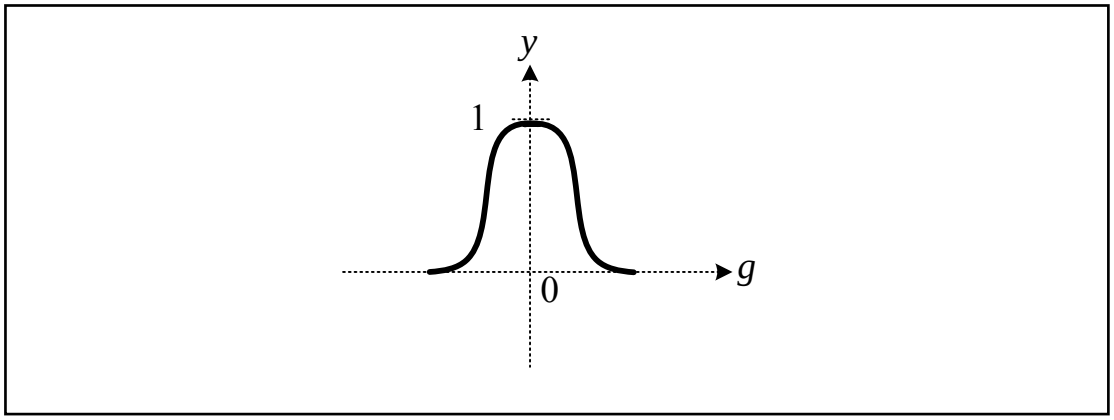
ภาพประกอบ 2.13 ลักษณะของฟังก์ชันเชิงเส้น (Linear Function)
(ดัดแปลงจาก: Hagan *et al.*, 1996)

ในการออกแบบโครงข่ายประสาทเทียม ที่มีการนำฟังก์ชันกระตุ้นแบบฟังก์ชันสเตป (Step Function) ฟังก์ชันซิกมอยด์ (Sigmoid Function) ฟังก์ชันลิเนียร์ (Linear Function) เป็นองค์ประกอบภายในนิวรอน จะเรียกว่า “โครงข่ายประสาทเทียมแบบหลายชั้น” (Multilayer Perceptron: MLP)

2.5.1.2.4 ฟังก์ชันเกาเซียน (Gaussian Function)

ผลลัพธ์ที่ได้จากฟังก์ชันจะมีค่าอยู่ในช่วงระหว่าง 0 ถึง 1 ซึ่งคำนวณได้ดังสมการ (2.14) แสดงได้ดังภาพประกอบ 2.14

$$y = e^{-g^2} \tag{2.14}$$



ภาพประกอบ 2.14 ลักษณะของฟังก์ชันเกาเซียน (Gaussian Function)
(ดัดแปลงจาก: Hagan *et al.*, 1996)

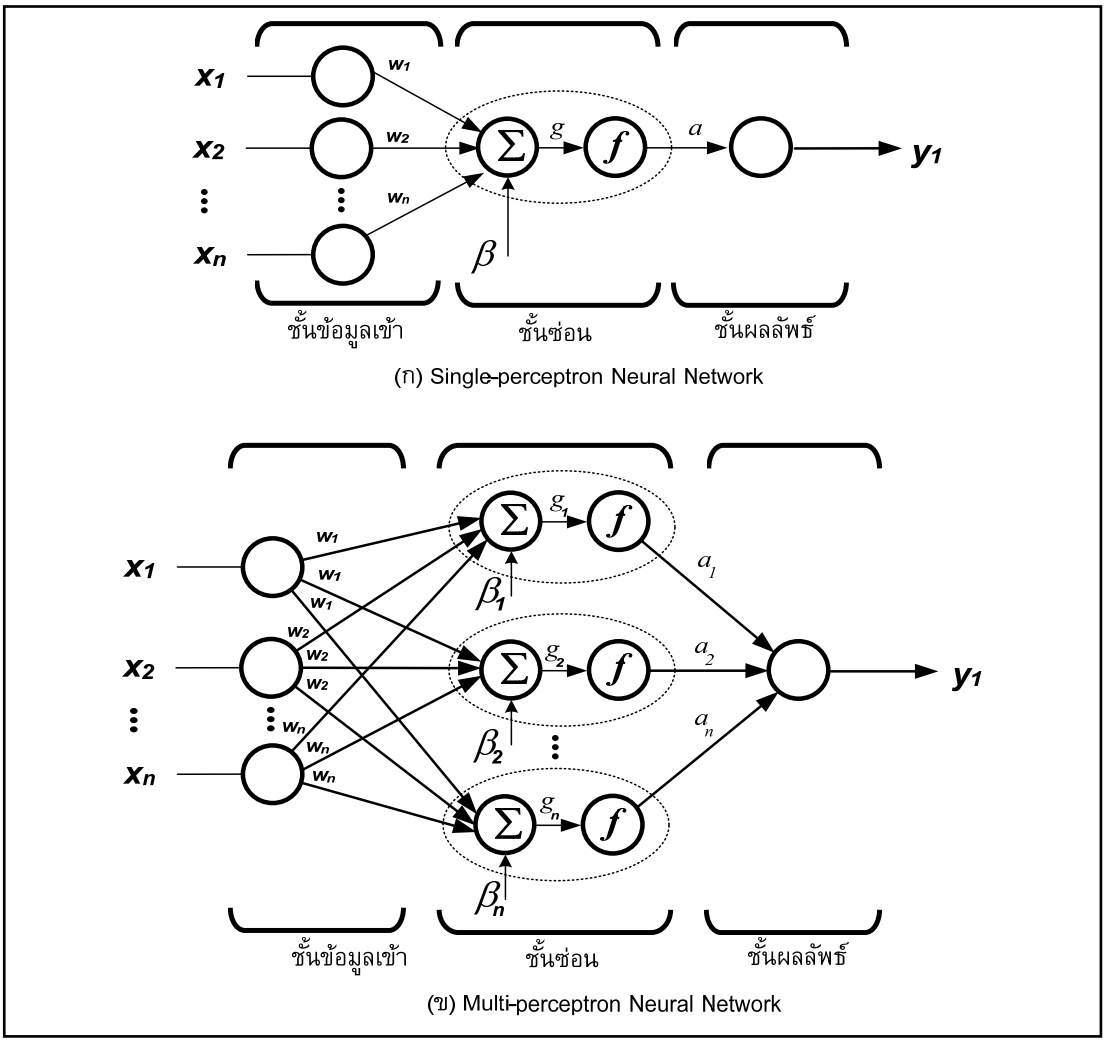
ในการออกแบบโครงข่ายประสาทเทียม ที่มีการนำฟังก์ชันกระตุ้นแบบฟังก์ชันเกาเซียน (Gaussian Function) มาเป็นองค์ประกอบภายในนิวรอน จะเรียกว่า “โครงข่ายประสาทเทียมแบบเรเดียลเบสิสฟังก์ชัน” (Radial Basis Function: RBF)

2.5.2 การคำนวณภายในโครงข่ายประสาท

กระบวนการทำงานของโครงข่ายประสาทเทียม เมื่อมีข้อมูลนำเข้า (Input Data) ในชั้นข้อมูลนำเข้า (Input Layer) ของโครงข่ายประสาท จะนำค่าข้อมูลเข้ามาคูณกับค่าน้ำหนัก (Weight) ของแต่ละขาของข้อมูลนำเข้า ผลที่ได้จากโหนดในชั้นข้อมูลนำเข้าทุก ๆ ขา จากนั้นค่าที่ได้จะถูกนำเข้าสู่นิวรอนสำหรับรวมค่าข้อมูลด้วยฟังก์ชันผลรวม (Summation Function) จากนั้นนำผลลัพธ์จากฟังก์ชันผลรวมมาปรับค่าให้อยู่ในช่วงที่ต้องการด้วยฟังก์ชันกระตุ้น การกระตุ้นจากฟังก์ชันทำให้เกิดการตอบสนอง (Threshold) ตามที่กำหนดไว้แต่ละฟังก์ชัน โดยถ้าผลรวมมีค่ามากกว่า Threshold แล้วนิวรอนก็จะส่งผลลัพธ์ออกไปเพื่อเป็นข้อมูลนำเข้าของนิวรอนที่เชื่อมกันในชั้นถัดไปของโครงข่ายประสาท แต่หากค่าน้อยกว่า Threshold ที่กำหนดไว้ นิวรอนก็จะไม่เกิดผลลัพธ์ โดยจะส่งค่าออกมาเป็นค่าซึ่งเป็นค่าต่ำสุดของฟังก์ชันนั้นออกมาโดยส่วนใหญ่จะเป็นค่าศูนย์ (Hagan *et al.*, 1996)

2.5.3 นิวรอนเพอร์เซฟตรอน

เนื่องจากการแก้ไขปัญหา (Problem Solving) บางปัญหามีความยากในการทำความเข้าใจจำนวนเพอร์เซฟตรอนนั้นสามารถช่วยในกระบวนแก้ปัญหาได้ การออกแบบลักษณะของเพอร์เซฟตอนสามารถแบ่งเป็น 2 แบบ คือ แบบหนึ่งเพอร์เซฟตรอน (Single-Perceptron) และแบบหลายเพอร์เซฟตรอน (Multi-Perceptron) (Hagan *et al.*, 1996) แสดงได้ดังภาพประกอบ 2.15



ภาพประกอบ 2.15 การออกแบบเพอร์เซฟตรอน
(ดัดแปลงจาก: Hagan *et al.*, 1996)

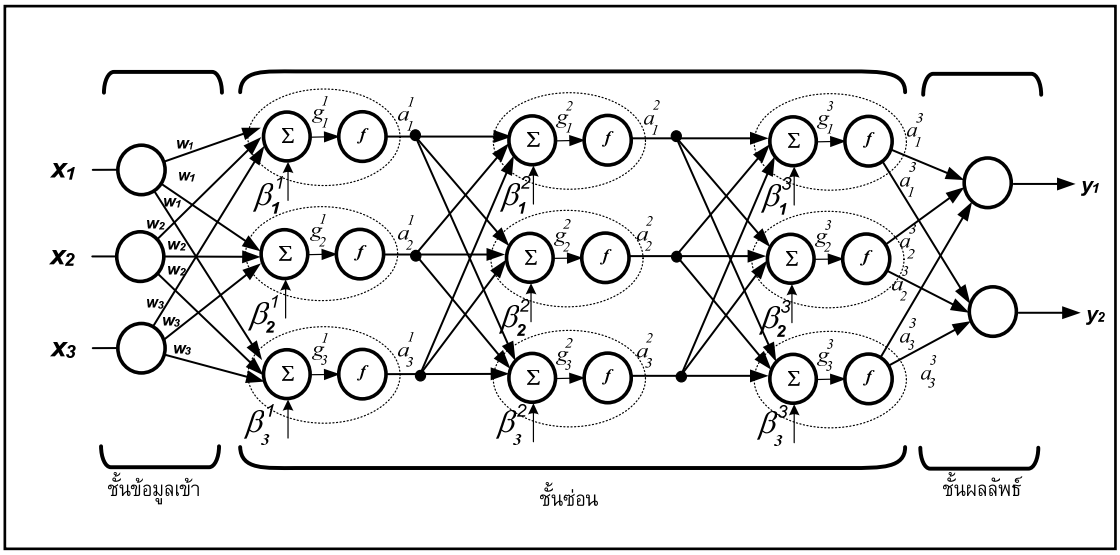
จากภาพประกอบ 2.15 จะเห็นได้ว่า ในชั้นซ่อนจะมีการปรับเพิ่มจำนวนของนิวรอนได้มากกว่า 1 ตัว ขึ้นอยู่กับลักษณะของปัญหาที่ต้องการใช้โครงข่ายประสาทเทียมในการแก้ปัญหา โดยสามารถกำหนดจำนวนนิวรอนได้ตามต้องการ

2.5.4 ประเภทของโครงข่ายประสาท (Architecture of Neural Network)

ประเภทของโครงข่ายประสาทเทียมที่นิยมมี 2 ลักษณะ คือ

2.5.4.1 โครงข่ายประสาทเทียมแบบไปข้างหน้า (Feedforward Neural Network)

การออกแบบโครงข่ายประสาทเทียมแบบไปข้างหน้า จะมีทิศทางการไหลของข้อมูลไปทางเดียวกัน ซึ่งโดยทั่วไปจะเป็นการสร้างโครงข่ายประสาทเทียมเช่นเดียวกับในลักษณะของเพอร์เซฟตรอนแบบชั้นเดียวหรือแบบเพอร์เซฟตรอนแบบหลายหลายชั้น โดยในแต่ละชั้นจะมีนิวรอนได้ไม่จำกัดจำนวน และไม่จำกัดจำนวนชั้นซ่อน ภาพประกอบ 2.16 แสดงลักษณะของโครงข่ายประสาทเทียมแบบไปข้างหน้าที่ประกอบด้วยชั้นข้อมูลเข้า 1 ชั้น ชั้นซ่อน 3 ชั้น และชั้นผลลัพธ์ 1 ชั้น โดยที่ ในชั้นข้อมูลนำเข้า มีโหนดนำเข้าเท่ากับ 3 โหนด จำนวนนิวรอนแต่ละชั้นซ่อนมีจำนวน 3 นิวรอน และมีโหนดแสดงผลลัพธ์จำนวน 2 โหนด

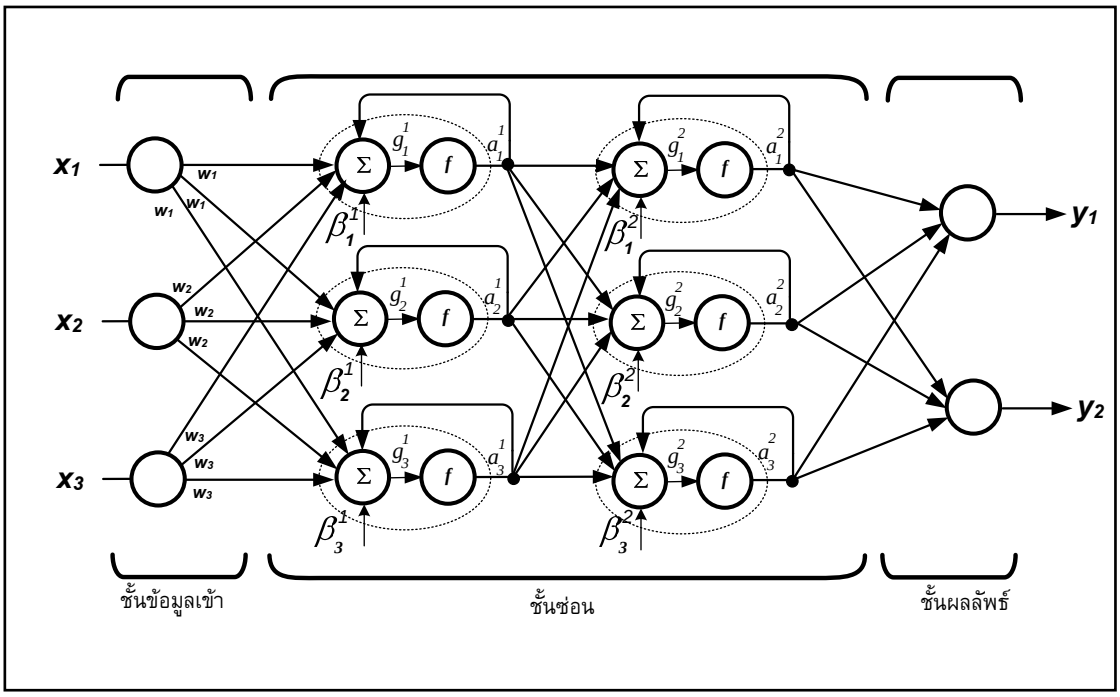


ภาพประกอบ 2.16 โครงข่ายประสาทเทียมแบบไปข้างหน้า (Feedforward Neural Network) (ดัดแปลงจาก: Hagan *et al.*, 1996)

จากภาพประกอบ 2.16 ฟังก์ชันกระตุ้นในการชั้นซ่อนแต่ละชั้นสามารถเลือกใช้ฟังก์ชันที่แตกต่างกัน โดยไม่จำเป็นต้องเป็นฟังก์ชันชนิดเดียวกัน แต่ฟังก์ชันกระตุ้นในชั้นเดียวกันจำเป็นต้องเป็นชนิดเดียวกัน

2.5.4.2 โครงข่ายประสาทแบบย้อนกลับ (Recurrent Neural Network)

ลักษณะการออกแบบโครงข่ายประสาทเทียมแบบย้อนกลับจะมีลักษณะคล้ายกับโครงข่ายประสาทเทียมแบบไปข้างหน้า โดยมีความแตกต่างกันในกรณีที่ผลลัพธ์ที่ได้จากชั้นซ่อนสามารถนำไปเป็นข้อมูลนำเข้าของชั้นซ่อนเดียวกันได้ ภาพประกอบ 2.17 แสดงการออกแบบโครงข่ายประสาทเทียมแบบย้อนกลับ ประกอบด้วย ชั้นข้อมูลเข้า 1 ชั้น ชั้นซ่อนจำนวน 2 ชั้น และ 1 ชั้นผลลัพธ์ โดยในชั้นข้อมูลนำเข้า มีโหนดนำเข้าเท่ากับ 3 โหนด จำนวนนิวรอนแต่ละชั้นซ่อนมีจำนวน 3 นิวรอน และมีโหนดแสดงผลลัพธ์จำนวน 2 โหนด



ภาพประกอบ 2.17 โครงข่ายประสาทเทียมแบบย้อนกลับ (Recurrent Neural Network) (ดัดแปลงจาก: Hagan *et al.*, 1996)

จากภาพประกอบ 2.17 การส่งข้อมูลแบบย้อนกลับนั้นมีการกำหนดเงื่อนไข การย้อนกลับจำนวนกี่รอบหรือกำหนดค่า Threshold สำหรับเป็นเงื่อนไขการหยุดการส่งค่าย้อนกลับของแต่ละชั้น (Rich and Knight, 1991; Hagan *et al.*, 1996)

2.6 ต้นไม้การตัดสินใจ

ต้นไม้การตัดสินใจ (Decision Trees) นับเป็นวิธีการเรียนรู้ที่นิยมใช้จำแนก (Classification) กลุ่มข้อมูลมากที่สุด เนื่องจากมีขั้นตอนการทำงานที่ง่าย ไม่ซับซ้อน และมีความรวดเร็ว การทำงานของต้นไม้การตัดสินใจเป็นวิธีการเรียนรู้ที่อาศัยการพิสูจน์ด้วยเหตุผลโดยการอนุมานความรู้ (Inference Engine) ที่ได้ในรูปของกฎ “*ถ้า...แล้ว...*” (Rich and Knight, 1991; Duda *et al.*, 2000; Mitra and Acharaya, 2003; Roiger and Greatz, 2003)

2.6.1 การแทนข้อมูลในต้นไม้การตัดสินใจ

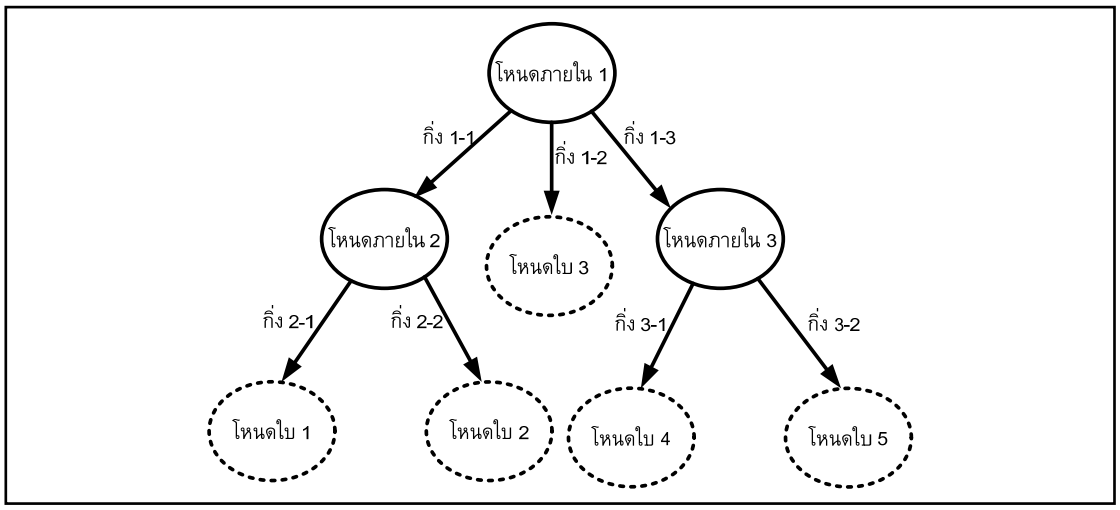
วิธีการแทนค่าข้อมูลผลลัพธ์จากการเรียนรู้ของต้นไม้ ประกอบด้วย

1) โหนดภายใน (Internal Node) คือ คุณสมบัติต่างๆ ของข้อมูล เมื่อข้อมูลใดตกลงมาที่โหนด จะใช้คุณสมบัติข้อมูลนี้เป็นตัวตัดสินใจว่าข้อมูลจะไปในทิศทางใด โดยโหนดภายในที่เป็นจุดเริ่มต้นของต้นไม้เรียกว่า “โหนดราก” (Root Node)

2) กิ่ง (Branch หรือ Link) คือ ค่าคุณสมบัติ (Value of Attribute) ของคุณสมบัติข้อมูลของโหนดภายในที่แตกกิ่งนี้ออกมา โหนดภายในจะแตกกิ่งเป็นจำนวนเท่ากับจำนวนค่าคุณสมบัติของโหนดภายในนั้น

3) โหนดใบ (Leaf Node) คือ กลุ่มต่าง ๆ ซึ่งเป็นผลลัพธ์สำหรับการจำแนกข้อมูลหรือการแบ่งกลุ่มข้อมูล

สามารถแสดงตัวอย่างการแทนข้อมูลในต้นไม้การตัดสินใจ ดังภาพประกอบ 2.18

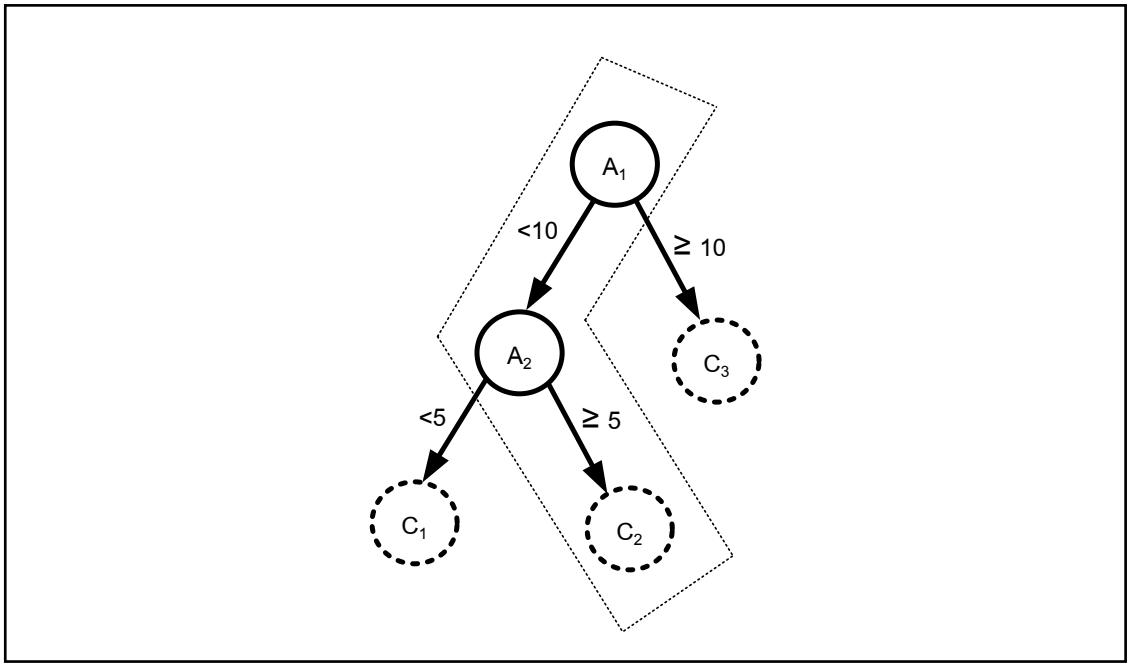


ภาพประกอบ 2.18 แสดงตัวอย่างการแทนข้อมูลในต้นไม้การตัดสินใจ

2.6.2 กฎจากต้นไม้การตัดสินใจ

การแทนกฎจากต้นไม้การตัดสินใจในรูปของกฎ “*ถ้า…แล้ว…*” หรือ “*IF…THEN…*” จะเริ่มจากโหนดรากไปยังโหนดใบ ซึ่งที่โหนดใบจะเป็นคำตอบของปัญหาหรือกลุ่มของปัญหานั้น

กำหนดให้ D = $\{t_1,\ldots,t_n\}$ เป็นฐานข้อมูล ประกอบด้วย เรคอร์ด (Records) ตั้งแต่ $t_1,t_2,\ldots,t_n$ เรคอร์ด ซึ่งในแต่ละเรคอร์ด ประกอบด้วย คุณสมบัติของข้อมูล (Attributes) ตั้งแต่ $A_1,A_2,\ldots,A_n$ ตามลำดับ และ D = $\{C_1,C_2,\ldots,C_n\}$ เป็นเซตของกลุ่ม (Class) ที่ต้องการแบ่งกลุ่มข้อมูลด้วยต้นไม้การตัดสินใจ โหนดภายในจะถูกกำกับ (Label) ด้วยคุณสมบัติข้อมูล A_i และแต่ละกิ่งจะกำกับด้วยเงื่อนไขการตัดสินใจที่ขึ้นต่อคุณสมบัติข้อมูล A_i และโหนดใบจะถูกกำกับด้วยกลุ่มข้อมูล C_j แสดงดังภาพประกอบ 2.19



ภาพประกอบ 2.19 การสร้างกฎจากต้นไม้การตัดสินใจ

จากภาพประกอบ 2.19 สามารถสร้างกฎตัวอย่างที่ได้จากต้นไม้การตัดสินใจ เช่น IF $A_1 < 10$ AND $A_2 \geq 5$ THEN Class = C_2 ซึ่งหมายถึง ถ้า A_1 มีค่าน้อยกว่า 10 และ A_2 มีค่ามากกว่าหรือเท่ากับ 5 แล้ว อยู่ในกลุ่ม C_2

2.6.3 ขั้นตอนวิธีของต้นไม้การตัดสินใจ

ขั้นตอนวิธีในการสร้างต้นไม้การตัดสินใจมีความแตกต่างกันไปและเหมาะสมกับการแก้ปัญหาซึ่งแตกต่างกัน (Witten and Frank, 2005) เช่น

1) ต้นไม้การตัดสินใจแบบ Alternating Decision Tree (ADTree)

สร้างต้นไม้ตัดสินใจโดยใช้ค่า Boosting Score เป็นวิธีการที่เหมาะสมกับปัญหาที่มีการจัดกลุ่มแบบ 2 กลุ่ม

2) ต้นไม้การตัดสินใจแบบ Pruned C4.5 Decision Tree (J48)

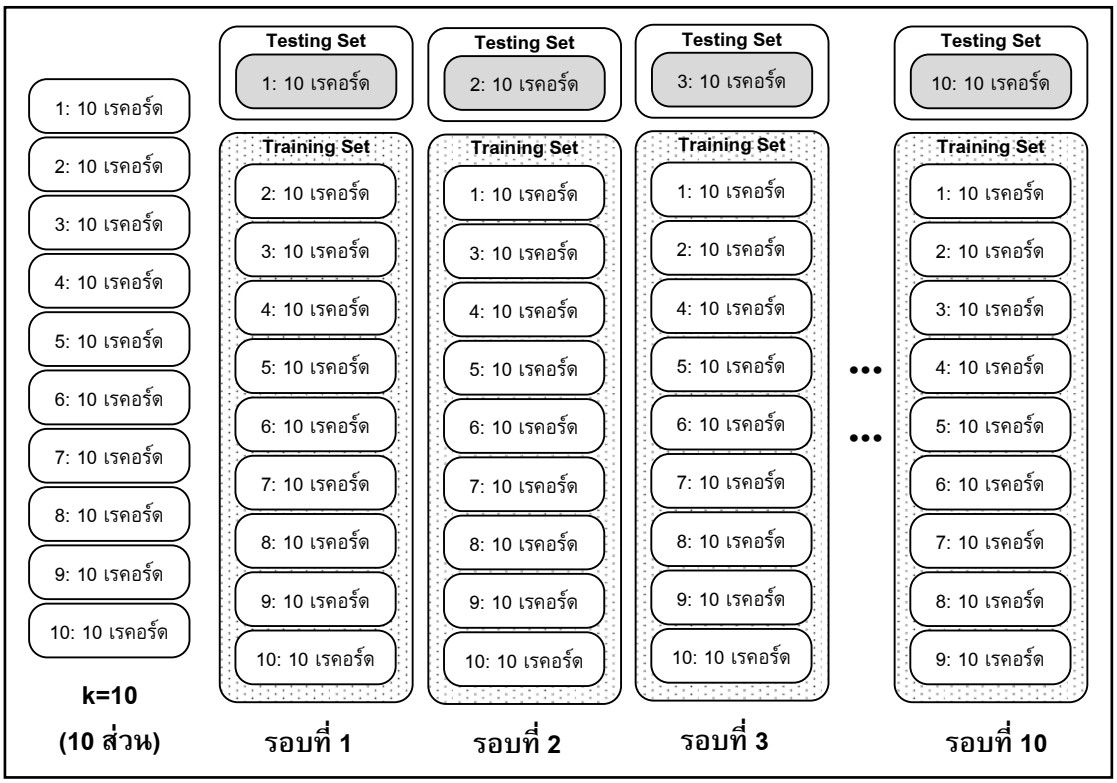
สร้างต้นไม้โดยใช้หลักการพื้นฐานของขั้นตอนวิธีแบบ C4.5 โดยปรับปรุงในส่วนการตัดโหนด (Pruning) ที่ไม่จำเป็นออก ทำให้ได้โหนดมีคุณสมบัติที่ดีและมีขนาดที่เหมาะสม

3) ต้นไม้การตัดสินใจแบบ K Random Features Decision Tree (RandomTree)

สร้างต้นไม้โดยอาศัยการสุ่มแอททริบิวต์สร้างโหนดราก (Root Node) จนได้โหนดที่มีคุณสมบัติที่ต้องการและจะทำเช่นเดียวกันในโหนดชั้นถัดไป

2.7 การทดสอบประสิทธิภาพแบบ k-Fold Cross Validation

การทดสอบประสิทธิภาพแบบ k-Fold Cross Validation คือ การแบ่งข้อมูลเป็นส่วนสำหรับสอน (Training Set) และสำหรับทดสอบ (Testing Set) ออกเป็น k ส่วนเท่า ๆ กัน ซึ่งค่า k ที่เหมาะสม คือ 10 (k = 10) หรือ 10-Fold Cross Validation (Roiger and Geatz, 2003; Written and Frank, 2005) ดังแสดงในภาพประกอบ 2.20



ภาพประกอบ 2.20 การแบ่งข้อมูลสำหรับทดสอบแบบ k-Fold Cross Validation

จากภาพประกอบ 2.20 สมมติมีข้อมูลทั้งสิ้น 100 เรคอร์ด และกำหนดค่า k = 10 สามารถแบ่งข้อมูลออกเป็น 10 ส่วนเท่าๆ กัน แต่ละส่วนประกอบด้วยข้อมูลจำนวน 10 เรคอร์ด และมีจำนวนรอบการทดสอบประสิทธิภาพทั้งสิ้น 10 รอบ โดยในการทำงานรอบที่ 1 จะสุ่มเลือกข้อมูล 1 ส่วนสำหรับเป็นข้อมูลชุดทดสอบ และ 9 ส่วนที่เหลือจะใช้เป็นข้อมูลชุดสอน การทำงานรอบที่ 2 จะใช้หลักการเดียวกันโดยจะสุ่มเลือกจากข้อมูล 9 ส่วนที่เหลือสำหรับสร้างเป็นข้อมูลทดสอบ และจะทำเช่นนี้จนกระทั่งข้อมูลทุกส่วนถูกนำมาเป็นชุดทดสอบ และนำค่าความถูกต้อง (Accuracy) ที่ได้ในแต่ละรอบมาหาค่าเฉลี่ยเพื่อเป็นค่าความถูกต้องโดยรวมของการจำแนกข้อมูลของแบบจำลอง

2.8 การวิเคราะห์ประสิทธิภาพ

ประสิทธิภาพ (Performance) การทำงานของขั้นตอนวิธีแต่ละวิธี สามารถวัดจากผลลัพธ์จัดกลุ่มข้อมูล (Classification) ค่าของผลลัพธ์ที่ได้จากการจัดกลุ่ม คือ ค่า True Positive (TP) ค่า True Negative (TN) ค่า False Positive (FP) และ ค่า False Negative (FN) ตามลำดับ สามารถแสดงดังตารางที่ 2.1 (Written and Frank, 2005)

ตารางที่ 2.1 ค่าของคอนฟิวชันเมทริกซ์ (Confusion Matrix) แบบ 2 กลุ่ม

ค่าที่แท้จริง (Actual Class)	**ค่าที่ทำนายได้ (Predicted Class)**	
	Class YES	**Class NO**
Class YES	True Positive : TP	False Negative : FN
Class NO	False Positive: FP	True Negative : TN

ค่าที่ได้จากการทำนาย (Prediction) ในตารางที่ 2.1 อธิบายรายละเอียดได้ดังนี้

1) ค่า True Positive (TP) คือ ค่าที่บอกความถูกต้องในการจำแนกข้อมูลซึ่งมีค่าที่แท้จริงอยู่ใน Class YES และมีการทำนายว่าอยู่ใน Class YES (ทำนายถูกต้อง)

2) ค่า False Negative (FN) คือ ค่าที่บอกความถูกต้องในการจำแนกข้อมูลซึ่งมีค่าที่แท้จริงอยู่ใน Class YES และมีการทำนายว่าอยู่ใน Class NO (ทำนายผิด)

3) ค่า False Positive (FP) คือ ค่าที่บอกความถูกต้องในการจำแนกข้อมูลซึ่งมีค่าที่แท้จริงอยู่ใน Class NO และมีการทำนายว่าอยู่ใน Class YES (ทำนายผิด)

4) ค่า True Negative (TN) คือ ค่าที่บอกความถูกต้องในการจำแนกข้อมูลซึ่งมีค่าที่แท้จริงอยู่ใน Class NO และมีการทำนายว่าอยู่ใน Class NO (ทำนายถูกต้อง)

จากค่าผลลัพธ์ที่ได้ทั้ง 4 ค่า สามารถนิยามการวัดประสิทธิภาพ คือ ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าถ่วงดุล (Tradeoff) แสดงรายละเอียดได้ดังนี้

2.8.1 ค่าความแม่นยำ (Precision: *p*)

คือ ค่าของการทำนายค่าที่จริงในกลุ่มที่พบและถูกต้อง ซึ่งพิจารณาจากจำนวนกลุ่มข้อมูลทั้งหมด คำนวณดังสมการ (2.15)

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.15}$$

เมื่อ TP + FN มีค่าเท่ากับ 0 ให้ตัวหาร (TP + FN) มีค่าเป็น 1

2.8.2 ค่าระลึกได้ (Recall: *r*)

คือ ค่าจำนวนกลุ่มข้อมูลที่พบ ซึ่งพิจารณาจากจำนวนกลุ่มข้อมูลทั้งหมด คำนวณดังสมการ (2.16)

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{TN}} \tag{2.16}$$

เมื่อ TP + TN มีค่าเท่ากับ 0 ให้ตัวหาร (TP + TN) มีค่าเป็น 1

2.8.3 ค่าถ่วงดุล (Tradeoff)

คือ ค่าถ่วงดุลระหว่างค่าความแม่นยำและค่าความระลึกได้ หรือเรียกว่า “F-Measure” โดยการนำค่าทั้ง 2 มาหาค่าเฉลี่ย คำนวณดังสมการ (2.18)

$$\text{F}(r, p) = \frac{2 \times p \times r}{p + r} \tag{2.18}$$

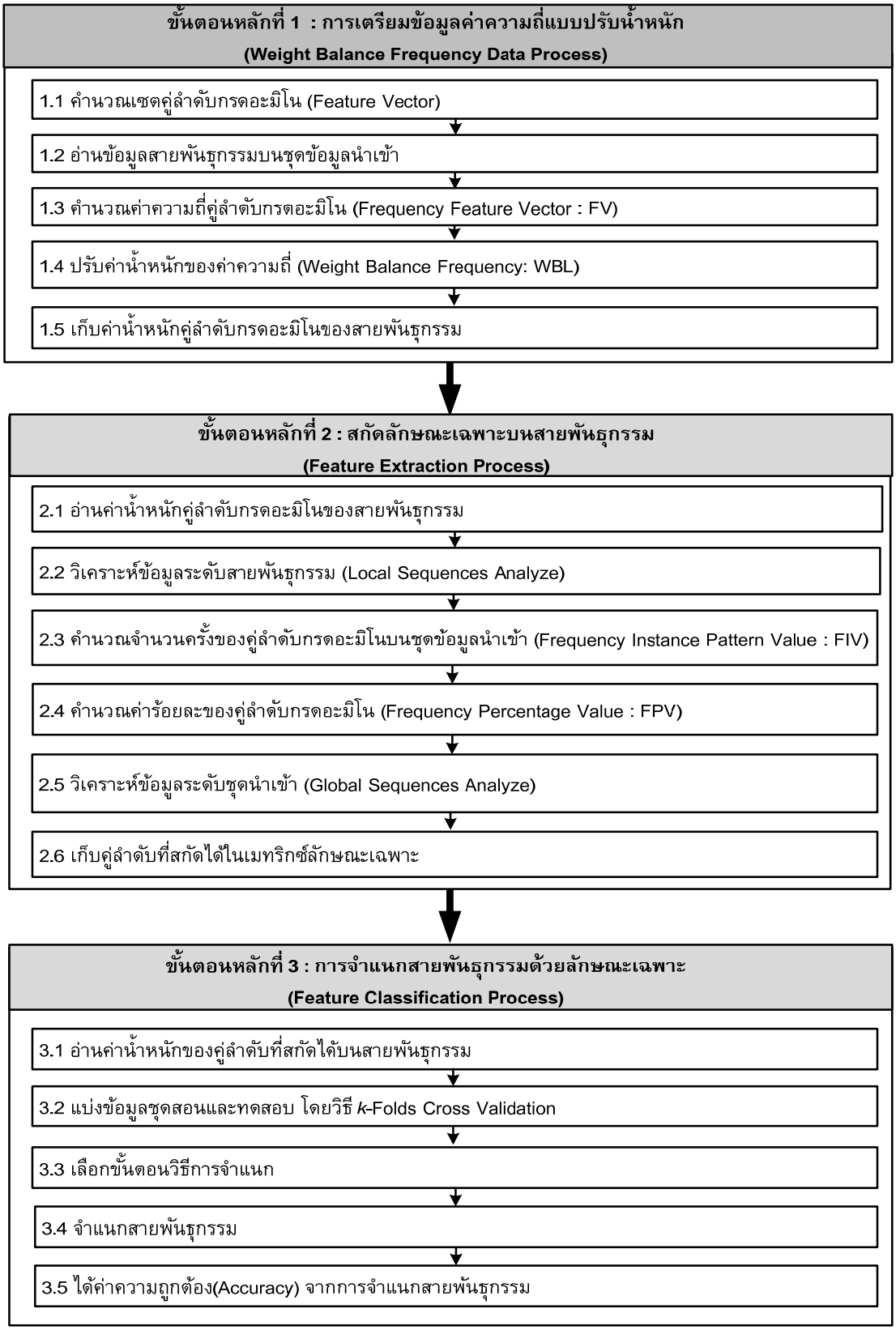
บทที่ 3

แบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

วิทยานิพนธ์นี้นำเสนอแบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม ด้วยขั้นตอนวิธีการสกัดลักษณะเฉพาะบนสายโปรตีน (FEPSA) ประกอบด้วย 3 ขั้นตอนหลัก คือ ขั้นตอนการเตรียมข้อมูลค่าความถี่แบบปรับน้ำหนัก ขั้นตอนการสกัดลักษณะเฉพาะบนสายพันธุกรรม และขั้นตอนการจำแนกสายพันธุกรรมด้วยลักษณะเฉพาะ สามารถอธิบายรายละเอียดตามลำดับดังนี้

3.1 แบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

แบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม ด้วยขั้นตอนวิธีการสกัดลักษณะเฉพาะบนสายโปรตีน (Feature Extraction of Protein Sequence Analysis: FEPSA) ประกอบด้วย 3 ขั้นตอนหลัก คือ ขั้นตอนหลักที่ 1 การเตรียมข้อมูลค่าความถี่แบบปรับน้ำหนัก (Weight Balance Frequency Data Process) เป็นขั้นตอนแปลงข้อมูลลำดับพันธุกรรมจากสายโปรตีน (Protein Sequences) ให้อยู่ในเมทริกซ์ค่าความถี่ที่มีการปรับค่าน้ำหนัก (Weight Balance Frequency) ด้วยเทคนิคเอ็นแกรม (n-Grams) ขั้นตอนหลักที่ 2 การสกัดลักษณะเฉพาะบนสายพันธุกรรม (Feature Extraction Process) เป็นขั้นตอนการสกัดลักษณะเฉพาะซึ่งปรากฏบนสายพันธุกรรมแต่ละสายด้วยเทคนิคการวิเคราะห์ระดับชุดข้อมูล (Global Sequences Analyze) และขั้นตอนหลักที่ 3 การจำแนกสายพันธุกรรมด้วยลักษณะเฉพาะ (Feature Classification Process) ที่สกัดได้จากสายพันธุกรรม และจำแนกด้วยขั้นตอนวิธีแบบโครงข่ายประสาทเทียม ได้แก่ โครงข่ายประสาทเทียมแบบหลายชั้น (Multilayer Perceptron: MLP) และแบบเรเดียลเบสิสฟังก์ชัน (Radial Basis Function: RBF) ขั้นตอนวิธีแบบต้นไม้การตัดสินใจ ได้แก่ แบบ Alternating Decision Tree (ADTree) แบบ Pruned C4.5 Decision Tree (J48) และแบบ K Random Features Decision Tree (RandomTree) ตามลำดับ สามารถแสดงรายละเอียดการทำงานของแบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม ดังภาพประกอบ 3.1



ภาพประกอบ 3.1 ขั้นตอนการทำงานแบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

ขั้นตอนหลักที่ 1 การเตรียมข้อมูลค่าความถี่แบบปรับน้ำหนัก (Weight Balance Frequency Data Process)
1.1 คำนวณเซตคู่ลำดับกรดอะมิโนที่เป็นได้ (Feature Vecter)
1.2 อ่านข้อมูลสายพันธุกรรมบนชุดข้อมูลนำเข้า
1.3 คำนวณค่าความถี่คู่ลำดับกรดอะมิโนบนสายพันธุกรรม (Frequency Feature Vector: FV)
1.4 ปรับค่าน้ำหนักของคู่ลำดับกรดอะมิโน (Weight Balance Frequency: WBL)
1.5 เก็บค่าน้ำหนักคู่ลำดับกรดอะมิโนของทุกสายพันธุกรรม

ภาพประกอบ 3.2 รายละเอียดขั้นตอนการเตรียมข้อมูลค่าความถี่แบบปรับน้ำหนัก

3.1.1 ขั้นตอนหลักที่ 1 การเตรียมข้อมูลค่าความถี่แบบปรับน้ำหนัก

จากภาพประกอบ 3.2 แสดงขั้นตอนการเตรียมข้อมูลค่าความถี่แบบปรับน้ำหนัก (Weight Balance Frequency Data Process) ของคู่ลำดับกรดอะมิโน โดยการแปลงรูปแบบของข้อมูลลำดับพันธุกรรมให้อยู่ในรูปของค่าสถิติ (Statistic Value) ด้วยการหาค่าความถี่ของคู่ลำดับกรดอะมิโน (Frequency Feature Vector: FV) บนสายพันธุกรรมทุกสายนำเข้าด้วยเทคนิคเอ็นแกรม โดยกำหนดค่า n = 2 ตัวอักษร (2-Grams) ดังรายละเอียดต่อไปนี้

3.1.1.1 คำนวณเซตของคู่ลำดับที่เป็นไปได้ (Feature Vector)

การคำนวณคู่ลำดับกรดอะมิโนที่เป็นไปได้สำหรับสร้างเป็นคุณสมบัติของข้อมูล (Attributes) ด้วยเทคนิคเอ็นแกรม โดยใช้สัญลักษณ์ย่อของกรดอะมิโนมาตรฐานจำนวน 20 ชนิด และกำหนดให้ตัวแปร T แทน เซตของสัญลักษณ์ย่อกรดอะมิโนมาตรฐานจำนวน 20 ชนิด เขียนแทนด้วย T = {A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V} และตัวแปร P แทน จำนวนคู่ลำดับกรดอะมิโนที่เป็นไปได้จากการจัดคู่ สามารถคำนวณรูปแบบที่เป็นไปได้ดังสมการ (3.1)

$$P = |T|^2 \tag{3.1}$$

เมื่อ ตัวแปร n มีค่าเท่ากับ 2 ตัวอักษร (หรือเรียกว่า 2-Grams) และจำนวนคู่ลำดับกรดอะมิโนที่ได้จากการจัดคือ $P = |20|^2$ ซึ่งมีค่าเท่ากับ 400 (400 คู่ลำดับ)

สามารถแสดงขั้นตอนวิธีการทำงานได้ ดังภาพประกอบที่ 3.3

ขั้นตอนวิธี : การคำนวณเซตของคู่ลำดับกรดอะมิโนที่เป็นไปได้แบบ n-Grams (n=2)

รายละเอียด : กำหนดให้ตัวแปร T เก็บชื่อย่อกรดอะมิโนมาตรฐาน 20 ชนิด, n มีค่าเท่ากับ 2, ตัวแปร P เก็บคู่ลำดับที่ได้จากการจัด, ตัวแปร i เก็บค่าแถว และตัวแปร j เก็บค่าหลักของเมทริกซ์ชื่อย่อกรดอะมิโนมาตรฐาน 20 ชนิด

การทำงาน : สร้างเซตคู่ลำดับที่เป็นไปได้ ดังนี้

```
FOR i = 1 TO |T|
        FOR j = 1 TO |T|
                P(i,j)= T(i) Combination with T(j)
                j=j+1
        END LOOP
        i=i+1
END LOOP
```

ภาพประกอบ 3.3 คำนวณเซตของคู่ลำดับที่เป็นได้ (Feature Vector) แบบ 2-Grams

จากภาพประกอบ 3.3 แสดงขั้นตอนการสร้างเซตคู่ลำดับกรดอะมิโนที่เป็นไปได้จากรายชื่อย่อของกรดอะมิโนมาตรฐาน 20 ชนิด สามารถแสดงเมทริกซ์คู่ลำดับกรดอะมิโนที่เป็นได้ ด้วยเมทริกซ์ขนาด $i \times j$ เมื่อกำหนดให้ $P_{i,j}$ แทน เมทริกซ์รูปแบบที่เป็นไปได้ คำนวณได้จากสมการ (3.2)

$$P_{i,j} = \begin{Bmatrix} P_{(1,1)}=AA & P_{(1,2)}=AR & P_{(1,3)}=AN & P_{(1,4)}=AD & \dots & P_{(1,20)}=AV \\ P_{(2,1)}=RA & P_{(2,2)}=RR & P_{(2,3)}=RN & P_{(2,4)}=RD & \dots & P_{(2,20)}=RV \\ P_{(3,1)}=NA & P_{(3,2)}=NR & P_{(3,3)}=NN & P_{(3,4)}=ND & \dots & P_{(3,20)}=NV \\ P_{(4,1)}=DA & P_{(4,2)}=DR & P_{(4,3)}=DN & P_{(4,4)}=DD & \dots & P_{(4,20)}=DV \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{(20,1)}=VA & P_{(20,2)}=VR & P_{(20,3)}=VN & P_{(20,4)}=VD & \dots & P_{(20,20)}=VV \end{Bmatrix} \quad (3.2)$$

เมื่อ ตัวแปร i แทน ค่าของแถว (Rows) และตัวแปร j แทน ค่าของหลัก (Columns) ของเมทริกซ์คู่ลำดับที่เป็นไปได้ ($P_{i,j}$) โดยตัวแปร i และ j มีค่าเท่ากับ |T| และ T แทน ชื่อย่อกรดอะมิโนมาตรฐาน 20 ชนิด

จากเมทริกซ์รูปแบบที่เป็นไปได้ คู่ลำดับ AA เก็บในตำแหน่งแถวที่ 1 หลักที่ 1 ของเมทริกซ์รูปแบบที่เป็นไปได้ (P) หรือ $P_{(1,1)}$ = AA คู่ลำดับ $P_{(1,2)}$ = AR, คู่ลำดับ $P_{(1,3)}$ = AN คู่ลำดับ $P_{(2,1)}$ = RA ... และคู่ลำดับ $P_{(20,20)}$ = VV ตามลำดับ

3.1.1.2 อ่านข้อมูลสายพันธุกรรมจากชุดข้อมูล

ในการอ่านข้อมูลสายพันธุกรรมจากชุดข้อมูล จะกำหนดให้ตัวแปร K เก็บค่าจำนวนสายพันธุกรรมทั้งหมดในชุดข้อมูล

3.1.1.3 คำนวณค่าความถี่คู่ลำดับกรดอะมิโนบนสายพันธุกรรม

การคำนวณค่าความถี่ที่พบคู่ลำดับกรดอะมิโนทุกรูปแบบบนสายพันธุกรรม (Frequency Feature Vector: FV) ดังนิยามที่ 3.1

นิยามที่ 3.1 : กำหนดให้ $FV(k,V_{i,j},F_{i,j})$ แทน ค่าความถี่ของคู่ลำดับกรดอะมิโนที่คำนวณได้เมื่อตัวแปร K แทน จำนวนสายพันธุกรรมทั้งหมด ตัวแปร k แทน ลำดับของสายพันธุกรรม (เมื่อ $1 \leq k \leq K$) ตัวแปร $V_{i,j}$ แทนตำแหน่งคู่ลำดับกรดอะมิโนบนเมทริกซ์คู่ลำดับกรดอะมิโนที่เป็นไปได้ และตัวแปร $F_{i,j}$ แทน ค่าความถี่ (Frequency Value) ที่พบบนสายพันธุกรรมลำดับที่ k

สามารถแสดงขั้นตอนวิธีการทำงานดังภาพประกอบ 3.4

ขั้นตอนวิธี : คำนวณความถี่คู่ลำดับที่เป็นไปได้บนสายพันธุกรรมแบบ 2-Grams

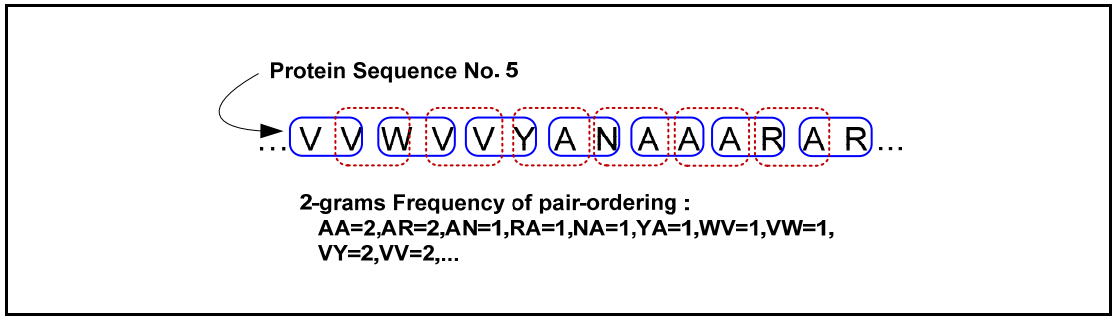
รายละเอียด : กำหนดให้ตัวแปร K แทนจำนวนสายพันธุกรรมทั้งหมด, ตัวแปร k แทนลำดับของสายพันธุกรรม, L(k) เก็บค่าความยาวของสายพันธุกรรมสายที่ k, ตัวแปร Sum เก็บจำนวนครั้งที่พบคู่ลำดับกรดอะมิโนบนสายพันธุกรรมสายที่ k , ตัวแปร SumWT เก็บผลรวม, ตัวแปร $FV(k,V_{i,j},F_{i,j})$ เก็บค่าความถี่ของคู่ลำดับกรดอะมิโน และตัวแปร Mean(k) เก็บค่าเฉลี่ยของค่าความถี่บนสายพันธุกรรมที่ k

การทำงาน : คำนวณความถี่ที่พบคู่ลำดับกรดอะมิโนบนสายพันธุกรรม ดังนี้

```
FOR k= 1 TO K
    FOR i = 1 TO  |T|
            FOR  j = 1 TO |T|
                    FOR  Arr1 = 0 TO L(k)-1
                        If P_{i,j} = Seq(k,Arr1&Arr1+1) Then Sum = Sum + 1
                    END LOOP
                   j = j + 1
             END LOOP
             FV(k,V_{i,j},F_{i,j}) = Sum,  SumWT = SumWT + FV(k,V_{i,j},F_{i,j}), i = i + 1
      END LOOP
      Mean(k) = SumWT / |P| , k = k+1
   END LOOP
```

ภาพประกอบ 3.4 ขั้นตอนวิธีการคำนวณความถี่คู่ลำดับกระอะมิโนบนสายพันธุกรรม

จากภาพประกอบ 3.4 วิธีการนับค่าความถี่ของคู่ลำดับกรดอะมิโนบนสายพันธุกรรม จะทำการนับตัวอักษรครั้งละ 2 ตัวอักษร หรือเป็นคู่ๆ โดยคู่แรกจะเริ่มจากตำแหน่งเริ่มต้นของของสายพันธุกรรม และมีการขยับไปทางขวาครั้งละ 1 ตำแหน่ง (หรือ 1 ตัวอักษร) เพื่อให้ได้คู่ลำดับกรดอะมิโนถัดไป สามารถแสดงวิธีการคำนวณดังภาพประกอบ 3.5



ภาพประกอบ 3.5 การหาค่าความถี่ด้วยเทคนิค n-Grams เมื่อกำหนดค่าของ n=2 ตัวอักษร

จากภาพประกอบ 3.5 สายพันธุกรรมที่เลือกพิจารณา คือ สายพันธุกรรมลำดับที่ 5 (Seq-5) และมีความยาวของสายพันธุกรรมเท่ากับ 73 ตัวอักษร โดยส่วนหนึ่งของสายพันธุกรรมปรากฏข้อมูลลำดับกรดอะมิโน คือ “**...VVWVVYANAAARAR...**” เมื่อทำการคำนวณจะได้คู่ลำดับกรดอะมิโนคู่แรก คือ VV คู่ถัดไป คือ VW คู่ลำดับ WV และคู่ลำดับกรดอะมิโน VV โดยจะทำเช่นนี้ไปเรื่อยๆ จนกระทั่งสิ้นสุดความยาวสายพันธุกรรม สามารถเขียนค่าความถี่ของคู่ลำดับกรดอะมิโนทั้งหมดคำนวณที่ได้ด้วยนิยามที่ 3.1 คือ FV(5,AA,2) FV(5,AR,2) FV(5,AN,1) FV(5,NA,1) FV(5,YA,1) FV(5,WV,1) ... FV(5,VW,1) FV(5,VY,1) และ FV(5,VV,2) ตามลำดับ ดังแสดงในตารางที่ 3.1

ตารางที่ 3.1 ค่าความถี่คู่ลำดับอะมิโนบนสายพันธุกรรม

Seq. No.	Seq. Length	Frequency Feature Vector (FV)							Mean
		AA	AR	AN	...	VW	VY	VV	
Seq-1	86	5	1	3	...	2	1	1	4.40
Seq-2	69	5	6	1	...	0	3	5	5.81
Seq-3	96	8	8	9	...	0	5	3	1.78
Seq-4	76	1	7	1	...	2	8	7	5.31
Seq-5	**73**	**2**	**2**	**1**	**...**	**1**	**1**	**2**	**4.98**
Seq-6	95	1	7	0	...	9	0	3	3.83
Seq-7	83	3	3	5	...	2	7	0	4.64
Seq-8	87	1	8	0	...	5	8	0	4.20
Seq-9	88	6	1	1	...	1	6	0	5.33
Seq-10	93	4	8	3	...	6	7	6	3.51

3.1.1.4 ปรับน้ำค่าหนักของคู่ลำดับกรดอะมิโน (Weight Balance Frequency: WBL)

ทำการปรับค่าน้ำหนักของทุกคู่ลำดับกรดอะมิโนบนสายพันธุกรรมที่นำเข้า เพื่อให้ได้ค่าข้อมูลความถี่ของคู่ลำดับกรดอะมิโนที่ปรากฏบนสายพันธุกรรมแต่ละสายที่เป็นมาตรฐานเดียวกัน โดยกำหนดให้ $WBL(k,V_{i,j},W_{i,j})$ แทน ค่าน้ำหนักใหม่ของคู่ลำดับกรดอะมิโน คำนวณได้ดังสมการ (3.3)

$$WBL(k,V_{i,j},W_{i,j}) = \frac{FV(k,V_{i,j},F_{i,j})}{L(k)-(n-1)} \times 100 \qquad (3.3)$$

เมื่อ ตัวแปร $W_{i,j}$ แทน ค่าน้ำหนักใหม่ของคู่ลำดับกรดอะมิโน ตัวแปร L(k) แทน ค่าความยาวของสายพันธุกรรมลำดับที่ k (เมื่อ $1 \leq k \leq K$) และตัวแปร n แทน ขนาดของ n-Grams

สามารถแสดงขั้นตอนวิธีการทำงาน ดังภาพประกอบที่ 3.6

ขั้นตอนวิธี : การปรับค่าน้ำหนักของคู่ลำดับกรดอะมิโนบนสายพันธุกรรม

รายละเอียด : กำหนดให้ตัวแปร K แทนจำนวนสายพันธุกรรมทั้งหมด, ตัวแปร n = 2 ,ตัวแปร k แทนลำดับของสายพันธุกรรม, ตัวแปร L(k) เก็บค่าความยาวสายพันธุกรรมที่ k, ตัวแปร $WBL(k,V_{i,j},W_{i,j})$ เก็บค่าน้ำหนักคู่ลำดับกรดอะมิโนที่ $V_{i,j}$ บนสายพันธุกรรมที่ k และตัวแปร Mean(k) เก็บค่าเฉลี่ยของค่าความถี่คู่ลำดับกรดอะมิโนบสายพันธุกรรมที่ k

การทำงาน : คำนวณค่าน้ำหนักของคู่ลำดับที่พบบนสายพันธุกรรม ดังนี้

```
FOR k = 1 TO  N
    Sum = 0, SumWT = 0
    FOR i = 1 TO  |T|
            FOR  j = 1 TO |T|
                    WBL(k,Vi,j,Wi,j) = (FV(k,Vi,j,Fi,j)) / (L(k) - (n - 1))) * 100
                    SumWT= SumWT + WBL(k,Vi,j,Wi,j)
                    j = j + 1
             END LOOP
        i = i + 1
     END LOOP
     Mean(k) = SumWT / |P|
     k=k + 1
END LOOP
```

ภาพประกอบ 3.6 ขั้นตอนวิธีการปรับน้ำหนักความถี่บนสายพันธุกรรม

เมื่อนำค่าจากตารางที่ 3.1 มาคำนวณด้วยสมการ (3.3) สามารถเขียนค่าน้ำหนักใหม่ของทุกคู่ลำดับกรดอะมิโนบนสายพันธุกรรมสายที่ 5 (Seq-5) ได้ดังนี้

$$FV(5,AA,2) = \frac{2}{73-(2-1)} \times 100 = 2.78$$

$$FV(5,AR,2) = \frac{2}{73-(2-1)} \times 100 = 2.78$$

$$FV(5,AN,1) = \frac{1}{73-(2-1)} \times 100 = 1.39$$

$$\vdots$$

$$FV(5,VW,1) = \frac{1}{73-(2-1)} \times 100 = 1.39$$

$$FV(5,VY,1) = \frac{1}{73-(2-1)} \times 100 = 1.39$$

$$FV(5,VV,2) = \frac{2}{73-(2-1)} \times 100 = 2.78$$

เมื่อนำค่าข้อมูลความถี่ซึ่งผ่านการปรับน้ำหนัก สามารถเขียนแทนด้วยสมการ $WBL(k,V_{i,j},W_{i,j})$ ได้คือ WBL(5,AA,2.78) WBL(5,AR,2.78) WBL(5,AN,1.39) … WBL(5,VW,1.39) WBL(5,VY,1.39) และ WBL(5,VV,2.78) ตามลำดับ แสดงได้ดังตารางที่ 3.2

ตารางที่ 3.2 ข้อมูลค่าความถี่คู่ลำดับอะมิโนแบบปรับค่าน้ำหนักบนสายพันธุกรรม

Seq. No.	**Seq. Length**	**Frequency Feature Vector (FV)**							**Mean**
		AA	**AR**	**AN**	**…**	**VW**	**VY**	**VV**	
Seq-1	86	5.88	1.18	3.53	…	2.35	1.18	1.18	2.42
Seq-2	69	7.35	8.82	1.47	…	0.00	4.41	7.35	2.04
Seq-3	96	8.42	8.42	9.47	…	0.00	5.26	3.16	3.01
Seq-4	76	1.33	9.33	1.33	…	2.67	10.67	9.33	2.85
Seq-5	**73**	**2.78**	**2.78**	**1.39**	**…**	**1.39**	**1.39**	**2.78**	**3.52**
Seq-6	95	1.06	7.45	0.00	…	9.57	0.00	3.19	2.11
Seq-7	83	3.66	3.66	6.10	…	2.44	8.54	0.00	3.19
Seq-8	87	1.16	9.30	0.00	…	5.81	9.30	0.00	1.67
Seq-9	88	6.90	1.15	1.15	…	1.15	6.90	0.00	3.10
Seq-10	93	4.35	8.70	3.26	…	6.52	7.61	6.52	3.35

3.1.1.5 เก็บค่าน้ำหนักคู่ลำดับกรดอะมิโนของทุกสายพันธุกรรม

ในการเก็บค่าน้ำหนักคู่ลำดับกรดอะมิโนสามารถทำได้ 2 กรณี คือ กรณีที่ไม่มีการปรับค่าน้ำหนักของคู่ลำดับกรดอะมิโน (Non-weight Balance Frequency: NBL) จะแทนด้วยค่าความถี่จริงที่นับได้ (FV) ลงในตารางค่าน้ำหนัก และกรณีที่มีการปรับค่าน้ำหนักของคู่ลำดับกรดอะมิโน (Weight Balance Frequency: WBL) จะแทนด้วยค่าข้อมูลที่ได้ลงในตารางค่าน้ำหนัก โดยข้อมูลในตารางจะถูกนำไปใช้สำหรับการสกัดหาลักษณะเฉพาะในขั้นตอนหลักที่ 2 ต่อไป

ขั้นตอนหลักที่ 2 การสกัดลักษณะเฉพาะบนสายพันธุกรรม
(Feature Extraction Process)

2.1 อ่านค่าน้ำหนักคู่ลำดับกรดอะมิโนของทุกสายพันธุกรรมที่นำเข้า
2.2 วิเคราะห์ข้อมูลระดับสายพันธุกรรม (Local Sequence Analyze)
  2.2.1 คำนวณค่าน้ำหนักเฉลี่ยของทุกสายพันธุกรรม (Average Frequency of All Sequence: μ)
  2.2.3 กำหนดสถานะคู่ลำดับกรดอะมิโน (Status of Feature Vector: ST) บนสายพันธุกรรม
    โดยมีเงื่อนไข คือ
    ถ้า $WBL(k,V_{i,j},W_{i,j}) \geq \mu$ แล้ว กำหนดสถานะคู่ลำดับ เป็น 1
    ถ้า $WBL(k,V_{i,j},W_{i,j}) < \mu$ แล้ว กำหนดสถานะคู่ลำดับ เป็น 0
2.3 คำนวณจำนวนครั้งของคู่ลำดับกรดอะมิโนบนชุดข้อมูลนำเข้า (Frequency Instances Value: FIV)
2.4 คำนวณค่าร้อยละของคู่ลำดับกรดอะมิโน (Frequency Percentage Value : FPV)
2.5 วิเคราะห์ข้อมูลระดับชุดข้อมูลนำเข้า (Global Sequence Analyze)
  2.5.1 กำหนดค่าร้อยละการยอมรับได้ (Support Threshold Point: λ)
  2.5.2 คำนวณสถานะการเลือกคู่ลำดับกรดอะมิโน (Accept Percentage Value: APV)
    โดยมีเงื่อนไข คือ
    ถ้า $FPV_{i,j} \geq \lambda$ แล้ว กำหนดสถานะการเลือกคู่ลำดับ เป็น 1
    ถ้า $FPV_{i,j} < \lambda$ แล้ว กำหนดสถานะการเลือกคู่ลำดับ เป็น 0
2.6 เก็บคู่ลำดับที่สกัดได้ในเมทริกซ์ลักษณะเฉพาะ

ภาพประกอบ 3.7 รายละเอียดขั้นตอนการสกัดลักษณะเฉพาะบนสายพันธุกรรม

3.1.2 ขั้นตอนหลักที่ 2 การสกัดลักษณะเฉพาะบนสายพันธุกรรม

จากภาพประกอบ 3.7 แสดงรายละเอียดการสกัดลักษณะเฉพาะ (Feature Extraction Process) ของคู่ลำดับกรดอะมิโนซึ่งปรากฏในสายพันธุกรรมแต่ละสายด้วยค่าน้ำหนักของคู่ลำดับที่ได้จากขั้นตอนหลักที่ 1 สามารถแสดงรายละเอียดตามลำดับดังนี้

3.1.2.1 อ่านค่าน้ำหนักของคู่ลำดับกรดอะมิโน

ทำการอ่านค่าน้ำหนักคู่ลำดับกรดอะมิโนของทุกสายพันธุกรรมบนชุดข้อมูลที่นำเข้า และกำหนดให้ตัวแปร K แทน จำนวนของสายพันธุกรรมทั้งหมดที่บนชุดข้อมูลนำเข้า

3.1.2.2 วิเคราะห์ข้อมูลระดับสายพันธุกรรม (Local Sequences Analyze)

การวิเคราะห์ข้อมูลในระดับสายพันธุกรรม เป็นขั้นตอนการกำหนดค่าสถานะของคู่ลำดับกรดอะมิโนบนสายพันธุกรรม ประกอบด้วย 2 ขั้นตอนย่อย คือ

3.1.2.2.1 คำนวณค่าน้ำหนักเฉลี่ยของสายพันธุกรรมทั้งหมดที่นำเข้า (Average Frequency of All Sequence: μ)

กำหนดให้ μ แทน ค่าน้ำหนักเฉลี่ยชุดข้อมูลนำเข้า คำนวณดังสมการ (3.4)

$$\mu = \frac{\sum_{k=1}^{K} M_k}{K} \qquad (3.4)$$

เมื่อ ตัวแปร K แทน จำนวนสายพันธุกรรมทั้งหมด ตัวแปร k แทน ลำดับของสายพันธุกรรม และตัวแปร M_k แทน ค่าน้ำหนักเฉลี่ยของสายลำดับที่ k

และสามารถแสดงขั้นตอนการคำนวณได้จากภาพประกอบ 3.8

ขั้นตอนวิธี	:	คำนวณค่าน้ำหนักเฉลี่ยของสายพันธุกรรม
รายละเอียด	:	กำหนดให้ตัวแปร K แทนจำนวนสายพันธุกรรมทั้งหมด, ตัวแปร k เก็บค่าลำดับของสายพันธุกรรม, ตัวแปร Mean(k) เก็บค่าน้ำหนักเฉลี่ยของสายพันธุกรรมที่ k, ตัวแปร SumMean เก็บค่าน้ำหนักรวมของสายพันธุกรรมทั้งหมด และตัวแปร μ เก็บค่าน้ำหนักเฉลี่ยของชุดข้อมูล
การทำงาน	:	คำนวณค่าน้ำหนักเฉลี่ยของสายพันธุกรรม ดังนี้ SumMean = 0 FOR k = 1 TO K SumMean = SumMean + Mean(k) k = k + 1 END LOOP μ = SumMean / K

ภาพประกอบ 3.8 ขั้นตอนวิธีการคำนวณค่าน้ำหนักเฉลี่ยของชุดข้อมูล

จากขั้นตอนการคำนวณในภาพประกอบ 3.8 และสมการ 3.4 สามารถคำนวณได้ 2 ลักษณะ คือ ค่าน้ำหนักเฉลี่ยของข้อมูลแบบไม่ปรับค่าน้ำหนัก (NBL) แทนด้วย μ_1 และค่าน้ำหนักเฉลี่ยของข้อมูลแบบปรับค่าน้ำหนัก (WBL) แทนด้วย μ_2 สมมติชุดข้อมูลนำเข้าที่ใช้ มีจำนวนสายพันธุกรรมทั้งสิ้น 10 สาย (K = 10) และใช้ข้อมูลค่าเฉลี่ยของสายพันธุกรรมแต่ละสายจากตารางที่ 3.2 ซึ่งเป็นข้อมูลซึ่งมีการปรับค่าน้ำหนัก สามารถคำนวณค่าน้ำหนักเฉลี่ยของชุดข้อมูล (μ_2) ได้เท่ากับ 2.73 หรือคำนวณจาก

$$\mu_2 = \frac{2.42 + 2.04 + 3.01 + ... + 3.35}{10}$$

$$\mu_2 = \frac{27.27}{10}$$

$$\mu_2 = 2.73$$

3.1.2.2.2 คำนวณค่าสถานะของคู่ลำดับกรดอะมิโนบนสายพันธุกรรม (Status of Feature Vector: ST)

ในการคำนวณค่าสถานะของคู่ลำดับกรดอะมิโนบนสายพันธุกรรม สามารถแบ่งเป็น 2 กรณี คือ กรณีใช้ข้อมูลค่าความถี่คู่ลำดับแบบไม่ปรับค่าน้ำหนัก (NBL) จะคำนวณด้วยสมการ (3.5) และกรณีใช้ข้อมูลค่าความถี่คู่ลำดับแบบปรับค่าน้ำหนัก (WBL) จะคำนวณด้วยสมการ (3.6) กำหนดให้ $ST(k,V_{i,j},S_{i,j})$ แทน ค่าสถานะของคู่ลำดับกรดอะมิโนบนสายพันธุกรรม

$$ST(k,V_{i,j},S_{i,j}) = \begin{cases} 1, \text{ if } FV(k,V_{i,j},F_{i,j}) \geq \mu_1 \\ 0, \text{Otherwise.} \end{cases} \quad (3.5)$$

$$ST(k,V_{i,j},S_{i,j}) = \begin{cases} 1, \text{if } WBL(k,V_{i,j},W_{i,j}) \geq \mu_2 \\ 0, \text{Otherwise.} \end{cases} \quad (3.6)$$

เมื่อ ตัวแปร $FV(k,V_{i,j},F_{i,j})$ แทน ค่าความถี่คู่ลำดับกรดอะมิโนแบบไม่ปรับค่าน้ำหนักตำแหน่งแถวที่ i หลักที่ j บนเมทริกซ์คู่ลำดับกรดอะมิโนของสายพันธุกรรมลำดับที่ k และ ตัวแปร $WBL(k,V_{i,j},W_{i,j})$ แทน ค่าค่าความถี่คู่ลำดับกรดอะมิโนแบบปรับค่าน้ำหนักตำแหน่งแถวที่ i หลักที่ j บนเมทริกซ์คู่ลำดับกรดอะมิโนของสายพันธุกรรมลำดับที่ k

สามารถแสดงขั้นตอนการคำนวณดังภาพประกอบ 3.9

ขั้นตอนวิธี : กำหนดค่าสถานะของคู่ลำดับกรดอะมิโนบนสายพันธุกรรม

รายละเอียด : กำหนดให้ตัวแปร K แทนจำนวนสายพันธุกรรมทั้งหมด, ตัวแปร k เก็บค่าลำดับของสายพันธุกรรม,ตัวแปร μ_1 เก็บค่าน้ำหนักเฉลี่ยของชุดข้อมูลแบบ NBL, μ_2 เก็บค่าน้ำหนักเฉลี่ยของชุดข้อมูลแบบ WBL, ตัวแปร $FV(k,V_{i,j},F_{i,j})$ เก็บค่าความถี่คู่ลำดับแบบ NBL ของสายพันธุกรรมที่ k, ตัวแปร $WBL(k,V_{i,j},W_{i,j})$ เก็บค่าความถี่คู่ลำดับแบบ WBL ของสายพันธุกรรมที่ k และตัวแปร $ST(k,V_{i,j},S_{i,j})$ เก็บค่าสถานะของแต่ละคู่ลำดับบนสายพันธุกรรมที่ k

การทำงาน : กำหนดค่าสถานะของคู่ลำดับกรดอะมิโน ดังนี้

Case : Using Non-weight Balance Frequency (NBL)

FOR k = 1 TO K
FOR i = 1 TO |T|
FOR j = 1 TO |T|
If $FV(k,V_{i,j},F_{i,j}) \geq \mu_1$ Then $ST(k,V_{i,j},S_{i,j}) = 1$
Else $ST(k,V_{i,j},S_{i,j}) = 0$
j = j+1
END LOOP
i = i + 1
END LOOP
k = k + 1
END LOOP

Case : Using Weight Balance Frequency (WBL)

FOR k = 1 TO K
FOR i = 1 TO |T|
FOR j = 1 TO |T|
If $WBL(k,V_{i,j},W_{i,j}) \geq \mu_2$ Then $ST(k,V_{i,j},S_{i,j}) = 1$
Else $ST(k,V_{i,j},S_{i,j}) = 0$
j = j + 1
END LOOP
i = i + 1
END LOOP
k = k + 1
END LOOP

ภาพประกอบ 3.9 ขั้นตอนวิธีการกำหนดค่าสถานะของคู่ลำดับกรดอะมิโน

ขั้นตอนวิธีการกำหนดค่าสถานะคู่ลำดับกรดอะมิโน จากภาพประกอบ 3.9 สามารถกำหนดค่าสถานะของคู่ลำดับบนสายพันธุกรรมด้วยค่าน้ำหนักเฉลี่ยของชุดข้อมูล (μ) เป็น 2 กรณี คือ

กรณีใช้ข้อมูลค่าความถี่คู่ลำดับแบบไม่ปรับค่าน้ำหนัก (NBL) คำนวณด้วยเงื่อนไขของสมการ (3.5) คือ

เงื่อนไขที่ 1: ถ้า $FV(k,V_{i,j},F_{i,j})$ มีค่ามากกว่าหรือเท่ากับ μ_1
แล้ว กำหนดค่า $ST(k,V_{i,j},S_{i,j})$ ของคู่ลำดับ เป็น 1

เงื่อนไขที่ 2: ถ้า $FV(k,V_{i,j},F_{i,j})$ มีค่าน้อยกว่า μ_1
แล้ว กำหนดค่า $ST(k,V_{i,j},S_{i,j})$ ของคู่ลำดับ เป็น 0

กรณีใช้ข้อมูลค่าความถี่คู่ลำดับแบบปรับค่าน้ำหนัก (WBL) คำนวณด้วยเงื่อนไขของสมการ (3.6) คือ

เงื่อนไขที่ 1: ถ้า $WBL(k,V_{i,j},W_{i,j})$ มีค่ามากกว่าหรือเท่ากับ μ_2
แล้ว กำหนดค่า $ST(k,V_{i,j},S_{i,j})$ ของคู่ลำดับ เป็น 1

เงื่อนไขที่ 2: ถ้า $WBL(k,V_{i,j},W_{i,j})$ มีค่าน้อยกว่า μ_2
แล้ว กำหนดค่า $ST(k,V_{i,j},S_{i,j})$ ของคู่ลำดับ เป็น 0

จากข้อมูลในตารางที่ 3.2 จะเลือกใช้เงื่อนไขของสมการ (3.6) เนื่องจากเป็นข้อมูลซึ่งมีการปรับค่าน้ำหนัก และสามารถแสดงการคำนวณค่าสถานะของคู่ลำดับกรดอะมิโนบนสายพันธุกรรมลำดับที่ 5 ด้วยค่าน้ำหนักเฉลี่ยของชุดข้อมูลเท่ากับ 2.73 (μ_2=2.73) พบว่า ค่าน้ำหนักของคู่ลำดับ AA ซึ่งมีค่าเท่ากับ 2.78 หรือ WBL(5,AA,2.78) มีค่ามากกว่าค่า μ_2 ดังนั้นค่าสถานะของคู่ลำดับกรดอะมิโน AA จะมีค่าเป็น 1 หรือ ST(5,AA,1) และจากเงื่อนไขเดียวกัน จะได้ ST(5,AR,1) ST(5,AN,0) … ST(5,VW,0) ST(5,VY,0) และ ST(5,VV,1) ตามลำดับ โดยจะเห็นว่าค่าสถานะของคู่ลำดับ AN VW และ VY มีค่าสถานะเป็น 0 เนื่องจาก ค่าน้ำหนักของคู่ลำดับกรดอะมิโนเหล่านี้มีค่าน้อยกว่าค่า μ_2 และค่าสถานะของคู่ลำดับ VY และ VV มีค่าสถานะเป็น 1 เนื่องจากมีค่าน้ำหนักของคู่ลำดับมากกว่าค่า μ_2 นั้นเอง และสามารถแสดงค่าสถานะของคู่ลำดับบนสายพันธุกรรมทั้งหมด และสามารถแสดงค่าสถานะของคู่ลำดับกรดอะมิโนบนสายพันธุกรรมแต่ละสาย ดังตารางที่ 3.3

ตารางที่ 3.3 ค่าสถานะของคู่ลำดับกรดอะมิโน เมื่อค่าน้ำหนักเฉลี่ยของชุดข้อมูลเท่ากับ 2.73

Seq. No.	Status of Feature Vector (ST)						
	AA	**AR**	**AN**	**...**	**VW**	**VY**	**VV**
Seq-1	1	0	1	…	0	0	0
Seq-2	1	1	0	…	0	1	1
Seq-3	1	1	1	…	0	1	1
Seq-4	0	1	0	…	0	1	1
Seq-5	**1**	**1**	**0**	**…**	**0**	**0**	**1**
Seq-6	0	1	0	…	1	0	1
Seq-7	1	1	1	…	0	1	0
Seq-8	0	1	0	…	1	1	0
Seq-9	1	0	0	…	0	1	0
Seq-10	1	1	1	…	1	1	1

3.1.2.3 คำนวณจำนวนครั้งของคู่ลำดับกรดอะมิโนบนสายพันธุกรรม (Frequency Instances Value: FIV)

การคำนวณจำนวนครั้งของคู่ลำดับกรดอะมิโน ซึ่งที่มีค่าสถานะ (ST) เท่ากับ 1 บนสายพันธุกรรมทุกสายในชุดข้อมูลนำเข้า โดยกำหนดให้ $FIV_{i,j}$ แทนผลรวมของคู่ลำดับกรดอะมิโนที่มีค่าสถานะเท่ากับ 1 คำนวณดังสมการ (3.7)

$$FIV_{i,j} = \sum_{k=1}^{K} ST(k, V_{i,j}, S_{i,j}) \tag{3.7}$$

เมื่อ ตัวแปร K แทน จำนวนสายพันธุกรรมทั้งหมด ตัวแปร k แทน ค่าลำดับของสายพันธุกรรม ตัวแปร $ST(k,V_{i,j},S_{i,j})$ แทน ค่าสถานะของคู่ลำดับตำแหน่งแถวที่ i หลักที่ j บนตารางค่าสถานะคู่ลำดับกรดอะมิโน

สามารถแสดงขั้นตอนวิธีการคำนวณได้ดังภาพประกอบ 3.10

ขั้นตอนวิธี : คำนวณผลรวมค่าสถานะของคู่ลำดับกรดอะมิโนบนสายพันธุกรรม

รายละเอียด : กำหนดให้ตัวแปร K แทนจำนวนสายพันธุกรรมทั้งหมด, ตัวแปร k เก็บค่าลำดับของสายพันธุกรรม, ตัวแปร $ST(k,V_{i,j},S_{i,j})$ ค่าสถานะของแต่ละคู่ลำดับบนสายพันธุกรรมที่ k และตัวแปร $FIV_{i,j}$ เก็บผลรวมของคู่ลำดับที่มีค่าสถานะเป็น 1 บนสายพันธุกรรมที่ k

การทำงาน : คำนวณผลรวมค่าสถานะของคู่ลำดับ ดังนี้

```
FOR i = 1 TO |T|
    FOR j = 1 TO |T|
        FIV(i,j) = 0
        FOR k = 1 TO K
            If ST(k,V(i,j),S(i,j)) = 1 Then FIV(i,j) = FIV(i,j) + 1
            k = k + 1
        END LOOP
        j = j + 1
    END LOOP
    i = i + 1
END LOOP
```

ภาพประกอบ 3.10 ขั้นตอนวิธีการคำนวณผลรวมค่าสถานะของคู่ลำดับกรดอะมิโน

จากขั้นตอนการคำนวณในภาพประกอบ 3.10 สามารถแสดงผลลัพธ์จากการคำนวณผลรวมของคู่ลำดับที่มีสถานะเป็น 1 บนชุดข้อมูล ได้ดังตารางที่ 3.4

ตารางที่ 3.4 ค่าผลรวมคู่ลำดับกรดอะมิโนที่มีค่าสถานะสถานะเท่ากับ 1

Seq. No.	**Status of Feature Vector (ST)**						
	AA	**AR**	**AN**	**…**	**VW**	**VY**	**VV**
Seq-1	1	0	1	…	0	0	0
Seq-2	1	1	0	…	0	1	1
Seq-3	1	1	1	…	0	1	1
Seq-4	0	1	0	…	0	1	1
Seq-5	1	1	0	…	0	0	1
Seq-6	0	1	0	…	1	0	1
Seq-7	1	1	1	…	0	1	0
Seq-8	0	1	0	…	1	1	0
Seq-9	1	0	0	…	0	1	0
Seq-10	1	1	1	…	1	1	1
FIV	**7**	**8**	**4**	**…**	**3**	**7**	**6**

จากตารางที่ 3.4 พบว่า คู่ลำดับกรดอะมิโน AA มีผลรวมค่าสถานะเท่ากับ 7 หรือเขียนแทนด้วย FIV(AA) = 7 ซึ่งมีความหมายว่ามีสายพันธุกรรมที่มีค่าสถานะของคู่ลำดับ AA เท่ากับ 1 จำนวน 7 สาย และจากการคำนวณแบบเดียวกัน จะได้ FIV(AR) = 8 FIV(AN) = 4 … FIV(VW) = 3 FIV(VY) = 7 และ FIV(VV) = 6 ตามลำดับ

3.1.2.4 คำนวณค่าร้อยละของคู่ลำดับกรดอะมิโน (Frequency Percentage Value: FPV)

การคำนวณร้อยละของคู่ลำดับกรดอะมิโน จะคำนวณด้วยค่าผลรวมจำนวนครั้ง (FIV) ของคู่ลำดับที่มีค่าสถานะเท่ากับ 1 กำหนดให้ $FPV_{i,j}$ แทน ค่าร้อยละของคู่ลำดับกรดอะมิโน คำนวณดังสมการ (3.8)

$$FPV_{i,j} = \frac{FIV_{i,j}}{K} \times 100 \tag{3.8}$$

เมื่อ ตัวแปร $FIV_{i,j}$ แทน ผลรวมของคู่ลำดับกรดอะมิโนแถวที่ i หลักที่ j ซึ่งมีค่าสถานะเท่ากับ 1 และตัวแปร K แทน จำนวนของสายพันธุกรรมทั้งหมด

สามารถแสดงขั้นตอนวิธีการคำนวณได้จากภาพประกอบ 3.11

ขั้นตอนวิธี : การคำนวณค่าร้อยละของคู่ลำดับกรดอะมิโน (FPV)

รายละเอียด : กำหนดให้ตัวแปร K แทนจำนวนสายพันธุกรรมทั้งหมด, ตัวแปร k เก็บค่าลำดับของสายพันธุกรรม, ตัวแปร $FIV_{i,j}$ เก็บผลรวมของคู่ลำดับที่มีค่าสถานะเป็น 1 บนสายพันธุกรรมสายที่ k, ตัวแปร $FPV_{i,j}$ เก็บค่าร้อยละของคู่ลำดับ

การทำงาน : วิเคราะห์ระดับสายพันธุกรรม ดังนี้

```
FOR i = 1  TO |T|
    FOR j = 1  TO |T|
            FPV_{i,j} = (FIV_{i,j} / K) * 100
            j = j+1
        END LOOP
        i = i+1
END LOOP
```

ภาพประกอบ 3.11 ขั้นตอนวิธีการคำนวณค่าร้อยละของคู่ลำดับกรดอะมิโน

จากภาพประกอบ 3.11 สามารถคำนวณค่าร้อยละของคู่ลำดับที่มีสถานะเป็น 1 ด้วยค่า FIV ของคู่ลำดับกรดอะมิโนจากตารางที่ 3.4 ได้ดังนี้

$$FPV(AA) = \frac{7}{10} \times 100 = 70\%$$

$$FPV(AR) = \frac{8}{10} \times 100 = 80\%$$

$$FPV(AN) = \frac{4}{10} \times 100 = 40\%$$

$$\vdots$$

$$FPV(VW) = \frac{3}{10} \times 100 = 30\%$$

$$FPV(VY) = \frac{7}{10} \times 100 = 70\%$$

$$FPV(VV) = \frac{6}{10} \times 100 = 60\%$$

จากการคำนวณจะได้ค่าร้อยละของคู่ลำดับ AA มีค่าเท่ากับ 70% หรือ FPV(AA) = 70% และจากการคำนวณแบบเดียวกัน จะได้ FPV(AR) = 80% FPV(AN) = 40% … FPV(VW) = 30% FPV(VY) = 70% และ FPV(VV) = 60% ตามลำดับ และแสดงตารางค่าร้อยละของคู่ลำดับได้ดังตารางที่ 3.5

ตารางที่ 3.5 ค่าร้อยละของคู่ลำดับกรดอะมิโน (FPV)

Seq. No.	**Status of Feature Vector (ST)**						
	AA	**AR**	**AN**	**…**	**VW**	**VY**	**VV**
Seq-1	1	0	1	…	0	0	0
Seq-2	1	1	0	…	0	1	1
Seq-3	1	1	1	…	0	1	1
Seq-4	0	1	0	…	0	1	1
Seq-5	1	1	0	…	0	0	1
Seq-6	0	1	0	…	1	0	1
Seq-7	1	1	1	…	0	1	0
Seq-8	0	1	0	…	1	1	0
Seq-9	1	0	0	…	0	1	0
Seq-10	1	1	1	…	1	1	1
FIV	**7**	**8**	**4**	**…**	**3**	**7**	**6**
FPV (%)	**70**	**80**	**40**	**…**	**30**	**70**	**60**

3.1.2.5 วิเคราะห์ข้อมูลระดับชุดข้อมูล (Global Sequences Analyze)

การวิเคราะห์ระดับชุดข้อมูล เป็นการพิจารณาค่าร้อยละของคู่ลำดับกรดอะมิโน (FPV) เพื่อสกัดคู่ลำดับกรดอะมิโนซึ่งเป็นลักษณะเฉพาะของชุดข้อมูลสำหรับสร้างเป็นคุณสมบัติของข้อมูล (Attributes) ประกอบด้วย 2 ขั้นตอนย่อย ดังนี้

3.1.2.5.1 กำหนดค่าร้อยละของการยอมรับได้ของคู่ลำดับบนชุดข้อมูลนำเข้า (Support Threshold Point: λ) โดยกำหนดให้ λ แทน ค่าร้อยละของการยอมรับได้ของคู่ลำดับบนชุดข้อมูลนำเข้า โดยกำหนดเป็นเงื่อนไขในการสกัดคู่ลำดับ ซึ่งเป็นลักษณะเฉพาะของคู่ลำดับที่ใช้สำหรับสร้างเป็นคุณสมบัติข้อมูล (Attributes)

3.1.2.5.2 คำนวณสถานะการเลือกคู่ลำดับกรดอะมิโน (Accept Percentage Value: APV) โดยกำหนดให้ $APV_{i,j}$ แทน ค่าสถานะการเลือกคู่ลำดับกรดอะมิโน คำนวณดังสมการ (3.9)

$$APV_{i,j} = \begin{cases} 1, \text{if } FPV_{i,j} \geq \lambda \\ 0, \text{Otherwise.} \end{cases} \quad (3.9)$$

เมื่อ λ แทน ค่าร้อยละของการยอมรับได้ของคู่ลำดับบนชุดข้อมูลนำเข้า ตัวแปร $FPV_{i,j}$ แทน ค่าร้อยละของคู่ลำดับตำแหน่งแถวที่ i หลักที่ j

และแสดงการคำนวณได้จากภาพประกอบ 3.12

ขั้นตอนวิธี : วิเคราะห์ข้อมูลระดับชุดข้อมูล (Global Sequences Analyze)

รายละเอียด : กำหนดให้ตัวแปร $FPV_{i,j}$ เก็บค่าร้อยละของคู่ลำดับกรดอะมิโน, ตัวแปร λ เก็บค่าร้อยละของการยอมรับได้ และตัวแปร $APV_{i,j}$ เก็บค่าสถานะการเลือกคู่ลำดับ

การทำงาน : วิเคราะห์ระดับชุดข้อมูล ดังนี้

```
Set Percentage of Support Threshold Point (λ)
FOR i = 1 TO |T|
    FOR j = 1 TO |T|
            If  FPV_{i,j} ≥ λ Then APV_{i,j} = 1
            Else APV_{i,j} = 0
            j = j + 1
     END LOOP
     i = i + 1
END LOOP
```

ภาพประกอบ 3.12 ขั้นตอนวิเคราะห์ข้อมูลระดับชุดข้อมูล

จากขั้นตอนการคำนวณในภาพประกอบ 3.12 การสกัดคู่ลำดับกรดอะมิโนแต่ละตำแหน่งบนตารางค่าร้อยละของคู่ลำดับกรดอะมิโน (FPV) ด้วยค่าร้อยละของการยอมรับได้ของคู่ลำดับกรดอะมิโนบนชุดข้อมูลนำเข้า (λ) โดยกำหนดเงื่อนไขการคำนวณจากสมการ 3.9 คือ

เงื่อนไขที่ 1: ถ้าค่า $FPV_{i,j}$ มีค่ามากกว่าหรือเท่ากับ λ
แล้ว กำหนดค่าสถานะ $APV_{i,j}$ ของคู่ลำดับ เป็น 1

เงื่อนไขที่ 2: ถ้าค่า $FPV_{i,j}$ มีค่าน้อยกว่า λ
แล้ว กำหนดค่าสถานะ $APV_{i,j}$ ของคู่ลำดับ เป็น 0

จากเงื่อนไขสามารถสกัดคู่ลำดับกรดอะมิโน โดยถ้ากำหนดค่าร้อยละของคู่ลำดับบนชุดข้อมูลนำเข้า (λ) เท่ากับ 60% หรือ $\lambda = 60\%$ ผลลัพธ์ที่สกัดได้สามารถแสดงดังตารางที่ 3.6

ตารางที่ 3.6 คู่ลำดับกรดอะมิโนที่สกัดได้ด้วยค่าร้อยละการยอมรับได้ (λ) เท่ากับ 60%

Seq. No.	**Status of Feature Vector (ST)**						
	AA	AR	AN	...	VW	VY	VV
Seq-1	1	0	1	…	0	0	0
Seq-2	1	1	0	…	0	1	1
Seq-3	1	1	1	…	0	1	1
Seq-4	0	1	0	…	0	1	1
Seq-5	1	1	0	…	0	0	1
Seq-6	0	1	0	…	1	0	1
Seq-7	1	1	1	…	0	1	0
Seq-8	0	1	0	…	1	1	0
Seq-9	1	0	0	…	0	1	0
Seq-10	1	1	1	…	1	1	1
FIV	**7**	**8**	**4**	**…**	**3**	**7**	**6**
FPV (%)	**70**	**80**	**40**	**…**	**30**	**70**	**60**
APV	**1**	**1**	**0**	**…**	**0**	**1**	**1**

จากตารางที่ 3.6 สามารถอธิบายผลลัพธ์ที่ได้ คือ เมื่อกำหนดค่า $\lambda = 60\%$ จะหมายถึง ถ้าค่าร้อยละของคู่ลำดับกรดอะมิโน (FPV) บนตารางค่าสถานะของชุดข้อมูลมีค่ามากกว่าหรือเท่ากับ 60% หรือ ถ้าค่า $FPV \geq \lambda$ แล้วค่าสถานะการเลือกคู่ลำดับนั้นจะถูกกำหนดค่าเป็น 1 และ ถ้าค่า $FPV < \lambda$ แล้วค่าสถานะการเลือกคู่ลำดับนั้น

จะถูกกำหนดค่าเป็น 0 ในการคำนวณพบว่า ค่าร้อยละของคู่ลำดับ AA หรือ FPV(AA,70) มีค่ามากกว่าหรือเท่ากับ λ ดังนั้น ค่าสถานะการยอมรับของคู่ลำดับเป็น 1 จะเขียนแทนด้วย APV(AA) = 1 และจากเงื่อนไขเดียวกัน จะได้ APV(AR) = 1 APV(AN) = 0 … APV (VW) = 0 APV(VY) = 1 และ APV(VV) = 1 ตามลำดับ

ดังนั้นคู่ลำดับกรดอะมิโนที่สกัดได้และใช้เป็นคุณสมบัติของข้อมูล (Attributes) ในขั้นตอนการจำแนกสายพันธุกรรม ประกอบด้วย คู่ลำดับ AA AR … VY และ VV ตามลำดับ

ขั้นตอนหลักที่ 3 การจำแนกสายพันธุกรรมด้วยลักษณะเฉพาะ (Feature Classification Process)
3.1 อ่านค่าน้ำหนักของคู่ลำดับที่สกัดได้จากชุดสายพันธุกรรมนำเข้า
3.2 แบ่งข้อมูลชุดสอน (Training Set) และข้อมูลชุดทดสอบ (Testing Set) ด้วยวิธีการ k-Fold Cross Validation
3.3 เลือกขั้นตอนวิธีการจำแนก (Classifier) สายพันธุกรรมในโปรแกรม WEKA ดังนี้
3.3.1 โครงข่ายประสาทเทียมแบบ Multilayer Perceptron (MLP)
3.3.2 โครงข่ายประสาทเทียมแบบ Radial Basis Function (RBF)
3.3.3 ต้นไม้การตัดสินใจแบบ Alternating Decision Tree (ADTree)
3.3.4 ต้นไม้การตัดสินใจแบบ Pruned C4.5 Decision Tree (J48)
3.3.5 ต้นไม้การตัดสินใจแบบ K random features Decision Tree (RandomTree)
3.4 จำแนกสายพันธุกรรม (Classification)
3.5 ได้ค่าความถูกต้อง (Accuracy) จากการจำแนกสายพันธุกรรม

ภาพประกอบ 3.13 รายละเอียดขั้นตอนการจำแนกสายพันธุกรรม

3.1.3 ขั้นตอนหลักที่ 3 การจำแนกสายพันธุกรรมด้วยลักษณะเฉพาะ

จากภาพประกอบ 3.13 แสดงรายละเอียดการจำแนกสายพันธุกรรม (Feature Classification Process) จะใช้ค่าของคู่ลำดับกรดอะมิโนซึ่งได้จากผ่านการสกัดลักษณะเฉพาะ (Feature) ของสายพันธุกรรมแต่ละสาย ประกอบด้วย 5 ขั้นตอนย่อย ดังนี้

3.1.3.1 อ่านค่าน้ำหนักของคู่ลำดับที่สกัดได้ (คุณสมบัติข้อมูลที่มีค่าการเลือกเป็น 1) จากสายพันธุกรรมทั้งหมดในชุดข้อมูลนำเข้า

3.1.3.2 แบ่งข้อมูลชุดสอน (Training Set) และชุดทดสอบ (Testing Set) ด้วยวิธี k-Fold Cross Validation

3.1.3.3 เลือกขั้นตอนวิธีการจำแนก (Classifier) โดยใช้ชุดเครื่องบนโปรแกรม WEKA เวอร์ชัน 3.4.10

3.1.3.4 จำแนกสายพันธุกรรม (Classification) ข้อมูลสายพันธุกรรมตามขั้นตอนวิธีที่เลือก

3.1.3.5 ได้ค่าความถูกต้อง (Accuracy) จากการจำแนกสายพันธุกรรมสายพันธุกรรม

สามารถแสดงขั้นตอนวิธีได้ดังภาพประกอบ 3.14

ขั้นตอนวิธี	:	การจำแนกข้อมูล(Classify) ด้วยชุดเครื่องมือบนโปรแกรม WEKA
รายละเอียด	:	กำหนดให้ตัวแปร k เก็บค่าจำนวนการแบ่งชุดข้อมูลด้วยวิธี k-Fold Cross Validation และตัวแปร Choice เก็บค่าตัวเลือกขั้นตอนวิธีสำหรับการจำแนก
การทำงาน	:	จำแนกข้อมูล (Classify) ด้วยชุดเครื่องมือบนโปรแกรม WEKA ดังนี้ Set k Divide Data Set into K Set Using k-Fold Cross Validation Method SELECT Choice of Classifier CASE “MLP” : Classification by MLP Algorithm CASE “RBF” : Classification by RBF Algorithm CASE “ADTree” : Classification by ADTree Algorithm CASE “J48” : Classification by J48 Algorithm CASE “RandomTree” : Classification by RandomTree Algorithm END SELECT Classification with Classifier algorithm Give Accuracy from Classification

ภาพประกอบ 3.14 ขั้นตอนวิธีการจำแนกข้อมูล (Classification) ด้วยโปรแกรม WEKA

จากภาพประกอบ 3.14 หลังจากแบ่งข้อมูลด้วยวิธีการ k-Fold Cross Validation และเลือกขั้นตอนวิธีในการจำแนก ได้แก่ โครงข่ายประสาทเทียมแบบ MLP และแบบ RBF ต้นไม้การตัดสินใจแบบ ADTree J48 และ RandomTree ตามลำดับ

ผลลัพธ์สุดท้ายที่ได้จากการทำงาน คือ ความถูกต้อง (Accuracy) ในการจำแนกข้อมูลสายพันธุกรรม

บทที่ 4

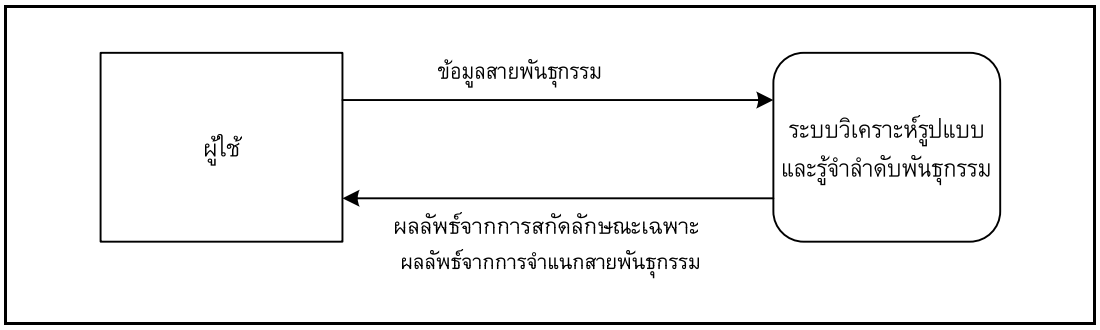
โปรแกรมระบบวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

การพัฒนาโปรแกรมระบบวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม จะใช้แนวคิดและขั้นตอนการทำงานจากแบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม ในบทนี้จะกล่าวถึงการออกแบบและการพัฒนาโปรแกรมวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม ประกอบด้วย แผนภาพกระแสข้อมูลของระบบการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม ผังงานระบบวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม และระบบวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม สามารถอธิบายรายละเอียดตามลำดับได้ดังนี้

4.1 แผนภาพกระแสข้อมูลโปรแกรมวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

ระบบวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม สามารถแสดงแผนภาพอธิบายความสัมพันธ์ของข้อมูลที่ใช้ด้วยแผนภาพกระแสข้อมูล (Data Flow Diagrams) มีรายละเอียดดังนี้

4.1.1 แผนภาพบริบท (Context Diagrams) ของระบบวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม ดังภาพประกอบ 4.1

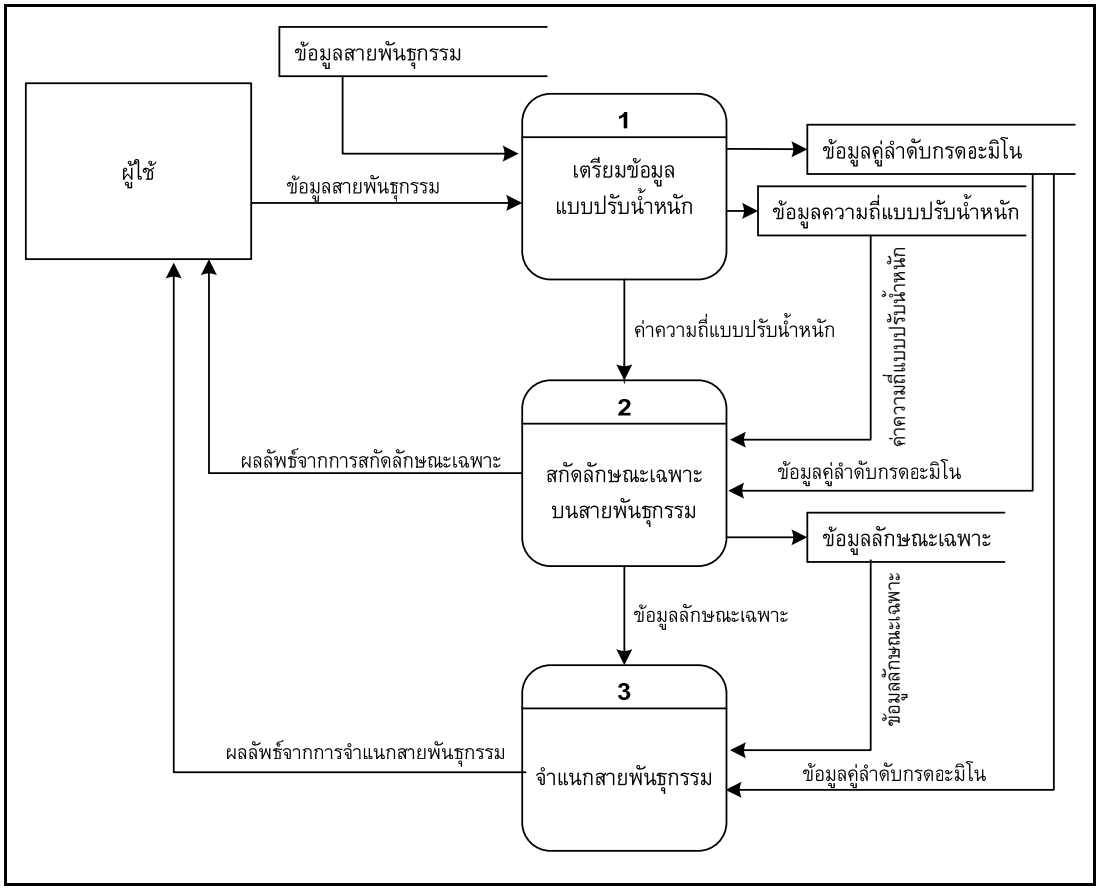


ภาพประกอบ 4.1 ภาพกระแสข้อมูลของระบบวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

จากภาพประกอบ 4.1 ผู้ใช้ (User) จะทำหน้าที่ในการนำข้อมูลสายพันธุกรรม (Import Sequences) เข้าสู่ระบบวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม ผลลัพธ์ที่ผู้ใช้ได้รับคือ ผลลัพธ์จากการสกัดลักษณะเฉพาะ (Feature) บนสายพันธุกรรมและค่าความถูกต้อง (Accuracy) จากการจำแนกสายพันธุกรรมด้วยระบบวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

4.1.2 แผนภาพกระแสข้อมูลระดับที่ 1 ระบบวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

ประกอบด้วย 3 ขั้นตอนหลัก คือ การเตรียมข้อมูลค่าความถี่แบบมีการปรับค่าน้ำหนัก (Weight Balance Frequency Data Process) การสกัดลักษณะเฉพาะบนสายพันธุกรรม (Feature Extraction Process) และการจำแนกสายพันธุกรรมด้วยลักษณะเฉพาะ (Feature Classification Process) ดังภาพประกอบ 4.2

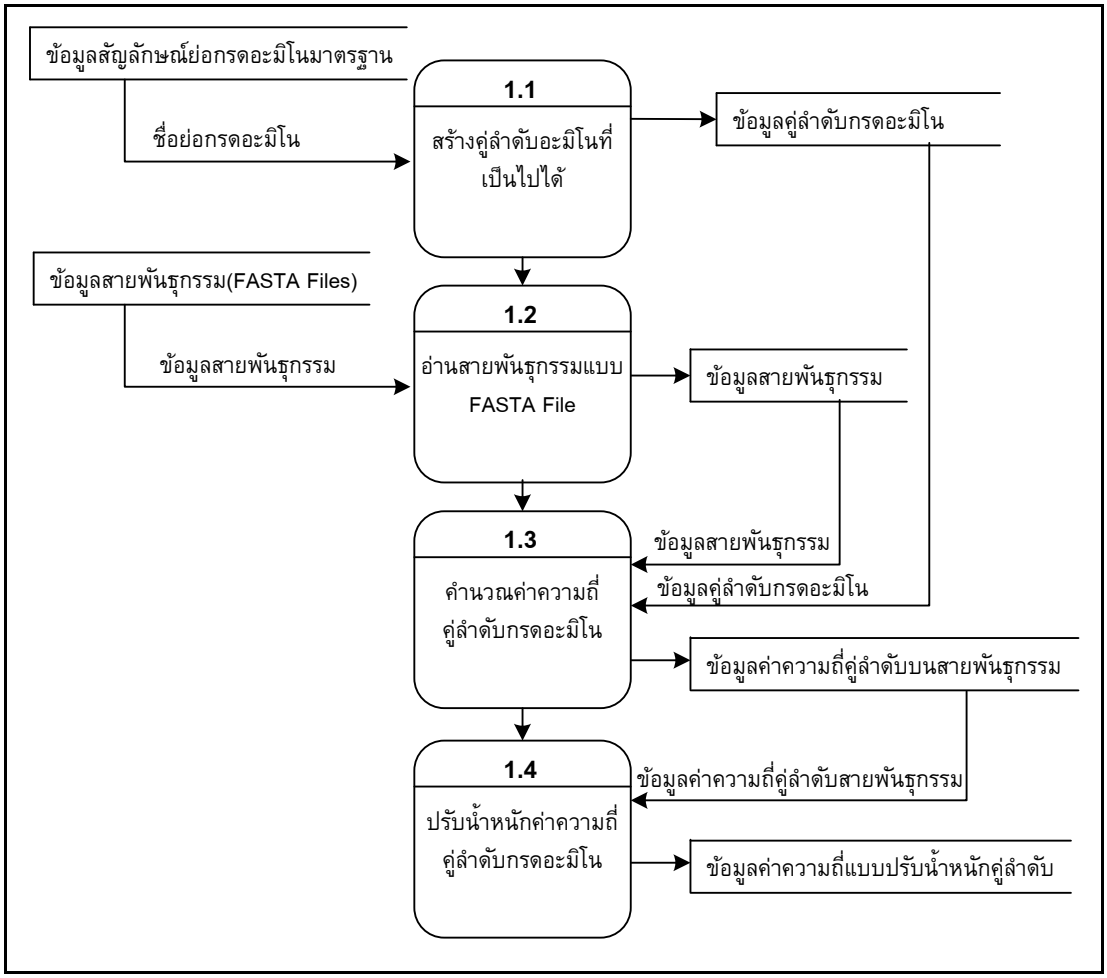


ภาพประกอบ 4.2 ภาพกระแสข้อมูลระดับที่ 1 ระบบวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

จากภาพประกอบ 4.2 หลังจากผู้ใช้นำข้อมูลสายพันธุกรรมเข้าสู่ระบบ (Import Sequences) ขั้นตอนการเตรียมข้อมูลแบบปรับน้ำหนักจะทำหน้าที่เปลี่ยนรูปแบบข้อมูลสายพันธุกรรมให้อยู่ในรูปของข้อมูลความค่าความถี่แบบปรับน้ำหนัก ขั้นตอนการสกัดลักษณะเฉพาะบนสายพันธุกรรม จะทำการสกัดลักษณะเฉพาะ (Feature) และในขั้นตอนการจำแนกสายพันธุกรรม จะนำข้อมูลลักษณะเฉพาะที่ได้มาทำการจำแนก โดยผลลัพธ์สุดท้ายที่ได้จากระบบ คือ ค่าความถูกต้องของการจำแนกสายพันธุกรรม

4.1.3 แผนภาพกระแสข้อมูลระดับที่ 2 ระบบวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

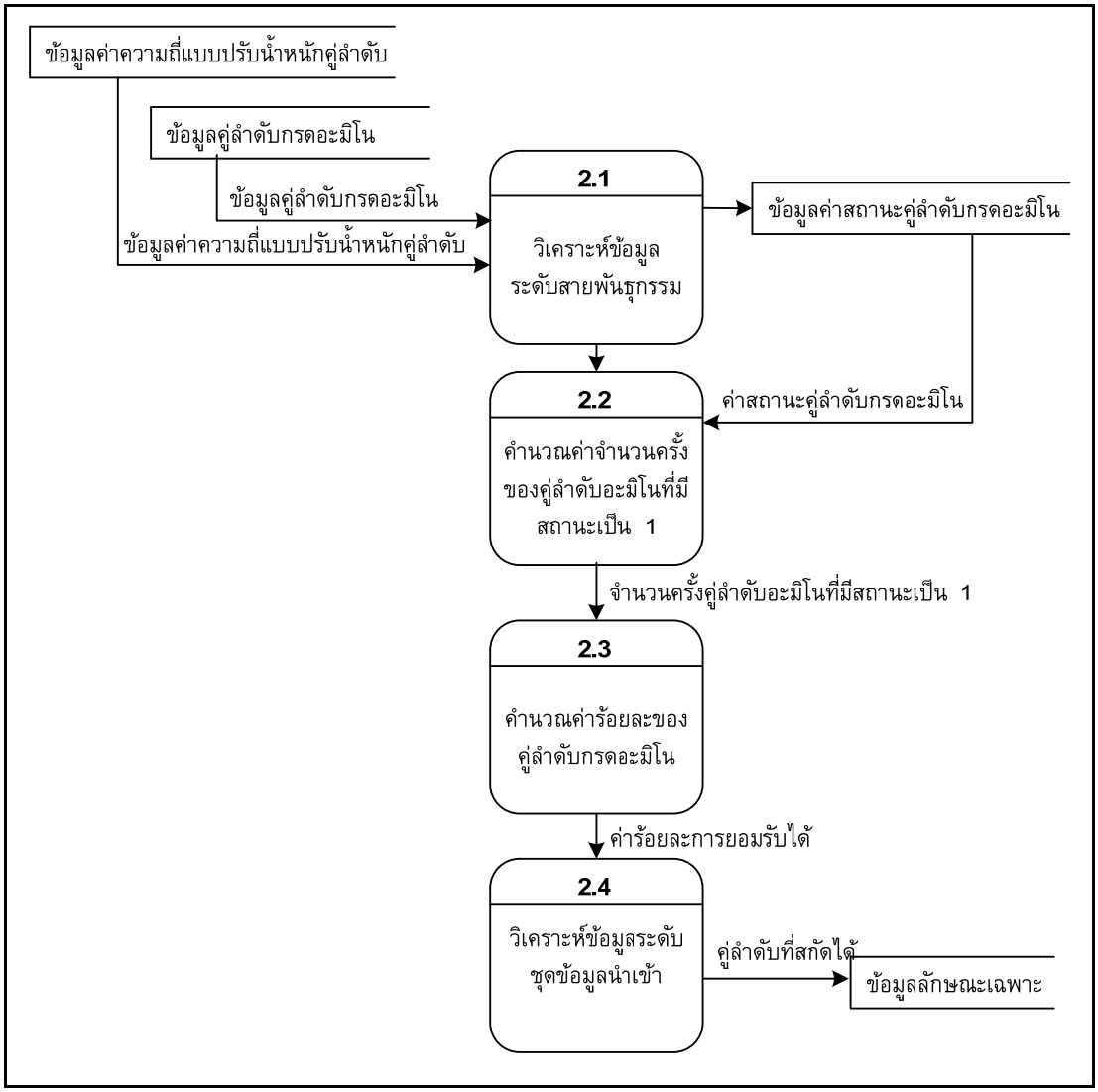
4.1.3.1 แผนภาพกระแสข้อมูลระดับที่ 2 การเตรียมข้อมูลค่าความถี่แบบปรับค่าน้ำหนัก (Weight Balance Frequency Data Process) ดังภาพประกอบ 4.3



ภาพประกอบ 4.3 ภาพกระแสข้อมูลระดับที่ 2 การเตรียมข้อมูลค่าความถี่แบบปรับน้ำหนัก

จากภาพประกอบ 4.3 เมื่อนำข้อมูลสายพันธุกรรมเข้าสู่ระบบ ในส่วนนี้จะทำหน้าที่แปลงข้อมูลสายพันธุกรรมซึ่งถูกเก็บในไฟล์ FASTA ที่อ่านเข้าด้วยการคำนวณค่าความถี่ของคู่ลำดับอะมิโนที่ได้บนสายพันธุกรรมแต่สาย และปรับคุณภาพของข้อมูลให้มีลักษณะเดียวกัน ด้วยวิธีปรับค่าน้ำหนักของค่าความถี่ของคู่ลำดับกรดอะมิโน (WBL)

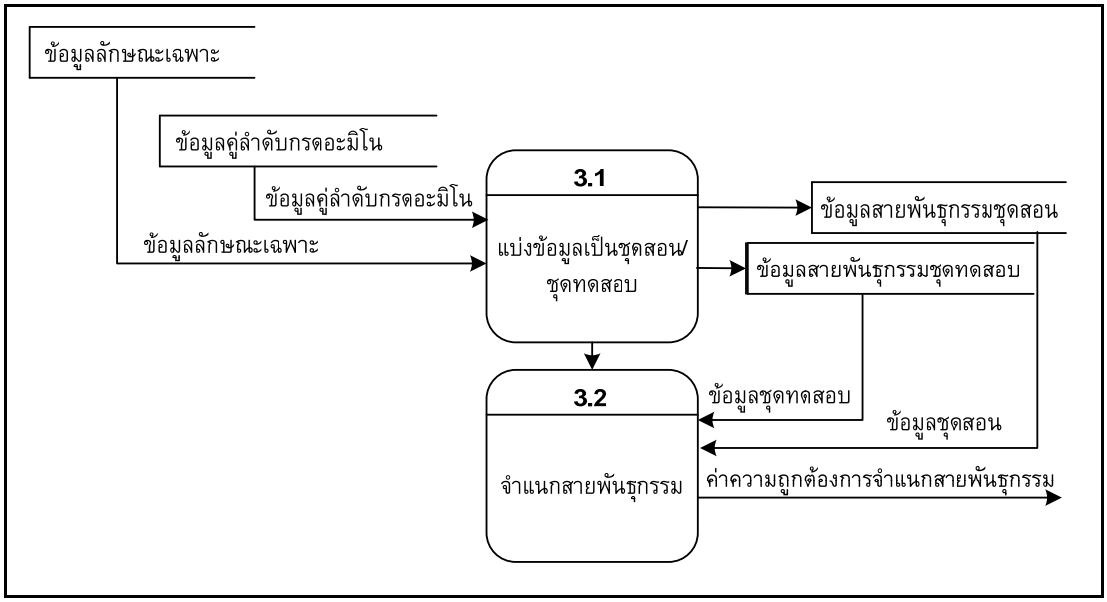
4.1.3.2 แผนภาพกระแสข้อมูลระดับที่ 2 การสกัดลักษณะเฉพาะบนสายพันธุกรรม (Feature Extraction Process) แสดงดังภาพประกอบ 4.4



ภาพประกอบ 4.4 ภาพกระแสข้อมูลระดับที่ 2 การสกัดลักษณะเฉพาะบนสายพันธุกรรม

จากภาพประกอบ 4.4 จะนำข้อมูลค่าความถี่แบบปรับน้ำหนักของคู่ลำดับกรดอะมิโนมาทำการวิเคราะห์ระดับสายพันธุกรรม โดยการหาค่าน้ำหนักเฉลี่ย (μ) และทำการกำหนดค่าสถานะของคู่ละดับที่มีค่ามากกว่าหรือเท่ากับค่าน้ำหนักเฉลี่ย (μ) ให้มีค่าเป็น 1 จากนั้นจึงคำนวณผลรวมจำนวนครั้งของคู่ลำดับที่มีค่าสถานะเป็น 1 (FIV) และคำนวณร้อยละการยอมรับได้ (FPV) และทำการวิเคราะห์ข้อมูลระดับชุดข้อมูลนำเข้าด้วยค่าร้อยละการยอมรับได้ (λ) เพื่อให้ได้ลักษณะเฉพาะ (Feature) ที่ต้องการ และนำข้อมูลที่ได้มาใช้ในขั้นตอนการจำแนกสายพันธุกรรมต่อไป

4.1.3.3 แผนภาพกระแสข้อมูลระดับที่ 2 การจำแนกสายพันธุกรรมด้วยลักษณะเฉพาะ (Feature Classification Process) ดังภาพประกอบ 4.5



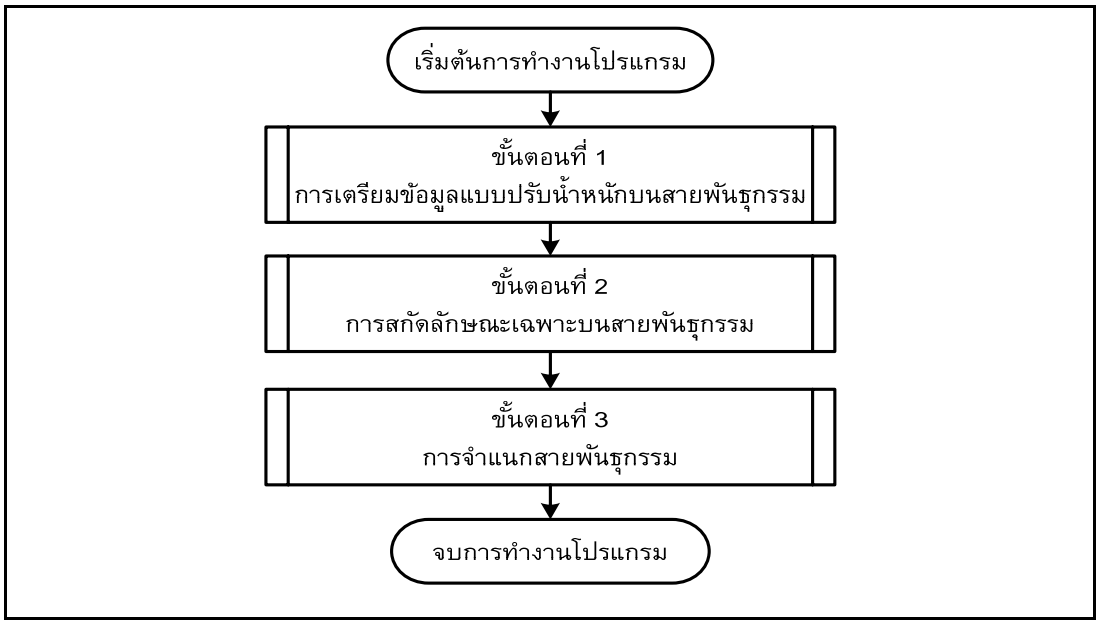
ภาพประกอบ 4.5 ภาพกระแสข้อมูลระดับที่ 2 การจำแนกสายพันธุกรรมด้วยลักษณะเฉพาะ

ภาพประกอบ 4.5 แสดงกระแสข้อมูลระดับที่ 2 ของขั้นตอนการจำแนกสายพันธุกรรมด้วยลักษณะเฉพาะที่สกัดได้ สำหรับเป็นคุณสมบัติของข้อมูล (Attributes) จากขั้นตอนหลักก่อนหน้า และใช้ค่าน้ำหนักของคู่ลำดับกรดอะมิโนบนสายพันธุกรรมแต่ละสายในการจำแนก (Classify) ด้วยขั้นตอนวิธีแบบต่าง ๆ ซึ่งจะอธิบายการทำงานในส่วนถัดไป ผลลัพธ์สุดท้ายที่ได้จากขั้นตอนนี้ คือ ค่าความถูกต้อง (Accuracy) ของการจำแนกสายพันธุกรรม

4.2 ผังงานโปรแกรมวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

ในการพัฒนาระบบวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม สามารถเขียนแผนภาพผังงานโปรแกรม (System Program) เพื่อแสดงรายละเอียดขั้นตอนการทำงานแต่ละส่วนได้ดังนี้

4.2.1 ผังงานโปรแกรมหลักของระบบวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม แสดงได้ดังภาพประกอบ 4.6



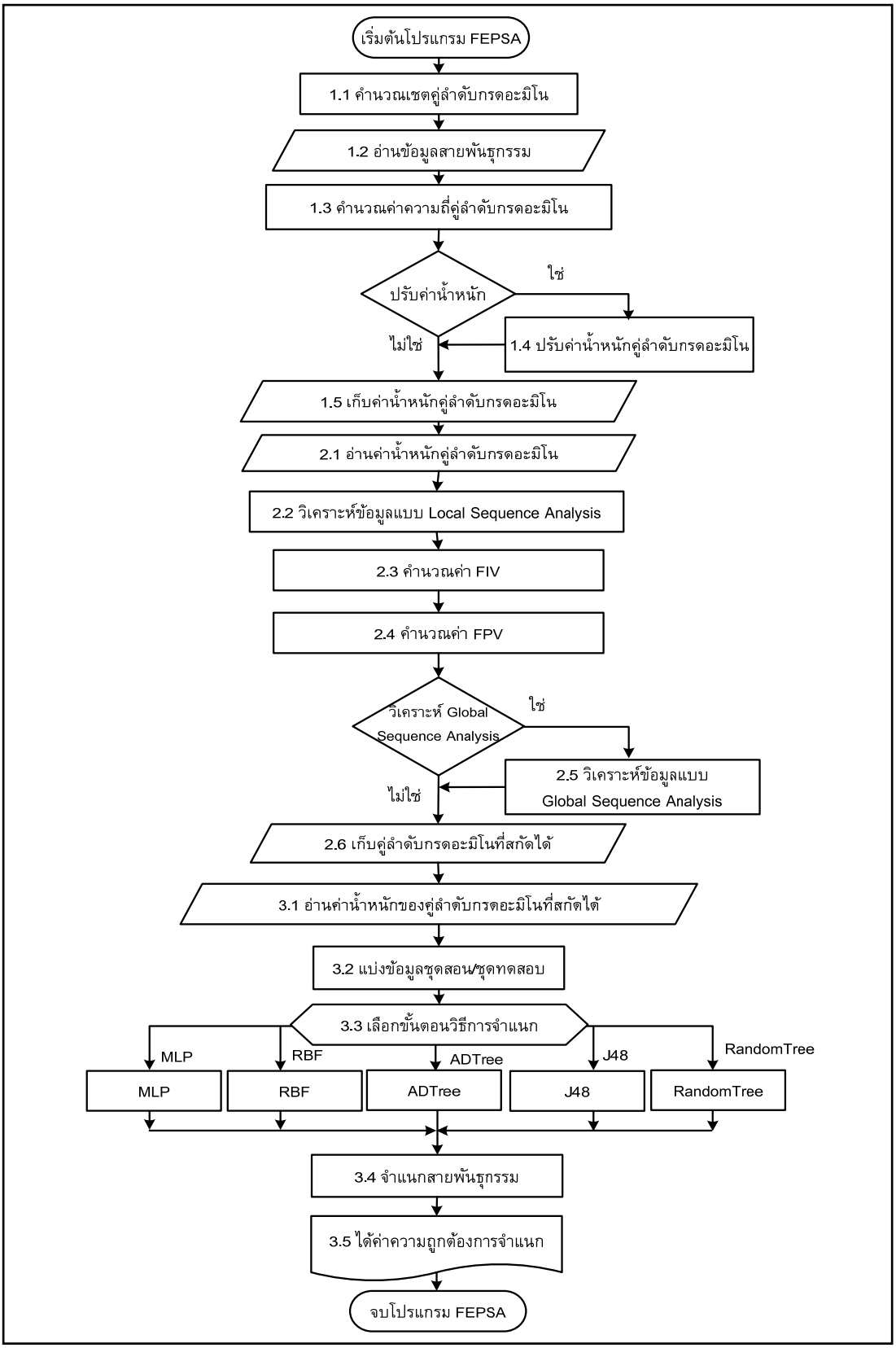
ภาพประกอบ 4.6 ผังงานโปรแกรมหลักระบบวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

จากผังงานโปรแกรมหลักระบบวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม (อ้างอิงจากภาพประกอบ 4.6) จะประกอบด้วย 3 โปรแกรมย่อย คือ

1) โปรแกรมย่อยที่ 1 การเตรียมข้อมูลค่าความถี่แบบปรับค่าน้ำหนัก (Weight Balance Frequency Data Program) ใช้สำหรับการแปลงข้อมูลสายพันธุกรรมให้อยู่ในรูปของค่าความถี่คู่ลำดับกรดอะมิโนแบบมีการปรับค่าน้ำหนัก (WBL)

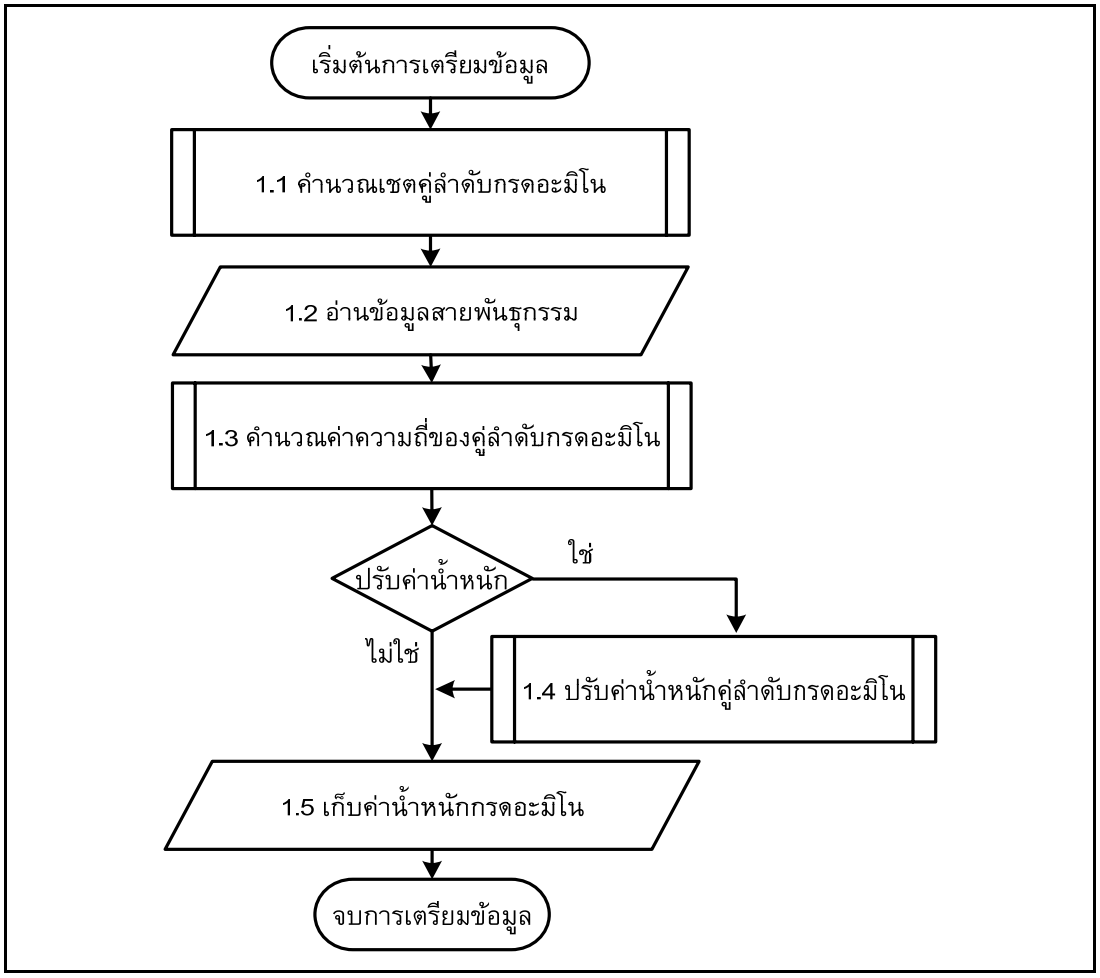
2) โปรแกรมย่อยที่ 2 การสกัดลักษณะเฉพาะบนสายพันธุกรรม (Feature Extraction Program) สำหรับสกัดหาลักษณะเฉพาะ (Feature) ซึ่งปรากฏอยู่บนสายพันธุกรรมแต่ละสาย ด้วยการวิเคราะห์ระดับชุดข้อมูล (Global Sequence Analysis)

3) โปรแกรมย่อยที่ 3 การจำแนกสายพันธุกรรมด้วยลักษณะเฉพาะ (Feature Classification Program) สำหรับการจำแนก (Classify) สายพันธุกรรมด้วยขั้นตอนวิธีโครงข่ายประสาทเทียมแบบ MLP และ RBF และต้นไม้การตัดสินใจแบบ ADTree J48 และ RandomTree ตามลำดับ สามารถอธิบายผังงานโปรแกรมโดยละเอียดได้จากภาพประกอบ 4.7



ภาพประกอบ 4.7 ผังงานโปรแกรมระบบวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

4.2.2 ผังงานของโปรแกรมย่อยที่ 1 การเตรียมข้อมูลค่าความถี่แบบมีการปรับน้ำหนัก (Weight Balance Frequency Data Program) แสดงดังภาพประกอบ 4.8

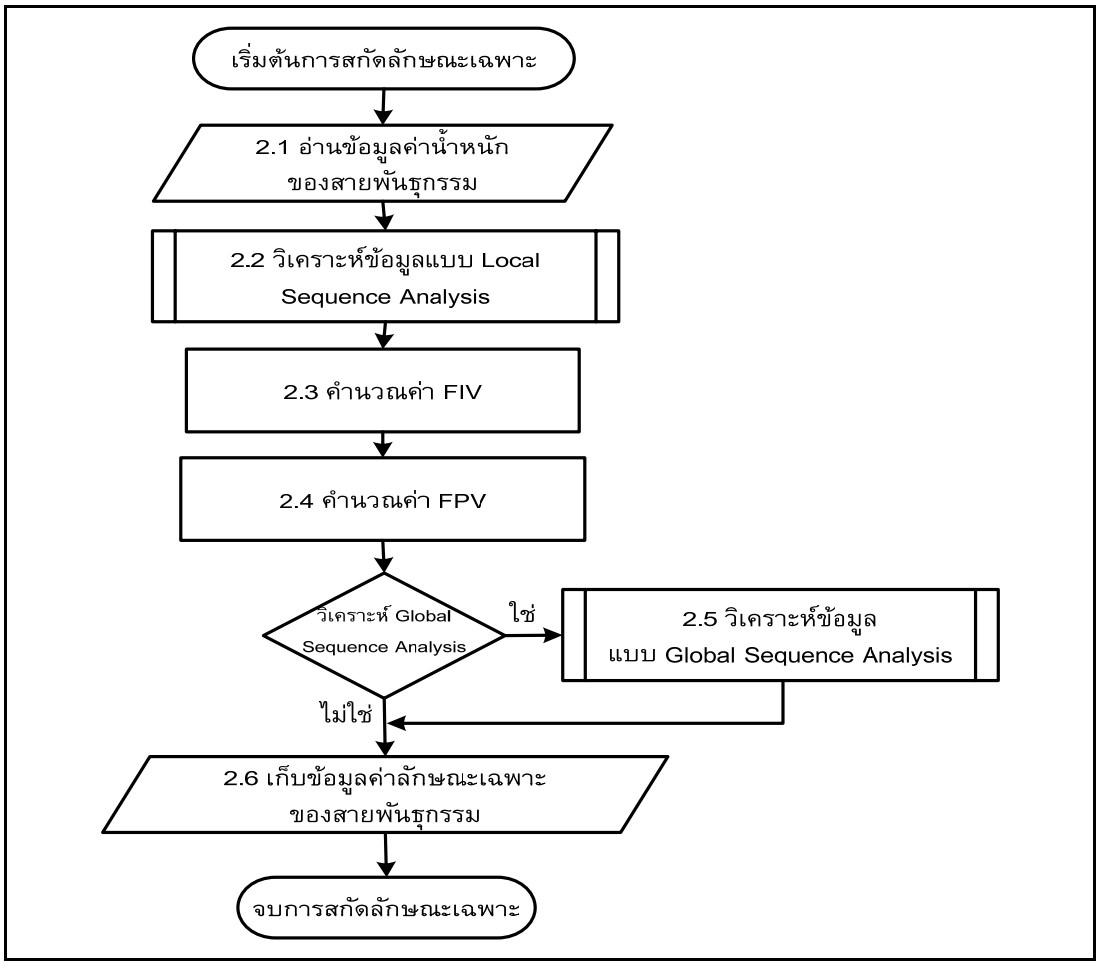


ภาพประกอบ 4.8 ผังงานโปรแกรมย่อยการเตรียมข้อมูลค่าความถี่แบบมีการปรับน้ำหนัก

จากผังงานโปรแกรมการเตรียมข้อมูลค่าความถี่แบบมีการปรับน้ำหนักประกอบด้วย 5 ขั้นตอน คือ 1) คำนวณเซตคู่ลำดับกรดอะมิโนจากรายชื่อย่อกรดอะมิโน 2) อ่านสายพันธุกรรมเข้าสู่ระบบ 3) คำนวนค่าความถี่ของคู่ลำดับกรดอะมิโนบนสายพันธุกรรมแต่ละสาย 4) ปรับคุณภาพของข้อมูลให้มีลักษณะเดียวกันด้วยวิธีปรับค่าน้ำหนักของค่าความถี่ (WBL) และ 5) บันทึกข้อมูลค่าความถี่ซึ่งมีการปรับค่าน้ำหนัก

โดยในขั้นตอนที่ 4 ผู้ใช้สามารถเลือก (Optional) ว่าจะมีการทำขั้นตอนการปรับค่าน้ำหนักหรือไม่ หากไม่ต้องการปรับค่าน้ำหนักในการจัดเก็บข้อมูลค่าน้ำหนักจะแทนด้วยค่าความถี่แบบจริงที่นับได้บนสายพันธุกรรม

4.2.3 ผังงานของโปรแกรมย่อยที่ 2 การสกัดลักษณะเฉพาะบนสายพันธุกรรม (Feature Extraction Program) แสดงดังภาพประกอบ 4.9

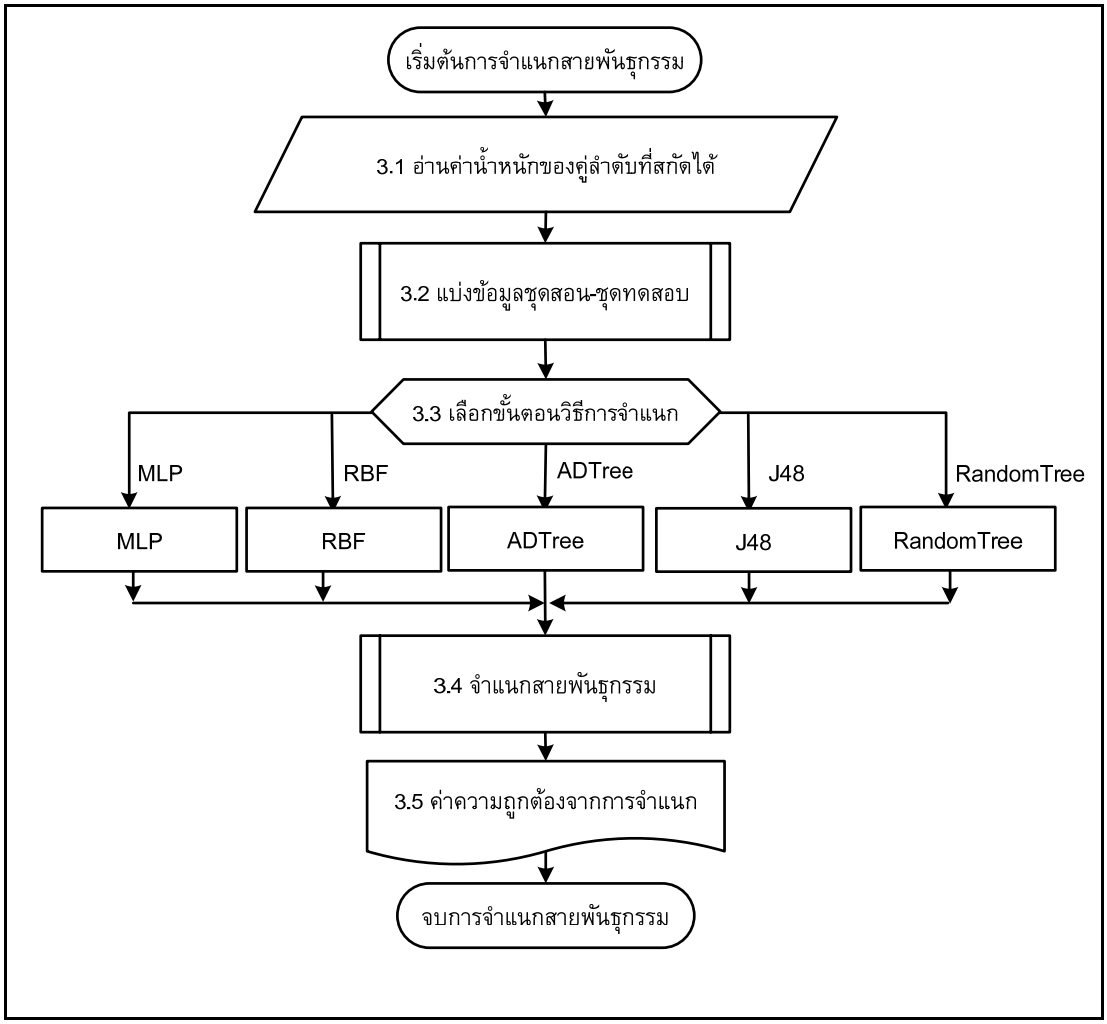


ภาพประกอบ 4.9 ผังงานโปรแกรมย่อยการสกัดลักษณะเฉพาะบนสายพันธุกรรม

จากภาพประกอบ 4.9 ขั้นตอนการทำงานของโปรแกรมการสกัดลักษณะเฉพาะบนสายพันธุกรรมจะประกอบด้วย 6 ขั้นตอน คือ 1) อ่านค่าน้ำหนักของคู่ลำดับ 2) วิเคราะห์ข้อมูลระดับสายพันธุกรรม (Local Sequences Analysis) 3) คำนวณจำนวนครั้งของคู่ลำดับ (FIV) 4) ขั้นตอนการคำนวณค่าร้อยละของคู่ลำดับดับ (FPV) 5) วิเคราะห์ข้อมูลระดับชุดข้อมูล (Global Sequence Analysis) และ 6) เก็บค่าลักษณะเฉพาะที่สกัดได้

โดยในขั้นตอนการวิเคราะห์ข้อมูลระดับชุดข้อมูล (Global Sequences Analysis) ผู้ใช้สามารถเลือกทำงาน (Optional) ได้ หากผู้ให้ประสงค์ที่จะไม่ทำงานในขั้นตอนนี้ ผลลัพธ์ของลักษณะเฉพาะที่ได้คือ จะมีทั้งสิ้น 400 คู่ลำดับ (ไม่มีการสกัดคู่ลำดับใดออก)

4.2.4 ผังงานของโปรแกรมย่อยที่ 3 การจำแนกสายพันธุกรรมด้วยลักษณะเฉพาะ (Feature Classification Program) แสดงดังภาพประกอบ 4.10



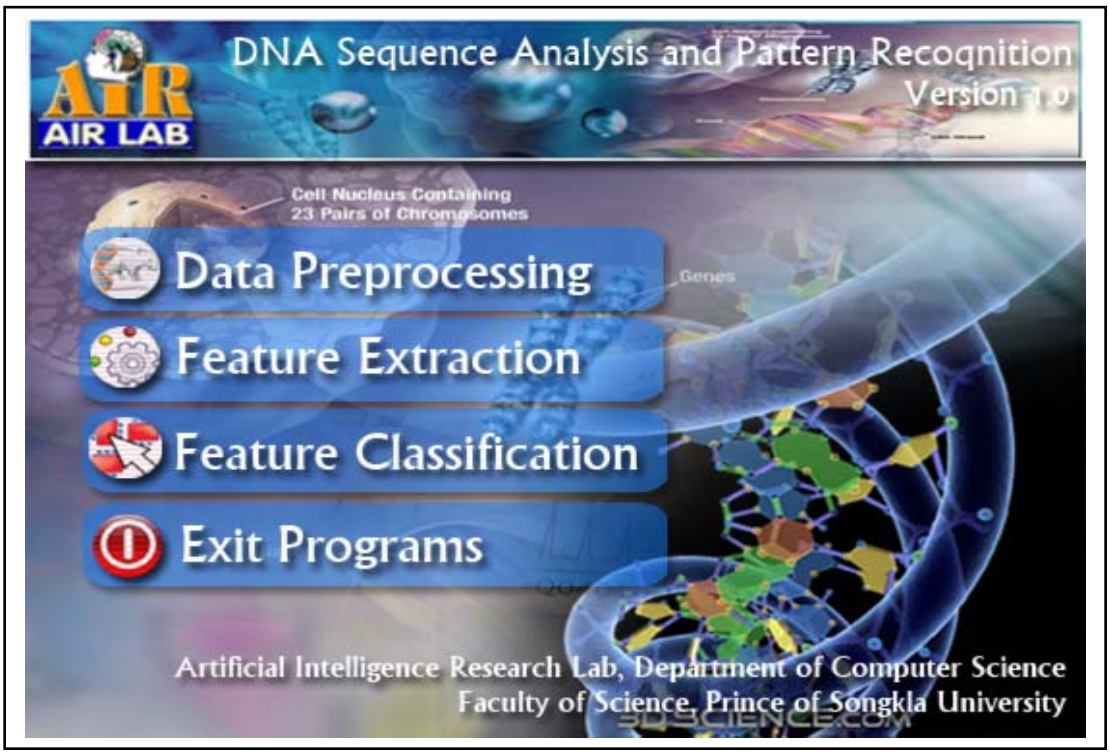
ภาพประกอบ 4.10 ผังงานโปรแกรมย่อยการจำแนกสายพันธุกรรมด้วยลักษณะเฉพาะ

จากภาพประกอบ 4.10 ขั้นตอนการจำแนกข้อมูลนั้นประกอบด้วย 5 ขั้นตอนหลัก คือ 1) อ่านค่าน้ำหนักของคู่ลำดับที่สกัดได้ 2) แบ่งข้อมูลสำหรับชุดสอนและชุดทดสอบด้วยวิธี k-Fold Cross Validation ซึ่งในส่วนนี้ผู้ใช้สามารถระบุจำนวน Fold ที่ต้องการได้ 3) เลือกขั้นตอนการจำแนกสายพันธุกรรม มีขั้นตอนวิธีให้ผู้ใช้เลือก 5 ขั้นตอนวิธี คือ MLP RBF ADTree J48 และ RandomTree ตามลำดับ 4) จำแนกสายพันธุกรรมจะเรียกใช้งานโปรแกรม WEKA ด้วยขั้นตอนวิธีที่เลือก และ 5) แสดงค่าความถูกต้องจากการจำแนกสายพันธุกรรม

4.3 การพัฒนาโปรแกรมวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

ในการพัฒนาระบบวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม มีการออกแบบหน้าต่างโปรแกรม (Program Screens) สำหรับติดต่อกับผู้ใช้ (User Interface) ด้วยโปรแกรม Matlab 2007a ซึ่งมีรายละเอียดหน้าที่และขั้นตอนการทำงานแต่ละส่วน ดังนี้

4.3.1 เมนูหลักของโปรแกรมวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม แสดงได้ดังภาพประกอบ 4.11



ภาพประกอบ 4.11 เมนูหลักของโปรแกรมวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

จากภาพประกอบ 4.11 เมนูหลักของโปรแกรมจะถูกสร้างในลักษณะของแท็บเมนู (Tab Panel) ประกอบด้วย 4 เมนูหลัก คือ

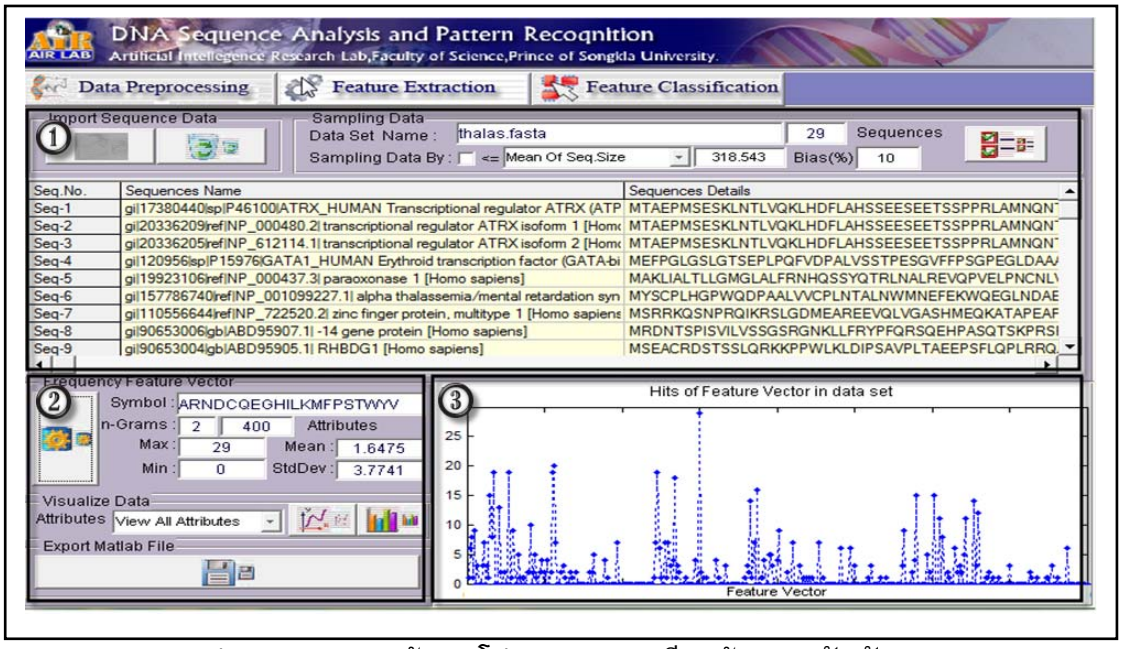
1) เมนูหลักที่ 1 ทำหน้าที่เรียกใช้โปรแกรมเตรียมข้อมูลค่าความถี่แบบมีการปรับน้ำหนัก (Weight Balance Frequency Data Program)

2) เมนูหลักที่ 2 ทำหน้าที่เรียกใช้โปรแกรมสกัดลักษณะเฉพาะบนสายพันธุกรรม (Feature Extraction Program)

3) เมนูหลักที่ 3 ทำหน้าที่เรียกใช้โปรแกรมจำแนกสายพันธุกรรมด้วยลักษณะเฉพาะ (Feature Classification Program)

4) เมนูหลักที่ 4 ทำหน้าที่สำหรับจบการทำงานโปรแกรม (Exit Program)

4.3.2 หน้าต่างโปรแกรมการเตรียมข้อมูลค่าความถี่แบบมีการปรับน้ำหนัก (Weight Balance Frequency Data Program) แสดงได้ดังภาพประกอบ 4.12



ภาพประกอบ 4.12 หน้าต่างโปรแกรมการเตรียมข้อมูลลำดับพันธุกรรม

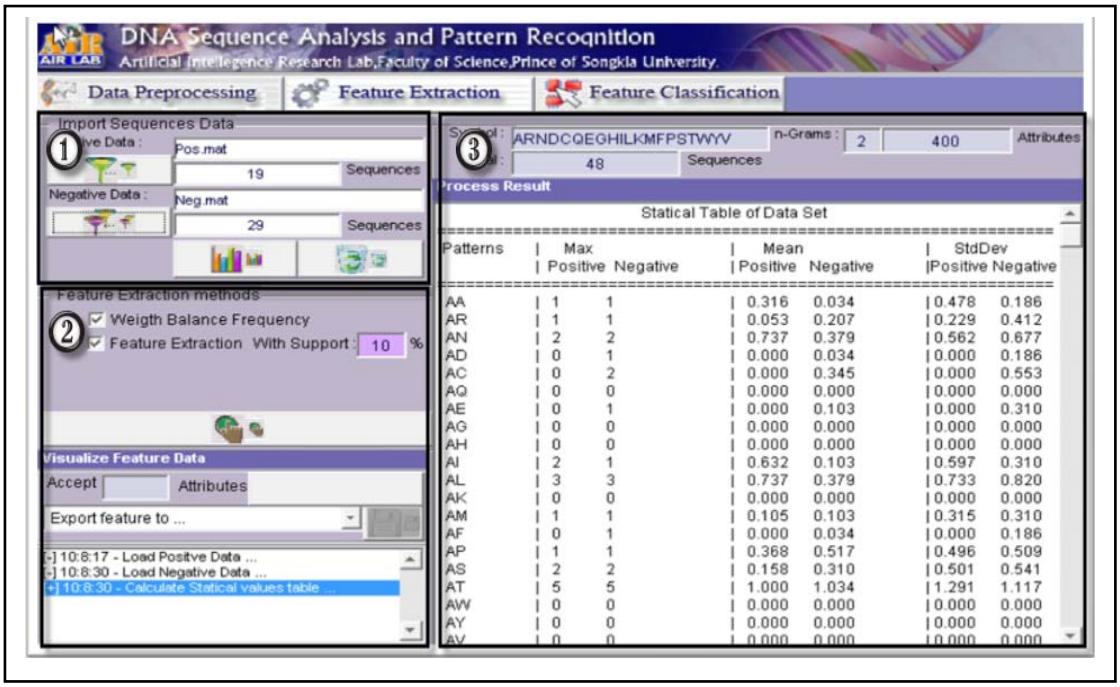
จากภาพประกอบ 4.12 ประกอบด้วย 3 ส่วนหลัก คือ

1) ส่วนการนำเข้าข้อมูลสายพันธุกรรม และการเลือกตัวอย่างข้อมูล (Sampling Data) ผู้ใช้สามารถเลือกค่าสถิติในการเลือกกลุ่มตัวอย่าง เช่น ค่าเฉลี่ยเลขคณิต (Mean) ค่าต่ำสุด (Minimum) ค่าสูงสุด (Maximum) และนอกจากนี้ผู้ใช้สามารถเพิ่มค่าความเอนเอียง (Bias) ได้หากต้องการ

2) ส่วนแสดงค่าความถี่ของสายพันธุกรรมที่นำเข้าสู่โปรแกรมวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม โดยจะแสดงค่าสถิติ เช่น ค่าสูงสุด (Maximum) ค่าต่ำสุด (Minimum) ค่าเฉลี่ยเลขคณิต (Mean) ค่าส่วนเบี่ยงเบนมาตรฐาน (Standard Derivation) และแผนภูมิแสดงค่าความถี่ของคู่ลำดับกรดอะมิโนที่คำนวณได้

3) ส่วนแสดงกราฟลักษณะของข้อมูลซึ่งข้อมูลที่ใช้เป็นค่าความถี่ของคู่ลำดับที่คำนวณได้ และนอกจากนี้ผู้ใช้ยังสามารถเรียกดูกราฟค่าความถี่ของคู่ลำดับกรดอะมิโนแต่ละคู่ลำดับกรดอะมิโนในสายพันธุกรรมแต่ละสายได้

4.3.3 หน้าต่างโปรแกรมการสกัดลักษณะเฉพาะบนสายพันธุกรรม (Feature Extraction Program) แสดงได้ดังภาพประกอบ 4.13



ภาพประกอบ 4.13 หน้าต่างโปรแกรมการสกัดลักษณะเฉพาะบนสายพันธุกรรม

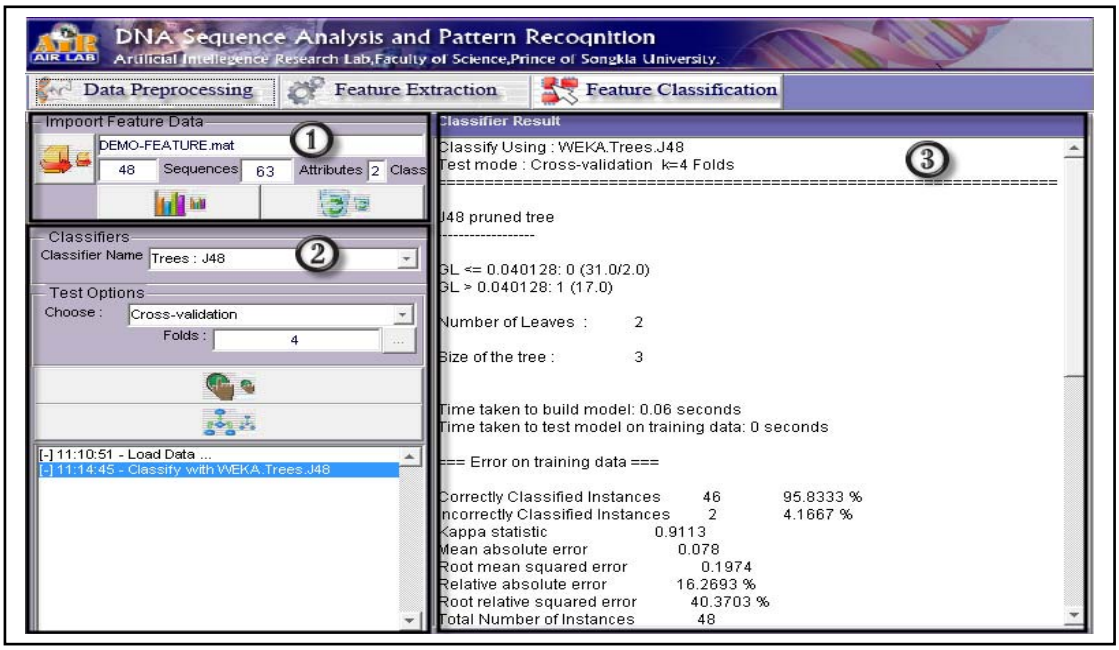
จากภาพประกอบ 4.13 ประกอบด้วย 3 ส่วนหลัก คือ

1) ส่วนการนำเข้าข้อมูลสายพันธุกรรม โดยผู้ใช้จำเป็นต้องมีการนำเข้าข้อมูล 2 ลักษณะคือ ข้อมูลซึ่งแสดงอาการของโรค (Positive Sequence) และข้อมูลซึ่งไม่แสดงอาการของโรค (Negative Sequence)

2) ส่วนการสกัดลักษณะเฉพาะบนสายพันธุกรรม ประกอบด้วยการวิเคราะห์ระดับสายพันธุกรรม (Local Sequences Analysis) และส่วนของการวิเคราะห์ระดับชุดข้อมูล (Global Sequences Analysis) และนอกจากนี้ด้านล่างจะแสดงส่วนของประวัติการทำงาน (History) ซึ่งสามารถเรียกดูประวัติย้อนหลังได้

3) ส่วนแสดงค่าสถิติค่าความถี่ของสายพันธุกรรมที่นำเข้าสู่โปรแกรมวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม ซึ่งผ่านขั้นตอนการลักษณะเฉพาะ (Feature Extraction) โดยจะแสดงค่าสถิติ เช่น ค่าความนิยม (Hits Ratio) ของข้อมูลที่แสดงอาการของโรค (Positive Data) กลุ่มข้อมูลที่ไม่แสดงอาการของโรค (Negative Data) และแผนภูมิแสดงค่าสถิติของคู่ลำดับกรดอะมิโนที่สกัดได้

4.3.4 หน้าต่างโปรแกรมการจำแนกสายพันธุกรรมด้วยลักษณะเฉพาะ (Feature Classification Program) แสดงได้ดังภาพประกอบ 4.14



ภาพประกอบ 4.14 หน้าต่างโปรแกรมการจำแนกสายพันธุกรรมด้วยลักษณะเฉพาะ

จากภาพประกอบ 4.14 ประกอบด้วย 3 ส่วนหลัก คือ

1) ส่วนของการนำเข้าค่าน้ำหนักของคู่ลำดับที่สกัดลักษณะเฉพาะได้บนสายพันธุกรรมทุกสายในชุดข้อมูล

2) ส่วนของการเลือกขั้นตอนวิธี (Algorithms) ในการจำแนกสายพันธุกรรมประกอบด้วย 5 ขั้นตอนวิธี คือ MLP RBF ADTree J48 และ RandomTree ตามลำดับ และกำหนดจำนวน Fold ที่ต้องการทดสอบแบบ k-Fold Cross Validation

3) ส่วนของการแสดงค่าสถิติของข้อมูลและผลลัพธ์ที่ได้จากการจำแนกข้อมูลสายพันธุกรรมด้วยลักษณะเฉพาะ ซึ่งประกอบด้วย ค่าความถูกต้อง (Accuracy) ตารางการจำแนก (Confusion Matrix) และกราฟค่าสถิติของข้อมูล เป็นต้น

โดยรายละเอียดการใช้งานโปรแกรมแสดงในคู่มือผู้ใช้ระบบการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม ดังภาคผนวก ก และการติดตั้งระบบแสดงในคู่มือการติดตั้งระบบการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม ดังภาคผนวก ข ตามลำดับ

บทที่ 5

ผลการทดลองและบทวิจารณ์

บทนี้จะนำเสนอผลลัพธ์จากการทดลองตามแบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรมด้วยเทคนิค FEPSA ในการจำแนกสายพันธุกรรมของผู้ป่วยซึ่งรวบรวมจาก The National Center for Biotechnology Information (NCBI) ประกอบด้วย ข้อมูลผู้ป่วยธาลัสซีเมีย (Thalassemia Disease) เบาหวาน (Diabetes Disease) และดาวน์ซินโดรม (Downsyndrome Disease) ตามลำดับ ซึ่งแสดงผลการทดลองที่ได้ตามลำดับดังนี้

5.1 การออกแบบวิธีการทดลอง

จากขั้นตอนวิธีของแบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรมโดยใช้เทคนิคการสกัดลักษณะเฉพาะบนสายโปรตีน (Feature Extraction of Protein Sequence Analysis: FEPSA) สามารถออกแบบการทดลองเป็น 4 ลักษณะ คือ

5.1.1 การทดลองแบบ A (Model A)

มีรายละเอียด คือ ใช้ข้อมูลค่าความถี่ของคู่ลำดับแบบไม่ปรับค่าน้ำหนัก (Non-weight Balance Frequency: NBL) และไม่มีการสกัดลักษณะเฉพาะ

5.1.2 การทดลองแบบ B (Model B)

มีรายละเอียด คือ ใช้ข้อมูลค่าความถี่ของคู่ลำดับแบบปรับค่าน้ำหนัก (Weight Balance Frequency: WBL) และไม่มีการสกัดลักษณะเฉพาะ

5.1.3 การทดลองแบบ C (Model C)

มีรายละเอียด คือ ใช้ข้อมูลค่าความถี่ของคู่ลำดับแบบไม่ปรับค่าน้ำหนัก (NBL) และมีการสกัดลักษณะเฉพาะด้วยขั้นตอนการวิเคราะห์ระดับชุดข้อมูล

5.1.4 การทดลองแบบ D (Model D)

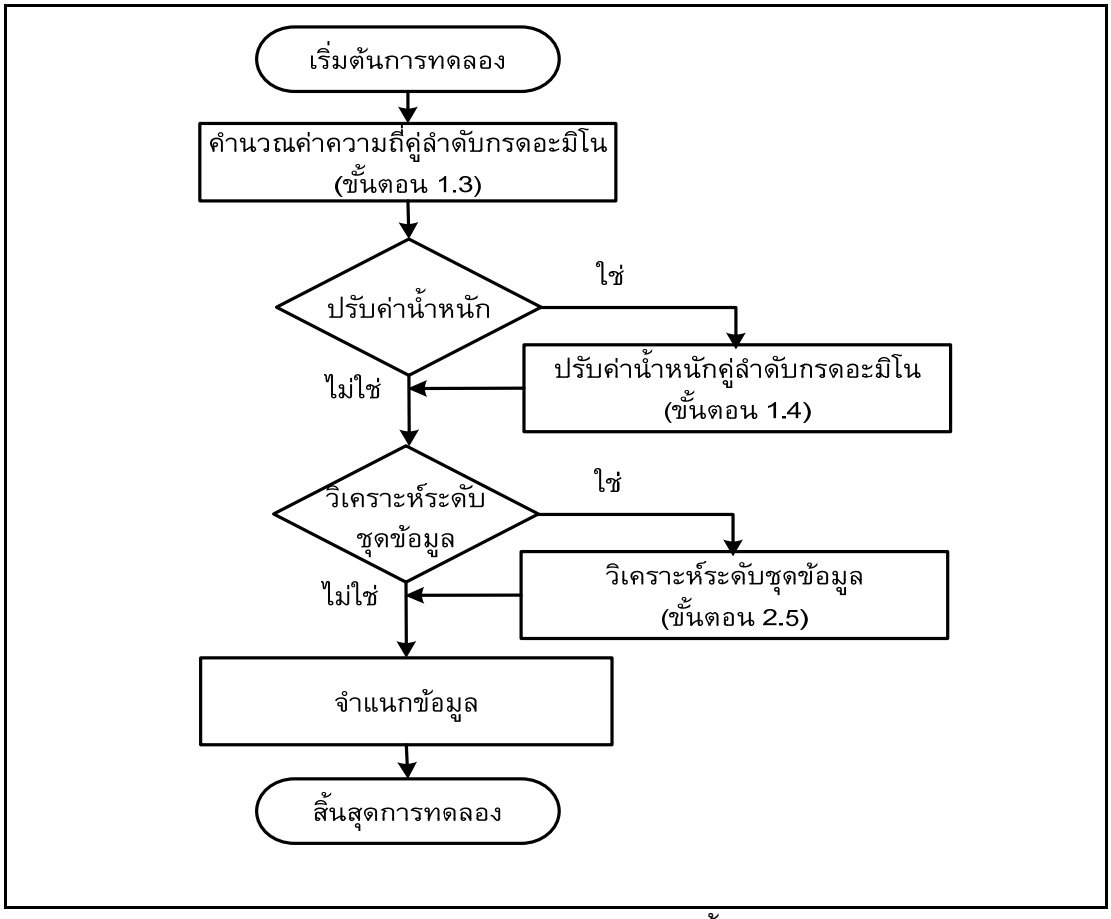
มีรายละเอียด คือ ใช้ข้อมูลค่าความถี่ของคู่ลำดับแบบปรับค่าน้ำหนัก (WBL) และมีการสกัดลักษณะเฉพาะด้วยขั้นตอนการวิเคราะห์ระดับชุดข้อมูล

สามารถแสดงรายละเอียดการออกแบบการทดลองด้วยขั้นตอนวิธี FEPSA ดังแสดงในตารางที่ 5.1

ตารางที่ 5.1 รายละเอียดการออกแบบการทดลองตามแบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรมด้วยขั้นตอนวิธี FEPSA

การทดลอง	รายละเอียดการทดลอง	
	ปรับน้ำหนักข้อมูลค่าความถี่	สกัดลักษณะเฉพาะด้วยการวิเคราะห์ระดับชุดข้อมูล
แบบ A	✘	✘
แบบ B	✓	✘
แบบ C	✘	✓
แบบ D	✓	✓

จากตารางที่ 5.1 สัญลักษณ์ ✘ แทน ไม่มีการทำงานในขั้นตอนนั้น และสัญลักษณ์ ✓ แทน มีการทำงานในขั้นตอนนั้น (อ้างอิงจากภาพประกอบ 4.7 ในบทที่ 4) และแสดงภาพรวมการทดลอง ได้ดังภาพประกอบ 5.1



ภาพประกอบ 5.1 ภาพรวมการทดลองตามขั้นตอนวิธี FEPSA

5.2 การเตรียมเซตของคู่ลำดับกรดอะมิโนที่เป็นได้ด้วยเทคนิคเอ็นแกรม

เซตของคู่ลำดับกรดอะมิโนที่เป็นได้ (Feature Vector) ด้วยเทคนิค 2-Grams จากสัญลักษณ์ย่อกรดอะมิโนมาตรฐาน 20 ชนิด คือ A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y และ V สามารถคำนวณคู่ลำดับที่เป็นไปได้ทั้งสิ้น 400 คู่ลำดับ ดังนี้ AA, AR, AN, AD, AC, AQ, AE, AG, AH, AI, AL, AK, AM, AF, AP, AS, AT, AW, AY, AV, RA, RR, RN, RD, RC, RQ, RE, RG, RH, RI, RL, RK, RM, RF, RP, RS, RT, RW, RY, RV, NA, NR, NN, ND, NC, NQ, NE, NG, NH, NI, NL, NK, NM, NF, NP, NS, NT, NW, NY, NV, DA, DR, DN, DD, DC, DQ, DE, DG, DH, DI, DL, DK, DM, DF, DP, DS, DT, DW, DY, DV, CA, CR, CN, CD, CC, CQ, CE, CG, CH, CI, CL, CK, CM, CF, CP, CS, CT, CW, CY, CV, QA, QR, QN, QD, QC, QQ, QE, QG, QH, QI, QL, QK, QM, QF, QP, QS, QT, QW, QY, QV, EA, ER, EN, ED, EC, EQ, EE, EG, EH, EI, EL, EK, EM, EF, EP, ES, ET, EW, EY, EV, GA, GR, GN, GD, GC, GQ, GE, GG, GH, GI, GL, GK, GM, GF, GP, GS, GT, GW, GY, GV, HA, HR , HN, HD, HC, HQ, HE, HG, HH, HI, HL, HK, HM, HF, HP, HS, HT, HW, HY, HV, IA, IR, IN, ID, IC, IQ, IE, IG, IH, II, IL, IK, IM, IF, IP, IS, IT, IW, IY, IV, LA, LR, LN, LD, LC, LQ, LE, LG, LH, LI, LL, LK, LM, LF, LP, LS, LT, LW, LY, LV, KA, KR, KN, KD, KC, KQ, KE, KG, KH, KI, KL, KK, KM, KF, KP, KS, KT, KW, KY, KV, MA, MR, MN, MD, MC, MQ, ME, MG, MH, MI, ML, MK, MM, MF, MP, MS, MT, MW, MY, MV, FA, FR, FN, FD, FC, FQ, FE, FG, FH, FI, FL, FK, FM, FF, FP, FS, FT, FW, FY, FV, PA, PR, PN, PD, PC, PQ, PE, PG, PH, PI, PL, PK, PM, PF, PP, PS, PT, PW, PY, PV, SA, SR, SN, SD, SC, SQ, SE, SG, SH, SI, SL, SK, SM, SF, SP, SS, ST, SW, SY, SV, TA, TR, TN, TD, TC, TQ, TE, TG, TH, TI, TL, TK, TM, TF, TP, TS, TT, TW, TY, TV, WA, WR, WN, WD, WC, WQ, WE, WG, WH, WI, WL, WK, WM, WF, WP, WS, WT, WW, WY, WV, YA, YR, YN, YD, YC, YQ, YE, YG, YH, YI, YL, YK, YM, YF, YP, YS, YT, YW, YY, YV, VA, VR, VN, VD, VC, VQ, VE, VG, VH, VI, VL, VK, VM, VF, VP, VS, VT, VW, VY และ VV ตามลำดับ

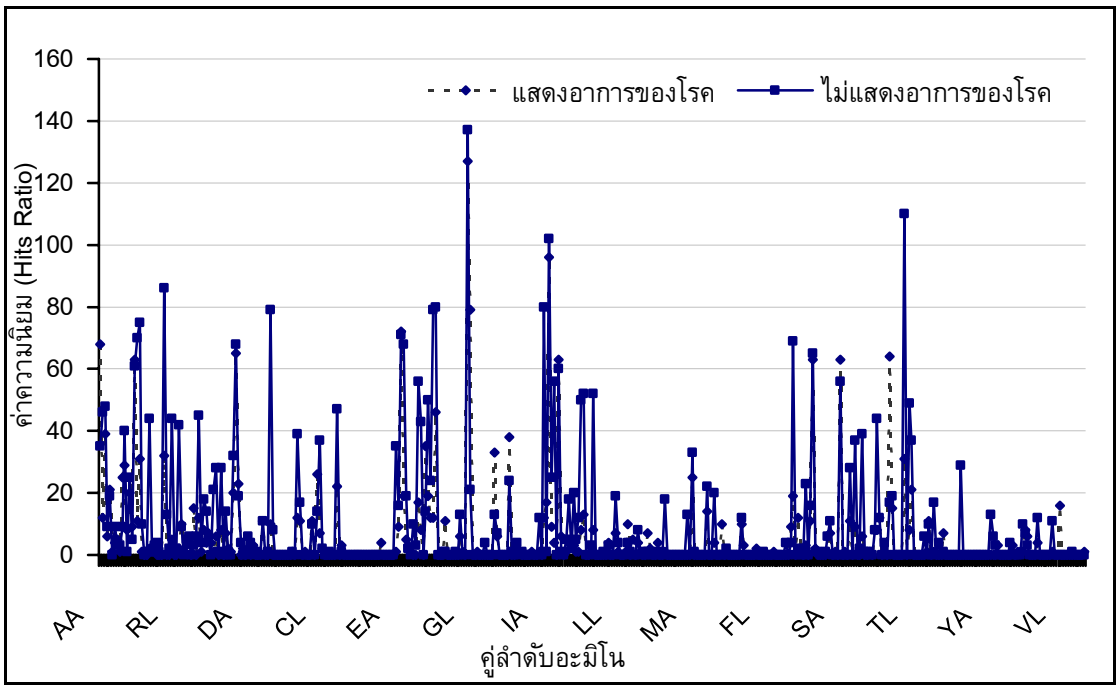
ในการทดสอบประสิทธิภาพของการทดลองทั้ง 4 ลักษณะ จะใช้ข้อมูลสายพันธุกรรมของผู้ป่วย ประกอบด้วย 3 ชุดข้อมูล คือ ข้อมูลผู้ป่วยธาลัสซีเมีย (Thalassemia Disease) เบาหวาน (Diabetes Disease) และดาวน์ซินโดรม (Downsyndrome Disease) ซึ่งรวบรวมจาก The National Center for Biotechnology Information (NCBI) สามารถแสดงรายละเอียดของการทดลองได้ตามลำดับ ดังนี้

5.3 ข้อมูลผู้ป่วยธาลัสซีเมีย (Thalassemia Disease)

ข้อมูลผู้ป่วยธาลัสซีเมียประกอบด้วยสายโปรตีน (Protein Sequence) ซึ่งสุ่มเลือก (Sampling) ด้วยใช้ค่าเฉลี่ยเลขคณิต (Mean Value) โดยสายโปรตีนมีความยาวอยู่ระหว่าง 30 (±5%) ตัวอักษร ถึง 463.496 (±5%) ตัวอักษร ทั้งสิ้น 264 สาย แบ่งเป็นข้อมูลซึ่งแสดงอาการของโรค (Positive Sequence) จำนวน 127 สาย และไม่แสดงอาการของโรค (Negative Sequence) จำนวน 137 สาย ประกอบด้วยขั้นตอนในการทดลอง 3 ขั้นตอน คือ การเตรียมข้อมูลค่าความถี่แบบปรับน้ำหนัก การสกัดลักษณะเฉพาะ และการจำแนกสายพันธุกรรม โดยสามารถอธิบายรายละเอียดการทดลองตามลำดับดังนี้

5.3.1 ขั้นตอนหลักที่ 1 การเตรียมข้อมูลผู้ป่วยธาลัสซีเมียแบบปรับน้ำหนัก

ค่าความถี่ของคู่ลำดับกรดอะมิโนแบบ 2-Grams (Frequency Feature Vector) ซึ่งนับได้จากข้อมูลผู้ป่วยธาลัสซีเมีย สามารถแสดงลักษณะของข้อมูลด้วยค่าความนิยม (Hits Ratio) ของคู่ลำดับกรดอะมิโน ได้ดังภาพประกอบ 5.2



ภาพประกอบ 5.2 ลักษณะของข้อมูลผู้ป่วยธาลัสซีเมียด้วยค่าความนิยม

จากภาพประกอบ 5.2 พบว่า ในกลุ่มข้อมูลซึ่งแสดงอาการของโรค มีคู่ลำดับกรดอะมิโน GI มีค่าความนิยมสูงสุดเท่ากับ 127 และในกลุ่มข้อมูลซึ่งไม่แสดงอาการของโรคมีค่าความนิยมสูงสุด 137 และสามารถแสดงค่าความถี่ของคู่ลำดับกรดอะมิโนแบบ 2-Grams ทั้ง 400 คู่ลำดับของข้อมูลผู้ป่วยธาลัสซีเมีย ได้ดังตารางที่ 5.2

ตารางที่ 5.2 ค่าความถี่ของคู่ลำดับอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยธาลัสซีเมีย

คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV		คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV	
	Positive	Negative	Positive	Negative		Positive	Negative	Positive	Negative
AA	1	2	68	35	RM	1	2	2	42
AR	2	3	12	46	RF	1	1	10	9
AN	3	5	39	48	RP	1	0	1	0
AD	1	3	6	9	RS	1	1	3	5
AC	2	2	21	19	RT	1	2	4	6
AQ	1	0	1	0	RW	0	0	0	0
AE	1	1	4	9	RY	2	1	15	3
AG	1	1	5	3	RV	0	1	0	6
AH	1	1	1	3	NA	2	4	12	45
AI	2	1	25	9	NR	1	1	1	4
AL	3	3	29	40	NN	1	3	8	18
AK	1	1	8	1	ND	2	2	4	14
AM	5	1	20	25	NC	2	1	7	5
AF	2	2	9	5	NQ	1	0	4	0
AP	2	2	63	61	NE	1	4	1	21
AS	2	4	11	70	NG	2	2	6	28
AT	3	3	31	75	NH	0	1	0	1
AW	1	1	1	10	NI	1	3	8	28
AY	0	0	0	0	NL	0	0	0	0
AV	0	1	0	1	NK	1	1	7	14
RA	1	5	2	44	NM	1	0	2	0
RR	1	1	1	1	NF	1	0	1	0
RN	1	1	3	3	NP	1	1	20	32
RD	0	3	0	4	NS	2	3	65	68
RC	0	1	0	1	NT	1	3	23	19
RQ	0	0	0	0	NW	0	3	0	4
RE	4	4	32	86	NY	1	0	2	0
RG	0	1	0	13	NV	0	0	0	0
RH	0	1	0	2	DA	1	1	2	6
RI	1	2	5	44	DR	0	2	0	4
RL	0	1	0	1	DN	1	1	3	2

ตารางที่ 5.2 ค่าความถี่ของคู่ลำดับอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยธาลัสซีเมีย (ต่อ)

คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV		คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV	
	Positive	Negative	Positive	Negative		Positive	Negative	Positive	Negative
RK	0	1	0	2	DD	0	0	0	0
DC	0	1	0	1	CT	2	3	22	47
DQ	0	0	0	0	CW	0	0	0	0
DE	3	3	11	11	CY	1	1	3	1
DG	0	0	0	0	CV	0	0	0	0
DH	0	1	0	1	QA	0	0	0	0
DI	2	3	10	79	QR	0	0	0	0
DL	0	1	0	8	QN	0	0	0	0
DK	0	0	0	0	QD	0	0	0	0
DM	0	0	0	0	QC	0	0	0	0
DF	0	0	0	0	QQ	0	0	0	0
DP	0	0	0	0	QE	0	0	0	0
DS	0	0	0	0	QG	0	0	0	0
DT	0	0	0	0	QH	0	0	0	0
DW	0	0	0	0	QI	0	0	0	0
DY	2	1	1	1	QL	0	0	0	0
DV	0	0	0	0	QK	0	0	0	0
CA	2	2	12	39	QM	0	0	0	0
CR	1	1	11	17	QF	0	0	0	0
CN	0	0	0	0	QP	1	0	4	0
CD	1	0	1	0	QS	0	0	0	0
CC	0	0	0	0	QT	0	0	0	0
CQ	0	0	0	0	QW	0	0	0	0
CE	2	1	11	10	QY	0	0	0	0
CG	0	0	0	0	QV	0	0	0	0
CH	2	3	26	14	EA	1	2	1	35
CI	1	1	7	37	ER	3	3	9	16
CL	1	1	2	2	EN	2	4	72	71
CK	0	0	0	0	ED	2	4	19	68
CM	0	0	0	0	EC	2	3	5	19
CF	0	1	0	1	EQ	1	1	3	2

ตารางที่ 5.2 ค่าความถี่ของคู่ลำดับอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยธาลัสซีเมีย (ต่อ)

คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV		คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV	
	Positive	Negative	Positive	Negative		Positive	Negative	Positive	Negative
CP	0	0	0	0	EE	1	0	1	0
CS	0	0	0	0	EG	0	1	0	10
EH	0	1	0	3	HA	4	3	33	13
EI	1	3	17	56	HR	2	1	6	7
EL	3	3	8	43	HN	0	0	0	0
EK	0	0	0	0	HD	0	0	0	0
EM	3	1	35	14	HC	0	0	0	0
EF	1	1	19	50	HQ	0	0	0	0
EP	2	3	12	24	HE	3	3	38	24
ES	2	2	12	79	HG	0	1	0	1
ET	3	3	46	80	HH	0	0	0	0
EW	0	0	0	0	HI	1	1	1	4
EY	0	0	0	0	HL	1	0	1	0
EV	0	1	0	1	HK	0	2	0	1
GA	5	1	11	1	HM	0	0	0	0
GR	0	0	0	0	HF	0	0	0	0
GN	0	0	0	0	HP	0	0	0	0
GD	0	0	0	0	HS	1	0	1	0
GC	0	1	0	1	HT	0	0	0	0
GQ	0	0	0	0	HW	0	0	0	0
GE	1	3	6	13	HY	1	3	1	12
GG	1	0	1	0	HV	0	0	0	0
GH	0	0	0	0	IA	2	3	12	80
GI	1	2	127	137	IR	1	2	17	1
GL	4	2	79	21	IN	6	5	96	102
GK	0	0	0	0	ID	2	2	9	25
GM	0	0	0	0	IC	1	3	4	56
GF	1	0	1	0	IQ	0	0	0	0
GP	0	0	0	0	IE	1	2	63	60
GS	0	0	0	0	IG	1	1	6	1
GT	0	1	0	4	IH	0	0	0	0

ตารางที่ 5.2 ค่าความถี่ของคู่ลำดับอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยธาลัสซีเมีย (ต่อ)

คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV		คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV	
	Positive	Negative	Positive	Negative		Positive	Negative	Positive	Negative
GW	0	0	0	0	II	1	1	1	5
GY	0	0	0	0	IL	2	2	4	18
GV	0	0	0	0	IK	1	1	4	2
IM	1	1	3	20	KC	0	0	0	0
IF	1	3	1	5	KQ	0	0	0	0
IP	1	1	1	12	KE	1	1	4	1
IS	2	3	8	50	KG	0	0	0	0
IT	1	2	13	52	KH	0	0	0	0
IW	0	0	0	0	KI	1	4	1	18
IY	0	0	0	0	KL	0	0	0	0
IV	1	2	1	3	KK	0	0	0	0
LA	2	3	8	52	KM	0	0	0	0
LR	0	0	0	0	KF	0	0	0	0
LN	0	0	0	0	KP	0	0	0	0
LD	0	0	0	0	KS	0	0	0	0
LC	0	1	0	1	KT	0	0	0	0
LQ	0	0	0	0	KW	0	0	0	0
LE	1	2	4	3	KY	0	1	0	13
LG	0	0	0	0	KV	0	0	0	0
LH	0	0	0	0	MA	6	3	25	33
LI	2	3	7	19	MR	0	1	0	1
LL	2	1	1	4	MN	0	0	0	0
LK	0	0	0	0	MD	0	0	0	0
LM	0	0	0	0	MC	0	0	0	0
LF	1	0	1	0	MQ	0	0	0	0
LP	3	1	10	4	ME	1	1	14	22
LS	0	1	0	1	MG	0	0	0	0
LT	3	0	5	0	MH	0	0	0	0
LW	0	0	0	0	MI	1	2	4	20
LY	2	2	4	8	ML	0	0	0	0
LV	1	0	1	0	MK	0	0	0	0

ตารางที่ 5.2 ค่าความถี่ของคู่ลำดับอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยธาลัสซีเมีย (ต่อ)

คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV		คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV	
	Positive	Negative	Positive	Negative		Positive	Negative	Positive	Negative
KA	0	1	0	1	MM	5	0	10	0
KR	0	0	0	0	MF	0	0	0	0
KN	1	1	7	1	MP	1	1	1	2
KD	1	0	2	0	MS	0	0	0	0
MT	0	0	0	0	PH	3	4	11	16
MW	0	0	0	0	PI	1	2	63	65
MY	0	0	0	0	PL	1	1	2	1
MV	0	0	0	0	PK	0	1	0	1
FA	2	1	10	12	PM	0	0	0	0
FR	1	0	3	0	PF	0	0	0	0
FN	0	0	0	0	PP	0	1	0	1
FD	0	0	0	0	PS	1	1	1	6
FC	0	0	0	0	PT	1	2	7	11
FQ	0	0	0	0	PW	0	0	0	0
FE	1	0	2	0	PY	1	0	1	0
FG	0	0	0	0	PV	0	0	0	0
FH	0	0	0	0	SA	1	1	63	56
FI	0	2	0	1	SR	0	0	0	0
FL	0	0	0	0	SN	1	0	1	0
FK	0	0	0	0	SD	0	0	0	0
FM	0	0	0	0	SC	1	1	11	28
FF	1	0	1	0	SQ	0	0	0	0
FP	0	0	0	0	SE	1	4	9	37
FS	0	0	0	0	SG	0	0	0	0
FT	0	0	0	0	SH	0	1	0	2
FW	0	0	0	0	SI	1	3	6	39
FY	0	1	0	4	SL	0	1	0	1
FV	0	0	0	0	SK	0	0	0	0
PA	1	4	9	4	SM	1	1	1	1
PR	2	4	19	69	SF	0	0	0	0
PN	0	0	0	0	SP	2	3	8	8

ตารางที่ 5.2 ค่าความถี่ของคู่ลำดับอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยธาลัสซีเมีย (ต่อ)

คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV		คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV	
	Positive	Negative	Positive	Negative		Positive	Negative	Positive	Negative
PD	1	1	12	2	SS	1	2	8	44
PC	1	0	2	0	ST	3	3	12	12
PQ	0	0	0	0	SW	0	0	0	0
PE	1	3	3	23	SY	1	1	1	4
PG	0	0	0	0	SV	0	0	0	0
TA	4	3	64	17	WM	0	0	0	0
TR	3	1	15	19	WF	0	0	0	0
TN	0	0	0	0	WP	0	0	0	0
TD	0	0	0	0	WS	0	0	0	0
TC	0	0	0	0	WT	0	0	0	0
TQ	0	0	0	0	WW	0	0	0	0
TE	3	6	31	110	WY	0	0	0	0
TG	0	0	0	0	WV	0	0	0	0
TH	2	2	8	49	YA	0	0	0	0
TI	3	4	21	37	YR	0	1	0	13
TL	0	0	0	0	YN	0	1	0	6
TK	0	0	0	0	YD	0	2	0	3
TM	0	0	0	0	YC	1	0	3	0
TF	0	0	0	0	YQ	0	0	0	0
TP	0	2	0	6	YE	0	0	0	0
TS	1	0	1	0	YG	0	0	0	0
TT	1	1	11	10	YH	0	0	0	0
TW	0	0	0	0	YI	0	1	0	4
TY	1	1	1	17	YL	2	0	3	0
TV	0	0	0	0	YK	0	0	0	0
WA	0	3	0	4	YM	1	0	1	0
WR	0	0	0	0	YF	1	0	1	0
WN	1	1	7	1	YP	1	1	1	10
WD	0	0	0	0	YS	1	2	8	7
WC	0	0	0	0	YT	2	1	6	3
WQ	0	0	0	0	YW	0	0	0	0

ตารางที่ 5.2 ค่าความถี่ของคู่ลำดับอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยธาลัสซีเมีย (ต่อ)

คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV		คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV	
	Positive	Negative	Positive	Negative		Positive	Negative	Positive	Negative
WE	0	0	0	0	YY	0	0	0	0
WG	0	0	0	0	YV	0	0	0	0
WH	0	0	0	0	VA	1	1	4	12
WI	0	1	0	29	VR	0	0	0	0
WL	0	0	0	0	VN	0	0	0	0
WK	0	0	0	0	VD	0	0	0	0
VC	0	0	0	0	VM	0	0	0	0
VQ	0	0	0	0	VF	0	0	0	0
VE	1	2	1	11	VP	0	1	0	1
VG	0	0	0	0	VS	0	0	0	0
VH	0	0	0	0	VT	0	0	0	0
VI	1	0	16	0	VW	0	0	0	0
VL	0	0	0	0	VY	0	0	0	0
VK	0	0	0	0	VV	1	0	1	0

จากตารางที่ 5.2 แสดงค่าความถี่ของคู่ลำดับกรดอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยธาลัสซีเมียโดยในกลุ่มสายพันธุกรรมซึ่งแสดงอาการของโรค (Positive Sequence) มีค่าความถี่ของคู่ลำดับกรดอะมิโน AA สูงสุดเท่ากับ 1 และมีค่าจำนวนครั้งของคู่ลำดับ (FIV) เท่ากับ 68 และในกลุ่มสายพันธุกรรมซึ่งไม่แสดงอาการของโรค (Negative Sequence) มีค่าค่าความถี่สูงสุดเท่ากับ 2 และมีค่าจำนวนครั้งของคู่ลำดับ (FIV) เท่ากับ 35

ตารางที่ 5.3 ค่าสถิติของข้อมูลผู้ป่วยธาลัสซีเมียแบบ 2-Grams ที่ไม่ปรับค่าน้ำหนัก (NBL)

ค่าสถิติของชุดข้อมูล	ลักษณะข้อมูล	
	แสดงอาการของโรค	ไม่แสดงอาการของโรค
ค่าความถี่สูงสุดของคู่ลำดับ (Max)	6.00	6.00
ค่าเฉลี่ยของความถี่คู่ลำดับ (Mean)	0.48	0.87
ค่าส่วนเบี่ยงเบนมาตรฐาน (StdDev)	0.19	0.26

จากตารางที่ 5.3 แสดงค่าสถิติของข้อมูลผู้ป่วยธาลัสซีเมียแบบ 2-Grams ที่ไม่ปรับค่าน้ำหนัก (NBL) มีความถี่สูงสุดของคู่ลำดับสำหรับกลุ่มที่แสดงอาการของโรคเท่ากับ 6.00 กลุ่มที่ไม่แสดงอาการของโรคเท่ากับ 6.00 มีค่าความถี่เฉลี่ยของคู่ลำดับกลุ่มที่แสดงอาการของโรคเท่ากับ 0.48 กลุ่มที่ไม่แสดงอาการของโรคเท่ากับ 0.87 มีค่าส่วนเบี่ยงเบนมาตรฐานของคู่ลำดับกลุ่มที่แสดงอาการของโรคเท่ากับ 0.19 กลุ่มที่ไม่แสดงอาการของโรคเท่ากับ 0.26 โดยเมื่อพิจารณาทั้งชุดข้อมูล พบว่า มีค่าความถี่สูงสุดเท่ากับ 6.00 ค่าความถี่เฉลี่ยเท่ากับ 0.07 และมีค่าส่วนเบี่ยงเบนมาตรฐานเท่ากับ 0.23

ตารางที่ 5.4 ค่าสถิติของข้อมูลผู้ป่วยธาลัสซีเมียแบบ 2-Grams ที่ปรับค่าน้ำหนัก (WBL)

ค่าสถิติของชุดข้อมูล	**ลักษณะข้อมูล**	
	แสดงอาการของโรค	**ไม่แสดงอาการของโรค**
ค่าความถี่สูงสุดของคู่ลำดับ (Max)	8.33	6.94
ค่าเฉลี่ยของความถี่คู่ลำดับ (Mean)	0.05	0.04
ค่าส่วนเบี่ยงเบนมาตรฐาน (StdDev)	0.23	0.14

จากตารางที่ 5.4 แสดงค่าสถิติของข้อมูลผู้ป่วยธาลัสซีเมียแบบ 2-Grams ที่ปรับค่าน้ำหนัก (WBL) ด้วยสมการ (3.3) โดยมีความถี่สูงสุดของคู่ลำดับสำหรับกลุ่มที่แสดงอาการของโรคเท่ากับ 8.33 กลุ่มที่ไม่แสดงอาการของโรคเท่ากับ 6.94 มีค่าความถี่เฉลี่ยสำหรับกลุ่มที่แสดงอาการของโรคเท่ากับ 0.05 กลุ่มที่ไม่แสดงอาการของโรคเท่ากับ 0.04 มีค่าส่วนเบี่ยงเบนมาตรฐานสำหรับกลุ่มที่แสดงอาการของโรคเท่ากับ 0.23 กลุ่มที่ไม่แสดงอาการของโรคเท่ากับ 0.14 เมื่อพิจารณาทั้งชุดข้อมูล พบว่า มีค่าความถี่สูงสุดเท่ากับ 8.33 ค่าความถี่เฉลี่ยเท่ากับ 0.42 และมีค่าส่วนเบี่ยงเบนมาตรฐานเท่ากับ 0.19

จะสังเกตเห็นได้ว่า ลักษณะของค่าสถิติที่ได้จากตารางที่ 5.3 และตารางที่ 5.4 ซึ่งเป็นข้อมูลแบบไม่ปรับค่าน้ำหนัก (NBL) และแบบปรับค่าน้ำหนัก (WBL) มีความแตกต่างกัน

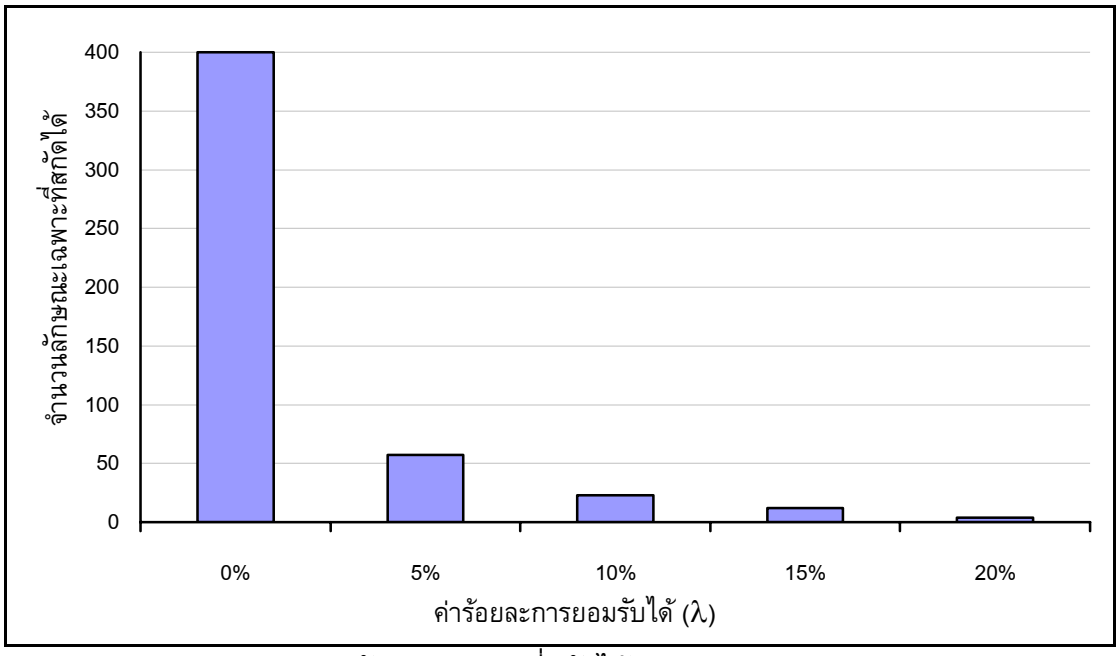
5.3.2 ขั้นตอนหลักที่ 2 สกัดลักษณะเฉพาะบนข้อมูลผู้ป่วยธาลัสซีเมีย

จำนวนลักษณะเฉพาะที่สกัดได้จากการทดลองแบบ C และ D โดยกำหนดค่าร้อยละการยอมรับได้ (Support Threshold Point: λ) ที่แตกต่างกันในขั้นตอนการวิเคราะห์ระดับชุดข้อมูล (Global Sequences Analyze) ด้วยค่า λ เท่ากับ 0%, 5%, 10%, 15% และ 20% ตามลำดับ และสามารถแสดงจำนวนของลักษณะเฉพาะที่สกัดได้ ดังตารางที่ 5.5

ตารางที่ 5.5 จำนวนลักษณะเฉพาะที่สกัดได้จากข้อมูลผู้ป่วยธาลัสซีเมีย จากการทดลองแบบ C และ D เมื่อกำหนดค่าร้อยละการยอมรับได้ (λ) ที่แตกต่างกัน

วิธีการทดลอง	จำนวนลักษณะเฉพาะที่สกัดได้				
	$\lambda \geq 0\%$	$\lambda \geq 5\%$	$\lambda \geq 10\%$	$\lambda \geq 15\%$	$\lambda \geq 20\%$
แบบ C และ D	400	57	23	12	4

จากตารางที่ 5.5 พบว่า เมื่อกำหนดค่า $\lambda \geq 0\%$ ได้จำนวนคู่ลำดับกรดอะมิโนที่ใช้สำหรับใช้เป็นคุณสมบัติของข้อมูล (Attributes) ทั้งสิ้น 400 คุณสมบัติ หรือหมายถึงใช้คู่ลำดับกรดอะมิโนทั้งหมด 400 คู่ลำดับโดยไม่มีการสกัดออก ที่ค่า $\lambda \geq 5\%$ สามารถสกัดได้ 57 คุณสมบัติ ได้แก่ AA, AR, AN, AC, AL, AM, AP, AS, AT, RA, RE, RI, RM, NA, NI, NP, NS, NT, DI, CA, CH, CI, CT, EA, EN, ED, EI, EL, EM, EF, EP, ES, ET, GI, GL, HA, HE, IA, IN, IC, IE, IS, IT, LA, MA, ME, PR, PI, SA, SC, SE, SI, SS, TA, TE, TH และ TI ที่ค่า $\lambda \geq 10\%$ สามารถสกัดได้ 23 คุณสมบัติ ได้แก่ AA, AN, AP, AS, AT, RE, NS, DI, EN, ED, EI, ES, ET, GI, GL, IA, IN, IE, PR, PI, SA, TA และ TE ตามลำดับ ที่ค่า $\lambda \geq 15\%$ สามารถสกัดได้ 12 คุณสมบัติ ได้แก่ AP, AT, RE, NS, EN, ET, GI, IN, IE, PI, SA และ TE และที่ค่า $\lambda \geq 20\%$ สามารถสกัดได้ 4 คุณสมบัติ ได้แก่ EN, GI, IN และ TE สามารถแสดงแผนภูมิเปรียบเทียบจำนวนลักษณะเฉพาะที่สกัดได้ ดังภาพประกอบ 5.3



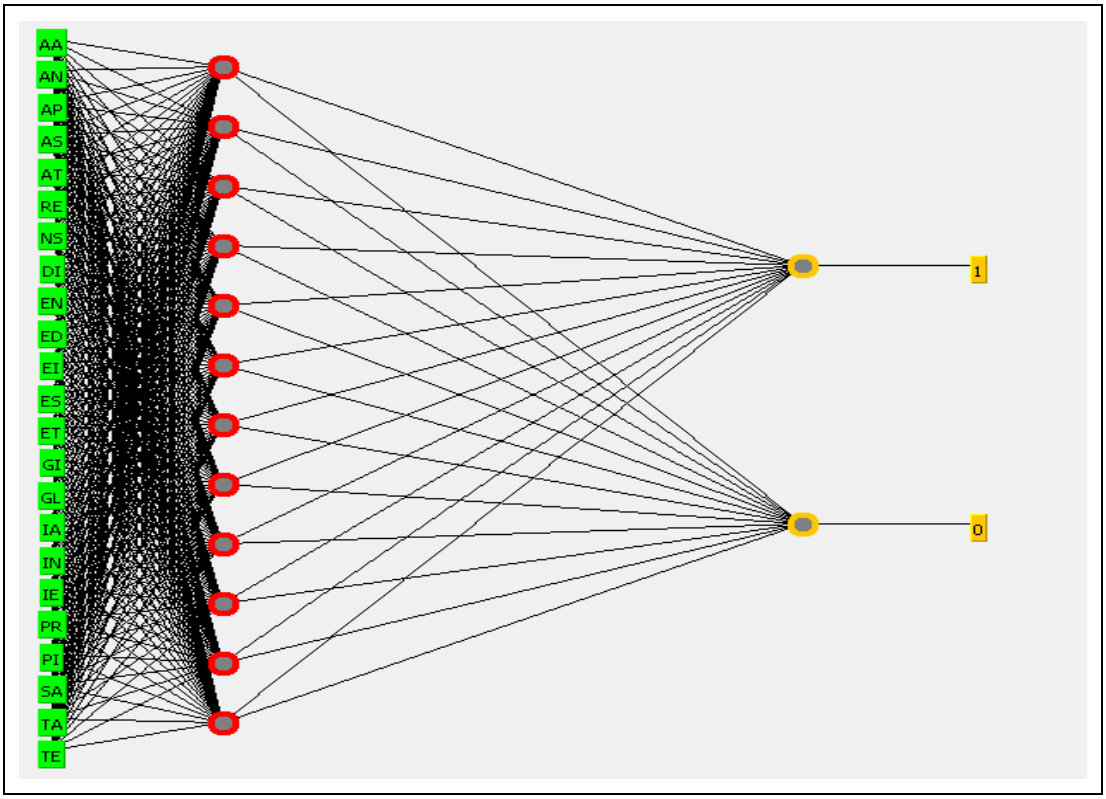
ภาพประกอบ 5.3 จำนวนลักษณะเฉพาะที่สกัดได้จากการทดลองแบบ C และ D บนชุดข้อมูลผู้ป่วยธาลัสซีเมียเมื่อกำหนดค่าร้อยละการยอมรับได้ (λ) ที่แตกต่างกัน

จากภาพประกอบ 5.3 พบว่า จำนวนลักษณะเฉพาะที่สกัดได้จะมีจำนวนลดลง เมื่อกำหนดค่าร้อยละของการยอมรับได้ (λ) ที่สูงขึ้น

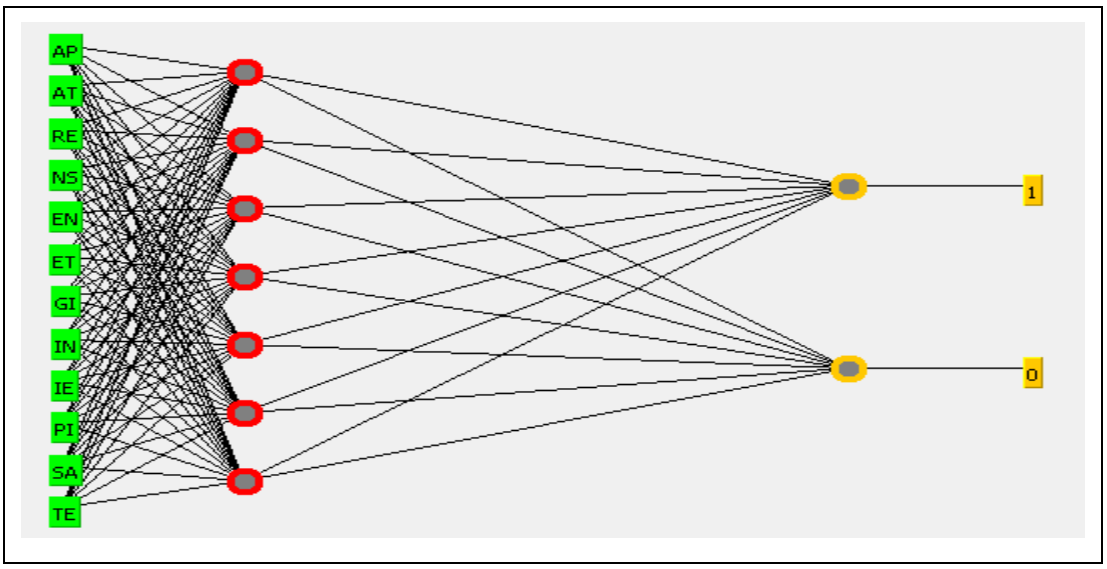
5.3.3 ขั้นตอนหลักที่ 3 จำแนกข้อมูลผู้ป่วยธาลัสซีเมียด้วยลักษณะเฉพาะ

ทำการจำแนกโดยใช้ชุดเครื่องมือ (Tools Box) ของโปรแกรม WEKA เวอร์ชัน 3.4.10 ประกอบด้วยขั้นตอนวิธี 2 กลุ่ม คือ กลุ่มที่ 1 โครงข่ายประสาทเทียมหลายชั้น (MLP) และโครงข่ายประสาทเทียมเรเดียลเบสิสฟังก์ชัน (RBF) สำหรับกลุ่มที่ 2 ได้แก่ ต้นไม้การตัดสินใจแบบ Alternating Decision Tree (ADTree) แบบ Pruned C4.5 Decision Tree (J48) และแบบ K Random Features Decision Tree (RandomTree) ตามลำดับ

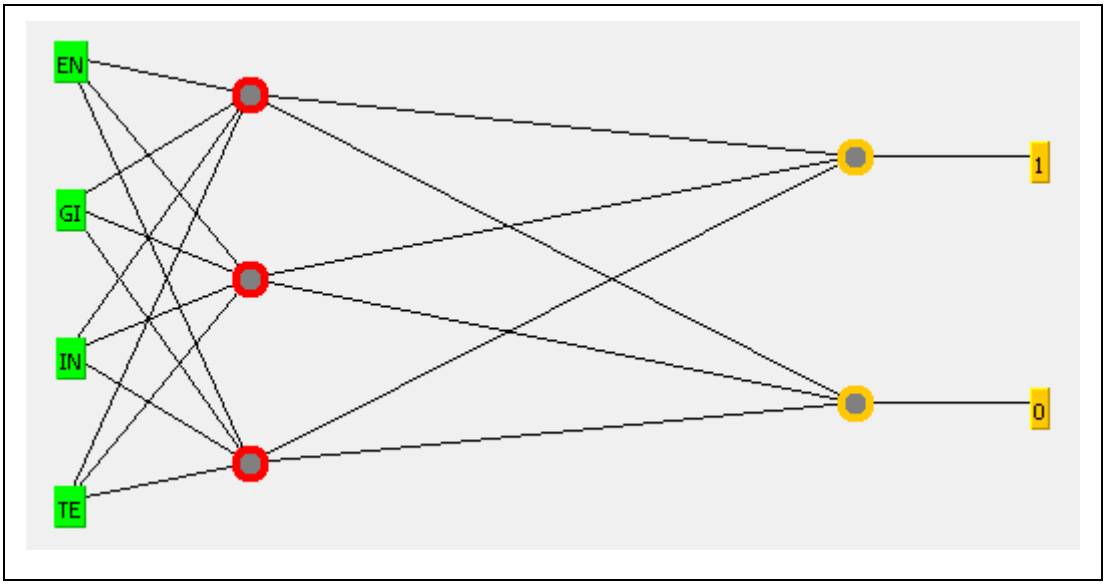
โดยสามารถแสดงลักษณะตัวอย่างโครงข่ายประสาทเทียมแบบหลายชั้น (MLP) ที่ใช้ในการจำแนกข้อมูลได้ดังภาพประกอบ 5.4 ถึงภาพประกอบ 5.6 ตามลำดับ



ภาพประกอบ 5.4 โครงข่ายประสาทเทียมแบบ MLP สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยธาลัสซีเมีย เมื่อกำหนดค่า $\lambda \geq 10\%$ (Input = 23 ตัว)



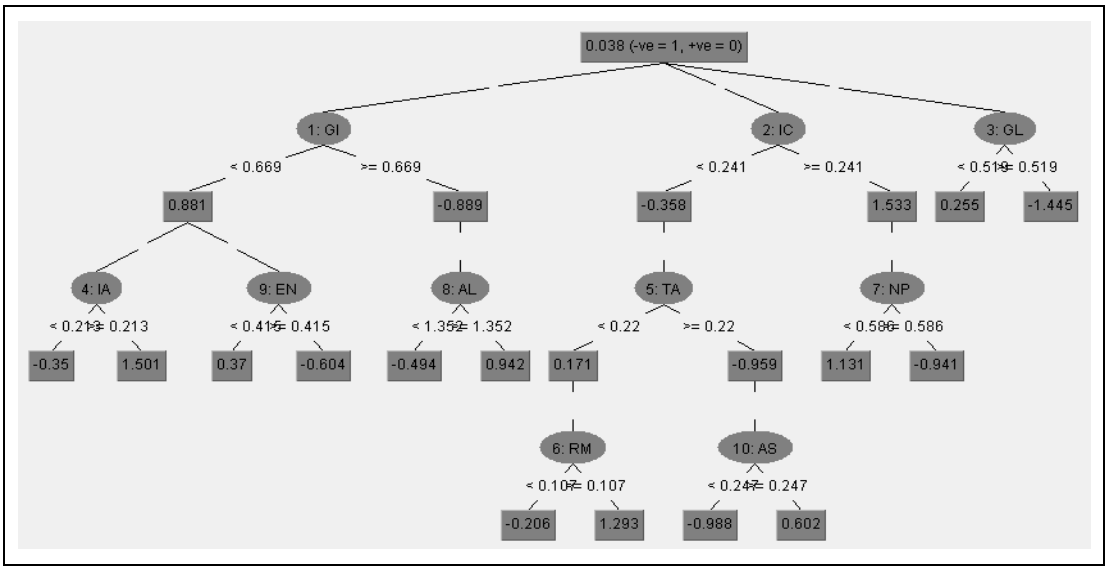
ภาพประกอบ 5.5 โครงข่ายประสาทเทียมแบบ MLP สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยธาลัสซีเมีย เมื่อกำหนดค่า $\lambda \geq 15\%$ (Input = 12 ตัว)



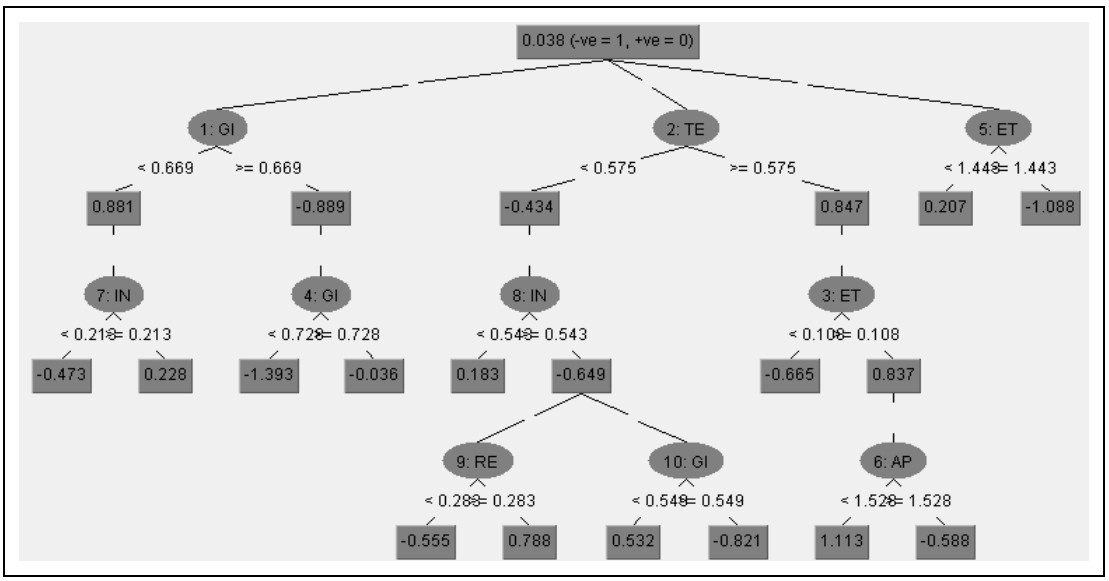
ภาพประกอบ 5.6 โครงข่ายประสาทเทียมแบบ MLP สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยธาลัสซีเมีย เมื่อกำหนดค่า $\lambda \geq 20\%$ (Input = 4 ตัว)

จากภาพประกอบ 5.4 ถึงภาพประกอบ 5.6 ลักษณะของโครงข่ายประสาทเทียมแบบหลายชั้น (MLP) คือ ชั้นที่ 1 ชั้นข้อมูลนำเข้าผ่านทางคุณสมบัติของข้อมูลซึ่งสกัดได้ ชั้นที่ 2 เป็นชั้นซ่อน และในชั้นที่ 3 เป็นชั้นผลลัพธ์ ประกอบด้วยผลลัพธ์จำนวน 2 ค่า คือ 1 แทน ความหมายว่าแสดงอาการของโรค (Positive) และค่า 0 แทน ความหมายว่าไม่แสดงอาการของโรค (Negative)

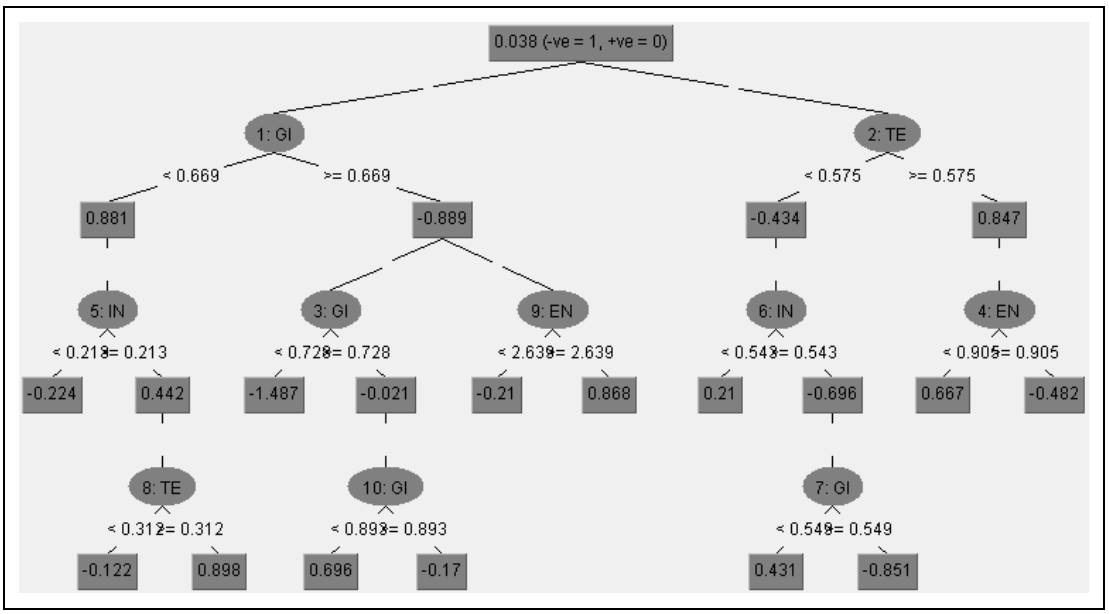
และสามารถแสดงตัวอย่างของต้นไม้การตัดสินใจสำหรับการจำแนกข้อมูลด้วยขั้นตอนวิธีแบบ Alternating Decision Tree (ADTree) ดังภาพประกอบ 5.7 ถึงภาพประกอบ 5.9 ตามลำดับ



ภาพประกอบ 5.7 ต้นไม้การตัดสินใจแบบ ADTree สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยธาลัสซีเมีย เมื่อกำหนดค่า $\lambda \geq 10\%$

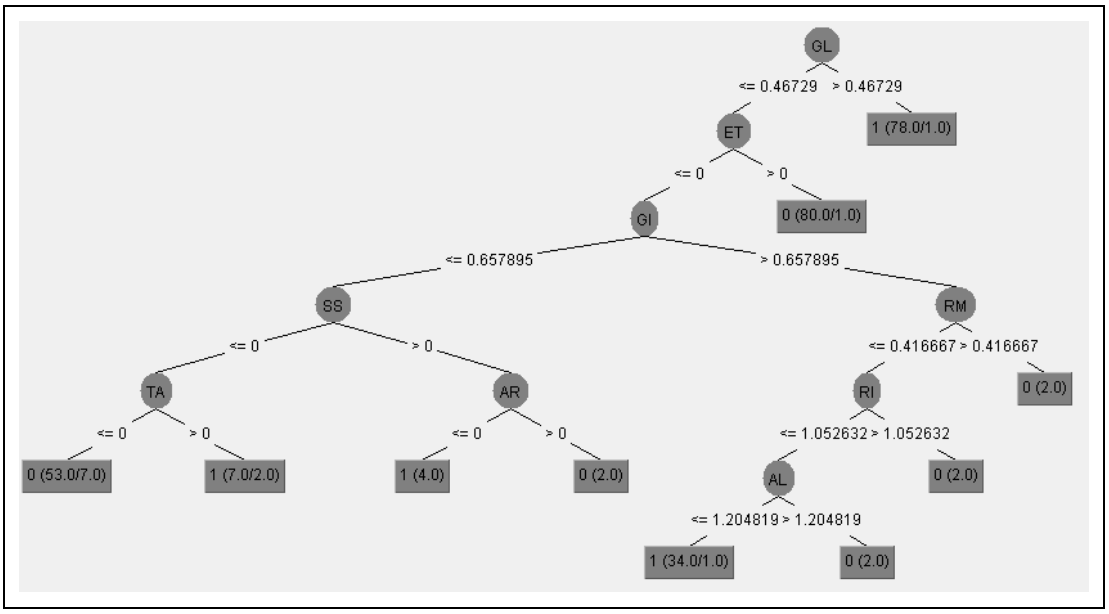


ภาพประกอบ 5.8 ต้นไม้การตัดสินใจแบบ ADTree สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยธาลัสซีเมีย เมื่อกำหนดค่า $\lambda \geq 15\%$

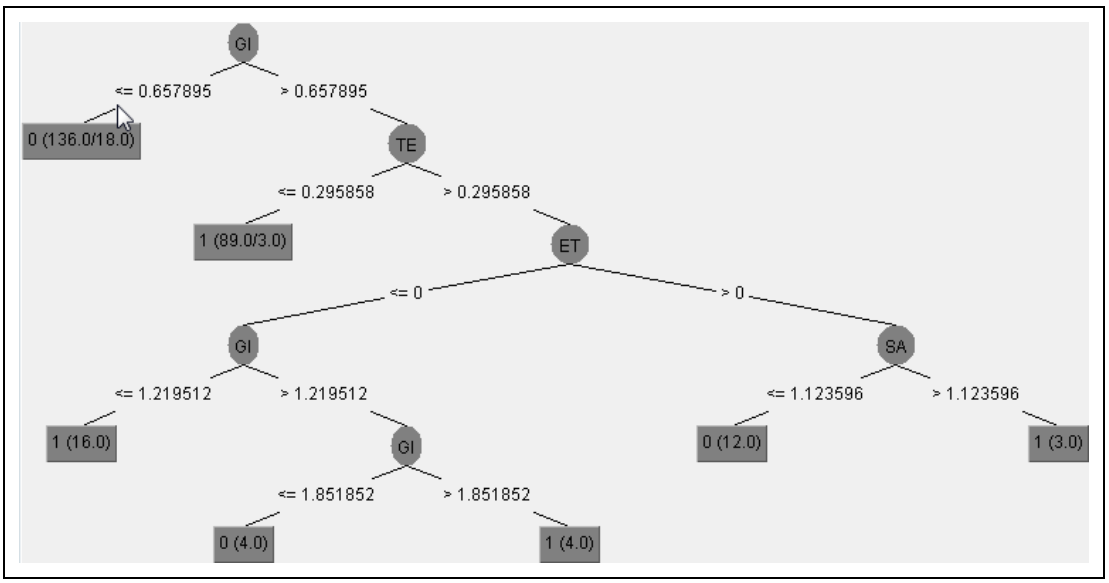


ภาพประกอบ 5.9 ต้นไม้การตัดสินใจแบบ ADTree สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยธาลัสซีเมีย เมื่อกำหนดค่า $\lambda \geq 20\%$

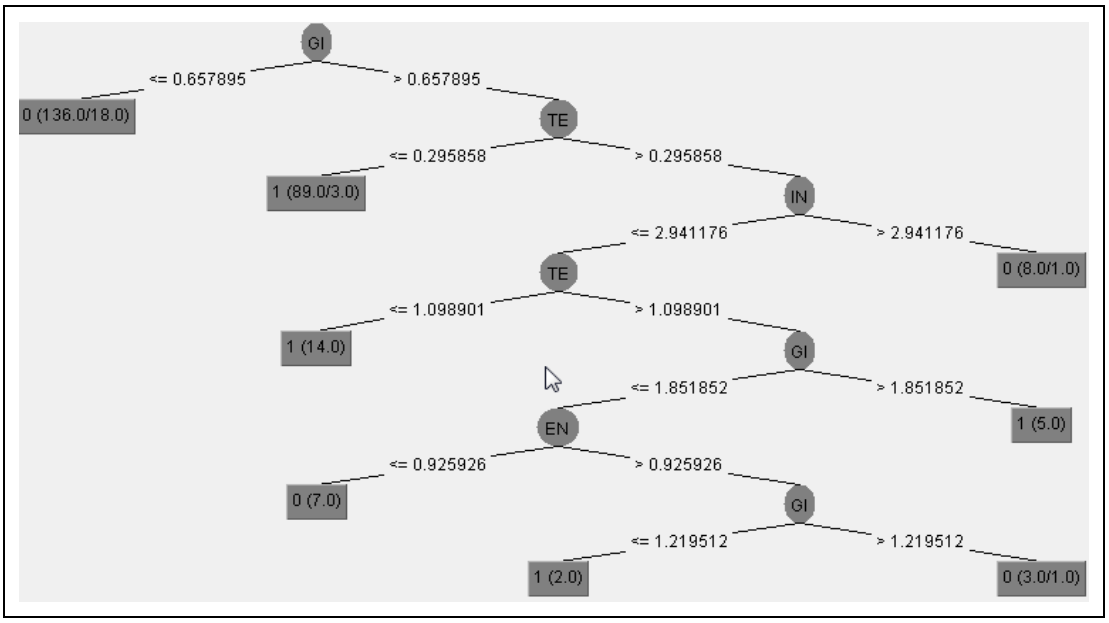
และสามารถแสดงตัวอย่างของต้นไม้การตัดสินใจสำหรับการจำแนกข้อมูลด้วยขั้นตอนวิธีแบบ Pruned C4.5 Decision Tree (J48) ดังภาพประกอบ 5.10 ถึงภาพประกอบ 5.12 ตามลำดับ



ภาพประกอบ 5.10 ต้นไม้การตัดสินใจแบบ J48 สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยธาลัสซีเมีย เมื่อกำหนดค่า $\lambda \geq 10\%$



ภาพประกอบ 5.11 ต้นไม้การตัดสินใจแบบ J48 สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยธาลัสซีเมีย เมื่อกำหนดค่า $\lambda \geq 15\%$



ภาพประกอบ 5.12 ต้นไม้การตัดสินใจแบบ J48 สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยธาลัสซีเมีย เมื่อกำหนดค่า $\lambda \geq 20\%$

จากภาพประกอบ 5.7 ถึงภาพประกอบ 5.12 พบว่า ลักษณะต้นไม้การตัดสินใจแบบ ADTree มีความสมดุล (Balance) มากกว่าต้นไม้การตัดสินใจแบบ J48 และพบว่าต้นไม้การตัดสินใจแบบ J48 มีจำนวนชั้นของต้นไม้มากกว่าแบบ ADTree

5.3.4 เปรียบเทียบผลการทดลองข้อมูลผู้ป่วยธาลัสซีเมีย

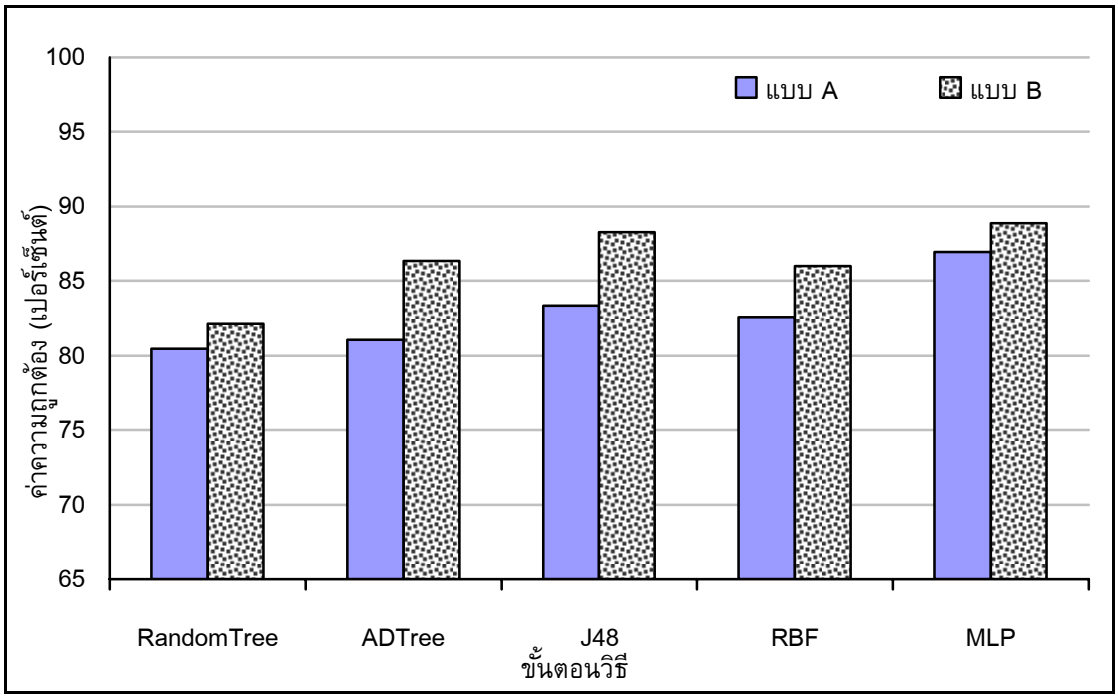
5.3.4.1 เปรียบเทียบค่าความถูกต้องของการจำแนกข้อมูล

ค่าความถูกต้องของการจำแนกข้อมูลตามแบบจำลองการวิเคราะห์รูปและรู้จำลำดับพันธุกรรม สามารถแสดงดังตารางที่ 5.6 และตารางที่ 5.7 ตามลำดับ

ตารางที่ 5.6 ความถูกต้องการจำแนกข้อมูลผู้ป่วยธาลัสซีเมียของการทดลองแบบ A และ B

การทดลอง	จำนวนคุณสมบัติ	ค่าความถูกต้อง (เปอร์เซ็นต์)				
		RandomTree	ADTree	J48	RBF	MLP
แบบ A	400	80.47%	81.06%	83.33%	82.58%	86.96%
แบบ B	400	82.12%	86.36%	88.26%	85.98%	88.87%

จากตารางที่ 5.5 สามารถแสดงแผนภูมิเปรียบเทียบค่าความถูกต้องการจำแนกข้อมูลผู้ป่วยธาลัสซีเมียด้วยขั้นตอนวิธีแบบต่าง ๆ ของการทดลองแบบ A และ B แสดงดังภาพประกอบ 5.13



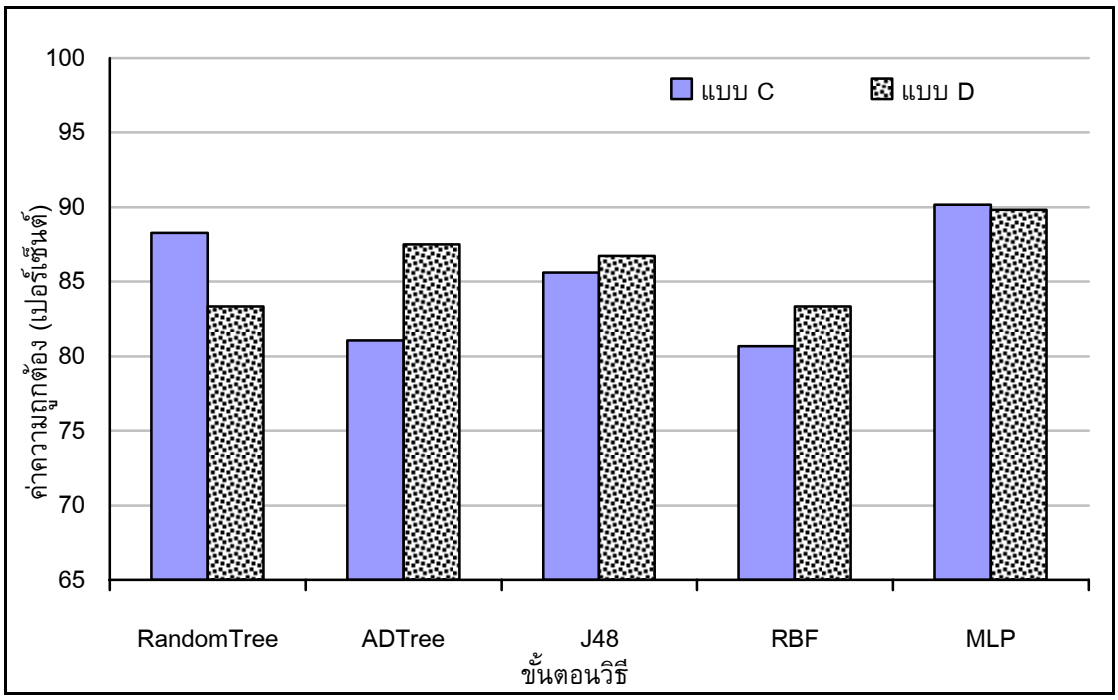
ภาพประกอบ 5.13 ความถูกต้องการจำแนกข้อมูลผู้ป่วยธาลัสซีเมียของการทดลองแบบ AและB

จากตารางที่ 5.6 และภาพประกอบ 5.13 พบว่า ในการจำแนกข้อมูลของการทดลองซึ่งมีการปรับค่าน้ำหนักข้อมูล (WBL) จะให้ค่าความถูกต้อง (Accuracy) สูงกว่าการทดลองแบบไม่ปรับค่าน้ำหนักของข้อมูล (NBL) โดยขั้นตอนวิธีแบบโครงข่ายประสาทเทียมหลายชั้น (MLP) ให้ค่าความถูกต้องในการจำแนกสูงที่สุด เท่ากับ 88.87% รองลงมาคือ J48 เท่ากับ 88.26% ADTree เท่ากับ 86.36% RBF เท่ากับ 85.98% และ RandomTree เท่ากับ 82.12% ตามลำดับ และสามารถแสดงค่าความถูกต้องการจำแนกของการทดลองแบบ C และ D เมื่อกำหนดค่าร้อยละการยอมรับได้ (λ) ที่แตกต่างกัน ดังตารางที่ 5.7

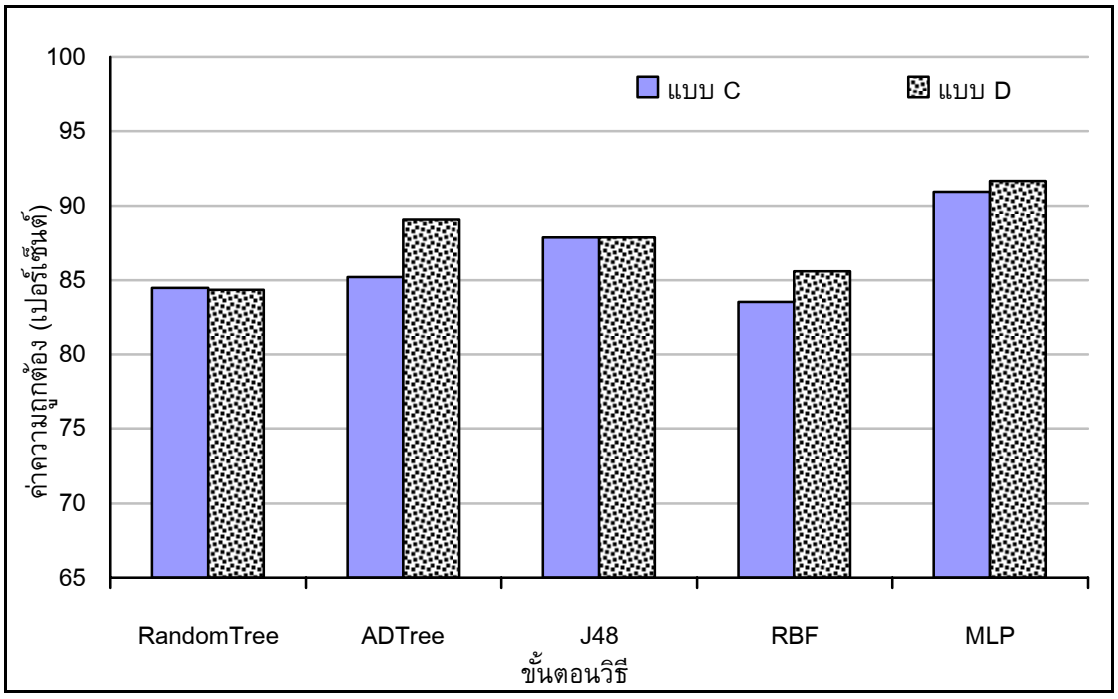
ตารางที่ 5.7 ความถูกต้องการจำแนกข้อมูลผู้ป่วยธาลัสซีเมียของการทดลองแบบ C และ D เมื่อกำหนดค่าร้อยละการยอมรับได้ (λ) ที่แตกต่างกัน

วิธีการทดลอง	λ	จำนวนคุณสมบัติ	ค่าความถูกต้อง (เปอร์เซ็นต์)				
			RandomTree	ADTree	J48	RBF	MLP
แบบ C	0%	400	80.47%	81.06%	83.33%	82.58%	86.96%
	5%	57	88.26%	81.06%	85.61%	80.68%	90.15%
	10%	23	84.47%	85.23%	87.88%	83.55%	90.91%
	15%	12	83.33%	76.52%	79.92%	73.11%	78.03%
	20%	4	80.30%	79.17%	78.41%	72.73%	78.03%
แบบ D	0%	400	82.12%	86.36%	88.26%	85.98%	88.87%
	5%	57	83.33%	87.50%	86.74%	83.33%	89.82%
	10%	23	84.33%	89.08%	87.88%	85.61%	91.64%
	15%	12	84.46%	84.09%	87.30%	75.00%	85.23%
	20%	4	83.33%	85.61%	86.36%	80.68%	85.04%

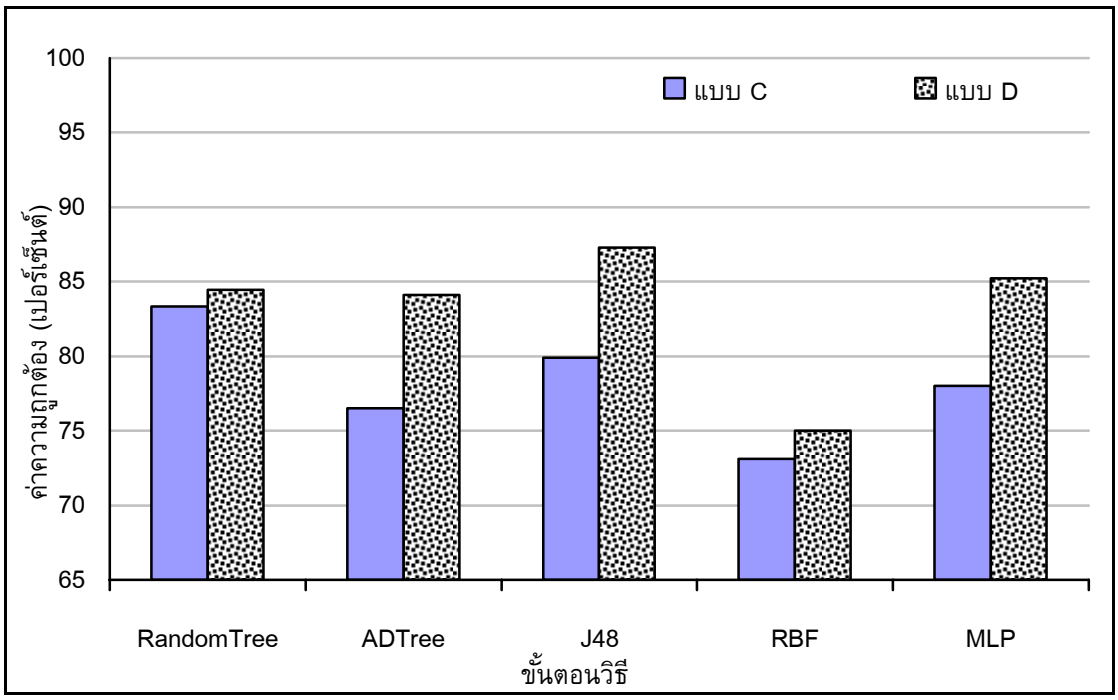
จากตารางที่ 5.7 สามารถแสดงแผนภูมิเปรียบเทียบความถูกต้องการจำแนกข้อมูลสายพันธุกรรมผู้ป่วยธาลัสซีเมียด้วยการทดลองแบบ C และ D ดังแสดงในภาพประกอบ 5.14 ถึงภาพประกอบ 5.17 ตามลำดับ



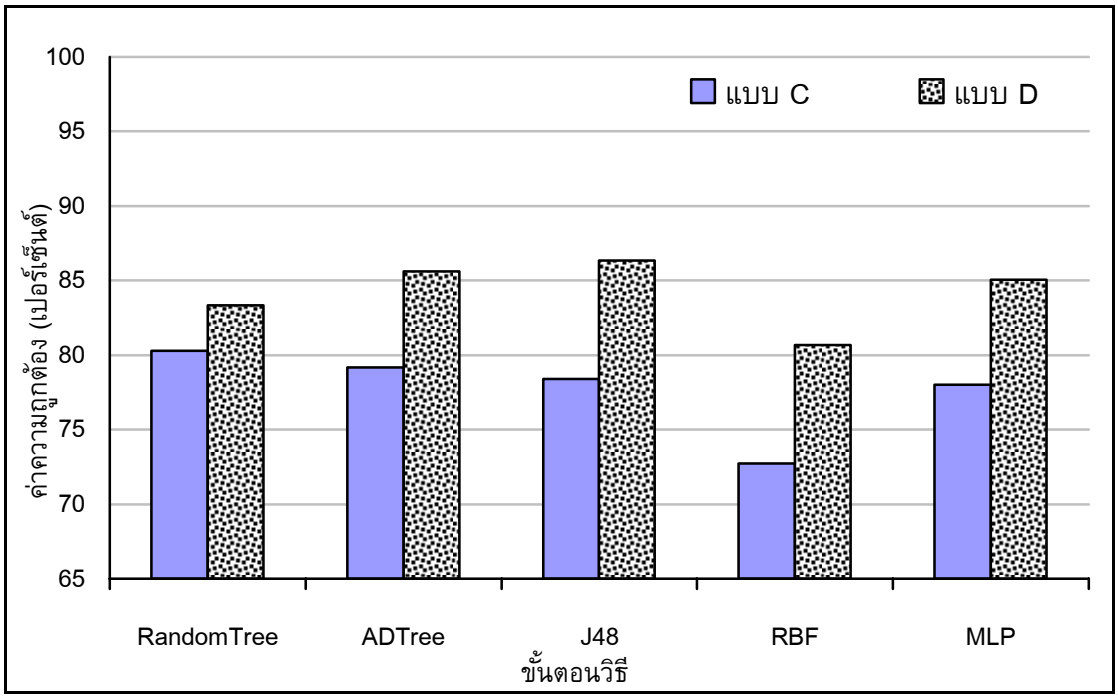
ภาพประกอบ 5.14 ความถูกต้องการจำแนกข้อมูลผู้ป่วยธาลัสซีเมียของการทดลองแบบ C และ D เมื่อกำหนดค่า $\lambda \geq 5\%$



ภาพประกอบ 5.15 ความถูกต้องการจำแนกข้อมูลผู้ป่วยธาลัสซีเมียของการทดลองแบบ C และ D เมื่อกำหนดค่า $\lambda \geq 10\%$



ภาพประกอบ 5.16 ความถูกต้องการจำแนกข้อมูลผู้ป่วยธาลัสซีเมียของการทดลองแบบ C และ D เมื่อกำหนดค่า $\lambda \geq 15\%$



ภาพประกอบ 5.17 ความถูกต้องการจำแนกข้อมูลผู้ป่วยธาลัสซีเมียของการทดลองแบบ C และ D เมื่อกำหนดค่า $\lambda \geq 20\%$

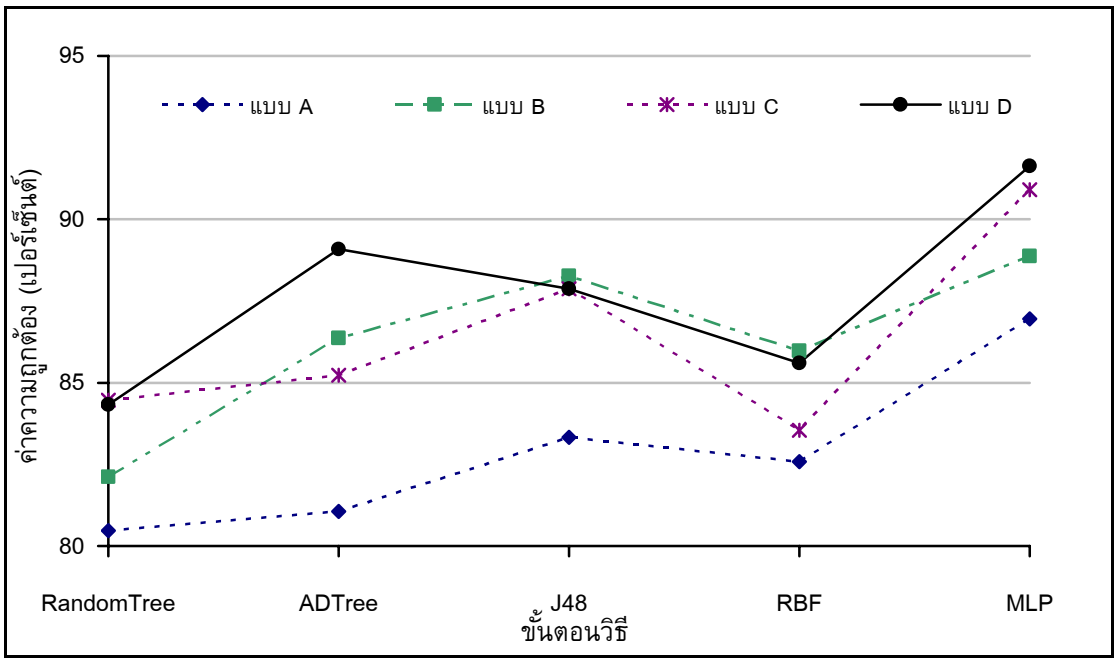
จากภาพประกอบ 5.14 ถึงภาพประกอบ 5.17 พบว่า โดยรวมการทดลองแบบปรับค่าน้ำหนัก (WBL) ให้ค่าความถูกต้องในการจำแนกข้อมูลสูงกว่าการทดลองแบบ ไม่ปรับค่าน้ำหนัก (NBL) และจากภาพประกอบ 5.9 ซึ่งมีการกำหนดค่า $\lambda \geq 10\%$ ให้ค่าความถูกต้องสูงที่สุด และเมื่อเปรียบเทียบขั้นตอนในการจำแนกข้อมูลพบว่าขั้นตอนวิธีแบบ MLP ให้ค่าความถูกต้องสูงสุดเท่ากับ 90.87% ADTree เท่ากับ 89.40% J48 เท่ากับ 87.88% RBF เท่ากับ 85.61% และ RandomTree เท่ากับ 84.85% ตามลำดับ

โดยเมื่อนำผลการจำแนกข้อมูลของการทดลองแบบ A และ B เปรียบเทียบกับการทดลองแบบ C และ D ซึ่งกำหนดค่า $\lambda \geq 10\%$ สามารถแสดงได้ดังตารางที่ 5.8 และภาพประกอบ 5.18

ตารางที่ 5.8 เปรียบเทียบค่าความถูกต้องการจำแนกข้อมูลผู้ป่วยธาลัสซีเมีย

การทดลอง	จำนวนคุณสมบัติ	ค่าความถูกต้อง (เปอร์เซ็นต์)				
		RandomTree	ADTree	J48	RBF	MLP
แบบ A	400	80.47%	81.06%	83.33%	82.58%	86.96%
แบบ B	400	82.12%	86.36%	88.26%	85.98%	88.87%
แบบ C	23	84.47%	85.23%	87.88%	83.55%	90.91%
แบบ D	23	84.33%	89.08%	87.88%	85.61%	91.64%

*หมายเหตุ การทดลองแบบ C และ D กำหนดค่าร้อยละการยอมรับได้เท่ากับ 10 ($\lambda \geq 10\%$)



ภาพประกอบ 5.18 เปรียบเทียบค่าความถูกต้องการจำแนกข้อมูลผู้ป่วยธาลัสซีเมีย

5.3.4.2 เปรียบเทียบเวลาที่ใช้ในการจำแนกข้อมูล

สามารถแสดงเวลาที่ใช้ในการจำแนกข้อมูลผู้ป่วยธาลัสซีเมียของการทดลองแบบ A และ B ดังตารางที่ 5.9 และการทดลองแบบ C และ D ดังตารางที่ 5.10 ตามลำดับ

ตารางที่ 5.9 เวลาที่ใช้ในการจำแนกข้อมูลผู้ป่วยธาลัสซีเมียของการทดลองแบบ A และ B

การทดลอง	จำนวนคุณสมบัติ	เวลาที่ใช้ (วินาที)				
		RandomTree	ADTree	J48	RBF	MLP
แบบ A	400	1.36	2.72	1.28	1.25	84.92
แบบ B	400	1.36	2.81	1.34	1.48	85.66

ตารางที่ 5.10 เวลาที่ใช้ในการจำแนกข้อมูลผู้ป่วยธาลัสซีเมียของการทดลองแบบ C และ D เมื่อกำหนดค่าร้อยละการยอมรับได้ (λ) ที่แตกต่างกัน

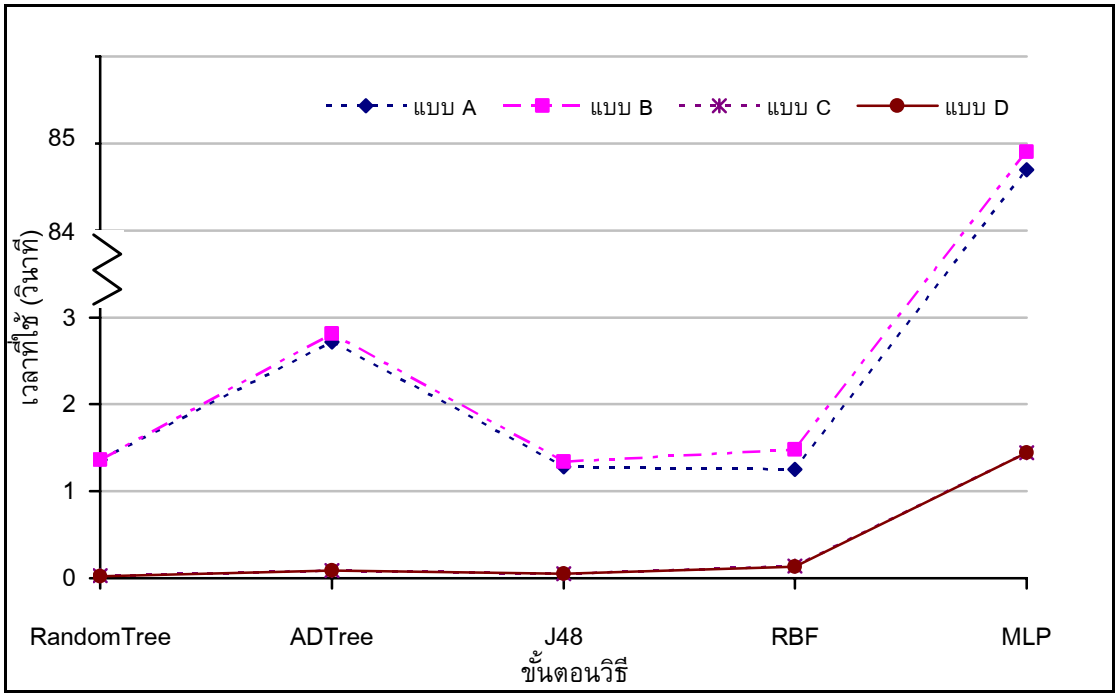
การทดลอง	λ	จำนวนคุณสมบัติ	เวลาที่ใช้ (วินาที)				
			RandomTree	ADTree	J48	RBF	MLP
แบบ C	0%	400	1.36	2.72	1.28	1.25	84.92
	5%	57	0.06	0.24	0.10	0.26	4.70
	10%	23	0.03	0.09	0.05	0.14	1.44
	15%	12	0.05	0.07	0.07	0.23	0.61
	20%	4	0.02	0.05	0.05	0.47	0.22
แบบ D	0%	400	1.36	2.81	1.34	1.48	85.66
	5%	57	0.05	0.89	0.10	0.35	5.80
	10%	23	0.02	0.09	0.05	0.13	1.44
	15%	12	0.03	0.06	0.05	0.44	0.69
	20%	4	0.04	0.03	0.03	0.27	0.24

จากตารางที่ 5.9 และตารางที่ 5.10 พบว่า การทดลองแบบ A และ B ใช้เวลาในการจำแนกสูงกว่าการทดลองแบบ C และ D และขั้นตอนวิธีแบบ MLP ใช้เวลาในการจำแนกสูงที่สุด และขั้นตอนวิธีแบบ RandomTree ใช้เวลาในการจำแนกน้อยที่สุด ตารางที่ 5.11 และภาพประกอบ 5.19 แสดงการเปรียบเทียบเวลาที่ใช้ในการจำแนกข้อมูลของการทดลองแบบ A และ B กับการทดลองแบบ C และ D เมื่อกำหนดค่า $\lambda \geq 10\%$ สามารถแสดงรายละเอียดดังตารางที่ 5.11 และภาพประกอบ 5.19 ตามลำดับ

ตารางที่ 5.11 เปรียบเทียบเวลาที่ใช้จำแนกข้อมูลผู้ป่วยธาลัสซีเมีย

การทดลอง	จำนวนคุณสมบัติ	เวลาที่ใช้ (วินาที)				
		RandomTree	ADTree	J48	RBF	MLP
แบบ A	400	1.36	2.72	1.28	1.25	84.92
แบบ B	400	1.36	2.81	1.34	1.48	85.66
แบบ C	23	0.03	0.09	0.05	0.14	1.44
แบบ D	23	0.02	0.09	0.05	0.13	1.44

*หมายเหตุ การทดลองแบบ C และ D กำหนดค่าร้อยละการยอมรับได้เท่ากับ 10 ($\lambda \geq 10\%$)



ภาพประกอบ 5.19 เปรียบเทียบเวลาที่ใช้บนชุดข้อมูลผู้ป่วยธาลัสซีเมีย

5.3.4.3 การวิจารณ์ผลทดลองข้อมูลผู้ป่วยธาลัสซีเมีย

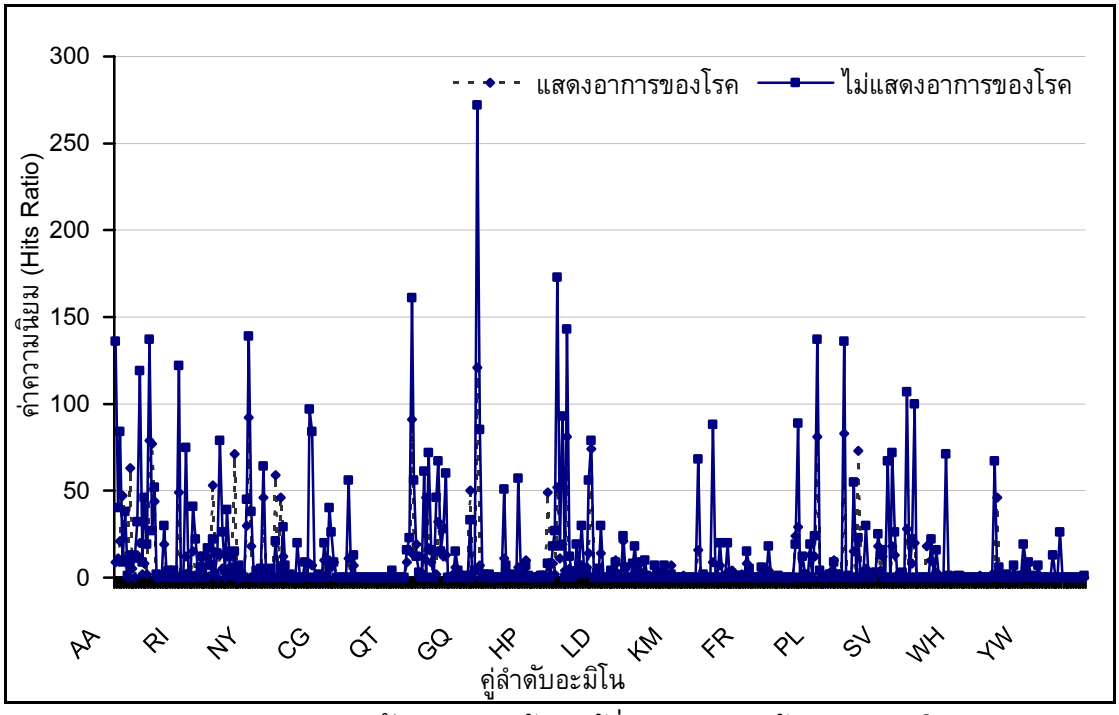
จากการทดลองแบบ A, B, C และ D (ตารางที่ 5.1) พบว่าการจำแนกข้อมูลผู้ป่วยธาลัสซีเมีย โดยทดลองด้วยการปรับค่าน้ำหนักข้อมูล (WBL) ให้ค่าความถูกต้องสูงกว่าการทดลองแบบไม่ปรับค่าน้ำหนักข้อมูล (NBL) เมื่อพิจารณาขั้นตอนวิธีที่ใช้ พบว่า ขั้นตอนวิธีโครงข่ายประสาทเทียมแบบหลายชั้น (MLP) ให้ค่าความถูกต้องในการจำแนกสูงที่สุด รองลงมาคือ ADTree J48 RBF และ RandomTree ตามลำดับ และพบว่าขั้นตอนวิธีแบบ RandomTree ใช้เวลาน้อยที่สุด และโครงข่ายประสาทเทียมแบบหลายชั้น (MLP) ใช้เวลา

มากที่สุด และพบว่าในขั้นตอนการสกัดลักษณะเฉพาะของการทดลองแบบ C และ D สามารถลดจำนวนคุณสมบัติของข้อมูลที่ใช้ในการจำแนกลงได้ โดยที่ค่า $\lambda \geq 5\%$ สามารถสกัดได้ 57 คุณสมบัติ ที่ค่า $\lambda \geq 10\%$ สกัดได้ 23 คุณสมบัติ ที่ค่า $\lambda \geq 15\%$ สกัดได้ 12 คุณสมบัติ และที่ค่า $\lambda \geq 20\%$ สกัดได้ 4 คุณสมบัติ ตามลำดับ ในการจำแนกข้อมูลผู้ป่วยธาลัสซีเมีย เมื่อกำหนดที่ค่า $\lambda \geq 10\%$ จะให้ค่าความถูกต้องการจำแนกข้อมูลที่สูง

5.4 ข้อมูลผู้ป่วยเบาหวาน (Diabetes Disease)

ข้อมูลผู้ป่วยเบาหวานประกอบด้วยสายโปรตีนทั้งสิ้น 393 สาย โดยมีความยาวที่ใช้อยู่ระหว่าง 30 (±5%) ตัวอักษร ถึง 152.58 (±5%) ตัวอักษร แบ่งเป็นข้อมูลซึ่งแสดงอาการของโรคจำนวน 121 สาย และไม่แสดงอาการของโรคจำนวน 272 สาย ประกอบด้วยขั้นตอนในการทดลอง 3 ขั้นตอน คือ การเตรียมข้อมูลค่าความถี่แบบปรับน้ำหนัก การสกัดลักษณะเฉพาะ และการจำแนกสายพันธุกรรม โดยสามารถอธิบายรายละเอียดการทดลองตามลำดับ ดังนี้

5.4.1 ขั้นตอนที่ 1 การเตรียมข้อมูลผู้ป่วยเบาหวานแบบปรับน้ำหนักค่าความถี่ของคู่ลำดับกรดอะมิโนแบบ 2-Grams จากข้อมูลผู้ป่วยเบาหวาน สามารถแสดงลักษณะของข้อมูลด้วยค่าความนิยม (Hits Ratio) ของคู่ลำดับกรดอะมิโน ดังภาพประกอบ 5.20



ภาพประกอบ 5.20 ลักษณะของข้อมูลผู้ป่วยเบาหวานด้วยค่าความนิยม

จากภาพประกอบ 5.20 พบว่า ในกลุ่มข้อมูลซึ่งแสดงอาการของโรค มีคู่ลำดับกรดอะมิโน GI มีค่าความนิยมสูงสุดเท่ากับ 127 สำหรับ และในกลุ่มข้อมูลซึ่งไม่แสดงอาการของโรค มีค่าความนิยมสูงสุด 272 และสามารถแสดงค่าความถี่ของคู่ลำดับกรดอะมิโนแบบ 2-Grams ทั้ง 400 คู่ลำดับของข้อมูลข้อมูลผู้ป่วยเบาหวาน ได้ดังตารางที่ 5.12

ตารางที่ 5.12 ค่าความถี่ของคู่ลำดับอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยเบาหวาน

คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV		คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV	
	Positive	Negative	Positive	Negative		Positive	Negative	Positive	Negative
AA	2	1	9	136	RM	2	2	15	41
AR	3	2	11	40	RF	0	1	0	22
AN	5	5	21	84	RP	1	1	1	1
AD	1	1	47	9	RS	1	1	12	6
AC	3	2	10	38	RT	3	1	6	12
AQ	0	1	0	1	RW	0	0	0	0
AE	1	1	63	9	RY	0	2	0	17
AG	1	1	5	13	RV	1	1	8	7
AH	0	1	0	11	NA	3	2	53	22
AI	3	3	13	32	NR	1	1	3	2
AL	4	3	20	119	NN	0	1	0	14
AK	1	1	2	9	ND	2	2	13	79
AM	2	5	8	46	NC	1	2	5	26
AF	1	2	2	19	NQ	0	1	0	4
AP	1	2	79	137	NE	2	1	15	39
AS	2	2	77	27	NG	3	1	6	12
AT	3	3	44	52	NH	1	1	4	1
AW	0	1	0	2	NI	3	1	71	15
AY	0	0	0	0	NL	0	0	0	0
AV	0	0	0	0	NK	0	1	0	7
RA	3	2	19	30	NM	0	1	0	2
RR	1	0	1	0	NF	1	0	1	0
RN	0	1	0	4	NP	1	1	30	45
RD	0	0	0	0	NS	5	2	92	139
RC	0	1	0	4	NT	3	2	18	38

ตารางที่ 5.12 ค่าความถี่ของคู่ลำดับอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยเบาหวาน (ต่อ)

คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV		คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV	
	Positive	Negative	Positive	Negative		Positive	Negative	Positive	Negative
RQ	0	0	0	0	NW	0	0	0	0
RE	5	3	49	122	NY	0	1	0	4
RG	0	0	0	0	NV	0	0	0	0
RH	0	1	0	1	DA	0	1	0	5
RI	2	3	12	75	DR	1	2	46	64
RL	1	0	3	0	DN	1	1	1	1
RK	0	0	0	0	DD	0	0	0	0
DC	0	1	0	5	CT	3	2	11	56
DQ	0	0	0	0	CW	0	0	0	0
DE	3	3	59	21	CY	1	1	7	13
DG	0	0	0	0	CV	0	0	0	0
DH	1	1	46	5	QA	0	0	0	0
DI	2	1	12	29	QR	0	0	0	0
DL	1	1	1	7	QN	0	0	0	0
DK	0	0	0	0	QD	0	0	0	0
DM	0	1	0	2	QC	0	0	0	0
DF	0	0	0	0	QQ	0	0	0	0
DP	0	0	0	0	QE	0	0	0	0
DS	0	1	0	20	QG	0	0	0	0
DT	0	0	0	0	QH	0	0	0	0
DW	0	0	0	0	QI	0	0	0	0
DY	1	2	1	9	QL	0	0	0	0
DV	0	0	0	0	QK	0	0	0	0
CA	4	3	8	97	QM	0	0	0	0
CR	1	3	7	84	QF	0	0	0	0
CN	0	0	0	0	QP	0	1	0	4
CD	0	0	0	0	QS	0	0	0	0
CC	0	1	0	2	QT	0	0	0	0
CQ	0	0	0	0	QW	0	0	0	0
CE	2	2	10	20	QY	0	0	0	0
CG	0	1	0	1	QV	0	0	0	0

ตารางที่ 5.12 ค่าความถี่ของคู่ลำดับอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยเบาหวาน (ต่อ)

คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV		คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV	
	Positive	Negative	Positive	Negative		Positive	Negative	Positive	Negative
CH	2	2	10	40	EA	2	1	9	16
CI	1	1	6	26	ER	3	3	16	23
CL	1	1	9	9	EN	4	2	91	161
CK	0	0	0	0	ED	3	3	13	56
CM	0	0	0	0	EC	2	2	19	12
CF	0	0	0	0	EQ	1	1	1	3
CP	0	0	0	0	EE	2	0	3	0
CS	0	0	0	0	EG	2	2	11	61
EH	1	0	46	0	HA	2	4	11	51
EI	2	3	17	72	HR	2	2	4	7
EL	1	3	9	16	HN	1	0	1	0
EK	0	0	0	0	HD	0	1	0	2
EM	1	3	2	46	HC	0	0	0	0
EF	1	1	32	67	HQ	0	0	0	0
EP	3	2	16	15	HE	1	3	6	57
ES	2	2	12	28	HG	0	1	0	2
ET	2	3	13	60	HH	0	0	0	0
EW	0	0	0	0	HI	2	1	10	7
EY	0	0	0	0	HL	0	1	0	1
EV	1	0	2	0	HK	0	0	0	0
GA	4	5	6	15	HM	0	0	0	0
GR	1	1	3	4	HF	0	0	0	0
GN	1	1	2	1	HP	0	0	0	0
GD	0	0	0	0	HS	0	1	0	1
GC	0	1	0	1	HT	1	1	1	1
GQ	0	0	0	0	HW	0	0	0	0
GE	3	1	50	33	HY	2	1	49	8
GG	0	1	0	1	HV	0	0	0	0
GH	1	1	4	4	IA	2	2	8	18
GI	2	3	121	272	IR	1	1	2	27
GL	2	4	7	85	IN	10	6	52	173

ตารางที่ 5.12 ค่าความถี่ของคู่ลำดับอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยเบาหวาน (ต่อ)

คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV		คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV	
	Positive	Negative	Positive	Negative		Positive	Negative	Positive	Negative
GK	0	0	0	0	ID	2	2	11	18
GM	1	1	2	2	IC	3	3	19	93
GF	1	0	1	0	IQ	2	0	4	0
GP	0	1	0	2	IE	1	1	81	143
GS	0	0	0	0	IG	1	2	5	12
GT	0	0	0	0	IH	0	0	0	0
GW	0	0	0	0	II	1	1	2	1
GY	0	0	0	0	IL	1	2	5	19
GV	0	0	0	0	IK	1	1	3	7
IM	2	1	4	30	KC	0	0	0	0
IF	1	1	2	3	KQ	0	0	0	0
IP	1	1	6	4	KE	1	1	3	7
IS	2	3	14	56	KG	0	0	0	0
IT	2	5	74	79	KH	0	0	0	0
IW	0	0	0	0	KI	1	1	7	2
IY	0	0	0	0	KL	0	0	0	0
IV	2	3	5	5	KK	0	0	0	0
LA	4	2	14	30	KM	0	0	0	0
LR	0	0	0	0	KF	0	0	0	0
LN	0	0	0	0	KP	1	0	1	0
LD	1	0	1	0	KS	0	0	0	0
LC	0	1	0	4	KT	0	0	0	0
LQ	0	0	0	0	KW	0	0	0	0
LE	2	2	10	9	KY	0	0	0	0
LG	0	1	0	2	KV	0	0	0	0
LH	0	0	0	0	MA	3	6	16	68
LI	3	2	21	24	MR	0	0	0	0
LL	2	2	5	6	MN	0	1	0	2
LK	0	0	0	0	MD	0	0	0	0
LM	0	0	0	0	MC	0	0	0	0
LF	1	1	1	8	MQ	0	0	0	0

ตารางที่ 5.12 ค่าความถี่ของคู่ลำดับอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยเบาหวาน (ต่อ)

คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV		คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV	
	Positive	Negative	Positive	Negative		Positive	Negative	Positive	Negative
LP	1	3	6	18	ME	2	2	9	88
LS	0	0	0	0	MG	0	0	0	0
LT	1	3	1	9	MH	0	0	0	0
LW	0	0	0	0	MI	2	3	7	20
LY	1	2	1	10	ML	0	0	0	0
LV	0	1	0	1	MK	0	0	0	0
KA	0	0	0	0	MM	1	5	3	20
KR	0	0	0	0	MF	0	0	0	0
KN	0	1	0	7	MP	1	1	4	2
KD	0	1	0	3	MS	1	0	1	0
MT	0	1	0	1	PH	4	4	12	24
MW	0	0	0	0	PI	1	2	81	137
MY	1	1	2	1	PL	1	1	3	4
MV	0	0	0	0	PK	0	0	0	0
FA	1	2	8	15	PM	0	0	0	0
FR	1	1	2	5	PF	0	0	0	0
FN	0	0	0	0	PP	1	1	2	2
FD	0	0	0	0	PS	1	1	4	2
FC	0	0	0	0	PT	2	1	10	9
FQ	0	0	0	0	PW	0	0	0	0
FE	1	1	2	6	PY	1	1	2	2
FG	0	0	0	0	PV	0	0	0	0
FH	0	0	0	0	SA	1	1	83	136
FI	1	1	4	18	SR	0	0	0	0
FL	0	0	0	0	SN	0	1	0	1
FK	0	0	0	0	SD	0	0	0	0
FM	0	0	0	0	SC	2	2	15	55
FF	0	1	0	1	SQ	0	0	0	0
FP	0	0	0	0	SE	2	2	73	23
FS	0	0	0	0	SG	0	0	0	0
FT	0	0	0	0	SH	0	1	0	3

ตารางที่ 5.12 ค่าความถี่ของคู่ลำดับอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยเบาหวาน (ต่อ)

คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV		คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV	
	Positive	Negative	Positive	Negative		Positive	Negative	Positive	Negative
FW	0	0	0	0	SI	2	3	5	30
FY	0	0	0	0	SL	1	0	2	0
FV	0	0	0	0	SK	0	0	0	0
PA	4	1	24	19	SM	2	0	3	0
PR	3	3	29	89	SF	1	1	2	3
PN	0	0	0	0	SP	3	3	18	25
PD	1	1	2	12	SS	2	1	4	13
PC	0	1	0	2	ST	1	3	12	16
PQ	0	0	0	0	SW	0	0	0	0
PE	3	1	13	19	SY	0	3	0	67
PG	0	0	0	0	SV	0	0	0	0
TA	2	4	18	72	WM	0	0	0	0
TR	2	3	13	26	WF	0	0	0	0
TN	0	0	0	0	WP	0	0	0	0
TD	0	1	0	1	WS	0	0	0	0
TC	0	1	0	3	WT	1	0	1	0
TQ	0	0	0	0	WW	0	0	0	0
TE	4	5	28	107	WY	0	0	0	0
TG	0	0	0	0	WV	0	0	0	0
TH	2	2	8	22	YA	0	0	0	0
TI	4	4	20	100	YR	2	0	2	0
TL	0	0	0	0	YN	1	3	1	67
TK	0	0	0	0	YD	1	0	46	0
TM	0	0	0	0	YC	1	1	3	6
TF	0	0	0	0	YQ	0	0	0	0
TP	1	0	18	0	YE	0	0	0	0
TS	0	0	0	0	YG	0	1	0	2
TT	1	1	10	22	YH	0	0	0	0
TW	0	0	0	0	YI	0	0	0	0
TY	1	1	3	16	YL	0	4	0	7
TV	0	0	0	0	YK	0	0	0	0

ตารางที่ 5.12 ค่าความถี่ของคู่ลำดับอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยเบาหวาน (ต่อ)

คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV		คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV	
	Positive	Negative	Positive	Negative		Positive	Negative	Positive	Negative
WA	0	0	0	0	YM	1	1	1	1
WR	0	0	0	0	YF	0	0	0	0
WN	0	2	0	71	YP	1	1	1	19
WD	0	1	0	1	YS	1	1	4	8
WC	0	0	0	0	YT	1	2	2	9
WQ	0	0	0	0	YW	0	0	0	0
WE	0	0	0	0	YY	0	1	0	1
WG	0	1	0	1	YV	0	0	0	0
WH	0	0	0	0	VA	1	2	2	7
WI	1	0	1	0	VR	0	0	0	0
WL	0	0	0	0	VN	0	0	0	0
WK	0	0	0	0	VD	0	0	0	0
VC	0	0	0	0	VM	0	0	0	0
VQ	0	0	0	0	VF	0	0	0	0
VE	2	1	13	13	VP	0	0	0	0
VG	0	0	0	0	VS	0	0	0	0
VH	0	0	0	0	VT	0	0	0	0
VI	0	1	0	26	VW	0	0	0	0
VL	0	0	0	0	VY	0	0	0	0
VK	0	0	0	0	VV	0	1	0	1

จากตารางที่ 5.12 แสดงค่าความถี่ของคู่ลำดับกรดอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยเบาหวานโดยในกลุ่มสายพันธุกรรมซึ่งแสดงอาการของโรค (Positive Sequence) คู่ลำดับกรดอะมิโน AA มีค่าความถี่สูงสุดเท่ากับ 2 และมีค่าจำนวนครั้งของคู่ลำดับ (FIV) เท่ากับ 9 และในกลุ่มสายพันธุกรรมซึ่งไม่แสดงอาการของโรค (Negative Sequence) มีค่าค่าความถี่สูงสุดเท่ากับ 1 และมีค่าจำนวนครั้งของคู่ลำดับ (FIV) เท่ากับ 136

ตารางที่ 5.13 ค่าสถิติของข้อมูลผู้ป่วยเบาหวานแบบ 2-Grams ที่ไม่ปรับค่าน้ำหนัก (NBL)

ค่าสถิติของชุดข้อมูล	ลักษณะข้อมูล	
	แสดงอาการของโรค	ไม่แสดงอาการของโรค
ค่าความถี่สูงสุดของคู่ลำดับ (Max)	10.00	6.00
ค่าเฉลี่ยของความถี่คู่ลำดับ (Mean)	0.07	0.06
ค่าส่วนเบี่ยงเบนมาตรฐาน (StdDev)	0.22	0.20

จากตารางที่ 5.13 ซึ่งแสดงค่าสถิติต่างๆ ของข้อมูลผู้ป่วยเบาหวานแบบ 2-Grams ที่ไม่ปรับค่าน้ำหนัก (NBL) มีความถี่สูงสุดของคู่ลำดับสำหรับกลุ่มที่แสดงอาการของโรคเท่ากับ 10.00 กลุ่มที่ไม่แสดงอาการของโรคเท่ากับ 6.00 มีค่าความถี่เฉลี่ยของคู่ลำดับกลุ่มที่แสดงอาการของโรคเท่ากับ 0.07 กลุ่มที่ไม่แสดงอาการของโรคเท่ากับ 0.06 มีค่าส่วนเบี่ยงเบนมาตรฐานของคู่ลำดับกลุ่มที่แสดงอาการของโรคเท่ากับ 0.22 กลุ่มที่ไม่แสดงอาการของโรคเท่ากับ 0.20 โดยเมื่อพิจารณาทั้งชุดข้อมูลพบว่าค่าความถี่สูงสุดเท่ากับ 6.00 ค่าความถี่เฉลี่ยเท่ากับ 0.06 และมีค่าส่วนเบี่ยงเบนมาตรฐานเท่ากับ 0.20

ตารางที่ 5.14 ค่าสถิติของข้อมูลผู้ป่วยเบาหวานแบบ 2-Grams ที่ปรับค่าน้ำหนัก (WBL)

ค่าสถิติของชุดข้อมูล	ลักษณะข้อมูล	
	แสดงอาการของโรค	ไม่แสดงอาการของโรค
ค่าความถี่สูงสุดของคู่ลำดับ (Max)	9.09	8.33
ค่าเฉลี่ยของความถี่คู่ลำดับ (Mean)	0.06	0.06
ค่าส่วนเบี่ยงเบนมาตรฐาน (StdDev)	0.21	0.24

จากตารางที่ 5.14 แสดงค่าสถิติของข้อมูลผู้ป่วยเบาหวานแบบ 2-Grams ที่ปรับค่าน้ำหนัก (WBL) มีความถี่สูงสุดของคู่ลำดับสำหรับกลุ่มที่แสดงอาการของโรคเท่ากับ 9.09 กลุ่มที่ไม่แสดงอาการของโรคเท่ากับ 8.33 มีค่าความถี่เฉลี่ยสำหรับกลุ่มที่แสดงอาการของโรคเท่ากับ 0.06 กลุ่มที่ไม่แสดงอาการของโรคเท่ากับ 0.06 มีค่าส่วนเบี่ยงเบนมาตรฐานสำหรับกลุ่มที่แสดงอาการของโรคเท่ากับ 0.21 กลุ่มที่ไม่แสดงอาการของโรคเท่ากับ 0.24 เมื่อพิจารณาทั้งชุดข้อมูล พบว่า ค่าความถี่สูงสุดเท่ากับ 9.09 ค่าความถี่เฉลี่ยเท่ากับ 0.06 และมีค่าส่วนเบี่ยงเบนมาตรฐานเท่ากับ 0.23

5.4.2 ขั้นตอนที่ 2 สกัดลักษณะเฉพาะบนข้อมูลผู้ป่วยเบาหวาน

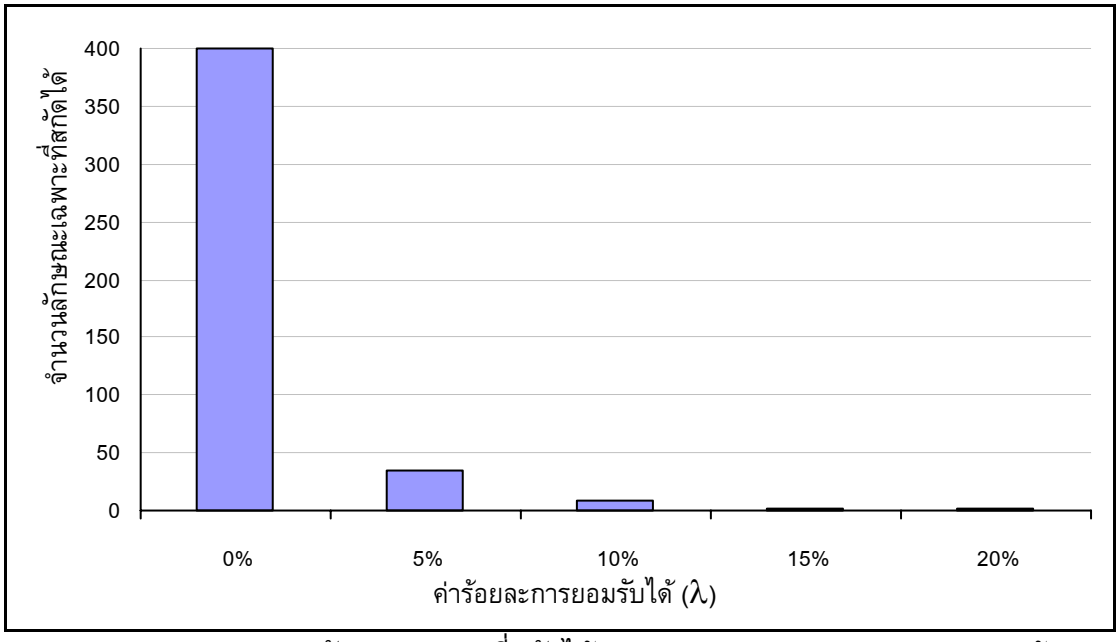
จำนวนลักษณะเฉพาะที่สกัดได้จากการทดลองแบบ C และ D โดยกำหนดค่าร้อยละการยอมรับได้ (λ) ที่แตกต่างกันในขั้นตอนการวิเคราะห์ระดับชุดข้อมูล

(Global Sequences Analyze) ด้วยค่า λ เท่ากับ 0%, 5%, 10%, 15% และ 20% ตามลำดับ สามารถแสดงจำนวนของลักษณะเฉพาะที่สกัดได้ ดังตารางที่ 5.15

ตารางที่ 5.15 จำนวนลักษณะเฉพาะที่สกัดได้จากข้อมูลผู้ป่วยเบาหวาน จากการทดลองแบบ C และ D เมื่อกำหนดค่าร้อยละการยอมรับได้ (λ) ที่แตกต่างกัน

วิธีการทดลอง	จำนวนลักษณะเฉพาะที่สกัดได้				
	λ ≥ 0%	**λ ≥ 5%**	**λ ≥ 10%**	**λ ≥ 15%**	**λ ≥ 20%**
แบบ C และ D	400	34	9	2	1

จากตารางที่ 5.15 พบว่า เมื่อกำหนดค่า λ ≥ 0% ได้จำนวนคู่ลำดับที่ใช้สำหรับใช้เป็นคุณสมบัติของข้อมูลทั้งสิ้น 400 คุณสมบัติ ที่ค่า λ ≥ 5% สามารถสกัดได้ 34 คุณสมบัติ ได้แก่ AA, AN, AL, AP, AS, AT, RE, RI, ND, NI, NS, DR, DE, CA, CR, EN, EI, EF, GE, GI, GL, IN, IC, IE, IT, MA, ME, PR, PI, SA, SE, TA, TE และ TI ที่ค่า λ ≥ 10% สามารถสกัดได้ 9 คุณสมบัติ ได้แก่ AP, RE, NS, EN, GI, IN, IE, PI และ SA ที่ค่า λ ≥ 15% สามารถสกัดได้ 2 คุณสมบัติ ได้แก่ EN และ GI และที่ค่า λ ≥ 20% สามารถสกัดได้ 1 คุณสมบัติ คือ GI สามารถแสดงแผนภูมิเปรียบเทียบจำนวนลักษณะเฉพาะที่สกัดได้ ดังภาพประกอบ 5.21

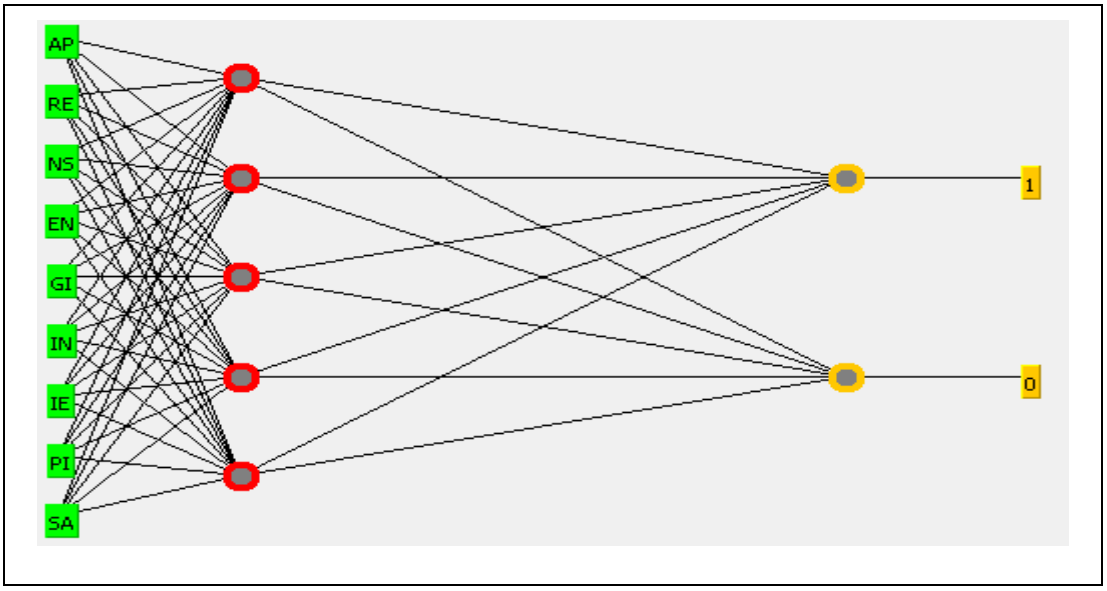


ภาพประกอบ 5.21 จำนวนลักษณะเฉพาะที่สกัดได้จากการทดลองแบบ C และ D บนชุดข้อมูลผู้ป่วยเบาหวาน เมื่อกำหนดค่าร้อยละการยอมรับได้ (λ) ที่แตกต่างกัน

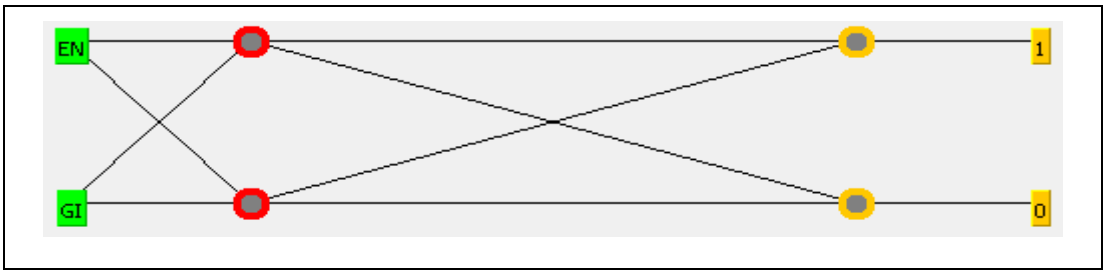
5.4.3 ขั้นตอนที่ 3 การจำแนกข้อมูลผู้ป่วยเบาหวานด้วยลักษณะเฉพาะ

ทำการจำแนกด้วยขั้นตอนวิธี โครงข่ายประสาทแบบ MLP และแบบ RBF และต้นไม้การตัดสินใจแบบ ADTree J48 และ RandomTree ตามลำดับ

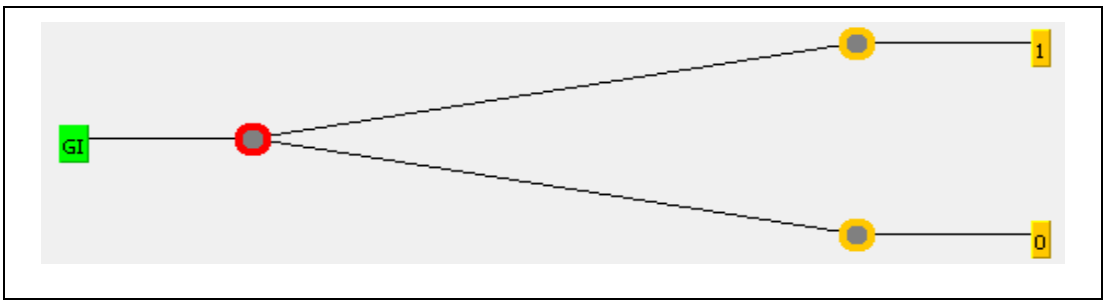
โดยสามารถแสดงลักษณะตัวอย่างโครงข่ายประสาทเทียมแบบ MLP ที่ใช้ในการจำแนกข้อมูลได้ดังภาพประกอบ 5.22 ถึงภาพประกอบ 5.24 ตามลำดับ



ภาพประกอบ 5.22 โครงข่ายประสาทเทียมแบบ MLP สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยเบาหวาน เมื่อกำหนดค่า $\lambda \geq 10\%$ (Input = 9 ตัว)



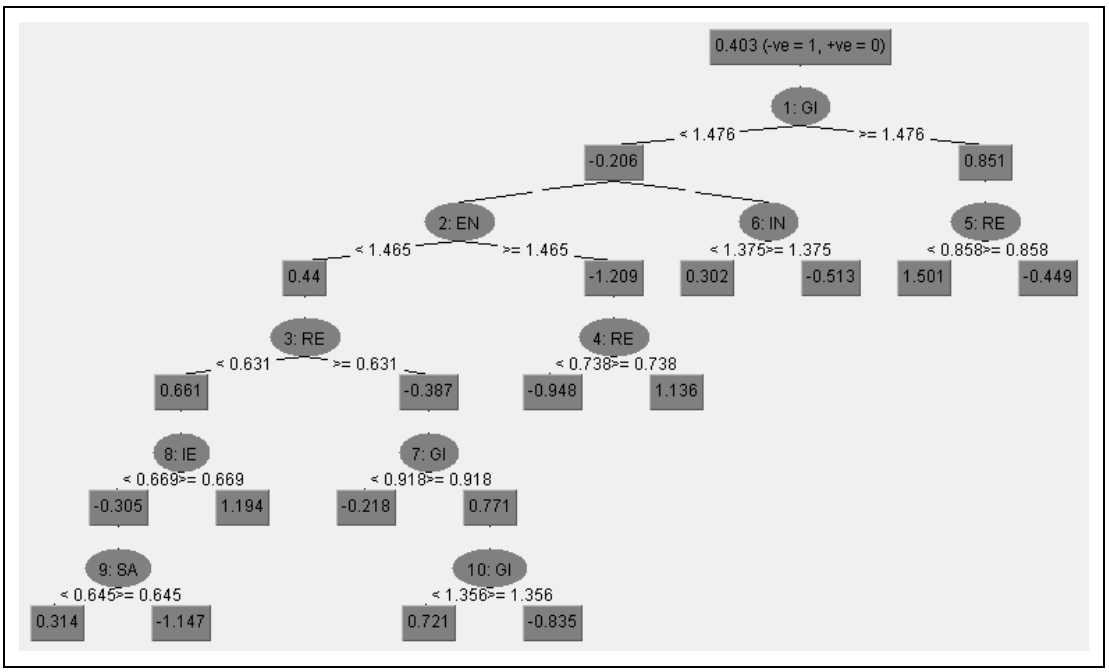
ภาพประกอบ 5.23 โครงข่ายประสาทเทียมแบบ MLP สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยเบาหวาน เมื่อกำหนดค่า $\lambda \geq 15\%$ (Input = 2 ตัว)



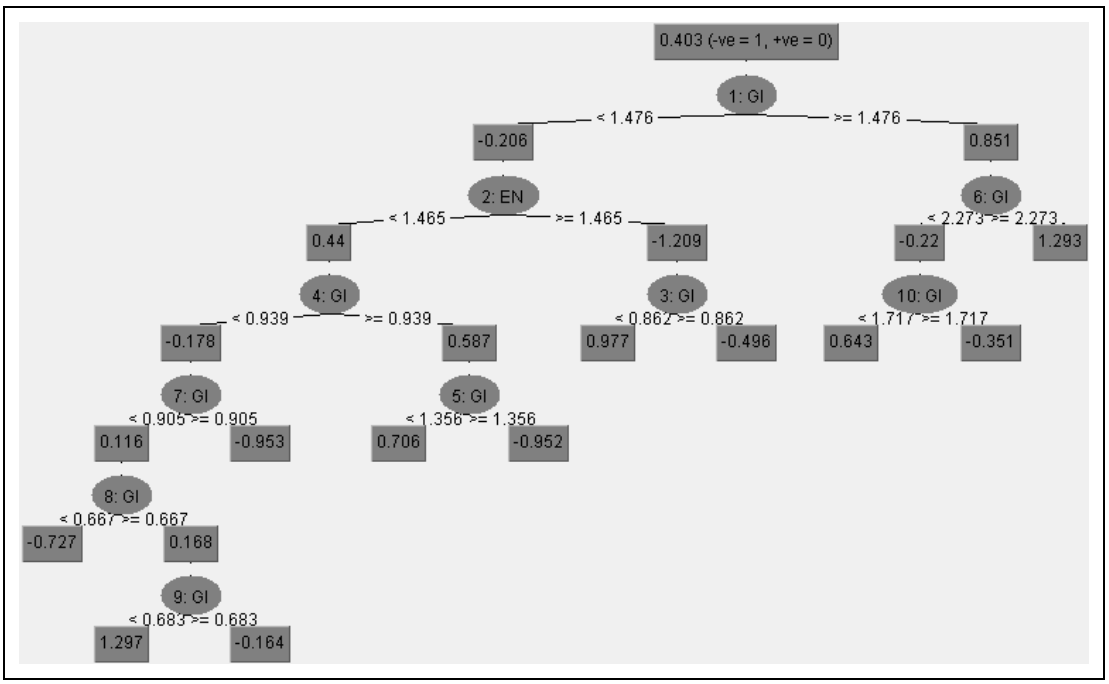
ภาพประกอบ 5.24 โครงข่ายประสาทเทียมแบบ MLP สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยเบาหวาน เมื่อกำหนดค่า $\lambda \geq 20\%$ (Input = 1 ตัว)

จากภาพประกอบ 5.22 ถึงภาพประกอบ 5.24 ลักษณะของโครงข่ายประสาทเทียมแบบหลายชั้น (MLP) คือ ชั้นที่ 1 ชั้นข้อมูลนำเข้าผ่านทางคุณสมบัติของข้อมูลซึ่งสกัดได้ ชั้นที่ 2 เป็นชั้นซ่อน และในชั้นที่ 3 เป็นชั้นผลลัพธ์ ประกอบด้วยผลลัพธ์จำนวน 2 ค่า คือ 1 แทนความหมายว่าแสดงอาการของโรค (Positive) และค่า 0 แทนความหมายว่าไม่แสดงอาการของโรค (Negative)

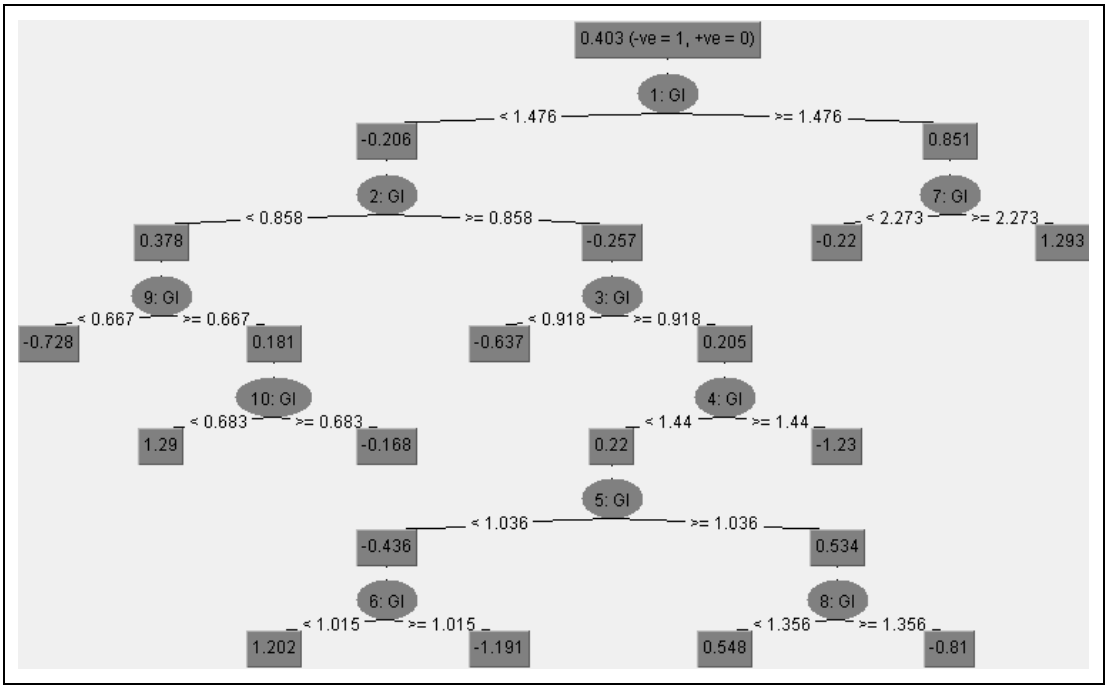
สามารถแสดงตัวอย่างของต้นไม้การตัดสินใจสำหรับการจำแนกข้อมูล ด้วยขั้นตอนวิธีแบบ ADTree ดังภาพประกอบ 5.25 ถึงภาพประกอบ 5.27 ตามลำดับ



ภาพประกอบ 5.25 ต้นไม้การตัดสินใจแบบ ADTree สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยเบาหวาน เมื่อกำหนดค่า $\lambda \geq 10\%$

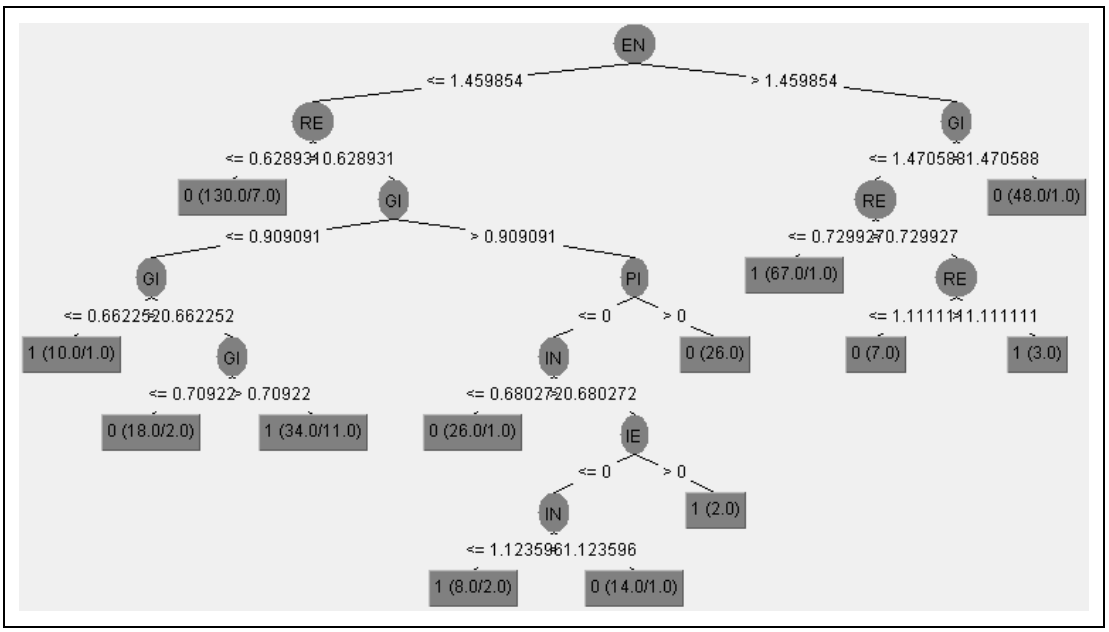


ภาพประกอบ 5.26 ต้นไม้การตัดสินใจแบบ ADTree สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยเบาหวาน เมื่อกำหนดค่า $\lambda \geq 15\%$

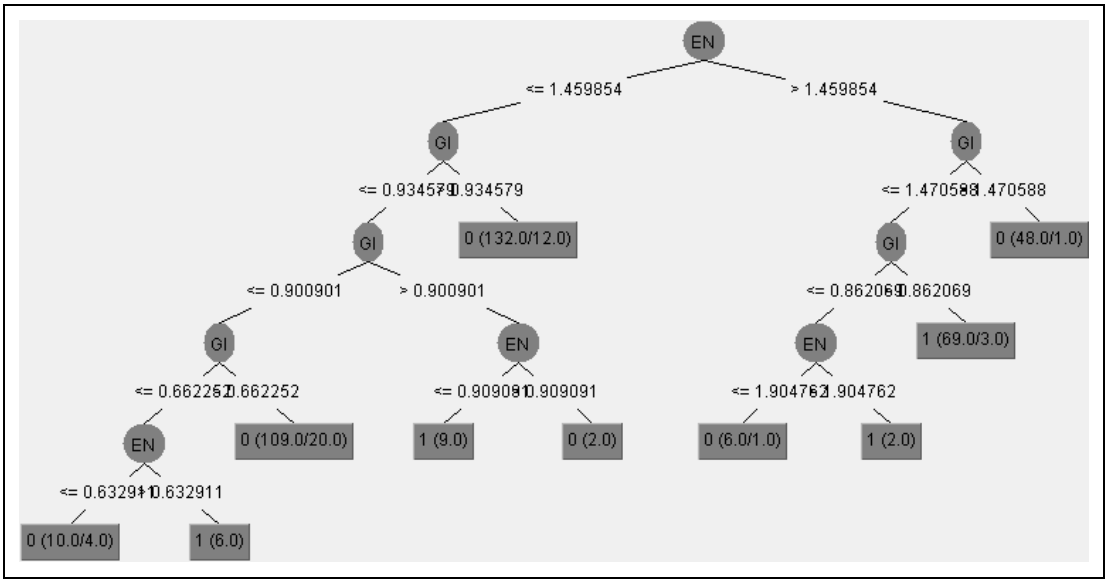


ภาพประกอบ 5.27 ต้นไม้การตัดสินใจแบบ ADTree สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยเบาหวาน เมื่อกำหนดค่า $\lambda \geq 20\%$

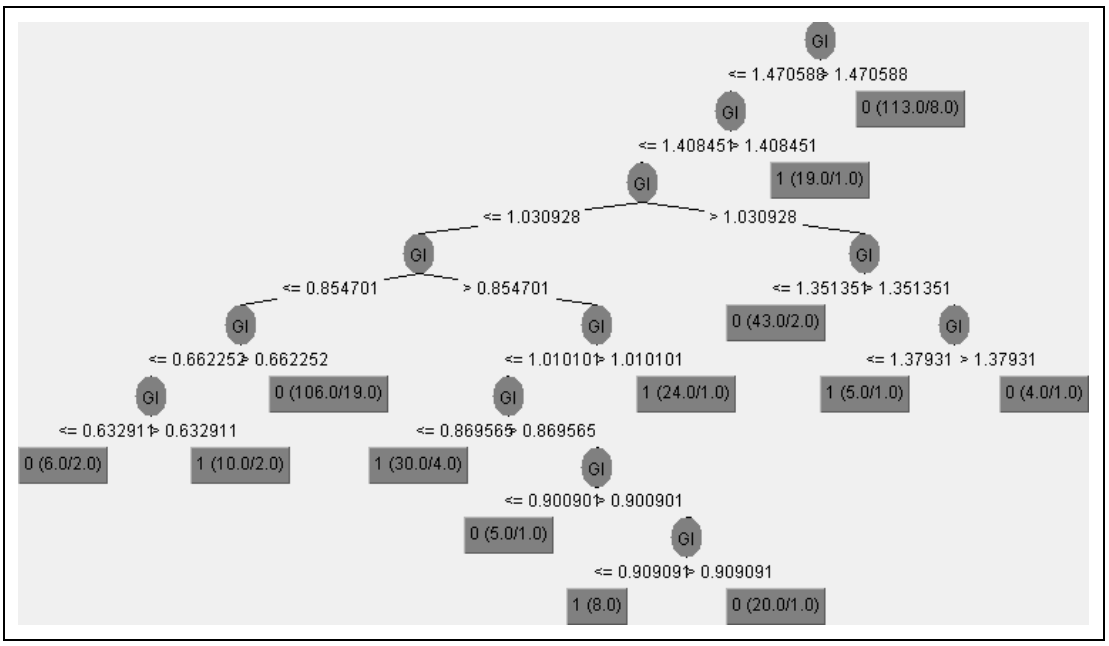
และสามารถแสดงตัวอย่างของต้นไม้การตัดสินใจสำหรับการจำแนกข้อมูลด้วยขั้นตอนวิธีแบบ J48 ดังภาพประกอบ 5.28 ถึงภาพประกอบ 5.30 ตามลำดับ



ภาพประกอบ 5.28 ต้นไม้การตัดสินใจแบบ J48 สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยเบาหวาน เมื่อกำหนดค่า $\lambda \geq 10\%$



ภาพประกอบ 5.29 ต้นไม้การตัดสินใจแบบ J48 สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยเบาหวาน เมื่อกำหนดค่า $\lambda \geq 15\%$



ภาพประกอบ 5.30 ต้นไม้การตัดสินใจแบบ J48 สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยเบาหวาน เมื่อกำหนดค่า $\lambda \geq 20\%$

จากภาพประกอบ 5.25 ถึงภาพประกอบ 5.30 พบว่า ลักษณะต้นไม้การตัดสินใจแบบ ADTree มีความสมดุลของต้นไม้มากกว่าต้นไม้การตัดสินใจแบบ J48 และพบว่า ต้นไม้การตัดสินใจแบบ J48 มีจำนวนชั้นของต้นไม้มากกว่าแบบ ADTree

5.4.4 เปรียบเทียบผลการทดลองข้อมูลผู้ป่วยเบาหวาน

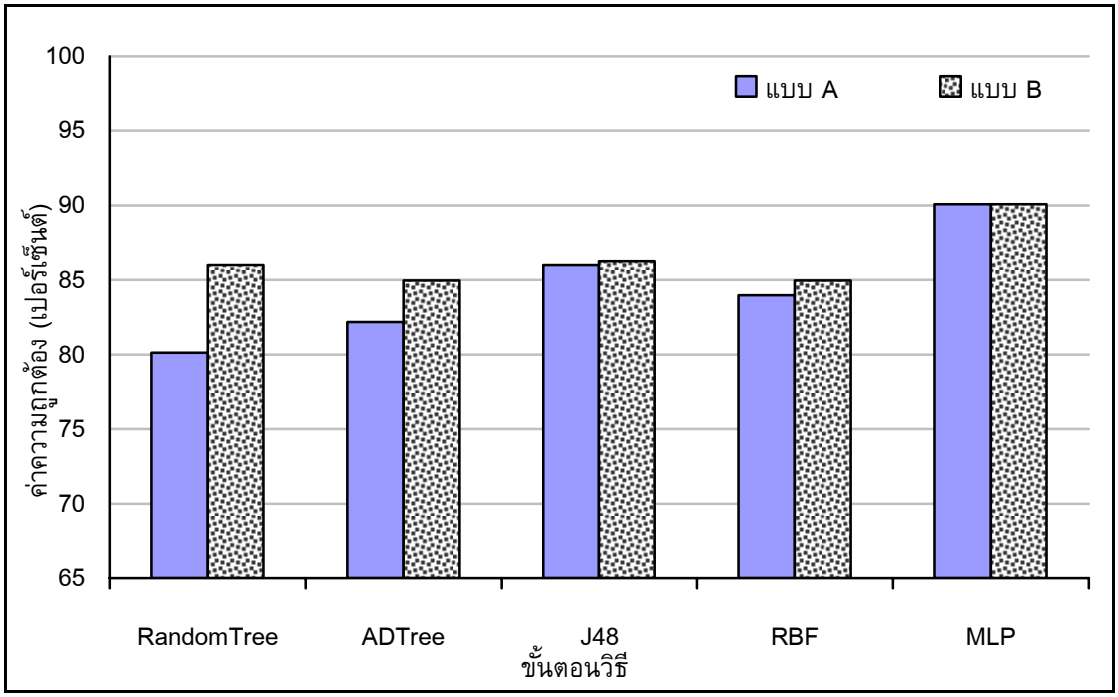
5.4.4.1 เปรียบเทียบค่าความถูกต้องของการจำแนกข้อมูล

ค่าความถูกต้องของการจำแนกข้อมูลตามแบบจำลองการวิเคราะห์รูปและรู้จำลำดับพันธุกรรม สามารถแสดงดังตารางที่ 5.16 และตารางที่ 5.17 ตามลำดับ

ตารางที่ 5.16 ความถูกต้องการจำแนกข้อมูลผู้ป่วยเบาหวานของการทดลองแบบ A และ B

การทดลอง	จำนวนคุณสมบัติ	ค่าความถูกต้อง (เปอร์เซ็นต์)				
		RandomTree	ADTree	J48	RBF	MLP
แบบ A	400	80.13%	82.19%	86.01%	83.97%	90.09%
แบบ B	400	86.01%	84.99%	86.26%	84.99%	90.08%

จากตารางที่ 5.16 สามารถแสดงแผนภูมิเปรียบเทียบค่าความถูกต้องการจำแนกข้อมูลผู้ป่วยเบาหวานด้วยขั้นตอนวิธีแบบต่างๆ ด้วยการทดลองแบบ A และ B แสดงดังภาพประกอบ 5.31



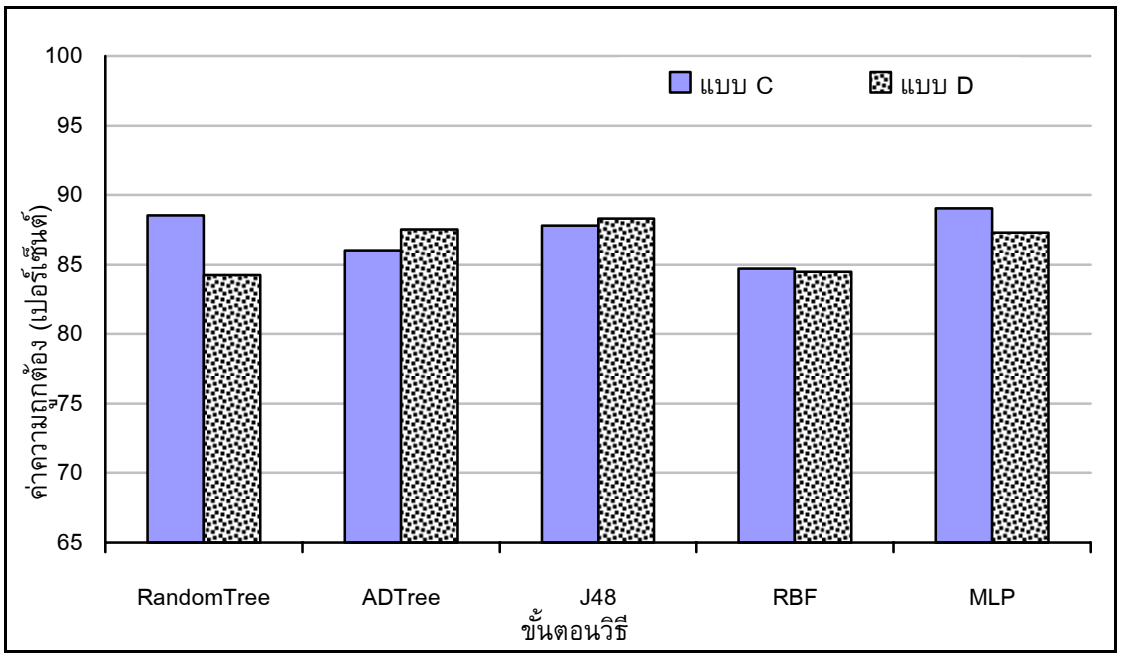
ภาพประกอบ 5.31 ความถูกต้องการจำแนกข้อมูลผู้ป่วยเบาหวานของการทดลองแบบ A และB

จากภาพประกอบ 5.31 พบว่า ในการจำแนกข้อมูลของการทดลองซึ่งมีการปรับค่าน้ำหนักข้อมูล (WBL) จะให้ค่าความถูกต้อง (Accuracy) สูงกว่าการทดลองแบบไม่ปรับค่าน้ำหนักข้อมูล (NBL) โดยขั้นตอนวิธีแบบ MLP ให้ค่าความถูกต้องในการจำแนกสูงที่สุด เท่ากับ 90.09% รองลงมาคือ J48 เท่ากับ 86.26% RandomTree เท่ากับ 86.01% ADTree เท่ากับ 84.99% และ RBF เท่ากับ 84.99% ตามลำดับ และสามารถแสดงค่าความถูกต้องการจำแนกข้อมูลของการทดลองแบบ C และ D เมื่อกำหนดค่าร้อยละการยอมรับได้ (λ) ที่แตกต่างกัน ดังตารางที่ 5.17

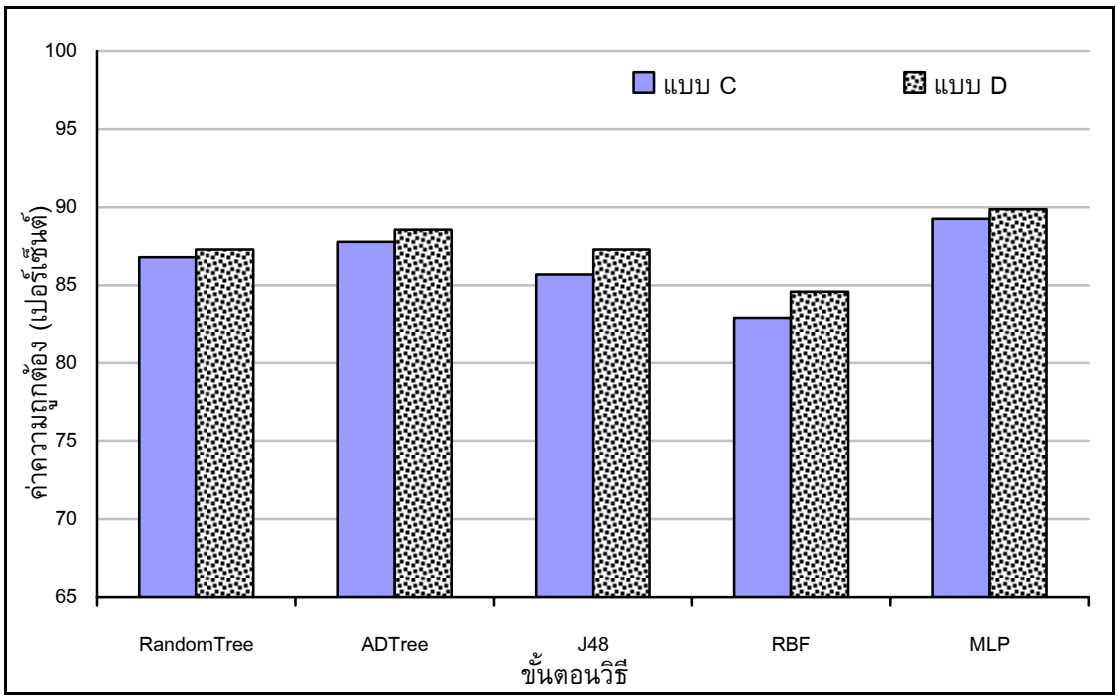
ตารางที่ 5.17 ความถูกต้องการจำแนกข้อมูลผู้ป่วยเบาหวานของการทดลองแบบ C และ D เมื่อกำหนดค่าร้อยละการยอมรับได้ (λ) ที่แตกต่างกัน

วิธีการทดลอง	λ	จำนวนคุณสมบัติ	ค่าความถูกต้อง (เปอร์เซ็นต์)				
			RandomTree	ADTree	J48	RBF	MLP
แบบ C	0%	400	80.13%	82.19%	86.01%	83.97%	90.09%
	5%	34	88.55%	86.01%	87.80%	84.73%	89.06%
	10%	9	86.77%	87.79%	85.70%	82.91%	89.24%
	15%	2	79.64%	79.64%	79.64%	79.64%	79.65%
	20%	1	69.21%	69.21%	69.21%	69.21%	69.21%
แบบ D	0%	400	86.01%	84.99%	86.26%	84.99%	90.08%
	5%	34	84.23%	87.53%	88.30%	84.48%	87.28%
	10%	9	87.28%	88.55%	87.28%	84.54%	89.87%
	15%	2	89.06%	85.50%	85.75%	78.88%	79.13%
	20%	1	85.24%	85.00%	85.50%	69.21%	69.21%

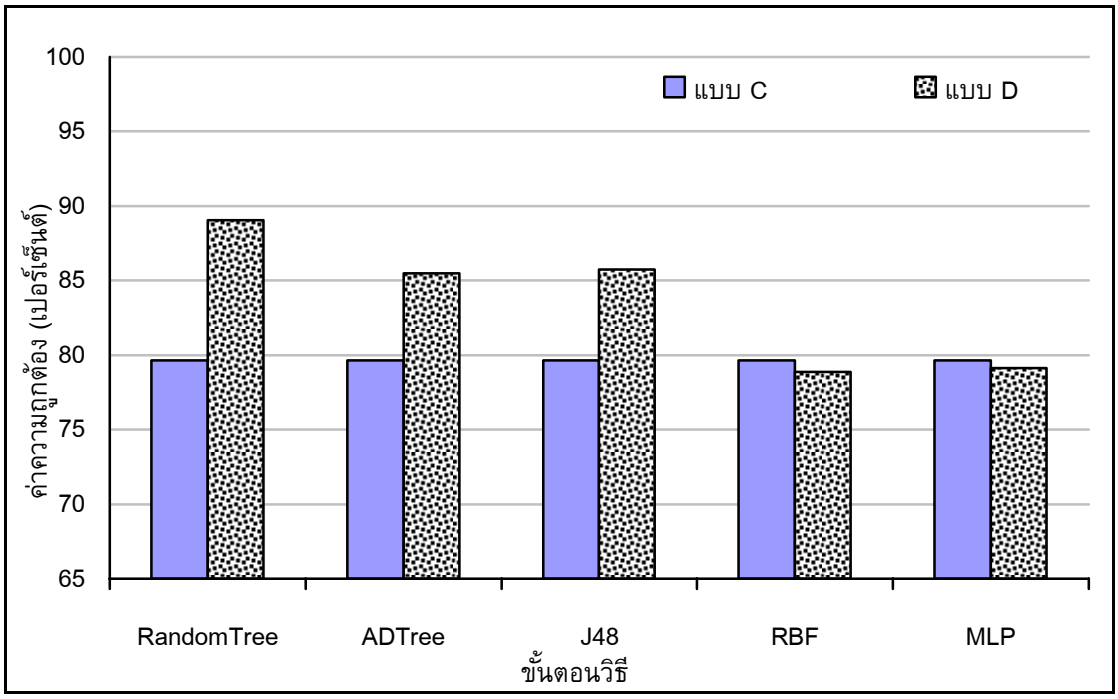
จากตารางที่ 5.17 สามารถแสดงแผนภูมิเปรียบเทียบความถูกต้องการจำแนกข้อมูลสายพันธุกรรมผู้ป่วยเบาหวานของการทดลองแบบ C และ D ดังแสดงในภาพประกอบ 5.32 ถึงภาพประกอบ 5.35 ตามลำดับ



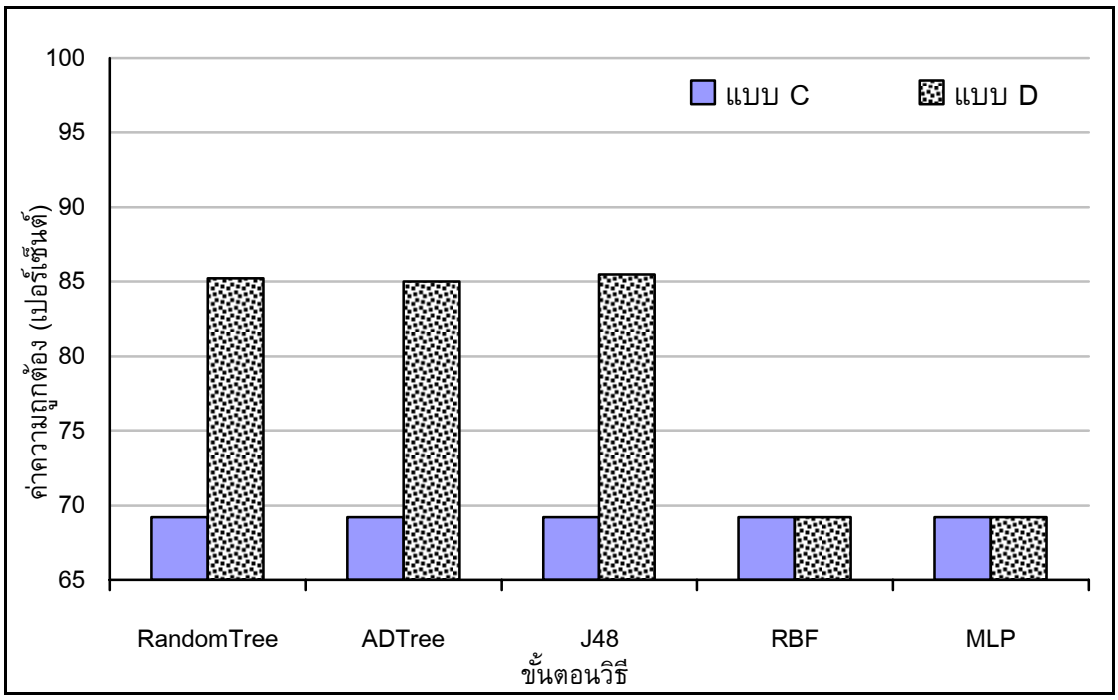
ภาพประกอบ 5.32 ความถูกต้องการจำแนกข้อมูลผู้ป่วยเบาหวานของการทดลองแบบ C และ D เมื่อกำหนดค่า $\lambda \geq 5\%$



ภาพประกอบ 5.33 ความถูกต้องการจำแนกข้อมูลผู้ป่วยเบาหวานของการทดลองแบบ C และ D เมื่อกำหนดค่า $\lambda \geq 10\%$



ภาพประกอบ 5.34 ความถูกต้องการจำแนกข้อมูลผู้ป่วยเบาหวานของการทดลองแบบ C และ D เมื่อกำหนดค่า $\lambda \geq 15\%$



ภาพประกอบ 5.35 ความถูกต้องการจำแนกข้อมูลผู้ป่วยเบาหวานของการทดลองแบบ C และ D เมื่อกำหนดค่า $\lambda \geq 20\%$

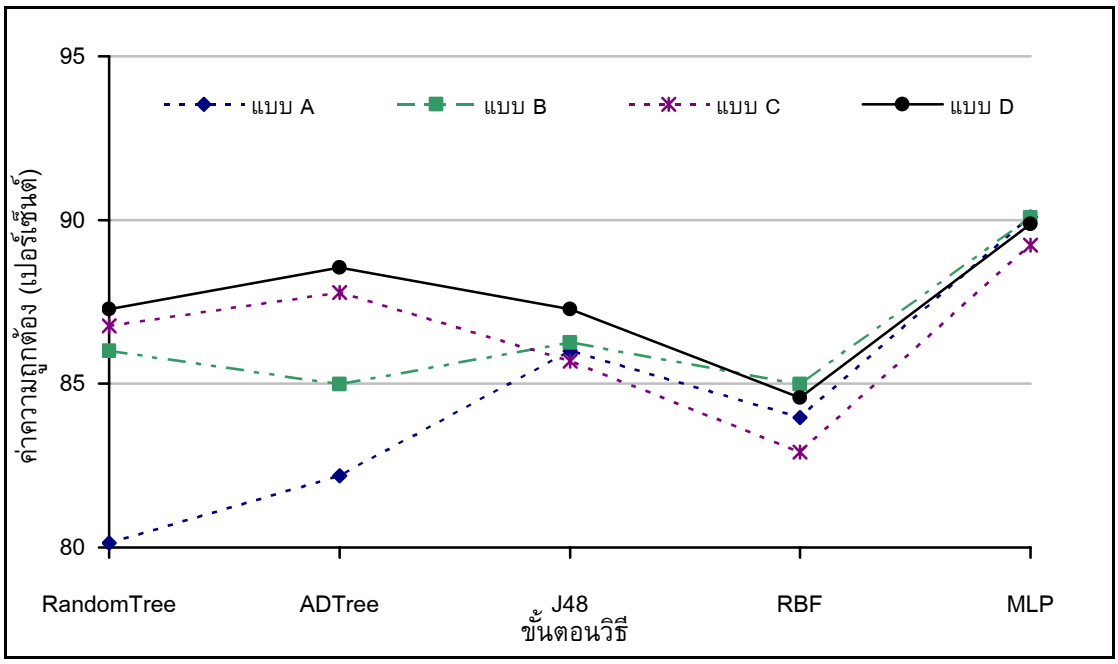
จากภาพประกอบ 5.32 ถึงภาพประกอบ 5.35 พบว่า โดยรวมการทดลองแบบ D ให้ค่าความถูกต้องในการจำแนกข้อมูลสูงกว่าการทดลองแบบ C และจากภาพประกอบ 5.33 ซึ่งมีการกำหนดค่า $\lambda \geq 10\%$ พบว่า มีจำนวนของคุณสมบัติของข้อมูลที่สกัดเพียง 9 คุณสมบัติ และยังคงให้ค่าความถูกต้องที่สูง และเมื่อเปรียบเทียบขั้นตอนวิธีในการจำแนกข้อมูล พบว่า ขั้นตอนวิธีแบบ MLP ให้ค่าความถูกต้องสูงสุดเท่ากับ 89.87% ADTree เท่ากับ 88.55% J48 เท่ากับ 87.28% RandomTree เท่ากับ 84.28% และ RBF เท่ากับ 84.57% ตามลำดับ และจากภาพประกอบ 5.35 พบว่าที่ค่า $\lambda \geq 20\%$ ได้จำนวนคุณสมบัติเพียง 1 คุณสมบัติ และมีค่าความถูกต้องที่ได้จากการจำแนกข้อมูลต่ำ โดยขั้นตอนวิธีแบบ MLP และ RBF ให้ค่าความถูกต้องเท่ากับ 69.21% J48 เท่ากับ 85.50% RandomTree เท่ากับ 85.24% และ ADTree เท่ากับ 85.00% ตามลำดับ

โดยเมื่อนำผลการจำแนกข้อมูลของการทดลองแบบ C และ D ซึ่งกำหนดค่า $\lambda \geq 10\%$ เปรียบเทียบกับการทดลองแบบ A และ B สามารถแสดงรายละเอียดดังตารางที่ 5.18 และภาพประกอบ 5.36 ตามลำดับ

ตารางที่ 5.18 เปรียบเทียบค่าความถูกต้องการจำแนกข้อมูลผู้ป่วยเบาหวาน

การทดลอง	จำนวนคุณสมบัติ	ค่าความถูกต้อง (เปอร์เซ็นต์)				
		RandomTree	ADTree	J48	RBF	MLP
แบบ A	400	80.13%	82.19%	86.01%	83.97%	90.09%
แบบ B	400	86.01%	84.99%	86.26%	84.99%	90.08%
แบบ C	9	86.77%	87.79%	85.70%	82.91%	89.24%
แบบ D	9	87.28%	88.55%	87.28%	84.54%	89.87%

*หมายเหตุ การทดลองแบบ C และ D กำหนดค่าร้อยละการยอมรับได้เท่ากับ 10 ($\lambda \geq 10\%$)



ภาพประกอบ 5.36 เปรียบเทียบค่าความถูกต้องการจำแนกข้อมูลผู้ป่วยเบาหวาน

5.4.4.2 เวลาที่ใช้ในการจำแนกข้อมูล

สามารถแสดงเวลาที่ใช้ในการจำแนกข้อมูลผู้ป่วยเบาหวานของการทดลองแบบ A และ B ดังตารางที่ 5.19 และการทดลองแบบ C และ D ดังตารางที่ 5.20 ตามลำดับ

ตารางที่ 5.19 เวลาที่ใช้ในการจำแนกข้อมูลผู้ป่วยเบาหวานของการทดลองแบบ A และ B

การทดลอง	จำนวนคุณสมบัติ	เวลาที่ใช้ (วินาที)				
		RandomTree	ADTree	J48	RBF	MLP
แบบ A	400	0.66	11.92	1.33	1.64	56.13
แบบ B	400	0.77	5.83	2.47	1.52	87.43

ตารางที่ 5.20 เวลาที่ใช้ในการจำแนกข้อมูลผู้ป่วยเบาหวาน ของการทดลองแบบ C และ D เมื่อกำหนดค่าร้อยละการยอมรับได้ (λ) ที่แตกต่างกัน

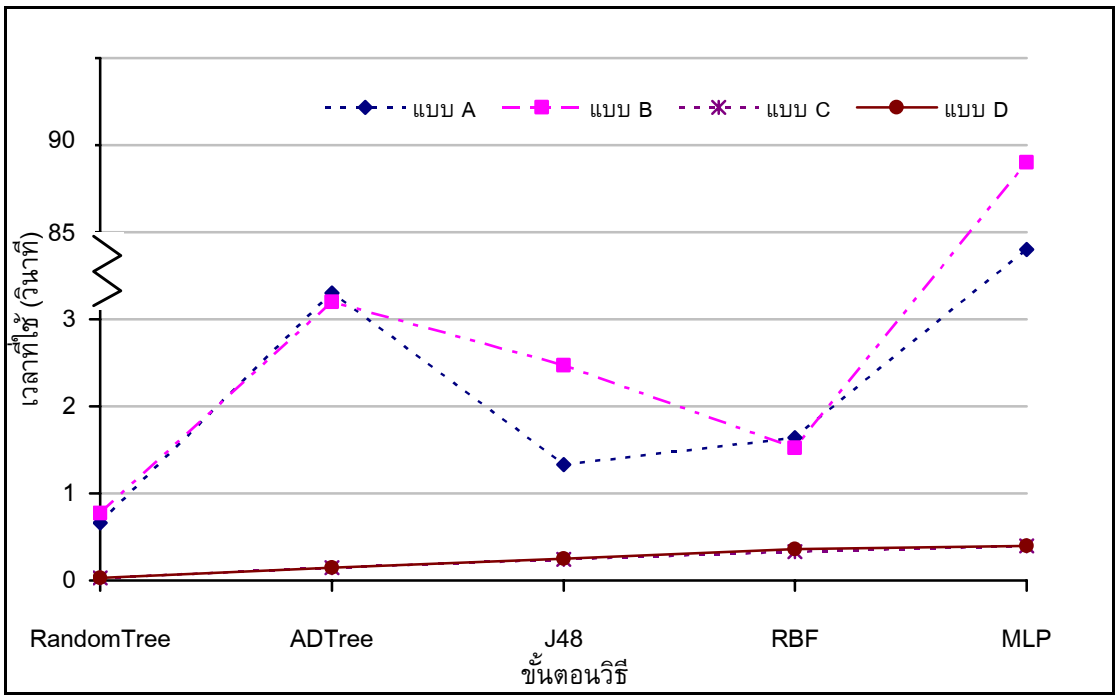
การทดลอง	λ	จำนวนคุณสมบัติ	เวลาที่ใช้ (วินาที)				
			RandomTree	ADTree	J48	RBF	MLP
แบบ C	0%	400	0.66	11.92	1.33	1.63	56.13
	5%	34	0.08	0.49	0.26	0.53	3.24
	10%	9	0.03	0.15	0.24	0.33	0.40
	15%	2	0.04	0.05	0.17	0.32	0.19
	20%	1	0.02	0.04	0.03	0.29	0.13
แบบ D	0%	400	0.74	5.36	2.79	2.33	87.43
	5%	34	0.11	0.33	0.28	0.58	2.33
	10%	9	0.03	0.15	0.25	0.36	0.40
	15%	2	0.09	0.05	0.08	0.21	0.19
	20%	1	0.04	0.05	0.19	0.52	0.15

จากตารางที่ 5.19 และตารางที่ 5.20 พบว่า การทดลองแบบ A และ B ใช้เวลาในการจำแนกสูงกว่าการทดลองแบบ C และ D และขั้นตอนวิธีแบบ MLP ใช้เวลาในการจำแนกมากที่สุด และขั้นตอนวิธีแบบ RandomTree ใช้เวลาในการจำแนกน้อยที่สุด ตารางที่ 5.21 และภาพประกอบ 5.37 แสดงการเปรียบเทียบเวลาที่ใช้ในการจำแนกข้อมูลของการทดลองแบบ A และ B กับการทดลองแบบ C และ D เมื่อกำหนดค่า $\lambda \geq 10\%$ สามารถแสดงรายละเอียดดังนี้

ตารางที่ 5.21 เปรียบเทียบเวลาที่ใช้จำแนกข้อมูลผู้ป่วยเบาหวาน

การทดลอง	จำนวนคุณสมบัติ	เวลาที่ใช้ (วินาที)				
		RandomTree	ADTree	J48	RBF	MLP
แบบ A	400	0.66	11.92	1.33	1.64	56.13
แบบ B	400	0.77	5.83	2.47	1.52	87.43
แบบ C	9	0.03	0.15	0.24	0.33	0.40
แบบ D	9	0.03	0.15	0.25	0.36	0.40

*หมายเหตุ การทดลองแบบ C และ D กำหนดค่าร้อยละการยอมรับได้เท่ากับ 10 ($\lambda \geq 10\%$)



ภาพประกอบ 5.37 เปรียบเทียบเวลาที่ใช้บนชุดข้อมูลผู้ป่วยเบาหวาน

5.4.4.3 การวิจารณ์ผลทดลองข้อมูลผู้ป่วยเบาหวาน

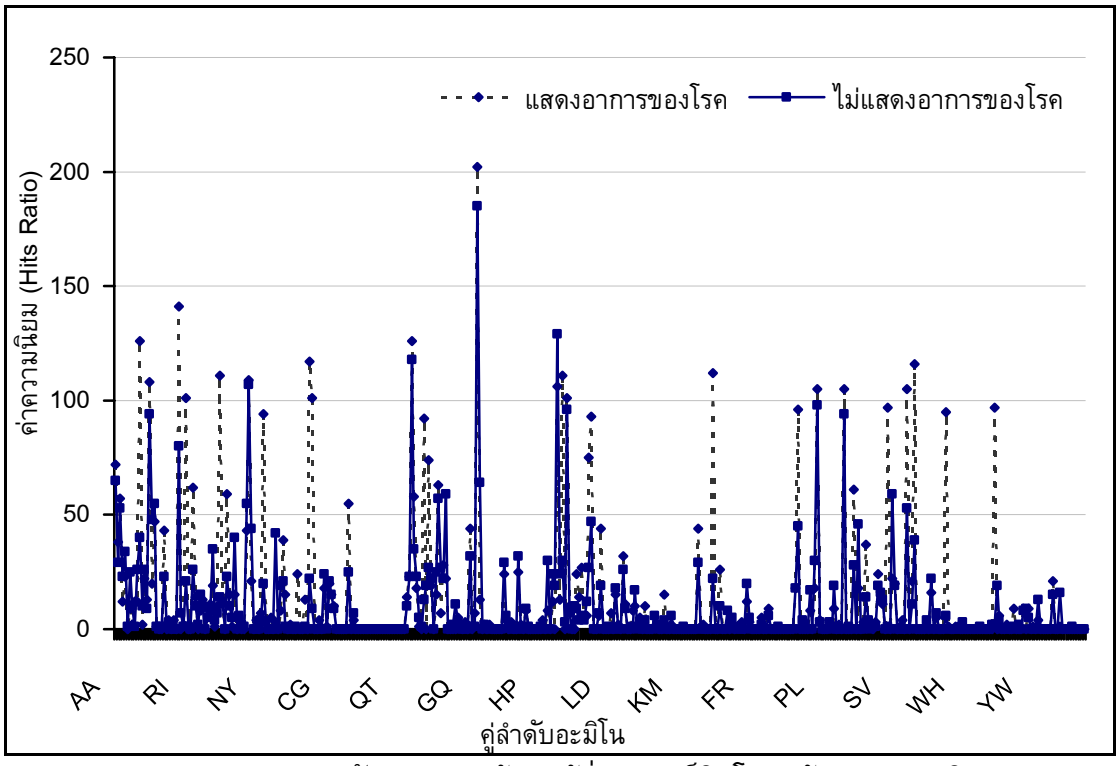
จากการทดลองแบบ A, B, C และ D (ตารางที่ 5.1) พบว่าการจำแนกข้อมูลผู้ป่วยเบาหวาน โดยการทดลองซึ่งมีการปรับค่าน้ำหนักข้อมูล (WBL) ให้ค่าความถูกต้องสูงกว่าการทดลองแบบไม่ปรับค่าน้ำหนัก (NBL) เมื่อพิจารณาขั้นตอนวิธีที่ใช้ พบว่า ขั้นตอนวิธีแบบ MLP ให้ค่าความถูกต้องในการจำแนกสูงที่สุด รองลงมาคือ ADTree J48 RBF และ RandomTree ตามลำดับ และพบว่า ขั้นตอนวิธีแบบ RandomTree ใช้เวลาน้อยที่สุด และขั้นตอนวิธีแบบ MLP ใช้เวลามากที่สุด และสังเกตุว่าในขั้นตอนการสกัดลักษณะเฉพาะของการทดลองแบบ C และ D สามารถลดจำนวนคุณสมบัติของข้อมูลที่ใช้ในการจำแนกลงได้ โดยที่ค่า $\lambda \geq 5\%$ สามารถสกัดได้ 34 คุณสมบัติ ที่ค่า $\lambda \geq 10\%$ สกัดได้ 9 คุณสมบัติ ที่ค่า $\lambda \geq 15\%$ สกัดได้ 2 คุณสมบัติ และที่ค่า $\lambda \geq 20\%$ สกัดได้ 1 คุณสมบัติ และพบว่าที่ค่า $\lambda \geq 10\%$ ให้ค่าความถูกต้องการจำแนกข้อมูลสูงที่สุด และที่ค่า $\lambda \geq 20\%$ ให้ค่าความถูกต้องการจำแนกข้อมูลต่ำที่สุด

5.5 ข้อมูลผู้ป่วยดาวน์ซินโดรม (Downsyndrome Disease)

ข้อมูลผู้ป่วยดาวน์ซินโดรมประกอบด้วยสายโปรตีนทั้งสิ้น 387 สาย โดยมีความยาวของสายที่ใช้อยู่ระหว่าง 30 (±5%) ตัวอักษร ถึง 187.50 (±5%) ตัวอักษร แบ่งเป็นข้อมูลซึ่งแสดงอาการของโรคจำนวน 202 สาย และไม่แสดงอาการของโรคจำนวน 185 สาย ประกอบด้วยขั้นตอนในการทดลอง 3 ขั้นตอน คือ การเตรียมข้อมูลค่าความถี่แบบปรับน้ำหนักการสกัดลักษณะเฉพาะ และการจำแนกสายพันธุกรรม โดยสามารถอธิบายรายละเอียดการทดลองตามลำดับ ดังนี้

5.5.1 ขั้นตอนที่ 1 การเตรียมข้อมูลผู้ป่วยดาวน์ซินโดรมแบบปรับน้ำหนัก

ค่าความถี่ของคู่ลำดับกรดอะมิโนซึ่งนับได้จากข้อมูลผู้ป่วยดาวน์ซินโดรมสามารถแสดงลักษณะของข้อมูลด้วยค่าความนิยม (Hits Ratio) ดังภาพประกอบ 5.38



ภาพประกอบ 5.38 ลักษณะของข้อมูลผู้ป่วยดาวน์ซินโดรมด้วยค่าความนิยม

จากภาพประกอบ 5.38 พบว่า ในกลุ่มข้อมูลซึ่งแสดงอาการของโรค คู่ลำดับกรดอะมิโน GI มีค่าความนิยมสูงสุดเท่ากับ 202 และกลุ่มข้อมูลซึ่งไม่แสดงอาการของโรค มีค่าความนิยมสูงสุด 185 สำหรับ และสามารถแสดงค่าความถี่ของคู่ลำดับกรดอะมิโนแบบ 2-Grams ทั้ง 400 คู่ลำดับ ของข้อมูลผู้ป่วยดาวน์ซินโดรม ได้ดังตารางที่ 5.22

ตารางที่ 5.22 ค่าความถี่ของคู่ลำดับอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยดาวน์ซินโดรม

คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV		คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV	
	Positive	Negative	Positive	Negative		Positive	Negative	Positive	Negative
AA	1	2	72	65	RM	2	2	62	26
AR	2	3	38	29	RF	1	1	14	10
AN	5	5	57	53	RP	0	1	0	2
AD	1	1	12	23	RS	1	1	8	15
AC	2	3	23	34	RT	1	3	13	10
AQ	0	1	0	1	RW	0	0	0	0
AE	1	1	9	25	RY	1	2	9	5
AG	1	6	2	10	RV	1	1	11	8
AH	1	1	12	1	NA	1	2	19	35
AI	3	3	12	26	NR	1	1	1	4
AL	2	4	126	40	NN	1	1	10	7
AK	1	1	2	10	ND	2	2	111	14
AM	2	5	22	26	NC	2	2	11	12
AF	2	2	13	9	NQ	0	0	0	0
AP	2	2	108	94	NE	5	4	59	23
AS	2	2	20	48	NG	3	3	15	10
AT	6	3	47	55	NH	1	1	1	5
AW	1	1	1	1	NI	2	3	15	40
AY	0	0	0	0	NL	0	0	0	0
AV	0	1	0	1	NK	1	1	1	6
RA	3	3	43	23	NM	0	1	0	2
RR	0	1	0	1	NF	0	1	0	1
RN	1	1	4	3	NP	1	1	43	55
RD	0	0	0	0	NS	2	5	109	107
RC	1	0	4	0	NT	2	4	21	44
RQ	0	0	0	0	NW	0	0	0	0
RE	9	4	141	80	NY	1	0	4	0
RG	0	1	0	7	NV	0	0	0	0
RH	1	1	2	1	DA	1	1	7	4
RI	3	2	101	21	DR	2	1	94	20
RL	0	1	0	1	DN	0	1	0	3

ตารางที่ 5.22 ค่าความถี่ของคู่ลำดับอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยดาวน์ซินโดรม (ต่อ)

คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV		คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV	
	Positive	Negative	Positive	Negative		Positive	Negative	Positive	Negative
RK	0	0	0	0	DD	0	0	0	0
DC	1	0	5	0	CT	1	3	55	25
DQ	0	0	0	0	CW	0	0	0	0
DE	1	5	3	42	CY	1	1	4	7
DG	0	0	0	0	CV	0	0	0	0
DH	1	1	8	19	QA	0	0	0	0
DI	2	3	39	21	QR	0	0	0	0
DL	1	1	15	1	QN	0	0	0	0
DK	0	0	0	0	QD	0	0	0	0
DM	1	0	2	0	QC	0	0	0	0
DF	0	0	0	0	QQ	0	0	0	0
DP	0	1	0	1	QE	0	0	0	0
DS	1	0	24	0	QG	0	0	0	0
DT	0	0	0	0	QH	0	0	0	0
DW	0	0	0	0	QI	0	0	0	0
DY	2	2	13	1	QL	0	0	0	0
DV	0	0	0	0	QK	0	0	0	0
CA	3	6	117	22	QM	0	0	0	0
CR	3	1	101	9	QF	0	0	0	0
CN	0	0	0	0	QP	0	0	0	0
CD	0	0	0	0	QS	0	0	0	0
CC	1	0	4	0	QT	0	0	0	0
CQ	0	0	0	0	QW	0	0	0	0
CE	3	2	18	24	QY	0	0	0	0
CG	1	0	1	0	QV	0	0	0	0
CH	2	2	20	21	EA	3	2	14	10
CI	2	2	15	15	ER	4	3	23	23
CL	1	1	10	9	EN	3	4	126	118
CK	0	0	0	0	ED	3	4	58	35
CM	0	0	0	0	EC	4	2	18	23
CF	0	0	0	0	EQ	1	1	1	5

ตารางที่ 5.22 ค่าความถี่ของคู่ลำดับอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยดาวน์ซินโดรม (ต่อ)

คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV		คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV	
	Positive	Negative	Positive	Negative		Positive	Negative	Positive	Negative
CP	0	0	0	0	EE	0	1	0	1
CS	0	0	0	0	EG	5	2	92	13
EH	0	1	0	19	HA	3	2	24	29
EI	4	2	74	27	HR	1	2	4	6
EL	2	1	19	22	HN	0	1	0	1
EK	0	0	0	0	HD	1	0	3	0
EM	1	3	15	25	HC	0	0	0	0
EF	1	1	63	57	HQ	0	0	0	0
EP	1	5	7	27	HE	1	3	25	32
ES	1	2	28	22	HG	1	0	2	0
ET	2	2	22	59	HH	0	0	0	0
EW	0	0	0	0	HI	1	2	8	9
EY	0	0	0	0	HL	0	1	0	1
EV	0	1	0	2	HK	0	0	0	0
GA	0	5	0	11	HM	0	0	0	0
GR	1	1	5	3	HF	0	0	0	0
GN	1	0	1	0	HP	0	0	0	0
GD	0	0	0	0	HS	0	1	0	1
GC	1	0	3	0	HT	1	0	4	0
GQ	0	0	0	0	HW	0	0	0	0
GE	4	3	44	32	HY	1	3	8	30
GG	0	1	0	1	HV	0	0	0	0
GH	1	1	7	1	IA	1	2	12	24
GI	3	2	202	185	IR	0	1	0	19
GL	2	13	13	64	IN	7	10	106	129
GK	0	0	0	0	ID	2	5	13	24
GM	1	1	2	2	IC	3	3	111	30
GF	0	1	0	1	IQ	0	2	0	3
GP	1	0	2	0	IE	1	1	101	96
GS	0	0	0	0	IG	2	1	10	7
GT	0	0	0	0	IH	0	0	0	0

ตารางที่ 5.22 ค่าความถี่ของคู่ลำดับอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยดาวน์ซินโดรม (ต่อ)

คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV		คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV	
	Positive	Negative	Positive	Negative		Positive	Negative	Positive	Negative
GW	0	0	0	0	II	0	2	0	10
GY	0	0	0	0	IL	2	2	24	9
GV	0	0	0	0	IK	1	4	14	4
IM	1	1	27	6	KC	0	0	0	0
IF	2	1	5	4	KQ	0	0	0	0
IP	1	1	6	12	KE	1	4	15	4
IS	3	2	75	27	KG	0	0	0	0
IT	5	2	93	47	KH	0	0	0	0
IW	0	0	0	0	KI	1	1	2	6
IY	0	0	0	0	KL	0	0	0	0
IV	3	1	6	7	KK	0	0	0	0
LA	2	4	44	19	KM	0	0	0	0
LR	0	0	0	0	KF	0	0	0	0
LN	0	0	0	0	KP	0	1	0	1
LD	0	1	0	1	KS	0	0	0	0
LC	1	0	7	0	KT	0	0	0	0
LQ	0	0	0	0	KW	0	0	0	0
LE	3	2	15	18	KY	0	0	0	0
LG	1	0	2	0	KV	0	0	0	0
LH	0	0	0	0	MA	4	6	44	29
LI	2	7	32	26	MR	1	0	1	0
LL	2	2	11	9	MN	1	0	2	0
LK	0	0	0	0	MD	0	0	0	0
LM	0	0	0	0	MC	0	0	0	0
LF	1	1	8	1	MQ	0	0	0	0
LP	2	4	10	17	ME	2	1	112	22
LS	1	0	1	0	MG	0	0	0	0
LT	1	1	5	3	MH	0	0	0	0
LW	0	0	0	0	MI	3	2	26	10
LY	1	2	10	4	ML	0	0	0	0
LV	0	0	0	0	MK	0	0	0	0

ตารางที่ 5.22 ค่าความถี่ของคู่ลำดับอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยดาวน์ซินโดรม (ต่อ)

คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV		คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV	
	Positive	Negative	Positive	Negative		Positive	Negative	Positive	Negative
KA	0	0	0	0	MM	1	5	1	8
KR	0	0	0	0	MF	0	0	0	0
KN	1	1	1	6	MP	1	1	1	5
KD	1	1	1	2	MS	0	1	0	1
MT	1	0	1	0	PH	4	4	18	30
MW	1	0	1	0	PI	2	2	105	98
MY	1	0	3	0	PL	1	1	4	3
MV	0	0	0	0	PK	0	0	0	0
FA	2	2	12	20	PM	0	0	0	0
FR	1	1	3	5	PF	0	0	0	0
FN	0	0	0	0	PP	1	1	3	3
FD	0	0	0	0	PS	1	0	2	0
FC	0	0	0	0	PT	1	5	9	19
FQ	0	0	0	0	PW	0	0	0	0
FE	1	1	5	2	PY	1	0	2	0
FG	0	0	0	0	PV	0	0	0	0
FH	0	0	0	0	SA	1	1	105	94
FI	2	1	9	6	SR	0	0	0	0
FL	0	0	0	0	SN	0	1	0	1
FK	0	0	0	0	SD	0	0	0	0
FM	0	0	0	0	SC	2	2	61	28
FF	0	1	0	1	SQ	0	0	0	0
FP	0	0	0	0	SE	2	2	18	46
FS	0	0	0	0	SG	0	0	0	0
FT	0	0	0	0	SH	1	0	3	0
FW	0	0	0	0	SI	3	5	37	14
FY	0	0	0	0	SL	2	1	1	4
FV	0	0	0	0	SK	1	0	1	0
PA	2	4	18	18	SM	0	2	0	2
PR	5	4	96	45	SF	1	1	3	1
PN	0	0	0	0	SP	3	3	24	19

ตารางที่ 5.22 ค่าความถี่ของคู่ลำดับอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยดาวน์ซินโดรม (ต่อ)

คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV		คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV	
	Positive	Negative	Positive	Negative		Positive	Negative	Positive	Negative
PD	0	1	0	4	SS	2	2	12	16
PC	0	1	0	2	ST	2	1	11	13
PQ	0	0	0	0	SW	0	0	0	0
PE	2	5	8	17	SY	3	1	97	1
PG	0	0	0	0	SV	0	0	0	0
TA	2	2	22	59	WM	0	0	0	0
TR	2	3	20	19	WF	0	0	0	0
TN	0	0	0	0	WP	0	0	0	0
TD	1	0	1	0	WS	0	0	0	0
TC	1	0	4	0	WT	0	1	0	1
TQ	0	0	0	0	WW	0	0	0	0
TE	5	4	105	53	WY	0	0	0	0
TG	0	0	0	0	WV	0	0	0	0
TH	3	2	21	11	YA	0	0	0	0
TI	5	7	116	39	YR	0	2	0	2
TL	0	0	0	0	YN	3	1	97	2
TK	0	0	0	0	YD	0	1	0	19
TM	0	0	0	0	YC	1	1	6	4
TF	0	0	0	0	YQ	0	0	0	0
TP	1	1	1	4	YE	0	0	0	0
TS	0	0	0	0	YG	1	0	2	0
TT	1	1	16	22	YH	0	0	0	0
TW	0	0	0	0	YI	0	0	0	0
TY	1	1	5	7	YL	4	1	9	1
TV	0	0	0	0	YK	0	0	0	0
WA	0	0	0	0	YM	0	1	0	2
WR	0	0	0	0	YF	0	0	0	0
WN	2	1	95	6	YP	1	1	9	8
WD	1	0	1	0	YS	0	3	0	9
WC	0	0	0	0	YT	1	2	9	5
WQ	0	0	0	0	YW	0	0	0	0

ตารางที่ 5.22 ค่าความถี่ของคู่ลำดับอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยดาวน์ซินโดรม (ต่อ)

คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV		คู่ลำดับ	ค่าความถี่สูงสุด		ค่า FIV	
	Positive	Negative	Positive	Negative		Positive	Negative	Positive	Negative
WE	1	0	1	0	YY	1	0	1	0
WG	1	0	1	0	YV	0	0	0	0
WH	0	0	0	0	VA	2	2	4	13
WI	0	1	0	3	VR	0	0	0	0
WL	0	0	0	0	VN	0	0	0	0
WK	0	0	0	0	VD	0	0	0	0
VC	0	0	0	0	VM	0	0	0	0
VQ	0	0	0	0	VF	0	0	0	0
VE	2	1	21	15	VP	0	1	0	1
VG	0	0	0	0	VS	0	0	0	0
VH	0	0	0	0	VT	0	0	0	0
VI	1	1	1	16	VW	0	0	0	0
VL	0	0	0	0	VY	0	0	0	0
VK	0	0	0	0	VV	0	0	0	0

จากตารางที่ 5.22 แสดงค่าความถี่ของคู่ลำดับกรดอะมิโนแบบ 2-Grams ของข้อมูลผู้ป่วยดาวน์ซินโดรมโดยในกลุ่มสายพันธุกรรมซึ่งแสดงอาการของโรค (Positive Sequence) คู่ลำดับกรดอะมิโน AA มีค่าความถี่สูงสุดเท่ากับ 1 และมีค่าจำนวนครั้งของคู่ลำดับ (FIV) เท่ากับ 72 และในกลุ่มสายพันธุกรรมซึ่งไม่แสดงอาการของโรค (Negative Sequence) มีค่าค่าความถี่สูงสุดเท่ากับ 2 และมีค่าจำนวนครั้งของคู่ลำดับ (FIV) เท่ากับ 65

ตารางที่ 5.23 ค่าสถิติของข้อมูลผู้ป่วยดาวน์ซินโดรมแบบ 2-Grams ที่ไม่ปรับค่าน้ำหนัก (NBL)

ค่าสถิติของชุดข้อมูล	ลักษณะข้อมูล	
	แสดงอาการของโรค	ไม่แสดงอาการของโรค
ค่าความถี่สูงสุดของคู่ลำดับ (Max)	9.00	13.00
ค่าเฉลี่ยของความถี่คู่ลำดับ (Mean)	0.07	0.06
ค่าส่วนเบี่ยงเบนมาตรฐาน (StdDev)	0.21	0.23

จากตารางที่ 5.23 แสดงค่าสถิติของข้อมูลผู้ป่วยดาวน์ซินโดรมแบบ 2-Grams ที่ไม่ปรับค่าน้ำหนัก (NBL) มีความถี่สูงสุดของคู่ลำดับสำหรับกลุ่มที่แสดงอาการของโรคเท่ากับ 9.00 กลุ่มที่ไม่แสดงอาการของโรคเท่ากับ 13.00 มีค่าความถี่เฉลี่ยของคู่ลำดับกลุ่มที่แสดงอาการของโรคเท่ากับ 0.07 กลุ่มที่ไม่แสดงอาการของโรคเท่ากับ 0.06 มีค่าส่วนเบี่ยงเบนมาตรฐานของคู่ลำดับกลุ่มที่แสดงอาการของโรคเท่ากับ 0.21 กลุ่มที่ไม่แสดงอาการของโรคเท่ากับ 0.23 โดยเมื่อพิจารณาทั้งชุดข้อมูล พบว่า มีค่าความถี่สูงสุดเท่ากับ 13.00 ค่าความถี่เฉลี่ยเท่ากับ 0.07 และมีค่าส่วนเบี่ยงเบนมาตรฐานเท่ากับ 0.22

ตารางที่ 5.24 ค่าสถิติของข้อมูลผู้ป่วยดาวน์ซินโดรมแบบ 2-Grams ที่ปรับค่าน้ำหนัก (WBL)

ค่าสถิติของชุดข้อมูล	**ลักษณะข้อมูล**	
	แสดงอาการของโรค	**ไม่แสดงอาการของโรค**
ค่าความถี่สูงสุดของคู่ลำดับ (Max)	6.67	9.09
ค่าเฉลี่ยของความถี่คู่ลำดับ (Mean)	0.07	0.06
ค่าส่วนเบี่ยงเบนมาตรฐาน (StdDev)	0.22	0.22

จากตารางที่ 5.24 แสดงค่าสถิติของข้อมูลผู้ป่วยดาวน์ซินโดรมแบบ 2-Grams ที่ปรับค่าน้ำหนัก (WBL) มีความถี่สูงสุดของคู่ลำดับสำหรับกลุ่มที่แสดงอาการของโรคเท่ากับ 6.67 กลุ่มที่ไม่แสดงอาการของโรคเท่ากับ 9.09 มีค่าความถี่เฉลี่ยสำหรับกลุ่มที่แสดงอาการของโรคเท่ากับ 0.07 กลุ่มที่ไม่แสดงอาการของโรคเท่ากับ 0.06 มีค่าส่วนเบี่ยงเบนมาตรฐานสำหรับกลุ่มที่แสดงอาการของโรคเท่ากับ 0.22 กลุ่มที่ไม่แสดงอาการของโรคเท่ากับ 0.22 เมื่อพิจารณาทั้งชุดข้อมูลพบว่า มีค่าความถี่สูงสุดเท่ากับ 9.09 มีค่าความถี่เฉลี่ยเท่ากับ 0.06 และมีค่าส่วนเบี่ยงเบนมาตรฐานเท่ากับ 0.22

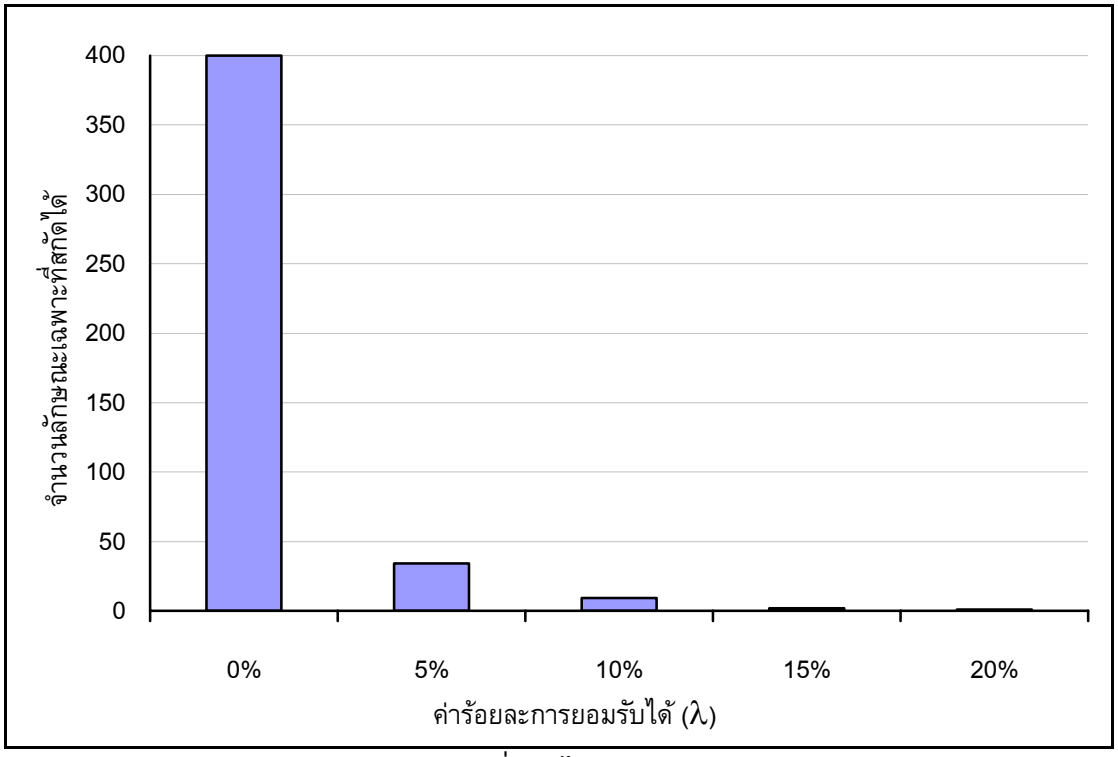
5.5.2 ขั้นตอนที่ 2 สกัดลักษณะเฉพาะบนข้อมูลผู้ป่วยดาวน์ซินโดรม

จำนวนลักษณะเฉพาะที่สกัดได้จากการทดลองแบบ C และ D โดยกำหนดค่าร้อยละการยอมรับได้ (λ) ที่แตกต่างกันในขั้นตอนการวิเคราะห์ระดับชุดข้อมูล (Global Sequences Analyze) ด้วยค่า λ เท่ากับ 0%, 5%, 10%, 15% และ 20% ตามลำดับ สามารถแสดงจำนวนของลักษณะเฉพาะที่สกัดได้ ดังตารางที่ 5.25

ตารางที่ 5.25 จำนวนลักษณะเฉพาะที่สกัดได้จากข้อมูลผู้ป่วยดาวน์ซินโดรม จากการทดลองแบบ C และ D เมื่อกำหนดค่าร้อยละการยอมรับได้ (λ) ที่แตกต่างกัน

วิธีการทดลอง	**จำนวนลักษณะเฉพาะที่สกัดได้**				
	$\lambda \geq 0\%$	**$\lambda \geq 5\%$**	**$\lambda \geq 10\%$**	**$\lambda \geq 15\%$**	**$\lambda \geq 20\%$**
แบบ C และ D	400	41	12	3	1

จากตารางที่ 5.25 พบว่า เมื่อกำหนดค่า $\lambda \geq 5\%$ สามารถสกัดได้ 41 คุณสมบัติ ได้แก่ AA, AN, AL, AP, AT, RE, RI, RM, ND, NE, NP, NS, DR, CA, CR, CT, EN, ED, EG, EI, EF, ET, GE, GI, GL, IN, IC, IE, IS, IT, ME, PR, PI, SA, SC, SY, TA, TE, TI, WN และ YN ที่ค่า $\lambda \geq 10\%$ สามารถสกัดได้ 12 คุณสมบัติ ได้แก่ AL, AP, RE, NS, EN, GI, IN, IE, PI, SA, TE และ TI ที่ค่า $\lambda \geq 15\%$ สามารถสกัดได้ 3 คุณสมบัติ ได้แก่ EN, GI และ IN และที่ค่า $\lambda \geq 20\%$ สามารถสกัดได้ 1 คุณสมบัติ คือ GI สามารถแสดงแผนภูมิเปรียบเทียบจำนวนลักษณะเฉพาะที่สกัดได้ ดังภาพประกอบ 5.39



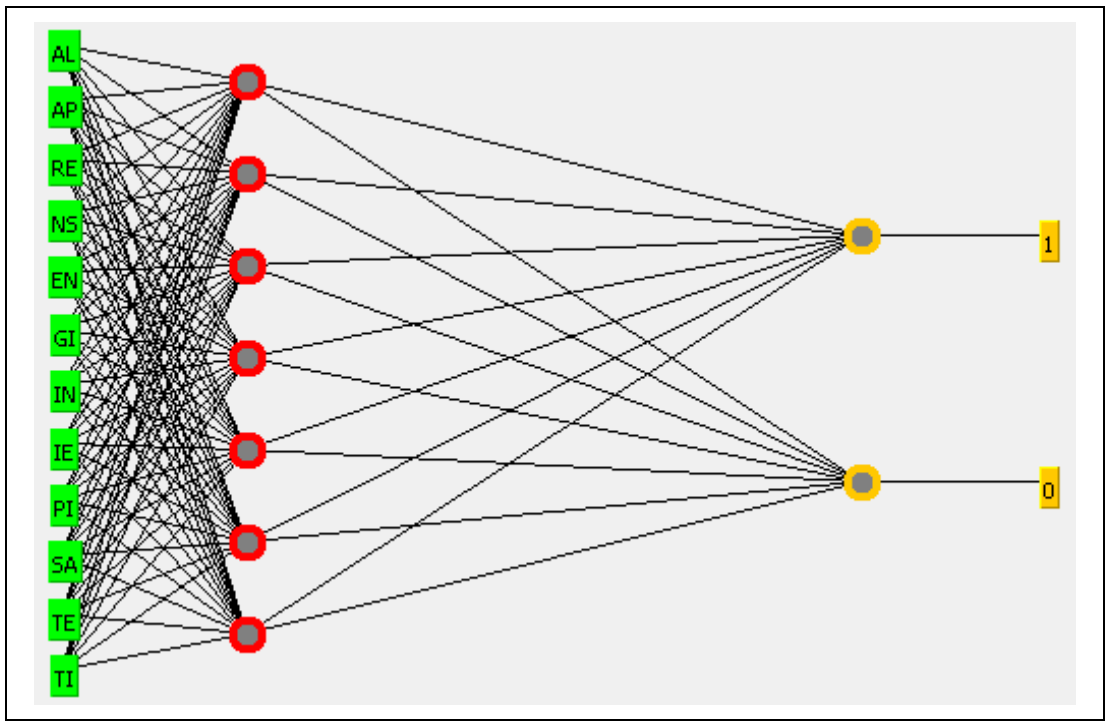
ภาพประกอบ 5.39 จำนวนลักษณะเฉพาะที่สกัดได้จากการทดลองแบบ C และ D บนชุดข้อมูลผู้ป่วยดาวน์ซินโดรมเมื่อกำหนดค่าร้อยละการยอมรับได้ (λ) ที่แตกต่างกัน

จากภาพประกอบ 5.39 พบว่า จำนวนลักษณะที่สกัดลักษณะเฉพาะได้จากการทดลองแบบ C และ D จะมีจำนวนลดลง เมื่อกำหนดค่าร้อยละของการยอมรับได้ (λ) ที่สูงขึ้น

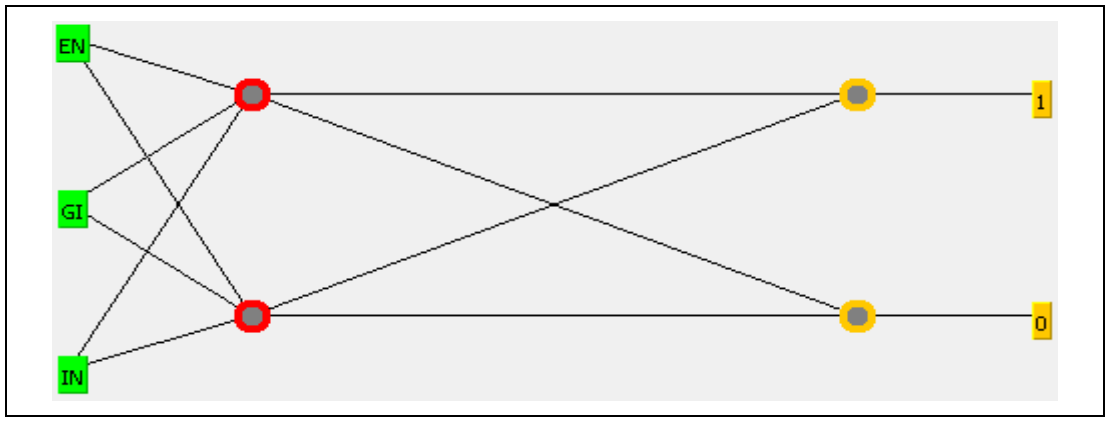
5.5.3 จำแนกข้อมูลผู้ป่วยดาวน์ซินโดรมด้วยลักษณะเฉพาะ

ทำการจำแนกขั้นตอนวิธี 2 กลุ่มคือ กลุ่มที่ 1 โครงข่ายประสาทเทียมแบบ MLP และแบบ RBF สำหรับกลุ่มที่ 2 ได้แก่ ต้นไม้การตัดสินใจแบบ ADTree J48 และ RandomTree ตามลำดับ

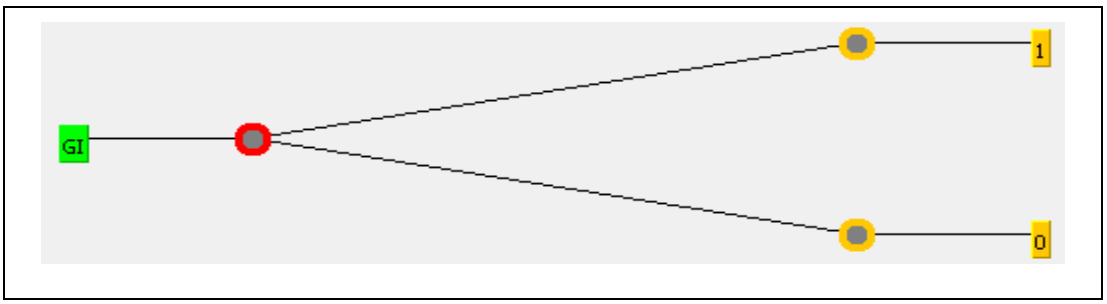
โดยสามารถแสดงลักษณะตัวอย่างโครงข่ายประสาทเทียมแบบ MLP ที่ใช้ในการจำแนกข้อมูลได้ดังภาพประกอบ 5.40 ถึงภาพประกอบ 5.42 ตามลำดับ



ภาพประกอบ 5.40 โครงข่ายประสาทเทียมแบบ MLP สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยดาวน์ซินโดรม เมื่อกำหนดค่า $\lambda \geq 10\%$ (Input = 12 ตัว)



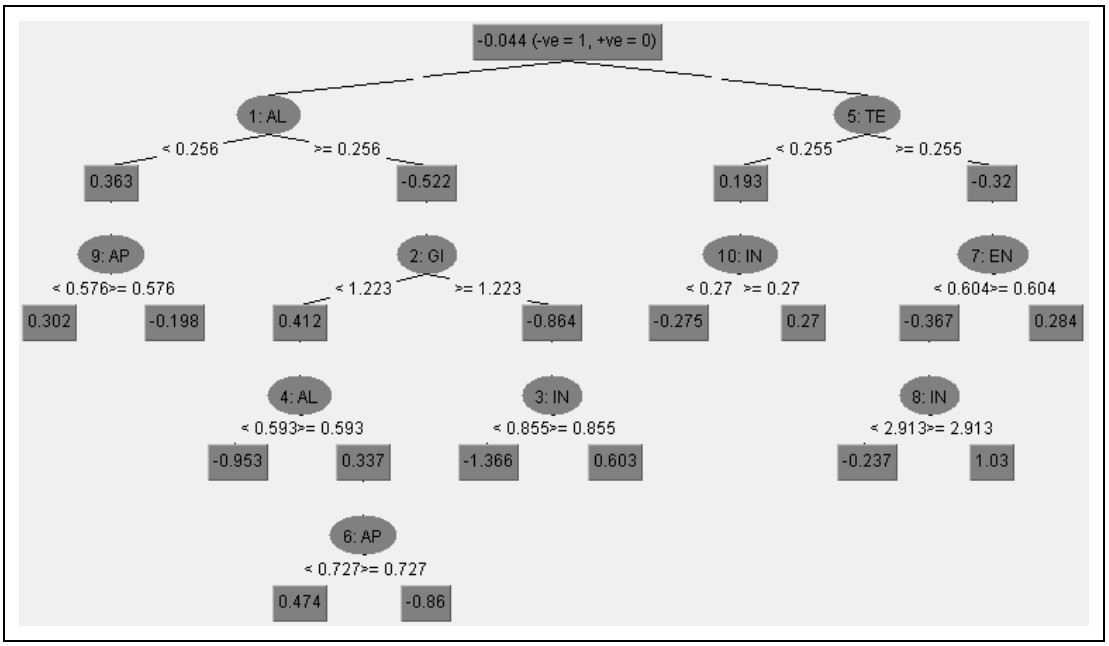
ภาพประกอบ 5.41 โครงข่ายประสาทเทียมแบบ MLP สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยดาวน์ซินโดรม เมื่อกำหนดค่า $\lambda \geq 15\%$ (Input = 3 ตัว)



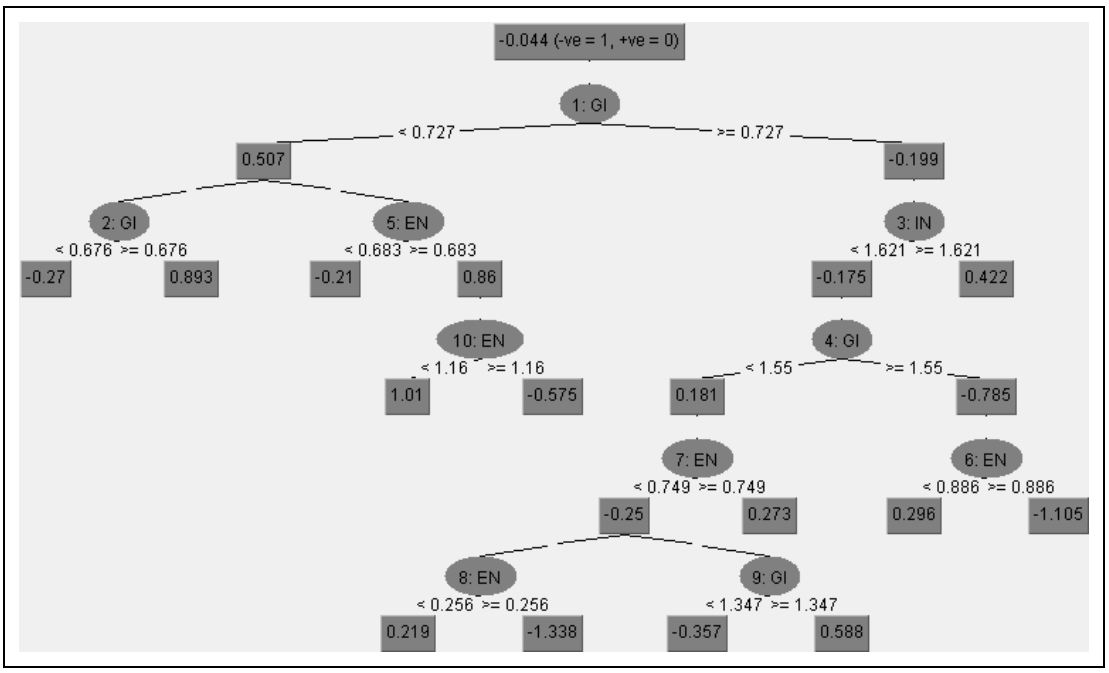
ภาพประกอบ 5.42 โครงข่ายประสาทเทียมแบบ MLP สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยดาวน์ซินโดรม เมื่อกำหนดค่า $\lambda \geq 20\%$ (Input = 1 ตัว)

จากภาพประกอบ 5.40 ถึงภาพประกอบ 5.42 ลักษณะของโครงข่ายประสาทเทียมแบบ MLP คือ ชั้นที่ 1 ชั้นข้อมูลนำเข้าผ่านทางคุณสมบัติของข้อมูลซึ่งสกัดได้ ชั้นที่ 2 เป็นชั้นซ่อน และในชั้นที่ 3 เป็นชั้นผลลัพธ์ ประกอบด้วยผลลัพธ์จำนวน 2 ค่า คือ 1 แทน ความหมายว่าแสดงอาการของโรค (Positive) และค่า 0 แทน ความหมายว่าไม่แสดงอาการของโรค (Negative)

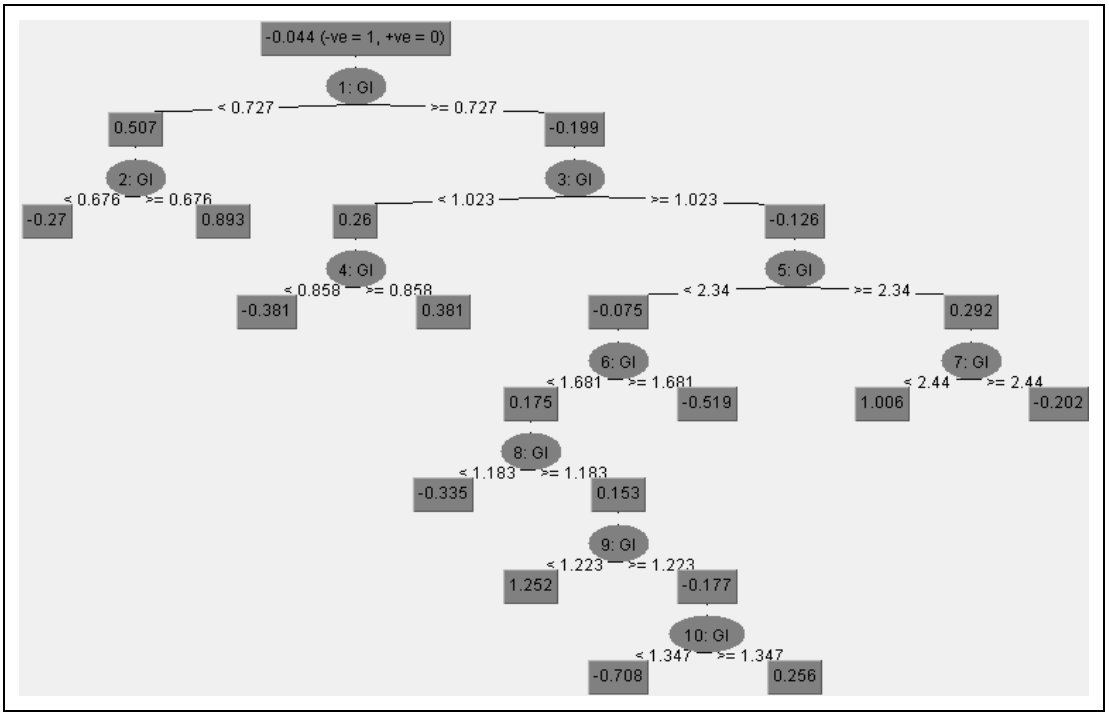
สามารถแสดงตัวอย่างของต้นไม้การตัดสินใจสำหรับการจำแนกข้อมูล ด้วยขั้นตอนวิธีแบบ ADTree ดังภาพประกอบ 5.43 ถึงภาพประกอบ 5.45 ตามลำดับ



ภาพประกอบ 5.43 ต้นไม้การตัดสินใจแบบ ADTree สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยดาวน์ซินโดรม เมื่อกำหนดค่า $\lambda \geq 10\%$

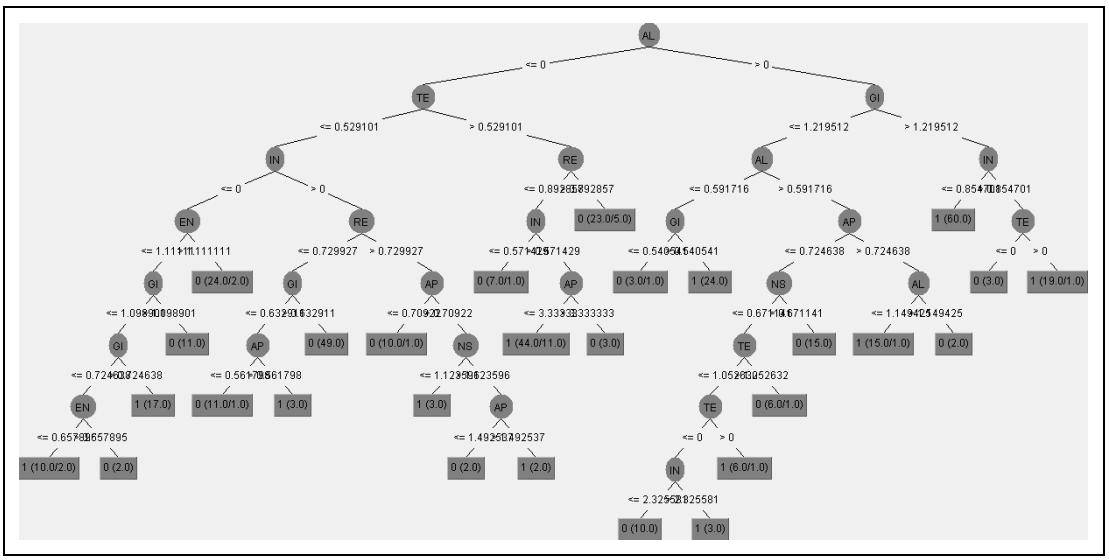


ภาพประกอบ 5.44 ต้นไม้การตัดสินใจแบบ ADTree สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยดาวน์ซินโดรม เมื่อกำหนดค่า $\lambda \geq 15\%$

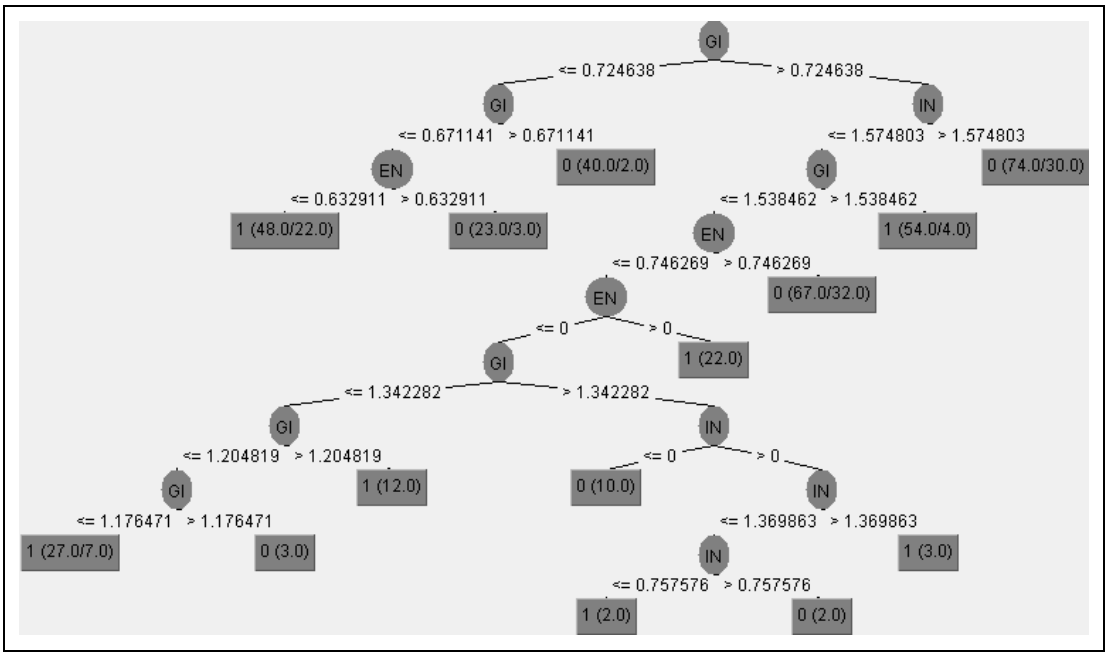


ภาพประกอบ 5.45 ต้นไม้การตัดสินใจแบบ ADTree สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยดาวน์ซินโดรม เมื่อกำหนดค่า $\lambda \geq 20\%$

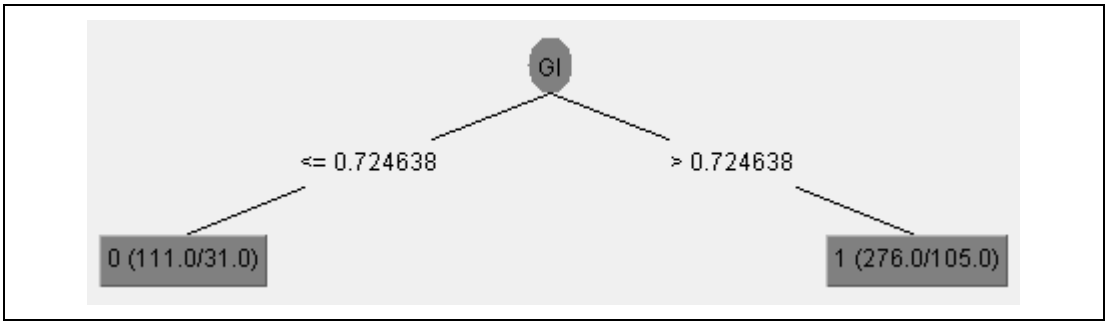
และสามารถแสดงตัวอย่างของต้นไม้การตัดสินใจสำหรับการจำแนกข้อมูล ด้วยขั้นตอนวิธีแบบ J48 ดังภาพประกอบ 5.46 และ 5.48 ตามลำดับ



ภาพประกอบ 5.46 ต้นไม้การตัดสินใจแบบ J48 สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยดาวน์ซินโดรม เมื่อกำหนดค่า $\lambda \geq 10\%$



ภาพประกอบ 5.47 ต้นไม้การตัดสินใจแบบ J48 สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยดาวน์ซินโดรม เมื่อกำหนดค่า $\lambda \geq 15\%$



ภาพประกอบ 5.48 ต้นไม้การตัดสินใจแบบ J48 สำหรับการทดลองแบบ D บนชุดข้อมูลผู้ป่วยดาวน์ซินโดรม เมื่อกำหนดค่า $\lambda \geq 20\%$

จากภาพประกอบ 5.46 ถึงภาพประกอบ 5.48 พบว่า ลักษณะต้นไม้การตัดสินใจแบบ ADTree โดยรวมมีความสมดุล (Balance) ของต้นไม้มากกว่าแบบ J48 และพบว่าต้นไม้การตัดสินใจแบบ J48 มีจำนวนชั้นของต้นไม้มากกว่าแบบ ADTree

5.5.4 เปรียบเทียบผลการทดลองข้อมูลผู้ป่วยดาวน์ซินโดรม

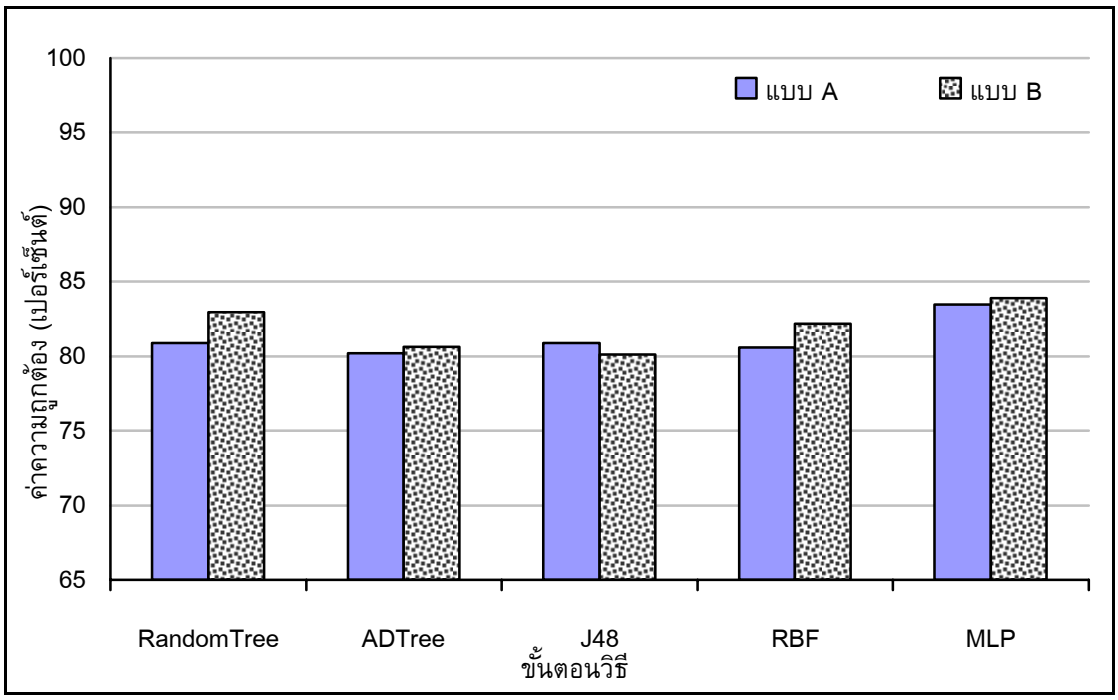
5.5.4.1 เปรียบเทียบค่าความถูกต้องของการจำแนกข้อมูล

ค่าความถูกต้องของการจำแนกข้อมูลตามแบบจำลองการวิเคราะห์รูปและรู้จำลำดับพันธุกรรม สามารถแสดงดังตารางที่ 5.26 และตารางที่ 5.27 ตามลำดับ

ตารางที่ 5.26 ความถูกต้องการจำแนกข้อมูลผู้ป่วยดาวน์ซินโดรมของการทดลองแบบ A และ B

การทดลอง	**จำนวนคุณสมบัติ**	**ค่าความถูกต้อง (เปอร์เซ็นต์)**				
		RandomTree	**ADTree**	**J48**	**RBF**	**MLP**
แบบ A	400	80.88%	80.19%	80.88%	80.59%	83.46%
แบบ B	400	82.95%	80.62%	80.10%	82.17%	83.90%

จากตารางที่ 5.26 สามารถแสดงแผนภูมิเปรียบเทียบค่าความถูกต้องการจำแนกข้อมูลผู้ป่วยดาวน์ซินโดรมด้วยขั้นตอนวิธีแบบต่างๆ ด้วยการทดลองแบบ A และ B บน แสดงดังภาพประกอบ 5.49



ภาพประกอบ 5.49 ความถูกต้องการจำแนกผู้ป่วยดาวน์ซินโดรมของการทดลอง A และ B

จากภาพประกอบ 5.49 พบว่า การทดลองซึ่งมีการปรับค่าน้ำหนักข้อมูล (WBL) ให้ค่าความถูกต้อง (Accuracy) สูงกว่าการทดลองแบบไม่ปรับค่าน้ำหนัก (NBL) ของข้อมูล โดยขั้นตอนวิธีแบบ MLP ให้ค่าความถูกต้องในการจำแนกสูงที่สุด เท่ากับ 83.90% รองลงมาคือ RandomTree เท่ากับ 82.95% RBF เท่ากับ 82.17% ADTree เท่ากับ 80.62% และ J48 เท่ากับ 80.10% ตามลำดับ และสามารถแสดงค่าความถูกต้องการจำแนกของการทดลองแบบ C และ D เมื่อกำหนดค่าร้อยละการยอมรับได้ (λ) ที่แตกต่างกัน ดังตารางที่ 5.27

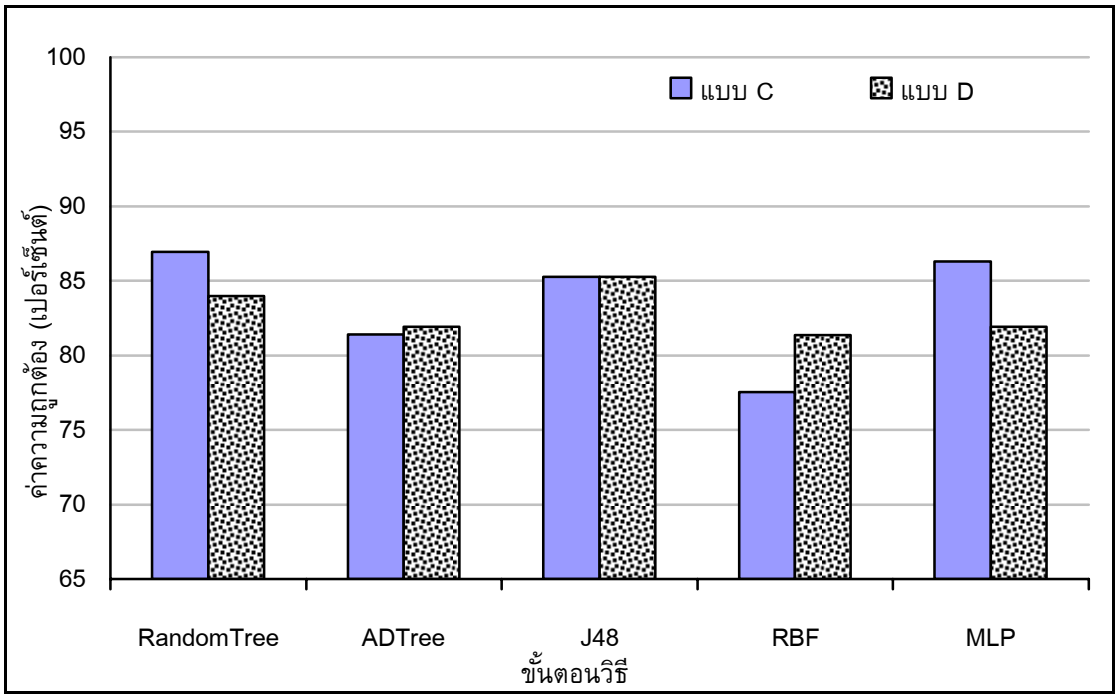
ตารางที่ 5.27 ความถูกต้องการจำแนกข้อมูลผู้ป่วยดาวน์ซินโดรมของการทดลองแบบ C และ D เมื่อกำหนดค่าร้อยละการยอมรับได้ (λ) ที่แตกต่างกัน

วิธีการทดลอง	λ	จำนวนคุณสมบัติ	ค่าความถูกต้อง (เปอร์เซ็นต์)				
			RandomTree	ADTree	J48	RBF	MLP
แบบ C	0%	400	80.88%	80.19%	80.88%	80.59%	83.46%
	5%	41	86.95%	81.39%	85.27%	77.52%	86.30%
	10%	12	81.91%	84.75%	84.24%	80.83%	80.87%
	15%	3	67.44%	69.77%	66.41%	67.70%	67.18%
	20%	1	66.41%	66.41%	66.41%	66.41%	66.41%

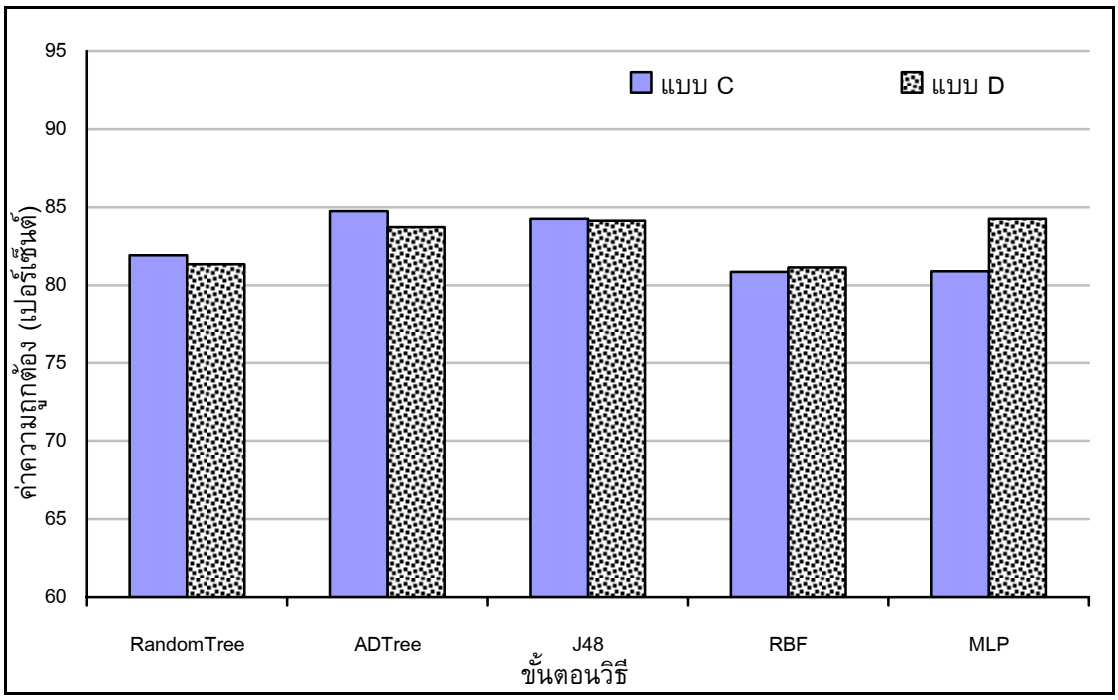
ตารางที่ 5.27 ความถูกต้องการจำแนกข้อมูลผู้ป่วยดาวน์ซินโดรมของการทดลองแบบ C และ D เมื่อกำหนดค่าร้อยละการยอมรับได้ (λ) ที่แตกต่างกัน (ต่อ)

วิธีการทดลอง	λ	จำนวนคุณสมบัติ	ค่าความถูกต้อง (เปอร์เซ็นต์)				
			RandomTree	**ADTree**	**J48**	**RBF**	**MLP**
แบบ D	0%	400	82.95%	80.62%	80.10%	82.17%	83.90%
	5%	41	83.98%	81.91%	85.27%	81.36%	81.91%
	10%	12	81.33%	83.72%	84.11%	81.14%	84.23%
	15%	3	72.87%	63.31%	64.34%	62.27%	67.18%
	20%	1	72.09%	63.82%	62.53%	61.24%	62.53%

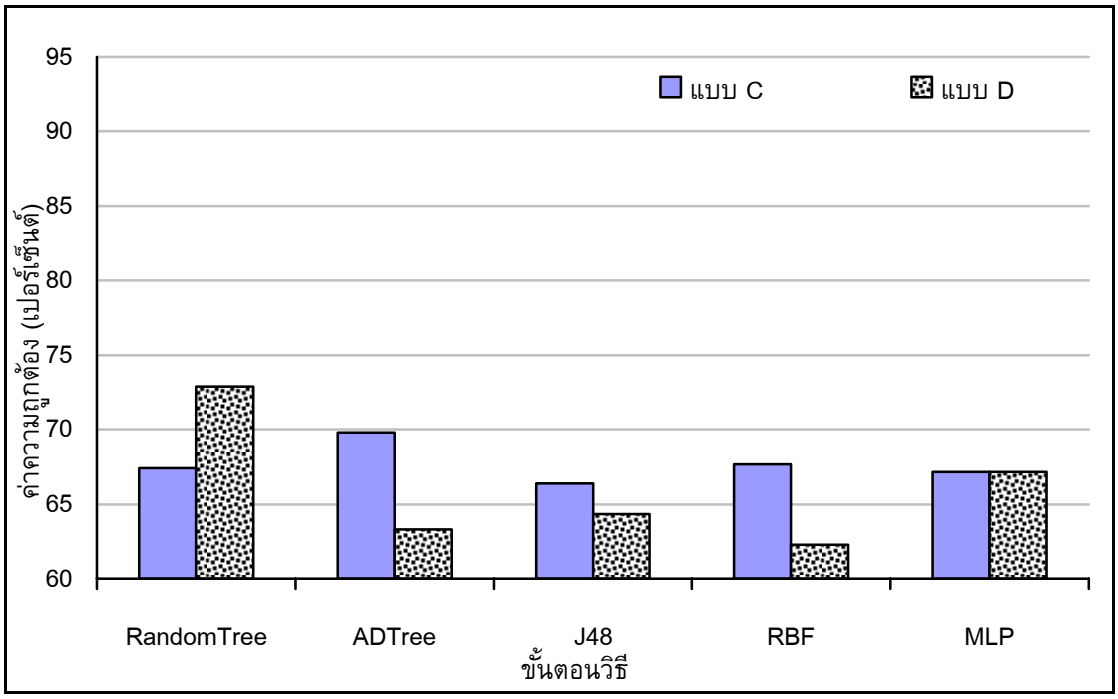
จากตารางที่ 5.27 สามารถแสดงแผนภูมิเปรียบเทียบความถูกต้องการจำแนกข้อมูลสายพันธุกรรมผู้ป่วยดาวน์ซินโดรมของการทดลองแบบ C และ D ดังแสดงในภาพประกอบ 5.50 ถึงภาพประกอบ 5.53 ตามลำดับ



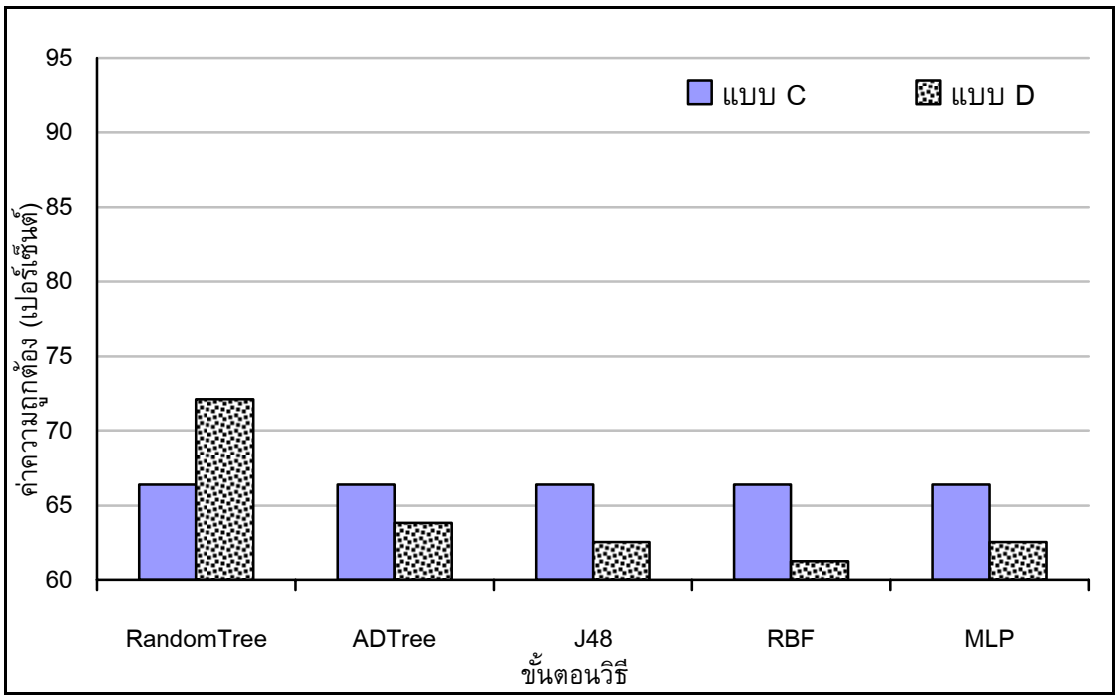
ภาพประกอบ 5.50 ความถูกต้องการจำแนกข้อมูลผู้ป่วยดาวน์ซินโดรมของการทดลองแบบ C และ D เมื่อกำหนดค่า $\lambda \geq 5\%$



ภาพประกอบ 5.51 ความถูกต้องการจำแนกข้อมูลผู้ป่วยดาวน์ซินโดรมของการทดลอง แบบ C และ D เมื่อกำหนดค่า $\lambda \geq 10\%$



ภาพประกอบ 5.52 ความถูกต้องการจำแนกข้อมูลผู้ป่วยดาวน์ซินโดรมของการทดลอง แบบ C และ D เมื่อกำหนดค่า $\lambda \geq 15\%$



ภาพประกอบ 5.53 ความถูกต้องการจำแนกข้อมูลผู้ป่วยดาวน์ซินโดรมของการทดลอง แบบ C และ D เมื่อกำหนดค่า $\lambda \geq 20\%$

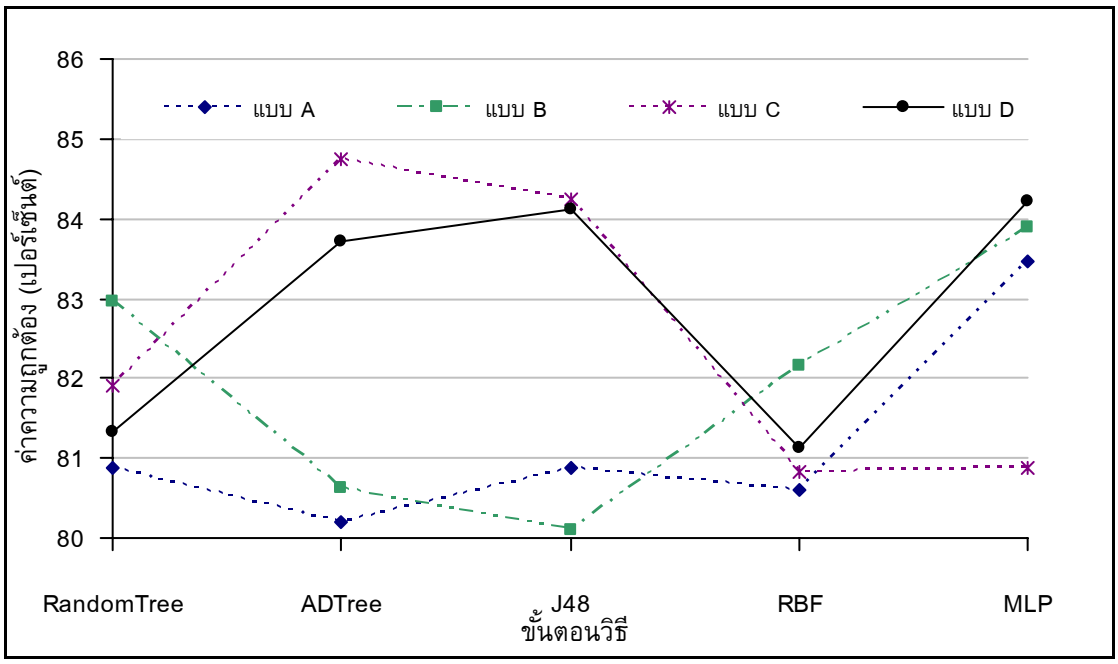
จากภาพประกอบ 5.50 ถึงภาพประกอบ 5.53 พบว่าการทดลองแบบ D ให้ค่าความถูกต้องในการจำแนกข้อมูลสูงกว่าการทดลองแบบ C และจากภาพประกอบ 5.33 ซึ่งมีการกำหนดค่า $\lambda \geq 10\%$ พบว่า มีจำนวนของคุณสมบัติของข้อมูลที่สกัด 12 คุณสมบัติ และให้ค่าความถูกต้องสูงที่สุด และเมื่อเปรียบเทียบขั้นตอนในการจำแนกข้อมูลพบว่าขั้นตอนวิธีแบบ MLP ให้ค่าความถูกต้องสูงสุดเท่ากับ 84.23% J48 เท่ากับ 84.11% RBF เท่ากับ 84.11% ADTree เท่ากับ 83.72% และ RandomTree เท่ากับ 81.33% ตามลำดับ และจากภาพประกอบ 5.35 พบว่าที่ค่า $\lambda \geq 20\%$ ได้จำนวนคุณสมบัติเพียง 1 คุณสมบัติ และค่าความถูกต้องที่ได้จากการจำแนกข้อมูลมีค่าต่ำ โดยขั้นตอนวิธีแบบ ADTree ให้ค่าความถูกต้องเท่ากับ 63.82% MLP เท่ากับ 62.53% J48 เท่ากับ 62.53% RBF เท่ากับ 61.24% และ RandomTree เท่ากับ 72.09% ตามลำดับ

โดยเมื่อนำผลการจำแนกข้อมูลของการทดลองแบบ A และ B เปรียบเทียบกับการทดลองแบบ C และ D ซึ่งกำหนดค่า $\lambda \geq 10\%$ แสดงดังตารางที่ 5.28 และภาพประกอบ 5.54

ตารางที่ 5.28 เปรียบเทียบค่าความถูกต้องการจำแนกข้อมูลผู้ป่วยดาวน์ซินโดรม

การทดลอง	จำนวนคุณสมบัติ	ค่าความถูกต้อง (เปอร์เซ็นต์)				
		RandomTree	ADTree	J48	RBF	MLP
แบบ A	400	80.88%	80.19%	80.88%	80.59%	83.46%
แบบ B	400	82.95%	80.62%	80.10%	82.17%	83.90%
แบบ C	12	81.91%	84.75%	84.24%	80.83%	80.87%
แบบ D	12	81.33%	83.72%	84.11%	81.14%	84.23%

*หมายเหตุ การทดลองแบบ C และ D กำหนดค่าร้อยละการยอมรับได้เท่ากับ 10 ($\lambda \geq 10\%$)



ภาพประกอบ 5.54 เปรียบเทียบค่าความถูกต้องการจำแนกข้อมูลผู้ป่วยดาวน์ซินโดรม

5.5.4.2 เวลาที่ใช้ในการจำแนกข้อมูล

สามารถแสดงเวลาที่ใช้ในการจำแนกข้อมูลผู้ป่วยดาวน์ซินโดรมของการทดลองแบบ A และ B ดังตารางที่ 5.29 และการทดลองแบบ C และ D ดังตารางที่ 5.30 ตามลำดับ

ตารางที่ 5.29 เวลาที่ใช้ในการจำแนกข้อมูลผู้ป่วยดาวน์ซินโดรมของการทดลองแบบ A และ B

การทดลอง	จำนวนคุณสมบัติ	เวลาที่ใช้ (วินาที)				
		RandomTree	ADTree	J48	RBF	MLP
แบบ A	400	0.71	8.97	1.44	2.11	61.80
แบบ B	400	0.83	9.64	2.52	1.41	54.43

ตารางที่ 5.30 เวลาที่ใช้ในการจำแนกข้อมูลผู้ป่วยดาวน์ซินโดรมของการทดลองแบบ C และ D เมื่อกำหนดค่าร้อยละการยอมรับได้ (λ) ที่แตกต่างกัน

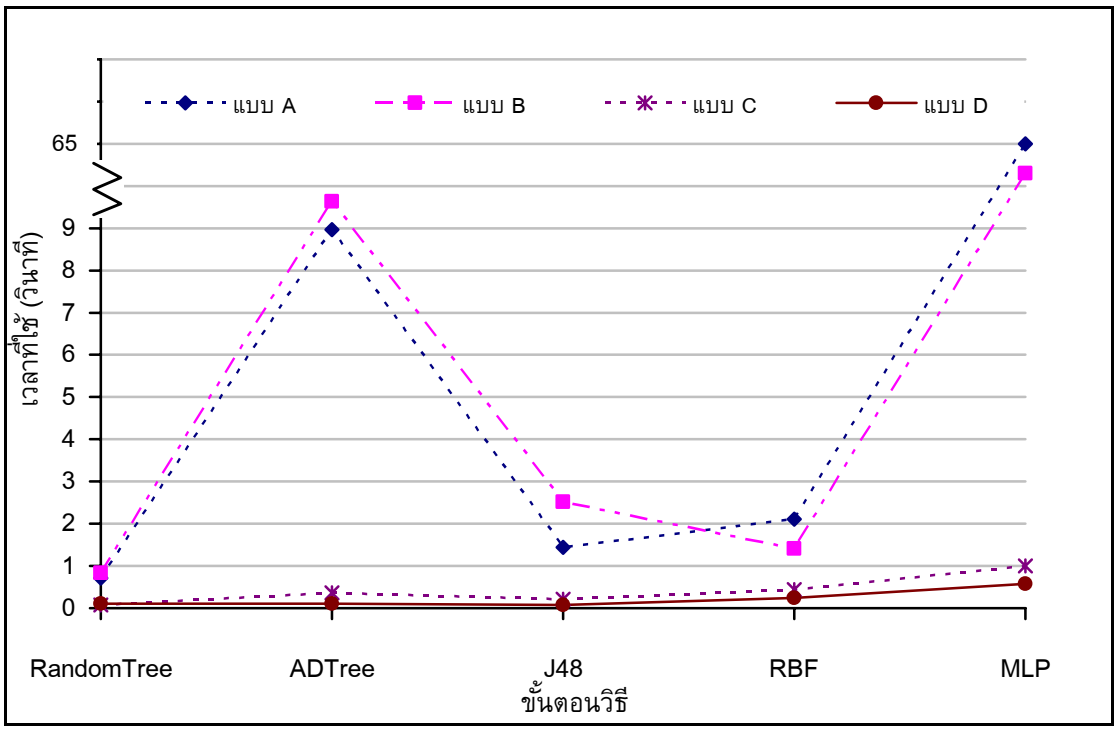
การทดลอง	λ	จำนวนคุณสมบัติ	เวลาที่ใช้ (วินาที)				
			RandomTree	ADTree	J48	RBF	MLP
แบบ C	0%	400	0.66	5.33	3.22	1.97	61.80
	5%	41	0.11	0.65	0.38	0.47	5.27
	10%	12	0.08	0.36	0.21	0.44	1.00
	15%	3	0.04	0.05	0.30	0.47	0.28
	20%	1	0.02	0.03	0.03	0.72	0.19
แบบ D	0%	400	0.83	4.38	2.44	2.27	54.43
	5%	41	0.28	0.35	0.30	0.36	3.01
	10%	12	0.10	0.11	0.08	0.25	0.57
	15%	3	0.02	0.05	0.28	0.25	0.17
	20%	1	0.02	0.03	0.15	0.19	0.13

จากตารางที่ 5.29 และตารางที่ 5.30 พบว่า การทดลองแบบ A และ B ใช้เวลาในการจำแนกสูงกว่าการทดลองแบบ C และ D และขั้นตอนวิธีแบบ MLP ใช้เวลาในการจำแนกมากที่สุด และขั้นตอนวิธีแบบ RandomTree ใช้เวลาในการจำแนกน้อยที่สุด ตารางที่ 5.31 และภาพประกอบ 5.55 แสดงการเปรียบเทียบเวลาที่ใช้ในการจำแนกข้อมูลของการทดลองแบบ A และ B กับการทดลองแบบ C และ D เมื่อกำหนดค่า $\lambda \geq 10\%$ ซึ่งมีรายละเอียดดังนี้

ตารางที่ 5.31 เปรียบเทียบเวลาที่ใช้จำแนกข้อมูลผู้ป่วยดาวน์ซินโดรม

การทดลอง	จำนวนคุณสมบัติ	เวลาที่ใช้ (วินาที)				
		RandomTree	ADTree	J48	RBF	MLP
แบบ A	400	0.71	8.97	1.44	2.11	61.80
แบบ B	400	0.83	9.64	2.52	1.41	54.43
แบบ C	12	0.08	0.36	0.21	0.44	1.00
แบบ D	12	0.10	0.11	0.08	0.25	0.57

*หมายเหตุ การทดลองแบบ C และ D กำหนดค่าร้อยละการยอมรับได้เท่ากับ 10 ($\lambda \geq 10\%$)



ภาพประกอบ 5.55 เปรียบเทียบเวลาที่ใช้บนชุดข้อมูลผู้ป่วยดาวน์ซินโดรม

5.5.4.3 การวิจารณ์ผลทดลองข้อมูลผู้ป่วยดาวน์ซินโดรม

จากการทดลองแบบ A, B, C และ D (ตารางที่ 5.1) พบว่าการจำแนกข้อมูลผู้ป่วยดาวน์ซินโดรม การทดลองซึ่งมีการปรับค่าน้ำหนักข้อมูล (WBL) ให้ค่าความถูกต้องสูงกว่าการทดลองแบบไม่ปรับค่าน้ำหนัก (NBL) เมื่อพิจารณาขั้นตอนวิธีที่ใช้ พบว่าขั้นตอนวิธีแบบ MLP ให้ค่าความถูกต้องในการจำแนกสูงที่สุด รองลงมาคือ ADTree J48 RBF และ RandomTree ตามลำดับ และพบว่าขั้นตอนวิธีแบบ RandomTree ใช้เวลาน้อยที่สุด และชั้นตอนวิธีแบบ MLP ใช้เวลามากที่สุด และพบว่า ในขั้นตอนการสกัดลักษณะเฉพาะของการทดลองแบบ C และ D สามารถลดจำนวนคุณสมบัติของข้อมูลที่ใช้ในการจำแนกลงได้ โดยที่ค่า $\lambda \geq 5\%$ สามารถสกัดได้ 41 คุณสมบัติ ที่ค่า $\lambda \geq 10\%$ สกัดได้ 12 คุณสมบัติ ที่ค่า $\lambda \geq 15\%$ สกัดได้ 3 คุณสมบัติ และที่ค่า $\lambda \geq 20\%$ สกัดได้ 1 คุณสมบัติ ตามลำดับ และโดยที่ค่า $\lambda \geq 10\%$ ให้ค่าความถูกต้องของการจำแนกข้อมูลที่สูง และและที่ค่า $\lambda \geq 20\%$ ให้ค่าความถูกต้องการจำแนกข้อมูลต่ำที่สุด

5.6 วิเคราะห์ผลการทดลอง

จากแบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรมด้วยเทคนิคการสกัดลักษณะเฉพาะบนสายโปรตีน (FEPSA) พบว่าสามารถสกัดหาคู่ลำดับกรดอะมิโนที่เป็นลักษณะเฉพาะ (Feature) ออกมาและนำคู่ลำดับที่สกัดได้สร้างเป็นคุณสมบัติของข้อมูล (Attributes) เพื่อใช้ในการจำแนกข้อมูลได้ โดยให้ค่าความถูกต้องจากการจำแนกข้อมูลที่สูงและใช้เวลาในการจำแนกน้อยเมื่อเปรียบเทียบกับการทดลองซึ่งไม่มีการสกัดลักษณะเฉพาะของข้อมูล และสามารถแสดงตารางสรุปผลการทดลองของทั้ง 3 ชุดข้อมูล ได้ดังตารางที่ 5.32 และตารางที่ 5.33 โดยสามารถแสดงแผนภูมิเปรียบเทียบได้ดังภาพประกอบ 5.56 และภาพประกอบ 5.57 ตามลำดับ

ตารางที่ 5.32 ความถูกต้องในการจำแนกข้อมูลสายพันธุกรรมของการทดลองแบบ D

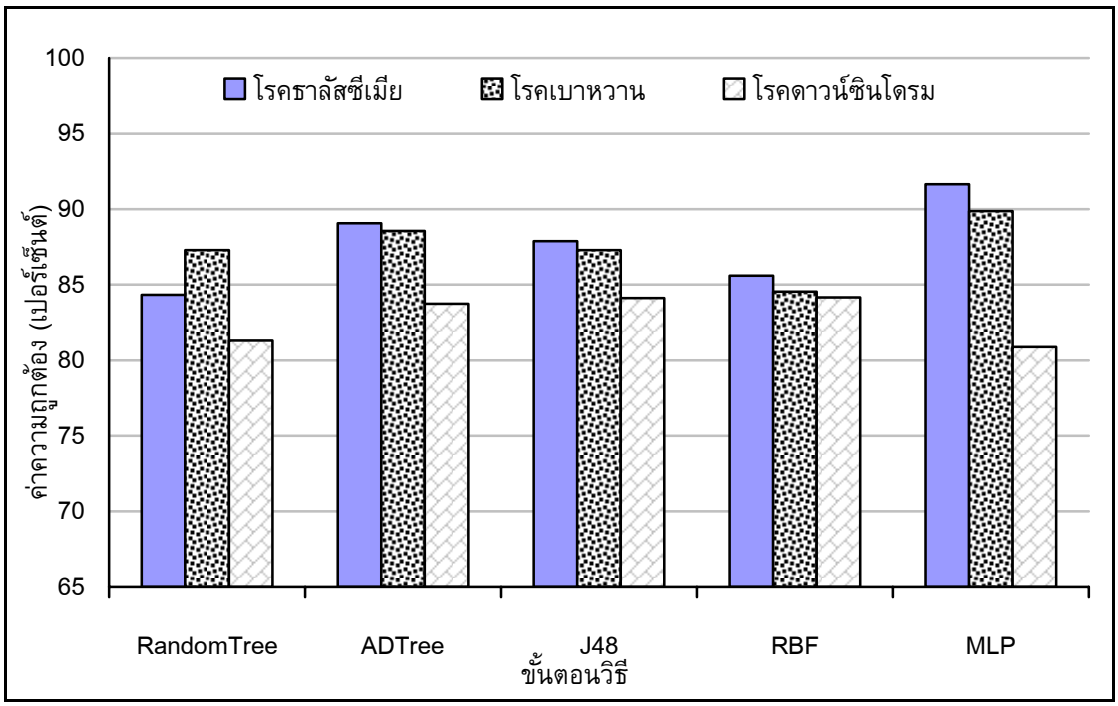
ข้อมูลผู้ป่วย	จำนวนคุณสมบัติ	ค่าความถูกต้อง (เปอร์เซ็นต์)				
		RandomTree	ADTree	J48	RBF	MLP
โรคธาลัสซีเมีย	23	84.33%	89.08%	87.88%	85.61%	91.64%
โรคเบาหวาน	9	87.28%	88.55%	87.28%	84.54%	89.87%
โรคดาวน์ซินโดรม	12	81.33%	84.72%	84.11%	81.14%	80.87%

*หมายเหตุ กำหนดค่าร้อยละการยอมรับได้เท่ากับ 10 ($\lambda \geq 10\%$)

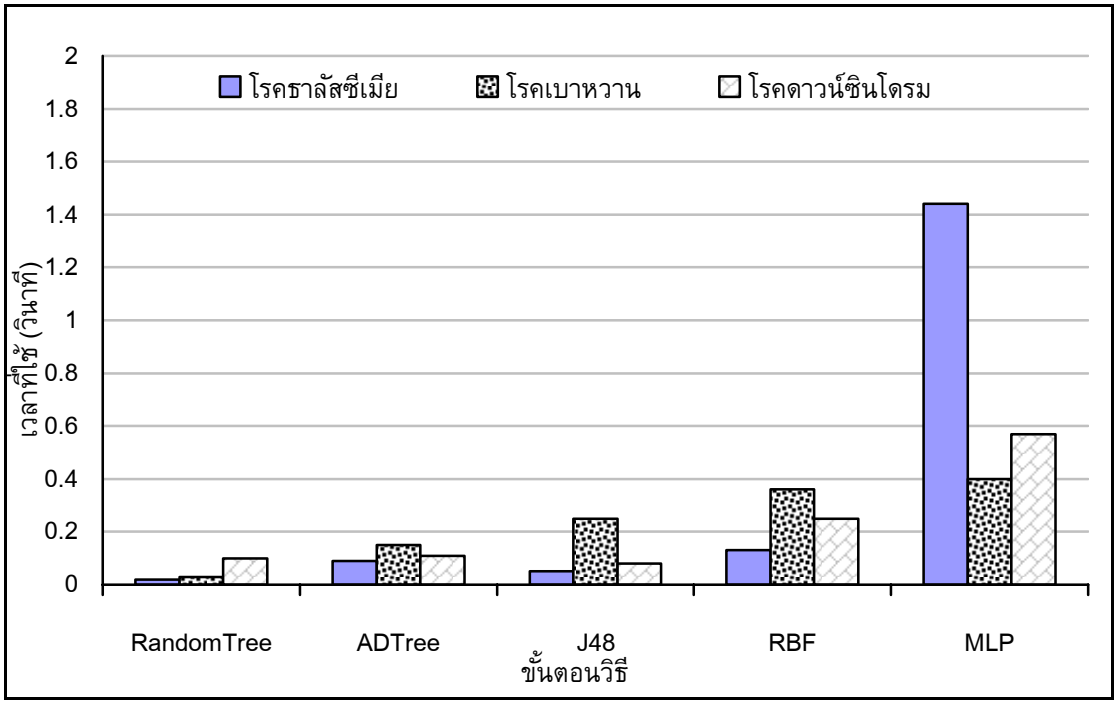
ตารางที่ 5.33 เวลาที่ใช้ในการจำแนกข้อมูลสายพันธุกรรมของการทดลองแบบ D

ข้อมูลผู้ป่วย	จำนวนคุณสมบัติ	ค่าความถูกต้อง (เปอร์เซ็นต์)				
		RandomTree	ADTree	J48	RBF	MLP
โรคธาลัสซีเมีย	23	0.02	0.09	0.05	0.13	1.44
โรคเบาหวาน	9	0.03	0.15	0.25	0.36	0.40
โรคดาวน์ซินโดรม	12	0.10	0.11	0.08	0.25	0.57

*หมายเหตุ กำหนดค่าร้อยละการยอมรับได้เท่ากับ 10 ($\lambda \geq 10\%$)



ภาพประกอบ 5.56 เปรียบเทียบค่าความถูกต้องการจำแนกข้อมูลสายพันธุกรรมของการทดลองแบบ D เมื่อกำหนดค่า $\lambda \geq 10\%$



ภาพประกอบ 5.57 เปรียบเทียบเวลาที่ใช้ในการจำแนกข้อมูลสายพันธุกรรมของการทดลองแบบ D เมื่อกำหนดค่า $\lambda \geq 10\%$

บทที่ 6

บทสรุปและข้อเสนอแนะ

งานวิจัยนี้ได้นำเสนอแบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรมด้วยเทคนิคการสกัดลักษณะเฉพาะบนสายโปรตีน (Feature Extraction of Protein Sequence Analysis: FEPSA) ซึ่งสามารถนำมาใช้เป็นเครื่องมือสำหรับการจำแนกสายพันธุกรรมร่วมกับขั้นตอนวิธีโครงข่ายประสาทเทียมและต้นไม้การตัดสินใจ โดยมีผลการทดลองที่ดีและมีค่าความถูกต้องในการจำแนกข้อมูลสูง

6.1 สรุปผลงานวิจัย

งานวิจัยนี้ได้บรรลุตามวัตถุประสงค์ โดยมีการออกแบบแบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม และพัฒนาโปรแกรมวิเคราะห์รูปแบบลำดับพันธุกรรมตามขั้นตอนวิธีของแบบจำลองซึ่งประกอบด้วย 3 ขั้นตอนหลัก คือ 1) การเตรียมข้อมูลค่าความถี่แบบปรับน้ำหนัก (Weight Balance Frequency Data Process) 2) การสกัดลักษณะเฉพาะบนสายพันธุกรรม (Feature Extraction Process) และ 3) การจำแนกสายพันธุกรรมด้วยลักษณะเฉพาะ (Feature Classification Process) ตามลำดับ ผลงานวิจัยที่ได้จากแบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม โดยได้รับการตีพิมพ์เรื่อง “n-Grams Protein Sequence Analysis for Disease Classification” ในการประชุมวิชาการ The 4th International Joint Conference on Computer Science and Software Engineering (JCSSE 2007) ระหว่างวันที่ 2-4 พฤษภาคม 2550 ณ โรงแรมโซฟิเทลราชาออคิด จังหวัดขอนแก่น ดังแสดงในภาคผนวก ค ซึ่งได้นำเสนอขั้นตอนการวิเคราะห์ลำดับของโปรตีนแบบเอ็นแกรมสำหรับการแบ่งกลุ่มโรค (n-Grams_PSA_DC) โดยใช้เทคนิคการหาค่าความถี่ของคู่ลำดับกรดอะมิโนแบบเอ็นแกรม (*n*-Grams) และใช้ขั้นตอนการเรียนรู้ของต้นไม้การตัดสินใจแบบ DecisionStump REPTree ADTree J48 และ RandomTree ในจำแนกข้อมูล

ผลจากการออกแบบและการทดลองตามขั้นตอนวิธีของแบบจำลองการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรมด้วยเทคนิคการสกัดลักษณะเฉพาะบนสายโปรตีน (FEPSA) ได้รับการตีพิมพ์เรื่อง “Feature Extraction of Protein Sequence Analysis Using n-Grams Weight Balance Frequency and Neural Networks” ในการประชุมวิชาการ 2008 IEEE International Conference on Research, Inovation and Vision for the Feature

ระหว่างวันที่ 13-17 กรกฎาคม 2551 ณ ประเทศเวียดนาม ดังแสดงในภาคผนวก ง และสามารถอธิบายผลลัพธ์จากการทดลองในส่วนต่าง ๆ ได้ดังนี้

6.1.1 การเตรียมข้อมูลค่าความถี่แบบปรับน้ำหนัก

ผลการทดลอง พบว่า การหาค่าความถี่ข้อมูลแบบเอ็นแกรมที่มีขนาด n = 2 ตัวอักษร ให้ค่าความถูกต้องของการจำแนกข้อมูลที่สูง และเมื่อนำค่าความถี่ข้อมูลแบบ 2-Grams ที่ได้มาทำการปรับค่าน้ำหนัก (Weight Balance Frequency: WBL) พบว่า ให้ค่าความถูกต้องการจำแนกข้อมูลสูงกว่าข้อมูลค่าความถี่แบบ 2-Grams ซึ่งไม่ปรับค่าน้ำหนัก (Non-weight Balance Frequency: NBL)

6.1.2 การสกัดลักษณะเฉพาะบนสายพันธุกรรม

จากการใช้เทคนิคการหาค่าความถี่ข้อมูลแบบ 2-Grams พบว่า จะได้คู่ลำดับซึ่งใช้สำหรับสร้างเป็นคุณสมบัติของข้อมูล (Attributes) ทั้งสิ้น 400 คู่ลำดับ ซึ่งใช้เวลาในการจำแนกข้อมูลเป็นเวลานาน การนำเทคนิคการสกัดลักษณะเฉพาะบนสายโปรตีน (FEPSA) มาทำการสกัดลักษณะเฉพาะซึ่งปรากฏบนสายพันธุกรรม สามารถสกัดหาคู่ลำดับกรดอะมิโนที่เป็นลักษณะเฉพาะ (Feature) ออกมาและใช้สำหรับการจำแนกข้อมูล โดยเมื่อกำหนดค่าร้อยละการยอมรับได้ของคู่ลำดับ (Support Threshold Point: λ) มีค่าสูงขึ้น จะทำให้จำนวนของคู่ลำดับที่ใช้ในการจำแนกข้อมูลมีจำนวนน้อยลง และทำให้เวลาในการจำแนกข้อมูลลดลง และยังคงให้ค่าความถูกต้องสูงเช่นเดิม

6.1.3 การจำแนกสายพันธุกรรมด้วยลักษณะเฉพาะ

ขั้นตอนวิธีที่ใช้ในการจำแนกข้อมูลสายพันธุกรรมจะเรียกใช้ชุดเครื่องมือ (Tools Box) ของโปรแกรม WEKA เวอร์ชัน 3.4.10 ด้วยขั้นตอนวิธีโครงข่ายประสาทเทียมแบบหลายชั้น (MLP) โครงข่ายประสาทเทียมเรเดียลเบสิสฟังก์ชัน (RBF) ต้นไม้การตัดสินใจแบบ Alternating Decision Tree (ADTree) แบบ Pruned C4.5 Decision Tree (J48) และแบบ K random features Decision Tree (RandomTree) ตามลำดับ เมื่อเปรียบเทียบค่าความถูกต้องที่ได้จากการจำแนกข้อมูล พบว่า โครงข่ายประสาทเทียมแบบหลายชั้น (MLP) ให้ค่าความถูกต้องสูงที่สุด และเมื่อเปรียบเทียบเวลาที่ใช้ในการจำแนกข้อมูลพบว่า ต้นไม้การตัดสินใจแบบ RandomTree ใช้เวลาน้อยที่สุด และโครงข่ายประสาทเทียมแบบหลายชั้นใช้เวลามากที่สุด

6.2 ปัญหาและอุปสรรค

ปัญหาและอุปสรรคที่พบในการทำงานวิจัย มีดังนี้

1) การพัฒนาโปรแกรมวิเคราะห์รูปแบบลำดับพันธุกรรมตามขั้นตอนวิธีของแบบจำลองด้วยโปรแกรม Matlab 2007a จำเป็นต้องใช้เครื่องคอมพิวเตอร์ที่มีสมรรถนะสูง

2) เนื่องจากข้อมูลที่ใช้ในการทดลองมีสายพันธุกรรมและคุณสมบัติของข้อมูล (Attributes) จำนวนมาก ดังนั้นจึงต้องใช้พื้นที่หน่วยความจำหลักจำนวนมากตามมา และส่งผลให้การทดลองตามขั้นตอนวิธีของแบบจำลองต้องใช้เวลานาน

6.3 ข้อเสนอแนะ

1) เนื่องจากข้อมูลสายพันธุกรรมที่ใช้ในการทดลองมีความแตกต่างในเรื่องความยาวของสายพันธุกรรม การเลือกสายพันธุกรรมที่มีความยาวที่ใกล้เคียงกันในการทดลองจึงเป็นสิ่งที่จำเป็น

2) ในขั้นตอนการเตรียมข้อมูลค่าความถี่คู่ลำดับกรดอะมิโนสายพันธุกรรมด้วยเทคนิคเอ็นแกรมแบบปรับค่าน้ำหนัก (WBL) ซึ่งมีสายพันธุกรรมที่ใช้ในการทดลองจำนวนมาก อาจนำเทคนิคการประมวลผลแบบขนาน (Parallel Processing) เช่น การประมวลผลแบบคลัสเตอร์ (Cluster Computing) หรือการประมวลผลแบบกริด (Grid Computing) ซึ่งมีเครื่องแม่ขายทำหน้าที่กระจายงานออกไปประมวลผลยังโหนดคำนวณหลายๆ โหนด มาใช้ในขั้นตอนนี้จะช่วยลดระยะเวลาในขั้นตอนการเตรียมข้อมูลลงได้

3) ปัจจุบันข้อมูลทางด้านวิทยาศาสตร์ชีวภาพในระดับโมเลกุล มีความหลากหลายและแตกต่างกัน เช่น สิ่งมีชีวิตที่ศึกษา โรคที่ศึกษา บริเวณหรือส่วนที่ทำการศึกษา หน้าที่การทำงานและรูปแบบของข้อมูลสายพันธุกรรม เป็นต้น ดังนั้นการเลือกข้อมูลสำหรับใช้ในการทดลองควรมีการระบุหรือเจาะจง เช่น การเลือกศึกษาเพียงกลุ่มยีนใดยีนหนึ่งที่มีความสัมพันธ์และเกี่ยวเนื่องกันต่อการแสดงออกของโรคแต่ละโรค จะช่วยให้ได้ความแม่นยำของการจำแนกข้อมูลมากยิ่งขึ้น

บรรณานุกรม

กัลยา วานิชย์บัญชา. 2546. การใช้งาน SPSS for Windows ในการวิเคราะห์ข้อมูล. ศูนย์หนังสือแห่งจุฬาลงกรณ์มหาวิทยาลัย: กรุงเทพฯ.

คมสัน จันมา และพันธุปิติ เปี่ยมสง่า. 2548. การเทียบเรียงกลุ่มลำดับข้อมูลชีวภาพโดยการพิจารณาความถี่ส่วนย่อยของลำดับ. Prodeeding of The 9th National Computer Science and Engineering Conference, University of The Thai Chamber of Commerce, October 27-28, 2005. pp.231-239.

ชัชวาลย์ เรืองประพันธ์. 2542. สถิติพื้นฐาน: พร้อมตัวอย่างการวิเคราะห์ด้วยโปรแกรม Minitab SPSS และ SAS. คณะวิทยาศาสตร์, มหาวิทยาลัยขอนแก่น.

มนตรี จุฬาวัฒนทล, ชิษณุสรร สวัสดิวัตน์, ยงยุทธ ยุทธวงศ์, ภิญโญ พานิชพันธ์, ประหยัด โกมารทัต,พิณทิพ รื่นวงษา, ธีรยศ วิทิตสุวรรณกุล, บุรชัย สนธยานนท์, สุมาลี ตั้งประดับกุล, และมธุรส พงษ์ลิขิตมงคล. 2543. ชีวเคมี. คณะวิทยาศาสตร์, มหาวิทยาลัยมหิดล.

วสันต์ จันทราทิตย์ และวีระพงศ์ ลุลิตานนท์. 2544. ชีวสารสนเทศศาสตร์. ศูนย์หนังสือแห่งจุฬาลงกรณ์มหาวิทยาลัย: กรุงเทพฯ.

วิสุทธ์ ใบไม้. 2538. พันธุศาสตร์. เอ็นซีซัพลายพริ้นติ้ง: กรุงเทพฯ.

สุนทร โสตถิพันธุ์. 2548. หลักชีววิทยา 1. ภาควิชาชีววิทยา คณะวิทยาศาสตร์. มหาวิทยาลัยสงขลานครินทร์.

อมรรัตน์ พงศ์ดารา. 2550. ชีวโมเลกุล: เพื่อการอยู่รอดของกุ้ง. ไอคิว มีเดีย: กรุงเทพฯ.

อุไรวรรณ จิจารณกุล. 2545. ดีเอ็นเอเทคโนโลยี. คณะวิทยาศาสตร์และเทคโนโลยี, สถาบันราชภัฏพิบูลสงคราม.

Addison Wesley Longman. 1999. http://www.informit.com/ (accessed 10/03/2007)

Aitkenhead, M.J. 2008. A Co-evolving Decision Tree Classification Method. Expert Systems With Applications: An International Journal. 34(1), 18-25.

Akay, M. 2004. Proteomics and Genomics Signal Processing. Proceeding of The 4th International Conference on Computer and Information Technology (CIT'04). pp.3.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic Local Alignment Search Tool. Journal of Molecular Biology. 215, 403-410.

Andrade-Cetto, L., and Manolakos, E. 2005. Feature Extraction for DNA Base-calling Using NNLS. Proceeding of The IEEE Workshop on Statistical Signal Processing, July 17-20, 2005. pp.1408-1413.

Andrija, T., Predrag, J., and Vlado, K. 2006. N-gram-based Classification and Unsupervised Hierarchical Clustering of Genome Sequences. Computer Methods and Programs in Biomedicine. 81(2), 137-153.

Atchley, W.R., and Zhao, J. 2007. Molecular Architecture of The DNA-binding Region and Its Relationship to Classification of Basic Helix–loop–helix Proteins. Molecular Biology and Evolution. 24(1), 192–202.

Bandyopadhyay, S. 2003. Feature extraction in Transformed Domain for Protein Superfamily Identification from Amino Acid Aequences. Proceeding of The 1st Indian International Conference on Artificial Intelligence 2003. pp. 610-620.

Bandyopadhyay, S. 2005. An Efficient Technique for Superfamily Classification of Amino Acid Sequences: Feature Extraction, Fuzzy Clustering and Prototype Selection. Fuzzy Sets and Systems. 152(1), 5-16.

Berry, E.A., Yang, Z.R., and Wu, X. 2003. A Biology Inspired Neural Learning Algorithm for Analysing Protein Sequence. Proceeding of The 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'2003). pp.18-25.

Browne, A., Hudson, B.D., Whitley, D.C., Ford, M.G., and Picton, P.D. 2004. Biological Data Mining with Neural Networks: Implementation and Application of A Flexible Decision Tree Extraction Algorithm to Genomic Problem Domains. Neurocomputing. 57, pp.275-293.

Burges, C.J. 2004. Geometric Methods for Feature Extraction and Dimensional Reduction: A Guided Tour. Technical Report, Microsoft Research, University of Toronto, June. 2004.

Camoglu O., Can T., Singh A.K., and Wang Y.F. 2005. Decision Tree Based Information Integration for Automated Protein. Journal of Bioinformatics and Computational Biology. 3(3), 717-742.

Cios, K.J., Pedrycz, W., Swiniarski, R.W., and Kurgan, L.A. 2007. Data Mining: Knowledge Discovery Approach. Springer: N/A.

Claverie, J.M., and Notredame, C. 2003. Bioinformatics For Dummies. John Wiley:N/A.

Cohen, J. 2004. Bioinformatics: An Introduction for Computer Scientists. ACM Computer Surveys, 36(2), 122-158.

Darnell, S.J., Page, D., and Mitchell, J.C. 2007. An Automated Decision-tree Approach to Predicting Protein Interaction Hot Spots. Proteins: Structure, Function, and Bioinformatics, 68(4), 813-23.

Devroye, L., Gyrfi, L., and Lugosi, G. 1996. A Probabilistic Theory of Pattern Recognition. Springer: New York.

Duda, R.O., Hart, P.E., and Stork, D.G. 2000. Pattern Classification, 2nd Editon. John Wiley: New York.

Eliasmith, C., and Anderson, C.H. 2003. Neural Engineering Computation, Representation, and Dynamics in Neurobiological Systems. The MIT Press.

Friedman, R. 2000. Lesson 5: Sequence Comparison and Alignment. Columbia University.

Fukui, K., Murase, H., Okayama, T., and Murakami, K. 2003. Features Extraction from DNA Sequences Using The Textural Analysis. American Society of Agricultural and Biological Engineers, Paper number: 033115.

Ganapathiraju, M., Weisser, D., Rosenfeld, R., Carbonell, J., Reddy, R., and Klein-Seetharaman, J. 2002. Comparative N-gram Analysis of Whole-genome Protein Sequences. Proceeding of The 2nd International Conference on Human Language Technology Research, San Diego, California, March 24-27, 2002. pp.76-81.

Golding, B., and Morton, D. 2003. Elementary Sequence Analysis. Department of Biology. McMaster University.

Hagan, M., Demuth, H., and Beale, M. 1996. Neural Network Design. University of Colorado Bookstore: United States of America.

Han, P., Zhang, X., Norton, R.S., and Feng, Z.P. 2006. Predicting Disordered Regions in Proteins Based on Decision Trees of Reduced Amino Acid Composition. Journal of Computational Biology. 13(10), 1723-1734.

Herrero, J., Valencia, A., and Dopazo, J. 2001. A Hierarchical Unsupervised Growing Neural Network for Clustering Gene Expression Patterns. Bioinformatics. 17(2), 126-136.

Human Genome Program in U.S. Department of Energy. 1992. http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml (accessed 10/03/2007)

Jiampojamarn, S., Cercone, N., and Keselj, V. 2005. Biological Named Entity Recognition Using N-grams and Classification Methods. Proceeding of The Conference Pacific Association for Computational Linguistics (PACLING'05), Meisei University, Hino Campus, Hino-shi, Tokyo, Japan. N/A.

Jiang, N., Wu, W.X., and Mitchell, I. 2005. Protein Fold Recognition by Mixed Environment Specific Amino Acid Substitution Mapping Using Neural Networks. Proceeding of The 11th InternationalConference on Parallel and Distributed Systems-Workshops (ICPADS'05). Fukuoka Institute of Technology (FIT), Fukuoka, Japan. pp. 341-345.

Kesmir, C. 2007. Course: Bioinformatics Pattern Analysis. Utrecht University.

Kim, J.Y., and Shawe-taylor, J. 1994. Fast String Matching Using An N-gram Algorithm. Software---Practice and Experience. 21(1), 79-88.

Kim, K.J., and Cho, S.B. 2004. Prediction of Colon Cancer Using Evolutionary Neural Network. Neurocomputing, 61, 361-379.

Kimball, R., and Ross, M. 2002. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. 2nd Edition. Wiley Computer Publishing: United States of America.

Kohavi, R., and John, G.H. 1997. Wrappers for Feature Subset Selection. Journal of Artificial Intelligence. 97(1-2), 273–324.

Litwin, W., and Mokadem, R. 2007. Fast nGram-based String Search Over Data Encoded Using Algebraic Signatures. Proceeding of The 33rd International Conference on Very Large Data Bases (VLDB ’07). Vienna, Austria, September 23-27, 2007. pp. 207-218.

Lumini, A., and Nanni, L. 2006. Identifying Splice-Junction Sequences by Hierarchical Multiclassifier. Artificial Intelligence and Applications. 27, 1390-1396.

M.A.H. Zahid, Ankush Mittal, and Joshi, R. C. 2005a. Use of Phylogenetic Network and Its Reconstruction Algorithms, Journal of Bioinformatics, India. 4, 47-58.

M.A.H. Zahid, Ankush Mittal, and Joshi, R. C. 2005b. Least Common Ancestor Based Efficient Method for Constructing Rooted Supertrees, Journal of Bioinformatics and Biomedical Engineering. 1(5), 1-6.

M.A.H. Zahid, Ankush Mittal, and Joshi, R. C. 2005c. Pattern Recognition Based Approach for Phylogenetic Network Construction with Constrained Recombination, Journal of Pattern Recognition. 39, 2312-2322.

Ma, Q., Wang, J.T.L., Shasha, D., and Wu, C.H. 2001. DNA Sequence Classification Via An Expectation Maximization Algorithm and Neural Networks: A Case Study. Systems, Man,and Cybernetics, IEEE Transactions. 31(4), 468-475.

Ma, Z., Zhou, C., Lu, L., Ma, Y., Sun, P., and Cui, Y. 2007. Predicting Protein-Protein Interactions Based on BP Neural Network. Proceeding of Bioinformatics and Biomedicine Workshops 2007 (BIBMW'2007). November 2-4, 2007. pp. 3-7.

Mahonya, S., Benosa, P.V., Smithd, T.J., and Goldend, A. 2006. Self-Organizing Neural Networks to Support The Discovery of DNA-Binding Motifs. Neural Networks. 19, 950–962.

Mhamdi, F., Rakotomalala, R., and Elloumi, M. 2006a. A Hierarchical N-grams Extraction Approach for Classification Problem. Proceeding of International Conference on Signal-ImageTechnology and Internet-Based System (SITIS'06). Hammamet, Tunisia, December 17-21, 2006. pp. 310-321.

Mhamdi, F., Rakotomalala, R., and Elloumi, M. 2006b. A Compromise Between N-gram Length and Classifier Characteristics for Protein Classification, International Journal of Computer Science and Network Security. 6(4), 82-87.

Mitra, S., and Acharya, T. 2003. Data Mining: Multimedia, Soft Computing and Bioinformatics. Wiley: New Jersey.

Nanni, L. 2005. Hyperplances for Predicting Protein-Protein Interactions. Neurocomputing, Elsevier. 69(1-3), 257-236.

Nanni, L. 2006a. Comparison Among Feature Extraction Methods for HIV-1 Protease Cleavage Site Prediction. Neurocomputing, Elsevier. 39(4), 495-497.

Nanni, L. 2006b. A Novel Ensemble of Classifiers for Protein Fold Recognition. Neurocomputing, Elsevier. 69(16-18), 2434-2437.

Nanni, L., and Lumini, A. 2006. A Reliable Method for HIV-1 Protease Cleavage Site Prediction Neurocomputing, Elsevier. 69, 838-841.

Natalia, V., and Cathy, H. 2006. Prediction of Catalytic Residues Using Support Vector Machine with Selected Protein Sequence and Structural Properties. BMC Bioinformatics. 7, 312-324.

National Center for Biotechnology Information. 2007. http://www.ncbi.nlm.nih.gov (accessed 12/12/2007)

Needleman, S.B., and Wunsch C.D. 1970. A General Method Applicable to The Search for Similarities in The Amino Acid Sequence of Two Proteins. Journal of Molecular Biology. 48, 443-453.

Nikolajewa, S., Pudimat, R., Hiller, M., Platzer, M., and Backofen, R. 2007. BioBayesNet: A Web Server for Feature Extraction and Bayesian Network Modeling of Biological Sequence Data. Nucleic Acids Research. 35(Web-Server-Issue), 688-693.

Nilsson, R., Bjorkegren, J., and Tegner, J. 2007. Consistent Feature Selection for Pattern Recognition in Polynomial Time. Journal of Machine Learning Research. 8, 589-612.

Pereira, M.S., Andrade, L., Difrawy S.E., Karger, B.L., and Manolakos, E.S. 2000. Statistical Learning Formulation of The DNA Base-Calling Problem and Its Solution in A Bayesion EM Framework. Discrete Applied Mathematics. 104(1-3), 229-258.

Peterson, L.E., Ozen, M., Erdem, H., Amini, A., Gomez, L., Nelson, C.C., and Ittmann, M. 2005. Artificial Neural Network Analysis of DNA Microarray-Based Prostate Cancer Recurrence. Proceeding of The 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, November 14-15, 2005. pp.1-8.

Poddar, A., Chandra, N., Ganapathiraju, M., Sekar, K., Klein-Seetharaman, J., Reddy, R., and Balakrishnan, N. 2007. Evolutionary Insights from Suffix Array Based Genome Sequence Analysis. Journal of Biosciences. 33, 871-881.

Rich, E., and Knight, K. 1991. Artificial Intelligence, 2nd Edition. McGraw-Hill Book Company: N/A.

Roiger, R., and Geatz, M. 2003. Data Mining: A Tutorial Based Primer. Addison Wesley: N/A.

Salzberg, S., 1995. Locating Protein Coding Regions in Human DNA Using A Decision Tree Algorithm. Journal of Computational Biology. 2, 473-485.

Salzberg, S., Delcher, A.L., Fasman, K.H., and Henderson, J. 1998. A Decision Tree System for Finding Genes in DNA. Journal of Computational Biology. 5, 667-680.

Smith, T.F., and Waterman, M.S. 1981. Identification of Common Molecular Subsequences. Journal of Computational Biology. 147, 195-197.

Tan, A., and Pan, H. 2005. Predictive Neural Networks for Gene Expression Data Analysis. Neural Networks, 18(3), 297-306.

Thompson, J.D., Higgins, D.G., and Gibson T.J. 1994. CLUSTAL W: Improving The Sensitivity of Progressive Multiple Sequence Alignment Through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. Nucleic Acids Research. 22(22), 4673-4680.

Wang, D., Lee, N.K., DILLON, T.S. , HOOGENRAAD, N.J., Bob, M., and John, S. 2002. Protein Sequences Classification Using Modular Radial Basis Function Neural Networks. Proceeding of The 9th International Conference on Neural Information Proceeding. Orchid Country Club, Singapore, November 18-22, 2002. pp. 764-768.

Wang, J.T.L., Ma, Q., Shasha, D., and Wu, C.H. 2000. Application of Neural Networks to Biological Data Mining: A Case Study in Protein Sequence Classification. Proceeding of The 6th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining. Massachusetts, United States. July 23-26, 2000, pp. 305-309.

Wiemann, S., Weil, B., and Wellenreuther, R. 2001. Toward a Catalog of Human Genes and Proteins: Sequencing and Analysis of 500 Novel Complete Protein Coding Human cDNAs. Genome Research. 11(3), 422-435.

Wikipedia. 2007.
http://en.wikipedia.org/wiki/Neron (accessed 01/03/2007)

Witten, I.H., and Frank, E. 2005. Data Mining: Practical Machine Learning Tools and Technique. 2nd Edition. Elsevier: N/A.

Wren, J.D., Hildebrand, W.H., Chandrasekaran, S., and Melcher U. 2005. Markov Model Recognition and Classification of DNA/Protein Sequences Within Large Text Databases. Bioinformatics. 21(21). 4046-4053.

Wu, C.H., Ermongkonchai, A., and Chang, T. 1991. Protein classification Using A Neural Network Protein Database System. Proceeding of Analysis of Neural Net Applications Conference. Fairfax, Virginia, United States, May 29-31, 1991. pp.29-41.

Wu, C.H., Whitson, C., Mclarty, J., Ermongkonchai, A., and Chang, T.C. 1992a. Protein Classification Artificial Neural System. Protein Science, 1(5), 667-677.

Wu, C.H., Whitson, G.M., Hsiao, C.T., and Huang, C.F. 1992b. Classification Artificial Neural Systems for Genome Research. Proceeding of The 1992 ACM/IEEE Conference onSupercomputing. Minnesota, United States, November 16-20, 1992, pp. 797-803.

Wu, Cathy H., and McLarty, J. 2000. Neural Networks and Genome Informatics. Methods in Computational Biology and Biochemistry, Elsevier Science: N/A.

Wu, Cathy H., Zhao, S., Simmons, K., and Shivakumar, S. 1997. Motif Neural Network Design Large-Scale Protein Family Identification. Proceeding of 1997 International Conference on Neural Networks. Houston, TX, USA, January 9-12, 1997. pp. 86-89.

Xie, Q., Ratnasinghe, L.D., Hong, H., Perkins, R., Tang, Z.-Z., Hu, N., Taylor, P.R., and Tong, W. 2005. Decision Forest Analysis of 61 Single Nucleotide Polymorphisms in a Case-Control Study of Esophageal Cancer; A Novel Method. BMC Bioinformatics. 6(2), S4.

Zhang, H., and Kinsner, W. 2001. Feature Extraction From DNA Sequences by Multifractal Analysis Proceeding of The 23rd Annual EMBS International Conference of the IEEE, N/A, pp. 1567-1572.

Zhang, L.V., Wong, S.L., King, O.D., and Roth, F.P. 2004. Predicting Co-Complexed Protein Pairs Using Genomic and Proteomic Data Integration. BMC Bioinformatics. 5(38), N/A.

Zhong, W., He, J., Harrison, R., Tai, P.C., and Pan, Y. 2007. Clustering Support Vector Machines for Protein Local Structure Prediction. Expert System. 32, 518-526.

ภาคผนวก ก

คู่มือผู้ใช้ระบบการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

คู่มือผู้ใช้ระบบการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม จะประกอบด้วยส่วนของการอธิบายการใช้งาน 3 ส่วน คือ การใช้งานโปรแกรมเตรียมข้อมูลค่าความถี่ การใช้งานโปรแกรมสกัดลักษณะเฉพาะ และการใช้งานโปรแกรมจำแนกสายพันธุกรรม ซึ่งสามารถอธิบายได้ตามลำดับดังนี้

1. การเรียกใช้งานเมนูหลักโปรแกรมระบบการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

ในการเรียกใช้งานเมนูหลักของโปรแกรมระบบการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรมนั้น โดยไฟล์โปรแกรมจะถูกบันทึกในชื่อ Mainmenu.fig เมื่อต้องการเรียกใช้งานให้ทำการเปิดไฟล์โปรแกรมดังกล่าว จากนั้นกดปุ่ม Run บนแถบเครื่องมือของโปรแกรม Matlab จากนั้นหน้าต่างเมนูหลักของโปรแกรมระบบการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรมจะปรากฏขึ้น ดังแสดงในภาพประกอบ ก.1

ภาพประกอบ ก.1 เมนูหลักโปรแกรมระบบการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

จากภาพประกอบ ก.1 เมนูหลักของโปรแกรมจะถูกสร้างในลักษณะของแท็บเมนู (Tab Panel) ซึ่งประกอบด้วย 4 เมนูหลัก คือ

1) เมนูหลักที่ 1 ทำหน้าที่เรียกใช้โปรแกรมการเตรียมข้อมูลค่าความถี่ (Data Preprocessing Program)

2) เมนูหลักที่ 2 ทำหน้าที่เรียกใช้โปรแกรมการสกัดลักษณะเฉพาะบนสายพันธุกรรม (Feature Extraction Program) สำหรับสกัดหาลักษณะเฉพาะ (Feature)

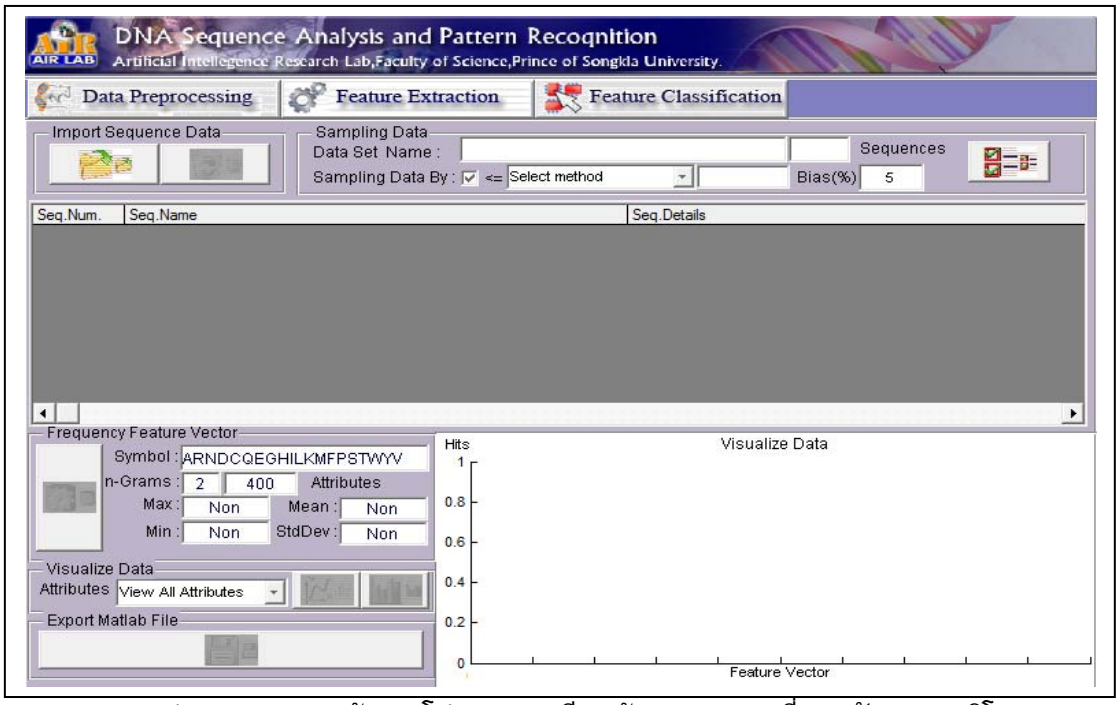
3) เมนูหลักที่ 3 ทำหน้าที่เรียกใช้โปรแกรมการจำแนกสายพันธุกรรมด้วยลักษณะเฉพาะ (Feature Classification Program)

4) เมนูหลักที่ 4 ทำหน้าที่สำหรับจบการทำงานโปรแกรม

ทั้งนี้จะเห็นว่าโปรแกรมระบบการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรมสามารถแบ่งงานหลักออกเป็น 3 ส่วน ซึ่งช่วยให้ผู้ใช้มีความสะดวกในการทำงานมากขึ้น เนื่องจากผู้ใช้อาจเลือกทำงานเฉพาะบางขั้นตอนของโปรแกรมและสามารถนำผลลัพธ์ที่ได้ไปใช้งานร่วมกับโปรแกรมอื่นๆ เช่น โปรแกรม Microsoft Excel โปรแกรม NotePad หรือ EditPlus เป็นต้น โดยจะอธิบายรายละเอียดการใช้งานในส่วนถัดไป

2. การเรียกใช้งานโปรแกรมเตรียมข้อมูลค่าความถี่คู่ลำดับกรดอะมิโน

ผู้ใช้สามารถเรียกใช้งานโปรแกรมเตรียมข้อมูลค่าความถี่ ได้จากเมนู Data Preprocessing จากหน้าต่างเมนูหลักของระบบ ซึ่งมีหน้าต่างโปรแกรมดังภาพประกอบ ก.2



ภาพประกอบ ก.2 หน้าต่างโปรแกรมเตรียมข้อมูลค่าความถี่คู่ลำดับกรดอะมิโน

จากภาพประกอบ ก.2 ประกอบด้วยส่วนหลัก ๆ คือ ส่วนของการนำเข้าข้อมูลสายพันธุกรรม และการสุ่มตัวอย่างข้อมูล (Sampling Data) และส่วนของการแสดงค่าข้อมูลความถี่ของสายพันธุกรรมที่นำเข้าสู่โปรแกรมวิเคราะห์รูปแบบลำดับพันธุกรรม โดยจะแสดงค่าสถิติ เช่น ค่าสูงสุด (Maximum) ค่าต่ำสุด (Minimum) ค่าเฉลี่ยเลขคณิต (Mean) ค่าส่วนเบี่ยงเบนมาตรฐาน (Standard Derivation) และแผนภูมิแสดงค่าความถี่ของคู่ลำดับกรดอะมิโนที่คำนวณได้

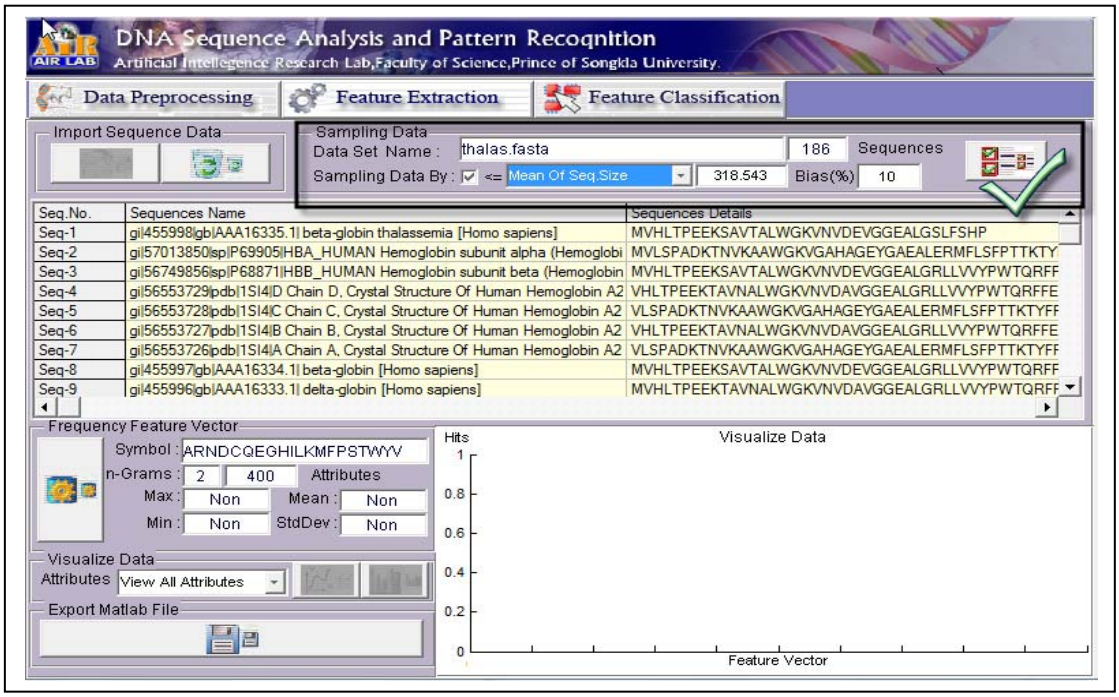
ขั้นตอนการเตรียมข้อมูลค่าความถี่ของคู่ลำดับกรดอะมิโน มีรายละเอียดดังนี้

1) การอ่านข้อมูลสายพันธุกรรมเข้าสู่ระบบไฟล์ข้อมูลนำเข้าใช้ไฟล์ข้อมูลซึ่งถูกบันทึกในรูปแบบของไฟล์ FASTA (*.fasta) โดยกดปุ่ม  (Import) และทำการเลือกไฟล์ข้อมูลที่ต้องการ

2) หลังจากโปรแกรมทำการอ่านข้อมูลเข้าสู่ระบบแล้ว ผู้ใช้สามารถทำความสะอาดข้อมูล (Data Cleaning) ด้วยเทคนิคการสุ่มตัวอย่าง (Sampling) โดยใช้ค่าความยาวเฉลี่ยของสายพันธุกรรมของข้อมูลที่ใช้ ซึ่งมีเงื่อนไข คือ

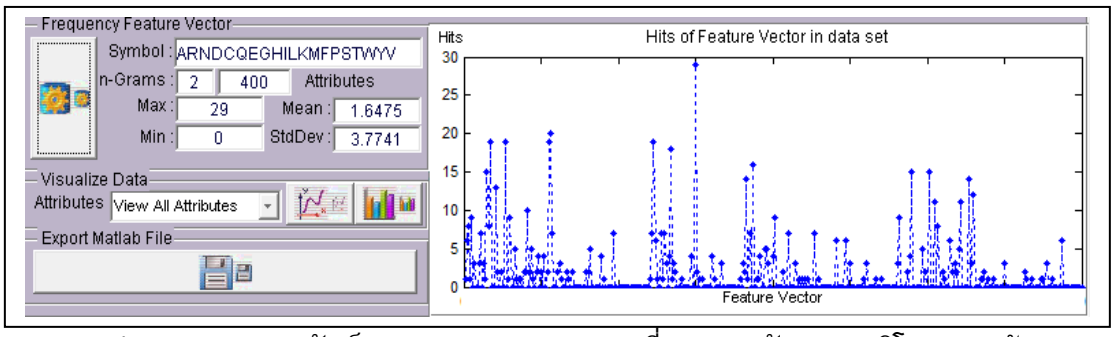
- เลือกข้อมูลจากที่มีค่าน้อยกว่าหรือเท่ากับค่าความยาวเฉลี่ย
- เลือกข้อมูลจากที่มีค่ามากกว่าค่าความยาวเฉลี่ย

และนอกจากนี้ผู้ใช้ยังสามารถค่าความเอนเอียง (Bias) ในเลือกได้ เช่น ±5% ของค่าความยาวเฉลี่ย จากนั้นกดปุ่ม Sampling ดังแสดงในภาพประกอบ ก.3



ภาพประกอบ ก.3 การสุ่มตัวอย่างข้อมูลด้วยค่าเฉลี่ยของสายพันธุกรรมที่ใช้

3) เมื่อผู้ใช้ทำการเลือกกลุ่มตัวอย่าง หรืออาจไม่มีการเลือกกลุ่มตัวอย่างได้ตามความต้องการของผู้ใช้ การคำนวณค่าความถี่ของคู่ลำดับอะมิโน ซึ่งปรากฏอยู่บนสายพันธุกรรมแต่ละสายในชุดข้อมูลนำเข้าทำโดยการกดปุ่ม  (Execute) จากนั้นรอจนกระทั่งโปรแกรมทำการคำนวณค่าความถี่ของคู่ลำดับกรดอะมิโนจนทุกสายพันธุกรรมที่มีการนำเข้า และเมื่อทำการคำนวณเรียบร้อยโปรแกรมจะแสดงค่าลักษณะของค่าความถี่คู่ลำดับกรดอะมิโนบนชุดข้อมูลด้วยแผนภาพ Histograms โดยใช้ค่าความนิยม (Hits Ratio) ของคู่ลำดับแต่ละคู่ดังแสดงในภาพประกอบ ก.4



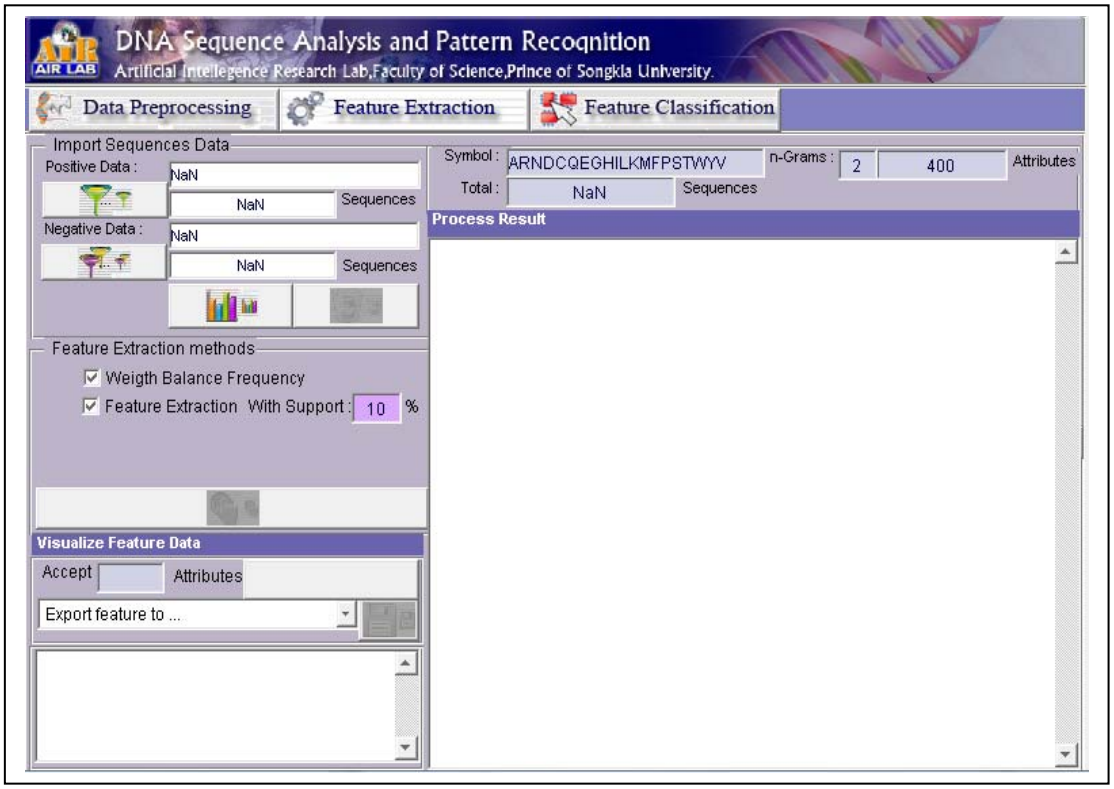
ภาพประกอบ ก.4 ผลลัพธ์จากการคำนวณค่าความถี่ของคู่ลำดับกรดอะมิโนบนชุดข้อมูล

จากภาพประกอบ ก.4 นอกจากค่าความนิยมแล้ว โปรแกรมจะแสดงค่าความถี่สูงสุดของคู่ลำดับ (Max) ค่าเฉลี่ยของคู่ลำดับ (Mean) ค่าส่วนเบี่ยงเบนมาตรฐาน (StdDev) และนอกจากนี้ผู้ใช้ยังสามารถเลือกดูลักษณะของค่าความถี่ของคู่คำลับกรดอะมิโนแต่ละคู่ได้จากปุ่ม Visualize Data โดยสามารถเลือกชนิดของแผนภูมิที่จะใช้ได้ 2 ชนิด คือ แบบแท่ง (Bar) และแบบเส้น (Line)

4) ขั้นตอนการบันทึกข้อมูล กรณีที่ผู้ใช้ต้องการบันทึกข้อมูลค่าความถี่ของคู่ลำดับที่คำนวณได้ สามารถทำได้โดยกดปุ่ม  (Export) โดยบันทึกไฟล์ในลักษณะของตารางข้อมูลของ Matlab มีนามสกุลเป็น *.mat ซึ่งเป็นขั้นตอนสุดท้ายของโปรแกรมการเตรียมข้อมูลค่าความถี่ของคู่ลำดับกรดอะมิโน

3. การเรียกใช้งานโปรแกรมสกัดลักษณะเฉพาะบนสายพันธุกรรม

การเรียกใช้งานโปรแกรมสกัดลักษณะเฉพาะบนสายพันุกรรม สามารถเรียกใช้จากเมนู Feature Extractionประกอบด้วยส่วนต่าง ๆ คือ ส่วนของการนำเข้าข้อมูลค่าความถี่คู่ลำดับกรดอะมิโนของสายพันธุกรรมในชุดข้อมูลนำเข้า ส่วนของการสกัดลักษณะเฉพาะบนสายพันธุกรรม และส่วนของการแสดงค่าสถิติค่าความถี่ของสายพันธุกรรมที่นำเข้าสู่โปรแกรมวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม ดังแสดงในภาพประกอบ ก.5

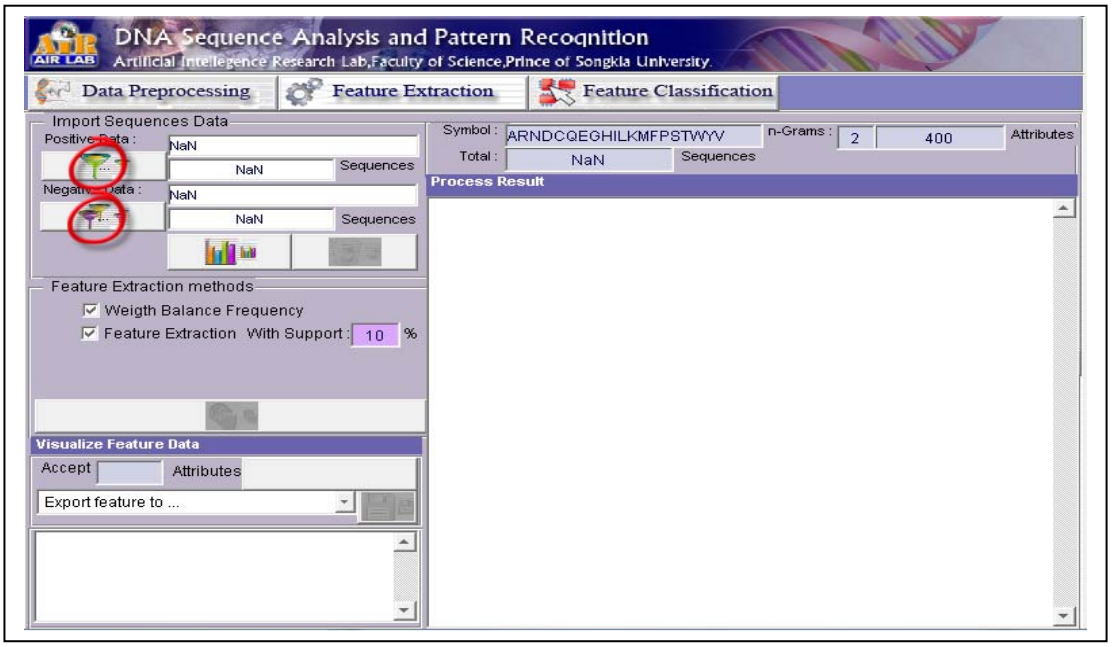


ภาพประกอบ ก.5 หน้าต่างโปรแกรมสกัดลักษณะเฉพาะบนสายพันธุกรรม

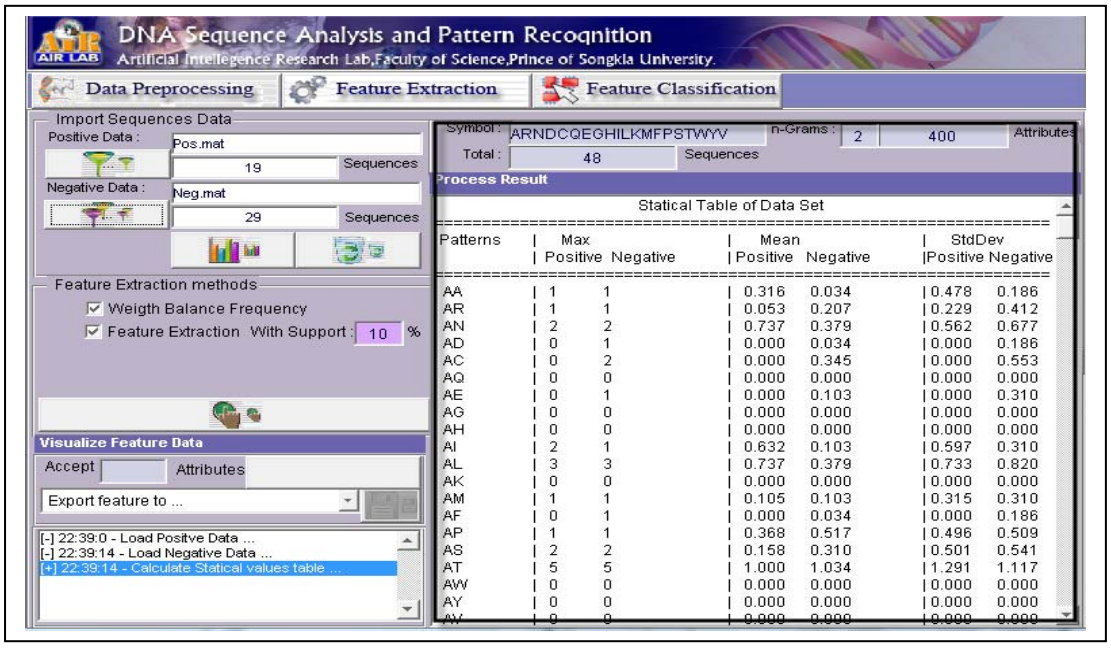
จากภาพประกอบ ก.5 ในส่วนของการแสดงค่าสถิติซึ่งผ่านขั้นตอนการลักษณะเฉพาะ (Feature Extraction) ประกอบด้วยค่าสถิติ เช่น ค่าความนิยม (Hits Ratio) ของข้อมูลที่แสดงอาการของโรค (Positive Data) กลุ่มข้อมูลที่ไม่แสดงอาการของโรค (Negative Data) และแผนภูมิแสดงค่าสถิติของคู่ลำดับกรดอะมิโนที่สกัดได้

ขั้นตอนการสกัดลักษณะเฉพาะบนสายพันธุกรรม มีรายละเอียดดังนี้

1) การนำเข้าข้อมูลค่าความถี่ของคู่ลำดับกรดอะมิโน ซึ่งข้อมูลค่าความถี่ที่นำเข้าสำหรับการสกัดลักษณะเฉพาะประกอบด้วย 2 ลักษณะ คือ ข้อมูลซึ่งแสดงอาการของโรค (Positive Data) และ กลุ่มที่ไม่แสดงอาการของโรค (Negative Data) สามารถทำได้โดยการกดปุ่ม Import Sequence Data ดังแสดงในภาพประกอบ ก.6 และ ก.7 ตามลำดับ

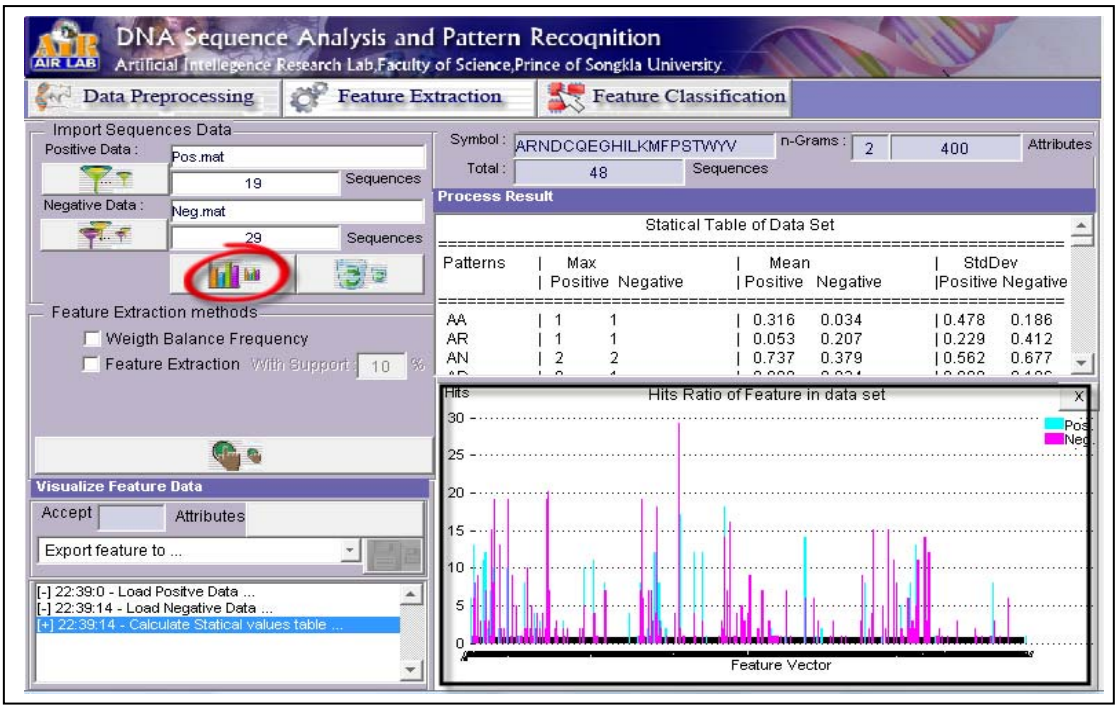


ภาพประกอบ ก.6 การนำเข้าข้อมูลค่าความถี่คู่ลำดับกรดอะมิโน



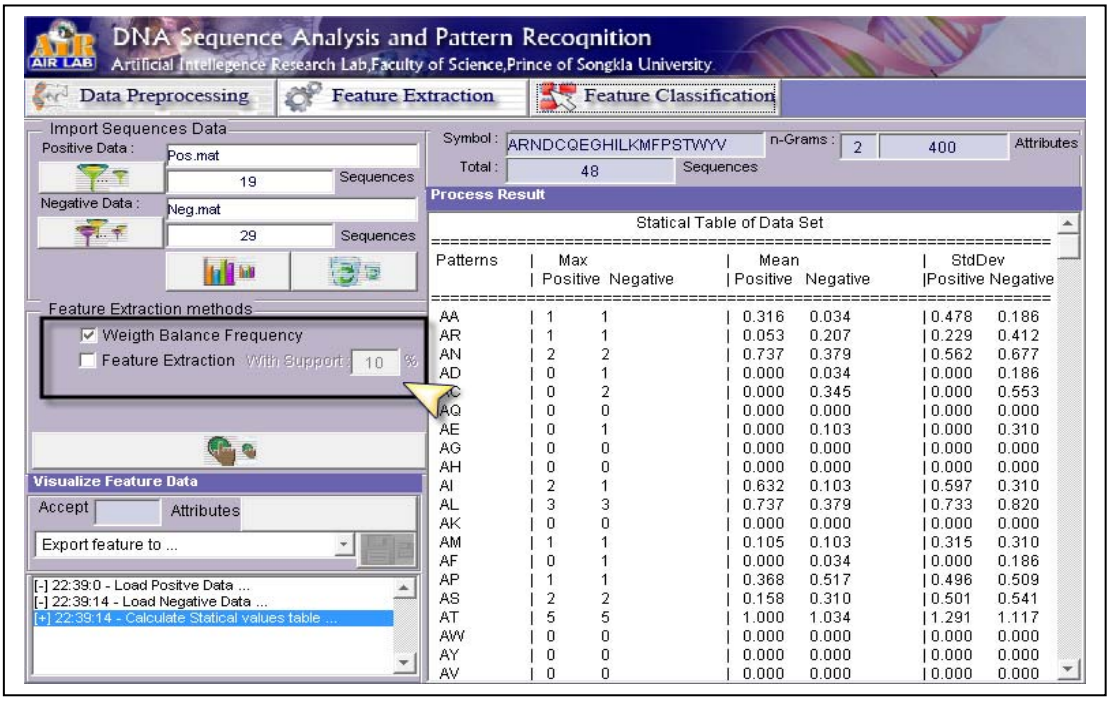
ภาพประกอบ ก.7 ค่าสถิติของข้อมูลค่าความถี่กรดอะมิโนในชุดข้อมูลนำเข้า

จากภาพประกอบ ก.6 และ ก.7 หลังจากมีการนำเข้าข้อมูลทั้ง 2 กลุ่ม โปรแกรม จะแสดงค่าความนิยมของทั้ง 2 ค่าสถิติของคู่ลำดับทั้ง 400 คู่ลำดับ ผู้ใช้สามารถทำการกดปุ่ม View Graph กรณีที่ต้องการดูลักษณะของข้อมูล ดังแสดงในภาพประกอบ ก.8



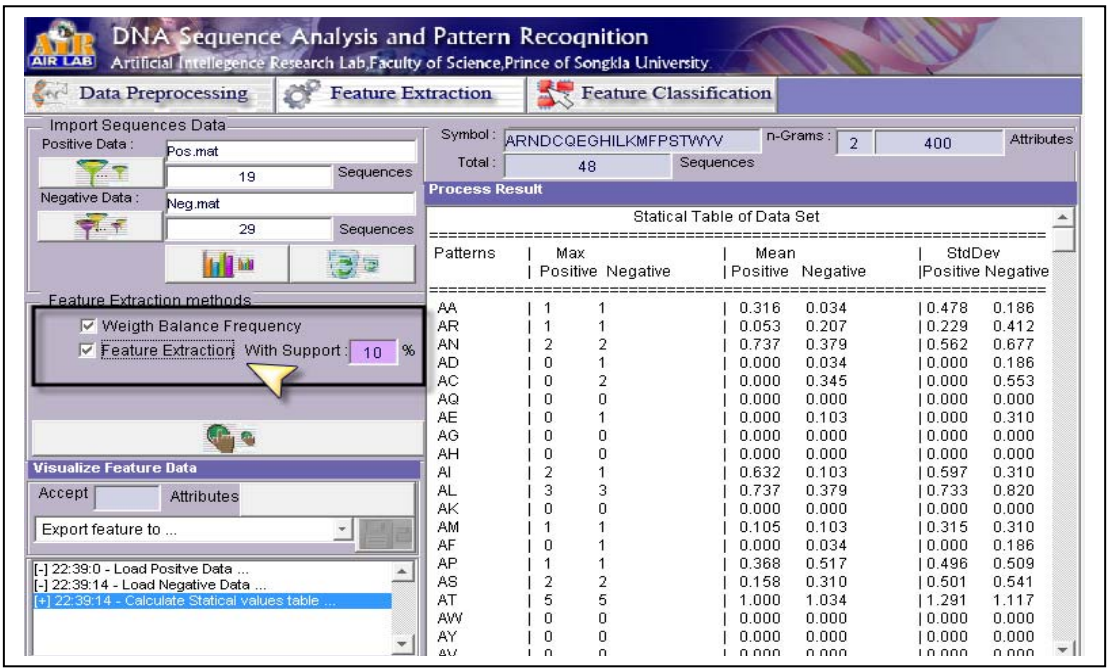
ภาพประกอบ ก.8 แผนภูมิลักษณะของข้อมูลนำเข้า

2) การปรับค่าน้ำหนักของค่าความถี่คู่ลำดับกรดอะมิโน สามารถทำได้โดยการกดปุ่มรายการ Weight Balance ผลลัพธ์แสดงดังภาพประกอบ ก.9



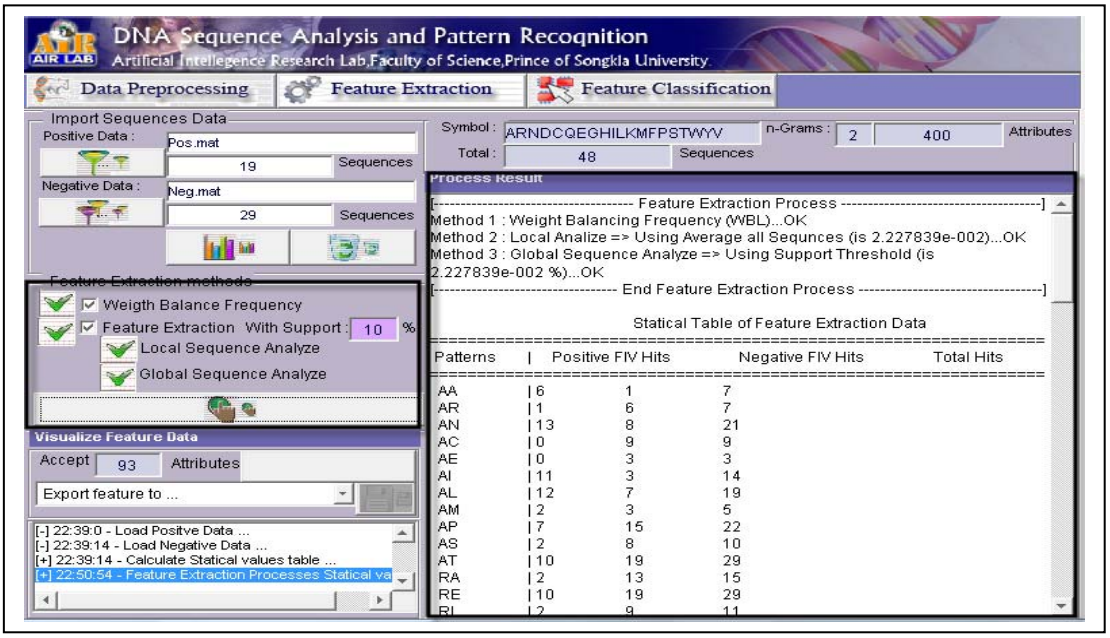
ภาพประกอบ ก.9 การเลือกปุ่มรายการปรับค่าน้ำหนักข้อมูลค่าความถี่

3) ทำการสกัดลักษณะเฉพาะบนสายพันธุกรรม (Feature Extraction) ดังแสดงในภาพประกอบ ก.10



ภาพประกอบ ก.10 การสกัดลักษณะเฉพาะบนสายพันธุกรรม

จากภาพประกอบ ก.10 ผลลัพธ์สุดท้ายของขั้นตอนการสกัดลักษณะเฉพาะคือคู่ลำดับซึ่งผ่านตามเงื่อนไข ที่กำหนดไว้ กล่าวคือมีค่า FPV มากกว่าหรือเท่ากับ 10% และสามารถแสดงผลลัพธ์จากการสกัดลักษณะเฉพาะ ดังภาพประกอบ ก.11

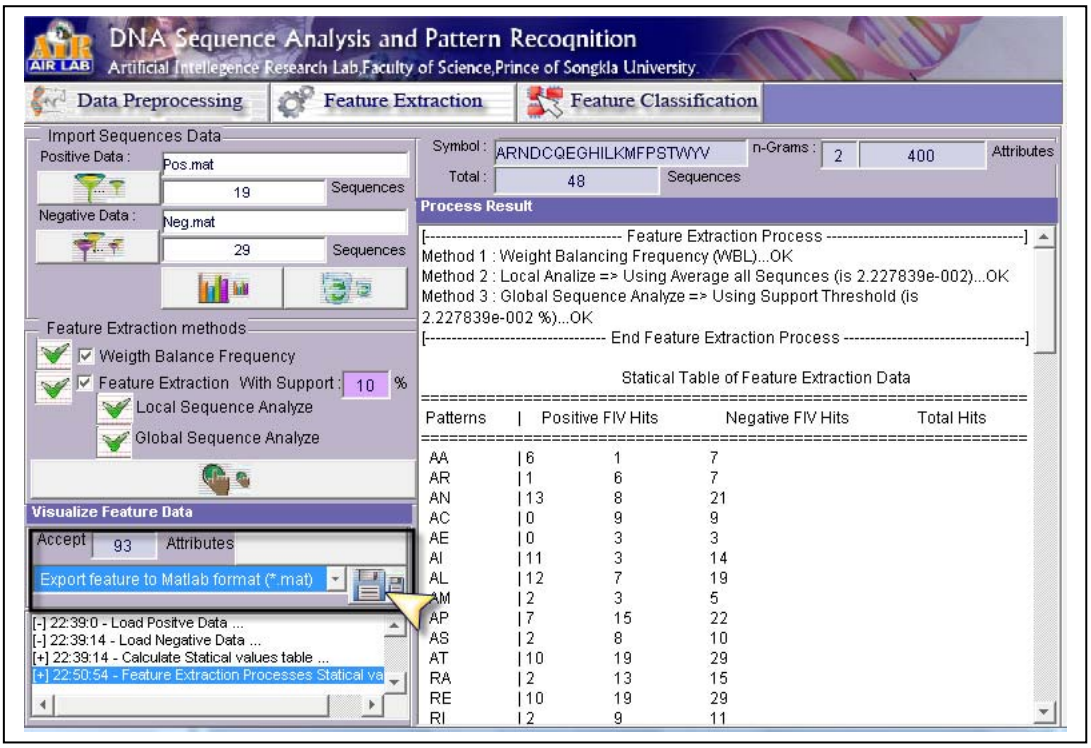


ภาพประกอบ ก.11 ผลลัพธ์จากการสกัดลักษณะเฉพาะบนสายพันธุกรรม

5) การบันทึกผลข้อมูลลักษณะเฉพาะที่สกัดได้ โดยสามารถเลือกบันทึกข้อมูลได้ 3 ลักษณะ คือ

- เก็บในรูปแบบไฟล์ข้อความ (Text File) โดยมีนามสกุล *.txt ซึ่งสามารถนำไปใช้งานบนโปรแกรม Microsoft Excel, EditPlus และ NotePad เป็นต้น
- เก็บในรูปแบบไฟล์ ARFF ซึ่งสามารถนำไปใช้งานร่วมกับโปรแกรม WEKA โดยมีนามสกุล *.arff
- เก็บในรูปแบบไฟล์ตารางข้อมูลของ Matlab ซึ่งมีนามสกุลเป็น *.mat

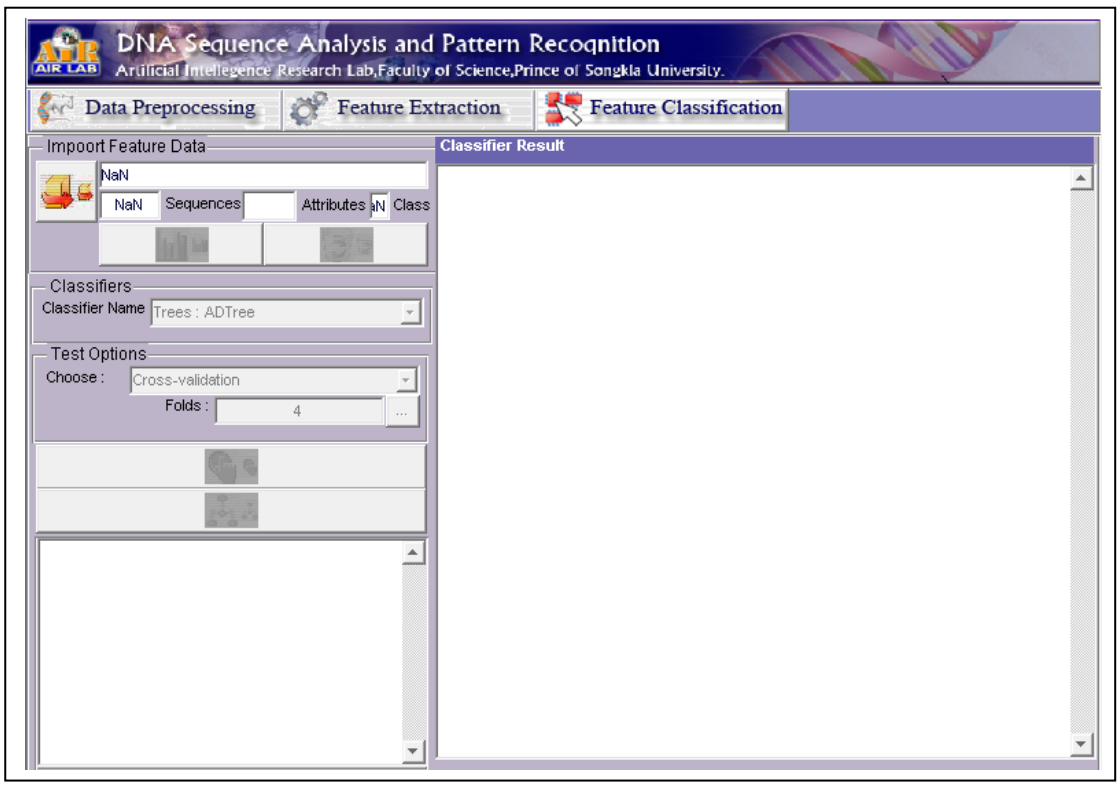
โดยในขั้นตอนนี้ จะขอเลือกบันทึกไฟล์เป็นไฟล์ตารางข้อมูลของ Matlab ซึ่งมีนามสกุลเป็น *.mat เพื่อนำไปใช้งานในโปรแกรมการจำแนกสายพันธุกรรมด้วยลักษณะเฉพาะ โดยการกดปุ่ม  ดังแสดงในภาพประกอบ ก.12



ภาพประกอบ ก.12 การบักทึกผลลัพธ์การสกัดลักษณะเฉพาะ

4. การเรียกใช้งานโปรแกรมการจำแนกสายพันธุกรรมด้วยลักษณะเฉพาะ

ในการจำแนกสายพันธุกรรมผู้ใช้สามารถเรียกใช้โปรแกรม ได้จากเมนู Feature Classification ดังแสดงหน้าต่างโปรแกรมในภาพประกอบ ก.13

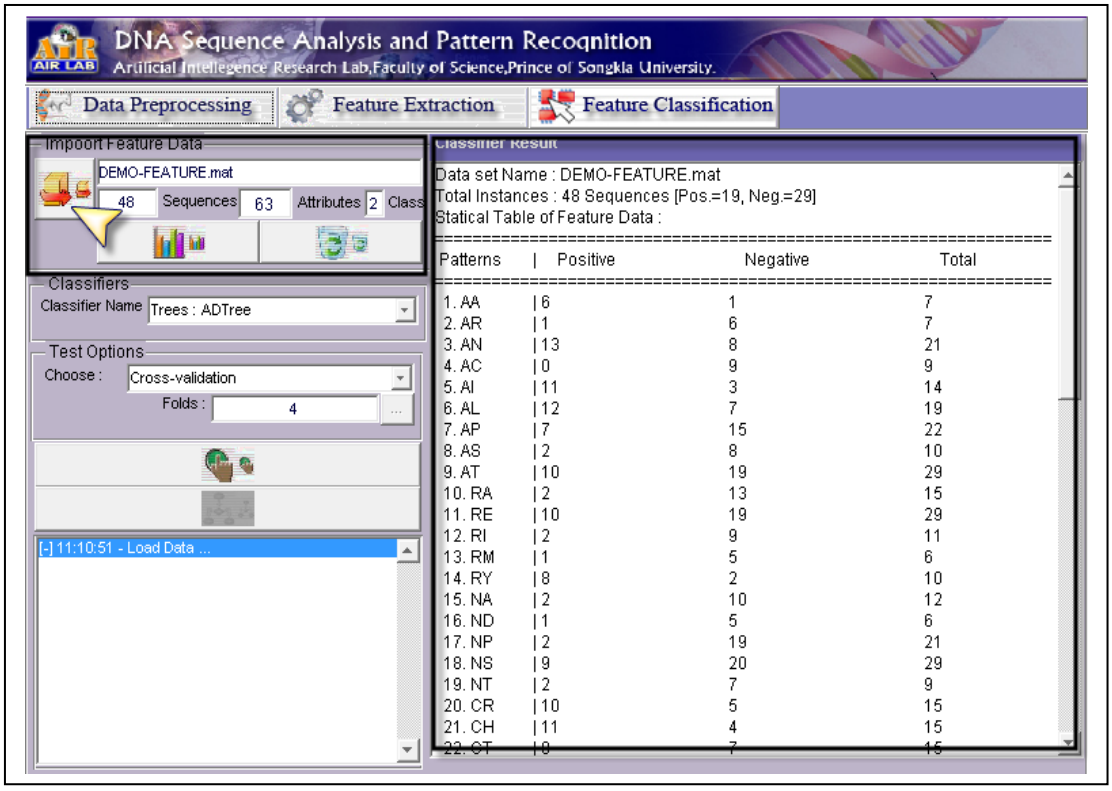


ภาพประกอบ ก.13 หน้าต่างโปรแกรมการจำแนกสายพันธุกรรมด้วยลักษณะเฉพาะ

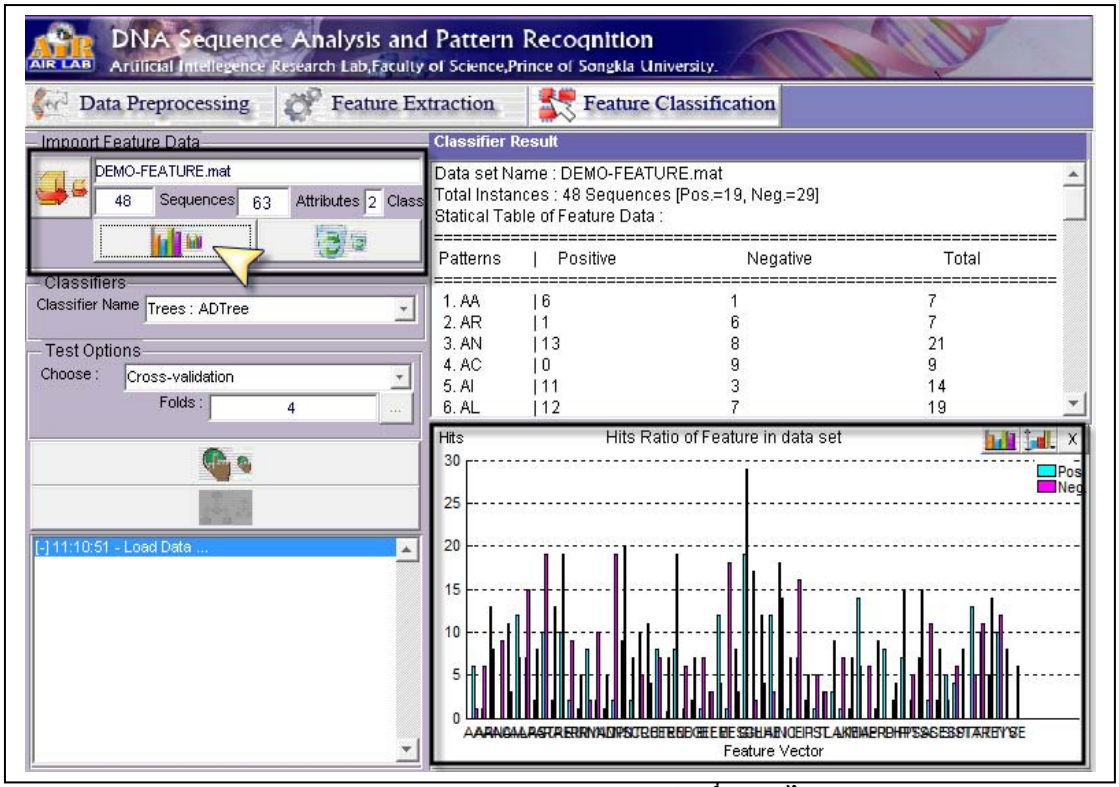
จากภาพประกอบ ก.13 ประกอบด้วย 3 ส่วนหลัก คือ ส่วนของการนำเข้าค่าน้ำหนักของคู่ลำดับที่สกัดลักษณะเฉพาะได้ของข้อมูลสายพันธุกรรมทุกสายในชุดข้อมูลนำเข้า ส่วนของการเลือกขั้นตอนวิธี (Algorithms) ในการจำแนกสายพันธุกรรม และส่วนของการแสดงค่าสถิติของข้อมูลและผลลัพธ์ที่ได้จากการจำแนกข้อมูลสายพันธุกรรมด้วยลักษณะเฉพาะ ซึ่งประกอบด้วยค่าความถูกต้อง (Accuracy) ตารางการจำแนก (Confusion Matrix) เป็นต้น

ขั้นตอนในการจำแนกสายพันธุกรรมด้วยลักษณะเฉพาะ มีรายละเอียดดังนี้

1) อ่านค่าน้ำหนักของคู่ลำดับกรดอะมิโนซึ่งสกัดได้ของสายพันธุกรรมทุกสายในไฟล์ข้อมูล โดยเมื่อทำการนำเข้าข้อมูลโปรแกรมจะแสดงลักษณะของข้อมูลด้วยค่าความนิยม ดังแสดงในภาพประกอบ ก.14 และ ก.15 ตามลำดับ



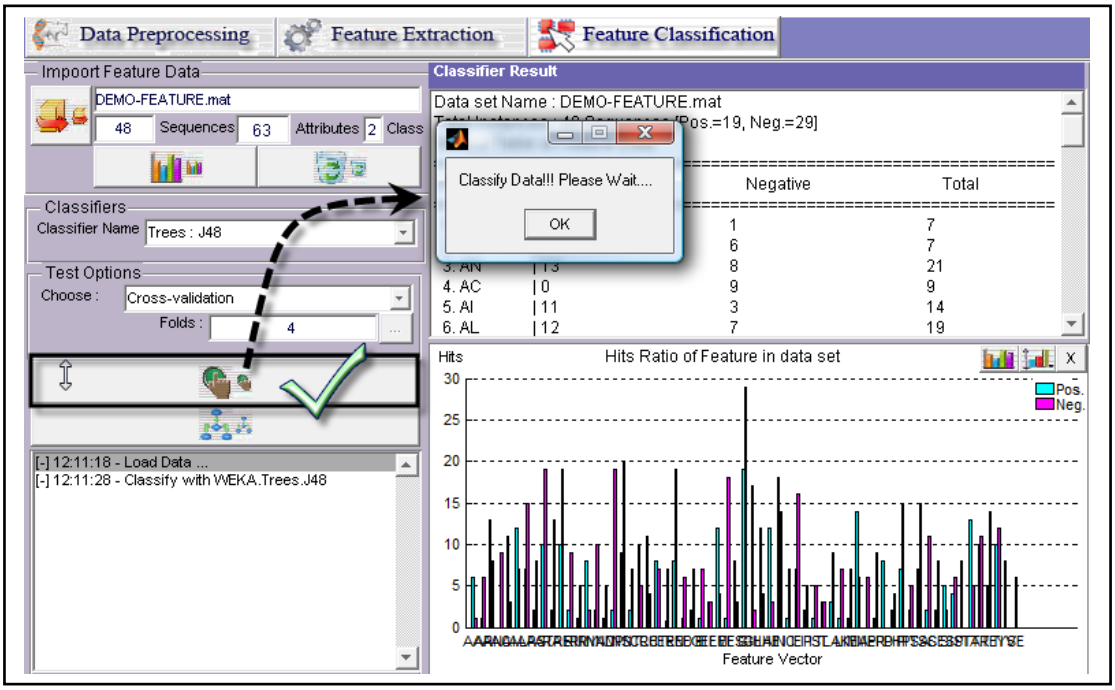
ภาพประกอบ ก.14 ค่าความนิยมของคู่ลำดับลำดับกรดอะมิโนซึ่งสกัดได้บนสายพันธุกรรม



ภาพประกอบ ก.15 แผนภูมิค่าความนิยมของคู่ลำดับซึ่งสกัดได้บนสายพันธุกรรม

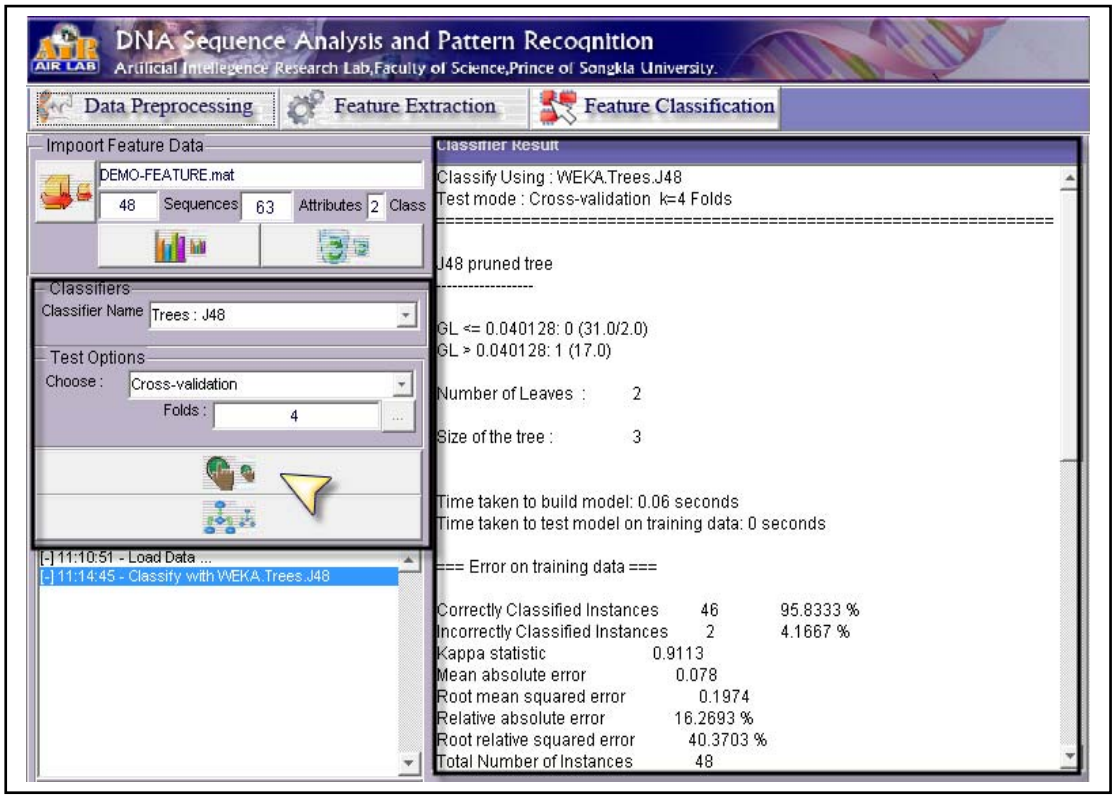
2) เลือกขั้นตอนวิธีสำหรับการจำแนกข้อมูลสายพันธุกรรม ประกอบด้วย โครงข่ายประสาทเทียบแบบหลายชั้น (Multilayer Perceptron: MLP) แบบเรเดียลเบสิสก์ฟังชัน (Radial Basis Function: RBF) และต้นไม้การตัดสินใจแบบ Alternating Decision Tree (ADTree) แบบ Pruned C4.5 Decision Tree (J48) และแบบ K Random Features Decision Tree (RandomTree) ตามลำดับ

หลังจากเลือกขั้นตอนวิธีที่ต้องการ จำเป็นต้องมีการกำหนดจำนวนแบ่งกลุ่มข้อมูลสำหรับการทดสอบแบบ k-Fold Cross Validation จากนั้นทำการกดปุ่ม Classify ดังแสดงในภาพประกอบ ก.16



ภาพประกอบ ก.16 การเลือกขั้นตอนวิธีสำหรับการจำแนกข้อมูล

และสามารถแสดงผลลัพธ์จากการจำแนกข้อมูลสายพันธุกรรมบนชุดข้อมูลได้ ดังภาพประกอบ ก.17



ภาพประกอบ ก.17 ผลลัพธ์จากการจำแนกข้อมูลสายพันธุกรรม

จากภาพประกอบ ก.17 ผลลัพธ์จากการจำแนกข้อมูลด้วยขั้นตอนวิธีแบบ J48 ซึ่งประกอบด้วย ข้อมูลต้นไม้การตัดสินใจ เวลาที่ใช้ในการทำงานในส่วนของการสอนและการทดสอบ ค่าความถูกต้อง (Accuracy) และคอนฟิวชันเมทริกซ์ ด้วยวิธีทดสอบแบบ k Fold Cross Validation เมื่อค่า k=4

ผลลัพธ์ที่ได้จากการทำงานของขั้นตอนวิธีอื่น ๆ จะสามารถแสดงได้ในลักษณะเดียวกัน

ภาคผนวก ข

คู่มือการติดตั้งระบบการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม

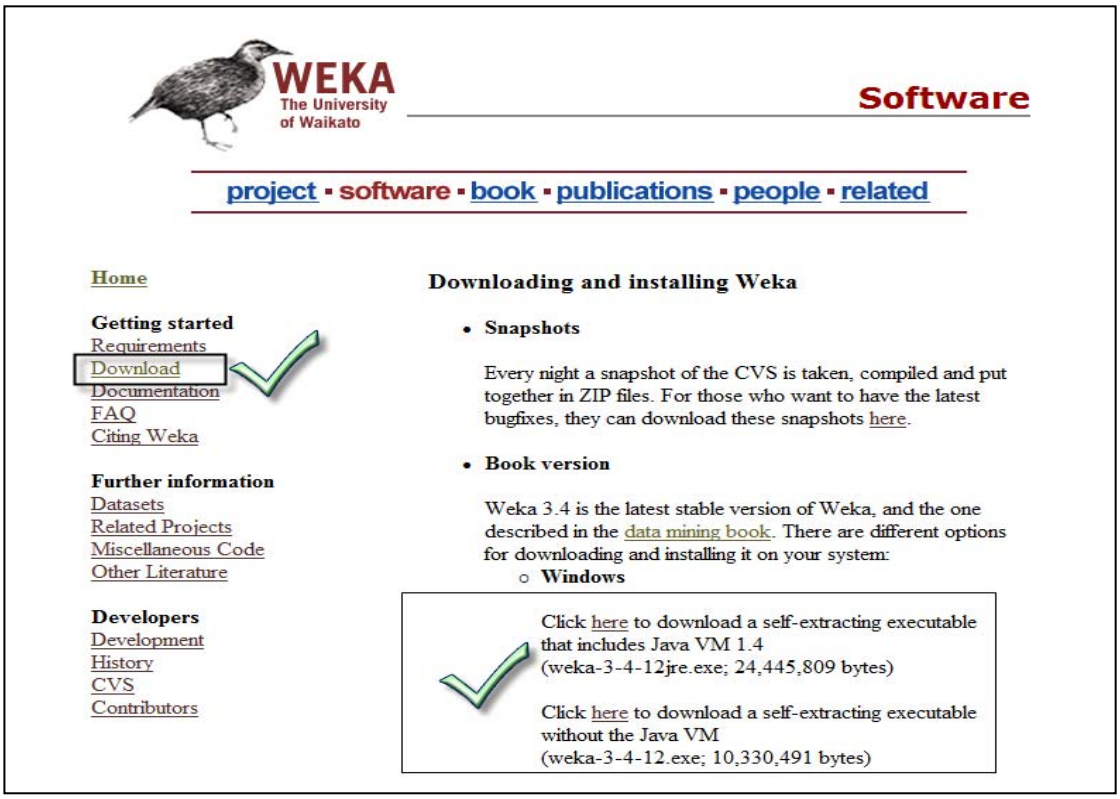
คู่มือการติดตั้งระบบ ประกอบด้วย ส่วนของการติดตั้งระบบ รูปแบบการสั่งงานผ่านทางบรรทัดคำสั่ง และตัวอย่างคำสั่งงานทางบรรทัดคำสั่ง สามารถอธิบายได้ตามลำดับดังนี้

1. การเตรียมระบบ

ขั้นตอนการเตรียมระบบเพื่อรองรับการใช้งานโปรแกรมการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม ประกอบด้วย การติดตั้งโปรแกรม WEKA เวอร์ชัน 3.4.10 และการกำหนดค่าสภาพแวดล้อมของระบบปฏิบัติการ (Environment) มีรายละเอียด ดังนี้

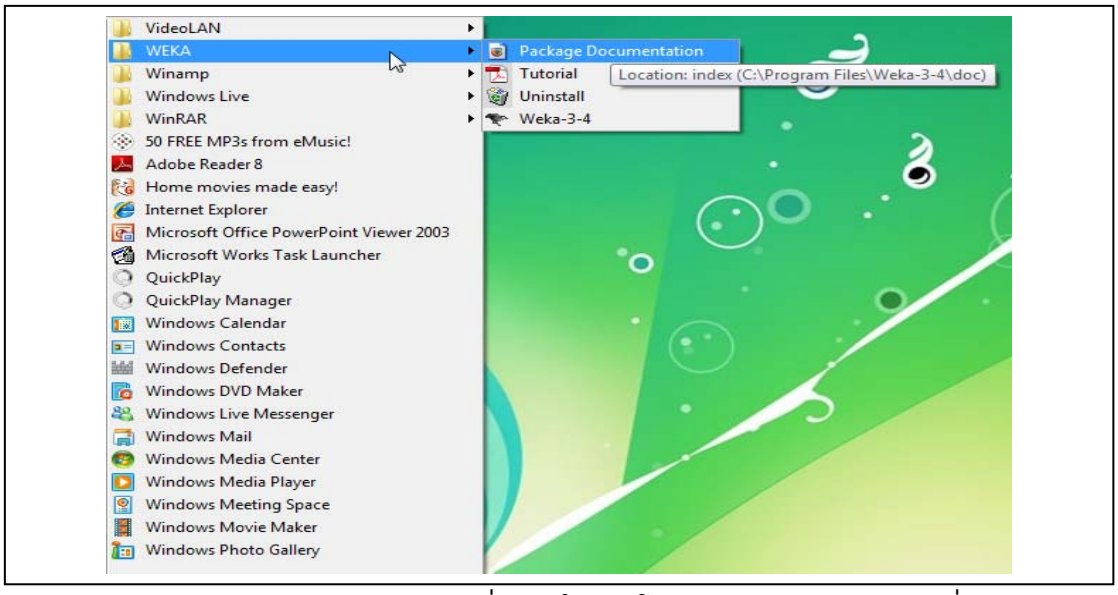
1.1 การติดตั้งโปรแกรม WEKA เวอร์ชัน 3.4.10

ผู้ใช้สามารถดาวน์โหลดโปรแกรม WEKA เวอร์ชัน 3.4.10 ได้จากเว็บไซต์ Weka Machine Learning Project (http://www.cs.waikato.ac.nz/~ml/index.html) โดยกดลิงค์ไปยังหน้าดาวน์โหลด (Download) ดังแสดงในภาพประกอบ ข.1



ภาพประกอบ ข.1 ลิงค์ดาวน์โหลดโปรแกรม WEKA

จากภาพประกอบ ข.1 หลังจากทำการดาวน์โหลดไฟล์ติดตั้ง (Setup File) ของโปรแกรม WEKA และทำการติดตั้งบนระบบ โดยในงานวิจัยนี้จะเลือกใช้ไฟล์ติดตั้ง ซึ่งติดตั้งบนระบบปฏิบัติการ Windows เมื่อทำการติดตั้งเรียบร้อยแล้วจะปรากฏแถบคำสั่งสำหรับเรียกใช้งาน บนคำสั่งเริ่มต้นของระบบปฏิบัติการ Windows (Start Menu) ดังแสดงในกาพประกอบ ข.2



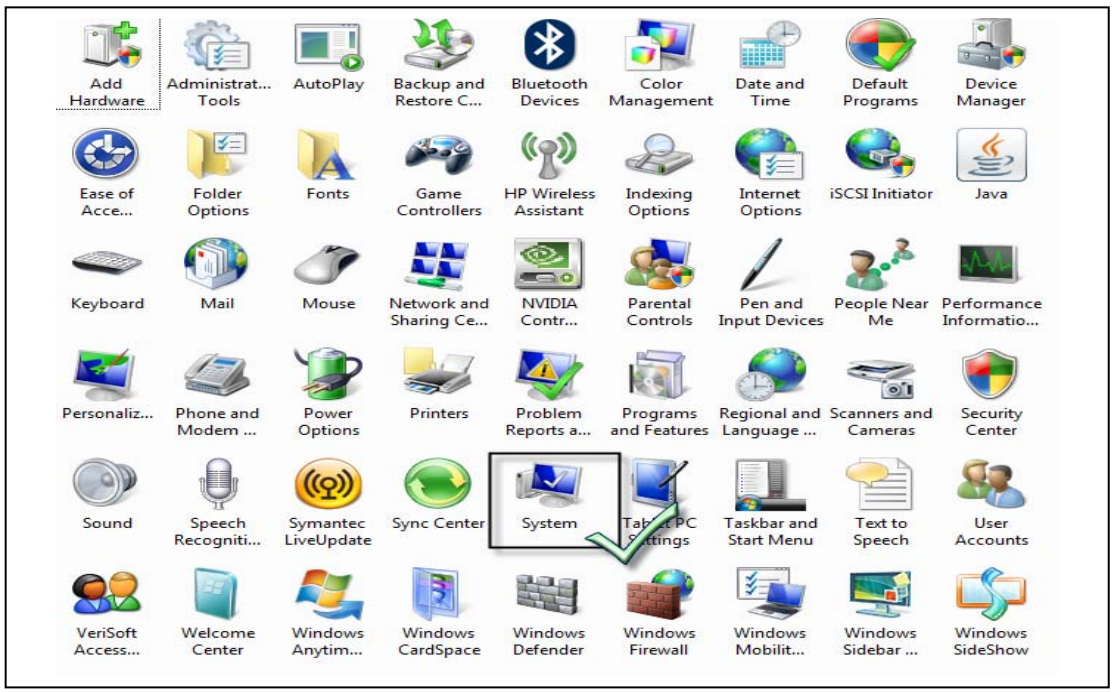
ภาพประกอบ ข.2 แสดงแถบคำสั่งเรียกใช้งานโปรแกรม WEKA บนเมนูเริ่มต้น

จากภาพประกอบ ข.2 เมื่อต้องการเรียกใช้งานโปรแกรม WEKA ผู้ใช้สามารถเรียกใช้งานโปรแกรมผ่านทางแถบคำสั่งได้ โดยเมื่อทำการกดแถบคำสั่งเพื่อเรียกใช้โปรแกรมจะปรากฏหน้าต่างโปรแกรมสำหรับใช้งาน นอกจากนี้หากผู้ใช้ต้องการอ่านเอกสารประกอบการใช้งานโปรแกรม WEKA นั้นผู้ใช้สามารถเรียกอ่านได้จากเมนูคำสั่ง Tutorial

1.2 การกำหนดสถานะแวดล้อมให้กับโปรแกรม WEKA

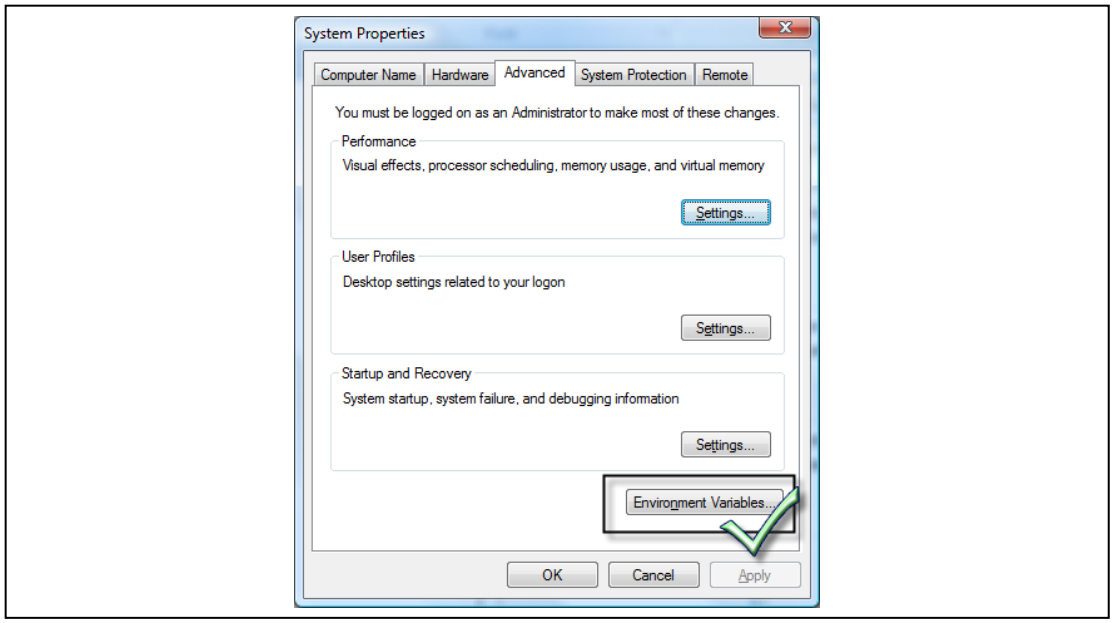
เนื่องจากโปรแกรม WEKA ซึ่งถูกพัฒนาด้วยภาษาจาวา (JAVA) จึงจำเป็นต้องมีการกำหนดสภาพแวดล้อม (Environment) ที่เหมาะสมกับการใช้งานโปรแกรม ซึ่งการกำหนดค่าสภาพแวดล้อม มีขั้นตอนดังนี้

1) เรียกใช้เมนูคำสั่ง Setting
2) เลือกคำสั่ง Control Panel
3) เลือกคำสั่ง System ดังแสดงในภาพประกอบ ข.3



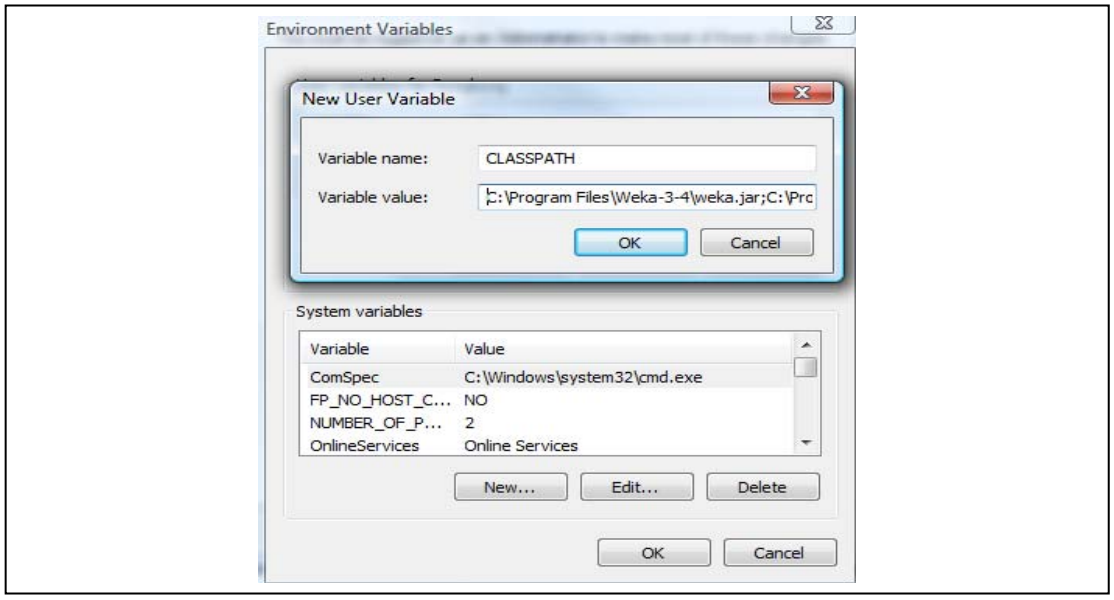
ภาพประกอบ ข.3 การเรียกใช้คำสั่ง System ในกลุ่มคำสั่งของ Control Panel

หลังจากกดปุ่มคำสั่ง System จะปรากฏหน้าต่าง System Properties แล้วให้ผู้ใช้เลือกคำสั่ง Advanced จากนั้นกดปุ่มเลือกคำสั่ง Environment Variables ดังแสดงในภาพประกอบ ข.4



ภาพประกอบ ข.4 ปุ่มคำสั่ง Environment Variables

จากนั้นให้กดปุ่ม New สร้างตัวแปรใหม่ ชื่อ CLASSPATH ลงในช่อง Variable Name และโดยป้อนคำสั่งสำหรับกำหนดค่าระบบ คือ C:\Program Files\Weka-3-4\weka.jar; C:\Program Files\Weka-3-4\weka-scr.jar ลงในช่อง Variable Value ดังภาพประกอบ ข.5



ภาพประกอบ ข.5 การกำหนดค่า Environment Variables

เมื่อผู้ใช้ทำการป้อนคำสั่งเรียบร้อยแล้ว ให้กดปุ่ม OK เพื่อทำการบันทึกค่าตัวแปร CLASSPATH เข้าสู่ระบบ จากนั้นทำการเริ่มต้นระบบใหม่ (System Reboot) แล้วจึงเรียกใช้งานโปรแกรม WEKA ได้ตามต้องการ

2 การสั่งงานผ่านทางบรรทัดคำสั่ง (Command Line)

ในการทำงานของระบบการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม สำหรับการจำแนกข้อมูลนั้น จะมีคำสั่งในการทำงานบางส่วนของระบบ ซึ่งเป็นการประมวลข้อมูลแบบชุด (Batch Processing) ด้วยโปรแกรม WEKA ดังนั้นจึงจำเป็นต้องอาศัยการทำงานผ่านทางบรรทัดคำสั่งและนำผลลัพธ์ที่ได้ไปใช้ในส่วนของการแสดงผลของระบบ โดยการเขียนชุดคำสั่งบนบรรทัดคำสั่งเพื่อสั่งให้โปรแกรม WEKA ทำงาน จะมีรูปแบบโดยการกำหนดค่าพารามิเตอร์ (Parameters) ผ่านทางบรรทัดคำสั่ง (Command Line) ดังภาพประกอบ ข.6

java {ชื่อขั้นตอนวิธีที่ต้องการเรียกใช้งาน} {ค่าพารามิเตอร์} {ข้อมูลนำเข้า} >> {ผลลัพธ์}
❶ ❷ ❸ ❹

ภาพประกอบ ข.6 แสดงรูปแบบของคำสั่งทางบรรทัดคำสั่ง (Command Line)

จากภาพประกอบ ข.6 บรรทัดคำสั่งที่ต้องการสั่งให้โปรแกรม WEKA ทำงานประกอบด้วย 4 ส่วน คือ

ส่วนที่ 1 คือ ส่วนของชุดคำสั่ง (Package) ที่ต้องการเรียกใช้

ส่วนที่ 2 คือ ส่วนของการกำหนดค่าพารามิเตอร์ชุดคำสั่งที่เรียกใช้

ส่วนที่ 3 คือ ส่วนของไฟล์ข้อมูลนำเข้าที่ต้องการโหลดเข้ามาใช้งาน

ส่วนที่ 4 คือ ส่วนของไฟล์ผลลัพธ์จากการทำงาน

3. ตัวอย่างคำสั่งบนบรรทัดคำสั่ง

ในการทำงานของระบบการวิเคราะห์รูปแบบและรู้จำลำดับพันธุกรรม นั้นจะใช้ชุดคำสั่งซึ่งเกี่ยวข้องกับขั้นตอนวิธี (Algorithms) ได้แก่ โครงข่ายประสาทเทียมแบบหลายชั้น (Multilayer Perceptron: MLP) แบบเรเดียลเบสิสฟังก์ชัน (Radial Basis Function: RBF) และต้นไม้การตัดสินใจแบบ Alternating Decision Tree (ADTree) แบบ Pruned C4.5 Decision Tree (J48) และแบบ K Random Features Decision Tree (RandomTree) ตามลำดับ ซึ่งสามารถอธิบายรายละเอียดการกำหนดค่าพารามิเตอร์ของแต่ละขั้นตอนวิธีได้ ดังนี้

3.1 พารามิเตอร์ของโครงข่ายประสาทเทียมแบบ MLP

ค่าพารามิเตอร์สำหรับโครงข่ายประสาทเทียมแบบหลายชั้น (MLP) มีการกำหนดค่าอัตราการเรียนรู้ (Learning Rate: L) เท่ากับ 0.3 ค่าโมเมนตัม (Momentum: M) เท่ากับ 0.2 ค่าจำนวนรอบในทำงาน (Number of Epoch: N) เท่ากับ 100 รอบ ค่าการสุ่ม (Random Seed: S) เท่ากับ 0 ค่าร้อยละของการทดสอบ (Percentage Size of The Validation Set: V) เท่ากับ 0 ค่าความผิดพลาดที่ยอมรับได้จากการทดสอบ (The Threshold for The Number of Consecutive Errors: E) เท่ากับ 20 และค่าของนิวรอนในชั้นซ้อน (Number of Nodes: H) แบบอัตโนมัติ (Auto) โดยคำนวณจำนวนของนิวรอนในแต่ละชั้น ดังสมการ (ข.1)

$$\text{จำนวนนิวรอนในชั้นซ่อน} = \frac{\text{จำนวนคุณสมบัตินำเข้าจากชั้นก่อนหน้า}}{2} \quad \text{(ข.1)}$$

โดยสามารถแสดงตัวอย่างการกำหนดค่าพารามิเตอร์สำหรับการเรียนรู้ด้วยโครงข่ายประสาทเทียมแบบ MLP ทางบรรทัดคำสั่งดังภาพประกอบ 5.11

```
java weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 100 -V 0 -S 0
    -E 20 -H a -x 4 -t  DataSet.arff  >> ResultMLP.txt
```

ภาพประกอบ ข.7 การกำหนดค่าพารามิเตอร์ของโครงข่ายประสาทเทียมแบบ MLP

3.2 ค่าพารามิเตอร์สำหรับโครงข่ายประสาทเทียมแบบ RBF

ค่าพารามิเตอร์สำหรับโครงข่ายประสาทเทียมแบบเรเดียลเบสิสฟังก์ชัน (RBF) ประกอบด้วยการกำหนดจำนวนกลุ่มที่ต้องการแบ่ง (Number of Clusters: B) เท่ากับ 2 รูปแบบของการเรียนรู้ (Ridge Parameter: R) มี 2 แบบคือ Logistic Regression และ Linear Regression โดยค่าเริ่มต้นจะเป็นแบบ Logistic Regression และกำหนดอัตราการเรียนรู้ (Learning Rate: R) เท่ากับ 1.0E-8 ค่าการสุ่ม (Random Seed: S) เท่ากับ 1 ค่าสูงสุดสำหรับการดำเนินการ (Maximum Number of Iterations: M) เท่ากับ -1 ซึ่งเป็นค่าเริ่มต้นกรณีเลือกรูปแบบการเรียนรู้แบบ Logistic Regression ซึ่งให้ขั้นตอนวิธีกำหนดค่าแบบอัตโนมัติ และต่ำสุดของส่วนเบี่ยงเบนมาตรฐาน (Minimum Standard Deviation: W) เท่ากับ 0.1 สามารถแสดงตัวอย่างการทำงานทางบรรทัดคำสั่ง (Command Line) ดังภาพประกอบ ข.8

```
java weka.classifiers.functions.RBFNetwork -B 2 -S 1 -R 1.0E-8 -M -1 -W 0.1 -x 4
    –t DataSet.arff  >> ResultRBF.txt
```

ภาพประกอบ ข.8 การกำหนดค่าพารามิเตอร์ของโครงข่ายประสาทเทียมแบบ RBF

3.3 พารามิเตอร์ต้นไม้การตัดสินใจแบบ ADTree, RandomTree และ J48

ในการกำหนดค่าพารามิเตอร์ของต้นไม้การตัดสินใจ มีจำนวนของพารามิเตอร์ที่ใช้งานไม่ซับซ้อน และมีลักษณะการกำหนดที่คล้ายคลึงกัน ต่างกันเพียงแค่ในส่วนของการเรียกใช้ ชุดคำสั่ง (Package) ที่ต้องการเรียกใช้เท่านั้น ดังแสดงในภาพประกอบ ข.9 ถึงภาพประกอบ ข.11 คือ

```
java weka.classifiers.trees.ADTree -x 4 -t DataSet.arff >> ResultADTree.txt
```

ภาพประกอบ ข.9 การกำหนดค่าพารามิเตอร์ของต้นไม้การตัดสินใจแบบ ADTree

```
java weka.classifiers.trees.J48 -x 4 -t DataSet.arff >> ResultJ48.txt
```

ภาพประกอบ ข.10 การกำหนดค่าพารามิเตอร์ของต้นไม้การตัดสินใจแบบ J48

```
java weka.classifiers.trees.RandomTree -x 4 -t DataSet.arff >> ResultRandom.txt
```

ภาพประกอบ ข.11 การกำหนดค่าพารามิเตอร์ของต้นไม้การตัดสินใจแบบ RandomTree

ภาคผนวก ค

ผลงานวิจัยที่ได้รับการตีพิมพ์ JCSSE 2007

เรื่อง n-Grams Protein Sequences Analysis for Disease Classification

สถานที่ The 4th International Joint Conference on Computer Science and Software Engineering (JCSSE 2007)

วันที่ ระหว่างวันที่ 2-4 พฤษภาคม 2550

ณ โรงแรมโซฟิเทลราชาออคิด จังหวัดขอนแก่น

n-Grams Protein Sequences Analysis for Disease Classification

Tanate Panrat, Ladda Preechaveerakul and Wiphada Wettayaprasit
Department of Computer Science, Prince of Songkla University, Thailand, 90112
Email : tanatepanrat@hotmail.com, ladda.p@psu.ac.th, wwettayaprasit@yahoo.com

บทคัดย่อ

การวิเคราะห์ลำดับโปรตีนบนสายพันธุกรรมสำหรับจัดกลุ่มโรค ช่วยให้สามารถหาลักษณะเด่นของโรคแต่ละชนิดออกมาและทราบถึงรูปแบบของลำดับโปรตีนบนสายพันธุกรรม งานวิจัยนี้ได้นำเสนอการวิเคราะห์ข้อมูลสายพันธุกรรมแบบ n-Grams มาใช้ในการแก้ปัญหา โดยใช้เทคนิคต้นไม้การตัดสินใจ (Decision Tree) เช่น ADTree, J48, Decision Stump, RandomTree และ REPTree ใช้ข้อมูลลำดับพันธุกรรมของผู้ป่วยโรคโรคเอดส์, ธาลัสซีเมีย, เบาหวาน และดาว์นซินโดรม ได้ผลการทดลองการแบ่งกลุ่มข้อมูลที่มีค่าความถูกต้องสูง

คำสำคัญ : การวิเคราะห์ลำดับโปรตีน, การแบ่งกลุ่มข้อมูล, n-Grams

1.บทนำ

ลำดับพันธุกรรมของสิ่งมีชีวิตเกิดจากการเรียงตัวของโปรตีน การวิเคราะห์ข้อมูลบนสายพันธุกรรมสามารถทำได้โดยวิธีการทำนายจากรูปแบบของการเกิดปฏิกิริยาความสัมพันธ์ระหว่างโปรตีนกับโปรตีน (Protein-Protein Interaction) [1] ซึ่งเป็นปฏิกิริยาความสัมพันธ์ในระดับโมเลกุลซึ่งสามารถบ่งบอกถึงระดับของวิวัฒนาการและกลไกในทำงานภายในสิ่งมีชีวิต ในปัจจุบันข้อมูลลำดับพันธุกรรมเกี่ยวกับการแสดงออกของโรคแต่ละชนิดมีจำนวนมาก ข้อมูลเหล่านี้สามารถนำไปวิเคราะห์เพื่อหาลักษณะเด่นของโรคแต่ละชนิดว่ามีความสัมพันธ์กับรูปแบบของลำดับโปรตีนบนสายพันธุกรรมอย่างไร

เทคนิคในการจัดกลุ่มข้อมูลที่นิยมใช้ในการวิเคราะห์ลำดับโปรตีนบนสายพันธุกรรม เช่น Support Vector Machine: SVM เป็นเทคนิคที่สร้างตัวจำแนกกลุ่มของข้อมูลจากการเรียนรู้ในกลุ่มต่าง ๆ ในการเลือกกลุ่มของคู่ลำดับโปรตีนที่มีความคล้ายคลึงกันของคู่ลำดับโปรตีนบนสายพันธุกรรม [2] SVM เหมาะสำหรับปัญหาการวิเคราะห์รูปแบบของลำดับที่มีจำนวนของลำดับพันธุกรรมจำนวนมากได้ดี เช่นขั้นตอนวิธี CSVM (Clustering Support Vector Machine) [3] เป็นต้น นอกจากนี้โครงข่ายประสาทเทียมเป็นกระบวนการในการประมวลผลซึ่งเลียนแบบการทำงานของ

สมองมนุษย์โดยที่จะมีการข้อมูลอินพุต-เอาท์พุตเข้าไปยังโครงข่ายประสาทเพื่อทำการเรียนรู้ ตัวอย่างการเรียนรู้เกี่ยวกับรูปแบบของโปรตีนบนสายพันธุกรรมช่วยให้สามารถจัดกลุ่มของโปรตีนได้อย่างมีประสิทธิภาพ [4] เช่นขั้นตอนวิธี MESSM (Mixed Environment-Specific Substitution Mapping) [5] เป็นต้น

ปัญหาหลักที่พบในการดำเนินการเกี่ยวกับลำดับพันธุกรรมคือความยาวของลำดับพันธุกรรมที่มีขนาดแตกต่างกัน การนำเทคนิคการหาค่าความถี่ของข้อมูลแบบ n-Grams สามารถช่วยในการทำนายลักษณะเด่นของโครงสร้างของลำดับพันธุกรรมได้สะดวกขึ้น [6, 7] นอกจากนี้ยังสามารถทำนายลักษณะรูปแบบของลำดับพันธุกรรมได้ด้วย [8, 9]

งานวิจัยนี้ได้ทำการจัดกลุ่มข้อมูลโดยใช้เทคนิคต้นไม้การตัดสินใจ (Decision Tree) และใช้หลักการจัดลำดับโปรตีนแบบ n-Grams ของผู้ป่วยที่เป็นโรคเอดส์, ธาลัสซีเมีย, เบาหวาน และคาว์นซินโดรม ที่รวบรวมจากฐานข้อมูล National Center for Biotechnology Information : NCBI [14] ส่วนที่ 2 ของบทความนี้จะกล่าวถึงเทคนิค n-Grams และต้นไม้การตัดสินใจ ส่วนที่ 3 กล่าวถึงขั้นตอนวิธีการวิเคราะห์ลำดับของโปรตีนแบบ n-Grams สำหรับการแบ่งกลุ่มโรค (n-Grams Protein Sequences Analysis for Disease Classification: n-Gram_PSA_DC) ส่วนที่ 4 แสดงผลการทดลอง และส่วนที่ 5 คือบทสรุป

2. เทคนิค n-Grams และต้นไม้การตัดสินใจ

2.1 เทคนิค n-Grams

n-Grams ใช้หลักการในการแบ่งข้อมูลของคู่ลำดับให้อยู่ในรูปแบบของเซตค่าความถี่ของคู่ลำดับ โดยทำการแบ่งข้อมูลออกเป็นส่วนย่อย ๆ ซึ่งขนาดของส่วนย่อยที่ทำการแบ่งในแต่ละครั้งนั้นจะมีขนาดเท่ากับ n ตัว (เรียกว่า n-Grams) [11] โดยสามารถเขียนแทนคู่อันดับได้ดังนี้ คือ

$$\text{คู่อันดับ } (V_i, C_i) \qquad (1)$$

โดยที่ V_i แทนคุณสมบัติที่ทำการเลือกขนาด n ตัว และ C_i แทนความถี่คู่ลำดับของข้อมูลที่นับได้บนชุดข้อมูลหรือสายอักขระ

2.2 ต้นไม้การตัดสินใจ

ต้นไม้การตัดสินใจแบบต่าง ๆ คือ ADTree, J48, Decision Stump, Random Tree และ REPTree ตามลำดับ [12] สามารถแสดงรายละเอียดได้ดังนี้

2.2.1) ADTree เป็นขั้นตอนการสร้างต้นไม้ตัดสินใจโดยใช้ค่า Boosting เป็นวิธีการที่เหมาะสมกับการปัญหาที่มีการจัดกลุ่มแบบ 2 กลุ่ม

2.2.2) J48 จะทำการสร้างต้นไม้โดยอาศัยพื้นฐานของขั้นวิธีแบบ C4.5 Decision Tree โดยปรับปรุงเพิ่มเติมโดยการเลือกโหนด (pruning) ที่ไม่จำเป็นออก

2.2.3) Decision Stump เป็นขั้นตอนวิธีการสร้างต้นไม้โดยพื้นฐานของอาศัยค่า Mean-Squared Error ร่วมกับเทคนิคของ Boosting Algorithm

2.2.4) Random Tree การสร้างต้นไม้จะอาศัยการสร้างต้นไม้ที่มีการสุ่มแอททริบิวที่เป็นได้ใน

การสร้าง Root Node โดยไม่ได้ทำการตัดแอททริบิวที่ไม่จำเป็นทิ้งในช่วงแรก

2.2.5) REPTree ใช้เทคนิคในการสร้างต้นไม้โดยใช้ค่า Information Gain และค่า Variance โดยทำการตัดแอททริบิวที่ไม่จำเป็นก่อนนำมาสร้างต้นไม้

3. ขั้นตอนวิธีการวิเคราะห์ลำดับของโปรตีนแบบ n-Grams สำหรับการแบ่งกลุ่มโรค

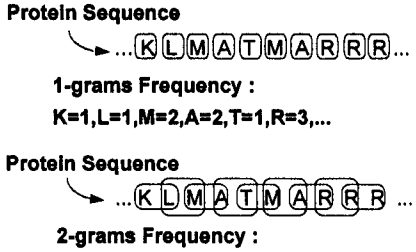
ขั้นตอนวิธีการวิเคราะห์ลำดับของโปรตีนแบบ n-Grams สำหรับการแบ่งกลุ่มโรค (n-Grams Protein Sequences Analysis for Disease Classification: n-Gram_PSA_DC) ประกอบด้วย 2 ขั้นตอนดังนี้คือ ขั้นตอนการเตรียมข้อมูล และขั้นตอนการแบ่งกลุ่มข้อมูล แสดงได้ดังรูปที่ 1

ขั้นตอนที่ 1 : ขั้นตอนการเตรียมข้อมูล
1.1 แปลงข้อมูลลำดับพันธุกรรมจากให้อยู่ในเซตของคู่อันดับโปรตีนบนสายพันธุกรรม ในรูปของ (V,C) โดยใช้เทคนิค n-Grams ที่ขนาดของ n=1, n=2
ขั้นตอนที่ 2 : ขั้นตอนการแบ่งกลุ่มข้อมูล
2.1 ข้อมูลลำดับพันธุกรรมมี 2 กลุ่ม คือกลุ่มที่เป็นโรค และกลุ่มที่ไม่เป็นโรค 2.2 ใช้การทดสอบแบบ 10-Folds Cross Validation 2.3 เลือกเทคนิคการแบ่งกลุ่มด้วยโปรแกรม WEKA กรณีที่ 1 : แบบ ADTree กรณีที่ 2 : แบบ J48 กรณีที่ 3 : แบบ Decision Stump กรณีที่ 4 : แบบ Random Tree กรณีที่ 5 : แบบ REPTree 2.4 ได้ค่าความถูกต้องการแบ่งกลุ่มข้อมูล

รูปที่ 1. ภาพแสดงขั้นตอนวิธีการวิเคราะห์ลำดับของโปรตีนแบบ n-Grams สำหรับการแบ่งกลุ่มโรค (n-Gram_PSA_DC)

3.1) ขั้นตอนการเตรียมข้อมูล

การเตรียมข้อมูลจะทำโดยแปลงรูปแบบของข้อมูลจากลำดับโปรตีนบนสายพันธุกรรม ให้อยู่ในรูปของเซตค่าความถี่ โดยใช้เทคนิค n-Grams โดยกำหนดให้ขนาดของ n เท่ากับ 1 (1-Grams) และ n เท่ากับ 2 (2-Grams) คือพิจารณาค่าความถี่ของลำดับโปรตีนเป็น 1 ตัวและ 2 ตัว ตามลำดับ เพื่อเป็นตัวแทนของลำดับพันธุกรรม การทดลองจะพิจารณาจากโปรตีนซึ่งพบในลำดับพันธุกรรมของสิ่งมีจำนวน 20 ชนิดประกอบด้วย A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y และ V ตามลำดับ รูปที่ 2 แสดงตัวอย่างการแปลงข้อมูลจากสายพันธุกรรม "...**KLMATMARRR**..." สำหรับการคำนวณแบบ 1-Grams คือการนับความถี่ของโปรตีนแต่ละค่า เช่นค่าโปรตีน M มีความถี่คือ 2 และ ค่าโปรตีน R มีความถี่คือ 3 เป็นต้น การคำนวณแบบ 2-Grams คือ การนับค่าความถี่ของโปรตีนครั้งละ 2 ค่าหรือเป็นคู่ๆโดยเริ่มจากคู่แรกและมีการขยับไปทางขวาทีละ 1 ตำแหน่ง เพื่อให้ได้คู่โปรตีนถัดไปจากรูปที่ 2 คู่โปรตีนคู่แรกคือ KL คู่ถัดไปคือ LM และคู่ถัดไปคือ MA เป็นต้น ในการพิจารณานับความถี่บนสายพันธุกรรมคู่โปรตีน RR มีความถี่คือ 2 เป็นต้น



รูปที่ 2. แสดงเทคนิค n-Grams ที่ n=1และn=2 ตามลำดับ

3.2) ขั้นตอนการแบ่งกลุ่มข้อมูล

ข้อมูลลำดับพันธุกรรมมี 2 กลุ่มคือกลุ่มที่เป็นโรคและกลุ่มที่ไม่เป็นโรค ใช้โปรแกรม WEKA ทำการทดสอบแบบ 10-Folds Cross Validation โดยใช้เทคนิค ADTree, J48, Decision Stump, Random Tree และ REPTree ตามลำดับ

4. ผลการทดลอง

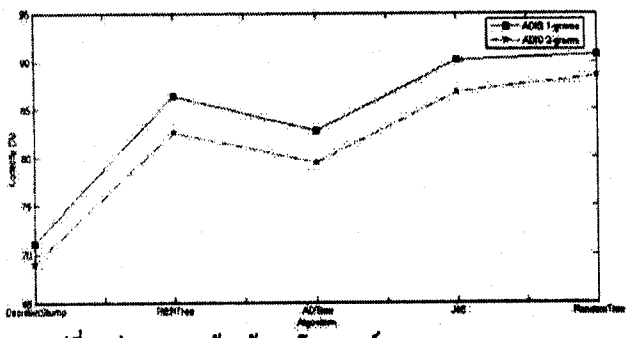
ในการทดลองทำการทดลองโดยใช้ข้อมูลผู้ป่วย 4 ชุดซึ่งประกอบด้วย ข้อมูลลำดับพันธุกรรมของผู้ป่วยเอดส์, ผู้ป่วยดาวน์ซินโดรม, ผู้ป่วยธาลัสซีเมียและผู้ป่วยโรคเบาหวาน จากฐานข้อมูล National Center for Biotechnology Information : NCBI [10] และโปรแกรม WEKA ทำการทดสอบข้อมูลแบบ 10-Folds Cross Validation

4.1) ข้อมูลผู้ป่วยโรคเอดส์: ในการทดลองประกอบด้วยข้อมูล 2 กลุ่มคือกลุ่มที่เป็นโรค (Disease Sequences) จำนวน 400 ชุดและไม่เป็นโรค (Normal Sequences) จำนวน 400 ชุดรวมเป็น 800 ชุด โดยทำการเตรียมข้อมูลแบบ 1-Grams (n=1) และ 2-Grams (n=2) จากนั้นแบ่งกลุ่มข้อมูลโดยใช้เทคนิคต้นไม้การตัดสินใจแบบ ADTree, J48, Decision Stump, RandomTree และ REPTree ค่าความถูกต้องแสดงได้ดังตารางที่ 1 และรูปที่ 3 ตามลำดับ จากการทดลองพบว่าเทคนิคที่ให้ค่าความถูกต้องสูงสุดคือเทคนิค Random Tree 1-Grams เท่ากับ 90.75% และ 2-Grams เท่ากับ 88.50% รองลงมาคือเทคนิค J48 ค่าความถูกต้องในแบบ 1-Grams เท่ากับ 90.12% และแบบ 2-Grams เท่ากับ 86.75%เทคนิคที่ให้ค่าความถูกต้องน้อยที่สุดคือเทคนิค Decision Stump ซึ่งสำหรับข้อมูลแบบ 1-Grams เท่ากับ 71.12%และแบบ 2-Grams เท่ากับ 69.00% ระยะเวลาในการทำงานแสดงได้ดังตารางที่ 2 และรูปที่ 4 ตามลำดับเมื่อเปรียบเทียบระยะเวลาในการทำงานพบข้อมูลแบบ 2-Grams ใช้เวลามากกว่าแบบ 1-Grams เทคนิค Decision Stump ใช้เวลาในการทำงานน้อยที่สุด และเทคนิคที่ใช้เวลาในการทำงานมากที่สุดคือเทคนิค ADTree

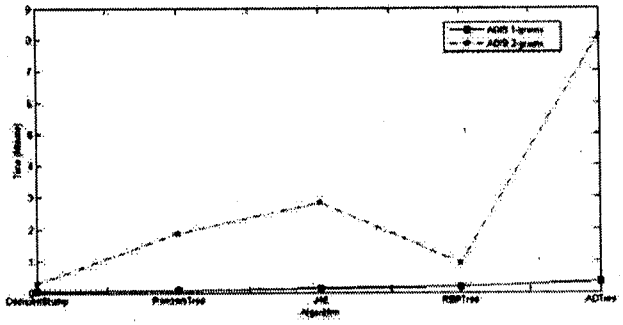
4.2) ข้อมูลผู้ป่วยโรคดาวน์ซินโดรม: ประกอบด้วยข้อมูลกลุ่มที่เป็นโรคจำนวน 476 ชุดและไม่เป็นโรคจำนวน 476 ชุด รวมเป็น 952 ชุด โดยทำการเตรียมข้อมูลแบบ 1-Grams (n=1) และ 2-Grams (n=2) ตามลำดับ จากนั้นใช้เทคนิคสำหรับการ

ตารางที่ 1. ค่าความถูกต้องในแบ่งกลุ่มข้อมูลของเทคนิคต่างๆ

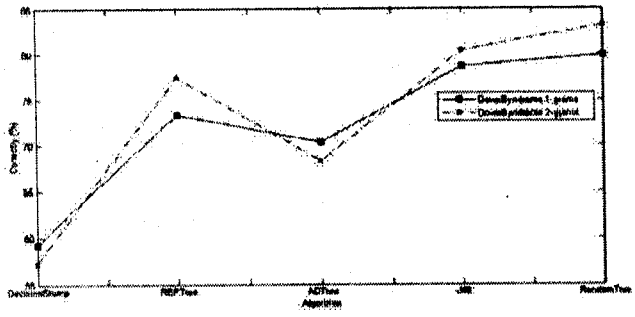
ชุดข้อมูล	ความถูกต้องในการจัดกลุ่ม (เปอร์เซ็นต์)				
	DecisionStump	REPTree	ADTree	J48	Random Tree
1) ผู้ป่วยเอดส์					
1-Grams	71.12	86.37	82.75	90.12	**90.75**
2-Grams	69.00	82.63	79.50	86.75	**88.50**
2) ผู้ป่วยดาวน์ซินโดรม					
1-Grams	59.24	73.32	70.38	78.67	**79.84**
2-Grams	57.23	77.43	68.28	80.36	**83.14**
3) ผู้ป่วยธาลัสซีเมีย					
1-Grams	76.00	79.67	81.33	80.67	**81.30**
2-Grams	70.67	79.00	79.67	81.67	**80.33**
4) ผู้ป่วยเบาหวาน					
1-Grams	64.93	77.00	73.64	82.43	**79.43**
2-Grams	61.22	77.43	73.36	80.57	**83.14**



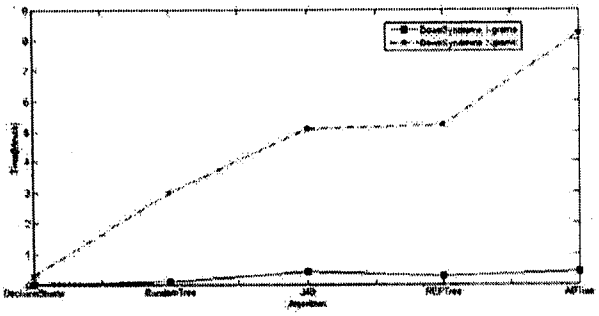
รูปที่ 3. ค่าความถูกต้องข้อมูลโรคเอดส์ 1-Grams และ 2-Grams



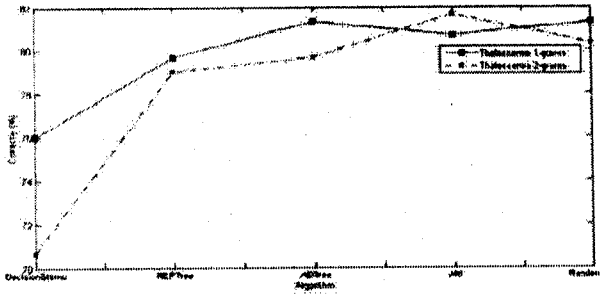
รูปที่ 4. เวลาที่ใช้ข้อมูลโรคเอดส์ 1-Grams และ 2-Grams



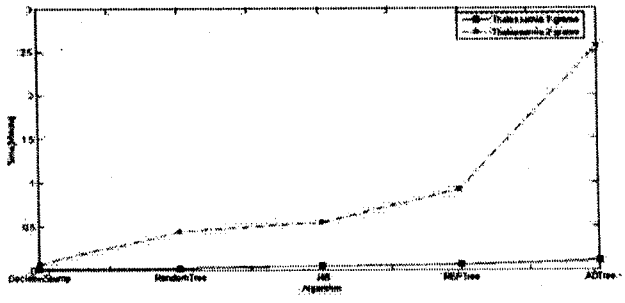
รูปที่ 5. ค่าความถูกต้องข้อมูลโรคดาว์นซินโดรม 1-Grams และ 2-Grams



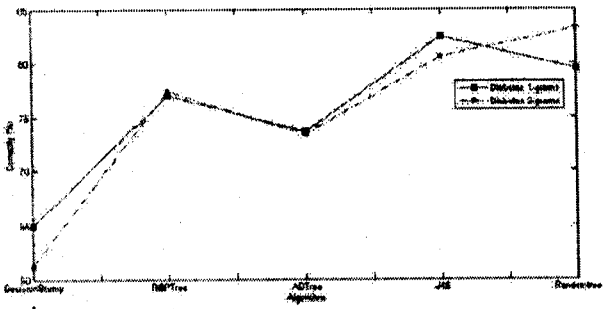
รูปที่ 6. เวลาที่ใช้ข้อมูลโรคดาว์นซินโดรม 1-Grams และ 2-Grams



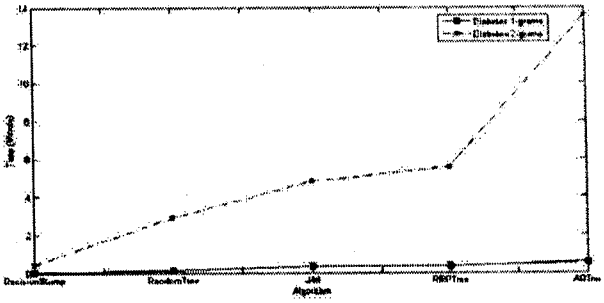
รูปที่ 7. ค่าความถูกต้องข้อมูลโรคธาลัสซีเมีย 1-Grams และ 2-Grams



รูปที่8. เวลาที่ใช้ข้อมูลโรคธาลัสซีเมีย 1-Grams และ 2-Grams



รูปที่ 9. ค่าความถูกต้องข้อมูลโรคเบาหวาน 1-Grams และ 2-Grams



รูปที่ 10. เวลาที่ใช้ข้อมูลโรคเบาหวาน1-Grams และ 2-Grams

ตารางที่ 2. เวลาในการแบ่งกลุ่มข้อมูลของเทคนิค

ชุดข้อมูล	เวลา (นาที)				
	DecisionStump	REPTree	ADTree	J48	Random Tree
1) ผู้ป่วยเอดส์					
1-Grams	**0.03**	0.17	0.33	0.13	**0.06**
2-Grams	**0.28**	0.92	8.16	2.83	**1.86**
2) ผู้ป่วยดาว์นซินโดรม					
1-Grams	**0.02**	0.27	0.41	0.42	**0.11**
2-Grams	**0.30**	5.22	8.20	5.08	**2.98**
3) ผู้ป่วยธาลัสซีเมีย					
1-Grams	**0.02**	0.06	0.11	0.05	**0.02**
2-Grams	**0.08**	0.92	2.56	0.55	**0.44**
4) ผู้ป่วยเบาหวาน					
1-Grams	**0.03**	0.34	0.53	0.34	**0.13**
2-Grams	**0.42**	5.53	13.56	4.81	**2.91**

แบ่งกลุ่มข้อมูลแบบต้นไม้ เพื่อเปรียบเทียบค่าความถูกต้องแสดงได้ในตารางที่ 1 และรูปที่ 5 ตามลำดับ

จากการทดลองพบว่าเทคนิคที่ให้ค่าความถูกต้องของการแบ่งกลุ่มสูงที่สุดคือเทคนิค Random Tree แบบ 1-Grams มีค่าความถูกต้องเท่ากับ 79.84% แบบ 2-Grams เท่ากับ 83.14% รองลงมาคือเทคนิค J48 แบบ 1-Grams มีค่าเท่ากับ 78.67% และแบบ 2-Grams มีค่าความถูกต้องเท่ากับ 80.36% เทคนิคที่ให้ค่าความถูกต้องน้อยที่สุดคือเทคนิค Decision Stump แบบ 1-Grams มีค่าเท่ากับ 59.24% และแบบ 2-Grams เท่ากับ 57.23% เมื่อเปรียบเทียบระยะเวลาในการทำงานพบว่าเทคนิค Decision Stump ใช้เวลาในการทำงานน้อยที่สุดและเทคนิคที่ใช้เวลาในการทำงานมากที่สุดคือเทคนิค ADTree แสดงได้ดังตารางที่ 2 และรูปที่ 6 ตามลำดับ

4.3) ข้อมูลผู้ป่วยโรคธาลัสซีเมีย: ประกอบด้วยข้อมูลกลุ่มที่เป็นโรคจำนวน 150 ชุดและไม่เป็นโรคจำนวน 150 ชุด รวมเป็น 300 ชุด โดยทำการเตรียมข้อมูลแบบ 1-Grams (n=1) และ 2-Grams (n=2) ตามลำดับ

จากการทดลองแสดงดังตารางที่ 1 และรูปที่ 7 พบว่าเทคนิคที่ให้ค่าความถูกต้องของการแบ่งกลุ่มข้อมูลสูงที่สุดคือเทคนิค Random Tree และเทคนิค J48 โดยสำหรับเทคนิค Random Tree ให้ค่าความถูกต้องข้อมูลแบบ 1-Grams ที่ 81.30% และแบบ 2-Grams เท่ากับ 80.33% และเทคนิค J48 แบบ 1-Grams เท่ากับ 80.76% และแบบ 2-Grams เท่ากับ 81.76% และพบว่าเทคนิคที่ให้ค่าความถูกต้องน้อยที่สุดคือเทคนิค Decision Stump แบบ 1-Grams เท่ากับ 76.00% และแบบ 2-Grams เท่ากับ 70.67% เมื่อเปรียบเทียบระยะเวลาในการทำงานพบว่าเทคนิค Decision Stump ใช้เวลาในการทำงานน้อยที่สุดและเทคนิคที่ใช้เวลาในการทำงานมากที่สุดคือเทคนิค ADTree แสดงได้ดังตารางที่ 2และรูปที่ 8 ตามลำดับ

4.4) ข้อมูลผู้ป่วยโรคเบาหวาน : ประกอบด้วยข้อมูลกลุ่มที่เป็นโรคจำนวน 700 ชุดและไม่เป็นโรคจำนวน 700 ชุด รวมเป็น 1,400 ชุด ประสิทธิภาพในการเรียนรู้และแบ่งกลุ่มข้อมูลของเทคนิคแต่ละวิธีซึ่งสามารถแสดงได้ในตารางที่ 1 และรูปที่ 9 จากการทดลองพบว่าเทคนิคที่ให้ค่าความถูกต้องของการแบ่งกลุ่มข้อมูลสูงที่สุดคือ

เทคนิค Random Tree โดยให้ค่าความถูกต้องข้อมูลแบบ 1-Grams เท่ากับ 79.43% และแบบ 2-Grams เท่ากับ 83.14% รองลงมาคือเทคนิค J48 แบบ 1-Grams เท่ากับ 82.43% และแบบ 2-Grams เท่ากับ 87.43% เทคนิคที่ให้ค่าความถูกต้องน้อยที่สุดคือเทคนิค Decision Stump ซึ่งสำหรับข้อมูลแบบ 1-Grams มีค่าความถูกต้องเท่ากับ 64.93% และแบบ 2-Grams เท่ากับ 61.22% เมื่อเปรียบเทียบระยะเวลาในการทำงานพบว่าเทคนิค Decision Stump ใช้เวลาในการทำงานน้อยที่สุดและเทคนิคที่ใช้เวลาในการทำงานมากที่สุดคือเทคนิค ADTree แสดงได้ดังตารางที่ 2 และรูปที่ 10 ตามลำดับ

5. สรุปผลการทดลอง

บทความนี้ได้นำเสนอเทคนิคการวิเคราะห์ลำดับของโปรตีนสำหรับแบ่งกลุ่มโรค (n-Grams_PSA_DC) เปรียบเทียบระหว่างข้อมูลแบบ 1-Grams และข้อมูลแบบ 2-Grams จากข้อมูลลำดับพันธุกรรมของผู้ป่วยเอดส์, ผู้ป่วยดาวน์ซินโดรม, ผู้ป่วยธาลัสซีเมีย และผู้ป่วยโรคเบาหวานโดยใช้ต้นไม้การตัดสินใจเทคนิค ADTree, J48, Decision Stump, Random Tree และ REPTree ตามลำดับ พบว่าเทคนิค Random Tree ให้ค่าความถูกต้องในการแบ่งกลุ่มข้อมูลสูงสุด รองลงมาคือเทคนิค J48 โดยที่เทคนิค Random Tree ใช้เวลาในการทำงานน้อยกว่าเทคนิค J48 และเทคนิคที่ให้ความความถูกต้องต่ำสุดคือเทคนิค Decision Stump โดยที่เวลาที่ใช้ในการทำงานพบว่าเทคนิคที่ใช้เวลาในการทำงานต่ำที่สุดเช่นกัน สำหรับเทคนิคที่ใช้เวลามากที่สุดคือ ADTree

ข้อมูลแบบ 2-Grams ใช้เวลาในการทำงานมากกว่าข้อมูลแบบ 1-Grams เนื่องจากมีคู่โปรตีนที่ใช้ในการวิเคราะห์ข้อมูลมากกว่ากล่าวคือ ถ้ามีโปรตีน 20 ชนิด ข้อมูลแบบ 1-Grams จะมีรูปแบบที่เป็นไปได้ทั้งสิ้น 20 แบบ แต่สำหรับข้อมูลแบบ 2-Grams จะมีรูปแบบที่เป็นไปได้ทั้งสิ้น 400 แบบ หรือ 20^2 แบบจึงทำให้ใช้เวลาในการทำงานมากขึ้น อย่างไรก็ตามข้อมูลแบบ 2-Grams สามารถให้ค่าความถูกต้องสูงกว่าข้อมูลแบบ 1-Grams ทั้งนี้ขึ้นอยู่กับข้อมูลดิบตามชนิดของโรคในสายพันธุกรรม

6. งานที่จะทำในอนาคต

เนื่องจากในกระบวนการวิเคราะห์ลำดับของโปรตีนสำหรับแบ่งกลุ่มโรคโดยใช้ต้นไม้การตัดสินใจได้กฎซึ่งมีจำนวนมาก งานวิจัยที่จะศึกษาและพัฒนาต่อในอนาคตเป็นส่วนของการลดจำนวนของกฎที่ใช้ในการวิเคราะห์ให้น้อยลง และได้กฎซึ่งให้ผลของค่าความถูกต้องสูงขึ้น

7. เอกสารอ้างอิง

[1] Jacques C., "Bioinformatics : An Introduction for Computer Scientists", ACM Computing Surveys, 2004, Vol.36 No.2, pp: 122-158.

[2] Natalia V. and Cathy H., "Prediction of catalytic residues using Support Vector Machine with Selected Protein Sequence and Structure Properties", BMC Bioinformatics, 2006.

[3] Wei Zhong, Jieyue He, Robert Harrison, Phnag C.Tai and Yi Pan, " Clustering support vector machine for protein local structure prediction", Science Direct, Expert system with Application 32, 2007, pp: 518-526.

[4] Wu C., Whitson G., Ermongkonchai A., Mclarty J. and Chang T. C., "Protein Classification Artificial Neural System", Protein Scicnce, 2006.

[5] Nan J.g, Wendy Xinyu W. and Lan M., "Protein Fold Recognition by Mixed Environment-Specific Amino Acid Substitution Mappang Using Nueral Networks", in Proceeding, 11th International Conference on Parallel and Distributed Systems (ICPADS'05), 2005.

[6] Faouzi M., Ricco R. and Mourad E., " A Compromise between N-gram Length and Classifier Characteristics for Protein Classification",IJCSNS International Journal of computer Science and Network Security, Vol.6 No.4, April 2006.

[7] Ganapathiraju M.,Weisser D., Rosenfeld R., Carbonell J., Reddy R. and Klein-Sestharaman, "Comparative n-gram Analysis of Whole-Genome Protein Sequences", Bioinformatics, 2006, Vol.22 No.3 , pp: 285-290.

[8] Faouzi M., Ricco R. and Mourad E., "A Hierarchical n-Grams Extraction Approach for Classification Problem", in Proceeding, The IEEE International Conference On Signal-Image Technology & Internet–Based Systems 2006 , Tunisia.

[9] Jiampojamarn S., Nick C. and Vlado K., "Biological Name Entity Recognition Using n-Grams and Classification Methods", http://users.cs.dal.ca/~vlado/papers/pacling 05a.pdf, accessed Jan 2007.

[10] The National Center for Biotechnology Information (NCBI), http://www.ncbi.nlm. nih.gov, accessed Nov 2006.

[11] Nanni L., "Hyperplances for Predicting Protein-Protein Interactions", Science Direct, Neurocomputing 69, 2005, pp: 254-263.

[12] Ian H. and Eile F., *Data Mining : Practical Machine Learning Tools and Techniques*, 2nd ed, 2005, p.403-425.

ภาคผนวก ง

ผลงานวิจัยที่ตีพิมพ์ IEEE (RIVF 2008)

เรื่อง Feature Extraction of Protein Sequence Analysis Using n-Grams Weight Balance Frequency and Neural Networks

สถานที่ 2008 IEEE International Conference on Research, Innovation and Vision for the Feature (RIVF'08)

วันที่ ระหว่างวันที่ 13-17 กรกฎาคม 2551
ณ ประเทศเวียดนาม

Feature Extraction of Protein Sequence Analysis Using n-Grams Weight Balance Frequency and Neural Networks

Wiphada Wettayaprasit
Artificial Intelligent Research Laboratary
Department of Computer Science
Prince of Songkla University
Songkhla, Thailand, 90112
wwettayaprasit@yahoo.com

Tanate Panrat
Artificial Intelligent Research Laboratary
Department of Computer Science
Prince of Songkla University
Songkhla, Thailand, 90112
tanatepanrat@gmail.com

Ladda Preecharveerakul
iSTAR Research Laboratary
Department of Computer Science
Prince of Songkla University
Songkhla, Thailand, 90112
ladda.p@psu.ac.th

Abstract—The study of protein-protein interaction on genetic sequences will be benefit to understand the dominant feature of disease that occurs on that genetic sequences. This study proposes feature extraction of protein sequence analysis model using n-grams data type that has weight adjusted with the proportion of dominant feature. Feature extraction of protein sequence analysis (FEPSA) using n-Grams weight balance frequency with neural networks composes of 3 main steps that are data preprocessing using weight balance frequency, feature extraction process, and feature classification process. The experiment uses patients' protein sequence of Thalassemia, Hepatites, and Downsyndrome. The study had found that Multilayer Perceptron neural networks gave the highest accuracy when compare with RBF, J48, and ADTree.

Keywords- Feature extraction; Protein sequences analysis; n-Grams; Neural networks; Decision tree.

I. INTRODUCTION

Genetic sequences of living organisms occur from the connection of at least 2 proteins with the bonding of peptide. Protein data analysis on genetic sequences can be tested by the method of prediction from the format of reaction occurrence between protein-protein interactions [1] which occurs at the molecule level that can indicate the feature of that particular genetic sequences. At present, there are a large number of data of genetic sequences that relate to the appearance of each disease. The data can be used for the analysis to find the dominant feature of each disease whether the disease has relationship with the protein-protein interaction of the genetic sequences. The major problem found in working with the genetic sequences is the different length of sequences. Then the technique of computing frequency value of *n*-Grams data type is used for the prediction. For examples, predicting the feature of genetic sequences [2, 3], predicting the feature of genetic sequences for Hyperplances type [4], and predicting the feature structure of genetic sequences for Hierarchical type [5, 6].

The famous techniques used for the sequence analysis of protein on genetic sequences are Support Vector Machine (SVM) and Artificial Neural Networks (ANN). SVM technique can be used for the selection of pair-order protein that has similar of pair-order protein on genetic sequences [7]. This technique is proper for the sequence analysis format that has a large number of sequences such as Clustering Support Vector Machine (CSVM) [8]. The neural networks use processing technique that imitate the working of human brain. The protein classification uses the neural networks such as ProCANS [9] and Mixed Environment-Specific Substitution Mapping (MESSM) [10].

Section 2 of this study will discuss on *n*-Grams technique, Neural Networks, and Decision Tree. Section 3 discusses on the steps of Feature Extraction of Protein Sequences Analysis (FEPSA) using *n*-Grams Weight Balance Frequency and Neural Networks. Section 4 discusses on the result of the study. Section 5 is conclusion.

II. N-GRAMS TECHNIQUE, NN AND DECISION TREE

A. *n-Grams Technique*

N-Grams technique uses the principle of arranging data into the format of frequency into pair-order set. The data will be divided into sub-data which the size equals to n (Call n-Grams) as (1).

$$\text{Pair}(v_i, c_i) \tag{1}$$

where v_i represents the feature property and c_i represents the frequency counts of the feature of the sequence.

B. *Neural Networks and Decision Tree*

Neural Networks are proper for problem solving that has complex data or has a nonlinear feature [11]. Multilayer Perceptron (MLP) Neural Networks uses the neural networks architecture of feedforward Networks which composes of input layer, hidden layer, and output layer. The logistic activation function will be used for this method while Radial Basis Function (RBF) Neural Networks use the Gaussian Function activation function. Decision tree is a learning method that uses reasoning. This method is a proper solution for problem which non-consecutive data by inferenceing the knowledge into the "*if...then...*" rule that is easy to understand [12, 13]. The well known methods are such as ADTree, J48, and RandomTree. ADTree will construct a decision tree by using Booting score which is proper for problem which has 2 sets of data. J48 will construct a tree by using the basic principle of C.45 method. C.45 will improve the pruning of the unnecessary nodes. RandomTree constructs a tree by using random sampling of attributes to construct the root node to receive a suitable node in the next layer.

III. PROTEIN SEQUENCES ANALYSIS MODEL

FEPSA using n-Grams weight balance frequency with neural networks composes of 3 main steps that are Data Preprocessing using Weight Balance Frequency, Feature Extraction Process, and Feature Classification Process. The FEPSA algorithm are shown in Fig 1.

Step 1 : Data Preprocessing Using Weight Balance Frequency

Compute the frequency using *n*-Grams Technique. Convert the protein equences into the format of *2*-Grams pair-order as shows in (1).
Calculate the Frequency Feature Vector of each sequence as shows in (2).
Compute the Weight Balance Frequency Feature Vector as shows in (3).

Step 2 : Feature Extraction Process

Compute the Local Analyze as shows in (4).
Calculate the Frequency Instance Value (FIV) as shows in (7).
Calculate Frequency Percentage Value (FPV) as shows in (8).
Compute the Global Analyze by Threshold point (λ) as shows in (9).

Step 3 : Feature Classification Process

Set k-Folds Cross Validation.
Select algorithms: MLP, RBF, ADTree, J48, or RandomTree.
Calculate the classification accuracy.

Figure 1. The proposed FEPSA Model

Data Preprocessing Using Weight Balance Frequency

.1 Compute the frequency using n-Grams technique

his step is to compute the frequency value from the ein sequences in the file format using *2*-Grams technique. epresent the protein sequences, let T be the symbol of o acid of living organisms (20 kinds), where T={A, R, N, , Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V} and $P_{i,j}$ be number of all possible features of *n*-Grams. Let $k,V_{i,j},F_{i,j}$) be the frequency feature vector of all feature as ws in (2)

$$\mathrm{FV}(k,V_{i,j},F_{i,j}) \qquad (2)$$

re k is the sequence number, $V_{i,j}$ is the *2*-Grams feature, $F_{i,j}$ e frequency of *2*-Grams feature, i is the row of the feature or, and j is the column of feature vector. Let n be 2, then umbers of formats received is equal to 20^2 or 400 possible rns as follows.

$$P_{i,j}=\begin{Bmatrix} AA & AR & AN & \dots & AV \\ RA & RR & RN & \dots & RV \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ VA & VR & VN & \dots & VV \end{Bmatrix}$$

or example, $P_{1,1}$ is AA, $P_{1,2}$ is AR, …, and $P_{20,20}$ is VV.

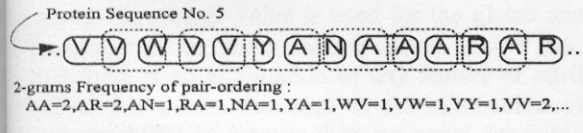


Figure 2. Shows 2-Grams Technique.

or example, Fig 2 shows the protein sequence no. 5 of VWWVVYANAAARAR...", using *2*-Grams technique. frequencies of feature vector are as follows: FV(5,AA,2), ,AR,2), FV(5,AN,1), …, FV(5,VW,1), FV(5,VY,1), and ,VV,2) as shows in Table 1.

TABLE 1. THE EXAMPLE OF FREQUENCY OF FEATURE VECTOR

No.	Seq. Length	Frequency of Feature Vector (FV)							Mean
		AA	AR	AN	...	VW	VY	VV	
q-1	86	5	1	3	...	2	1	1	4.40
q-2	69	5	6	1	...	0	3	5	5.81
q-3	96	8	8	9	...	0	5	3	1.78
q-4	76	1	7	1	...	2	8	7	5.31
q-5	73	2	2	1	...	1	1	2	4.98

1.2 Compute the Weight Balance Frequency Feature Vector

Since the protein sequences of the data set have difference size of length, the Weight Balance of Frequency (WBL) will compute the new frequency value of each input feature into a percentage format between 0 to 100. Therefore, the new WBL value can be appropriate when compare between each protein sequence. Let $WBL(k,V_{i,j},W_{i,j})$ be a weight balance feature vector of each sequence as shows in (3).

$$\mathrm{WBL}(k,V_{i,j},W_{i,j})=\frac{\mathrm{FV}(k,V_{i,j},F_{i,j})}{\mathrm{L}(k)-(n-1)}\times 100 \qquad (3)$$

where $W_{i,j}$ is the new weight balance value, $L(k)$ is the length of sequence number k, and n is the number of n-Grams value.

From (3), suppose that the sequence no. 5 has the length equal to 73. The examples of WBL Feature Vector are WBL(5,AA,2.78), WBL(5,AR,2.78), WBL(5,AN,1.39), WBL(5,VW,1.39), WBL(5,VY,1.39), and WBL(5,VV,2.78) as shows in Table 2.

TABLE 2. THE EXAMPLE OF WEIGHT BALANCE FREQUENCY

Seq. No.	Seq. Length	Weight Balance Feature Vector (WBL)							Mean
		AA	AR	AN	...	VW	VY	VV	
Seq-1	86	5.88	1.18	3.53	...	2.35	1.18	1.18	2.42
Seq-2	69	7.35	8.82	1.47	...	0.00	4.41	7.35	2.04
Seq-3	96	8.42	8.42	9.47	...	0.00	5.26	3.16	3.01
Seq-4	76	1.33	9.33	1.33	...	2.67	10.67	9.33	2.85
Seq-5	*73*	*2.78*	*2.78*	*1.39*	*...*	*1.39*	*1.39*	*2.78*	*3.52*

Note that, if the user does not use the WBL, but user uses the frequency feature vector (FV), then it will call Non-Weight Balance (NBL) approach.

B. Feature Extraction Process

2.1 Compute the Local Analyze

1) Compute the average frequency of all sequences. Let μ be the average frequency of all sequences as shows in (4).

$$\mu=\frac{\sum_{k=1}^{K} M_k}{K} \qquad (4)$$

where M_k is the average frequency of each sequence k. K is the total number of protein sequences.

Note that, the average frequency of each sequence (M_k) can be calculated by using the frequency feature vector of the $\mathrm{FV}(k,V_{i,j},F_{i,j})$ of NBL for μ_1 or $\mathrm{WBL}(k,V_{i,j},W_{i,j})$ of WBL for μ_2. For example, in Table 2, the average value of WBL of all sequences (μ_2) is equal to 2.13 or (2.42+2.04+3.01+2.85+3.52)/5, where the total number of sequence K is equal to 5.

2) Identify status of each feature. Let $\mathrm{ST}(k,V_{i,j},S_{i,j})$ be the status of feature vector as shows in (5) for NBL approach and for WBL approach in (6).

$$\mathrm{ST}(k,V_{i,j},S_{i,j})=\begin{cases} 1, If\ FV(k,V_{i,j},F_{i,j}) \geq \mu_1 \\ 0, Otherwise. \end{cases} \qquad (5)$$

$$\mathrm{ST}(k,V_{i,j},S_{i,j})=\begin{cases} 1, If\ WBL(k,V_{i,j},W_{i,j}) \geq \mu_2 \\ 0, Otherwise. \end{cases} \qquad (6)$$

Table 3 represents the status of each feature. In Table 2 sequence no. 5, μ_2 is equal to 2.13. For WBL(5,AA,2.78) weight balance value of AA is 2.78. Because 2.78 is greater the value of μ_2 then the status representation is 5,AA,1). For WBL(5,VY,1.39) the weight balance value of is 1.39. Because 1.39 is less then the value of μ_2, then is representation is as ST(5,VY,0).

TABLE 3. THE EXAMPLE OF STATUS FEATURE VECTOR.

No.	Status of Feature Vector (ST)						
	AA	AR	AN	...	VW	VY	VV
q-1	1	0	1	...	1	0	0
q-2	1	1	0	...	0	1	1
q-3	1	1	1	...	0	1	1
q-4	0	1	0	...	1	1	1
q-5	*1*	*1*	*0*	...	*0*	*0*	*1*
IV	4	4	2	...	2	3	4
V(%)	80.00	80.00	40.00	...	40.00	60.00	80.00

2.2 Calculate the Frequency Instance Value (FIV) of status ture vector as shows in (7).

$$FIV_j = \sum_{k=1}^{K} ST(k, V_{i,j}, S_{i,j}) \quad (7)$$

ere j is the number of feature vector, k is the number uence. The next step is to calculate the percentage of FIV as ws in (8).

$$FPV_j = \frac{FIV_j}{K} \times 100 \quad (8)$$

For example, from Table 3, the value of FIV_1 is equal and the value of FPV_1 is equal to 80%.

2.3 Compute Global Analyze

Let λ be the support threshold between 0 to 100, APV_j he Accept Percentage Value (APV) as shows in (9).

$$APV_j = \begin{cases} 1, If\ FPV_j \geq \lambda \\ 0, Otherwise. \end{cases} \quad (9)$$

The accept percentage value is used for the global analyze he frequency of the status feature vector. If λ = 0%, this ans that there is no elimination of any feature or attribute. λ = 60%, this means that then is the elimination of the ture where APV_j is greater than or equal to 60%. For mple, in Table 3, if set the value of λ is equal to 60%, then features that will be accepted are AA, AR, VY, and VV ause APV value are 80%, 80%, 60%, and 80%, pectively. Note that, AN and VW will be eliminated.

Feature Classification Process

This step is used to classify the input features by separating data set into training set and testing set. The protein uence should have two classes. The first class is the positive a, and the second class is the negative data. Next is to ection the classification algorithm from WEKA, which are LP, RBF, ADTree, J48, and RandomTree.

IV. EXPERIMENTAL RESULTS

The study designs the experiment into 4. Model A uses the data type of NBL without Feature Extraction. Model B uses the data type of WBL without Feature Extraction. Model C uses the data type of NBL with Feature Extraction. The last Model D uses the data type of WBL with Feature Extraction.

The experiment uses their data sets from NCBI [14] from the patients protein sequence who are sick with the disease caused by the genetics transferring. These data sets are genetic sequences of Thalassemia, Diabetes, and Downsyndrome. The 4-Folds cross validation is used. The teaching data set and the testing data set will be different from each other. The experiments with three- data sets are as follows:

D. The genetic sequence of patients with Thalassemia

There are two sets of data used for these patients. The first data set composes of 127 sequences of positive data. The second data set composes of 137 sequences of negative data. Then The total of data set is 264 sequences. The average length of genetic sequence is 463.50 (±5%). In the process of global analyze, when specified the percentage value of acceptance equal to 10% (λ=10%) for Model C and Model D, the accuracy of Model D gave highest accuracy when compared with all algorithms as shows in Fig 3. The experiment also found that MLP algorithm gave highest accuracy equal to 91.64% when ADTree, J48, RBF, and RandomTree gave less accuracy equal to 89.08%, 87.88%, 85.61%, and 84.33%, respectively. The number of attributes received from Model C and Model D can reduce the input number of attributes received from 400 attributes (in Model A and Model B) to 23 attributes (in Model C and Model D). These attributes are pair-order proteins as follows : AA, AN, AP, AS, AT, RE, NS, DI, EN, ED, EI, ES, ET, GI, GL, IA, IN, IE, PR, PI, SA, TA, and TE.

Fig 4 shows the time used from classification process. The experimental result related to time indicated that Model D of all algorithms used the lowest time when compared with Model A, Model B, and Model C. The algorithm of Random Tree of Model D used lowest time equal to 0.02 sec when Model C, Model A, and Model B used 0.03 sec, 1.36 sec, and 1.36 sec, respectively. Model C and Model D used less time because of fewer numbers of attributes.

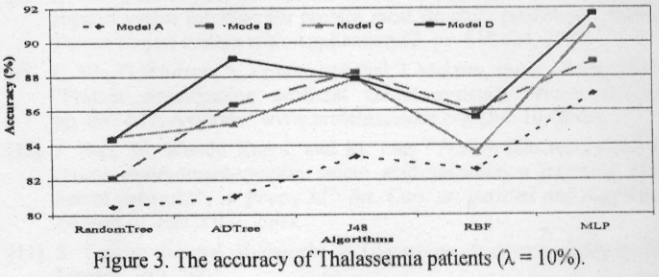


Figure 3. The accuracy of Thalassemia patients (λ = 10%).

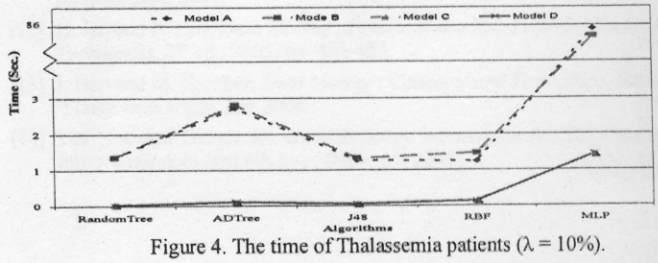
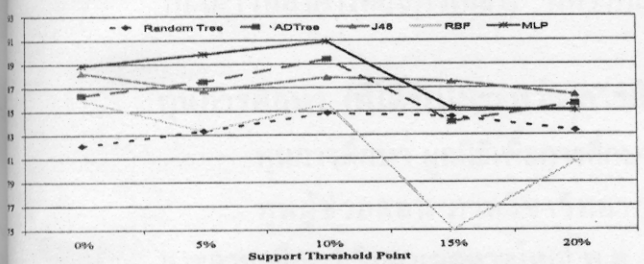


Figure 4. The time of Thalassemia patients (λ = 10%).

When select the variable Support Threshlod Point (λ), the eriment shows that λ has effect on the number of attributes ved. If λ=0%, this means no elimination for any of ibutes. In specifying the value of λ, if λ is increasing, then number of attributes received will be decreased as well. e that where λ=0%, 5%, 10%, 15%, and 20% the number ttributes received is 400, 57, 23, 12, 4, receptivity. From 5, when compare the accuracy of different λ values, at 0% the number of attributes received is 23. The experiment ws that MLP algorithm giving its highest accuracy at 87%, ADTree algorithm at λ=10% will give its highest uracy at 89.40%, RandomTree algorithm will give its hest accuracy at 84.54%. The classification accuracy ends on the proper value of λ. If λ is a small number, then attributes received will be large number. In contrast, if λ is ge, then the attributes received will be small.



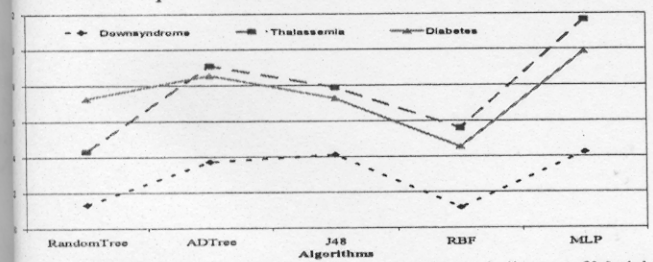
igure 5. Model D of Thalassemia patients compare with different value of λ.

The genetic sequence of patients with Diabetes

There are 2 sets of data used for these patients. The first ta set composes of 121 sequences of positive data. The cond data set composes of 272 sequences of negative data. e total of data set is 393 sequences. The average length of netic sequence data set is 152.58 (±5%). The experiment also nd that MLP algorithm gave highest accuracy equal to 87% when ADTree, J48, RBF, and RandomTree gave aller values equal to 88.55%, 87.28%, 87.28%, and 84.57%, pectively. The number of attributes received from Model C d Model D can reduce the input number of attributes to be 9 ributes.

The genetic sequence of patients with Downsyndrome

There are 2 sets of data used for these patients. The first ta set composes of 202 sequences of positive data. The cond data set composes of 185 sequences of negative data. e total of data set is 387 sequences. The average length of netic sequences is 187.50 (±5%). The experiment found that LP algorithm gave highest accuracy equal to 84.23% when 8, ADTree, RBF, and RandomTree gave smaller values ual to 84.11%, 83.72%, 81.33%, and 81.14%, respectively. e number of attributes received from Model C and Model D n reduce the input number of attributes to be 12 attributes.



igure 6. Show the results comparing of accuracy for each disease of Model D

From Fig 6, the comparison of classification accuracy of each disease of Model D indicated that MLP algorithm gave the highest accuracy.

V. Conclusion

This study proposed techniques for of protein sequences analysis of disease classification by using the genetic sequences of patients who were sick with the causes of Diabetes, Thalassemia, and Downsyndrome. There are four issues to be considered. 1) Comparing the efficiency of using frequency value of pair-order amino acids on genetic sequence for NBL and WBL. The study found that the WBL data gave higher accuracy than the NBL data, 2) The efficiency of classification of algorithms for Neural Networks (MLP and RBF) and Decision Tree (RandomTree, ADTree, and J48) indicated that MLP gave highest accuracy, 3) The Using of classification time indicated that Neural Networks used more time than Decision Tree, and 4) The WBL and the a Feature extraction gave small number of attributes. In conclusion, the proposed mode l using WBL with global analyze and MLP algorithm of feature extraction techniques gives the best classification accuracy.

References

[1] C. Jacques, "Bioinformatics : an introduction for computer scientists", *ACM Computing Surveys*, vol. 36, no. 2, pp. 122-158, 2004.

[2] S. Jiampojamarn, C. Nick, and K. Vlado, "Biological name entity recognition using n-grams and classification methods", Internet: http://users.cs.dal.ca/~vlado/papers/pacling05a.pdf, [Jan. 08, 2007].

[3] M. Faouzi, R. Ricco, and E. Mourad, " A compromise between n-gram length and classifier characteristics for protein classification", *IJCSNS Int. Journal of computer science and network security*, vol. 6, no. 4, Apr. 2006.

[4] L. Nanni, "Hyperplances for predicting protein-protein interactions", *Science Direct*, Neurocomputing 69, pp. 254-263, 2005.

[5] M. Ganapathiraju, D. Weisser, R. Rosenfeld, J. Carbonell, R. Reddy and K. Sestharaman, "Comparative n-gram analysis of whole-genome protein sequences", *Bioinformatics*, vol. 22, no. 3 , pp. 285-290, 2006.

[6] M. Faouzi, R. Ricco, and E. Mourad, "A hierarchical n-grams extraction approach for classification problem", in *Proc., the IEEE Int. Con. on signal-image technology & internet–based systems*, 2006, Tunisia.

[7] V. Natalia and H. Cathy, "Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structure properties", *BMC Bioinformatics*, 2006.

[8] Z. Wei , H. Jieyue, H. Robert, C. Tai Phnag, and P. Yi, "Clustering support vector machine for protein local structure prediction", *Science Direct*, Expert system with Application 32, pp. 518-526, 2007.

[9] C. Wu, G.Whitson, A. Ermongkonchai, J. Mclarty, and T. Chang (1992), "Protein classification artificial neural system", *Protein Scicnce*, pp. 667-677, Avilable : www.proteinscience.org, [Jul. 10, 2006].

[10] J. Nan, W. Wendy Xinyu, and M. Lan, "Protein fold recognition by mixed environment-specific amino acid substitution mappang using nueral networks", in *Proc., 11th Int. Con. on parallel and distributed systems (ICPADS'05)*, 2005.

[11] S. Kumar, *Neural Netoworks A Classroom Approach*, Singgapore : Mcgraw Hill, 2005.

[12] H. Ian and F. Eile, *Data Mining : Practical Machine Learning Tools and Techniques*, 2nd ed., 2005, pp. 403-425.

[13] J. Han and M. Kamber, *Data Mining : Concepts and Techniques*, Simon Fraser University, Jan. 2000.

[14] The National Center for Biotechnology Information (NCBI), Internet: http://www.ncbi.nlm.nih.gov , [Nov. 15, 2007].

ประวัติผู้เขียน

ชื่อ-สกุล นายธเนศ ปานรัตน์

รหัสประจำตัวนักศึกษา 4822030

วุฒิการศึกษา

วุฒิ	ชื่อสถาบัน	ปีที่สำเร็จการศึกษา
วท.บ. (วิทยาการคอมพิวเตอร์)	มหาวิทยาลัยราชภัฏสุราษฎร์ธานี	2546

ทุนการศึกษา (ที่ได้รับในระหว่างการศึกษา)

ทุนการศึกษา มูลนิธิเพื่อการศึกษาคอมพิวเตอร์และการสื่อสาร พ.ศ. 2548

ทุนผู้ช่วยสอน ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ พ.ศ. 2548-2550

การตีพิมพ์เผยแพร่ผลงาน

1. Panrat, T., Preechaveerakul, L., and Wettayaprasit, W. 2007. n-Grams Protein Sequences Anaysis for Disease Classification. Proceeding of The 4th International Joint Conference on Computer Science and Software Engineering (JCSSE 2007). Khon Kaen, Thailand, May 2-4, 2007, pp. 349-356.
2. Wettayaprasit, W., Panrat, T., and Preechaveerakul, L. 2008. Feature Extraction of Protein Sequence Analysis Using n-Grams Weight Balance Frequency and Neural Networks. Proceeding of The 2008 IEEE International Conference on Research Innovation and Vision for the Feature (RIVF 2008). Ho Chi Minh City, Vietnam, July 13-17, 2008.

ประวัติผู้เขียน

ชื่อ-สกุล นายธเนศ ปานรัตน์

รหัสประจำตัวนักศึกษา 4822030

วุฒิการศึกษา

วุฒิ	ชื่อสถาบัน	ปีที่สำเร็จการศึกษา
วท.บ. (วิทยาการคอมพิวเตอร์)	มหาวิทยาลัยราชภัฏสุราษฎร์ธานี	2546

ทุนการศึกษา (ที่ได้รับในระหว่างการศึกษา)

ทุนการศึกษา มูลนิธิเพื่อการศึกษาคอมพิวเตอร์และการสื่อสาร พ.ศ. 2548

ทุนผู้ช่วยสอน ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ พ.ศ. 2548-2550

การตีพิมพ์เผยแพร่ผลงาน

1. Panrat, T., Preechaveerakul, L., and Wettayaprasit, W. 2007. n-Grams Protein Sequences Anaysis for Disease Classification. Proceeding of The 4th International Joint Conference on Computer Science and Software Engineering (JCSSE 2007). Khon Kaen, Thailand, May 2-4, 2007, pp. 349-356.
2. Wettayaprasit, W., Panrat, T., and Preechaveerakul, L. 2008. Feature Extraction of Protein Sequence Analysis Using n-Grams Weight Balance Frequency and Neural Networks. Proceeding of The 2008 IEEE International Conference on Research Innovation and Vision for the Feature (RIVF 2008). Ho Chi Minh City, Vietnam, July 13-17, 2008.