การวิเคราะห์สายจีโนมโดยใช้โครงข่ายประสาทเทียม

Genome Sequences Analysis Using Neural Networks

ฏรินทร คงมณี

Tarintorn Kongmanee

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
2552

ชื่อวิทยานิพนธ์ การวิเคราะห์สายจีโนมโดยใช้โครงข่ายประสาทเทียม
ผู้เขียน นางสาวฏรินทร คงมณี
สาขาวิชา วิทยาการคอมพิวเตอร์

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	**คณะกรรมการสอบ**
…….....	……..ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.วิภาดา เวทย์ประสิทธิ์)	(ดร.ฐิมาพร เพชรแก้ว)
	……..กรรมการ
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	(ดร.ลัดดา ปรีชาวีรกุล)
…….....	……..กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ศิริรัตน์ วณิชโยบล)	(ผู้ช่วยศาสตราจารย์ ดร.ศิริรัตน์ วณิชโยบล)
	……..กรรมการ
	(ผู้ช่วยศาสตราจารย์ ดร.วิภาดา เวทย์ประสิทธิ์)

บัณฑิตวิทยาลัย มหาวิทยาลัยสงขลานครินทร์ อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษา ตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์

……
(รองศาสตราจารย์ ดร.เกริกชัย ทองหนู)
คณบดีบัณฑิตวิทยาลัย

ชื่อวิทยานิพนธ์ การวิเคราะห์สายจีโนมโดยใช้โครงข่ายประสาทเทียม
ผู้เขียน นางสาวฏรินทร คงมณี
สาขาวิชา วิทยาการคอมพิวเตอร์
ปีการศึกษา 2551

บทคัดย่อ

ความแม่นยำการทำนายจุดเริ่มต้นการแปลรหัสเป็นงานสำคัญในงานวิจัยการวิเคราะห์สายพันธุกรรม วิทยานิพนธ์นี้จึงนำเสนอแบบจำลองการทำนายจุดเริ่มต้นการแปลรหัสโดยใช้วิธี TF-IDF และโครงข่ายประสาทเทียม (The TF-IDF and Neural Networks Approach for Translation Initiation Site Prediction: TF-IDF-NN-TIS) เพื่อเพิ่มค่าความถูกต้องการทำนายจุดเริ่มต้นการแปลรหัส ซึ่งแบบจำลองแบ่งสายพันธุกรรมเป็นสายพันธุกรรมย่อย จากนั้นสร้างลักษณะเฉพาะโดยใช้เทคนิค n-แกรม ทั้งในส่วนอัพสตรีมและดาวน์สตรีมแยกจากกัน กำหนดค่าลักษณะเฉพาะด้วยค่า TF-IDF เลือกลักษณะเฉพาะด้วยเทคนิคการเลือกลักษณะเฉพาะสหสัมพันธ์ ประเมินผลการทดสอบแบบไขว้เปลี่ยน k กลุ่ม โดยแบบจำลองที่นำเสนอมี 5 ขั้นตอน คือ 1) การแบ่งสายพันธุกรรม 2) การสร้างลักษณะเฉพาะ n-แกรม 3) การเลือกลักษณะเฉพาะ 4) การสร้างลักษณะเฉพาะรูปแบบคอนเซนซัส และ 5) การทำนายผลลัพธ์ วิทยานิพนธ์นี้ได้ทำการพัฒนาโปรแกรมโดยใช้ภาษา MATLAB ชุดข้อมูลที่ใช้ในการทดลองประกอบด้วย ชุดข้อมูล Vertebrate ชุดข้อมูล Arabidopsis thaliana และชุดข้อมูล TIS+50 ผลการทดลองแสดงให้เห็นว่าแบบจำลองที่นำเสนอมีประสิทธิภาพในการทำนายจุดเริ่มต้นการแปลรหัส โดยให้ค่าความถูกต้องสูงกว่างานวิจัยที่ศึกษาก่อนหน้า และใช้เวลาในการทำงานน้อย

Thesis Title	Genome Sequences Analysis Using Neural Networks
Author	Miss Tarintorn Kongmanee
Major Program	Computer Science
Academic Year	2008

ABSTRACT

The precise prediction of translation initiation site is an important task for the analysis of genomic sequence. This study aims to increase the accuracy for the prediction of translation initiation site using a TF-IDF-NN-TIS model (The TF-IDF and Neural Networks Approach for Translation Initiation Site Prediction). This study deals with segment genome sequence to subsequence. Then the study creates feature using n-gram techniques for both upstream and downstream. Determining feature value uses TF-IDF approach and feature selection by correlation-based feature selection method. Evaluation prediction results use k-fold cross validation. The TF-IDF-NN-TIS composes of 5 steps; step 1) sequence segmentation, step 2) feature generation with n-gram technique, step 3) feature selection, step 4) feature generation from consensus pattern, and step 5) translation initiation sites prediction. MATLAB has been used for the developing of the program. This study performs experiments on three different datasets that are Vertebrate, Arabidopsis thaliana, and TIS+50. The experimental result indicates that the proposed model has the good efficiency of translation initiation site prediction that the accuracy received has higher value than the previous studied with less processing time.

สารบัญ

หน้า

สารบัญ.....(6)
รายการตาราง.....(9)
รายการภาพประกอบ.....(11)
บทที่ 1 บทนำ..... 1
1.1 การตรวจเอกสาร.....2
1.1.1 อณูชีววิทยา.....2
1.1.2 เทคนิค n-แกรม.....3
1.1.3 การเลือกลักษณะเฉพาะ.....4
1.1.4 โครงข่ายประสาทเทียม.....5
1.2 วัตถุประสงค์ของโครงการ.....5
1.3 ขอบเขตการดำเนินงานและวิธีการดำเนินการวิจัย.....5
1.4 ขั้นตอนการดำเนินงานและระยะเวลาการดำเนินงาน.....6
1.5 สถานที่ และเครื่องมือที่ใช้ในงานวิจัย.....7
1.6 ประโยชน์ที่คาดว่าจะได้รับ.....8
บทที่ 2 ทฤษฎีที่เกี่ยวข้องกับการสร้างแบบจำลองการทำนายจุดเริ่มต้นการแปลรหัสโดยใช้วิธี TF-IDF และโครงข่ายประสาทเทียม.....9
2.1 อณูชีววิทยา.....9
2.1.1 ดีเอ็นเอ.....9
2.1.2 โปรตีน.....11
2.2 ฟาสต์-เอ (FASTA).....12
2.3 เทคนิค n- แกรม (n-gram).....13
2.4 TF-IDF (Term Frequency and Inverse Document Frequency).....13
2.5 เทคนิคการเลือกลักษณะเฉพาะ.....14
2.5.1 เทคนิคการเลือกลักษณะเฉพาะสหสัมพันธ์.....15
2.5.2 เทคนิคไคสแควร์ (Chi-Square).....16
2.5.3 เทคนิคอัตราส่วนเกน (Gain Ratio).....17
2.5.4 เทคนิครีลีฟ-เอฟ (ReliefF).....18
2.6 โครงข่ายประสาทเทียม.....20
2.6.1 สถาปัตยกรรมโครงข่ายประสาทเทียม.....20

สารบัญ (ต่อ)

หน้า

2.6.2 ประเภทของโครงข่ายประสาทเทียม 24
2.6.3 รูปแบบการเรียนรู้ของโครงข่ายประสาทเทียม 25
2.6.4 ขั้นตอนวิธีการส่งค่าย้อนกลับ (Backpropagation Algorithm) 26
2.7 การทดสอบแบบไขว้เปลี่ยน *k* กลุ่ม 27
2.8 การประเมินประสิทธิภาพ 28
บทที่ 3 แบบจำลองการทำนายจุดเริ่มต้นการแปลรหัสโดยใช้วิธี TF-IDF
และโครงข่ายประสาทเทียม 30
3.1 ขั้นตอนการแบ่งสายพันธุกรรม 31
3.2 ขั้นตอนการสร้างลักษณะเฉพาะ n-แกรม 34
3.3 ขั้นตอนการเลือกลักษณะเฉพาะ 37
3.4 ขั้นตอนการสร้างลักษณะเฉพาะรูปแบบคอนเซนซัส 37
3.5 ขั้นตอนการทำนายผลลัพธ์ 38
บทที่ 4 โปรแกรมการทำนายจุดเริ่มต้นการแปลรหัสโดยใช้วิธี TF-IDF
และโครงข่ายประสาทเทียม 40
4.1 ผังการทำงานของโปรแกรม 40
4.1.1 ผังงานโปรแกรมหลัก 40
4.1.2 ผังงานโปรแกรมการแบ่งสายพันธุกรรม 42
4.1.3 ผังงานโปรแกรมการสร้างลักษณะเฉพาะ n-แกรม 43
4.1.4 ผังงานโปรแกรมการเลือกลักษณะเฉพาะ 44
4.1.5 ผังงานโปรแกรมการสร้างลักษณะเฉพาะรูปแบบคอนเซนซัส 45
4.1.6 ผังงานโปรแกรมการทำนายผลลัพธ์ 46
4.2 การทำงานของโปรแกรม 46
4.2.1 ส่วนการนำเข้าข้อมูล 47
4.2.2 ส่วนการสร้างลักษณะเฉพาะ n-แกรม 49
4.2.3 ส่วนการเลือกลักษณะเฉพาะ 50
4.2.4 ส่วนการทำนายผลลัพธ์ 52
4.2.5 ส่วนการแสดงผลการทำงาน 52

สารบัญ (ต่อ)

หน้า

บทที่ 5 ผลการทดลองและบทวิจารณ์ 54
5.1 ชุดข้อมูลที่ใช้ในการทดลอง 54
5.2 การทดลอง 56
5.2.1 การออกแบบการทดลอง 56
5.2.2 ผลการทดลอง 57
บทที่ 6 บทสรุปและข้อเสนอแนะ 93
6.1 สรุปผลงานวิจัย 93
6.2 ปัญหาและอุปสรรค 95
6.3 ข้อเสนอแนะ 96
บรรณานุกรม 97
ภาคผนวก 100
ก ผลงานวิจัยที่ได้รับการตีพิมพ์ในงานประชุมวิชาการ NCSEC 2008 101
ข ผลงานวิจัยที่ได้รับการตีพิมพ์ในงานประชุมวิชาการ IEEE ICCSIT 2009 110
ประวัติผู้เขียน 116

รายการตาราง

ตาราง		หน้า
1.1	ระยะเวลาการดำเนินงาน	7
2.1	ค่าของคอนฟิวชันเมตริกซ์ (Confusion Matrix) แบบ 2 กลุ่ม	28
3.1	ลักษณะเฉพาะที่สร้างจาก 1-แกรม และ 2-แกรม	35
3.2	จำนวนลักษณะเฉพาะสำหรับเทคนิค 1-แกรม 2-แกรม และ 3-แกรม	35
3.3	ลักษณะเฉพาะที่สร้างจากรูปแบบคอนเซนซัส	38
5.1	ขั้นตอนการทดลองของการทดลองแบบ A B C และ D	56
5.2	การออกแบบการทดลอง และสถาปัตยกรรมของโครงข่ายประสาทเทียม	57
5.3	คุณลักษณะของชุดข้อมูล	58
5.4	ลักษณะเฉพาะ 1-แกรม 2-แกรม หรือ 3-แกรมสำหรับการทดลอง	58
5.5	ค่าความถี่ตัวอย่างกลุ่มบวกที่ 1 ถึง 10 ของชุดข้อมูล Vertebrate	60
5.6	ค่า TF-IDF ตัวอย่างกลุ่มบวกที่ 1 ถึง 10 ของชุดข้อมูล Vertebrate	61
5.7	ลักษณะเฉพาะที่ผ่านการเลือกสำหรับค่าความถี่ และค่า TF-IDF ของชุดข้อมูล Vertebrate	63
5.8	ลักษณะเฉพาะที่ผ่านการเลือกสำหรับค่าความถี่ และค่า TF-IDF ของชุดข้อมูล A.thaliana	63
5.9	ลักษณะเฉพาะที่ผ่านการเลือกสำหรับค่าความถี่ และค่า TF-IDF ของชุดข้อมูล TIS+50	64
5.10	ลักษณะเฉพาะที่ผ่านการเลือกสำหรับค่า TF-IDF ของชุดข้อมูล Vertebrate	65
5.11	ลักษณะเฉพาะที่ผ่านการเลือกสำหรับค่า TF-IDF ของชุดข้อมูล A.thaliana	66
5.12	ลักษณะเฉพาะที่ผ่านการเลือกสำหรับค่า TF-IDF ของชุดข้อมูล TIS+50	67
5.13	ลักษณะเฉพาะที่สร้างจากรูปแบบคอนเซนซัส	69
5.14	ค่าความถูกต้องของขนาดหน้าต่างที่แตกต่างกันสำหรับชุดข้อมูล Vertebrate	70
5.15	ค่าความถูกต้องของขนาดหน้าต่างที่แตกต่างกันสำหรับชุดข้อมูล A.thaliana	70
5.16	ค่าความถูกต้องของขนาดหน้าต่างที่แตกต่างกันสำหรับชุดข้อมูล TIS+50	71
5.17	ค่าความถูกต้องลักษณะเฉพาะ n-แกรมสำหรับข้อมูล Vertebrate	72
5.18	ค่าการตอบสนองไวของลักษณะเฉพาะ n-แกรม	73
5.19	ค่าความเฉพาะเจาะจงของลักษณะเฉพาะ n-แกรม	74
5.20	ค่าความถูกต้องของจำนวนลักษณะเฉพาะที่มีลำดับนัยสำคัญแตกต่างกันของเทคนิคการเลือกลักษณะเฉพาะแบบตัวกรองสำหรับข้อมูล Vertebrate	82

รายการตาราง (ต่อ)

ตาราง **หน้า**

5.21 ค่าความถูกต้องของจำนวนลักษณะเฉพาะที่มีลำดับนัยสำคัญแตกต่างกัน ของเทคนิคการเลือกลักษณะเฉพาะแบบตัวกรองสำหรับข้อมูล A.thaliana 82

5.22 ค่าความถูกต้องของจำนวนลักษณะเฉพาะที่มีลำดับนัยสำคัญแตกต่างกัน ของเทคนิคการเลือกลักษณะเฉพาะแบบตัวกรองสำหรับข้อมูล TIS+50 83

5.23 ค่าความถูกต้องของเทคนิคการเลือกลักษณะเฉพาะ 83

5.24 ค่าการตอบสนองไวของเทคนิคการเลือกลักษณะเฉพาะ 84

5.25 ค่าความเฉพาะเจาะจงของเทคนิคการเลือกลักษณะเฉพาะ 85

5.26 ค่าความถูกต้องของการทดลองแบบ A B C และ D 90

5.27 ค่าการตอบสนองไวของการทดลองแบบ A B C และ D 90

5.28 ค่าความเฉพาะเจาะจงของการทดลองแบบ A B C และ D 91

5.29 ค่าความถูกต้องของแบบจำลอง TF-IDF-NN-TIS และงานวิจัยที่ศึกษาก่อนหน้า 92

รายการภาพประกอบ

ภาพประกอบ		หน้า
1.1	แบบจำลองการทำนายจุดเริ่มต้นการแปลรหัสโดยใช้โครงข่ายประสาทเทียม	6
2.1	โครงสร้างของดีเอ็นเอ	10
2.2	โครงสร้างพื้นฐานการสร้างโปรตีนของเซลล์ยูคารีโอต	12
2.3	บริบทโดยรอบจุดเริ่มต้นการแปลรหัส	12
2.4	โครงสร้างการจัดเก็บข้อมูลแบบฟาสต์-เอ	13
2.5	ขั้นตอนวิธีรีลีฟพื้นฐาน	18
2.6	ขั้นตอนวิธีรีลีฟ-เอฟ	19
2.7	องค์ประกอบของหน่วยประมวลผลย่อย	21
2.8	ฟังก์ชันสเตป	22
2.9	ฟังก์ชันเชิงเส้น	22
2.10	ฟังก์ชันล็อกซิกมอยด์	23
2.11	สถาปัตยกรรมแบบหน่วยประมวลผลย่อยหลายชั้น	24
2.12	โครงข่ายแบบไปข้างหน้าชั้นเดียว	24
2.13	โครงข่ายแบบไปข้างหน้าหลายชั้น	25
2.14	โครงข่ายแบบย้อนกลับ	25
2.15	ขั้นตอนวิธีการส่งค่าย้อนกลับ	27
3.1	แบบจำลองการทำนายจุดเริ่มต้นการแปลรหัสโดยใช้วิธี TF-IDF และโครงข่ายประสาทเทียม	31
3.2	ขั้นตอนการแบ่งสายพันธุกรรม	32
3.3	ตัวอย่างข้อมูลสายพันธุกรรม	33
3.4	การแบ่งสายพันธุกรรมด้วยขนาดหน้าต่าง 303 นิวคลีโอไทด์	33
3.5	ขั้นตอนการสร้างลักษณะเฉพาะ n-แกรม	34
3.6	ขั้นตอนการเลือกลักษณะเฉพาะ	37
3.7	ขั้นตอนการสร้างลักษณะเฉพาะรูปแบบคอนเซนซัส	38
3.8	ขั้นตอนการทำนายผลลัพธ์	39
4.1	ผังงานโปรแกรมหลัก	41
4.2	ผังงานโปรแกรมการแบ่งสายพันธุกรรม	42
4.3	ผังงานโปรแกรมการสร้างลักษณะเฉพาะ n-แกรม	43
4.4	ผังงานโปรแกรมการเลือกลักษณะเฉพาะ	44

รายการภาพประกอบ (ต่อ)

ภาพประกอบ **หน้า**

4.5 ผังงานโปรแกรมการสร้างลักษณะเฉพาะรูปแบบคอนเซนซัส 46
4.6 ผังงานโปรแกรมการทำนายผลลัพธ์ 46
4.7 หน้าจอหลักของโปรแกรมการทำนายจุดเริ่มต้นการแปลรหัสโดยใช้
วิธี TF-IDF และโครงข่ายประสาทเทียม 47
4.8 หน้าต่างการเลือกแหล่งข้อมูลสายพันธุกรรม 48
4.9 ข้อมูลสายพันธุกรรม 48
4.10 รูปแบบสายพันธุกรรม 49
4.11 ส่วนการสร้างลักษณะเฉพาะ 49
4.12 รูปแบบและค่าของลักษณะเฉพาะ 50
4.13 ส่วนการเลือกลักษณะเฉพาะ 51
4.14 ข้อมูลเข้าของโครงข่ายประสาทเทียม 51
4.15 ส่วนการทำนายผลลัพธ์ 52
4.16 ผลลัพธ์การทำนายของโครงข่ายประสาทเทียม 53
5.1 ตัวอย่างสายพันธุกรรมของชุดข้อมูล Vertebrate 54
5.2 ตัวอย่างสายพันธุกรรมของชุดข้อมูล Arabidopsis thaliana 55
5.3 ตัวอย่างสายพันธุกรรมของชุดข้อมูล TIS+50 55
5.4 เปรียบเทียบค่าความถูกต้องของขนาดหน้าต่างที่แตกต่างกัน
สำหรับชุดข้อมูล Vertebrate 70
5.5 เปรียบเทียบค่าความถูกต้องของขนาดหน้าต่างที่แตกต่างกัน
สำหรับชุดข้อมูล A.thaliana 71
5.6 เปรียบเทียบค่าความถูกต้องของขนาดหน้าต่างที่แตกต่างกัน
สำหรับชุดข้อมูล TIS+50 72
5.7 เปรียบเทียบค่าความถูกต้องของลักษณะเฉพาะ n-แกรม 73
5.8 เปรียบเทียบค่าการตอบสนองไวของลักษณะเฉพาะ n-แกรม 74
5.9 เปรียบเทียบค่าความเฉพาะเจาะจงของลักษณะเฉพาะ n-แกรม 75
5.10 เปรียบเทียบค่าความถูกต้อง ระหว่างค่าความถี่ และค่า TF-IDF
สำหรับชุดข้อมูล Vertebrate 76
5.11 เปรียบเทียบค่าการตอบสนองไวระหว่างค่าความถี่ และค่า TF-IDF
สำหรับชุดข้อมูล Vertebrate 76

รายการภาพประกอบ (ต่อ)

ภาพประกอบ **หน้า**

5.12 เปรียบเทียบค่าความเฉพาะเจาะจงระหว่างค่าความถี่ และค่า TF-IDF สำหรับชุดข้อมูล Vertebrate 76
5.13 เปรียบเทียบค่าความถูกต้องระหว่างค่าความถี่ และค่า TF-IDF สำหรับชุดข้อมูล A.thaliana 77
5.14 เปรียบเทียบค่าการตอบสนองไวระหว่าง ค่าความถี่ และค่า TF-IDF สำหรับชุดข้อมูล A.thaliana 77
5.15 เปรียบเทียบค่าความเฉพาะเจาะจงระหว่างค่าความถี่ และค่า TF-IDF สำหรับชุดข้อมูล A.thaliana 78
5.16 เปรียบเทียบค่าความถูกต้องระหว่างค่าความถี่ และค่า TF-IDF สำหรับชุดข้อมูล TIS+50 78
5.17 เปรียบเทียบค่าการตอบสนองไวระหว่าง ค่าความถี่ และค่า TF-IDF สำหรับชุดข้อมูล TIS+50 79
5.18 เปรียบเทียบค่าความเฉพาะเจาะจงระหว่างค่าความถี่ และค่า TF-IDF สำหรับชุดข้อมูล TIS+50 79
5.19 เปรียบเทียบเวลาการสร้างแบบจำลองระหว่างค่าความถี่ และค่า TF-IDF สำหรับชุดข้อมูล Vertebrate 80
5.20 เปรียบเทียบเวลาการสร้างแบบจำลองระหว่างค่าความถี่ และค่า TF-IDF สำหรับชุดข้อมูล A.thaliana 80
5.21 เปรียบเทียบเวลาการสร้างแบบจำลองระหว่างค่าความถี่ และค่า TF-IDF สำหรับชุดข้อมูล TIS+50 81
5.22 เปรียบเทียบค่าความถูกต้องของเทคนิคการเลือกลักษณะเฉพาะ 84
5.23 เปรียบเทียบค่าการตอบสนองไวของเทคนิคการเลือกลักษณะเฉพาะ 84
5.24 เปรียบเทียบค่าความเฉพาะเจาะจงของเทคนิคการเลือกลักษณะเฉพาะ 85
5.25 เปรียบเทียบค่าความถูกต้องระหว่างการทดลองแบบ A และ C 86
5.26 เปรียบเทียบค่าการตอบสนองไวระหว่างการทดลองแบบ A และ C 87
5.27 เปรียบเทียบค่าความเฉพาะเจาะจงระหว่างการทดลองแบบ A และ C 87
5.28 เปรียบเทียบค่าความถูกต้องระหว่างการทดลองแบบ B และ D 88
5.29 เปรียบเทียบค่าการตอบสนองไวระหว่างการทดลองแบบ B และ D 88
5.30 เปรียบเทียบค่าความเฉพาะเจาะจงระหว่างการทดลองแบบ B และ D 89

รายการภาพประกอบ (ต่อ)

ภาพประกอบ **หน้า**

5.31 เปรียบเทียบค่าความถูกต้องระหว่างการทดลองแบบ A B C และ D 90
5.32 เปรียบเทียบค่าการตอบสนองไวระหว่างการทดลองแบบ A B C และ D 91
5.33 เปรียบเทียบค่าความเฉพาะเจาะจงระหว่างการทดลองแบบ A B C และ D 91

บทที่ 1

บทนำ

ปัจจุบันคอมพิวเตอร์ได้เข้ามามีบทบาทในการพัฒนาและช่วยเพิ่มประสิทธิภาพการทำงานของมนุษย์โดยเฉพาะอย่างยิ่งงานทางด้านวิทยาศาสตร์ชีวภาพ ปัญหาสำคัญของงานทางด้านวิทยาศาสตร์ชีวภาพที่เกี่ยวข้องกับข้อมูลพันธุศาสตร์ และอณูชีววิทยา คือ ใช้เวลาการทำงานในห้องปฏิบัติการ (Wet Lab) นาน ใช้งบประมาณในการวิจัยสูง มีขั้นตอนในการศึกษาที่ซับซ้อน และต้องการความชำนาญของผู้เชี่ยวชาญเฉพาะด้าน เพื่อให้ได้ข้อมูลหรือผลการศึกษาที่ต้องการ จึงได้มีการนำศาสตร์ทางด้านชีวสารสนเทศศาสตร์ (Bioinformatics) เข้ามาช่วยในขั้นตอนการดำเนินงานในห้องปฏิบัติการ ชีวสารสนเทศศาสตร์เป็นสาขาที่ใช้ความรู้จากคณิตศาสตร์ สถิติ สารสนเทศ และวิทยาการคอมพิวเตอร์ เพื่อแก้ปัญหาทางชีววิทยา (Dana, 2006; Ray et *al.*, 2005)

สิ่งมีชีวิตทั้งหมดมีโปรตีนเป็นองค์ประกอบส่วนใหญ่ โปรตีนมีบทบาทสำคัญยิ่งในการเป็นตัวกำหนดหน้าที่ และรูปร่างของเซลล์ในแต่ละส่วนของ การวิเคราะห์สายพันธุกรรม (Genome Sequence Analysis) ทำให้ทราบบริเวณที่มีรหัสสำหรับการสร้างโปรตีน (Coding Region) จึงเป็นงานที่นักวิจัยให้ความสำคัญ การหาบริเวณที่มีรหัสสำหรับการสร้างโปรตีนสามารถทำได้โดยการหา Open Reading Frame หรือ ORF ที่เริ่มต้นด้วยโคดอนเริ่มต้น (Start Codon) หรือจุดเริ่มต้นการแปลรหัส (Translation Initiation Sites: TIS) และสิ้นสุดที่โคดอนหยุด (Stop Codon) จากนั้นต้องการทำการวิเคราะห์ว่า ORF ที่ได้เป็นยีนที่สามารถแปลรหัสเป็นโปรตีน หรือเป็นเพียงลำดับนิวคลีโอไทด์ช่วงหนึ่งที่ตั้งต้นด้วยโคดอนเริ่มต้น และสิ้นสุดที่โคดอนหยุด (บุรชัย สนธยานนท์, 2542) การตรวจสอบว่า ORF ที่หาได้นั้นเป็นบริเวณที่มีรหัสสำหรับการสร้างโปรตีนอาจทำได้โดยการหาบริเวณที่เกี่ยวข้องกับการควบคุมการแสดงออกของยีน ตัวอย่างเช่น จุดเริ่มต้นการแปลรหัส บริเวณโปรโมเตอร์ (Promoter) หรือบริเวณที่เกี่ยวข้องกับขั้นตอนการเชื่อมต่ออาร์เอ็นเอ (Splice Junction Site)

การประยุกต์ใช้เทคนิคโครงข่ายประสาทเทียม (Artificial Neural Networks) เป็นแนวทางหนึ่งที่นำมาช่วยในการดำเนินการกับข้อมูลปริมาณมากที่อยู่ในรูปของตัวเลขได้อย่างมีประสิทธิภาพ เนื่องจากเทคนิคโครงข่ายประสาทเทียมมีการทำงานโดยเลียนแบบการประมวลในสมองมนุษย์ การทำงานภายในเป็นแบบขนาน สามารถทำการเรียนรู้ชุดข้อมูลตัวอย่างทั้งการเรียนรู้แบบมีผู้สอน (Supervised Learning) หรือ การเรียนรู้แบบไม่มีผู้สอน

(Unsupervised Learning) โดยที่ไม่จำเป็นต้องทราบถึงความสัมพันธ์ของข้อมูลต่างๆ อีกทั้งยังสามารถแบ่งกลุ่มข้อมูลและทำการทำนายผลลัพธ์ที่จะเกิดขึ้นได้ (Kantardzic, 2003)

งานวิจัยนี้ประยุกต์ใช้โครงข่ายประสาทเทียมสำหรับการทำนายจุดเริ่มต้นการแปลรหัสของสายพันธุกรรมโดยอัตโนมัติ เพื่อให้ได้ค่าความถูกต้องสูง และใช้เวลาประมวลผลน้อย

1.1 การตรวจเอกสาร

งานวิจัยที่เกี่ยวข้องกับงานวิจัยนี้แบ่งออกได้เป็น 3 ส่วน คือ อณูชีววิทยา (Molecular Biology) เทคนิค n-แกรม (n-Gram) การเลือกลักษณะเฉพาะ (Feature Selection) และโครงข่ายประสาทเทียม (Neural Networks)

1.1.1 อณูชีววิทยา

อณูชีววิทยา เน้นการศึกษาโครงสร้างและการทำงานของยีน (Gene) ซึ่งเป็นรหัสพันธุกรรมบนสายพันธุกรรม ในระดับต่างๆ จนได้เป็นโปรตีน

1) ดีเอ็นเอ (DNA) คือ ชื่อย่อของสารพันธุกรรมที่มีชื่อทางวิทยาศาสตร์ว่า “กรดดีออกซีไรโบนิวคลีอิก” (Deoxyribonucleic Acid: DNA) เป็นสารพันธุกรรมที่กำหนดรหัสพันธุกรรมของสิ่งมีชีวิตทุกชนิด ได้แก่ คน สัตว์ พืช เชื้อรา แบคทีเรีย และไวรัส เป็นต้น ดีเอ็นเอประกอบด้วย นิวคลีโอไทด์หลาย ๆ นิวคลีโอไทด์มาเรียงต่อกัน แต่ละนิวคลีโอไทด์ประกอบด้วยน้ำตาล ฟอสเฟต และเบส ซึ่งเบสในดีเอ็นเอมี 4 ชนิด ได้แก่ เบสอะดีนิน (Adenine: A) เบสไทมีน (Thymine: T) เบสไซโตซีน (Cytosine: C) และเบสกัวนีน (Guanine: G) ตามลำดับ (อมรา คัมภิรานนท์, 2542; ประดิษฐ์ พงศ์ทองคำ, 2543) ยีนเป็นสายดีเอ็นเอสั้น ๆ ที่กำหนดและควบคุมลักษณะต่าง ๆ ของสิ่งมีชีวิต ทำหน้าที่กำหนดชนิดของโปรตีนที่เซลล์สังเคราะห์ขึ้นเพื่อนำไปใช้ในกิจกรรมภายในเซลล์

2) โปรตีน (Protein) เป็นสารอินทรีย์ซึ่งพบได้ในสิ่งมีชีวิตทุกชนิด มีโครงสร้างซับซ้อนและมีมวลโมเลกุลมากโปรตีนมีหน่วยย่อยคือ กรดอะมิโน เรียงต่อกันด้วยพันธะเปปไทด์ โปรตีนมีหน้าที่สำคัญต่อโครงสร้างและกิจกรรมภายในเซลล์ของสิ่งมีชีวิตทุกชนิด กระบวนการสร้างโปรตีนมี 2 กระบวนการคือ กระบวนการถ่ายสำเนา (Transcription) และกระบวนการแปลรหัส (Translation) โดยเริ่มต้นเป็นการนำข้อมูลที่เก็บอยู่ในดีเอ็นเอถ่ายสำเนาไปเป็นอาร์เอ็นสื่อสาร หรือเอ็มอาร์เอ็นเอ (messenger RNA: mRNA) จากนั้น mRNA ก็เข้าสู่กระบวนการแปลรหัสไปเป็นโปรตีน (อมรา คัมภิรานนท์, 2542) รหัสพันธุกรรมของกรดอะมิโนหนึ่งตัวประกอบด้วย 3 นิวคลีโอไทด์ เรียกว่า โคดอน (Codon) กระบวนการแปลรหัสพันธุกรรม

บนสาย mRNA ไปเป็นโปรตีนจะเริ่มต้นที่จุดเริ่มต้นการแปลรหัส (บุรชัย และคณะ, 2542; อุไรวรรณ วิจารณกุล, 2545) มีรหัสพันธุกรรมเป็น โคดอน ATG สำหรับกระบวนการแปลรหัสพันธุกรรมตามแบบจำลองการตรวจสอบไรโบโซม (Kozak, 1978; Cigan *et al.*, 1988; Kozak, 1989) สมมติให้ไรโบโซมเริ่มต้นตรวจสอบสาย mRNA จากปลาย *5′* ไปยังปลาย *3′* จนกระทั่งเจอจุดเริ่มต้นการแปลรหัสพันธุกรรมจึงเริ่มต้นแปลรหัสพันธุกรรม และหยุดการแปลรหัสพันธุกรรมเมื่อเจอโคดอนหยุด (Stop Codon) ได้แก่ โคดอน TAA TAG หรือ TGA

การดำเนินการวิเคราะห์สายพันธุกรรมเพื่อระบุจุดเริ่มต้นการแปลรหัสเป็นงานที่นักวิจัยให้ความสำคัญ งานวิจัยแรกนำเสนอการระบุจุดเริ่มต้นการแปลรหัสของสายดีเอ็นเอ E. coli (Stormo *et al.*, 1982) ต่อมาประยุกต์ใช้เทคนิคเมตริกซ์น้ำหนักเพื่อค้นหารูปแบบโดยทั่วไปที่อยู่รอบ ๆ จุดเริ่มต้นการแปลรหัสในเซลล์ยูคารีโอต โดยค้นพบลักษณะเด่น คือ GCC[**AG**]CC*ATG***G** เรียกว่ารูปแบบคอนเซนซัส (Kozak, 1987) ซึ่งเป็นจุดเริ่มต้นของการนำวิธีการทางสถิติเข้ามาช่วยในการทำนายจุดเริ่มต้นการแปลรหัส ปี ค.ศ. 1997 เริ่มมีการประยุกต์โครงข่ายประสาทเทียมสำหรับการทำนายจุดเริ่มต้นการแปลรหัสในสายดีเอ็นเอสัตว์มีกระดูกสันหลังและสายดีเอ็นเอของพืช โดยสร้างชุดข้อมูล Vertebrate และชุดข้อมูล Arabidopsis thaliana (Pedersen and Neilsen, 1997) ตามลำดับ ชุดข้อมูลทั้งสองได้จากฐานข้อมูล GenBank ของ National Center of Biotechnology Information หรือ NCBI (NCBI, 1988) Hatxigeorgiou (Hatzigeorgiou, 2002) ประยุกต์โครงข่ายประสาทเทียมตามแบบจำลองการตรวจสอบไรโบโซมเพื่อทำนายจุดเริ่มต้นการแปลรหัสของมนุษย์ สำหรับการประยุกต์ขั้นตอนวิธีเรียนรู้ของเครื่องอื่นๆ เพื่อการทำนายจุดเริ่มต้นการแปลรหัสด้วย ได้แก่ การประยุกต์ Support Vector Machine สำหรับการทำนายจุดเริ่มต้นการแปลรหัสของสัตว์มีกระดูกสันหลังโดยใช้ชุดข้อมูล Vertebrate (Zien *et al.*, 2000) การประยุกต์เทคนิคเหมืองข้อมูลหลายๆ เทคนิคเพื่อปรับปรุงประสิทธิภาพการทำนายจุดเริ่มต้นการแปลรหัสของสัตว์มีกระดูกสันหลังโดยใช้ชุดข้อมูล Vertebrate (Zeng *et al.*, 2002) เพื่อพัฒนาประสิทธิภาพการทำนายจุดเริ่มต้นการแปลรหัสให้ดียิ่งขึ้นจึงได้มีการพัฒนาเทคนิคการเข้ารหัสแบบ n-แกรม ของสายโปรตีนร่วมกับเทคนิค Support Vector Machine และต้นไม้ตัดสินใจเพื่อทำนายจุดเริ่มต้นการแปลรหัส (Liu *et al.*, 2004) Tzanis และคณะ (Tzanis *et al.*, 2005; Tzanis *et al.*, 2006; Tzanis and Vlahavas, 2006) นำเสนอลักษณะเฉพาะใหม่ซึ่งแตกต่างจากลักษณะเฉพาะของงานวิจัยที่ศึกษาก่อนหน้าเพื่อการทำนายจุดเริ่มต้นการแปลรหัสด้วยขนาดหน้าต่าง 201 นิวคลีโอไทด์ เทคนิคอื่นๆ สำหรับการทำนายจุดเริ่มต้นการแปลรหัส ตัวอย่างเช่น แบบจำลองเกาส์เซียนผสม (Li *et al.*, 2005) เป็นต้น

1.1.2 เทคนิค n-แกรม (n-gram Technique)

เนื่องจากข้อมูลสายพันธุกรรมอยู่ในรูปของสายอักขระ (Text Sequences) นั่นคือ เป็นการเรียงของลำดับนิวคลีโอไทด์ A C G และ T บางครั้งอาจไม่สามารถนำมาใช้ในการ

วิเคราะห์ หรือการประมวลผลข้อมูลด้วยขั้นตอนวิธีทางคอมพิวเตอร์บางขั้นตอนวิธีจึงจำเป็นต้องแปลงข้อมูลให้อยู่ในรูปที่เหมาะสมสำหรับการคำนวณในขั้นตอนวิธีทางคอมพิวเตอร์ การเปลี่ยนรูปแบบลำดับนิวคลีโอไทด์เป็นข้อมูลเชิงตัวเลขทำโดยสร้างรูปแบบของ n-แกรมที่เป็นไปได้จากนั้นนับความถี่ของแต่ละแกรมที่ปรากฏในสายพันธุกรรม

นักวิจัย (Zeng *et al.*, 2002) ประยุกต์เทคนิค n-แกรมของสายดีเอ็นเอเพื่อระบุจุดเริ่มต้นการแปลรหัสสำหรับชุดข้อมูล Vertebrate โดยกำหนดค่า *n* เท่ากับ 3, 4, และ 5 ได้ค่าความถูกต้องเท่ากับ 90.00% งานวิจัยนี้ถูกพัฒนาต่อโดย Liu และคณะ (Liu *et al.*, 2004) ซึ่งใช้วิธีเข้ารหัส n-แกรมของกรดอะมิโนร่วมกับ Support Vector Machine เพื่อทำนายจุดเริ่มต้นการแปลรหัสของชุดข้อมูล Vertebrate ได้ค่าความถูกต้องเท่ากับ 92.45% ในปี ค.ศ. 2005 Tzanis และคณะ (Tzanis, 2005) ประยุกต์เทคนิค n-แกรม กำหนดค่า *n* เท่ากับ 3 ถึง 6 พบว่าค่า *n* ที่มากกว่าหรือเท่ากับ 3 ไม่ช่วยเพิ่มประสิทธิภาพการทำนายจุดเริ่มต้นการแปลรหัสของชุดข้อมูล Vertebrate

1.1.3 การเลือกลักษณะเฉพาะ

ข้อมูลเชิงตัวเลขจากลักษณะเฉพาะที่แตกต่างกันสามารถวัดความแตกต่างระหว่างข้อมูลกลุ่มบวกและกลุ่มลบได้ เมื่อข้อมูลกลุ่มบวกหมายถึงข้อมูลที่เป็นจริง และข้อมูลกลุ่มลบหมายถึงข้อมูลที่เป็นเท็จ ซึ่งสามารถวัดได้ชัดเจนถ้าใช้ลักษณะเฉพาะจำนวนมากขึ้น อย่างไรก็ตามเมื่อจำนวนลักษณะเฉพาะเพิ่ม ก็อาจจะมีข้อมูลรบกวน และข้อมูลซ้ำซ้อนเพิ่มด้วยเช่นกัน ทำให้ประสิทธิภาพการวัดความแตกต่างระหว่างข้อมูลกลุ่มบวกและกลุ่มลบลดลง และการประมวลช้าลง เพื่อลดข้อมูลรบกวนและข้อมูลซ้ำซ้อนและใช้ประโยชน์จากลักษณะเฉพาะได้อย่างเต็มที่ จึงประยุกต์เทคนิคการเลือกลักษณะเฉพาะเพื่อรักษาลักษณะเฉพาะที่มีนัยสำคัญสูงสุดต่อการวัดความแตกต่างระหว่างกลุ่มข้อมูล ข้อมูลที่ผ่านการเลือกลักษณะเฉพาะแล้วจะถูกแบ่งออกเป็นสองส่วน ได้แก่ ส่วนแรกใช้ในกระบวนการค้นหารูปแบบ หรือความสัมพันธ์จากข้อมูล เรียกข้อมูลส่วนนี้ว่า ชุดสอน (Training Set) และส่วนที่สองใช้ตรวจสอบความถูกต้องของรูปแบบ เรียกข้อมูลส่วนนี้ว่า ชุดทดสอบ (Test Set) เทคนิคที่นำมาใช้ในการเลือกลักษณะเฉพาะได้แก่ เทคนิคการเลือกลักษณะเฉพาะสหสัมพันธ์ (Hall and Smith, 1997; Hall, 2000) เทคนิคไคสแควร์ (Li and Leong, 2005) เทคนิคอัตราส่วนเกน (Kantardzic, 2003) และเทคนิครีลีฟ-เอฟ (Marko and Igor, 2003)

ตัวอย่างงานวิจัยของเทคนิคการเลือกลักษณะเฉพาะ เช่น การประยุกต์เทคนิคการเลือกลักษณะเฉพาะสหสัมพันธ์เพื่อค้นหาลักษณะเฉพาะสำคัญที่ได้จากเทคนิค n-แกรม เพื่อทำนายจุดเริ่มต้นการแปลรหัสในสัตว์มีกระดูกสันหลังโดยใช้ชุดข้อมูล Vertebrates (Zeng *et al.*, 2002) การประยุกต์เทคนิคไคสแควร์ และเทคนิคอัตราส่วนเกน เพื่อเลือกลักษณะเฉพาะที่สร้างตามแบบจำลองต้นแบบการตรวจสอบไรโบโซม (Tzanis *et al.*, 2006) เป็นต้น

1.1.4 โครงข่ายประสาทเทียม

โครงข่ายประสาทเทียม คือ เทคนิคการทำเหมืองข้อมูลเพื่อค้นหาความรู้ที่แฝงอยู่ในฐานข้อมูล มีรูปแบบการประมวลผลที่เลียนแบบการทำงานของเซลล์ประสาทของมนุษย์ ประกอบด้วย หน่วยประมวลผลย่อยหลายหน่วยเชื่อมต่อกันเป็นโครงข่าย

โครงข่ายประสาทเทียมสำหรับการวิเคราะห์สายพันธุกรรม ตัวอย่างเช่น การใช้โครงข่ายประสาทเทียมเพื่อทำนายจุดเริ่มต้นการแปลรหัสสำหรับชุดข้อมูล Vertebrate (Pedersen and Nielsen, 1997) การพัฒนาซอฟต์แวร์ DIANA-TIS ซึ่งใช้โครงข่ายประสาทเทียม 2 โครงข่ายสำหรับทำนายจุดเริ่มต้นการแปลรหัสสำหรับสายพันธุกรรมมนุษย์ โดยใช้หลักการของต้นแบบการตรวจสอบไรโบโซม (Hatzigeorgiou, 2002) การใช้โครงข่ายประสาทเทียมร่วมกับเทคนิคการจำแนกประเภทอื่น เพื่อปรับปรุงการทำนายจุดเริ่มต้นการแปลรหัสสำหรับข้อมูล Vertebrate (Zeng *et al*., 2002) Liu และคณะ (Liu *et al.*, 2006) ศึกษาปัญหาการระบุลักษณะเด่นของสายพันธุกรรมและสายโปรตีนด้วยโครงข่ายประสาทเทียม Conilione และ Wang (Conilione and Wang, 2005) ใช้โครงข่ายประสาทเทียมเพื่อค้นหาลักษณะเด่นของตำแหน่งโปรโมเตอร์ในสายพันธุกรรมของ E. coli ซึ่งแบบจำลองที่ได้สามารถระบุลักษณะเด่นที่กลายพันธ์ได้ดี และทำงานได้ดีเมื่อสายดีเอ็นเอมีความยาวมากๆ นักวิจัย (Tan *et al.*, 2006) ประยุกต์โครงข่ายประสาทเทียมฟัซซี่เพื่อทำนายจุดเริ่มต้นการแปลรหัสสำหรับชุดข้อมูล Vertebrate

1.2 วัตถุประสงค์

1.2.1 สร้างแบบจำลองการวิเคราะห์สายจีโนมด้วยโครงข่ายประสาทเทียม

1.2.2 พัฒนาโปรแกรมการวิเคราะห์สายจีโนมด้วยโครงข่ายประสาทเทียม

1.3 ขอบเขตการดำเนินงาน

1.3.1 ออกแบบและสร้างแบบจำลองการวิเคราะห์สายจีโนมด้วยโครงข่ายประสาทเทียม

1.3.2 พัฒนาโปรแกรมเพื่อใช้ในการวิเคราะห์สายจีโนมด้วยโครงข่ายประสาทเทียม

1.3.3 ทดสอบโปรแกรมการวิเคราะห์สายจีโนมด้วยโครงข่ายประสาทเทียมด้วยชุดข้อมูล 3 ชุดข้อมูลได้แก่

1) ชุดข้อมูล Vertebrate

2) ชุดข้อมูล Arobidopsis thaliana

3) ชุดข้อมูล TIS+50

1.4 ขั้นตอนและระยะเวลาการดำเนินงาน

1.4.1 ขั้นตอนการดำเนินงาน มีรายละเอียดดังนี้

1) ศึกษางานวิจัยและเอกสารที่เกี่ยวข้องกับการวิเคราะห์สายจีโนมเพื่อระบุตำแหน่งยีนจากจุดเริ่มต้นการแปลรหัส

2) ศึกษาเทคโนโลยีและเครื่องมือสนับสนุน

3) วิเคราะห์และออกแบบแบบจำลองการทำนายจุดเริ่มต้นการแปลรหัสโดยใช้โครงข่ายประสาทเทียมซึ่งมีขั้นตอนวิธีดังภาพประกอบ 1.1

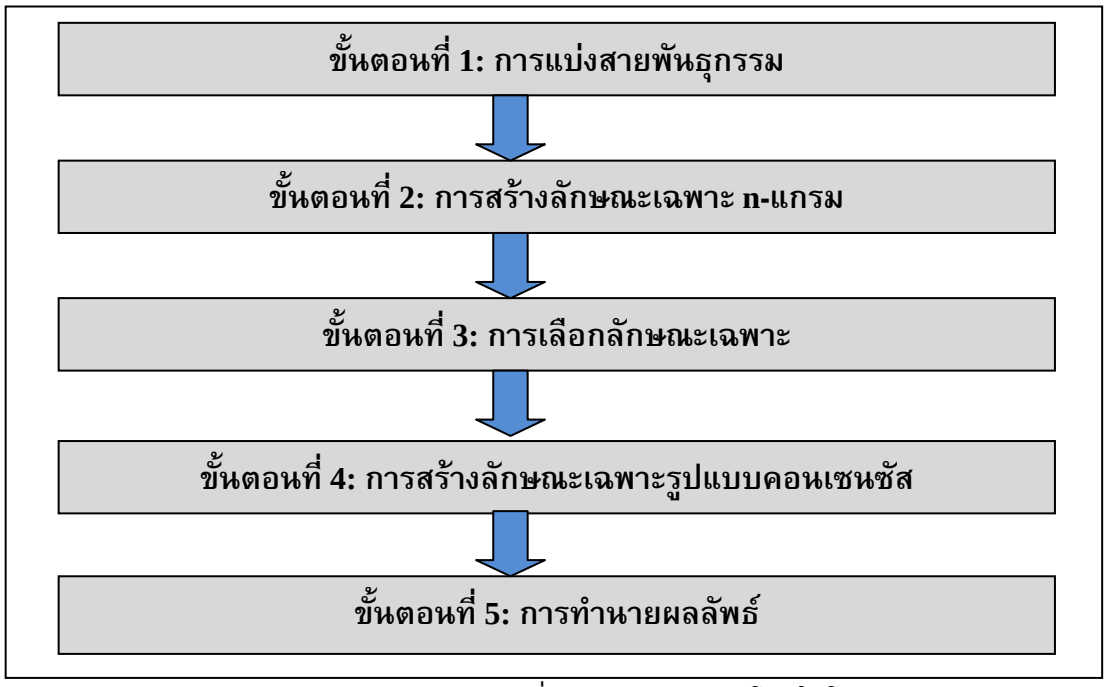
4) เตรียมข้อมูลสำหรับทดสอบโปรแกรมการทำนายจุดเริ่มต้นการแปลรหัสโดยใช้โครงข่ายประสาทเทียม

5) พัฒนาโปรแกรมการทำนายจุดเริ่มต้นการแปลรหัสโดยใช้โครงข่ายประสาทเทียม

6) ทดสอบและติดตั้งโปรแกรม

7) จัดทำเอกสารประกอบโปรแกรมและเขียนผลงานวิจัย

8) จัดทำเอกสารวิทยานิพนธ์



ภาพประกอบ 1.1 แบบจำลองการทำนายจุดเริ่มต้นการแปลรหัสโดยใช้โครงข่ายประสาทเทียม

1.4.2 ระยะเวลาการดำเนินงานสามารถแสดงดังตารางที่ 1.1

ตารางที่ 1.1 ระยะเวลาการดำเนินงาน

กิจกรรม/ขั้นตอนการดำเนินงาน	เดือน															
	2551												2552			
	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4
1. ศึกษางานวิจัยและเอกสารที่เกี่ยวข้อง	■	■	■	■	■	■	■	■	■	■	■	■				
2. ศึกษาเทคโนโลยีและเครื่องมือสนับสนุน			■	■	■	■	■	■	■	■	■	■				
3. วิเคราะห์และออกแบบ				■	■	■	■	■	■							
4. เตรียมข้อมูล				■	■	■	■	■								
5. พัฒนาโปรแกรม							■	■	■	■	■	■				
6. ทดสอบและติดตั้งโปรแกรม											■	■	■	■		
7. จัดทำเอกสารและเขียนผลงานวิจัย													■	■	■	
8. จัดทำเอกสารวิทยานิพนธ์													■	■	■	■

1.5 สถานที่ และเครื่องมือ

1.5.1 สถานที่

ห้องปฏิบัติการวิจัยปัญญาประดิษฐ์ CS207 ภาควิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่

1.5.2 เครื่องมือและอุปกรณ์

1) ด้านฮาร์ดแวร์

เครื่องคอมพิวเตอร์ส่วนบุคคล หน่วยความจำหลัก 1 กิกะไบต์ ฮาร์ดดิสก์ 80 กิกะไบต์ หน่วยประมวลผลกลางรุ่น Intel(R) Core(TM) 2 ความเร็วการประมวลผล 1.86 กิกะเฮิร์ต จำนวน 1 เครื่อง สำหรับพัฒนาโปรแกรม และทดสอบ

2) ด้านซอฟต์แวร์

- ระบบปฏิบัติการ Microsoft Windows XP
- โปรแกรมประยุกต์ MATLAB 7.0
- โปรแกรมประยุกต์ Microsoft Office 2007

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1.6.1 ได้แบบจำลองการวิเคราะห์สายจีโนมโดยใช้โครงข่ายประสาทเทียม

1.6.2 ได้โปรแกรมการวิเคราะห์สายจีโนมโดยใช้โครงข่ายประสาทเทียม

บทที่ 2

ทฤษฎีที่เกี่ยวข้องกับการสร้างแบบจำลองการทำนายจุดเริ่มต้นการแปลรหัส โดยใช้วิธี TF-IDF และโครงข่ายประสาทเทียม

ในบทนี้จะกล่าวถึงทฤษฎีต่าง ๆ ที่ใช้ในการพัฒนาแบบจำลองการทำนายจุดเริ่มต้นการแปลรหัสพันธุกรรมโดยใช้วิธี TF-IDF และโครงข่ายประสาทเทียม ประกอบด้วย 1) อณูชีววิทยา ซึ่งเป็นความรู้พื้นฐานเกี่ยวกับดีเอ็นเอ ยีน และโปรตีน 2) ฟาสต์-เอ เป็นรูปแบบการจัดเรียงสายพันธุกรรมในรูปข้อความ 3) n-แกรม (n-gram) เป็นวิธีการแปลงสายอักขระให้อยู่ในรูปแบบที่สามารถนำไปใช้คำนวณกับโครงข่ายประสาท-เทียมได้ 4) TF-IDF (Term Frequency and Inverse Document Frequency) เป็นวิธีการกำหนดค่าความสามารถในการแบ่งแยกระหว่างข้อมูลกลุ่มบวกและลบ 5) เทคนิคการเลือกลักษณะเฉพาะ เป็นการลดมิติข้อมูลโดยเลือกลักษณะเฉพาะที่มีนัยสำคัญสำหรับการสร้างแบบจำลองโดยวิทยานิพนธ์นี้เลือกใช้ทั้งหมด 4 เทคนิค คือ เทคนิคการเลือกลักษณะเฉพาะสหสัมพันธ์ (Correlation-based Feature Selection: CFS) เทคนิคไคสแควร์ (Chi-Square) เทคนิคอัตราส่วนเกน (Gain Raio) และเทคนิครีลีฟ-เอฟ (ReleifF) 6) โครงข่ายประสาทเทียม (Artificial Neural Networks) ประกอบด้วยสถาปัตยกรรม ประเภท รูปแบบการเรียนรู้ของโครงข่ายประสาทเทียม โดยใช้ขั้นตอนการวิธีการส่งค่าย้อนกลับ 7) การทดสอบแบบไขว้เปลี่ยนแบบ k กลุ่มเป็นการทดสอบประสิทธิภาพของโครงข่ายประสาทเทียม และ 8) การประเมินค่าประสิทธิภาพที่ได้จากการทำนายได้แก่ ค่าความถูกต้อง ค่าการตอบสนองไว และค่าความเฉพาะเจาะจง

2.1 อณูชีววิทยา

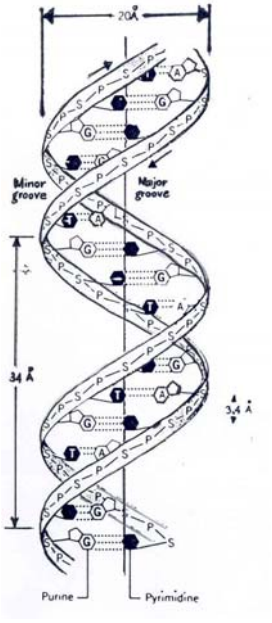
เป็นการศึกษาโครงสร้างของดีเอ็นเอ โปรตีน และการทำงานของยีน (Gene) ซึ่งเป็นรหัสพันธุกรรมบนสายดีเอ็นเอ

2.1.1 ดีเอ็นเอ

เป็นชื่อย่อของสารพันธุกรรมที่มีชื่อวิทยาศาสตร์ว่า “กรดดีออกซีไรโบนิวคลีอิก” (Deoxyribonucleic Acid: DNA) ซึ่งเป็นกรดนิวคลีอิกที่พบในในกลางของเซลล์ทุกชนิด เช่น มนุษย์ สัตว์ พืช เชื้อรา แบคทีเรีย ไวรัส เป็นต้น ผู้ค้นพบดีเอ็นเอ คือ ฟรีดิช มีสเชอร์ ในปี พ.ศ.

2412 (ค.ศ. 1869) แต่ไม่ทราบว่ามีโครงสร้างอย่างไร จนกระทั่งในปี พ.ศ. 2469 (ค.ศ. 1953) เจมส์ ดี. วัตสัน และ ฟรานซิส คริก เป็นผู้ค้นพบโครงสร้างของดีเอ็นเอและนับเป็นจุดเริ่มต้นของยุคเทคโนโลยีทางดีเอ็นเอ ดีเอ็นเอจะบรรจุข้อมูลทางพันธุกรรมของสิ่งมีชีวิตชนิดนั้นไว้ ซึ่งมีลักษณะที่เกิดจากการผสมผสานมาจากลักษณะพันธุกรรมของสิ่งมีชีวิตรุ่นบรรพบุรุษ เช่น ปู่และย่า ตาและยาย หรือพ่อและแม่ และถ่ายทอดลักษณะทางพันธุกรรมซึ่งเกิดการผสมผสานไปยังสิ่งมีชีวิตรุ่นถัดไป คือ ลูกและหลาน

ลักษณะโครงสร้างของดีเอ็นเอมีรูปร่างเป็นสายเกลียวคู่ (Double Helix) แสดงดังภาพประกอบ 2.1 แต่ละข้างเป็นการเรียงตัวของลำดับนิวคลีโอไทด์ (Nucleotide) หลายๆนิวคลีโอไทด์มาต่อกัน แต่ละ นิวคลีโอไทด์ประกอบด้วยหน่วยย่อย 3 หน่วย คือ 1) น้ำตาลดีออกซีไรโบส (Deoxyribose Sugar) 2) กรดฟอสโฟริก และ 3) ไนโตรจีนัสเบส (Nitrogenous Base) สำหรับไนโตรจีนัสเบส แบ่งออกเป็น 2 กลุ่ม คือ a) เบสพิวรีน มี 2 ชนิดได้แก่ อะดินิน (Adenin) อักษรย่อเป็น A และกัวนิน (Guanine) อักษรย่อเป็น G และ b) ไพริมิดิน (Pyrimedine Base) มี 2 ชนิดได้แก่ ไทมิน (Thymine) อักษรย่อเป็น T และไซโตซิน (Cytosine) อักษรย่อเป็น C ในปฏิกิริยาการจับคู่ของเบสนั้น G จับคู่กับ C ขณะที่ A จับคู่กับ T (ประดิษฐ์ พงศ์ทองคำ, 2543)



ภาพประกอบ 2.1 โครงสร้างของดีเอ็นเอ

ยีน คือ ดีเอ็นเอความยาวช่วงหนึ่งๆที่ก่อให้เกิดลักษณะต่างๆของสิ่งมีชีวิตขึ้นมาในสภาพแวดล้อมหนึ่ง ยีนมีสมบัติในการ ควบคุมลักษณะกรรมพันธุ์ต่างๆในร่างกายของ

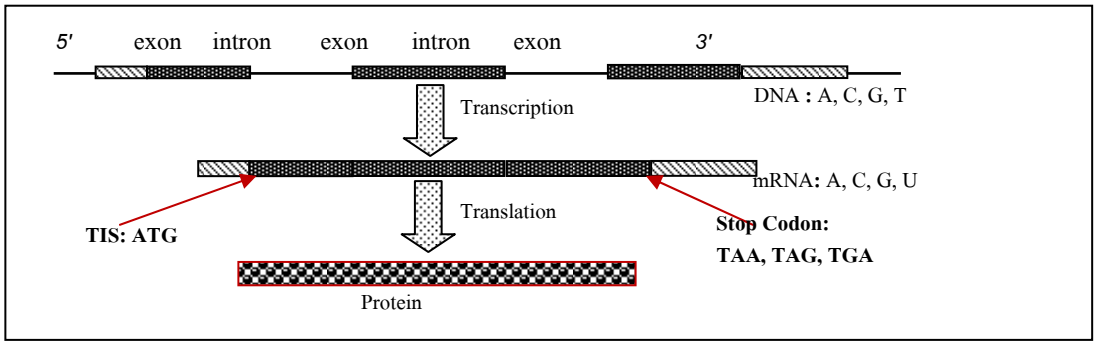
สิ่งมีชีวิต และมีสมบัติถ่ายทอดจากพ่อแม่ไปสู่ลูกได้อย่างไม่มีที่สิ้นสุด ตราบใดที่สิ่งมีชีวิตนั้นยังคงมีสมบัติการสืบพันธุ์ที่สมบูรณ์แบบต่อไปได้อีก ยีนทำหน้าที่กำหนดชนิดของโปรตีนที่เซลล์สังเคราะห์ขึ้นเพื่อนำไปใช้ในกิจกรรมต่าง ๆ ภายในเซลล์ ลำดับนิวคลีโอไทด์ในยีนหนึ่ง ๆ เป็นตัวกำหนดการเรียงตัวของกรดอะมิโนชนิดต่าง ๆ ของโปรตีน รหัสพันธุกรรมของกรดอะมิโน 1 ตัวประกอบด้วยนิวคลีโอไทด์จำนวน 3 นิวคลีโอไทด์ เรียกว่า โคดอน (Codon) สำหรับรหัสพันธุกรรมที่เป็นไปได้ทั้งหมดของกรดอะมิโนเท่ากับ 64 ชนิด โดยการแปลรหัสพันธุกรรมมีคุณสมบัติ ได้แก่ 1) ไม่มีการเหลื่อมหรือซ้อนกันของเบส ไม่เว้นหรือข้ามเบส 2) รหัสพันธุกรรมหลายรหัสสามารถกำหนดกรดอะมิโนตัวเดียวกันได้ 3) รหัสพันธุกรรมบางรหัสไม่ทำหน้าที่กำหนดกรดอะมิโนตัวใดเลย 4) รหัสพันธุกรรมที่กำหนดกรดอะมิโนแต่ละชนิดนั้นเหมือนกันในสิ่งมีชีวิตทุกชนิด และ 5) การแปลรหัสพันธุกรรมบนเอ็มอาร์เอ็นเอมีทิศทางจากปลาย *5′* ไปยังปลาย *3′*

2.1.2 โปรตีน

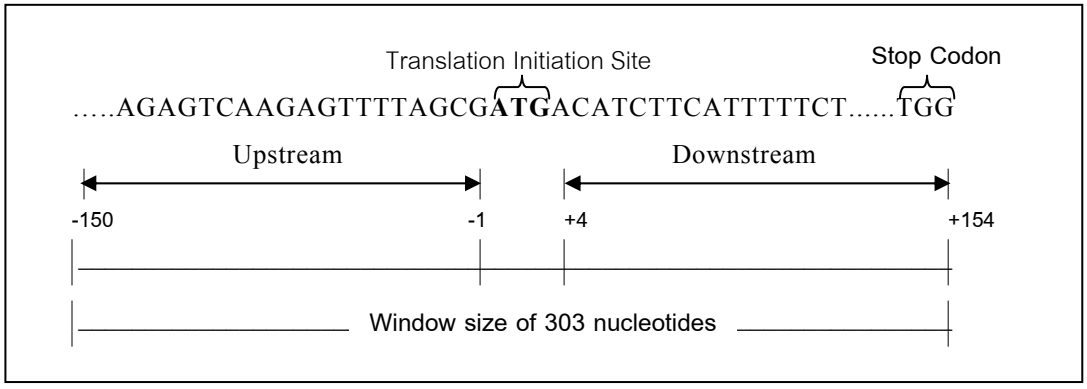
โปรตีน เป็นสารอินทรีย์ซึ่งพบได้ในสิ่งมีชีวิตทุกชนิด มีโครงสร้างซับซ้อนและมีมวลโมเลกุลมากโปรตีนมีหน่วยย่อยคือ กรดอะมิโน เรียงต่อกันด้วยพันธะเปปไทด์ โปรตีนมีหน้าที่สำคัญต่อโครงสร้างและกิจกรรมภายในเซลล์ของสิ่งมีชีวิตทุกชนิด รวมทั้งไวรัสด้วย โปรตีนหลายชนิดทำหน้าที่เป็นเอนไซม์หรือหน่วยย่อยของเอนไซม์ ส่วนโปรตีนอื่นทำหน้าที่ทางด้านโครงสร้าง เช่น โครงสร้างภายในเซลล์ (Cytoskeleton) กลไกทางกายภาพ และบางชนิดยังมีหน้าที่เป็นภูมิคุ้มกันคอยปกป้องร่างกายจากสิ่งแวดล้อม นอกจากนั้นยังมีหน้าที่เป็นขนส่งสารภายในระบบร่างกายและเป็นแหล่งสำรองพลังงานยามขาดแคลนอีกด้วย โปรตีนในอาหารนั้นเป็นแหล่งของกรดอะมิโน ให้แก่สิ่งมีชีวิตแต่ไม่สามารถสังเคราะห์กรดอะมิโนเหล่านั้นได้เอง (อมรา คัมภิรานนท์, 2542) สำหรับเซลล์สิ่งมีชีวิตชั้นสูง (Eukaryote Cell)

กระบวนการสร้างโปรตีนมี 2 กระบวนการคือ กระบวนการถ่ายสำเนา (Transcription) และกระบวนการแปลรหัส (Translation) โดยเริ่มต้นเป็นการนำข้อมูลที่เก็บอยู่ในดีเอ็นเอถ่ายสำเนาไปเป็นอาร์เอ็นสื่อสาร หรือเอ็มอาร์เอ็นเอ (messenger RNA หรือ mRNA) จากนั้นเอ็มอาร์เอ็นเอก็เข้าสู่กระบวนการแปลรหัสไปเป็นโปรตีน (อมรา คัมภิรานนท์, 2542) แสดงดังภาพประกอบ 2.2 กระบวนการแปลรหัสพันธุกรรมบนลำดับเอ็มอาร์เอ็นเอไปเป็นโปรตีนจะเริ่มต้นที่จุดเริ่มต้นการแปลรหัส (Translation Initiation Sites หรือ TIS) (บุรชัย และคณะ, 2542; อุไรวรรณ คัมภิรานนท์, 2545) ซึ่งมีรหัสพันธุกรรมเป็นโคดอน ATG (หรือ AUG บนสายเอ็มอาร์เอ็นเอ) สำหรับกระบวนการแปลรหัสพันธุกรรมตามแบบจำลองการตรวจสอบไรโบโซม (Kozak, 1979; Cigan, 1988; Kozak, 1989) นั้นสมมติให้ไรโบโซมเริ่มต้นตรวจสอบลำดับเอ็มอาร์เอ็นเอจากปลาย *5′* ไปยังปลาย *3′* จนกระทั่งเจอตำแหน่งเริ่มต้นการแปลรหัสพันธุกรรมจึงเริ่มต้นแปลรหัสพันธุกรรม และหยุดการแปลรหัสพันธุกรรมเมื่อเจอโคดอนหยุด

(Stop Codon) ได้แก่โคดอน TAA TAG หรือ TGA สำหรับนิวคลีโอไทด์ด้านซ้ายของจุดเริ่มต้นการแปลรหัสเรียกว่า อัพสตรีม (Upstream) และ นิวคลีโอไทด์ด้านขวาของจุดเริ่มต้นการแปลรหัส เรียกว่า ดาวน์สตรีม (Downstream) แสดงดังภาพประกอบ 2.3



ภาพประกอบ 2.2 โครงสร้างพื้นฐานการสร้างโปรตีนของเซลล์ยูคารีโอต
(ดัดแปลงจาก: Computational Systems Biology Laboratory, 2004)

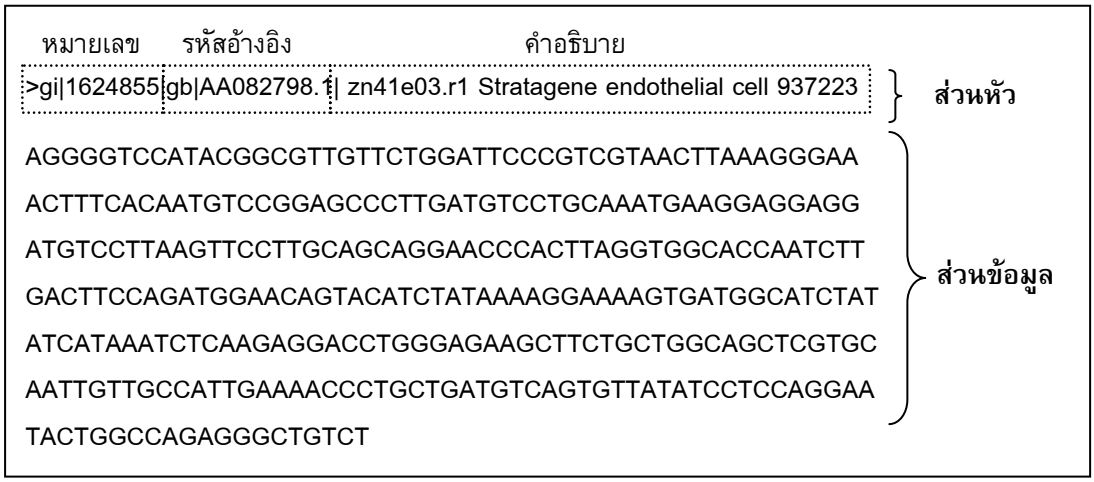


ภาพประกอบ 2.3 การแบ่งสายพันธุกรรมด้วยขนาดหน้าต่าง 303 นิวคลีโอไทด์

จากภาพประกอบ 2.3 แสดงลำดับดีเอ็นเอย่อยของหน้าต่างขนาด 303 นิวคลีโอไทด์ที่มีโคดอน ATG อยู่กึ่งกลาง กำหนดให้นิวคลีโอไทด์ A ของ โคดอน ATG เป็นตำแหน่ง +1 สำหรับส่วนอัพสตรีมจะเริ่มต้นที่ตำแหน่ง -1 และลดลงเรื่อยๆไปด้านซ้าย สำหรับส่วนดาวน์สตรีมจะเริ่มต้นที่ตำแหน่ง +4 และเพิ่มขึ้นเรื่อยๆไปด้านขวา ทั้งนี้สายดีเอ็นเอมีโคดอน ATG หลายตำแหน่ง ซึ่งโคดอน ATG ที่เป็นจุดเริ่มต้นการแปลรหัสส่วนใหญ่จะอยู่ตำแหน่งแรกใกล้ปลาย *5′* อย่างไรก็ตามมีข้อยกเว้นที่โคดอน ATG แรกไม่เป็นจุดเริ่มต้นการแปลรหัส (Kozak, 1996; Kozak, 2002; Dever, 2002) เนื่องจากโคดอน ATG แรกมีบริบทรอบๆ ไม่เหมาะสม หรือ ขนาดของ Open Reading Frame (ORF) มีความยาวน้อย หรือ ไรโบโซมผูกติดครั้งแรกที่โคดอน ATG ที่เป็นจุดเริ่มต้นการแปลรหัสโดยตรง

2.2 ฟาสต์-เอ (FASTA)

ฟาสต์-เอ (FASTA) ย่อมาจากคำว่า FAST-All โดยคำว่า “All” หมายถึง การรวมการจัดเรียงฟาสต์-พี (FAST-P) ซึ่งเป็นการจัดเรียงของโปรตีน และการจัดเรียงแบบฟาสต์-เอ็น (FAST-N) ซึ่งเป็นการจัดเรียงของนิวคลีโอไทด์ โดยรูปแบบฟาสต์-เอจะจัดเก็บข้อมูลแบบไฟล์ข้อความ นามสกุลของฟาสต์-เอ คือ *.fasta โครงสร้างการเก็บข้อมูลแบบฟาสต์-เอ ประกอบด้วย 2 ส่วนหลัก คือ ส่วนหัวของสายพันธุกรรม (Header of Sequences) ขึ้นต้นด้วยสัญลักษณ์ “>” ตามด้วยชื่อสายพันธุกรรม (Sequence Name) และส่วนของข้อมูลหรือลักษณะของข้อมูลสายพันธุกรรม (Sequence Detail) จะขึ้นบรรทัดใหม่ (Claveric and Notredame, 2003) แสดงได้ดังภาพประกอบ 2.4



ภาพประกอบ 2.4 โครงสร้างการจัดเก็บข้อมูลแบบฟาสต์-เอ

จากภาพประกอบ 2.4 แสดงตัวอย่างการเก็บข้อมูลแบบฟาสต์-เอ ของสายพันธุกรรม gi|1624855|gb|AA082798.1| ประกอบด้วยส่วนหัวของสายพันธุกรรมจะระบุหมายเลขสายพันธุกรรม (Sequence ID) คือ “gi|1624855” รหัสอ้างอิง (Accession Number) คือ “gb|AA082798.1” และคำอธิบายสายพันธุกรรม (Description) คือ “zn41e03.r1 Stratagene endothelial cell 937223” และสามารถดาวน์โหลดสายพันธุกรรมได้จากฐานข้อมูล GenBank (National Center for Biotechnology Information, 1988)

2.3 เทคนิค n- แกรม (n-Gram)

กำหนดสายอักขระ $S = s_1, s_2, \ldots, s_N$ และอักขระ $s_i \in \{A_1, A_2, \ldots, A_m\}$ เมื่อ $i = 1,2,\ldots,N$ กำหนด N คือ ความยาวของสายอักขระ S และ m คือ จำนวนของอักขระ รูปแบบ n-แกรมของสายอักขระ S คือ สายอักขระย่อยของอักขระ s_i ที่เรียงต่อกัน ซึ่งมีความยาวเท่ากับ n โดยรูปแบบที่เป็นไปได้ทั้งหมดของ n-แกรม เท่ากับ m^n รูปแบบ ตัวอย่างเช่น กำหนดสายอักขระ S = AACCAGT และอักขระ $s_i \in$ {A, C, G, T} จะได้ N เท่ากับ 7 และ m เท่ากับ 4 ดังนั้น 1-แกรม คือ A C G และ T และ 2-แกรม คือ AA AC CC CA AG และ GT โดย 2-แกรมมีรูปแบบที่เป็นไปได้ทั้งหมดเท่ากับ 16 (4^2 = 16) รูปแบบ

การทำงานของ n-แกรม คือ เลื่อนหน้าต่างของอักขระไปบนสายอักขระ S ทั้งหมด n ตัว และในแต่ละครั้งที่ทำการเลื่อนหน้าต่างนั้นลำดับย่อยขนาด n อักขระจะถูกสกัดออกมา เทคนิค n-แกรมมี 2 แบบได้แก่เทคนิค n-แกรมแบบทุกเฟรม (Any-frame) และเทคนิค n-แกรมแบบอิน-เฟรม (In-frame)

สำหรับเทคนิค n-แกรม แบบทุกเฟรม มีหลักการทำงาน คือ เลื่อนหน้าต่างไปบนสายอักขระ S ทุกอักขระ รูปแบบที่สกัดได้จะมีค่าเท่ากับ $N - n + 1$ รูปแบบ (Saidi, 2007) ตัวอย่างเช่นกำหนดสายอักขระ S = AAGGGCCTAG ซึ่งมีความยาวเท่ากับ 10 (N = 10) อักขระ ถ้าต้องการสร้าง 2-แกรม (n = 2) ดังนั้นจะมีชุดข้อมูลที่สกัดได้เท่ากับ 9 (10-2+1 = 9) รูปแบบ ได้แก่ AA AG GG GG GC CC CT TA และ AG

สำหรับเทคนิค n-แกรม แบบอิน-เฟรม มีหลักการทำงาน คือ เลื่อนหน้าต่างไปบนสายอักขระ S โดยข้ามอักขระที่หน้าต่างเลื่อนผ่าน รูปแบบที่สกัดได้จะมีค่าเท่ากับ N/n รูปแบบ ตัวอย่างเช่นกำหนดให้ลำดับคือ “AAGGGCCTAG” ซึ่งมีความยาวเท่ากับ 10 (N = 10) อักขระ ถ้าต้องการสร้าง 2-แกรม (n = 2) ดังนั้นจะมีชุดข้อมูลที่สกัดได้เท่ากับ 5 (10/2 = 5) รูปแบบ ได้แก่ AA GG GC CT และ AG

2.4 TF-IDF (Term Frequency and Inverse Document Frequency)

การจำแนกหมวดหมู่เอกสาร (Text Categorization) (Joachims, 2003) คือ กิจกรรมในการแยกเอกสารซึ่งประกอบด้วยภาษาธรรมชาติให้อยู่ภายใต้หมวดหมู่ที่กำหนดไว้ก่อน โดยใช้ใจความสำคัญของเอกสาร เนื่องจากคอมพิวเตอร์ไม่สามารถจำแนกหมวดหมู่ของเอกสารซึ่งเป็นภาษาธรรมชาติโดยตรงได้ ดังนั้นจึงต้องแปลงเอกสารให้อยู่ในรูปแบบที่คอมพิวเตอร์สามารถใช้ในการเรียนรู้ได้ ขั้นตอนในการแปลงเอกสาร เรียกว่า การทำดัชนี (Indexing) เพื่อสร้างตัวแทนเนื้อหาของเอกสาร (Document Representation) สำหรับใช้ใน

กระบวนการเรียนรู้ ลักษณะของตัวแทนเอกสารขึ้นอยู่กับสิ่งที่ต้องการพิจารณา หรือต้องการพิจารณาความหมายตามกฎของภาษา สำหรับการจำแนกหมวดหมู่ด้วยวิธีการทางด้านการเรียนรู้ด้วยคอมพิวเตอร์นิยมใช้ลักษณะของตัวแทนเอกสารที่สนใจความหมายของคำ โดยไม่สนใจตำแหน่งของคำ

กำหนดให้เซตเอกสาร (Document) $D = \{d_1, d_2, ..., d_N\}$ มีจำนวนเอกสารทั้งหมดเท่ากับ N และเซตของคำ (Term) $T = \{t_1, t_2, ..., t_M\}$ การแสดงความสัมพันธ์ระหว่างแต่ละเอกสาร d_i และคำ t_j แทนด้วยเวคเตอร์ (Vector) $V_i = \{v_{i,1}, v_{i,2}, ..., v_{i,M}\}$ สำหรับวิธีการหาค่า $v_{i,j}$ ที่นิยมใช้ ได้แก่ ค่าความถี่ (Frequency) และ ค่า TF-IDF

1) ค่าความถี่ เป็นการนับจำนวนครั้งของ t_j ที่ปรากฏในเอกสาร d_i ดังสมการที่ (2.1) หมายความว่า ถ้ารูปแบบ t_j ปรากฏในเอกสาร d_i บ่อยก็จะมีความเกี่ยวข้องกับเรื่องในเอกสาร d_i มาก

$$v_{i,j} = f_{i,j} \tag{2.1}$$

กำหนดให้ $f_{i,j}$ คือ ความถี่ของคำ t_j ที่ปรากฏในเอกสาร d_i

2) ค่า TF-IDF เป็นการพิจารณาความสามารถในการแบ่งแยกของคำ t_j ในเอกสาร d_i จากความถี่ของคำ t_j ในเอกสาร d_i และ อัตราส่วนกลับของจำนวนเอกสารทั้งหมดกับจำนวนเอกสารที่มีคำ t_j ดังสมการที่ (2.2) หมายความว่า หากรูปแบบ t_j มีปรากฏในทุกเอกสาร d_i แล้วรูปแบบ t_j ก็ไม่สามารถแบ่งแยกเอกสารได้

$$v_{i,j} = f_{i,j} \times log_2 \left(\frac{N}{n_j} \right) \tag{2.2}$$

กำหนดให้ N คือ จำนวนเอกสารทั้งหมด

n_j คือ จำนวนของเอกสารซึ่งมีคำ t_j ที่ปรากฏ

2.5 เทคนิคการเลือกลักษณะเฉพาะ

2.5.1 เทคนิคการเลือกลักษณะเฉพาะสหสัมพันธ์

เทคนิคการเลือกลักษณะเฉพาะสหสัมพันธ์ (Correlation-based Feature Selection: CFS) (Hall and Smith, 1997; Hall, 2000) เป็นการเลือกลักษณะเฉพาะที่เกี่ยวข้องกับคลาสโดยใช้การค้นหาฮิวริสติก หลักการทำงาน คือ สร้างกลุ่มลักษณะเฉพาะย่อย k คำนวณหาค่าความสัมพันธ์ระหว่างลักษณะเฉพาะกับลักษณะเฉพาะ และความสัมพันธ์ระหว่างลักษณะเฉพาะกับคลาสด้วยค่าสหสัมพันธ์ของเพียร์สัน ดังสมการที่ (2.3) จากนั้นคำนวณหาค่าฮิวริสติกสำหรับแต่ละกลุ่ม ดังสมการที่ (2.4) ถ้าค่าฮิวริสติกสูง หมายถึง ลักษณะเฉพาะมีความสัมพันธ์กับคลาสสูง และมีความสัมพันธ์ระหว่างลักษณะเฉพาะกับลักษณะเฉพาะน้อย

$$r_{xy} = \frac{\sum xy - n\overline{xy}}{\sqrt{(\sum x^2 - nx^2)(\sum y^2 - ny^2)}} \tag{2.3}$$

กำหนด r_{xy} คือ สัมประสิทธิ์สหสัมพันธ์สำหรับตัวแปร x และ y
x คือ ข้อมูลต่อเนื่องตัวแปรที่ 1
y คือ ข้อมูลต่อเนื่องตัวแปรที่ 2
n คือ จำนวนข้อมูลของตัวแปรที่ 1 และตัวแปรที่ 2

$$Merit_S = \frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}} \tag{2.4}$$

กำหนด $Merit_S$ คือ ค่าฮิวริสติกของกลุ่มลักษณะเฉพาะ S
$\overline{r_{cf}}$ คือ ค่าเฉลี่ยสหสัมพันธ์ระหว่างลักษณะเฉพาะกับคลาส
$\overline{r_{ff}}$ คือ ค่าเฉลี่ยสหสัมพันธ์ระหว่างลักษณะเฉพาะกับลักษณะเฉพาะ

2.5.2 เทคนิคไคสแควร์ (Chi-Square)

เป็นเทคนิคที่ใช้ค่าไคส์แควร์ซึ่งเป็นค่าทางสถิติหาความสัมพันธ์ระหว่างลักษณะเฉพาะกับคลาส เพื่อจัดลำดับลักษณะเฉพาะตามค่านัยสำคัญทางสถิติ (Li and Leong, 2005) โดยค่าไคสแควร์ของแต่ละลักษณะเฉพาะหาได้ดังสมการที่ (2.5)

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(A_{ij} - E_{ij})^2}{E_{ij}} \quad (2.5)$$

กำหนด A_{ij} คือ ความถี่จริงของตัวอย่างที่มีค่าที่ i และคลาสที่ j

E_{ij} คือ ความถี่คาดหวังของตัวอย่างที่มีค่าที่ i และคลาสที่ j หรือความถี่คาดหวังของ A_{ij} คำนวณดังสมการที่ (2.6)

m คือ จำนวนค่าของลักษณะเฉพาะ

n คือ จำนวนของคลาส

$$E_{ij} = R_i \times \frac{C_j}{N} \quad (2.6)$$

กำหนด R_i คือ จำนวนตัวอย่างทั้งหมดที่มีค่าลักษณะเฉพาะที่ i

C_j คือ จำนวนตัวอย่างทั้งหมดที่อยู่ในคลาสที่ j

N คือ จำนวนของตัวอย่างทั้งหมด

ค่าไคส์แควร์สำหรับแต่ละลักษณะเฉพาะหาค่าได้จากความแตกต่างระหว่างค่าความถี่คาดหวังและค่าความถี่จริง ลักษณะเฉพาะที่มีค่าไคส์แควร์มากจะมีนัยสำคัญสูง โดยลักษณะเฉพาะจะถูกจัดเรียงตามค่าไคส์แควร์จากค่ามากไปน้อย

2.5.3 เทคนิคอัตราส่วนเกน (Gain Ratio)

เป็นเทคนิคการเลือกลักษณะเฉพาะที่ใช้การประเมินค่าลักษณะเฉพาะด้วยค่าอัตราส่วนเกน ซึ่งวัดความสัมพันธ์ของลักษณะเฉพาะที่จะปรับสเกลตามค่าลักษณะเฉพาะที่สนใจกับคลาส การหาค่าอัตราส่วนเกนคำนวณได้ดังสมการที่ (2.7)

$$Gain\ Ratio = \frac{H(Y) - H(Y|X)}{H(X)} \quad (2.7)$$

กำหนด X คือ ลักษณะเฉพาะ

Y คือ คลาส

$H(X)$ คือ ค่าเอนโทรปีของ X

$H(Y)$ คือ ค่าเอนโทรปีของ Y

$H(Y|X)$ คือ ค่าเอนโทรปีของ Y ภายใต้เงื่อนไข X

การหาค่า *H(Y)* แสดงได้ดังสมการที่ (2.8) และการหาค่า *H(Y|X)* แสดงได้ดังสมการที่ (2.9)

$$H(Y) = -\sum_{y \in Y} p(y) \log_2 (p(y)) \tag{2.8}$$

$$H(Y \mid X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y \mid x) \log_2 (p(y \mid x)) \tag{2.9}$$

กำหนด *p(y)* คือ ความน่าจะเป็นของ *y*
p(x) คือ ความน่าจะเป็นของ *x*
p(y|x) คือ ความน่าจะเป็นของ *y* เมื่อรู้ *x*

ค่าอัตราส่วนเกนได้จากการหารด้วยค่าเอนโทรปีของลักษณะเฉพาะทำให้ค่าที่ได้อยู่ระหว่าง [0,1] ถ้าค่าอัตราส่วนเกนเท่ากับ 0 หมายถึง ลักษณะเฉพาะ *X* ไม่มีความสัมพันธ์กับคลาส *Y* ถ้าค่าอัตราส่วนเกนมีค่าเท่ากับ 1 แสดงว่าลักษณะเฉพาะ *X* มีความสัมพันธ์กับคลาส *Y* มากที่สุด ข้อสังเกตจากสมการที่ (2.7) ค่าอัตราส่วนเกนมีค่าน้อยเมื่อเทียบกับค่าเกนสารสนเทศ

2.5.4 เทคนิครีลีฟ-เอฟ (ReliefF)

ขั้นตอนวิธีรีลีฟ (Relief Algorithm) (Marko and Igor, 2003) เป็นเทคนิคการกรองเพื่อเลือกลักษณะเฉพาะที่เกี่ยวข้องโดยอาศัยวิธีการเชิงสถิติ สมมติให้ $I_1, I_2, \ldots, I_n$ เป็นกลุ่มตัวอย่างของลักษณะเฉพาะ A_i เมื่อ i = 1, 2,..., a โดยที่ a เป็นจำนวนลักษณะเฉพาะทั้งหมด ซึ่งขั้นตอนวิธีรีลีฟพื้นฐานดังภาพประกอบ 2.5

จากภาพประกอบ 2.5 เริ่มต้นสุ่มตัวอย่าง R_i (บรรทัด 3) จากนั้นค้นหาตัวอย่างใกล้สุดซึ่งมีคลาสเดียวกับ R_i แทนด้วย *H* และตัวอย่างใกล้สุดซึ่งต่างคลาสกับ R_i แทนด้วย *M* (บรรทัด 4) ตามลำดับ การปรับค่าน้ำหนัก *W[A]* ขึ้นอยู่กับค่าของตัวอย่าง R_i *M* และ *H* โดยหาความแตกต่างระหว่างค่าน้ำหนักเดิมกับความแตกต่างของค่าลักษณะเฉพาะจากฟังก์ชัน $diff(A, I_1, I_2)$ ซึ่งใช้คำนวณความแตกต่างของค่าลักษณะเฉพาะ *A* ระหว่างตัวอย่าง I_1 และตัวอย่าง I_2 (บรรทัด 5 และ 6) สำหรับค่าลักษณะเฉพาะแบบ nominal คำนวณได้ดังสมการที่ (2.9)

Algorithm Relief

Input: for each training instance a vector of attribute values and the value

Output: the vector W of estimations of the qualities of attributes

1 set all weights $W[A]$:= 0.0;

2 **for** i := 1 **to** m **do begin**

3 randomly select an instance R_i;

4 find nearest hit H and nearest miss M;

5 **for** A := 1 **to** a **do**

6 $W[A] := W[A] - diff(A,R_i,H)/m + diff(A,R_i,M)/m$

7 **end**;

ภาพประกอบ 2.5 ขั้นตอนวิธีรีลีฟพื้นฐาน

$$diff(A,I_1,I_2) = \begin{cases} 0 \ ; \ value(A,I_1)=value(A,I_2) \\ 1 \ ; \ otherwise \end{cases} \quad (2.9)$$

สำหรับค่าลักษณะเฉพาะเชิงตัวเลข ฟังก์ชัน $diff(A,I_1,I_2)$ คำนวณได้ดังสมการที่ (2.10)

$$diff(A,I_1,I_2) = \frac{|value(A,I_1) - value(A,I_2)|}{max(A) - min(A)} \quad (2.10)$$

กระบวนการปรับค่าน้ำหนักนี้จะถูกทำซ้ำ m ครั้งตามที่ผู้ใช้กำหนด สำหรับขั้นตอนวิธีรีลีฟ-เอฟ (ReleifF Algorithm) มีหลักการทำงานคล้ายกับขั้นตอนวิธีรีลีฟ ข้อดีของขั้นตอนวิธีรีลีฟ-เอฟ คือ สามารถใช้กับข้อมูลที่มีมากกว่า 2 คลาส และทนทานต่อข้อมูลรบกวน หรือ ข้อมูลที่ไม่สมบูรณ์มากขึ้น โดยขั้นตอนวิธีรีลีฟ-เอฟแสดงดังภาพประกอบ 2.6

จากภาพประกอบ 2.6 เริ่มต้นสุ่มตัวอย่าง R_i จากนั้นค้นหาตัวอย่างใกล้สุดที่มีคลาสเดียวกับ R_i จำนวน k ตัวอย่าง (บรรทัด 3 และ 4) แทนด้วย H_j และตัวอย่างใกล้สุดที่มีคลาสแตกต่างจาก R_i จำนวน k ตัวอย่าง แทนด้วย $M_j(C)$ (บรรทัด 5 และ 6) การปรับค่าน้ำหนัก $W[A]$ ขึ้นอยู่กับค่าลักษณะเฉพาะ A ของตัวอย่าง R_i H_j และ $M_j(C)$ โดยแต่ละคลาสของ $M_j(C)$ ต้องหาความน่าจะเป็นของคลาส $P(C)$ จากตัวอย่างสอนที่มีอยู่ (บรรทัด 7 และ 8)

Algorithm ReliefF

Input: for each training instance a vector of attribute values and the class value

Output: the vector W of estimations the qualities of attributes

1 set all weights $W[A]$:= 0.0;

2 **for** i := 1 **to** m **do begin**

3 randomly select an instance R_i;

4 find k nearest hits H_j;

5 **for** each class $C \neq class(R_i)$ **do**

6 from class C find k nearest misses $M_j(C)$;

7 **for** A := 1 **to** a **do**

8 $$W[A] := W[A] - \sum_{j=1}^{k} diff(A, R_i, H_j)/(m \times k) + \sum_{C \neq class(R_i)} \left[\frac{P(C)}{1 - (P(class(R_i)))} \sum_{j=1}^{k} diff(A, R_i, M_j(C)) \right]/(m \times k);$$

9 **end;**

ภาพประกอบ 2.6 ขั้นตอนวิธีรีลีฟ-เอฟ

2.6 โครงข่ายประสาทเทียม

โครงข่ายประสาทเทียม (Artificial Neural Networks) (Roiger and Geatz, 2003; Kantardzic, 2003) เป็นสาขาหนึ่งของปัญญาประดิษฐ์ ซึ่งจำลองการทำงานของสมองมนุษย์ซึ่งประกอบด้วยเซลล์ประสาท (Neuron) จำนวนมาก แต่ละเซลล์ประสาทประกอบด้วยนิวเคลียส (Nucleus) ตัวเซลล์ (Cell Body) ใยประสาทนำเข้า (Dendrite) และแกนประสาทนำออก (Axon) โดยใยประสาทนำเข้าทำหน้าที่รับสัญญาณจากเซลล์ประสาทอื่น และแกนประสาทนำออกทำหน้าที่นำสัญญาณจากตัวเซลล์ส่งต่อไปให้เซลล์ประสาทอื่นต่อไป คุณลักษณะสมองมนุษย์ ตัวอย่างเช่น การเรียนรู้จากประสบการณ์ แก้ไขปัญหาใหม่โดยอ้างอิงจากปัญหาที่เคยเจอ และให้คำตอบได้แม้ข้อมูลผิดพลาดและไม่สมบูรณ์ และประมวลผลได้รวดเร็ว ข้อดีของโครงข่ายประสาทเทียม คือ มีความแม่นยำสูง ทนทานต่อความผิดพลาด และสามารถรองรับข้อมูลที่ไม่สมบูรณ์ หรือมีสิ่งรบกวนได้

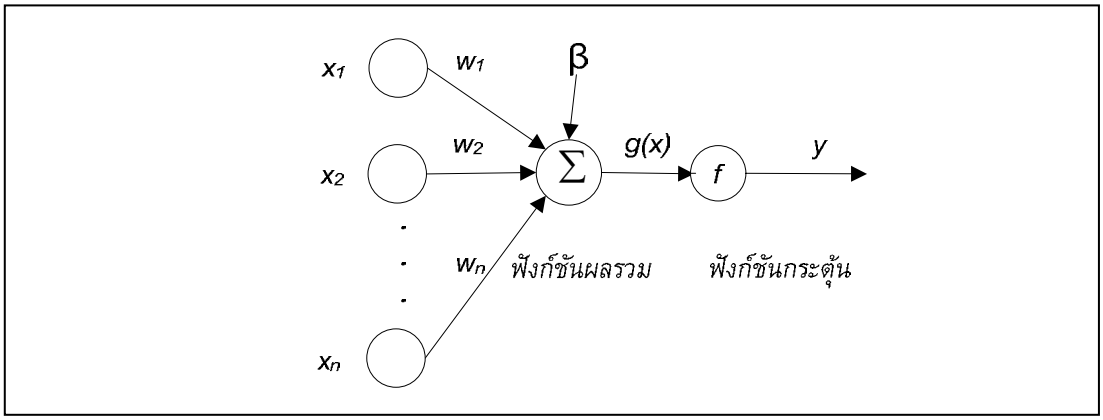
กระบวนการสร้างแบบจำลองโครงข่ายประสาทเทียมมี 5 ขั้นตอน คือ 1) การแทนค่าข้อมูลเข้า 2) การกำหนดสถาปัตยกรรม 3) การกำหนดวิธีการเรียนรู้ 4) การสอนของโครงข่าย และ 5) การทดสอบเพื่อประเมินผลการสอนโครงข่าย การแทนค่าข้อมูลเข้าของโครงข่ายประสาทเทียมส่งผลกระทบสำคัญต่อการสอนของโครงข่ายและสมรรถภาพผลลัพธ์ของ

แบบจำลอง นอกจากนี้ยังส่งผลต่อเวลาการประมวลผลและการเรียนรู้ด้วย ตัวแปรข้อมูลเข้ามี 2 ชนิด ได้แก่ ตัวแปรตาม หรือตัวแปรเชิงปริมาณ และตัวแปรอิสระ หรือคลาส โดยตัวแปรเชิงปริมาณอาจประยุกต์ใช้ฟังก์ชันซิกมอยด์ หรือการนอมอลไลซ์ให้เป็นข้อมูลต่อเนื่องในช่วงค่า 0 ถึง 1 หรือ -1 ถึง 1 เป็นต้น ตัวแปรคลาสใช้การแทนค่าแบบไบนารี่ ตัวอย่างเช่นผลลัพธ์แบบ 1 ไบนารี ได้แก่ รูปภาพดำและขาว คำตอบใช่หรือไม่ใช่ หรือสวิตซ์เปิดหรือปิด เป็นต้น สำหรับตัวแปรคลาสแบบหลายตัวแปรจำนวนโหนดผลลัพธ์ไม่ควรจะเท่ากับจำนวนตัวแปรคลาส การกำหนดค่า 1 สำหรับคลาสที่สนใจ ส่วนคลาสอื่น ๆ กำหนดให้เป็น 0 จะทำให้เวลาการประมวลผลลดลง

2.6.1 สถาปัตยกรรมโครงข่ายประสาทเทียม

สถาปัตยกรรมของโครงข่ายประสาทเทียมมี 2 ชนิด ได้แก่ สถาปัตยกรรมหน่วยประมวลผลย่อยเดียว (Perceptron) และสถาปัตยกรรมหน่วยประมวลผลย่อยหลายชั้น (Multilayer Perceptron)

1) สถาปัตยกรรมหน่วยประมวลผลย่อยเดียวเป็นโครงข่ายประสาทเทียมแบบง่ายมีหน่วยประมวลผลย่อยเดียวประกอบด้วย ชั้นข้อมูลเข้า ฟังก์ชันผลรวม (Summation Function) ทำหน้าที่หาผลรวมของผลคูณระหว่างค่าข้อมูลเข้า $x_1,x_2,\ldots,x_n$ กับค่าน้ำหนักข้อมูลเข้า $w_1,w_2,\ldots,w_n$ และฟังก์ชันกระตุ้น (Activation Function) ทำหน้าที่แปลงผลลัพธ์จากฟังก์ชันผลรวมให้อยู่ในช่วงค่าที่ต้องการ ตัวอย่างฟังก์ชันกระตุ้นเช่น ฟังก์ชันซิกมอยด์ มีค่าอยู่ในช่วง 0 ถึง 1 และค่าของฟังก์ชันซิกมอยด์จะเพิ่มขณะที่ค่าฟังก์ชันผลรวมเพิ่มขึ้นแสดงภาพองค์ประกอบของหน่วยประมวลผลย่อยได้ดังภาพประกอบ 2.7



ภาพประกอบ 2.7 องค์ประกอบของหน่วยประมวลผลย่อย

การคำนวณของฟังก์ชันผลรวมสามารถแสดงได้ดังสมการที่ (2.12) ดังนี้

$$g(x) = \sum_{i=1}^{n} x_i w_i + \beta \qquad (2.12)$$

กำหนดให้ $g(x)$ คือ ผลลัพธ์ของฟังก์ชันผลรวม

x_i คือ ค่าข้อมูลเข้าตัวที่ i

n คือ จำนวนข้อมูลเข้าทั้งหมด

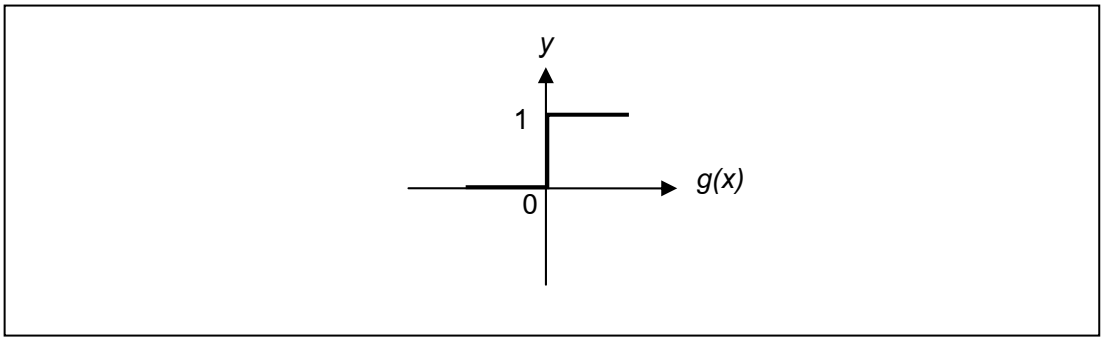
w_i คือ ค่าน้ำหนักของข้อมูลเข้าตัวที่ i

β คือ ค่าความโน้มเอียง

ฟังก์ชันกระตุ้น (f) ทำหน้าที่แปลผลลัพธ์ของฟังก์ชันผลรวม g(x) ให้อยู่ในช่วงค่าที่ต้องการ y ตัวอย่างฟังก์ชันกระตุ้นสามารถแสดงได้ ดังสมการที่ (2.13) ถึง (2.15)

1) ฟังก์ชันสเตป (Step Function) ผลลัพธ์ที่ได้จะเป็นค่า 0 และ 1 แสดงดังสมการที่ (2.13) และภาพประกอบ 2.8

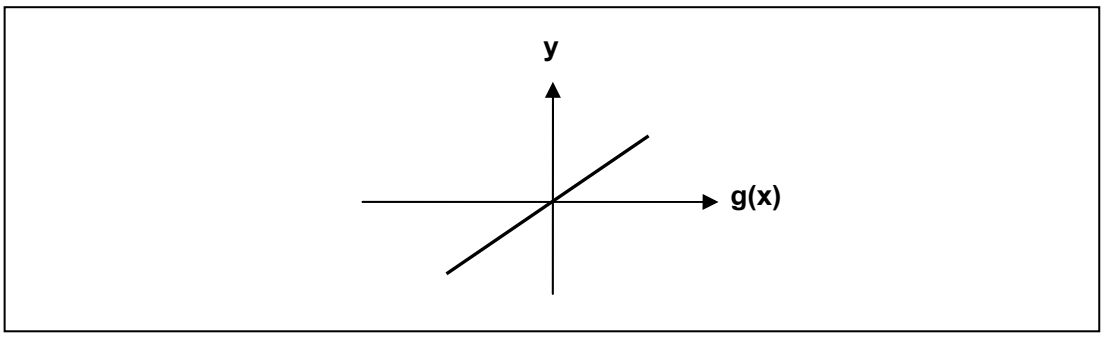
$$y = \begin{cases} 1 & ;\ if\ g(x) > 0 \\ 0 & ;\ otherwise \end{cases} \qquad (2.13)$$



ภาพประกอบ 2.8 ฟังก์ชันสเตป

2) ฟังก์ชันเชิงเส้น (Linear Function) ผลลัพธ์ที่ได้จะมีค่าเท่ากับข้อมูลเข้า แสดงดังสมการที่ (2.14) และภาพประกอบ 2.9

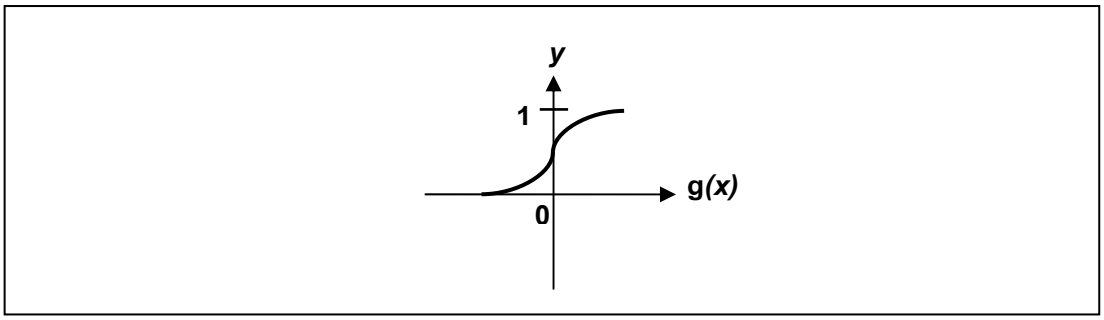
$$y = g(x) \qquad (2.14)$$



ภาพประกอบ 2.9 ฟังก์ชันเชิงเส้น

3) ฟังก์ชันล็อกซิกมอยด์ (Log-Sigmoid Function) ผลลัพธ์ที่ได้จะอยู่ในช่วง 0 ถึง 1 แสดงดังสมการที่ (2.15) และภาพประกอบ 2.10

$$y = \frac{1}{1 + e^{-g(x)}} \qquad (2.15)$$

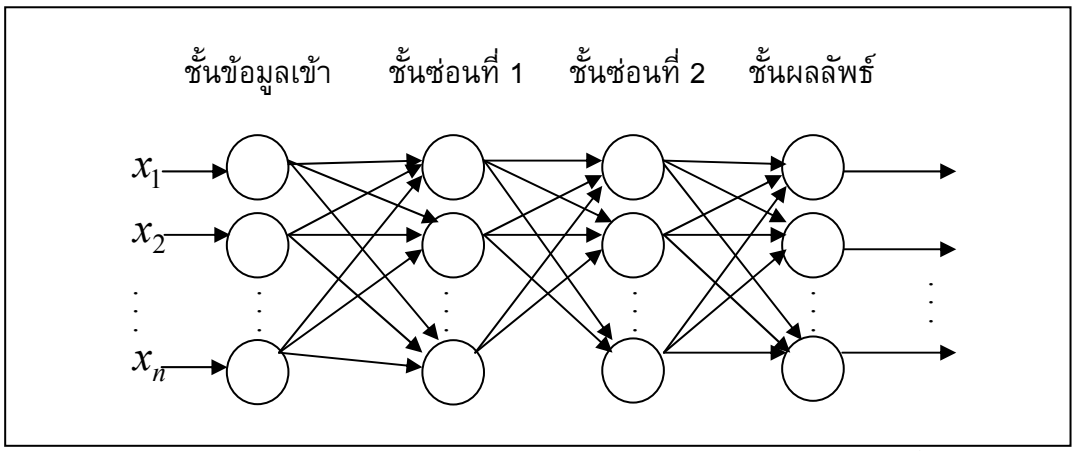


ภาพประกอบ 2.10 ฟังก์ชันล็อกซิกมอยด์

2) หน่วยประมวลผลย่อยหลายชั้นแสดงตัวอย่างดังภาพประกอบ 2.11 เป็นโครงข่ายประสาทเทียมที่มีหน่วยประมวลผลย่อยหลาย ๆ หน่วยมาเชื่อมต่อกันเป็นโครงข่ายเพื่อเพิ่มประสิทธิภาพในการทำนาย รูปแบบการเรียนรู้ของโครงข่ายประสาทเทียมแบบหน่วยประมวลผลย่อยหลายชั้นเป็นการเรียนรู้แบบมีผู้สอน สถาปัตยกรรมโครงข่ายประสาทเทียมแบบหน่วยประมวลผลย่อยหลายชั้นมี 3 ระดับ คือ ชั้นข้อมูลเข้า (Input Layer) ชั้นซ่อน (Hidden Layer) และชั้นผลลัพธ์ (Output Layer) โดยชั้นข้อมูลเข้ามี 1 ชั้น ชั้นซ่อนมีกี่ชั้นก็ได้ และชั้นผลลัพธ์มี 1 ชั้น ในแต่ละชั้นจะมีหน่วยประมวลผลย่อยกี่หน่วยก็ได้

ชั้นซ่อนหมายถึง ชั้นของของหน่วยประมวลผลย่อยที่อยู่ระหว่างชั้นข้อมูลเข้าและชั้นผลลัพธ์ซึ่งช่วยเพิ่มประสิทธิภาพการคำนวณ ในเชิงหลักการชั้นซ่อนสามารถมีได้มากกว่า 1 ชั้น โดยโครงข่ายจะสามารถทำงานที่ซับซ้อนได้ดีขึ้นถ้ามีโหนด

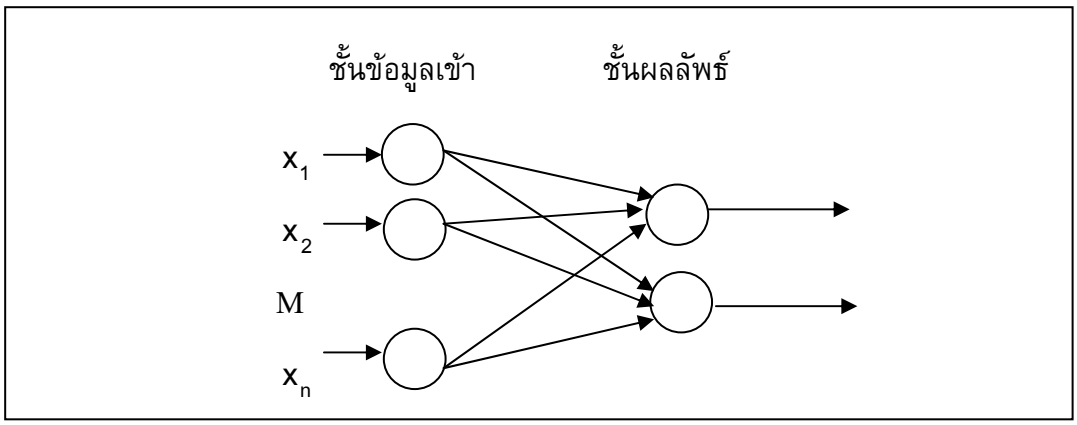
ซ่อนเพียงพอ ผลลัพธ์ของโหนดซ่อนจะเป็นข้อมูลเข้าให้แก่โหนดที่อยู่ในชั้นถัดไปหรือโหนดของชั้นผลลัพธ์ Berke และ Hajela (Berke and Hajela, 1991) แนะนำว่าจำนวนโหนดของชั้นซ่อนควรจะอยู่ระหว่างค่าเฉลี่ยระหว่างโหนดข้อมูลเข้าและโหนดผลลัพธ์ Soemardi (Soemardi, 1996) แนะนำว่าจำนวนโหนดซ่อนควรจะมีประมาณ 75% ของโหนดข้อมูลเข้า ดังนั้นจำนวนโหนดซ่อนควรมีมากสุดเท่ากับผลรวมของโหนดข้อมูลเข้าและโหนดผลลัพธ์ แต่ไม่ควรน้อยกว่า 75% ของโหนดข้อมูลเข้า หรือค่าเฉลี่ยของโหนดข้อมูลเข้าและผลลัพธ์



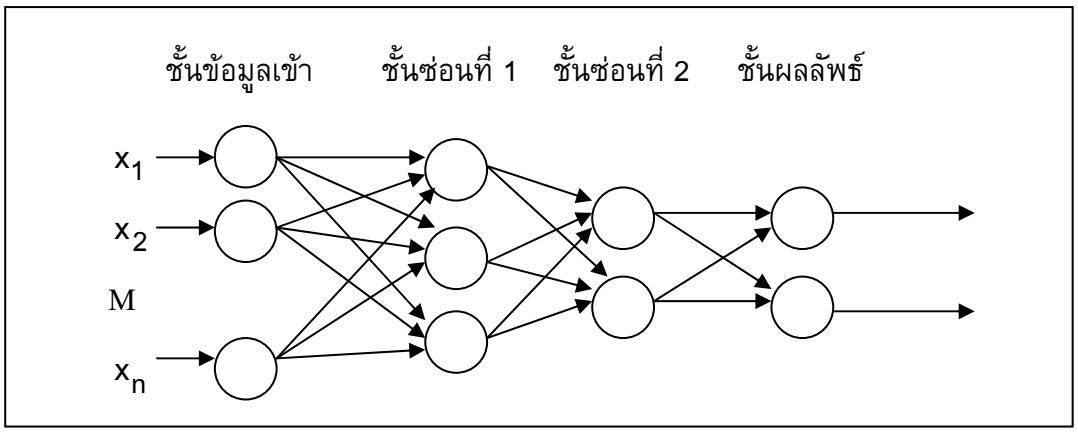
ภาพประกอบ 2.11 สถาปัตยกรรมแบบหน่วยประมวลผลย่อยหลายชั้น

2.6.2 ประเภทของโครงข่ายประสาทเทียม

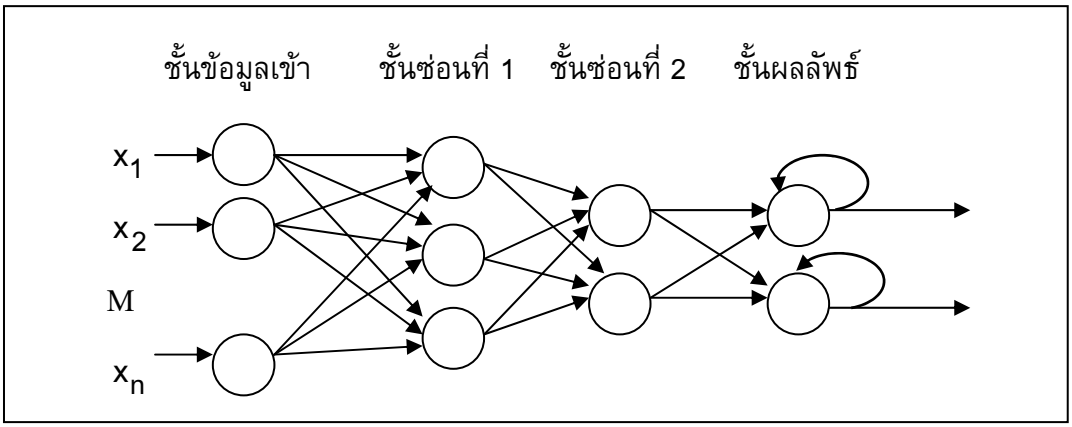
ประเภทของโครงข่ายประสาทเทียมมี 3 ประเภท ได้แก่ 1)โครงข่ายแบบไปข้างหน้าชั้นเดียว (Single-Layer Feedforward Networks) เป็นโครงข่ายที่มีทิศทางไปข้างหน้าและมีเฉพาะชั้นข้อมูลเข้าและชั้นผลลัพธ์ 2) โครงข่ายแบบไปข้างหน้าหลายชั้น (Multilayer Feedforward Networks) แสดงดังภาพประกอบ 2.12 เป็นโครงข่ายที่มีทิศทางไปข้างหน้า มีชั้นข้อมูลเข้า ชั้นซ่อน และชั้นแสดงผล แสดงดังภาพประกอบ 2.13 และ 3) โครงข่ายแบบย้อนกลับ (Recurrent Networks) เป็นโครงข่ายที่สามารถจะมีลูปในการวนกลับได้อย่างน้อย 1 วง นั่นคือผลลัพธ์ของหน่วยประมวลผลย่อยหนึ่งสามารถกลับไปเป็นข้อมูลเข้าของหน่วยประมวลผลย่อยในระดับที่ผ่านมาได้ ดังแสดงในภาพประกอบ 2.14



ภาพประกอบ 2.12 โครงข่ายแบบไปข้างหน้าชั้นเดียว



ภาพประกอบ 2.13 โครงข่ายแบบไปข้างหน้าหลายชั้น



ภาพประกอบ 2.14 โครงข่ายแบบย้อนกลับ

2.6.3 รูปแบบการเรียนรู้ของของโครงข่ายประสาทเทียม

รูปแบบการเรียนรู้ของโครงข่ายประสาทเทียมมี 2 รูปแบบ คือการเรียนรู้แบบมีการสอน (Supervised Learning) และการเรียนรู้แบบไม่มีการสอน (Unsupervised Learning)

1) การเรียนรู้แบบมีการสอน (Supervised Learning) เป็นการเรียนรู้ที่ต้องมีชุดข้อมูลสอน (Training Set) ซึ่งประกอบด้วยชุดข้อมูลเข้าและชุดข้อมูลผลลัพธ์ โดยในระหว่างการสอนนั้นโครงข่ายจะให้ผลลัพธ์จริงที่ได้จากการคำนวณ (Actual Output) และเปรียบเทียบระหว่างผลลัพธ์จริงที่จากการคำนวณกับชุดข้อมูลผลลัพธ์เป้าหมาย (Target Output) โดยผลต่างจากการเปรียบเทียบคือ ค่าความผิดพลาด หรือค่าความเคลื่อน ตัวอย่างการเรียนแบบมีการสอน คือ ขั้นตอนวิธีการส่งค่าย้อนกลับ มักนิยมใช้กับโครงข่ายประสาทเทียมแบบไปข้างหน้าหลายชั้น (Multilayer Feedforward Neural Networks)

2) การเรียนรู้แบบไม่มีการสอน (Unsupervised Learning) เป็นการเรียนโดยไม่ต้องอาศัยชุดข้อมูลผลลัพธ์เป้าหมาย (Target Output) ใช้เฉพาะชุดข้อมูลเข้าให้กับโครงข่ายเท่านั้น โดยโครงข่ายจะจัดเรียงโครงสร้างด้วยตัวเองตามลักษณะของข้อมูลผลลัพธ์ที่ได้ (เปรียบเทียบกับคน เช่น การที่เราสามารถแยกแยะพันธุ์พืช พันธุ์สัตว์ตามลักษณะรูปร่างของมันได้เองโดยไม่มีใครสอน)

2.6.4 ขั้นตอนวิธีการส่งค่าย้อนกลับ (Backpropagation Algorithm)

วิธีการเรียนรู้ด้วยขั้นตอนการส่งค่าย้อนกลับ (Backpropagation Algorithm) โดยส่วนใหญ่มักใช้เพื่อสอนโครงข่ายแบบไปข้างหน้าหลายชั้น (Multilayer Feedforward Neural Networks) ขั้นตอนวิธีการส่งค่าย้อนกลับดังภาพประกอบ 2.15 ข้อมูลสอนแต่ละตัวอย่างจะพิจารณาเป็นคู่ $<\vec{x},\vec{t}>$ เมื่อ $\vec{x}$ คือ เวคเตอร์ข้อมูลเข้า $\vec{t}$ คือ เวคเตอร์ข้อมูลผลลัพธ์ และ η คือ อัตราการเรียนรู้ กำหนดให้ n_{in} n_{out} และ n_{hidden} เป็นจำนวนโหนดข้อมูลเข้า โหนดผลลัพธ์ และโหนดซ่อนของโครงข่ายประสาทเทียม ตามลำดับ x_{ji} คือข้อมูลเข้าจากโหนด i ไปยังโหนด j และ w_{ji} คือน้ำหนักจากโหนด i ไปยังโหนด j กระบวนการส่งค่าย้อนกลับ ประกอบด้วย 2 ส่วนย่อยคือ การส่งผ่านไปข้างหน้า (Forward Pass) และการส่งผ่านย้อนกลับ (Backward Pass) สำหรับการส่งผ่านไปข้างหน้า ข้อมูลจะผ่านเข้าโครงข่ายประสาทเทียมที่ชั้นข้อมูลเข้า และจะส่งผ่าน จากอีกชั้นหนึ่งไปสู่อีกชั้นหนึ่งจนกระทั่งถึงชั้นข้อมูลออก (บรรทัด 5) ส่วนการส่งผ่านย้อนกลับค่าน้ำหนักการเชื่อมต่อจะถูกปรับเปลี่ยนให้สอดคล้องกับกฎการแก้ข้อผิดพลาด (Error-Correction) คือผลต่างของผลลัพธ์แท้จริง (Actual Response) กับผลลัพธ์เป้าหมาย (Target Response) เกิดเป็นสัญญาณผิดพลาด (Error Signal) สำหรับการคำนวณหาสัญญาณผิดพลาดโหนดผลลัพธ์แสดงดังบรรทัดที่ 7 และ 8 และการคำนวณหาสัญญาณผิดพลาดโหนดซ่อนแสดงดังบรรทัดที่ 9 และ 10 ซึ่งสัญญาณนี้จะถูกส่งย้อนกลับเข้าสู่

โครงข่ายประสาทเทียมในทิศทางตรงกันข้ามกับการเชื่อมต่อ จากนั้นค่าน้ำหนักของการเชื่อมต่อจะถูกปรับจนกระทั่งผลลัพธ์แท้จริงเข้าใกล้ผลลัพธ์เป้าหมายแสดงดังบรรทัดที่ 11 ถึง 13

Algorithm: Backpropagation

Input: training-examples, h, n_{in}, n_{out} and n_{hidden}

Output: neural networks model

Detail: Each training example is a pair $<\vec{x}, \vec{t}>$, where $\vec{x}$ is the input vector, $\vec{t}$ is the target output vector, η is the learning rate. n_{in}, n_{out}, and n_{hidden} are the number of network inputs, units in the hidden layer, and output units, respectively. The input from unit i into unit j and the weight from unit i to unit j are denoted x_{ji} and w_{ji}

1 Initialize all network weights to small random number
2 Until the termination condition is met do
3 For each $<\vec{x}, \vec{t}>$ in training-example do
4 /*Propagate input forward through the network*/
5 Input the instance $\vec{x}$ to the network, compute the output o_u of every unit u
6 /*Propagate errors backward through the network*/
7 For each network output unit *k*, calculate its error term δ_k
8 $\delta_k = o_k(1-o_k)(t_k-o_k)$
9 For each hidden unit *h*, calculate its error term δ_h
10 $\delta_h = o_h(1-o_h)\sum_{k \varepsilon outpus} w_{kh}\delta_k$
11 Update each network weight w_{ji}
12 $\Delta w_{ji} = \eta \delta_j x_{ji}$
13 $w_{ji} = w_{ji} + \Delta w_{ji}$

ภาพประกอบ 2.15 ขั้นตอนวิธีการส่งค่าย้อนกลับ

ขั้นตอนวิธีการส่งค่าย้อนกลับเป็นขั้นตอนวิธีที่ใช้เวลาการเรียนนานเมื่อขนาดโครงข่ายประสาทเทียมใหญ่ หรือจำนวนของชุดข้อมูลสอนปริมาณมาก ข้อจำกัดของขั้นตอนวิธีการส่งค่าย้อนกลับ คือ ไม่สามารถรับประกันได้ว่าโครงข่ายประสาทเทียมสอนภายในเวลาจำกัด อัตราการเรียนรู้จะส่งผลต่อการปรับค่าน้ำหนัก ถ้าอัตราการเรียนรู้สูงจะทำให้การเรียนรู้เร็ว แต่ก็อาจเรียนรู้ไม่สำเร็จเนื่องจากการปรับค่ามีความหยาบเกินไป อัตราการเรียนรู้ที่มีค่าน้อยก็จะทำให้การปรับน้ำหนักทำได้อย่างละเอียดแต่ก็อาจเสียเวลาในการเรียนรู้นาน สำหรับ

เกณฑ์การหยุดการเรียนรู้มี 2 เกณฑ์ คือ 1) หยุดการเรียนรู้เมื่อค่าความผิดพลาดอยู่ในระดับที่ยอมรับได้ หรือน้อยกว่าค่า Error Acceptance 2) หยุดการเรียนรู้เมื่อครบจำนวนรอบการสอน (Epoch)

2.7 การทดสอบแบบไขว้เปลี่ยน *k* กลุ่ม

การทดสอบแบบไขว้เปลี่ยน *k* กลุ่ม (k-Fold Cross Validation) (Written and Frank, 2005) เป็นการประเมินประสิทธิภาพของวิธีเรียนรู้ของเครื่อง หลักการทำงาน คือ แบ่งชุดข้อมูลขนาดเท่าๆกันจำนวน *k* กลุ่ม และไม่มีส่วนที่ทับกัน จากนั้นแบ่งชุดข้อมูลเป็น 2 กลุ่ม คือชุดสอน และชุดทดสอบ กระบวนการนี้ถูกทำซ้ำ *k* ครั้ง ด้วยชุดทดสอบที่แตกต่างในแต่ละครั้ง โดยรอบที่ 1 ข้อมูลกลุ่มที่ 1 เป็นชุดทดสอบ ข้อมูลกลุ่มที่ 2 ถึง *k* เป็นข้อมูลสอน รอบที่ 2 ข้อมูลกลุ่มที่ 2 เป็นข้อมูลทดสอบ ข้อมูลกลุ่มที่ 1 และกลุ่มที่ 3 ถึง *k* เป็นข้อมูลสอน สลับกันจนข้อมูลทุกกลุ่มได้เป็นชุดทดสอบ ตัวอย่างเช่น มีข้อมูล 100 ตัวอย่างแบ่งออกเป็น 10 กลุ่ม กลุ่มละ 10 ตัวอย่าง ในรอบที่ 1 จะเอาข้อมูลในกลุ่มที่ 1 เป็นชุดทดสอบ ข้อมูลกลุ่มที่ 2 ถึง 10 เป็นชุดสอน รอบที่ 2 จะใช้ข้อมูลกลุ่มที่ 2 เป็นชุดทดสอบ ข้อมูลกลุ่มที่ 1 และกลุ่มที่ 3 ถึง 10 เป็นชุดสอน สลับกันไปจนทุกกลุ่มได้มีโอกาสเป็นชุดทดสอบครบทุกกลุ่ม จากนั้นจะทำการหาค่าความถูกต้องเฉลี่ย

2.8 การประเมินค่าประสิทธิภาพ

การประเมินค่าประสิทธิภาพ (Performance Evaluation) การทำงานของขั้นตอนวิธีการเรียนรู้ของเครื่องสามารถวัดจากผลลัพธ์การทำนาย (Prediction) โดยค่าของผลลัพธ์ที่ได้จากการทำนาย คือ ค่า True Positive (TP) ค่า True Negative (TN) ค่า False Positive (FP) และค่า False Negative (FN) ตามลำดับ แสดงดังตารางที่ 2.1 (Written and Fank, 2005)

ตารางที่ 2.1 ค่าของคอนฟิวชันเมตริกซ์ (Confusion Matrix) แบบ 2 กลุ่ม

ค่าที่แท้จริง (Actual Class)	**ค่าที่ทำนายได้ (Predicted Class)**	
	Class YES	**Class NO**
Class YES	True Positive	False Negative
Class NO	False Positive	True Negative

ค่าที่ได้จากการทำนาย (Prediction) ในตารางที่ 2.1 อธิบายรายละเอียดได้ดังนี้

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

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

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

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

จากค่าผลลัพธ์ที่ได้สามารถนิยามการวัดค่าประเมิน คือ ค่าความถูกต้อง (Accuracy) เป็นร้อยละของตัวอย่างที่ทำนายถูกในตัวอย่างทั้งหมด คำนวณดังสมการที่ 2.16 ค่าการตอบสนองไว (Sensitivity) เป็นร้อยละของตัวอย่างกลุ่มบวกที่ทำนายถูกในตัวอย่างกลุ่มบวกทั้งหมด คำนวณดังสมการที่ 2.17 และค่าความเฉพาะเจาะจง (Specificity) เป็นร้อยละของตัวอย่างกลุ่มลบที่ทำนายถูกในตัวอย่างกลุ่มลบทั้งหมด คำนวณดังสมการที่ 2.18

$$Accuracy(\%) = \frac{TP + TN}{TP + FN + TN + FP} \times 100 \tag{2.16}$$

$$Sensitivity(\%) = \frac{TP}{TP + FN} \times 100 \tag{2.17}$$

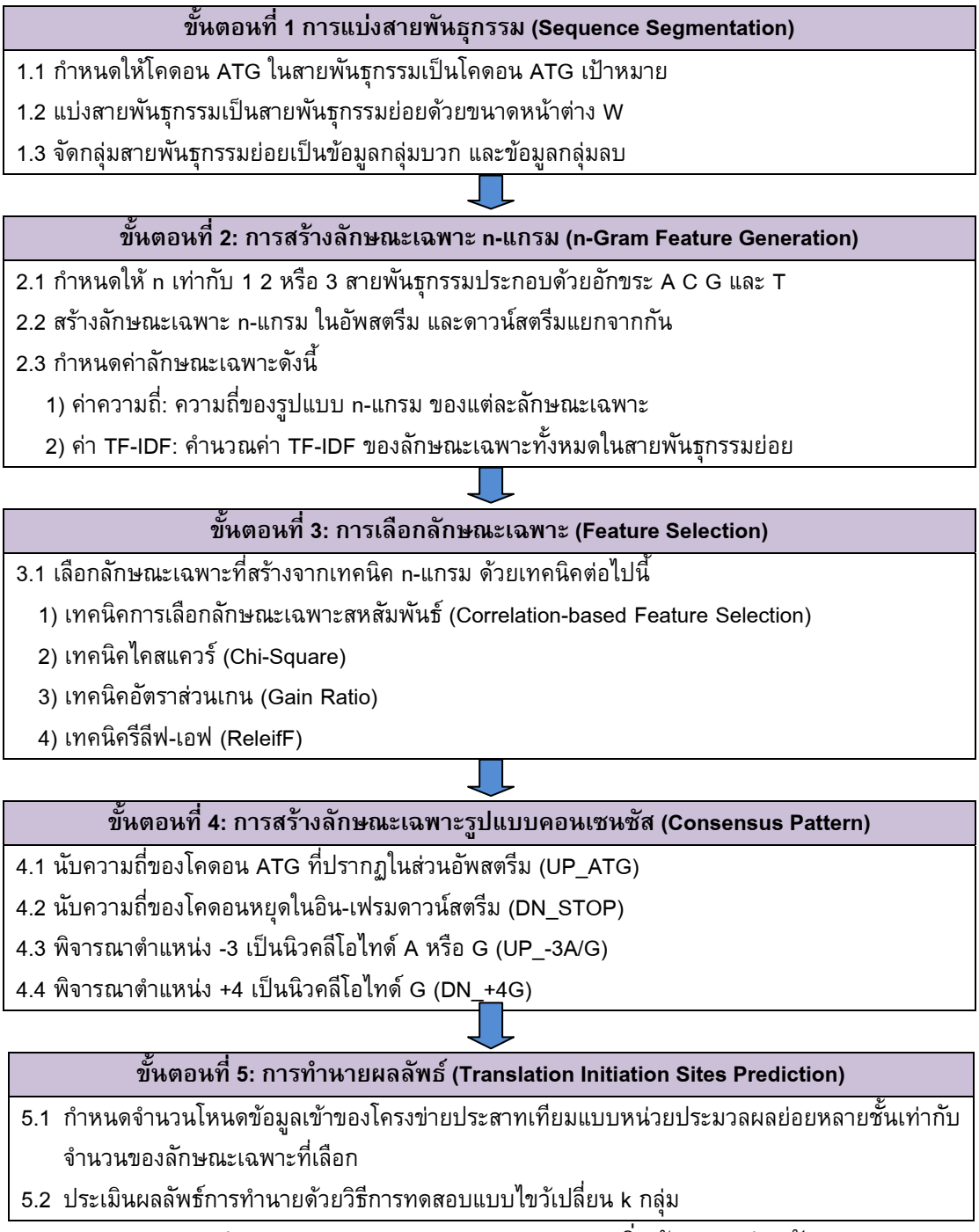
$$Specificity(\%) = \frac{TN}{TN + FP} \times 100 \tag{2.18}$$

บทที่ 3

แบบจำลองการทำนายจุดเริ่มต้นการแปลรหัสโดยใช้วิธี TF-IDF และโครงข่ายประสาทเทียม

วิทยานิพนธ์นี้นำเสนอแบบจำลองการวิเคราะห์สายจีโนมโดยใช้โครงข่ายประสาทเทียม โดยมุ่งเน้นไปที่การวิเคราะห์สายจีโนมเพื่อหาจุดเริ่มต้นการแปลรหัส (Translation Initiation Sites: TIS) โดยนำเสนอแบบจำลองการทำนายจุดเริ่มต้นการแปลรหัสโดยใช้วิธี TF-IDF และโครงข่ายประสาทเทียม (TF-IDF and Neural Networks Approach for Translation Initiation Sites Prediction: TF-IDF-NN-TIS) ซึ่งมีเป้าหมายเพื่อเพิ่มความถูกต้องการทำนายจุดเริ่มต้นการแปลรหัส และลดเวลาในการประมวลผลของโครงข่ายประสาทเทียม ในบทนี้จะกล่าวถึงการออกแบบจำลองการทำนายจุดเริ่มต้นแปลรหัสด้วยวิธี TF-IDF และโครงข่ายประสาทเทียม (TF-IDF-NN-TIS) แสดงดังภาพประกอบ 3.1 แบ่งการทำงานออกเป็น 5 ขั้นตอน คือ 1) การแบ่งสายพันธุกรรม (Sequence Segmentation) 2) การสร้างลักษณะเฉพาะ n-แกรม (n-Gram Feature Generation) 3) การเลือกลักษณะเฉพาะ (Feature Selection) 4) การสร้างลักษณะเฉพาะรูปแบบคอนเซนซัส (Consensus Pattern Feature Generation) และ 5) การทำนายผลลัพธ์ หรือการทำนายจุดเริ่มต้นการแปลรหัส (Translation Initiation Sites Prediction: TIS Prediction)

แบบจำลอง TF-IDF-NN-TIS ขั้นตอนที่ 1 คือ การแบ่งสายพันธุกรรม มีจุดประสงค์เพื่อแบ่งสายดีเอ็นเอในชุดข้อมูลเป็นสายดีเอ็นเอย่อย ขั้นตอนที่ 2 การสร้างลักษณะเฉพาะ n-แกรม มีจุดประสงค์เพื่อเปลี่ยนรูปแบบนิวคลีโอไทด์บนสายดีเอ็นเอด้วยวิธี n-แกรมให้อยู่ในรูปที่สามารถนำมาคำนวณด้วยวิธีเรียนรู้ของเครื่อง ขั้นตอนที่ 3 การเลือกลักษณะเฉพาะ มีจุดประสงค์เพื่อลดจำนวนข้อมูลเข้า และลดเวลาในการสร้างแบบจำลองของโครงข่ายประสาทเทียม เทคนิคการเลือกลักษณะเฉพาะที่นำมาใช้ในแบบจำลองนี้ได้แก่ เทคนิคการเลือกลักษณะเฉพาะสหสัมพันธ์ เทคนิคไคสแควร์ เทคนิคอัตราส่วนเกน และเทคนิครีลีฟ ขั้นตอนที่ 4 การสร้างลักษณะเฉพาะรูปแบบคอนเซนซัส มีจุดประสงค์เพื่อสร้างลักษณะเฉพาะจากลักษณะเด่นโดยทั่วไปรอบจุดเริ่มต้นการแปลรหัสที่ได้มีการนำเสนอในงานวิจัยก่อนหน้าและแบบจำลองการตรวจสอบไรโบโซม และขั้นตอนที่ 5 การทำนายผลลัพธ์ หรือการทำนายจุดเริ่มต้นการแปลรหัส มีจุดประสงค์เพื่อทำนายจุดเริ่มต้นการแปลรหัสโดยใช้โครงข่ายประสาทเทียมรายละเอียดการทำงานของแบบจำลอง TF-IDF-NN-TIS
ในแต่ละขั้นตอนมีดังนี้



ภาพประกอบ 3.1 แบบจำลองการทำนายจุดเริ่มต้นการแปลรหัส
โดยใช้วิธี TF-IDF และ โครงข่ายประสาทเทียม

3.1 ขั้นตอนการแบ่งสายพันธุกรรม

เนื่องจากสายดีเอ็นเอของชุดข้อมูลมีโคดอน ATG ปรากฏอยู่หลายตำแหน่งซึ่งทุกตำแหน่งจัดเป็นโคดอน ATG เป้าหมาย แต่ละสายดีเอ็นเอมีโคดอน ATG เป้าหมายเพียง 1

ตำแหน่งเท่านั้นที่จัดเป็นจุดเริ่มต้นการแปลรหัส หรือเป็นข้อมูลกลุ่มบวก ส่วนโคดอน ATG อื่นๆ จัดเป็นข้อมูลกลุ่มลบจึงต้องแบ่งสายดีเอ็นเอเป็นสายดีเอ็นเอย่อยตามโคดอน ATG เป้าหมาย เพื่อดูความแตกต่างของบริบทโดยรอบระหว่างโคดอน ATG เป้าหมายที่เป็นข้อมูลกลุ่มบวก กับโคดอน ATG เป้าหมายที่เป็นข้อมูลกลุ่มลบ ดังภาพประกอบ 3.2

ขั้นตอนที่ 1 การแบ่งสายพันธุกรรม (Sequence Segmentation)
1.1 กำหนดให้โคดอน ATG ในสายพันธุกรรมเป็นโคดอน ATG เป้าหมาย
1.2 แบ่งสายพันธุกรรมเป็นสายพันธุกรรมย่อยด้วยขนาดหน้าต่าง W
1.3 จัดกลุ่มสายพันธุกรรมย่อยเป็นกลุ่มบวก และกลุ่มลบ

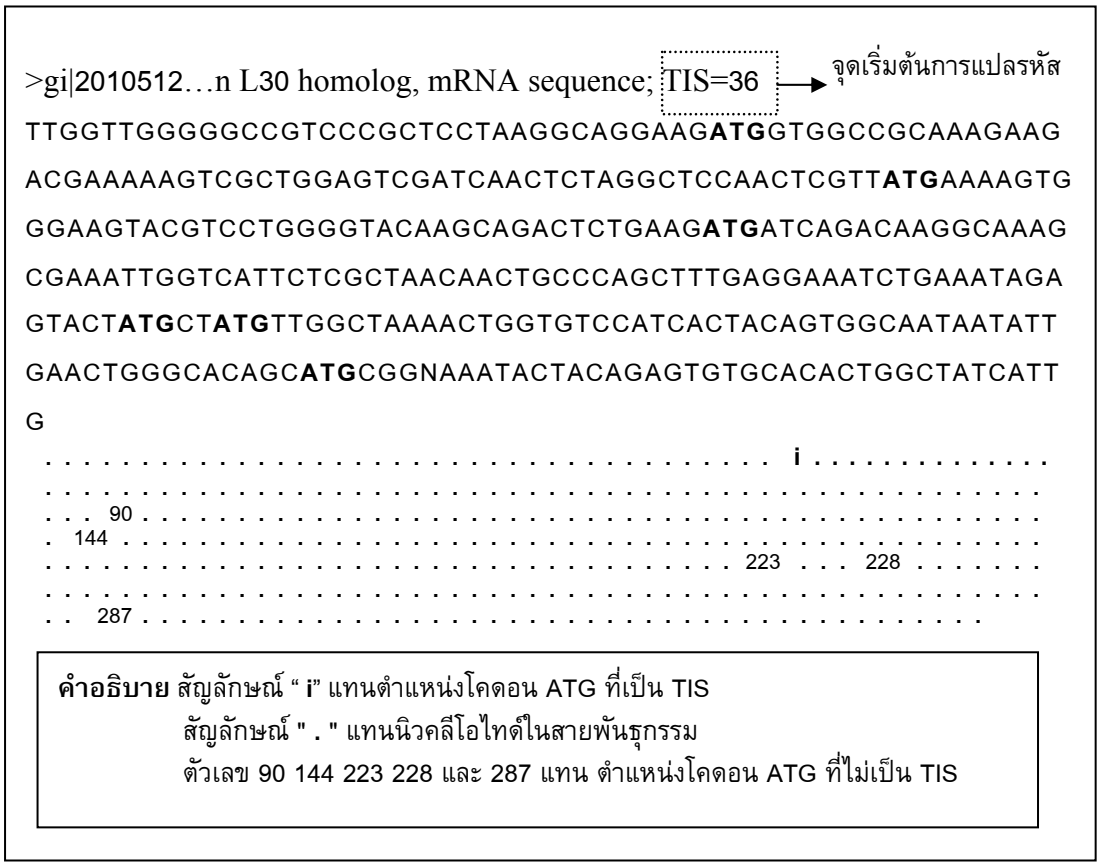
ภาพประกอบ 3.2 ขั้นตอนการแบ่งสายพันธุกรรม

ขั้นตอนที่ 1.1 เลือกโคดอน ATG ทั้งหมดในสายพันธุกรรมเป็นโคดอน ATG เป้าหมายที่พิจารณา ตัวอย่างเช่นสายพันธุกรรมมีความยาวเท่ากับ 327 นิวคลีโอไทด์ มีโคดอน ATG ทั้งหมดในสายพันธุกรรมเท่ากับ 6 ตำแหน่ง ดังนั้นจึงมีโคดอน ATG เป้าหมายที่พิจารณาเท่ากับ 6 ดังแสดงในภาพประกอบ 3.3 ข้อสังเกต สำหรับสายพันธุกรรม 1 สายจะมีโคดอน ATG เป้าหมายที่พิจารณาเป็นจุดเริ่มต้นการแปลเพียง 1 ตำแหน่งเท่านั้น จากภาพประกอบ 3.3 ประกอบด้วยข้อมูลนิวคลีโอไทด์ ตำแหน่งระบุจุดเริ่มต้นการแปลรหัส และสัญลักษณ์ “ . ”แสดงตำแหน่งของนิวคลีโอไทด์ในสายพันธุกรรม

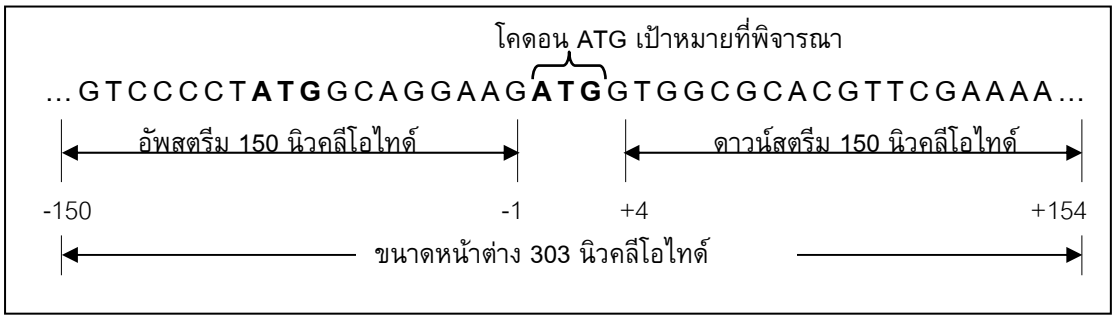
ขั้นตอนที่ 1.2 การแบ่งสายพันธุกรรมเป็นสายพันธุกรรมย่อยใช้ขนาดหน้าต่างของจำนวนนิวคลีโอไทด์รอบโคดอน ATG เป้าหมายที่พิจารณา ซึ่งขนาดหน้าต่าง W จะเท่ากับจำนวนนิวคลีโอไทด์ในส่วนอัพสตรีมบวกจำนวนนิวคลีโอไทด์ในส่วนดาวน์สตรีม โดยที่จำนวนนิวคลีโอไทด์ในส่วนอัพสตรีมและดาวน์สตรีมจะต้องมีค่าเท่ากับค่าเฉลี่ยของขนาดหน้าต่างลบจำนวนนิวคลีโอไทด์ของโคดอน ATG คือ (W -3)/2 ตัวอย่างแสดงดังภาพประกอบ 3.4 การแบ่งสายพันธุกรรมเป็นสายพันธุกรรมย่อยด้วยขนาดหน้าหน้าต่าง 303 นิวคลีโอไทด์ นั่นคือ แต่ละสายพันธุกรรมย่อยจะมีจำนวนของนิวคลีโอไทด์ในอัพสตรีม และดาวน์สตรีมมีค่าเท่ากับ (303-3)/2 =150 นิวคลีโอไทด์

ขั้นตอนที่ 1.3 แบ่งสายพันธุกรรมย่อยเป็น 2 กลุ่ม คือ กลุ่มบวก และกลุ่มลบ โดยสายพันธุกรรมย่อยที่มีตำแหน่งนิวคลีโอไทด์ A ของโคดอน ATG เป้าหมายที่พิจารณาตรงกับตำแหน่งระบุจุดเริ่มต้นการแปลรหัสจัดเป็นข้อมูลกลุ่มบวก ส่วนสายพันธุกรรมย่อยอื่น ๆ จัดเป็นข้อมูลกลุ่มลบ ตัวอย่างเช่นจากภาพประกอบ 3.3 สามารถแบ่งเป็นสายพันธุกรรมย่อยได้ทั้งหมดเท่ากับ 6 สาย ตำแหน่งนิวคลีโอไทด์ A ของโคดอน ATG เป้าหมายของสายพันธุกรรมย่อยที่ 1 2 3 4 5 และ 6 เท่ากับ 36 90 144 223 228 และ 287 ตามลำดับ พบว่า สายพันธุกรรมย่อยที่ 1 มีตำแหน่งนิวคลีโอไทด์ A ของโดคอน ATG เป้าหมายตรงกับตำแหน่งระบุ

จุดเริ่มต้นการแปลรหัส (TIS=36) จึงจัดสายพันธุกรรมย่อยที่ 1 เป็นกลุ่มบวก ส่วนสายพันธุกรรมย่อยที่ 2 3 4 5 และ 6 จัดเป็นกลุ่มลบ หมายเหตุ สำหรับโคดอน ATG เป้าหมายที่พิจารณาของขนาดหน้าต่าง 303 นิวคลีโอไทด์อาจจะมีโคดอน ATG อื่น ๆ ปรากฏอยู่ด้วยแสดงดังภาพประกอบ 3.4



ภาพประกอบ 3.3 ตัวอย่างข้อมูลสายพันธุกรรม



ภาพประกอบ 3.4 การแบ่งสายพันธุกรรมด้วยขนาดหน้าต่าง 303 นิวคลีโอไทด์

3.2 ขั้นตอนการสร้างลักษณะเฉพาะ n-แกรม

เป็นขั้นตอนการเปลี่ยนรูปแบบของลำดับนิวคลีโอไทด์บนสายพันธุกรรมย่อยให้อยู่ในรูปที่สามารถนำไปคำนวณหรือใช้กับวิธีการเรียนรู้ของเครื่องได้ มีรายละเอียดการทำงานดังภาพประกอบ 3.5

ขั้นตอนที่ 2: การสร้างลักษณะเฉพาะ n-แกรม (n-Gram Feature Generation)
2.1 กำหนดให้ n เท่ากับ 1 2 หรือ 3 สายพันธุกรรมประกอบด้วยอักขระ A C G และ T
2.2 สร้างลักษณะเฉพาะ n-แกรม ในอัพสตรีม และดาวน์สตรีมแยกจากกัน
2.3 กำหนดค่าลักษณะเฉพาะดังนี้ 1) ค่าความถี่: นับจำนวนของรูปแบบ n-แกรมของแต่ละลักษณะเฉพาะ 2) ค่า TF-IDF: คำนวณค่า TF-IDF ของลักษณะเฉพาะทั้งหมดในสายพันธุกรรมย่อย

ภาพประกอบ 3.5 ขั้นตอนการสร้างลักษณะเฉพาะ n-แกรม

ขั้นตอนที่ 2.1 ในงานวิจัยนี้กำหนดค่า n เท่ากับ 1 2 หรือ 3 จากงานวิจัยที่ศึกษาก่อนหน้าพบว่าค่า n ที่มากกว่า 3 ไม่สามารถปรับปรุงค่าความถูกต้อง (Tzanis et *al*, 2005) ทั้งนี้ยังเพิ่มความซับซ้อนในการคำนวณ และเพิ่มเวลาการประมวลผลลัพธ์ของโครงข่ายประสาทเทียม และสายพันธุกรรมประกอบด้วยอักขระ 4 ตัว คือ A C G และ T

ขั้นตอนที่ 2.2 1-แกรมมีรูปแบบที่เป็นไปได้ทั้งหมดเท่ากับ 4 (4^1 = 4) รูปแบบ ได้แก่ A C G และ T 2-แกรม มีรูปแบบที่เป็นไปได้ทั้งหมดเท่ากับ 16 (4^2 = 16) รูปแบบ ได้แก่ AA AC AG AT CA CC CG CT GA GC GG GT TA TC TG และ TT และ 3-แกรม มีรูปแบบที่เป็นไปได้ทั้งหมดเท่ากับ 64 (4^3 = 64) รูปแบบ ได้แก่ AAA AAC AAG AAT ACA ACC ACG ACT AGA AGC AGG AGT ATA ATC ATG ATT CAA CAC CAG CAT CCA CCC CCG CCT CGA CGC CGG CGT CTA CTC CTG CTT GAA GAC GAG GAT GCA GCC GCG GCT GGA GGC GGG GGT GTA GTC GTG GTT TAA TAC TAG TAT TCA TCC TCG TCT TGA TGC TGG TGT TTA TTC TTG และ TTT

สร้างลักษณะเฉพาะจากเทคนิค n-แกรม โดยในแต่ละสายพันธุกรรมย่อยแยกพิจารณาระหว่างอัพสตรีม และดาวน์สตรีม ตัวอย่างเช่น ลักษณะเฉพาะ 1-แกรม และ 2-แกรม สำหรับแต่ละสายพันธุกรรมย่อยจะประกอบด้วยลักษณะเฉพาะในส่วนอัพสตรีมทั้งหมดเท่ากับ 20 (4+16 = 20) และ ดาวน์สตรีมมีจำนวนลักษณะเฉพาะทั้งหมดเท่ากับ 20 (4+16 = 20) ดังตารางที่ 3.2 ดังนั้น แต่ละสายพันธุกรรมย่อยจึงมีลักษณะเฉพาะทั้งหมดเท่ากับ 40 ตารางที่ 3.1 แสดงตัวอย่าง ลักษณะเฉพาะ 1-แกรม ในอัพสตรีมแทนด้วยสัญลักษณ์ UP_A UP_C UP_G และUP_T ลักษณะเฉพาะ 1-แกรม ในดาวน์สตรีมแทนด้วยสัญลักษณ์ DN_A DN_C DN_G

และ DN_T ลักษณะเฉพาะ 2-แกรม ในอัพสตรีมแทนด้วยสัญลักษณ์ UP_AA UP_AC UP_AG เป็นต้น และลักษณะเฉพาะ 2-แกรม ในดาวน์สตรีมแทนด้วยสัญลักษณ์ DN_AA DN_AC DN_AG เป็นต้น

ตารางที่ 3.1 ลักษณะเฉพาะที่สร้างจาก 1-แกรม และ 2-แกรม

เทคนิคการสร้าง ลักษณะเฉพาะ	รูปแบบลักษณะเฉพาะ	จำนวน ลักษณะเฉพาะ
1-แกรมอัพสตรีม	UP_A UP_C UP_G UP_T	4
1-แกรมดาวน์สตรีม	DN_A DN_C DN_G DN_T	4
2-แกรมอัพสตรีม	UP_AA UP_AC UP_AG UP_AT UP_CA UP_CC UP_CG UP_CT UP_GA UP_GC UP_GG UP_GT UP_TA UP_TC UP_TG UP_TA	16
2-แกรมดาวน์สตรีม	DN_AA DN_AC DN_AG DN_AT DN_CA DN_CC DN_CG DN_CT DN_GA DN_GC DN_GG DN_GT DN_TA DN_TC DN_TG DN_TA	16
จำนวนลักษณะเฉพาะทั้งหมด		40

ขั้นตอนที่ 2.3 กำหนดให้ เซตของสายพันธุกรรมย่อย $S = \{s_1, s_2, \ldots, s_N\}$ มีจำนวนสายพันธุกรรมย่อยเท่ากับ N และ $T = \{t_1, t_2, \ldots, t_M\}$ เป็นเซตของรูปแบบ n-แกรมทั้งในส่วนอัพสตรีมและดาวน์สตรีมแยกจากกัน ดังนั้นจำนวนลักษณะเฉพาะ n-แกรมสำหรับแต่ละสายพันธุกรรมย่อยเท่ากับ M ทั้งนี้ค่า M จะแตกต่างกันเมื่อกำหนดค่า n แตกต่างกัน แสดงดังตารางที่ 3.2 ตัวอย่างเช่นจำนวนลักษณะเฉพาะ 1-แกรมเท่ากับ 8 จำนวนลักษณะเฉพาะ 2-แกรมเท่ากับ 32 จำนวนลักษณะเฉพาะ 3-แกรมเท่ากับ 128 จำนวนลักษณะเฉพาะ 1-แกรม และ 2-แกรม เท่ากับ 40 เป็นต้น

หมายเหตุ ตัวอย่างลักษณะเฉพาะ 2-แกรม และ 3-แกรม สำหรับแต่ละสายพันธุกรรมย่อยจะประกอบด้วยลักษณะเฉพาะในส่วนอัพสตรีมทั้งหมดเท่ากับ 80 (16+64 = 80) และ ดาวน์สตรีมมีจำนวนลักษณะเฉพาะทั้งหมดเท่ากับ 80 (16+64 = 80) ถ้ารวมอัพสตรีม และดาวน์สตรีมจะมีจำนวนลักษณะเฉพาะ M เท่ากับ 160 (80+80=160) ดังตารางที่ 3.2

ตารางที่ 3.2 จำนวนลักษณะเฉพาะสำหรับเทคนิค 1-แกรม 2-แกรม และ 3-แกรม

เทคนิค n-แกรม	จำนวนลักษณะเฉพาะจากอัพสตรีม	จำนวนลักษณะเฉพาะจากดาวน์สตรีม	จำนวนลักษณะเฉพาะทั้งหมด (M)
1-แกรม	4	4	8
2-แกรม	16	16	32
3-แกรม	64	64	128
1-แกรม และ 2-แกรม	4 + 16 = 20	4 + 16 = 20	40
2-แกรม และ 3-แกรม	16+64 = 80	16 + 64 = 80	160
1-แกรม และ 3-แกรม	4 + 64 = 68	4+ 64 = 68	136
1-แกรม 2-แกรม และ 3-แกรม	4 + 16 + 64 = 84	4 + 16 + 64 = 84	168

สำหรับค่าลักษณะเฉพาะที่ได้จากเทคนิค n-แกรม กำหนดค่าลักษณะเฉพาะดังนี้

1) ค่าความถี่: วิธีนี้ใช้ความถี่ของรูปแบบ t ในสายพันธุกรรมย่อย s ซึ่งสามารถคำนวณได้ดังสมการที่ (3.1)

$$v_{s,t} = f_{s,t} \tag{3.1}$$

กำหนดให้ $v_{s,t}$ คือ ค่าลักษณะเฉพาะของรูปแบบ t ในสายพันธุกรรมย่อย s

$f_{s,t}$ คือ ความถี่ของรูปแบบ t ในสายพันธุกรรมย่อย s

2) ค่า TF-IDF: วิธีนี้ใช้การกระจายของแต่ละรูปแบบ n-แกรมตลอดของทุกสายพันธุกรรมย่อยในชุดข้อมูลสอนซึ่งสามารถคำนวณได้ดังสมการที่ (3.2)

$$v_{s,t} = \begin{cases} f_{s,t} \times log\dfrac{N}{n_t}; & if\ f_{s,t} \geq 1 \\ 0 \quad ; & otherwises \end{cases} \tag{3.2}$$

กำหนดให้ $v_{s,t}$ คือ ค่าลักษณะเฉพาะของรูปแบบ t ในสายพันธุกรรมย่อย s

$f_{t,s}$ คือ ความถี่ของรูปแบบ t ในสายพันธุกรรมย่อย s

N คือ จำนวนสายพันธุกรรมย่อยทั้งหมดในชุดข้อมูลสอน

n_t คือ จำนวนครั้งทั้งหมดที่รูปแบบ t ปรากฏในชุดข้อมูลสอน

3.3 ขั้นตอนการเลือกลักษณะเฉพาะ

เป็นขั้นตอนการลดมิติของข้อมูลเพื่อเพิ่มประสิทธิภาพการทำนายผลลัพธ์ และลดเวลาการประมวลผล ในวิทยานิพนธ์นี้ใช้เทคนิคการเลือกลักษณะเฉพาะสหสัมพันธ์ เทคนิคไคสแควร์ เทคนิคอัตราส่วนเกน และเทคนิครีลีฟ-เอฟ แสดงดังภาพประกอบ 3.6

ขั้นตอนที่ 3: การเลือกลักษณะเฉพาะ (Feature Selection)
3.1 เลือกลักษณะเฉพาะที่สร้างจากเทคนิค n-แกรม ด้วยเทคนิคต่อไปนี้
1) เทคนิคการเลือกลักษณะเฉพาะสหสัมพันธ์ (Correlation-based Feature Selection)
2) เทคนิคไคสแควร์ (Chi-Square)
3) เทคนิคอัตราส่วนเกน (Gain Ratio)
4) เทคนิครีลีฟ-เอฟ (ReleifF)

ภาพประกอบ 3.6 ขั้นตอนการเลือกลักษณะเฉพาะ

3.4 ขั้นตอนการสร้างลักษณะเฉพาะรูปแบบคอนเซนซัส (Consensus Pattern Feature Generation)

ลักษณะเฉพาะรูปแบบคอนเซนซัส (Consensus Pattern) เป็นการสร้างลักษณะเฉพาะจากแบบจำลองการตรวจสอบไรโบโซม (Kozak, 1989) และลักษณะเด่นส่วนใหญ่รอบจุดเริ่มต้นการแปลรหัสของ Kozak (Kozak, 1987) สำหรับขั้นตอนการสร้างลักษณะเฉพาะรูปแบบคอนเซนซัสแสดงดังภาพประกอบ 3.7 โดยมีรายละเอียดดังนี้

ขั้นตอนที่ 4.1 พิจารณาสายพันธุกรรมย่อยถ้าตำแหน่ง -3 เป็นนิวคลีโอไทด์ A หรือ G กำหนดค่าลักษณะเฉพาะเป็น 1 อื่นๆ กำหนดค่าลักษณะเฉพาะเป็น 0 (Kozak, 1987)

ขั้นตอนที่ 4.2 พิจารณาสายพันธุกรรมย่อยถ้าตำแหน่ง +4 เป็นนิวคลีโอไทด์ A หรือ G กำหนดค่าลักษณะเฉพาะเป็น 1 อื่นๆ กำหนดค่าลักษณะเฉพาะเป็น 0 (Kozak, 1987)

ขั้นตอนที่ 4: การสร้างลักษณะเฉพาะรูปแบบคอนเซนซัส (Consensus Pattern Feature Generation)
4.1 พิจารณาตำแหน่ง -3 เป็นนิวคลีโอไทด์ A หรือ G (UP_-3A/G)
4.2 พิจารณาตำแหน่ง +4 เป็นนิวคลีโอไทด์ G (DN_+4G)
4.3 นับความถี่ของโคดอน ATG ที่ปรากฏในส่วนอัพสตรีม (UP_ATG)
4.4 นับความถี่ของโคดอนหยุดในอิน-เฟรมดาวน์สตรีม (DN_STOP)

ภาพประกอบ 3.7 ขั้นตอนการสร้างลักษณะเฉพาะรูปแบบคอนเซนซัส

ขั้นตอนที่ 4.3 นับความถี่ของโคดอน ATG แบบทุกเฟรมที่ปรากฏในส่วนอัพสตรีมของสายพันธุกรรมย่อย (Kozak, 1989)

ขั้นตอนที่ 4.4 นับความถี่ของโคดอนหยุดในอิน-เฟรมดาวน์สตรีมสายพันธุกรรมย่อย (Kozak, 1989)

ลักษณะเฉพาะที่สร้างรูปแบบคอนเซนซัสจึงมี 4 รูปแบบแสดงดังตารางที่ 3.2 เมื่อกำหนดให้นิวคลีโอไทด์ A ของโคดอน ATG เป้าหมายเป็นตำแหน่ง +1 โดยลักษณะเฉพาะที่สร้างจากรูปแบบคอนเซนซัสช่วยให้ระบุจุดเริ่มต้นการแปลรหัสได้ถูกต้องมากยิ่งขึ้น

ตารางที่ 3.3 ลักษณะเฉพาะที่สร้างจากรูปแบบคอนเซนซัส

รูปแบบ	คำอธิบาย	จำนวนลักษณะเฉพาะ
UP_−3A/G	ลักษณะเฉพาะบูลีนซึ่งเป็นจริงถ้ามีนิวคลีโอไทด์ A หรือ G ที่ตำแหน่ง -3	1
DN_+4G	ลักษณะเฉพาะบูลีนซึ่งเป็นจริงถ้ามีนิวคลีโอไทด์ G ที่ตำแหน่ง +4	1
UP_ATG	นับความถี่ของโคดอน ATG แบบทุกเฟรมที่ปรากฏในส่วนอัพสตรีม	1
DN_STOP	นับความถี่ของโดคอนหยุดแบบอิน-แฟรม (TAA TAG หรือ TGA) ที่ปรากฏในส่วนดาวน์สตรีม	1
จำนวนลักษณะเฉพาะทั้งหมด		4

3.5 การทำนายผลลัพธ์ (Translation Initiation Sites Prediction)

เป็นขั้นตอนการทำนายจุดเริ่มต้นการแปลรหัสด้วยโครงข่ายประสาทเทียมแบบหน่วยประมวลผลย่อยหลายชั้น ใช้ฟังก์ชันกระตุ้นซิกมอยด์ และใช้การประเมินประสิทธิภาพแบบไขว้เปลี่ยน *k* กลุ่ม (k-Fold Cross Validation) แสดงดังภาพประกอบ 3.8 มีรายละเอียดดังนี้

ขั้นตอนที่ 5: การทำนายผลลัพธ์ (Translation Initiation Sites Prediction)
5.1 กำหนดสถาปัตยกรรมของโครงข่ายประสาทเทียมแบบไปข้างหน้าหลายชั้น จำนวนโหนดข้อมูลเข้า เท่ากับ จำนวนลักษณะเฉพาะที่เลือก จำนวนโหนดซ่อน เท่ากับ จำนวนตัวแปรเข้า จำนวนโหนดผลลัพธ์ เท่ากับ 1
5.2 ประเมินผลลัพธ์การทำนายด้วยวิธีการทดสอบแบบไขว้เปลี่ยน k กลุ่ม

ภาพประกอบ 3.8 ขั้นตอนการทำนายผลลัพธ์

ขั้นตอนที่ 5.1 กำหนดใช้โครงข่ายประสาทเทียมแบบไปข้างหน้าหลายชั้น และฟังก์ชันกระตุ้นซิกมอยด์ สำหรับสถาปัตยกรรมโครงข่ายประสาทเทียมกำหนดให้จำนวนโหนดของชั้นข้อมูลเข้าเท่ากับ จำนวนของลักษณะเฉพาะที่เลือก ชั้นซ่อน (แทนด้วยสัญลักษณ์ H) ประกอบด้วยจำนวนโหนดเท่ากับจำนวนตัวแปรเข้า ชั้นผลลัพธ์ (แทนด้วยสัญลักษณ์ O) ประกอบด้วย 1 โหนด

ขั้นตอนที่ 5.2 การทดสอบไขว้เปลี่ยนแบบ k กลุ่ม เป็นวิธีการประเมินผลลัพธ์การทำนายที่ข้อมูลทุกตัวเป็นชุดข้อมูลสอน และชุดข้อมูลทดสอบ ในวิทยานิพนธ์นี้กำหนดค่า k เท่ากับ 10 จากนั้นประเมินค่าผลลัพธ์ด้วยค่าความถูกต้อง (Accuracy) ค่าการตอบสนองไว (Sensitivity) และค่าเฉพาะเจาะจง (Specificity) โดยหาค่าเฉลี่ยของค่าความถูกต้อง ค่าการตอบสนองไว และค่าเฉพาะเจาะจง จากชุดทดสอบทั้ง 10 กลุ่ม โดยทำการทดลองซ้ำ 3 ครั้ง และเลือกข้อมูลแบบสุ่ม

บทที่ 4

โปรแกรมการทำนายจุดเริ่มต้นการแปลรหัส โดยใช้วิธี TF-IDF และโครงข่ายประสาทเทียม

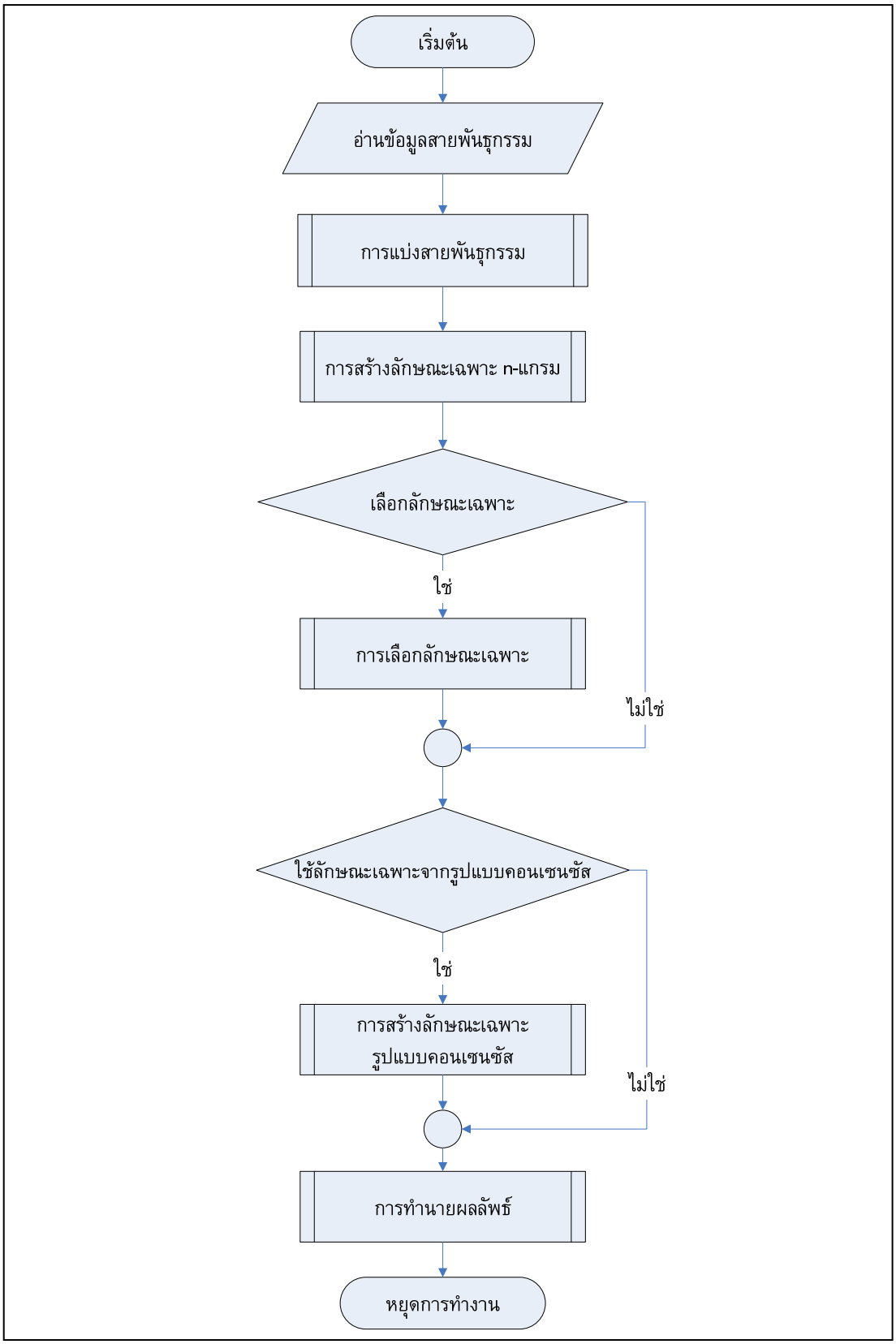
การพัฒนาโปรแกรมการทำนายจุดเริ่มต้นการแปลรหัสโดยใช้วิธี TF-IDF และโครงข่ายประสาทเทียมจะใช้แนวคิดและลำดับขั้นตอนการทำงานตามแบบจำลองการทำนายจุดเริ่มต้นการแปลรหัสโดยใช้วิธี TF-IDF และโครงข่ายประสาทเทียม (The TF-IDF and Neural Networks Approach for Translation Initiation Sites Prediction: TF-IDF-NN-TIS) การทำงานของโปรแกรมจะอธิบายด้วยผังการทำงานของโปรแกรม และแสดงตัวอย่างการทำงานของโปรแกรมด้วย

4.1 ผังการทำงานของโปรแกรม

ในการพัฒนาโปรแกรมสามารถแสดงผังงานโปรแกรม (Program Flow Chart) ของโปรแกรมการทำนายจุดเริ่มต้นการแปลรหัสโดยใช้วิธี TF-IDF และโครงข่ายประสาทเทียม ได้ดังนี้

4.1.1 ผังงานโปรแกรมหลัก

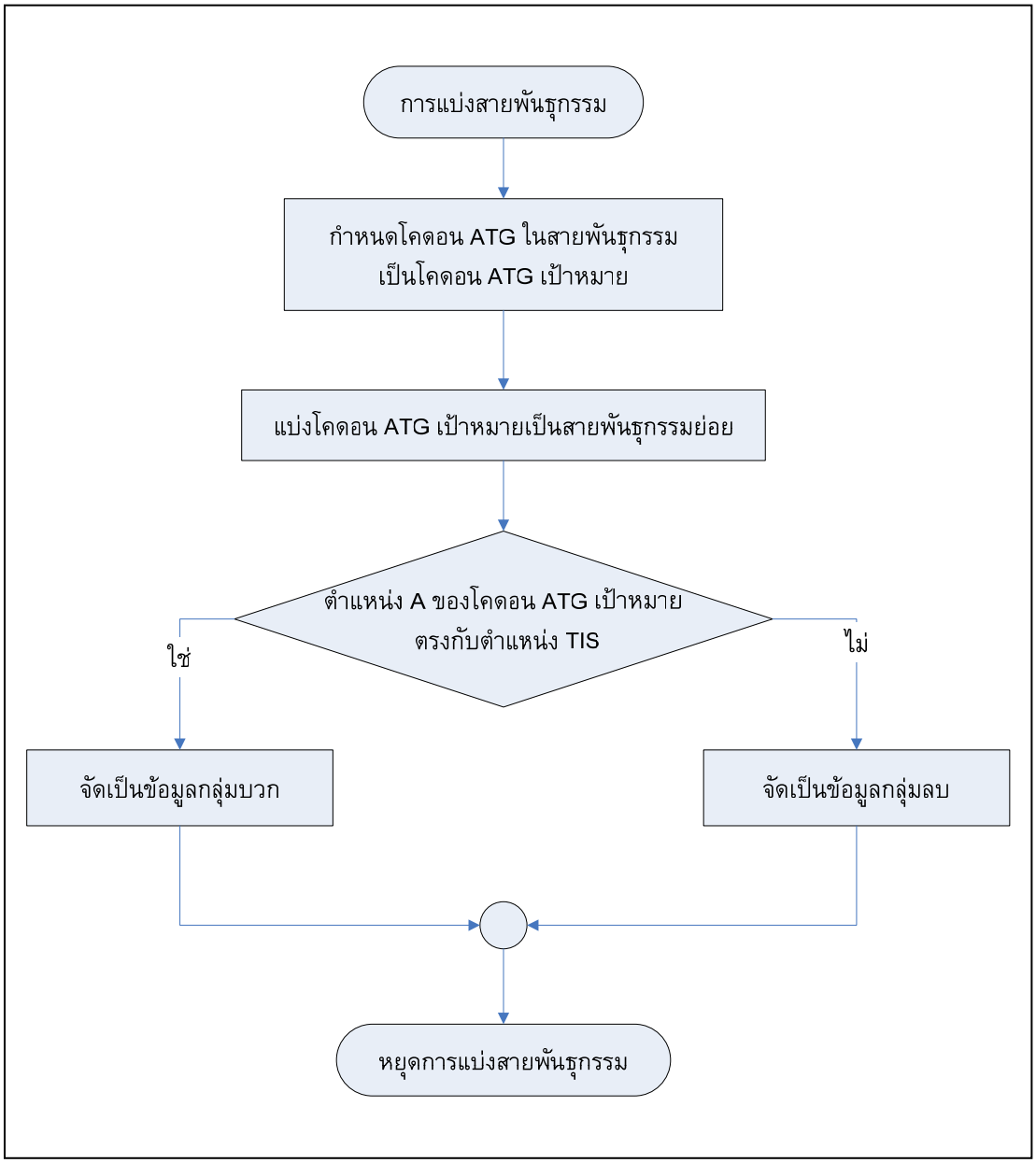
ผังงานโปรแกรมหลักของโปรแกรมการทำนายจุดเริ่มต้นการแปลรหัสโดยใช้วิธี TF-IDF และโครงข่ายประสาทเทียม แสดงดังภาพประกอบ 4.1



ภาพประกอบ 4.1 ผังงานโปรแกรมหลัก

4.1.2 ผังงานโปรแกรมการแบ่งสายพันธุกรรม

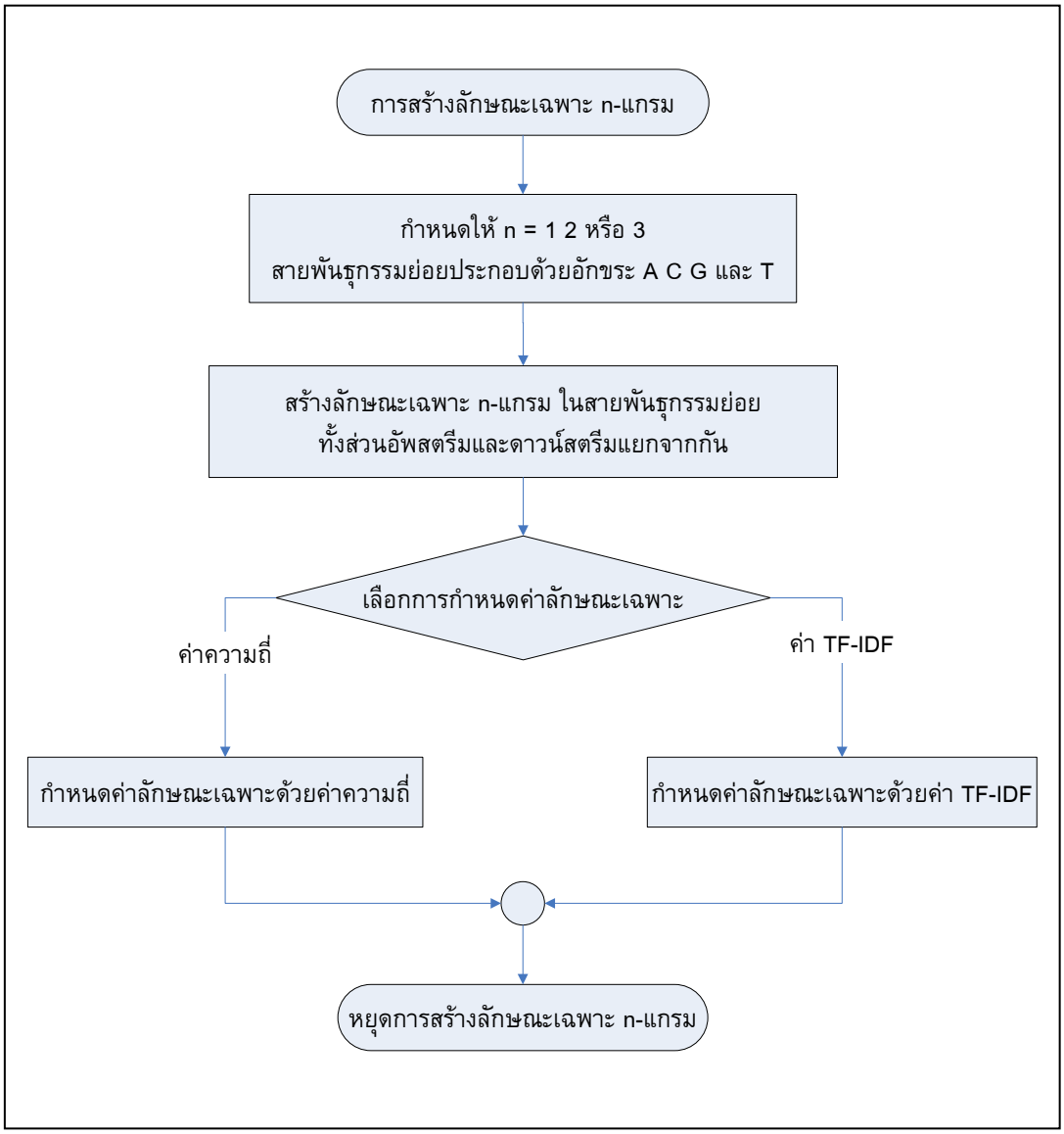
ผังงานโปรแกรมการแบ่งสายพันธุกรรมแสดงดังภาพประกอบ 4.2



ภาพประกอบ 4.2ผังงานโปรแกรมการแบ่งสายพันธุกรรม

4.1.3 ผังงานโปรแกรมการสร้างลักษณะเฉพาะ n-แกรม

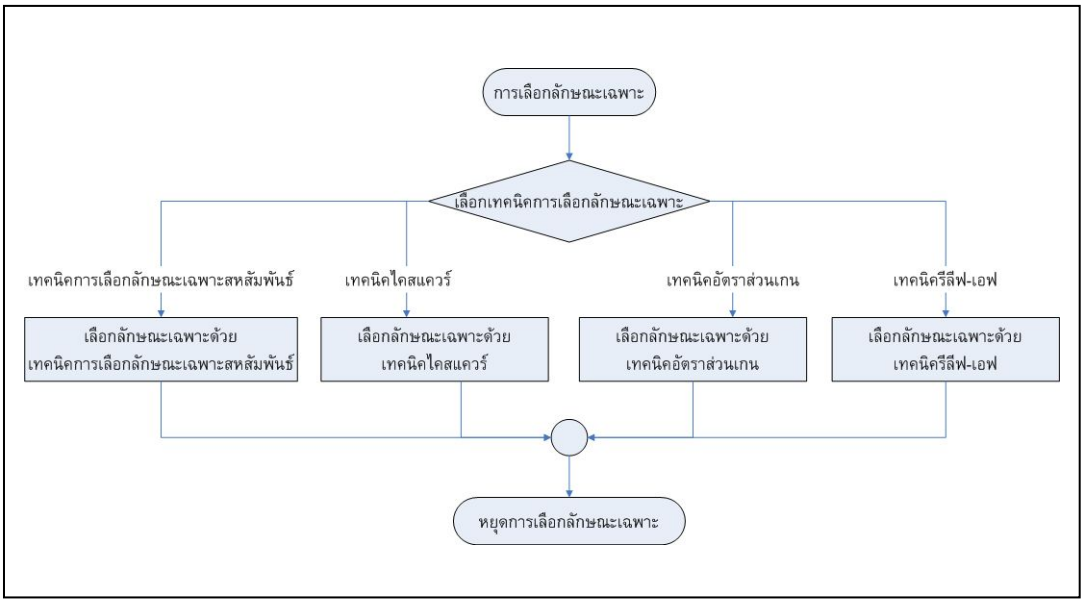
ผังงานโปรแกรมการสร้างลักษณะเฉพาะ n-แกรม แสดงดังภาพประกอบ 4.3



ภาพประกอบ 4.3 ผังงานโปรแกรมการสร้างลักษณะเฉพาะ n-แกรม

4.1.4 ผังงานโปรแกรมการเลือกลักษณะเฉพาะ

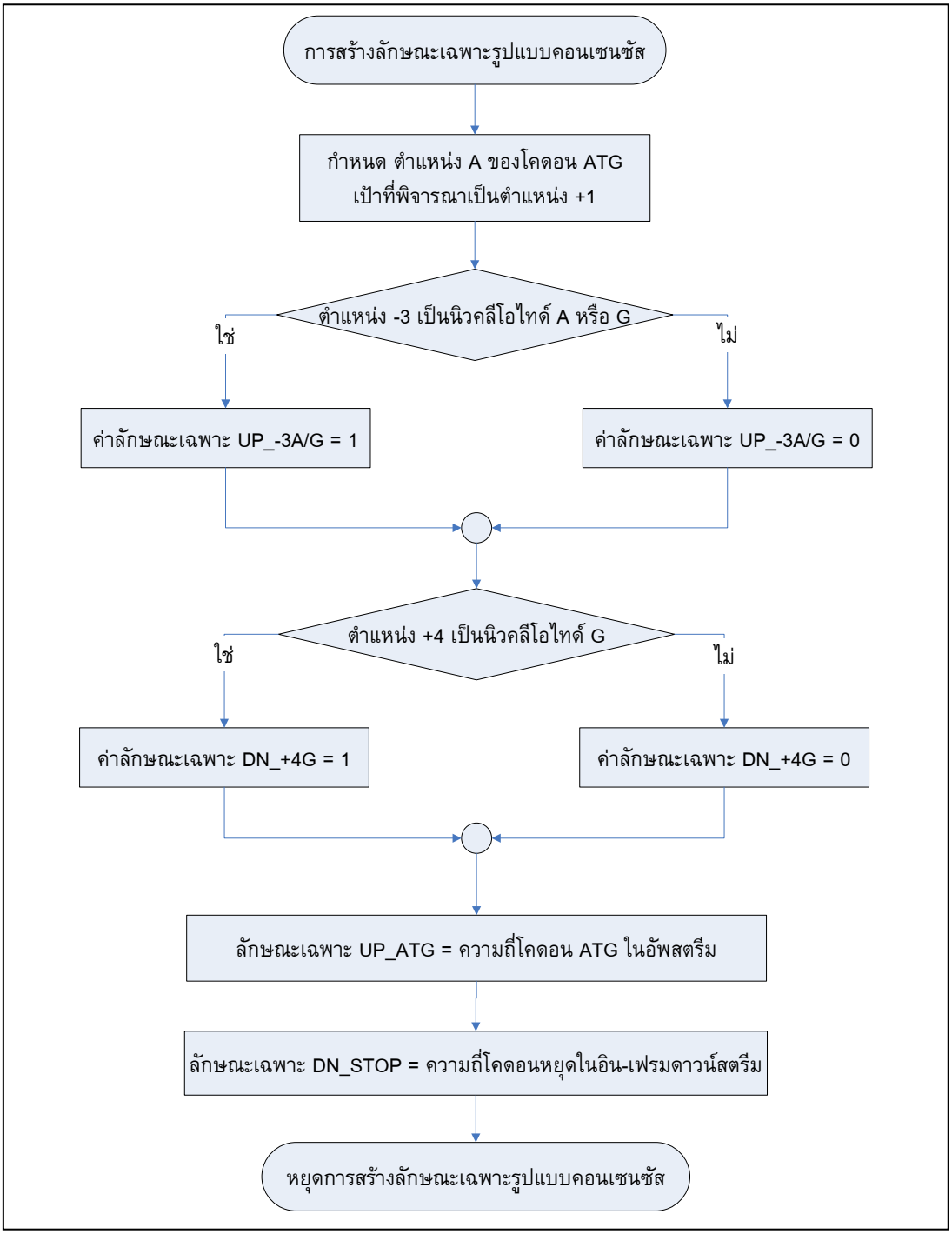
ผังงานโปรแกรมของการเลือกลักษณะเฉพาะแสดงดังภาพประกอบ 4.4



ภาพประกอบ 4.4 ผังงานโปรแกรมการเลือกลักษณะเฉพาะ

4.1.5 ผังงานโปรแกรมการสร้างลักษณะเฉพาะรูปแบบคอนเซนซัส

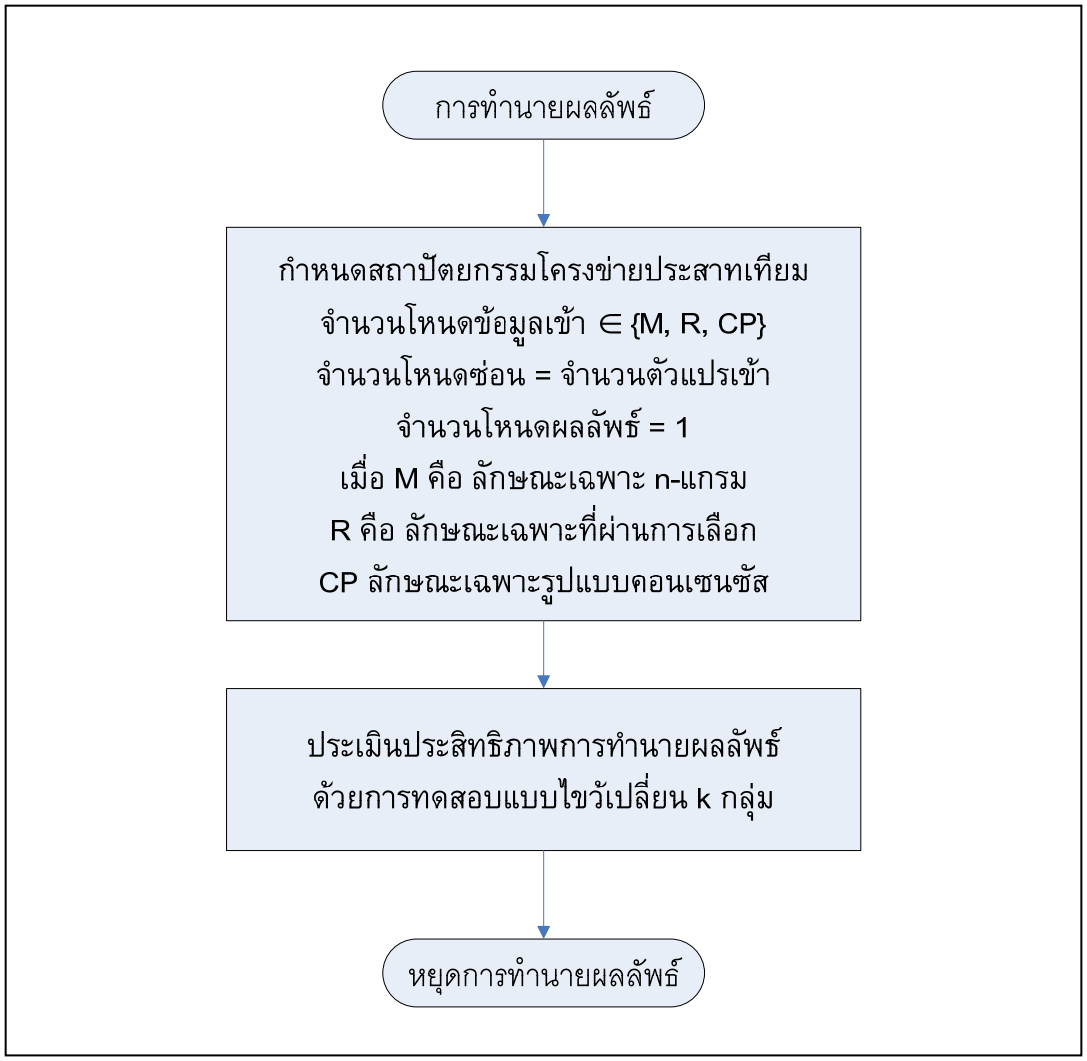
ผังงานโปรแกรมการสร้างลักษณะเฉพาะรูปแบบคอนเซนซัสแสดงดังภาพประกอบ 4.5



ภาพประกอบ 4.5 ผังงานโปรแกรมการสร้างลักษณะเฉพาะจากรูปแบบคอนเซนซัส

4.1.6 ผังงานโปรแกรมการทำนายผลลัพธ์

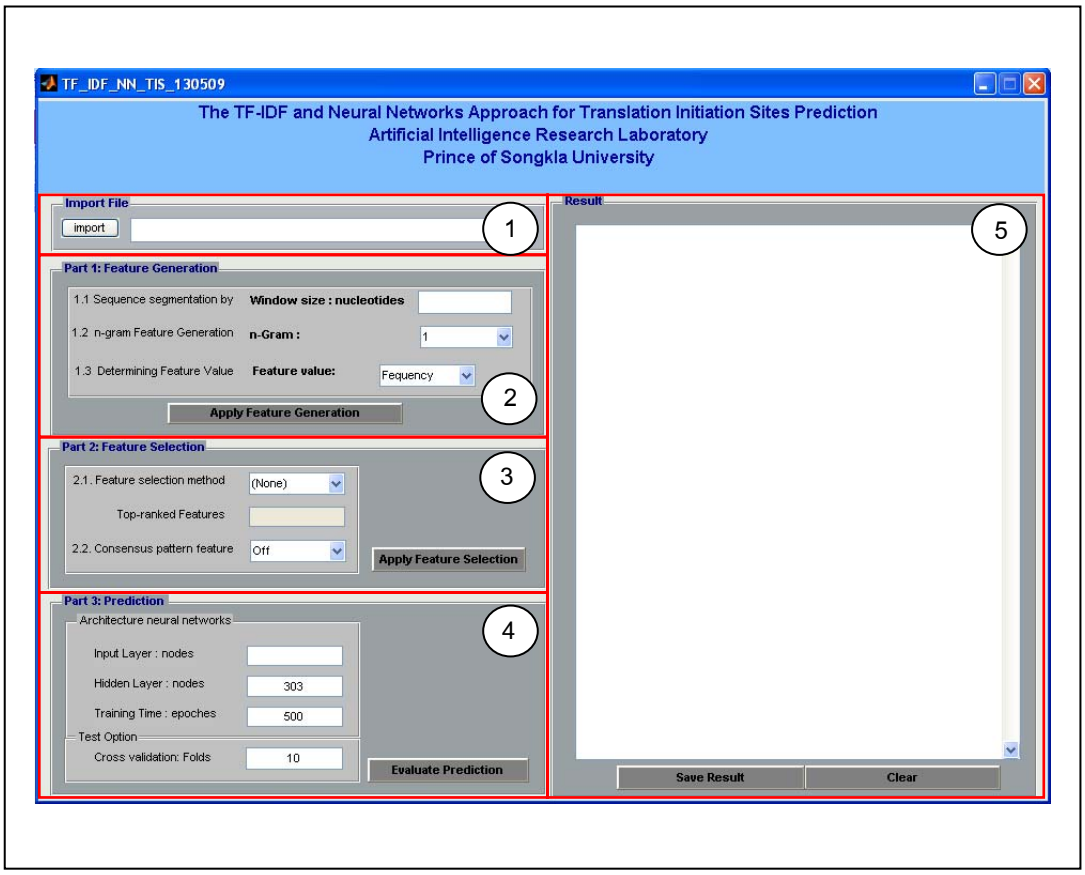
ผังงานโปรแกรมการทำนายผลลัพธ์แสดงดังภาพประกอบ 4.6



ภาพประกอบ 4.6 ผังงานโปรแกรมการทำนายผลลัพธ์

4.2 การทำงานของโปรแกรม

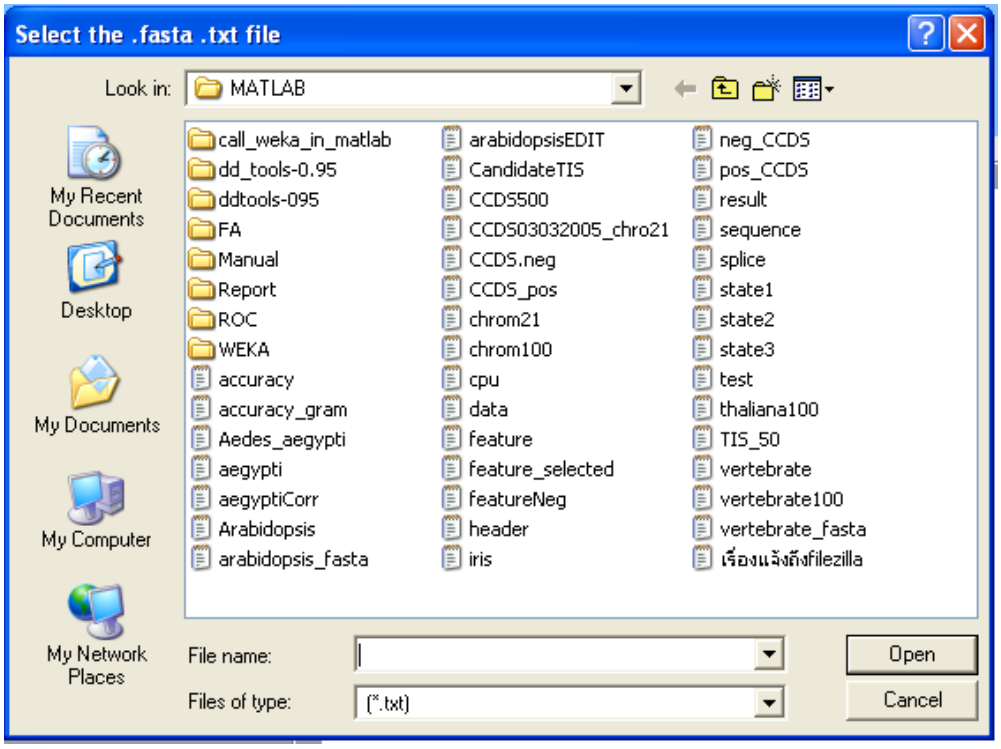
เมื่อเปิดโปรแกรมการทำนายจุดเริ่มต้นการแปลรหัสโดยใช้วิธี TF-IDF และโครงข่ายประสาทเทียม จะปรากฏหน้าจอให้ผู้ใช้ทำงาน หน้าจอหลักของโปรแกรมประกอบด้วยองค์ประกอบ 5 ส่วนหลัก คือ 1) ส่วนการนำเข้าข้อมูล 2) ส่วนสร้างลักษณะเฉพาะ n-แกรม 3) ส่วนการเลือกลักษณะเฉพาะ 4) ส่วนการทำนายผลลัพธ์ และ 5) ส่วนการแสดงผลการทำงาน แสดงดังภาพประกอบ 4.7 โดยแต่ละส่วนมีรายละเอียดการทำงานดังนี้



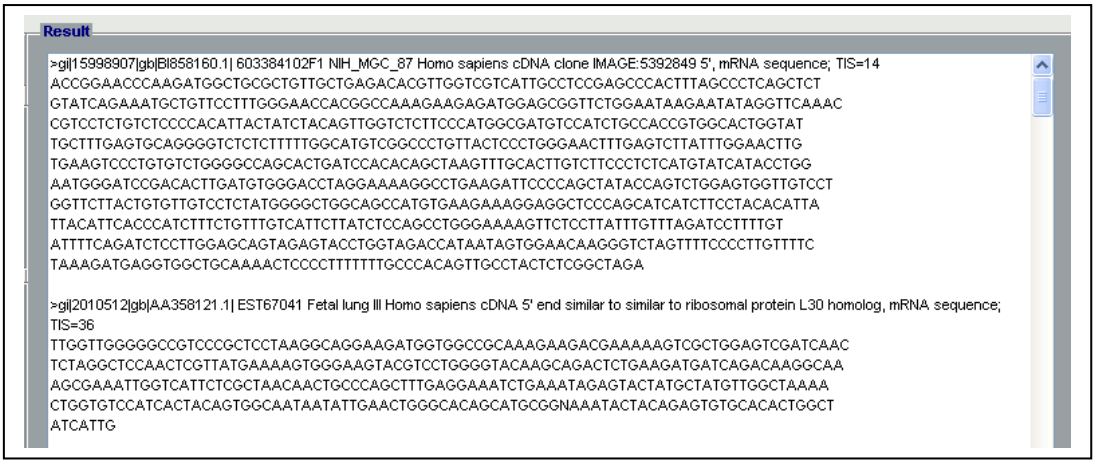
ภาพประกอบ 4.7 หน้าจอหลักของโปรแกรมการทำนายจุดเริ่มต้นการแปลรหัส
โดยใช้วิธี TF-IDF และโครงข่ายประสาทเทียม

4.2.1 ส่วนการนำเข้าข้อมูล

ไฟล์สายพันธุกรรมที่นำมาประมวลผลเป็นไฟล์ที่มีนามสกุล *.txt หรือ *.fasta วิธีการทำงาน คือ เริ่มต้นผู้ใช้กดปุ่ม “import” จะแสดงหน้าต่างสำหรับการเลือกแหล่งข้อมูล และไฟล์แสดงดังภาพประกอบ 4.8 โดยหน้าต่างจะแสดงเฉพาะรูปแบบไฟล์ที่มีนามสกุล *.txt หรือ *.fasta เท่านั้น จากนั้นโปรแกรมจะแสดงข้อมูลสายพันธุกรรมในส่วนแสดงผลการทำงานแสดงดังภาพประกอบ 4.9

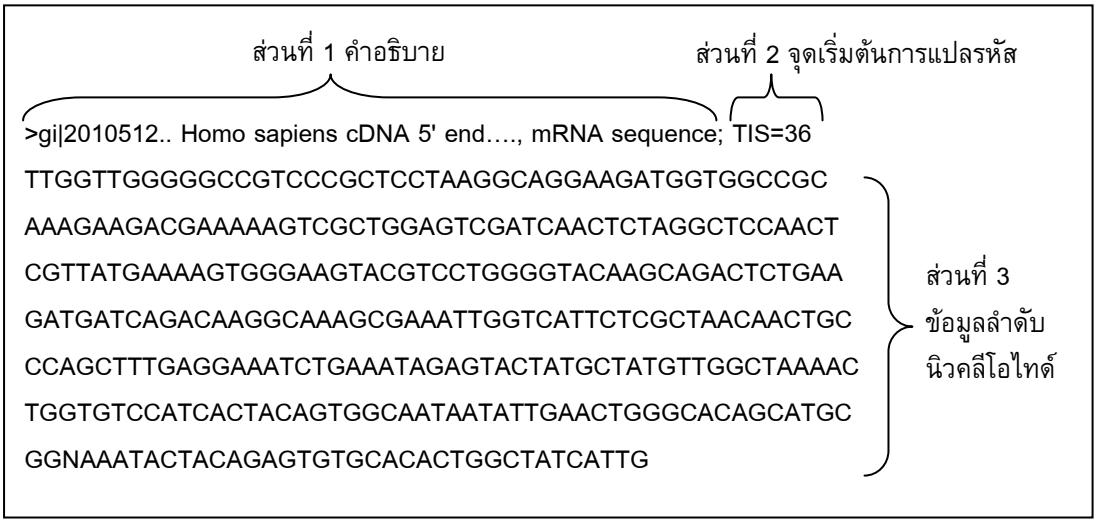


ภาพประกอบ 4.8 หน้าต่างการเลือกแหล่งข้อมูลสายพันธุกรรม



ภาพประกอบ 4.9 ข้อมูลสายพันธุกรรม

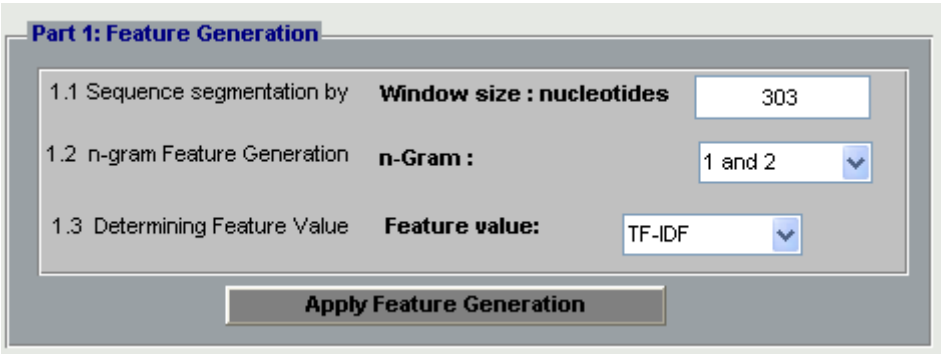
รูปแบบสายพันธุกรรมของข้อมูลนำเข้าอยู่ในรูปแบบฟาสต์-เอ (FASTA) โดยประกอบด้วย 3 ส่วนหลัก แสดงดังภาพประกอบ 4.10 ส่วนแรกเป็นคำอธิบายซึ่งเริ่มต้นด้วยเครื่องหมาย “>” ตามด้วยชื่อและแหล่งที่มาของข้อมูล ส่วนที่ 2 เป็นตำแหน่งจุดเริ่มต้นการแปลรหัส และส่วนที่ 3 เป็นข้อมูลลำดับนิวคลีโอไทด์



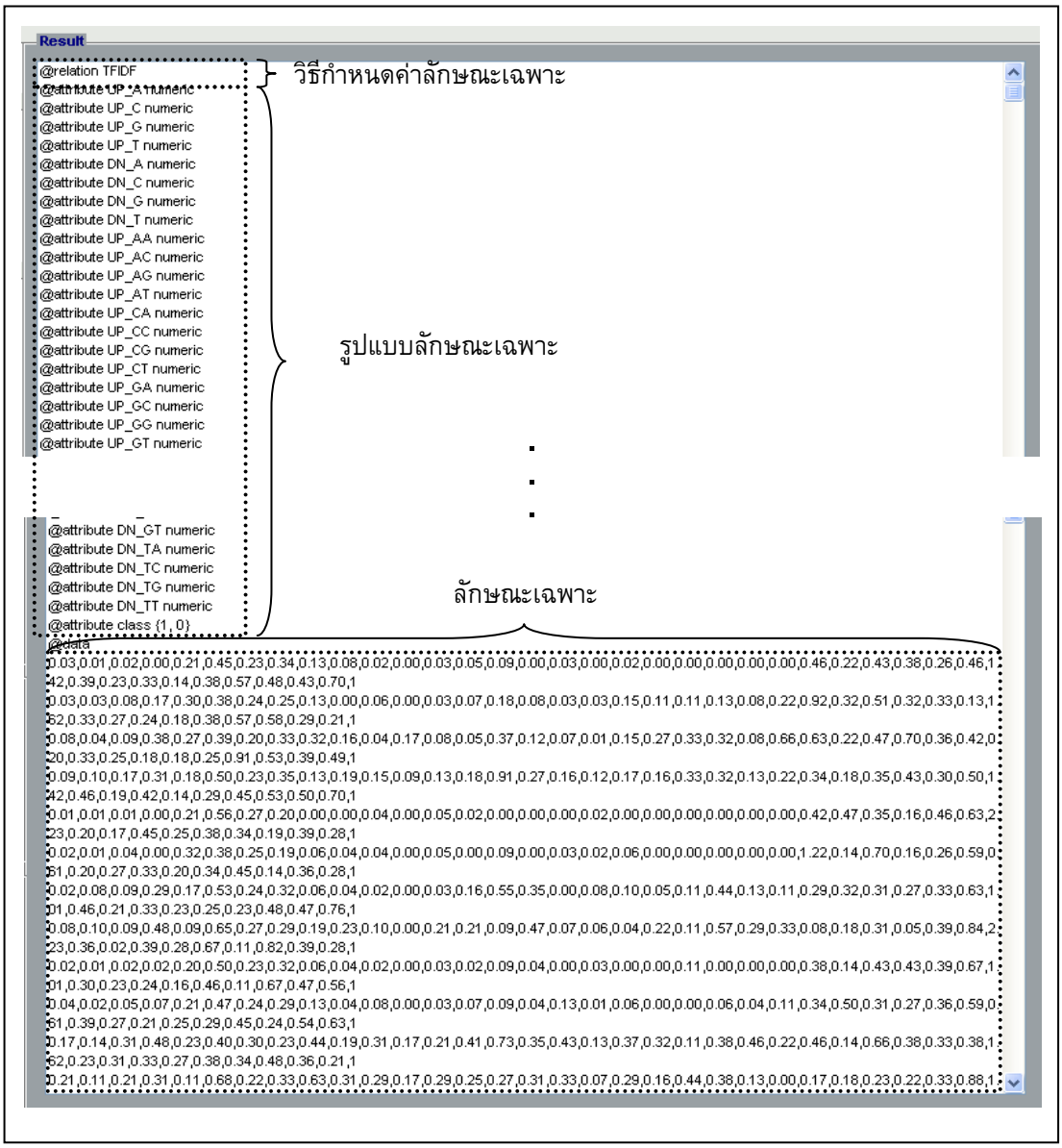
ภาพประกอบ 4.10 รูปแบบสายพันธุกรรม

4.2.2 ส่วนการสร้างลักษณะเฉพาะ n-แกรม

เป็นส่วนที่ประกอบด้วยการแบ่งสายพันธุกรรมเป็นสายพันธุกรรมย่อย โดยผู้ใช้สามารถระบุขนาดหน้าต่างที่ต้องการสำหรับการแบ่งสายพันธุกรรม และสามารถเลือกลักษณะเฉพาะ n-แกรม ได้โดยค่า n ที่เป็นไปได้ทั้งหมดเท่ากับ 7 ได้แก่ 1) 1-แกรม 2) 2-แกรม 3) 3-แกรม 4) 1-แกรมและ 2-แกรม 5) 1-แกรมและ 3-แกรม 6) 2-แกรมและ 3-แกรม และ 7) 1-แกรม 2-แกรม และ 3-แกรม สำหรับการกำหนดค่าลักษณะเฉพาะมี 2 วิธี คือ ค่าความถี่และค่า TF-IDF ตัวอย่างเช่น แบ่งสายพันธุกรรมด้วยขนาดหน้าต่างเท่ากับ 303 นิวคลีโอไดท์และสร้างลักษณะเฉพาะ 1-แกรม และ 2-แกรม ทั้งในส่วนอัพสตรีมและดาวน์สตรีมของสายพันธุกรรมย่อยแยกจากกัน กำหนดค่าลักษณะเฉพาะด้วยค่า TF-IDF ดังภาพประกอบ 4.11 จากนั้นกดปุ่ม “Apply Feature Generation” ผลการทำงานในส่วนนี้จะแสดงวิธีกำหนดค่าลักษณะเฉพาะ รูปแบบลักษณะเฉพาะและค่าลักษณะเฉพาะของตัวอย่างทั้งกลุ่มบวกและกลุ่มลบ แสดงดังภาพประกอบ 4.12



ภาพประกอบ 4.11 ส่วนการสร้างลักษณะเฉพาะ

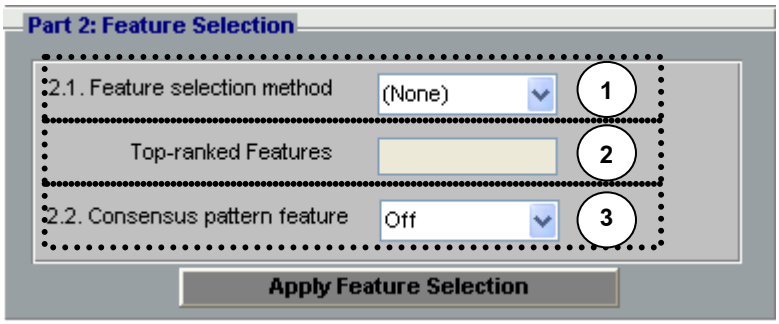


ภาพประกอบ 4.12 รูปแบบและค่าของลักษณะเฉพาะ

4.2.3 ส่วนการเลือกลักษณะเฉพาะ

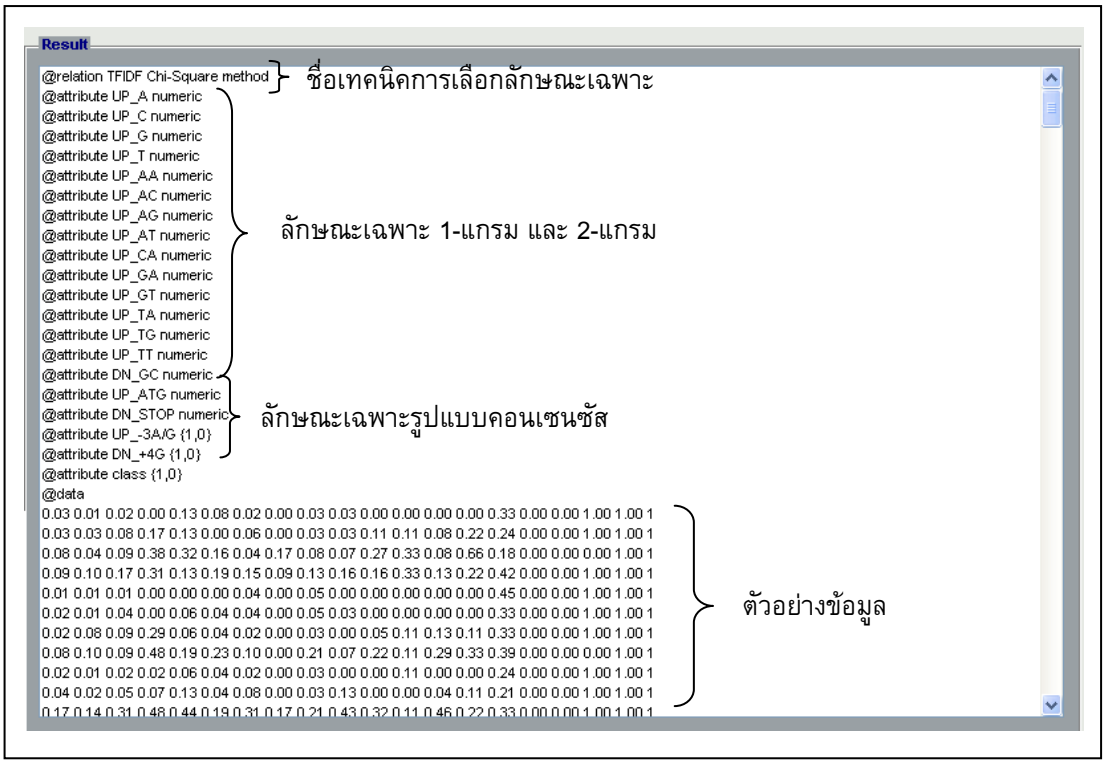
เป็นส่วนกำหนดข้อมูลเข้าของโครงข่ายประสาทเทียม ประกอบด้วย 3 ส่วน คือ 1) ผู้ใช้เลือกลักษณะเฉพาะ 1-แกรม และ 2-แกรม ด้วยเทคนิคการเลือกลักษณะเฉพาะ 2) ผู้ใช้กำหนดจำนวนลักษณะเฉพาะ และ 3) ผู้ใช้เลือกลักษณะเฉพาะรูปแบบคอนเซนซัส แสดงดังภาพประกอบ 4.13 สำหรับเทคนิคการเลือกลักษณะเฉพาะมี 4 ตัวเลือกได้แก่ None หมายถึง ลักษณะเฉพาะทั้งหมด CFS หมายถึง เทคนิคการเลือกลักษณะเฉพาะสหสัมพันธ์ Chi-Square หมายถึง เทคนิคไคส์แควร์ Gain Ratio หมายถึง เทคนิคอัตราส่วนเกน และ ReleifF หมายถึง

เทคนิคริลีฟ โดยเทคนิคไคส์แควร์ เทคนิคอัตราส่วนเกน และเทคนิคริลีฟ เป็นเทคนิคที่ผู้ใช้ต้องกำหนดจำนวนลักษณะเฉพาะ สำหรับลักษณะเฉพาะรูปแบบคอนเซนซัสมี 2 ตัวเลือกได้แก่ OFF หมายถึง ไม่เลือก และ ON หมายถึง เลือก



ภาพประกอบ 4.13 ส่วนการเลือกลักษณะเฉพาะ

เมื่อผู้ใช้กดปุ่ม “Apply Feature Selection” โปรแกรมจะแสดงข้อมูลเข้าของโครงข่ายประสาทเทียม รูปแบบข้อมูลประกอบด้วยชื่อเทคนิคการเลือกลักษณะเฉพาะ ลักษณะเฉพาะ และตัวอย่างข้อมูลแสดงดังภาพประกอบ 4.14

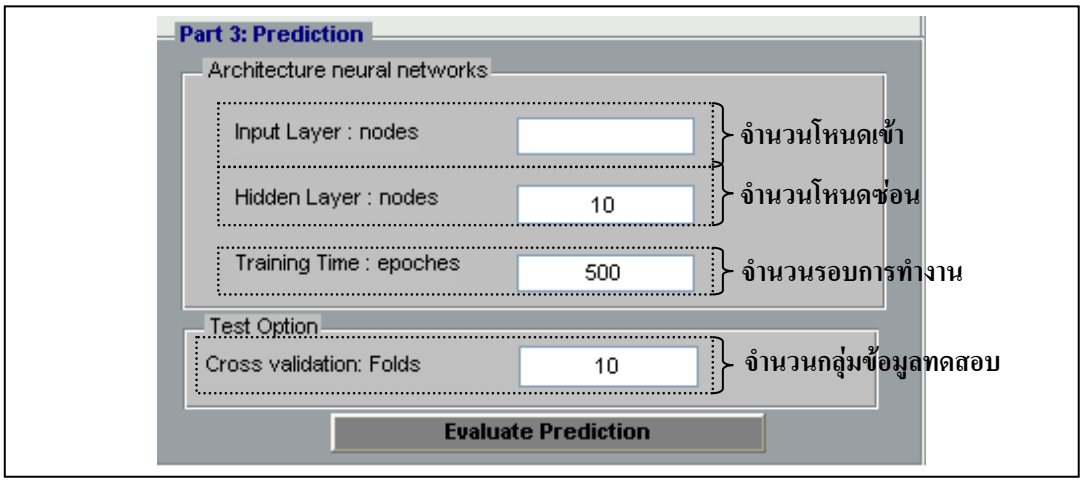


ภาพประกอบ 4.14 ข้อมูลเข้าของโครงข่ายประสาทเทียม

4.2.4 ส่วนการทำนายผลลัพธ์

เป็นการทำนายจุดเริ่มต้นการแปลรหัสด้วยโครงข่ายประสาทเทียม โดยส่วนการทำนายผลลัพธ์ประกอบด้วย 1) การกำหนดสถาปัตยกรรมโครงข่ายประสาทเทียมแบบหน่วยประมวลผลย่อยหลายชั้น และ 2) การทดสอบไขว้เปลี่ยน k กลุ่ม

ส่วนที่ 1 ประกอบด้วยการแสดงจำนวนโหนดชั้นข้อมูลเข้า (Input Nodes) จำนวนโหนดชั้นซ่อน (Hidden Nodes) และจำนวนรอบในการทำงาน (Epoch) โดยข้อมูลจำนวนโหนดชั้นซ่อน และจำนวนรอบการทำงานต้องเป็นจำนวนเต็มบวกเท่านั้น สำหรับการประเมินประสิทธิภาพโครงข่ายประสาทเทียมผู้ใช้สามารถกำหนดจำนวนกลุ่ม (k-Folds) ของการทดสอบไขว้เปลี่ยนได้ แสดงดังภาพประกอบ 4.15



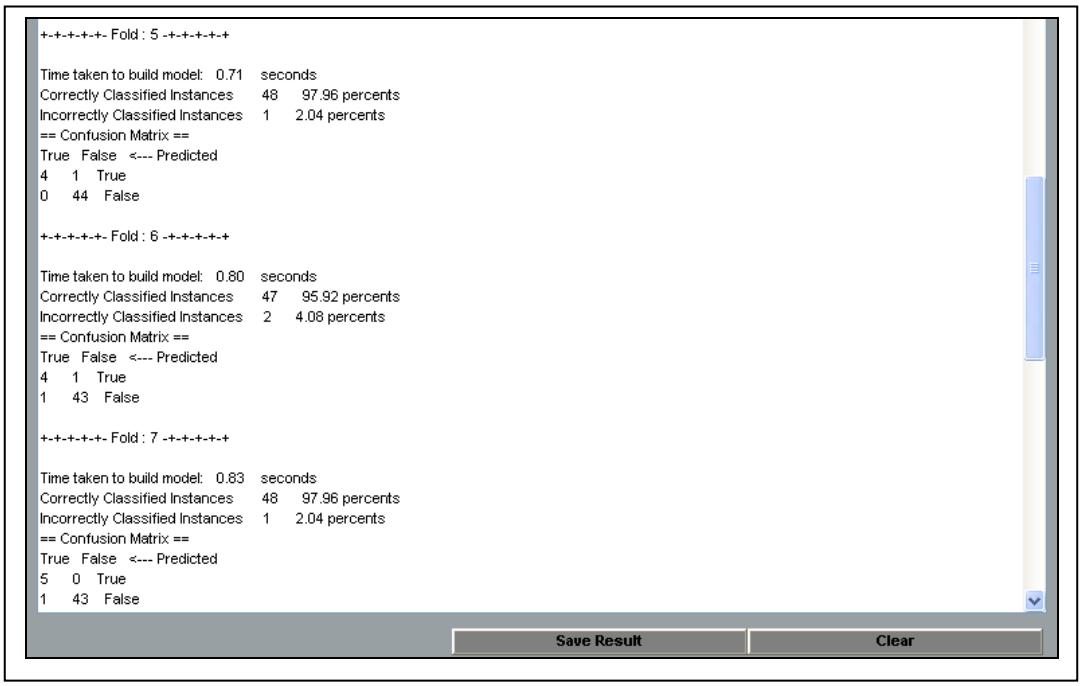
ภาพประกอบ 4.15 ส่วนการทำนายผลลัพธ์

เมื่อกดปุ่ม “Evaluate Prediction” โปรแกรมจะแสดงผลการทำนายของแต่ละกลุ่มข้อมูลทดสอบ แสดงดังภาพประกอบ 4.16 โดยผลการทำนายประกอบด้วยเวลาการสร้างแบบจำลอง จำนวนตัวอย่างที่ทำนายถูก จำนวนตัวอย่างที่ทำนายผิด และตารางคอนฟิวชั่นเมตริกซ์

จากภาพประกอบ 4.16 แสดงข้อมูลทดสอบกลุ่มที่ 5 ถึง 7 ตัวอย่างเช่น ข้อมูลทดสอบกลุ่มที่ 5 ใช้เวลาการสร้างแบบจำลองเท่ากับ 0.71 วินาที ตัวอย่างที่จัดกลุ่มถูกเท่ากับ 48 ตัวอย่าง คิดเป็น 97.96% และตัวอย่างที่จัดกลุ่มผิดเท่ากับ 1 ตัวอย่าง คิดเป็น 2.04%

4.2.5 ส่วนการแสดงผลการทำงาน

เป็นส่วนแสดงผลการทำงานจากแต่ละส่วนของโปรแกรม โดยผู้ใช้สามารถบันทึกผลลัพธ์เป็นไฟล์นามสกุล .mat ด้วยการกดปุ่ม “Save Result”



ภาพประกอบ 4.16 ผลลัพธ์การทำนายของโครงข่ายประสาทเทียม

บทที่ 5

ผลการทดลองและบทวิจารณ์

บทนี้จะนำเสนอผลลัพธ์ที่ได้จากการทดลองตามแบบจำลองการทำนายจุดเริ่มต้นการแปลรหัสโดยใช้วิธี TF-IDF และโครงข่ายประสาทเทียม (The TF-IDF and Neural Networks Approach for Translation Initiation Sites Prediction: TFIDF-NN-TIS) ในการทดลองใช้ข้อมูล 3 ชุดข้อมูล คือ ข้อมูล Vertebrate ข้อมูล Arabidopsis thaliana และข้อมูล TIS+50

5.1 ชุดข้อมูลที่ใช้ในการทดลอง

ข้อมูลที่ใช้ในการทดลองประกอบด้วย 3 ชุดข้อมูล คือ ชุดข้อมูล Vertebrate ชุดข้อมูล Arabidopsis thaliana และชุดข้อมูล TIS+50

5.1.1 ชุดข้อมูล Vertebrate (Pedersen and Nielsen, 1997) เป็นชุดข้อมูลของที่รวบรวมจากเนื้อเยื่อของสัตว์มีกระดูกสันหลังโดยแต่ละสายพันธุกรรมมีจุดเริ่มต้นการแปลรหัสเพียง 1 ตำแหน่ง สกัดชุดข้อมูลจากฐานข้อมูล GenBank จุดเริ่มต้นการแปลรหัสจะมีจำนวนนิวคลีโอไทด์ในอัพสตรีมอย่างน้อย 10 นิวคลีโอไทด์ และมีจำนวนนิวคลีโอไทด์ในดาวน์สตรีมอย่างน้อย 150 นิวคลีโอไทด์ ทั้งนี้ทุกสายพันธุกรรมจะผ่านกระบวนการกำจัดส่วนที่ไม่มีรหัส (Non-Coding Region) และกระบวนการเชื่อมต่อส่วนที่มีรหัส (Coding Region) เข้าด้วยกัน

```
>206 BBCALCB.1 CAT X71666 ;TIS=57
CCGTCAGAGCGCCGACACTCTTCTCTGTGCGAGCGAGCCGCCGACCGCCAAGCAAAATG
GGAAATGAGGCAAGTTATCCTTTGGAAATGTGCTCACACTTTGATGCAGATGAAATTAAAAG
GCTAGGAAAGAGATTTAAGAAGCTCGATTTGGACAATTCTGGTTCTTTGAGTGTGGAAGAG
TTCATGTCTCTACCTGAGTTACAA
```

ภาพประกอบ 5.1 ตัวอย่างสายพันธุกรรมของชุดข้อมูล Vertebrate

5.1.2 ชุดข้อมูล Arabidopsis thaliana หรือ A.thaliana (Pedersen and Nielsen, 1997) เป็นชุดข้อมูลพืชโดยแต่ละสายพันธุกรรมมีจุดเริ่มต้นการแปลรหัสเพียง 1 ตำแหน่ง สกัดชุดข้อมูลจากฐานข้อมูล GenBank จุดเริ่มต้นการแปลรหัสจะมีจำนวนนิวคลีโอไทด์ในอัพสตรีมอย่างน้อย 10 นิวคลีโอไทด์ และมีจำนวนนิวคลีโอไทด์ในดาวน์สตรีมอย่างน้อย 150 นิวคลีโอไทด์ ทั้งนี้ทุกสายพันธุกรรมจะผ่านกระบวนการกำจัดส่วนที่ไม่มีรหัส (Non-Coding Region) และกระบวนการเชื่อมต่อส่วนที่มีรหัส (Coding Region) เข้าด้วยกัน

```
>167 AT2A6.1 CAT X83096 Eukaryotae; …. Arabidopsis. Arabidopsis thaliana;TIS=18
CACGCGTCCGAAGCAAGATGGAGTCAAGTGATCGTTCAAGTCAAGCAAAAGCTTTCGACGA
GACAAAAACCGGCGTGAAAGGGCTTGTGGCTTCGGGAATCAAAGAGATTCCAGCCATGTT
CCATACACCTCCGGATACTCTAACAAGCCTGAAACAAACAGCACCA
```

ภาพประกอบ 5.2 ตัวอย่างสายพันธุกรรมของชุดข้อมูล Arabidopsis thaliana

5.1.3 ชุดข้อมูล TIS+50 (Nadershahi et *al.*, 2004) เป็นสายดีเอ็นเอสั้นๆ ประกอบด้วยจำนวน 50 สายดีเอ็นเอ โดยทุกสายดีเอ็นเอมี Open Reading Frame หรือ ORF นั้นคือ แต่ละสายมีส่วนซึ่งเป็นช่วงจากจุดเริ่มต้นการแปลรหัส (TIS) ไปยังโคดอนหยุด ได้แก่ โคดอน TAA TGA หรือ TAG แสดงดังภาพประกอบ 5.3 แต่ละสายพันธุกรรมมีจุดเริ่มต้นการแปลรหัสเพียง 1 ตำแหน่ง

```
>gi|2010512|gb|AA358121.1| EST67041 Fetal …. mRNA sequence; TIS=36
TTGGTTGGGGGCCGTCCCGCTCCTAAGGCAGGAAG**ATG**GTGGCCGCAAAGAAGACGAAA
AAGTCGCTGGAGTCGATCAACTCTAGGCTCCAACTCGTTATGAAAAGTGGGAAGTACGTCC
TGGGGTACAAGCAGACTCTGAAGATGATCAGACAAGGCAAAGCGAAATTGGTCATTCTCGC
TAACAACTGCCCAGCTTTGAGGAAATCTGAAA**TAG**AGTACTATGCTATGTTGGCTAAAA
CTGGTGTCCATCACTACAGTGGCAATAATATTGAACTGGGCACAGCATGCGGNAAATACTA
CAGAGTGTGCACACTGGCTATCATTG
```

ภาพประกอบ 5.3 ตัวอย่างสายพันธุกรรมของชุดข้อมูล TIS+50

5.2 การทดลอง

5.2.1 การออกแบบการทดลอง

ขั้นตอนของแบบจำลองการทำนายจุดเริ่มต้นการแปลรหัสโดยใช้วิธี TF-IDF และโครงข่ายประสาทเทียม (TF-IDF-NN-TIS) มี 5 ขั้นตอน คือ ขั้นตอนที่ 1 การแบ่งสายพันธุกรรม ขั้นตอนที่ 2 การสร้างลักษณะเฉพาะ n-แกรม ขั้นตอนที่ 3 การเลือกลักษณะเฉพาะ ขั้นตอนที่ 4 การสร้างลักษณะเฉพาะรูปแบบคอนเซนซัส และขั้นตอนที่ 5 การทำนายผลลัพธ์ ดังภาพประกอบ 3.1 มีการทดลองทั้งหมด 4 แบบ คือ การทดลอง A B C และ D การทดลองแต่ละแบบมีขั้นตอนการทดลองแตกต่างกันแสดงดังตารางที่ 5.1 โดยที่สัญลักษณ์ ✓ หมายถึง ทำขั้นตอนนั้น และสัญลักษณ์ ✕ หมายถึงไม่ทำขั้นตอนนั้น

ตารางที่ 5.1 ขั้นตอนการทดลองของการทดลองแต่ละแบบ

การทดลอง	ขั้นตอนที่ 1 การแบ่งสายพันธุกรรม	ขั้นตอนที่ 2 การสร้างลักษณะเฉพาะ n-แกรม (M)	ขั้นตอนที่ 3 การเลือกลักษณะเฉพาะ (R)	ขั้นตอนที่ 4 การสร้างลักษณะเฉพาะรูปแบบคอนเซนซัส (CP)	ขั้นตอนที่ 5 การทำนายผลลัพธ์
A	✓	✓	✕	✕	✓
B	✓	✓	✓	✕	✓
C	✓	✓	✕	✓	✓
D	✓	✓	✓	✓	✓

การทำนายจุดเริ่มต้นการแปลรหัสใช้โครงข่ายประสาทเทียมแบบไปข้างหน้าหลายชั้น (Feedforward Multilayer Neural Networks) ในขั้นตอนการสอนโครงข่ายนี้จะใช้วิธีการส่งค่าย้อนกลับ (Backpropagation Algorithm) โดยใช้ฟังก์ชันกระตุ้นซิกมอยด์ (Sigmoid Activation Function) เป็นฟังก์ชันการแปลงค่า (Transfer Function) เนื่องจากฟังก์ชันการแปลงค่ามีความสำคัญมากในโครงข่ายที่สอนด้วยวิธีการส่งค่าย้อนกลับ ซึ่งฟังก์ชันที่ใช้ควรมีความต่อเนื่อง ไม่เป็นเชิงเส้น สามารถหาค่าอนุพันธ์ได้ และง่ายต่อการคำนวณ กำหนดสถาปัตยกรรมของโครงข่ายประสาทเทียมเท่ากับ 3 ชั้น ได้แก่ชั้นข้อมูลเข้า ชั้นซ่อน และชั้นผลลัพธ์ โดยชั้นข้อมูลเข้ามีจำนวนโหนดเท่ากับจำนวนลักษณะเฉพาะที่ต้องการ จำนวนข้อมูลชั้นซ่อน (แทนด้วย H) มีจำนวน 10 โหนด และจำนวนข้อมูลชั้นแสดงผลลัพธ์ (แทนด้วย O) มีจำนวน 1 โหนด ประเมินผลลัพธ์ของโครงข่ายประสาทเทียมด้วยวิธีการทดสอบไขว้เปลี่ยนแบบ k กลุ่ม

การทดลองแต่ละแบบมีสถาปัตยกรรมโครงข่ายประสาทเทียมแบบหน่วยประมวลผลย่อยหลายชั้นในชั้นข้อมูลเข้าแตกต่างกันดังตารางที่ 5.2 โดยการทดลองแบบ A มีข้อมูลเข้าเป็นลักษณะเฉพาะ n-แกรม การทดลองแบบ B มีข้อมูลเข้าเป็นลักษณะเฉพาะ n-แกรม ที่ผ่านการเลือกด้วยเทคนิคการเลือกลักษณะเฉพาะ ได้แก่ เทคนิคการเลือกลักษณะเฉพาะสหสัมพันธ์ (CFS) เทคนิคไคสแควร์ เทคนิคอัตราส่วนเกน และเทคนิครีลีฟ-เอฟ การทดลองแบบ C มีข้อมูลเข้าเป็นลักษณะเฉพาะ n-แกรม ร่วมกับลักษณะเฉพาะที่สร้างจากรูปแบบคอนเซนซัส และการทดลองแบบ D มีข้อมูลเข้าเป็นลักษณะเฉพาะ n-แกรม ที่ผ่านการเลือกด้วยเทคนิคการเลือกลักษณะเฉพาะ และลักษณะเฉพาะที่สร้างจากรูปแบบคอนเซนซัส จากตารางที่ 5.2 กำหนดให้ M คือ ลักษณะเฉพาะ n-แกรม R คือ ลักษณะเฉพาะของ n-แกรม ที่ผ่านการเลือกด้วยเทคนิคการเลือกลักษณะเฉพาะ CP คือ ลักษณะเฉพาะของรูปแบบคอนเซนซัส H คือชั้นซ่อนมีจำนวนโหนดเท่ากับ 10 โหนด และ O คือ ชั้นผลลัพธ์มีจำนวนโหนดเท่ากับ 1 โหนด

ตารางที่ 5.2 การออกแบบการทดลอง และสถาปัตยกรรมของโครงข่ายประสาทเทียม

การทดลอง	ลักษณะเฉพาะที่ผ่านการเลือก (R)	ลักษณะเฉพาะรูปแบบคอนเซนซัส (CP)	สถาปัตยกรรมโครงข่ายประสาทเทียม
A	✗	✗	$M : H : O$
B	✓	✗	$R : H : O$
C	✗	✓	$(M+CP) : H : O$
D	✓	✓	$(R+CP) : H : O$

5.2.2 ผลการทดลอง

ขั้นตอนที่ 1 การแบ่งสายพันธุกรรม งานวิจัยนี้ใช้ชุดข้อมูล 3 ชุดข้อมูล ได้แก่ ชุดข้อมูล Vertebrate ชุดข้อมูล Arabidopsis thaliana และชุดข้อมูล TIS+50 แต่ละชุดข้อมูลมีคุณลักษณะแสดงดังตารางที่ 5.3 แต่ละสายพันธุกรรมมีจุดเริ่มต้นการแปลรหัสเพียง 1 ตำแหน่ง

ชุดข้อมูล Vertebrate ประกอบด้วยสายพันธุกรรมทั้งหมดเท่ากับ 3,312 สายพันธุกรรม มีโคดอน ATG ที่พิจารณาทั้งหมดเท่ากับ 13,503 ตำแหน่ง จากนั้นแบ่งสายพันธุกรรมเป็นสายพันธุกรรมย่อยด้วยขนาดหน้าต่างที่ต้องการ จะได้ตัวอย่างทั้งหมดเท่ากับ 13,503 ตัวอย่างแบ่งเป็นตัวอย่างกลุ่มบวกจำนวน 3,312 ตัวอย่าง คิดเป็น 24.5% และตัวอย่างกลุ่มลบจำนวน 10,191 ตัวอย่าง คิดเป็น 75% ดังนั้นอัตราส่วนระหว่างข้อมูลกลุ่มบวกต่อข้อมูลกลุ่มลบ คือ 1 ต่อ 3

ตารางที่ 5.3 คุณลักษณะของชุดข้อมูล

ชื่อ ชุดข้อมูล	ความยาวสาย พันธุกรรม สั้นสุด	ความยาวสาย พันธุกรรม ยาวสุด	ความยาวสาย พันธุกรรม ส่วนใหญ่	จำนวน กลุ่มบวก	จำนวน กลุ่มลบ	อัตราส่วน กลุ่มบวก ต่อ กลุ่มลบ
Vertebrate	169	299	299	3,312	10,191	1/3
A.thaliana	169	299	299	523	1,525	1/3
TIS+50	197	1,112	469	50	469	1/9

ชุดข้อมูล A.thaliana ประกอบด้วยสายพันธุกรรมทั้งหมด 523 สายพันธุกรรม มีโคดอน ATG ที่พิจารณาทั้งหมดเท่ากับ 2,048 ตำแหน่ง จากนั้นแบ่งสายพันธุกรรมเป็นสายพันธุกรรมย่อยด้วยขนาดหน้าต่างที่ต้องการ จะได้ตัวอย่างกลุ่มบวกจำนวน 523 ตัวอย่าง คิดเป็น 25.5% และตัวอย่างกลุ่มลบ จำนวน 1,525 ตัวอย่าง คิดเป็น 74.5% ดังนั้นอัตราส่วนระหว่างข้อมูลกลุ่มบวกต่อข้อมูลกลุ่มลบ คือ 1 ต่อ 3

ชุดข้อมูล TIS+50 ประกอบด้วยสายพันธุกรรมทั้งหมดเท่ากับ 50 สายพันธุกรรม มีโคดอน ATG ที่พิจารณาทั้งหมดเท่ากับ 519 ตำแหน่ง จากนั้นแบ่งสายพันธุกรรมเป็นสายพันธุกรรมย่อยด้วยขนาดหน้าต่างที่ต้องการจะได้ตัวอย่างกลุ่มบวกจำนวน 50 ตัวอย่าง คิดเป็น 9.6% และตัวอย่างกลุ่มลบจำนวน 469 ตัวอย่าง คิดเป็น 90.4% ดังนั้นอัตราส่วนระหว่างข้อมูลกลุ่มบวกต่อข้อมูลกลุ่มลบ คือ 1 ต่อ 9

ขั้นตอนที่ 2 การสร้างลักษณะเฉพาะ n-แกรม ลักษณะเฉพาะที่ได้จากเทคนิค 1-แกรม 2-แกรม หรือ 3-แกรม ทั้งในส่วนอัพสตรีมและดาวน์สตรีมแสดงดังตารางที่ 5.4

ตารางที่ 5.4 ลักษณะเฉพาะ 1-แกรม 2-แกรม หรือ 3-แกรมสำหรับการทดลอง

ลักษณะเฉพาะ	คำอธิบาย
UP_A UP_C UP_G UP_T	ลักษณะเฉพาะที่ได้จาก 1-แกรม ในส่วนอัพสตรีม
DN_A DN_C DN_G DN_T	ลักษณะเฉพาะที่ได้จาก 1-แกรม ในส่วนดาวน์สตรีม
UP_AA UP_AG UP_AC UP_AT UP_GA UP_GG UP_GC UP_GT UP_CA UP_CG UP_CC UP_CT UP_TA UP_TG UP_TC UP_TT	ลักษณะเฉพาะที่ได้จาก 2-แกรม ในส่วนอัพสตรีม
DN_AA DN_AG DN_AC DN_AT DN_GA DN_GG DN_GC DN_GT DN_CA DN_CG DN_CC DN_CT DN_TA DN_TG DN_TC DN_TT	ลักษณะเฉพาะที่ได้จาก 2-แกรม ในส่วนดาวน์สตรีม

ตารางที่ 5.4 ลักษณะเฉพาะ 1-แกรม 2-แกรม หรือ 3-แกรมสำหรับการทดลอง (ต่อ)

ลักษณะเฉพาะ	คำอธิบาย
UP_AAA UP_AAC UP_AAG UP_AAT UP_ACA UP_ACC UP_ACG UP_ACT UP_AGA UP_AGC UP_AGG UP_AGT UP_ATA UP_ATC UP_ATG UP_ATT UP_CAA UP_CAC UP_CAG UP_CAT UP_CCA UP_CCC UP_CCG UP_CCT UP_CGA UP_CGC UP_CGG UP_CGT UP_CTA UP_CTC UP_CTG UP_CTT UP_GAA UP_GAC UP_GAG UP_GAT UP_GCA UP_GCC UP_GCG UP_GCT UP_GGA UP_GGC UP_GGG UP_GGT UP_GTA UP_GTC UP_GTG UP_GTT UP_TAA UP_TAC UP_TAG UP_TAT UP_TCA UP_TCC UP_TCG UP_TCT UP_TGA UP_TGC UP_TGG UP_TGT UP_TTA UP_TTC UP_TTG UP_TTT	ลักษณะเฉพาะที่ได้จาก 3-แกรมในส่วนอัพสตรีม
DN_AAA DN_AAC DN_AAG DN_AAT DN_ACA DN_ACC DN_ACG DN_ACT DN_AGA DN_AGC DN_AGG DN_AGT DN_ATA DN_ATC DN_ATG DN_ATT DN_CAA DN_CAC DN_CAG DN_CAT DN_CCA DN_CCC DN_CCG DN_CCT DN_CGA DN_CGC DN_CGG DN_CGT DN_CTA DN_CTC DN_CTG DN_CTT DN_GAA DN_GAC DN_GAG DN_GAT DN_GCA DN_GCC DN_GCG DN_GCT DN_GGA DN_GGC DN_GGG DN_GGT DN_GTA DN_GTC DN_GTG DN_GTT DN_TAA DN_TAC DN_TAG DN_TAT DN_TCA DN_TCC DN_TCG DN_TCT DN_TGA DN_TGC DN_TGG DN_TGT DN_TTA DN_TTC DN_TTG DN_TTT	ลักษณะเฉพาะที่ได้จาก 3-แกรมในส่วนดาวน์สตรีม

สำหรับแต่ละลักษณะเฉพาะของตัวอย่างกำหนดค่าลักษณะเฉพาะด้วยค่าความถี่ และค่า TF-IDF โดยค่าความถี่เป็นการนับจำนวนลักษณะเฉพาะที่ปรากฏในตัวอย่าง ซึ่งค่าลักษณะเฉพาะที่ได้จะอยู่ในช่วงของจำนวนเต็มศูนย์ และจำนวนเต็มบวก สำหรับค่า TF-IDF ค่าที่ได้จะเป็นจำนวนจริงที่มีค่าตั้งแต่ 0 เป็นต้นไป ตัวอย่างเช่นชุดข้อมูล Vertebrate

แบ่งสายพันธุกรรมด้วยขนาดหน้าต่าง 303 นิวคลีโอไทด์ สร้างลักษณะเฉพาะ 1-แกรม และ 2-แกรม มีจำนวนลักษณะเฉพาะทั้งหมดเท่ากับ 40 พบว่า ค่าความถี่ของตัวอย่างบวกที่ 1 สำหรับลักษณะเฉพาะ UP_A UP_C UP_G UP_T DN_A DN_C DN_G และ DN_T มีค่าเท่ากับ 12 22 15 7 46 20 37 และ 44 ตามลำดับ แสดงดังตารางที่ 5.5 และค่า TF-IDF ของตัวอย่างบวกที่ 1 สำหรับลักษณะเฉพาะ UP_A UP_C UP_G UP_T DN_A DN_C DN_G และ DN_T มีค่าเท่ากับ 0.14 0.25 0.16 0.12 1.14 0.54 0.84 และ 1.24 ตามลำดับ แสดงดังตารางที่ 5.6

ตารางที่ 5.5 ค่าความถี่ตัวอย่างกลุ่มบวกที่ 1 ถึง 10 ของชุดข้อมูล Vertebrate

ลักษณะเฉพาะ	ค่าความถี่									
	1	2	3	4	5	6	7	8	9	10
UP_A	12	11	2	3	45	48	55	10	6	15
UP_C	22	40	27	6	23	21	16	8	6	7
UP_G	15	30	19	5	41	35	34	10	6	10
UP_T	7	21	6	5	40	45	44	2	6	8
DN_A	46	43	27	35	39	30	63	45	40	29
DN_C	20	29	41	26	35	44	25	35	40	46
DN_G	37	45	49	37	40	42	39	26	37	33
DN_T	44	30	30	49	33	31	20	41	30	39
UP_AA	3	0	0	2	20	16	19	2	2	7
UP_AC	3	8	0	0	6	3	7	2	0	3
UP_AG	5	3	2	0	11	13	15	4	3	3
UP_AT	0	0	0	0	8	16	13	1	1	1
UP_CA	4	3	2	0	6	11	5	4	2	3
UP_CC	6	18	13	2	2	3	3	1	1	0
UP_CG	8	8	9	2	9	2	2	3	0	1
UP_CT	4	10	2	2	6	4	6	0	3	3
UP_GA	5	7	0	1	8	11	18	3	2	4
UP_GC	8	6	10	1	10	9	2	4	2	1
UP_GG	0	12	5	2	12	7	5	3	1	2
UP_GT	2	5	4	1	10	8	9	0	0	3
UP_TA	0	1	0	0	10	10	12	0	0	1
UP_TC	4	8	3	3	5	5	4	1	2	3
UP_TG	2	6	3	1	9	13	12	0	2	3
UP_TT	1	6	0	1	16	17	16	1	2	1
DN_AA	17	19	5	11	11	10	25	15	10	3

ตารางที่ 5.5 ค่าความถี่ตัวอย่างกลุ่มบวกที่ 1 ถึง 10 ของชุดข้อมูล Vertebrate (ต่อ)

ลักษณะเฉพาะ	ค่าความถี่									
	1	2	3	4	5	6	7	8	9	10
DN_AC	5	7	11	4	9	8	4	8	11	10
DN_AG	13	9	10	10	10	9	25	11	12	8
DN_AT	10	8	1	10	8	3	8	11	7	8
DN_CA	7	6	8	5	13	10	14	13	12	10
DN_CC	2	5	8	3	9	15	4	10	10	17
DN_CG	1	5	9	5	6	4	2	2	5	1
DN_CT	10	13	16	13	7	14	5	10	13	17
DN_GA	16	13	11	10	14	6	21	8	15	8
DN_GC	5	10	15	8	8	13	10	6	9	9
DN_GG	8	16	16	8	13	15	5	3	9	10
DN_GT	8	6	6	11	5	8	3	9	3	6
DN_TA	6	5	3	9	1	4	3	9	2	8
DN_TC	8	7	7	11	9	8	7	11	10	10
DN_TG	14	14	14	13	10	13	6	10	11	13
DN_TT	16	3	6	15	13	6	4	10	7	8

ตารางที่ 5.6 ค่า TF-IDF ตัวอย่างกลุ่มบวกที่ 1 ถึง 10 ของชุดข้อมูล Vertebrate

ลักษณะเฉพาะ	ค่า TF-IDF									
	1	2	3	4	5	6	7	8	9	10
UP_A	0.14	0.13	0.02	0.04	0.54	0.58	0.66	0.12	0.07	0.18
UP_C	0.25	0.46	0.31	0.07	0.27	0.24	0.19	0.09	0.07	0.08
UP_G	0.16	0.32	0.20	0.05	0.44	0.37	0.36	0.11	0.06	0.11
UP_T	0.12	0.37	0.11	0.09	0.71	0.80	0.78	0.04	0.11	0.14
DN_A	1.14	1.06	0.67	0.86	0.96	0.74	1.55	1.11	0.99	0.72
DN_C	0.54	0.78	1.10	0.70	0.94	1.18	0.67	0.94	1.07	1.24
DN_G	0.84	1.03	1.12	0.84	0.91	0.96	0.89	0.59	0.84	0.75
DN_T	1.24	0.85	0.85	1.39	0.93	0.88	0.57	1.16	0.85	1.10
UP_AA	0.35	0.00	0.00	0.23	2.33	1.86	2.21	0.23	0.23	0.81
UP_AC	0.26	0.71	0.00	0.00	0.53	0.26	0.62	0.18	0.00	0.26
UP_AG	0.24	0.15	0.10	0.00	0.53	0.63	0.73	0.19	0.15	0.15
UP_AT	0.00	0.00	0.00	0.00	1.08	2.17	1.76	0.14	0.14	0.14
UP_CA	0.22	0.17	0.11	0.00	0.34	0.61	0.28	0.22	0.11	0.17
UP_CC	0.44	1.31	0.94	0.15	0.15	0.22	0.22	0.07	0.07	0.00

ตารางที่ 5.6 ค่า TF-IDF ตัวอย่างกลุ่มบวกที่ 1 ถึง 10 ของชุดข้อมูล Vertebrate (ต่อ)

ลักษณะเฉพาะ	ค่า TF-IDF									
	1	2	3	4	5	6	7	8	9	10
UP_CG	1.91	1.91	2.15	0.48	2.15	0.48	0.48	0.72	0.00	0.24
UP_CT	0.26	0.65	0.13	0.13	0.39	0.26	0.39	0.00	0.20	0.20
UP_GA	0.30	0.41	0.00	0.06	0.47	0.65	1.07	0.18	0.12	0.24
UP_GC	0.48	0.36	0.60	0.06	0.60	0.54	0.12	0.24	0.12	0.06
UP_GG	0.00	0.92	0.38	0.15	0.92	0.54	0.38	0.23	0.08	0.15
UP_GT	0.21	0.52	0.41	0.10	1.04	0.83	0.93	0.00	0.00	0.31
UP_TA	0.00	0.22	0.00	0.00	2.24	2.24	2.68	0.00	0.00	0.22
UP_TC	0.28	0.56	0.21	0.21	0.35	0.35	0.28	0.07	0.14	0.21
UP_TG	0.13	0.38	0.19	0.06	0.57	0.82	0.75	0.00	0.13	0.19
UP_TT	0.15	0.90	0.00	0.15	2.39	2.54	2.39	0.15	0.30	0.15
DN_AA	2.17	2.43	0.64	1.40	1.40	1.28	3.19	1.92	1.28	0.38
DN_AC	0.56	0.78	1.23	0.45	1.01	0.89	0.45	0.89	1.23	1.12
DN_AG	1.10	0.76	0.84	0.84	0.84	0.76	2.11	0.93	1.01	0.68
DN_AT	1.24	1.00	0.12	1.24	1.00	0.37	1.00	1.37	0.87	1.00
DN_CA	0.59	0.51	0.68	0.42	1.10	0.85	1.19	1.10	1.02	0.85
DN_CC	0.28	0.71	1.14	0.43	1.28	2.13	0.57	1.42	1.42	2.42
DN_CG	0.36	1.81	3.26	1.81	2.18	1.45	0.73	0.73	1.81	0.36
DN_CT	0.98	1.27	1.56	1.27	0.68	1.37	0.49	0.98	1.27	1.66
DN_GA	1.42	1.16	0.98	0.89	1.25	0.53	1.87	0.71	1.34	0.71
DN_GC	0.59	1.18	1.77	0.95	0.95	1.54	1.18	0.71	1.06	1.06
DN_GG	1.02	2.05	2.05	1.02	1.66	1.92	0.64	0.38	1.15	1.28
DN_GT	1.18	0.89	0.89	1.62	0.74	1.18	0.44	1.33	0.44	0.89
DN_TA	1.44	1.20	0.72	2.16	0.24	0.96	0.72	2.16	0.48	1.92
DN_TC	0.97	0.85	0.85	1.34	1.09	0.97	0.85	1.34	1.21	1.21
DN_TG	1.15	1.15	1.15	1.07	0.82	1.07	0.49	0.82	0.90	1.07
DN_TT	2.81	0.53	1.05	2.63	2.28	1.05	0.70	1.75	1.23	1.40

ขั้นตอนที่ 3 การเลือกลักษณะเฉพาะ ดำเนินการเลือกลักษณะเฉพาะสำคัญที่สามารถระบุจุดเริ่มต้นการแปลรหัสได้ถูกต้องมากขึ้น ด้วยเทคนิคการเลือกลักษณะเฉพาะสหสัมพันธ์ (CFS) เทคนิคไคสแควร์ เทคนิคอัตราส่วนเกน และเทคนิครีลีฟ-เอฟ เปรียบเทียบลักษณะเฉพาะที่ผ่านการเลือกสำหรับค่าความถี่และค่า TF-IDF พบว่าสำหรับชุดข้อมูล Vertebrate ชุดข้อมูล A.thaliana และชุดข้อมูล TIS+50 มีลักษณะเฉพาะที่ผ่านการเลือกสำหรับ

ค่าความถี่ และ ค่า TF-IDF เหมือนกัน ตัวอย่างเช่น แบ่งสายพันธุกรรมด้วยขนาดหน้าต่าง 303 นิวคลีโอไทด์ สร้างลักษณะเฉพาะ 1-แกรม และ 2-แกรม เลือกลักษณะเฉพาะ สำหรับชุดข้อมูล Vertebrate เลือกลักษณะเฉพาะด้วยเทคนิคไคสแควร์ พบว่า ลักษณะเฉพาะที่ผ่านการเลือก 5 ลำดับแรกที่มีนัยสำคัญสูงสุด สำหรับค่าความถี่ คือ DN_G DN_C DN_T DN_GC DN_CT ซึ่งเหมือนกับค่า TF-IDF คือ DN_G DN_C DN_T DN_GC DN_CT แสดงดังตารางที่ 5.7

ตารางที่ 5.7 ลักษณะเฉพาะที่ผ่านการเลือกสำหรับค่าความถี่ และค่า TF-IDF ของชุดข้อมูล Vertebrate

ชื่อเทคนิค การเลือกกษณะเฉพาะ	รูปแบบลักษณะเฉพาะที่ผ่านเลือก	
	ค่าความถี่	ค่า TF-IDF
เทคนิค CFS (ฮิวริสติกเลือกเท่ากับ 14 ลักษณะ)	DN_A DN_C DN_G DN_T UP_AT DN_AC DN_AG DN_CA DN_CG DN_CT DN_GC DN_GG DN_GT DN_TC	DN_A DN_C DN_G DN_T UP_AT DN_AC DN_AG DN_CA DN_CG DN_CT DN_GC DN_GG DN_GT DN_TC
เทคนิคไคสแควร์ (ผู้ใช้กำหนดเท่ากับ 5 ลักษณะ)	DN_G DN_C DN_T DN_GC DN_CT	DN_G DN_C DN_T DN_GC DN_CT
เทคนิคอัตราส่วนเกน (ผู้ใช้กำหนดเท่ากับ 5 ลักษณะ)	DN_G DN_C DN_T DN_A DN_CA	DN_G DN_C DN_T DN_A DN_CA
เทคนิครีลีฟ-เอฟ (ผู้ใช้กำหนดเท่ากับ 5 ลักษณะ)	DN_C DN_A DN_G DN_T DN_AG	DN_C DN_A DN_G DN_T DN_AG

ตัวอย่างชุดข้อมูล A.thaliana ลักษณะเฉพาะที่ผ่านการเลือกด้วยเทคนิค CFS สำหรับค่าความถี่ คือ DN_A DN_C DN_G DN_T DN_CG ซึ่งเหมือนกับค่า TF-IDF คือ DN_A DN_C DN_G DN_T DN_CG แสดงดังตารางที่ 5.8

ตารางที่ 5.8 ลักษณะเฉพาะที่ผ่านการเลือกสำหรับค่าความถี่ และค่า TF-IDF ของชุดข้อมูล A.thaliana

ชื่อเทคนิค การเลือกกษณะเฉพาะ	รูปแบบลักษณะเฉพาะที่ผ่านเลือก	
	ค่าความถี่	ค่า TF-IDF
เทคนิค CFS (ฮิวริสติกเลือกเท่ากับ 5 ลักษณะ)	DN_A DN_C DN_G DN_T DN_CG	DN_A DN_C DN_G DN_T DN_CG
เทคนิคไคสแควร์ (ผู้ใช้กำหนดเท่ากับ 5 ลักษณะ)	DN_C DN_G DN_T DN_A DN_GC	DN_C DN_G DN_T DN_A DN_GC

ตารางที่ 5.8 ลักษณะเฉพาะที่ผ่านการเลือกสำหรับค่าความถี่ และค่า TF-IDF ของชุดข้อมูล A.thaliana (ต่อ)

ชื่อเทคนิค การเลือกกษณะเฉพาะ	รูปแบบลักษณะเฉพาะที่ผ่านเลือก	
	ค่าความถี่	ค่า TF-IDF
เทคนิคอัตราส่วนเกน (ผู้ใช้กำหนดเท่ากับ 5 ลักษณะ)	DN_C DN_A DN_T DN_G DN_CT	DN_C DN_A DN_T DN_G DN_CT
เทคนิครีลีฟ-เอฟ (ผู้ใช้กำหนดเท่ากับ 5 ลักษณะ)	DN_T DN_A DN_C DN_G DN_TC	DN_T DN_A DN_C DN_G DN_TC

ตัวอย่างชุดข้อมูล TIS+50 เลือกลักษณะเฉพาะด้วยเทคนิครีลีฟ-เอฟ พบว่าลักษณะเฉพาะที่ผ่านการเลือก 5 ลำดับแรกที่มีนัยสำคัญสูงสุด สำหรับค่าความถี่ คือ DN_GC UP_T UP_AT UP_A DN_CT ซึ่งเหมือนกับค่า TF-IDF คือ DN_GC UP_T UP_AT UP_A DN_CT แสดงดังตารางที่ 5.9

ตารางที่ 5.9 ลักษณะเฉพาะที่ผ่านการเลือกสำหรับค่าความถี่ และค่า TF-IDF ของชุดข้อมูล TIS+50

ชื่อเทคนิคการเลือกลักษณะเฉพาะ	รูปแบบลักษณะเฉพาะที่ผ่านเลือก	
	ค่าความถี่	ค่า TF-IDF
เทคนิค CFS (ฮิวริสติกเลือกเท่ากับ 13 ลักษณะ)	UP_A UP_C UP_T DN_C UP_AG UP_AT UP_CA UP_GA UP_TA UP_TG UP_TT DN_CG DN_GC	UP_A UP_C UP_T DN_C UP_AG UP_AT UP_CA UP_GA UP_TA UP_TG UP_TT DN_CG DN_GC
เทคนิคไคสแควร์ (ผู้ใช้กำหนดเท่ากับ 5 ลักษณะ)	UP_AT UP_TG UP_A UP_T UP_GA	UP_AT UP_TG UP_A UP_T UP_C
เทคนิคอัตราส่วนเกน (ผู้ใช้กำหนดเท่ากับ 5 ลักษณะ)	UP_AG UP_GA UP_G UP_C UP_A	UP_AG UP_GA UP_G UP_C UP_A
เทคนิครีลีฟ-เอฟ (ผู้ใช้กำหนดเท่ากับ 5 ลักษณะ)	DN_GC UP_T UP_AT UP_A DN_CT	DN_GC UP_T UP_AT UP_A DN_CT

จากตารางที่ 5.7 ถึง 5.9 แสดงให้เห็นว่าจำนวนลักษณะเฉพาะที่ผ่านการเลือกด้วยเทคนิค CFS มีจำนวนลักษณะเฉพาะที่ผ่านการเลือกน้อยกว่า 40 เสมอ ตัวอย่างเช่น ชุดข้อมูล Vertebrate มีจำนวนลักษณะเฉพาะที่ผ่านการเลือกเท่ากับ 14 คือ DN_A DN_C DN_G DN_T UP_AT DN_AC DN_AG DN_CA DN_CG DN_CT DN_GC DN_GG DN_GT และ DN_TC แสดงดังตารางที่ 5.7 ชุดข้อมูล A.thaliana มีจำนวนลักษณะเฉพาะที่ผ่านการเลือกเท่ากับ 5 คือ DN_A DN_C DN_G DN_T และ DN_CG แสดงดังตารางที่ 5.8 และชุดข้อมูล

TIS+50 มีจำนวนลักษณะเฉพาะที่ผ่านการเลือกเท่ากับ 13 ลักษณะ คือ UP_A UP_C UP_T DN_C UP_AG UP_AT UP_CA UP_GA UP_TA UP_TG UP_TT DN_CG และ DN_GC

สำหรับเทคนิคไคสแควร์ เทคนิคอัตราส่วนเกน และเทคนิคริลีฟ-เอฟเป็นเทคนิคแบบตัวกรองที่เรียงลำดับลักษณะเฉพาะตามนัยสำคัญทางสถิติจากมากไปน้อย เปรียบเทียบจำนวนลักษณะเฉพาะที่เหมาะสมโดยเลือกลักษณะเฉพาะ 5 10 15 20 และ 30 ลำดับแรกที่มีนัยสำคัญสูงสุด พบว่าลักษณะเฉพาะที่ผ่านการเลือกด้วยเทคนิคไคสแควร์ เทคนิคอัตราส่วนเกน และเทคนิคริลีฟ-เอฟมีลักษณะเฉพาะที่ไม่เหมือนกัน ตัวอย่างการแบ่งสายพันธุกรรมด้วยขนาดหน้าต่าง 303 นิวคลีโอไทด์ สร้างลักษณะเฉพาะ 1-แกรม และ 2-แกรม กำหนดค่าลักษณะเฉพาะด้วยวิธี TF-IDF สำหรับชุดข้อมูล Vertebrate พบว่า ลักษณะเฉพาะที่ผ่านการเลือก 5 ลำดับแรกที่มีนัยสำคัญสูงสุดของเทคนิคไคสแควร์ คือ DN_G DN_C DN_T DN_GC และ DN_CT เทคนิคอัตราส่วนเกน คือ DN_G DN_C DN_T DN_A และ DN_CA และเทคนิคริลีฟ-เอฟ คือ DN_C DN_A DN_G DN_T และ DN_AG แสดงดังตารางที่ 5.10

ตารางที่ 5.10 ลักษณะเฉพาะที่ผ่านการเลือกสำหรับค่า TF-IDF ของชุดข้อมูล Vertebrate

จำนวนลักษณะเฉพาะ	รูปแบบลักษณะเฉพาะที่ผ่านการเลือก		
	เทคนิคไคสแควร์	เทคนิคอัตราส่วนเกน	เทคนิคริลีฟ-เอฟ
5	DN_G DN_C DN_T DN_GC DN_CT	DN_G DN_C DN_T DN_A DN_CA	DN_C DN_A DN_G DN_T DN_AG
10	DN_G DN_C DN_T DN_GC DN_CT DN_TC DN_A DN_TG DN_CA DN_GG	DN_G DN_C DN_T DN_A DN_CA DN_CT DN_TC DN_AG DN_TG DN_GC	DN_C DN_A DN_G DN_T DN_AG DN_CT DN_GA DN_CA DN_TG DN_TC
15	DN_G DN_C DN_T DN_GC DN_CT DN_TC DN_A DN_TG DN_CA DN_GG DN_AG DN_CC DN_GT DN_AC DN_GA	DN_G DN_C DN_T DN_A DN_CA DN_CT DN_TC DN_AG DN_TG DN_GC DN_AC DN_GG DN_GT DN_GA DN_CC	DN_C DN_A DN_G DN_T DN_AG DN_CT DN_GA DN_CA DN_TG DN_TC DN_GC DN_AC DN_CC DN_GT DN_AT
20	DN_G DN_C DN_T DN_GC DN_CT DN_TC DN_A DN_TG DN_CA DN_GG DN_AG DN_CC DN_GT DN_AC DN_GA DN_CG UP_AT DN_TT DN_AT DN_AA	DN_G DN_C DN_T DN_A DN_CA DN_CT DN_TC DN_AG DN_TG DN_GC DN_AC DN_GG DN_GT DN_GA DN_CC UP_AT DN_CG UP_CG DN_AT DN_TT	DN_C DN_A DN_G DN_T DN_AG DN_CT DN_GA DN_CA DN_TG DN_TC DN_GC DN_AC DN_CC DN_GT DN_AT DN_AA DN_GG DN_TT DN_TA DN_CG

ตารางที่ 5.10 ลักษณะเฉพาะที่ผ่านการเลือกสำหรับค่า TF-IDF ของชุดข้อมูล Vertebrate (ต่อ)

จำนวนลักษณะเฉพาะ	รูปแบบลักษณะเฉพาะที่ผ่านการเลือก		
	เทคนิคไคสแควร์	เทคนิคอัตราส่วนเกน	เทคนิครีลีฟ-เอฟ
30	DN_G DN_C DN_T DN_GC DN_CT DN_TC DN_A DN_TG DN_CA DN_GG DN_AG DN_CC DN_GT DN_AC DN_GA DN_CG UP_AT DN_TT DN_AT DN_AA UP_TG UP_A DN_TA UP_T UP_CA UP_AA UP_G UP_GA UP_C UP_TA	DN_G DN_C DN_T DN_A DN_CA DN_CT DN_TC DN_AG DN_TG DN_GC DN_AC DN_GG DN_GT DN_GA DN_CC UP_AT DN_CG UP_CG DN_AT DN_TT DN_AA UP_A DN_TA UP_TG UP_T UP_AA UP_CA UP_G UP_GA UP_AG	DN_C DN_A DN_G DN_T DN_AG DN_CT DN_GA DN_CA DN_TG DN_TC DN_GC DN_AC DN_CC DN_GT DN_AT DN_AA DN_GG DN_TT DN_TA DN_CG UP_AT UP_TG UP_TT UP_CG UP_AG UP_TA UP_AC UP_C UP_CA UP_CT

ตัวอย่างชุดข้อมูล A.thaliana ลักษณะเฉพาะที่ผ่านการเลือก 5 ลำดับแรกของเทคนิคไคส์แควร์ คือ DN_C DN_G DN_T DN_A และ DN_GC เทคนิคอัตราส่วนเกน คือ DN_C DN_A DN_T DN_G และ DN_CT และเทคนิครีลีฟ-เอฟ คือ DN_T DN_A DN_C DN_G และ DN_TC แสดงดังตารางที่ 5.11

ตารางที่ 5.11 ลักษณะเฉพาะที่ผ่านการเลือกสำหรับค่า TF-IDF ของชุดข้อมูล A.thaliana

จำนวนลักษณะเฉพาะ	รูปแบบลักษณะเฉพาะที่ผ่านการเลือก		
	เทคนิคไคสแควร์	เทคนิคอัตราส่วนเกน	เทคนิครีลีฟ-เอฟ
5	DN_C DN_G DN_T DN_A DN_GC	DN_C DN_A DN_T DN_G DN_CT	DN_T DN_A DN_C DN_G DN_TC
10	DN_C DN_G DN_T DN_A DN_GC DN_TC DN_CG DN_CT DN_GT DN_CA	DN_C DN_A DN_T DN_G DN_CT DN_TC DN_CG DN_GT DN_CA DN_GC	DN_T DN_A DN_C DN_G DN_TC DN_CA DN_GA DN_AG DN_AT DN_TT
15	DN_C DN_G DN_T DN_A DN_GC DN_TC DN_CG DN_CT DN_GT DN_CA DN_AC DN_GA DN_AG DN_CC DN_AA	DN_C DN_A DN_T DN_G DN_CT DN_TC DN_CG DN_GT DN_CA DN_GC DN_AT DN_AG DN_TT DN_GA DN_AA	DN_T DN_A DN_C DN_G DN_TC DN_CA DN_GA DN_AG DN_AT DN_TT DN_GT DN_CT DN_AA DN_TG DN_AC

ตารางที่ 5.11 ลักษณะเฉพาะที่ผ่านการเลือกสำหรับค่า TF-IDF ของชุดข้อมูล A.thaliana (ต่อ)

จำนวนลักษณะเฉพาะ	รูปแบบลักษณะเฉพาะที่ผ่านการเลือก		
	เทคนิคไคสแควร์	เทคนิคอัตราส่วนเกน	เทคนิครีลีฟ-เอฟ
20	DN_C DN_G DN_T DN_A DN_GC DN_TC DN_CG DN_CT DN_GT DN_CA DN_AC DN_GA DN_AG DN_CC DN_AA DN_TT UP_G DN_AT DN_GG UP_GG	DN_C DN_A DN_T DN_G DN_CT DN_TC DN_CG DN_GT DN_CA DN_GC DN_AT DN_AG DN_TT DN_GA DN_AA DN_AC DN_TA DN_CC DN_GG UP_G	DN_T DN_A DN_C DN_G DN_TC DN_CA DN_GA DN_AG DN_AT DN_TT DN_GT DN_CT DN_AA DN_TG DN_AC DN_CG DN_GC DN_GG DN_CC DN_TA
30	DN_C DN_G DN_T DN_A DN_GC DN_TC DN_CG DN_CT DN_GT DN_CA DN_AC DN_GA DN_AG DN_CC DN_AA DN_TT UP_G DN_AT DN_GG UP_GG DN_TG UP_TG DN_TA UP_GC UP_GA UP_T UP_GT UP_A UP_AT UP_AG	DN_C DN_A DN_T DN_G DN_CT DN_TC DN_CG DN_GT DN_CA DN_GC DN_AT DN_AG DN_TT DN_GA DN_AA DN_AC DN_TA DN_CC DN_GG UP_G DN_TG UP_GC UP_TG UP_A UP_GG UP_T UP_AT UP_GA UP_TA UP_GT	DN_T DN_A DN_C DN_G DN_TC DN_CA DN_GA DN_AG DN_AT DN_TT DN_GT DN_CT DN_AA DN_TG DN_AC DN_CG DN_GC DN_GG DN_CC DN_TA UP_TG UP_G UP_GG UP_TA UP_GC UP_TT UP_CA UP_CG UP_T UP_GA

ตัวอย่างชุดข้อมูล TIS+50 ลักษณะเฉพาะที่ผ่านการเลือก 5 ลำดับแรกของเทคนิคไคสแควร์ คือ UP_AT UP_TG UP_A UP_T และ UP_C เทคนิคอัตราส่วนเกน คือ UP_AG UP_GA UP_G UP_C และ UP_A และเทคนิครีลีฟ-เอฟ คือ DN_GC UP_T UP_AT UP_A และ DN_CT แสดงดังตารางที่ 5.12

ตารางที่ 5.12 ลักษณะเฉพาะที่ผ่านการเลือกสำหรับค่า TF-IDF ของชุดข้อมูล TIS+50

จำนวนลักษณะเฉพาะ	รูปแบบลักษณะเฉพาะที่ผ่านการเลือก		
	เทคนิคไคสแควร์	เทคนิคอัตราส่วนเกน	เทคนิครีลีฟ-เอฟ
5	UP_AT UP_TG UP_A UP_T UP_C	UP_AG UP_GA UP_G UP_C UP_A	DN_GC UP_T UP_AT UP_A DN_CT
10	UP_AT UP_TG UP_A UP_T UP_C UP_GA UP_G UP_AC UP_AG UP_CA	UP_AG UP_GA UP_G UP_C UP_A UP_T UP_AT UP_CA UP_GG UP_TG	DN_GC UP_T UP_AT UP_A DN_CT UP_TG DN_C UP_TA UP_CA DN_CG

ตารางที่ 5.12 ลักษณะเฉพาะที่ผ่านการเลือกสำหรับค่า TF-IDF ของชุดข้อมูล TIS+50 (ต่อ)

จำนวนลักษณะเฉพาะ	รูปแบบลักษณะเฉพาะที่ผ่านการเลือก		
	เทคนิคไคสแควร์	เทคนิคอัตราส่วนเกน	เทคนิครีลีฟ-เอฟ
15	UP_AT UP_TG UP_A UP_T UP_C UP_GA UP_G UP_AC UP_AG UP_CA UP_GT UP_TT UP_TA UP_AA DN_GC	UP_AG UP_GA UP_G UP_C UP_A UP_T UP_AT UP_CA UP_GG UP_TG UP_TT UP_CC UP_TA UP_AC UP_AA	DN_GC UP_T UP_AT UP_A DN_CT UP_TG DN_C UP_TA UP_CA DN_CG DN_G DN_TC UP_TC DN_CC UP_AA
20	UP_AT UP_TG UP_A UP_T UP_C UP_GA UP_G UP_AC UP_AG UP_CA UP_GT UP_TT UP_TA UP_AA DN_GC UP_GG UP_CT UP_TC DN_CG UP_CC	UP_AG UP_GA UP_G UP_C UP_A UP_T UP_AT UP_CA UP_GG UP_TG UP_TT UP_CC UP_TA UP_AC UP_AA UP_CT UP_TC DN_GC UP_GC UP_GT	DN_GC UP_T UP_AT UP_A DN_CT UP_TG DN_C UP_TA UP_CA DN_CG DN_G DN_TC UP_TC DN_CC UP_AA DN_AG UP_TT UP_GA UP_CC UP_AG
30	UP_AT UP_TG UP_A UP_T UP_C UP_GA UP_G UP_AC UP_AG UP_CA UP_GT UP_TT UP_TA UP_AA DN_GC UP_GG UP_CT UP_TC DN_CG UP_CC UP_GC DN_C DN_CT DN_TC DN_CC DN_G DN_T DN_TT DN_A DN_TG	UP_AG UP_GA UP_G UP_C UP_A UP_T UP_AT UP_CA UP_GG UP_TG UP_TT UP_CC UP_TA UP_AC UP_AA UP_CT UP_TC DN_GC UP_GC UP_GT DN_CG DN_C DN_G DN_CT DN_TC DN_CC DN_TA DN_T DN_TG DN_A	DN_GC UP_T UP_AT UP_A DN_CT UP_TG DN_C UP_TA UP_CA DN_CG DN_G DN_TC UP_TC DN_CC UP_AA DN_AG UP_TT UP_GA UP_CC UP_AG DN_CA DN_A UP_CT DN_T DN_TA DN_GA UP_CG UP_C DN_AT DN_TG

ขั้นตอนที่ 4 การสร้างลักษณะเฉพาะรูปแบบคอนเซนซัส

รูปแบบคอนเซนซัสเป็นลักษณะโดยทั่วไปรอบจุดเริ่มต้นการแปลรหัสพันธุกรรมที่มีงานวิจัยศึกษาก่อนหน้าได้แก่ ลักษณะเด่นรอบจุดเริ่มต้นการแปลรหัสพันธุกรรม (Kozak, 1987) และ แบบจำลองการตรวจสอบไรโบโซม (Kozak, 1989) ลักษณะเฉพาะที่สร้างจากรูปแบบคอนเซนซัสมี 4 ลักษณะเฉพาะแสดงดังตารางที่ 5.14 กำหนดให้นิวคลีโอไทด์ A ของโคดอน ATG เป้าหมายเป็นตำแหน่ง +1

ตารางที่ 5.13 ลักษณะเฉพาะที่สร้างจากรูปแบบคอนเซนซัส

ลักษณะเฉพาะ	คำอธิบาย
DN_+4G	ลักษณะเฉพาะบูลีนเป็นจริงถ้ามีนิวคลีโอไทด์ G ที่ตำแหน่ง +4 (Kozak, 1987)
UP_-3A/G	ลักษณะเฉพาะบูลีนเป็นจริงถ้ามีนิวคลีโอไทด์ A หรือ G ที่ตำแหน่ง -3 (Kozak, 1987)
UP_ATG	นับความถี่ของโคดอน ATG อัพสตรีม (Kozak, 1989)
DN_STOP	นับความถี่ของโคดอนหยุด (TAA, TAG, หรือ TGA) ในอิน-เฟรมดาวน์สตรีม (Kozak, 1989)

ขั้นตอนที่ 5 การทำนายผลลัพธ์

การทดลองแต่ละแบบมีสถาปัตยกรรมโครงข่ายประสาทเทียมแบบหน่วยประมวลผลย่อยหลายชั้นในชั้นข้อมูลเข้าแตกต่างกัน และประเมินผลด้วยวิธีการทดสอบไขว้เปลี่ยนแบบ k กลุ่ม นั้นคือ ข้อมูลจะแบ่งเป็น k กลุ่มเท่า ๆ กัน จากนั้น k -1 ส่วนจะใช้เป็นชุดข้อมูลสอน และ 1 ส่วนจะใช้เป็นชุดข้อมูลทดสอบ โดยชุดข้อมูลสอนและชุดข้อมูลทดสอบจะแตกต่างกันทั้ง k ครั้ง ผลการทดลองสามารถสรุปเป็น 7 ประเด็น ได้แก่ 1) ประสิทธิภาพขนาดหน้าต่าง 2) ประสิทธิภาพ n-แกรม 3) ประสิทธิภาพของค่า TF-IDF 4) ประสิทธิภาพของเวลาสำหรับ TF-IDF 5) ประสิทธิภาพของเทคนิคการเลือกลักษณะเฉพาะ 6) ประสิทธิภาพของรูปแบบคอนเซนซัส และ 7) ประสิทธิภาพของแบบจำลอง TF-IDF-NN-TIS โดยประสิทธิภาพการทำนายที่พิจารณาได้แก่ ค่าความถูกต้อง (Accuracy) คือร้อยละของตัวอย่างที่ทำนายถูกต้องทั้งหมด ค่าการตอบสนองไว (Sensitivity) คือร้อยละของตัวอย่างกลุ่มบวกที่ทำนายถูกต้อง และค่าความเฉพาะเจาะจง (Specificity) คือร้อยละของตัวอย่างกลุ่มลบที่ทำนายถูกต้อง

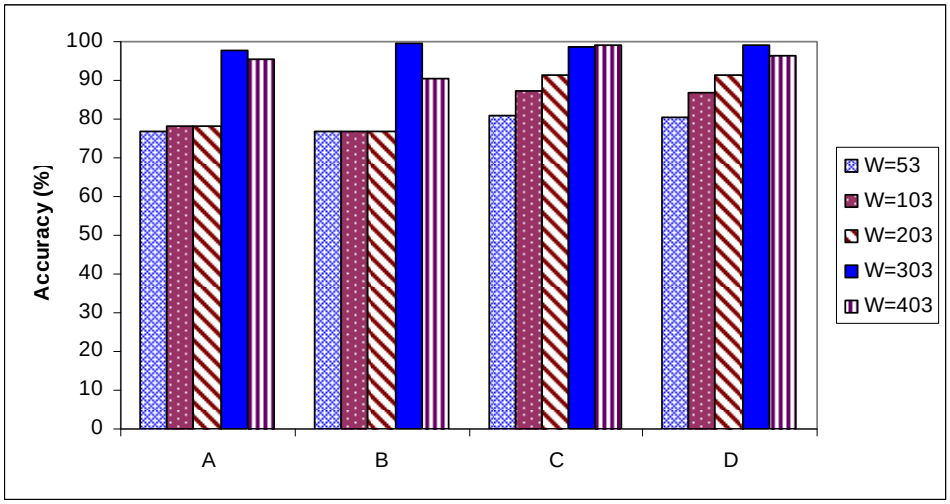
1) ประเด็นประสิทธิภาพขนาดหน้าต่าง

พิจารณาขนาดหน้าต่างที่แตกต่างกันสำหรับชุดข้อมูล Vertebrate ชุดข้อมูล A.thaliana และชุดข้อมูล TIS+50 ของการทดลองแบบ A B C และ D ผลการทดลองแสดงให้เห็นว่า ขนาดหน้าต่างที่มีจำนวนนิวคลีโอไทด์เหมาะสมจะให้ ค่าความถูกต้อง ค่าการตอบสนองไว และค่าความเฉพาะเจาะจงที่มีค่าสูง ตัวอย่างเช่น สร้างลักษณะเฉพาะ 1-แกรม กำหนดค่าลักษณะเฉพาะด้วยค่า TF-IDF ของการทดลอง A B C และ D เลือกลักษณะเฉพาะด้วยเทคนิค CFS

สำหรับชุดข้อมูล Vertebrate แสดงดังตารางที่ 5.14 และภาพประกอบ 5.4 การทดลองแบบ B ของขนาดหน้าต่าง 53 103 203 303 และ 403 ให้ค่าความถูกต้องเท่ากับ 76.67% 76.82% 76.61% 99.62% และ 90.37% ตามลำดับ ผลการทดลองแสดงให้เห็นว่า ขนาดหน้าต่าง 303 นิวคลีโอไทด์ ให้ค่าความถูกต้องสูงสุดเท่ากับ 99.62%

ตารางที่ 5.14 ค่าความถูกต้องของขนาดหน้าต่างที่แตกต่างกันสำหรับชุดข้อมูล Vertebrate

การทดลอง	ค่าความถูกต้อง (%)				
	W = 53	W = 103	W = 203	W = 303	W = 403
A	76.80	78.06	78.29	**97.78**	95.29
B	76.67	76.82	76.61	**99.62**	90.37
C	81.13	87.43	91.42	98.79	**98.90**
D	80.54	86.73	91.31	**98.89**	96.55

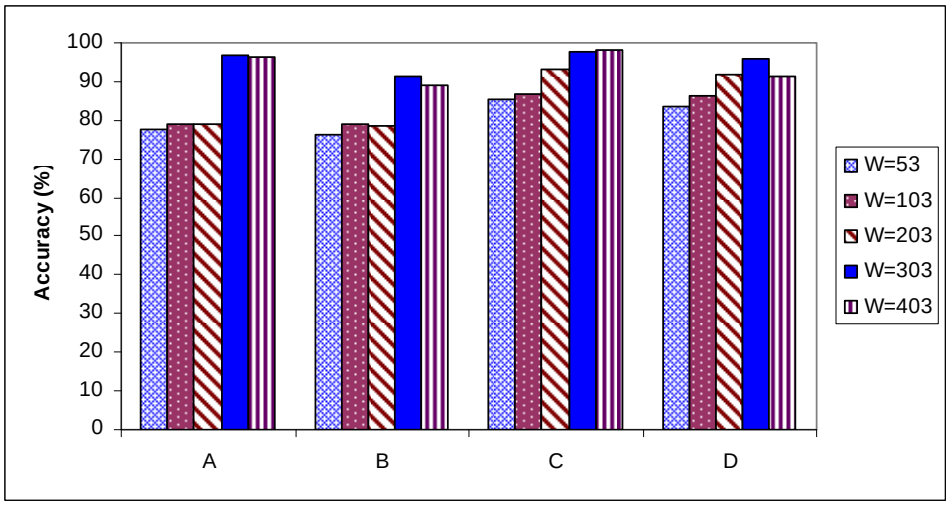


ภาพประกอบ 5.4 เปรียบเทียบค่าความถูกต้องของขนาดหน้าต่างที่แตกต่างกันสำหรับชุดข้อมูล Vertebrate

สำหรับชุดข้อมูล A.thaliana แสดงดังตารางที่ 5.15 และภาพประกอบ 5.5 การทดลองแบบ C ของขนาดหน้าต่าง 53 103 203 303 และ 403 ให้ค่าความถูกต้องเท่ากับ 85.50% 86.67% 92.97% 97.66% และ 98.10% ตามลำดับ ผลการทดลองแสดงให้เห็นว่าขนาดหน้าต่าง 403 นิวคลีโอไทด์ ให้ค่าความถูกต้องสูงสุดเท่ากับ 98.10% ทั้งนี้ขนาดหน้าต่าง 303 นิวคลีโอไทด์ก็ให้ค่าความถูกต้องสูงเช่นเดียวกัน

ตารางที่ 5.15 ค่าความถูกต้องของขนาดหน้าต่างที่แตกต่างกันสำหรับชุดข้อมูล A.thaliana

การทดลอง	ค่าความถูกต้อง (%)				
	W = 53	W = 103	W = 203	W = 303	W = 403
A	77.54	78.81	78.91	**96.83**	96.29
B	76.03	78.86	78.71	**91.11**	89.26
C	85.50	86.67	92.97	**97.66**	**98.10**
D	83.35	86.23	91.80	**95.85**	91.11

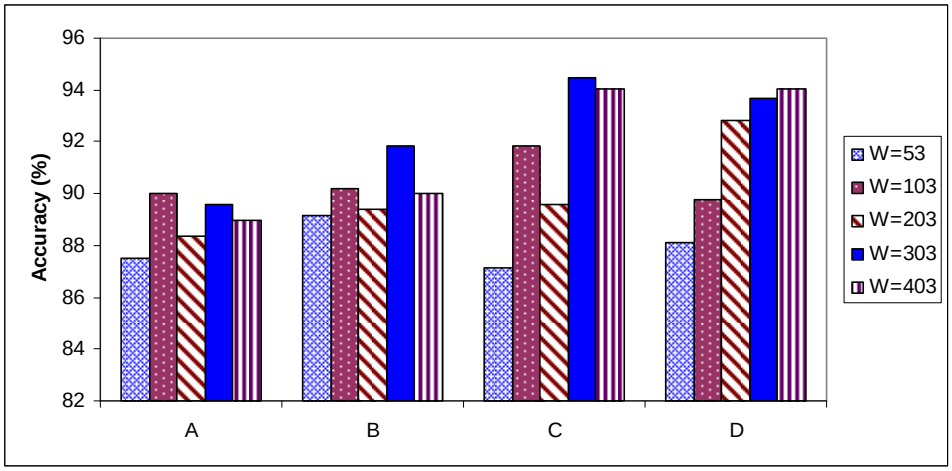


ภาพประกอบ 5.5 เปรียบเทียบค่าความถูกต้องของขนาดหน้าต่าง
ที่แตกต่างกันสำหรับชุดข้อมูล A.thaliana

สำหรับชุดข้อมูล TIS+50 แสดงดังตารางที่ 5.16 และภาพประกอบ 5.6 การทดลองแบบ C ของขนาดหน้าต่าง 53 103 203 303 และ 403 ให้ค่าความถูกต้องเท่ากับ 87.12% 91.82% 89.57% 94.48% และ 94.07% ตามลำดับ ผลการทดลองแสดงให้เห็นว่า ขนาดหน้าต่าง 303 นิวคลีโอไทด์ ให้ค่าความถูกต้องสูงสุดเท่ากับ 94.48%

ตารางที่ 5.16 ค่าความถูกต้องของขนาดหน้าต่างที่แตกต่างกันสำหรับชุดข้อมูล TIS+50

การทดลอง	ค่าความถูกต้อง (%)				
	W = 53	W = 103	W = 203	W = 303	W = 403
A	87.53	**89.98**	88.34	89.57	88.96
B	89.16	90.18	89.37	**91.82**	89.98
C	87.12	91.82	89.57	**94.48**	94.07
D	88.14	89.78	92.84	93.66	**94.07**



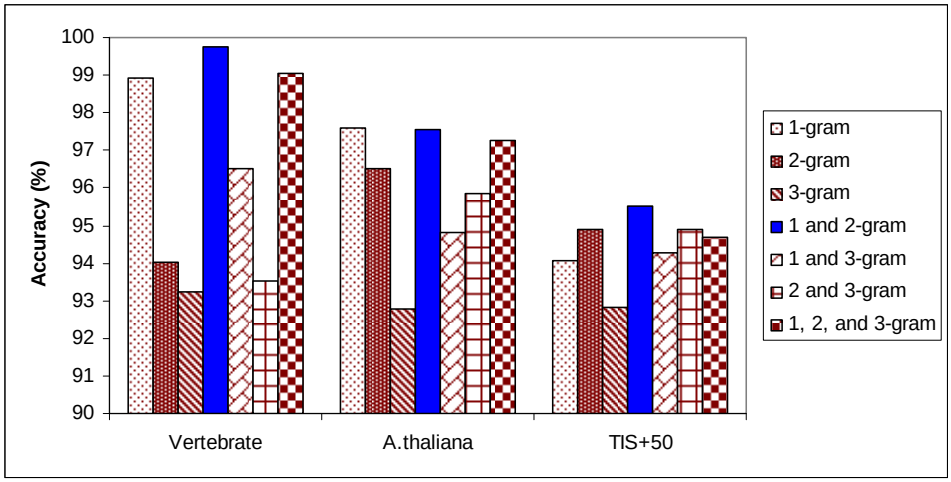
ภาพประกอบ 5.6 เปรียบเทียบค่าความถูกต้องของขนาดหน้าต่าง
ที่แตกต่างกันสำหรับชุดข้อมูล TIS+50

2) ประเด็นประสิทธิภาพ n-แกรม

พิจารณาลักษณะเฉพาะ n-แกรม กำหนด *n* เท่ากับ 1 2 หรือ 3 ผลการทดลองแสดงให้เห็นว่า การเลือก *n* ที่เหมาะสมจะช่วยเพิ่มความถูกต้องการทำนายจุดเริ่มต้นการแปลรหัส ตัวอย่างเช่น แบ่งสายพันธุกรรมด้วยขนาดหน้าต่าง 303 นิวคลีโอไทด์ เปรียบเทียบค่าความถูกต้องของลักษณะเฉพาะ n-แกรม 7 รูปแบบ คือ 1) ลักษณะ 1-แกรม 2) ลักษณะเฉพาะ 2-แกรม 3) ลักษณะเฉพาะ 3-แกรม 4) ลักษณะเฉพาะ 1-แกรม และ 2-แกรม 5) ลักษณะเฉพาะ 1-แกรม และ 3-แกรม 6) ลักษณะเฉพาะ 2-แกรม และ 3-แกรม และ 7) ลักษณะเฉพาะ 1-แกรม 2-แกรม และ 3-แกรม กำหนดค่าลักษณะเฉพาะด้วยค่า TF-IDF เลือกลักษณะเฉพาะด้วยเทคนิค CFS พิจารณาค่าความถูกต้องการทดลองแบบ D ของลักษณะเฉพาะ n-แกรมทั้ง 7 รูปแบบ ผลการทดลองแสดงให้เห็นว่า ชุดข้อมูล Vertebrate ลักษณะเฉพาะ 1-แกรม และ 2-แกรมให้ค่าความถูกต้องสูงสุดเท่ากับ 99.76% ชุดข้อมูล A.thaliana ลักษณะเฉพาะ 1-แกรมให้ค่าความถูกต้องสูงสุดเท่ากับ 97.61% และชุดข้อมูล TIS+50 ลักษณะเฉพาะ 1-แกรม และ 2-แกรม ให้ค่าความถูกต้องสูงสุดเท่ากับ 95.50% แสดงดังตารางที่ 5.17 และภาพประกอบ 5.7

ตารางที่ 5.17 ค่าความถูกต้องของลักษณะเฉพาะ n-แกรม

ชื่อชุดข้อมูล	ค่าความถูกต้อง (%)						
	1	2	3	1 และ 2	1 และ 3	2 และ 3	1 2 และ 3
Vertebrate	98.91	94.03	93.24	**99.76**	96.51	93.53	99.06
A.thaliana	**97.61**	96.53	92.77	97.56	94.82	95.85	97.27
TIS+50	94.07	94.89	92.84	**95.50**	94.27	94.89	94.68

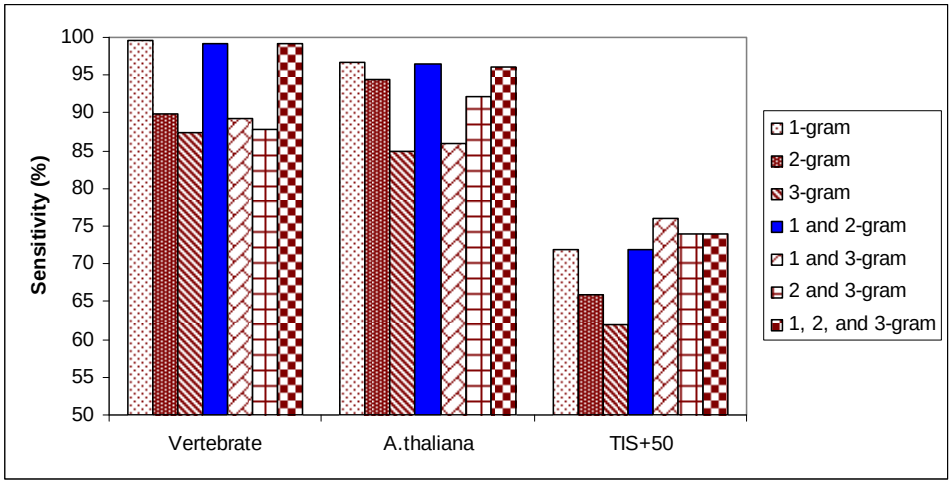


ภาพประกอบ 5.7 เปรียบเทียบค่าความถูกต้องของลักษณะเฉพาะ n-แกรม

พิจารณาค่าการตอบสนองไวการทดลองแบบ D ของลักษณะเฉพาะ n-แกรมทั้ง 7 รูปแบบ ผลการทดลองแสดงให้เห็นว่า ชุดข้อมูล Vertebrate ลักษณะเฉพาะ 1-แกรม ให้ค่าการตอบสนองไวสูงสุดเท่ากับ 99.49% ชุดข้อมูล A.thaliana ลักษณะเฉพาะ 1-แกรม ให้ค่าการตอบสนองไวสูงสุดเท่ากับ 96.75% และชุดข้อมูล TIS+50 ลักษณะเฉพาะ 1-แกรม และ 3-แกรมให้ค่าการตอบสนองไวสูงสุดเท่ากับ 76.00% แสดงดังตารางที่ 5.18 และภาพประกอบ 5.8

ตารางที่ 5.18 ค่าการตอบสนองไวของลักษณะเฉพาะ n-แกรม

ชื่อชุดข้อมูล	ค่าการตอบสนองไว (%)						
	1	2	3	1 และ 2	1 และ 3	2 และ 3	1 2 และ 3
Vertebrate	**99.49**	89.79	87.32	99.18	89.25	87.77	99.18
A.thaliana	**96.75**	94.46	84.89	96.56	85.85	92.16	96.18
TIS+50	72.00	66.00	62.00	72.00	**76.00**	74.00	74.00

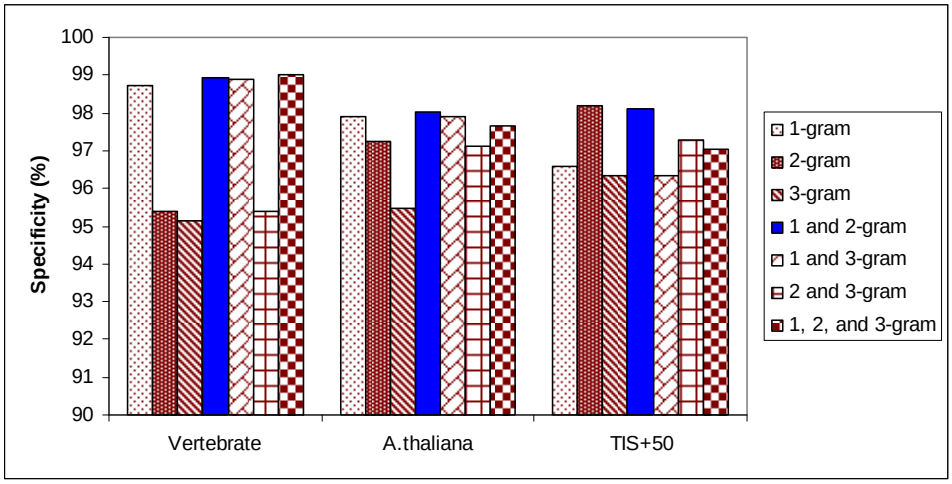


ภาพประกอบ 5.8 เปรียบเทียบค่าการตอบสนองไวของลักษณะเฉพาะ n-แกรม

พิจารณาค่าความเฉพาะเจาะจงการทดลองแบบ D ของลักษณะเฉพาะ n-แกรมทั้ง 7 รูปแบบ ผลการทดลองแสดงให้เห็นว่า ชุดข้อมูล Vertebrate ลักษณะเฉพาะ 1-แกรม 2-แกรม และ 3-แกรมให้ค่าความเฉพาะเจาะจงสูงสุดเท่ากับ 99.02% ชุดข้อมูล A.thaliana ลักษณะเฉพาะ 1-แกรม และ 2-แกรมให้ค่าความเฉพาะเจาะจงสูงสุดเท่ากับ 98.03% และชุดข้อมูล TIS+50 ลักษณะเฉพาะ 2-แกรมให้ค่าความเฉพาะเจาะจงสูงสุดเท่ากับ 98.18% แสดงดังตารางที่ 5.19 และภาพประกอบ 5.9

ตารางที่ 5.19 ค่าความเฉพาะเจาะจงของลักษณะเฉพาะ n-แกรม

ชื่อชุดข้อมูล	ค่าความเฉพาะเจาะจง (%)						
	1	2	3	1 และ 2	1 และ 3	2 และ 3	1 2 และ 3
Vertebrate	98.72	95.41	95.16	98.95	98.87	95.41	**99.02**
A.thaliana	97.90	97.25	95.48	**98.03**	97.90	97.11	97.64
TIS+50	96.58	**98.18**	96.36	98.12	96.36	97.27	97.04

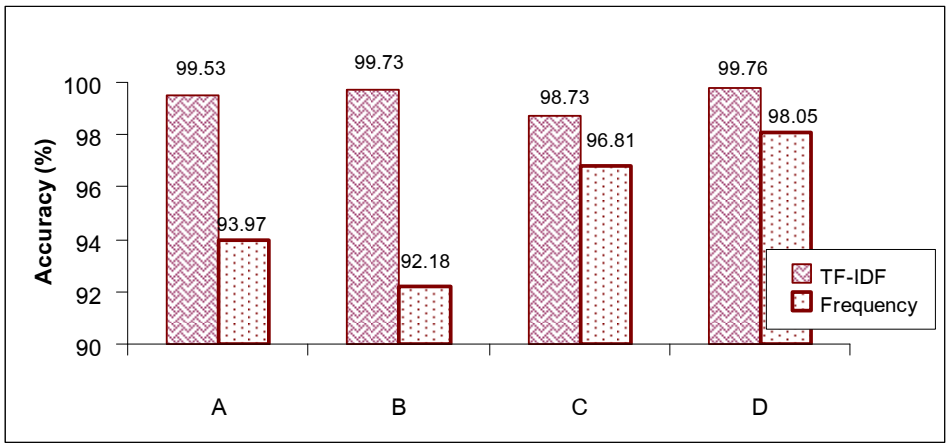


ภาพประกอบ 5.9 เปรียบเทียบค่าความเฉพาะเจาะจงของลักษณะเฉพาะ n-แกรม

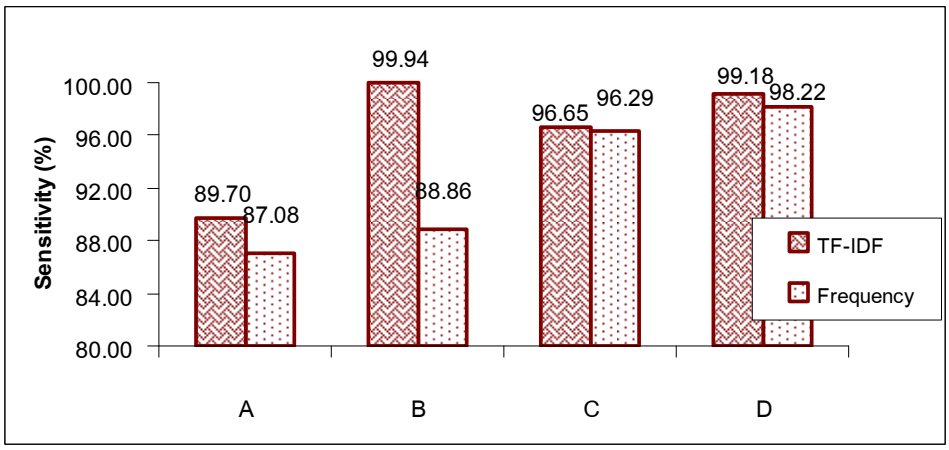
1) ประเด็นประสิทธิภาพของค่า TF-IDF

พิจารณาระหว่างค่า TF-IDF และค่าความถี่ ผลการทดลองแสดงให้เห็นว่า ค่า TF-IDF ให้ประสิทธิภาพการทำนายจุดเริ่มต้นการแปลรหัสสูงกว่า ค่าความถี่ สำหรับชุดข้อมูล Vertebrate ชุดข้อมูล A.thaliana และชุดข้อมูล TIS+50 ตัวอย่างเช่น แบ่งสายพันธุกรรมด้วยขนาดหน้าต่าง 303 นิวคลีโอไทด์ สร้างลักษณะเฉพาะ 1-แกรม และ 2-แกรม เลือกลักษณะเฉพาะด้วยเทคนิค CFS

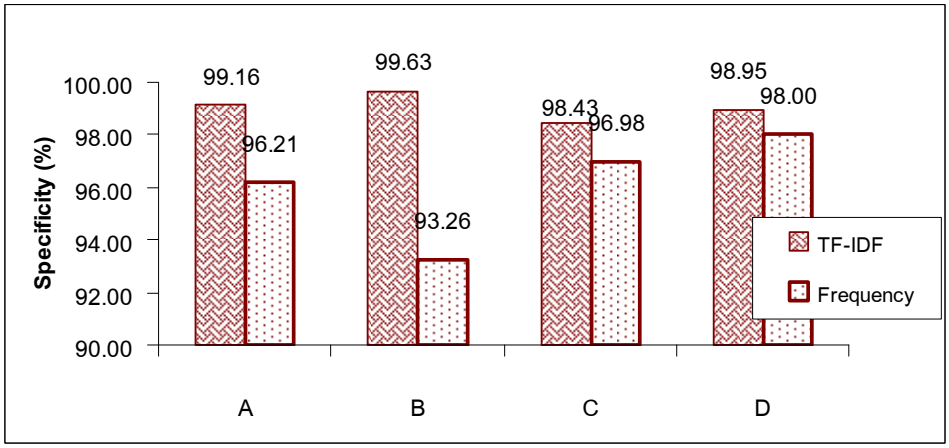
สำหรับชุดข้อมูล Vertebrate ของการทดลองแบบ A B C และ D ค่าความถูกต้องสำหรับค่า TF-IDF เท่ากับ 99.53% 99.73% 98.73% และ 99.76% ตามลำดับ สูงกว่าค่าความถี่ซึ่งให้ค่าความถูกต้องเท่ากับ 93.97% 92.18% 96.81% และ 98.05% ตามลำดับ แสดงดังภาพประกอบ 5.10 ค่าการตอบสนองไวสำหรับค่า TF-IDF เท่ากับ 89.70% 99.94% 96.65% และ 99.18% ตามลำดับ สูงกว่าค่าความถี่ซึ่งให้ค่าการตอบสนองไวเท่ากับ 87.08% 88.86% 96.29% และ 98.22% ตามลำดับ แสดงดังภาพประกอบ 5.11 และค่าความเฉพาะเจาะจงสำหรับค่า TF-IDF เท่ากับ 99.16% 99.63% 98.43% และ 98.95% ตามลำดับ สูงกว่าค่าความถี่ซึ่งให้ค่าความเฉพาะเจาะจงเท่ากับ 96.21% 93.26% 96.98% และ 98.00% ตามลำดับ แสดงดังภาพประกอบ 5.12



ภาพประกอบ 5.10 เปรียบเทียบค่าความถูกต้องของการทดลองแบบ A B C และ D ระหว่างค่าความถี่ และค่า TF-IDF สำหรับชุดข้อมูล Vertebrate

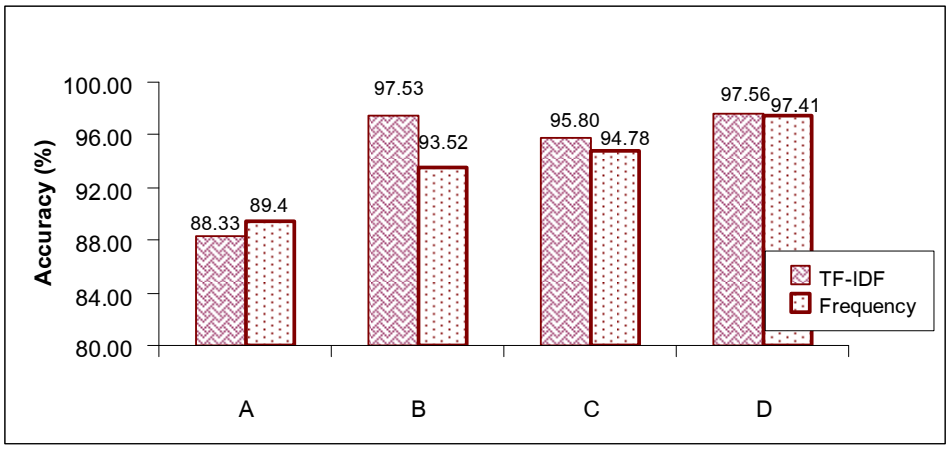


ภาพประกอบ 5.11 เปรียบเทียบค่าการตอบสนองไวของการทดลองแบบ A B C และ D ระหว่าง ค่าความถี่ กับ ค่า TF-IDF สำหรับชุดข้อมูล Vertebrate

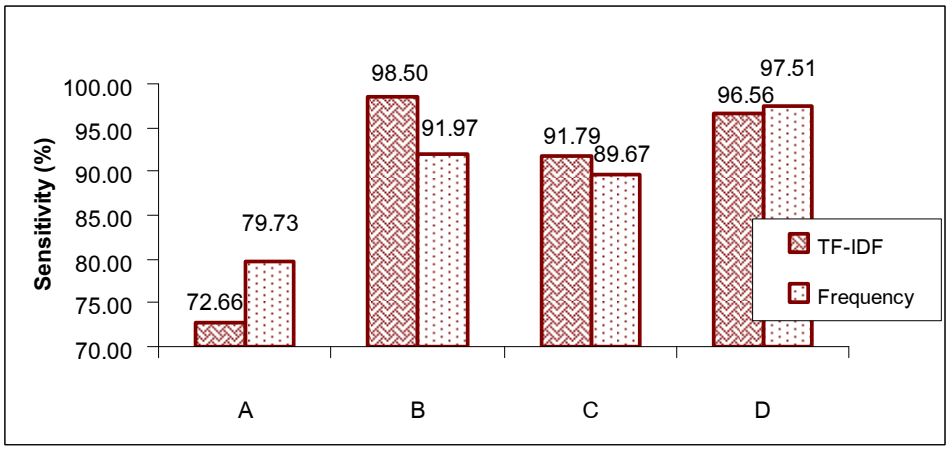


ภาพประกอบ 5.12 เปรียบเทียบค่าความเฉพาะเจาะจงของการทดลองแบบ A B C และ D ระหว่างค่าความถี่ กับ ค่า TF-IDF สำหรับชุดข้อมูล Vertebrate

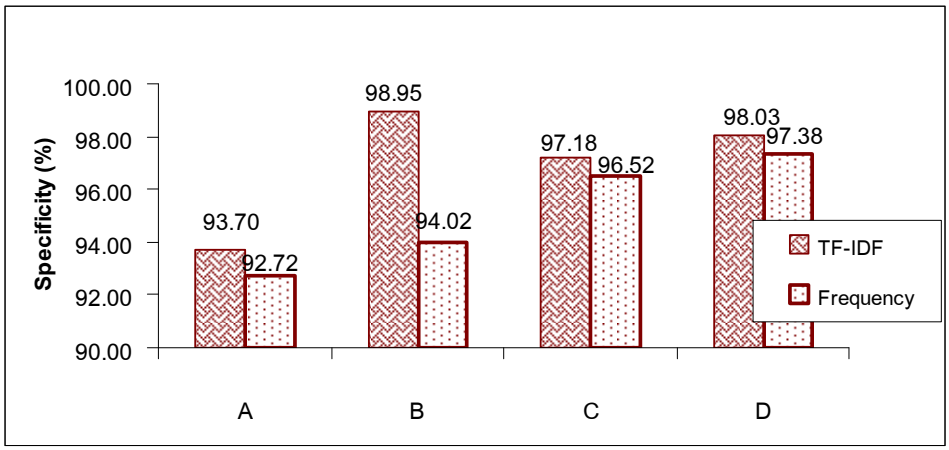
สำหรับชุดข้อมูล A.thaliana ของการทดลองแบบ A B C และ D ค่าความถูกต้องสำหรับค่า TF-IDF เท่ากับ 88.33% 97.53% 95.80% และ 97.56% ตามลำดับ สูงกว่าค่าความถี่ซึ่งให้ค่าความถูกต้องเท่ากับ 88.00% 93.52% 94.78% และ 97.41% ตามลำดับ แสดงดังภาพประกอบ 5.13 ค่าการตอบสนองไวสำหรับค่า TF-IDF เท่ากับ 72.66% 98.50% 91.79% และ 97.51% ตามลำดับ และค่าความถี่ให้ค่าการตอบสนองไวเท่ากับ 79.73% 91.97% 89.67% และ 96.65% ตามลำดับ แสดงดังภาพประกอบ 5.14 และค่าความเฉพาะเจาะจงสำหรับค่า TF-IDF เท่ากับ 93.70% 98.95% 97.18% และ 98.03% ตามลำดับ สูงกว่าค่าความถี่ซึ่งให้ค่าความเฉพาะเจาะจงเท่ากับ 92.72% 94.02% 96.52% และ 97.38% ตามลำดับ แสดงดังภาพประกอบ 5.15



ภาพประกอบ 5.13 เปรียบเทียบค่าความถูกต้องของการทดลองแบบ A B C และ D ระหว่างค่าความถี่ และค่า TF-IDF สำหรับชุดข้อมูล A.thaliana

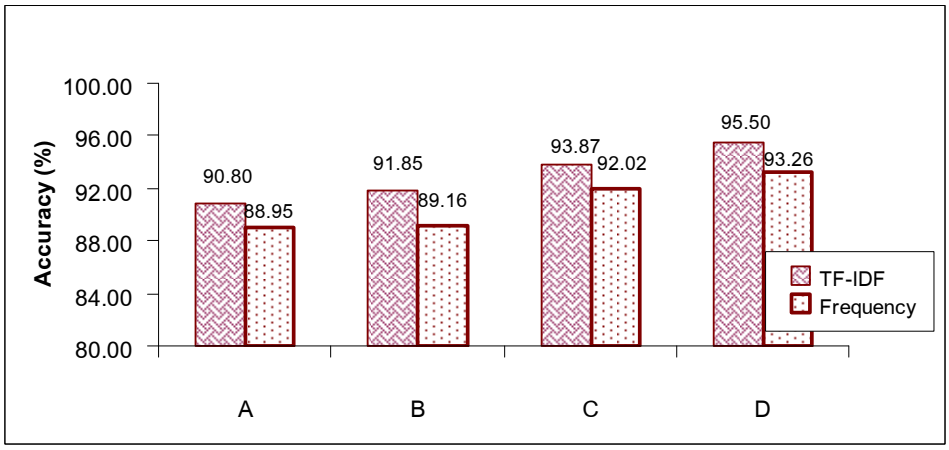


ภาพประกอบ 5.14 เปรียบเทียบค่าการตอบสนองไวของการทดลองแบบ A B C และ D ระหว่าง ค่าความถี่ กับ ค่า TF-IDF ของชุดข้อมูล A.thaliana

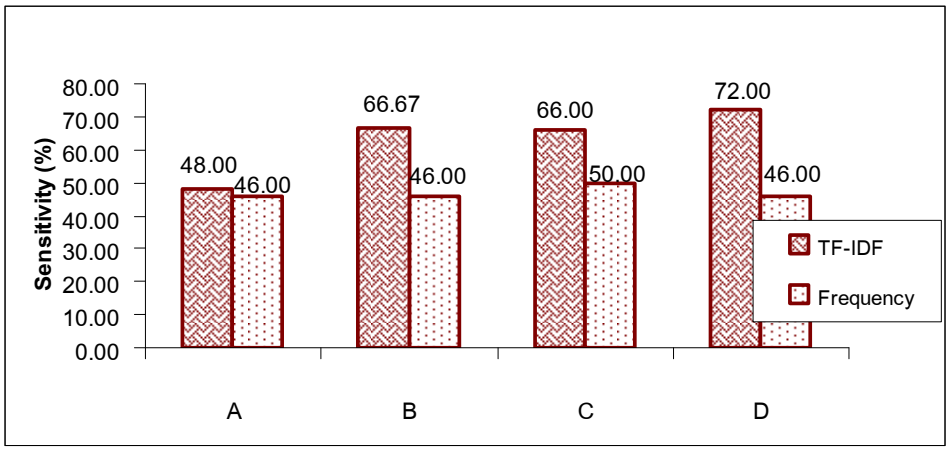


ภาพประกอบ 5.15 เปรียบเทียบค่าความเฉพาะเจาะจงของการทดลองแบบ A B C และ D ระหว่างค่าความถี่ กับ ค่า TF-IDF สำหรับชุดข้อมูล A.thaliana

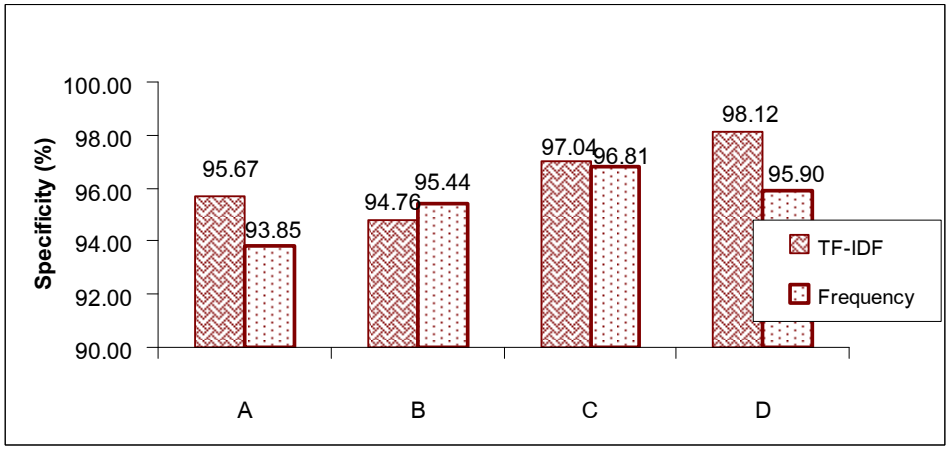
สำหรับชุดข้อมูล TIS+50 ของการทดลองแบบ A B C และ D ค่าความถูกต้องสำหรับค่า TF-IDF เท่ากับ 90.80% 91.85% 93.87% และ 95.50% ตามลำดับ สูงกว่าค่าความถี่ซึ่งให้ค่าความถูกต้องเท่ากับ 88.00% 93.52% 94.78% และ 97.41% ตามลำดับ แสดงดังภาพประกอบ 5.16 ค่าการตอบสนองไวสำหรับค่า TF-IDF เท่ากับ 48.00% 66.67% 66.00% และ 72.00% ตามลำดับ สูงกว่าค่าความถี่ซึ่งให้ค่าการตอบสนองไวเท่ากับ 46.00% 46.00% 50.00% และ 46.00% ตามลำดับ แสดงดังภาพประกอบ 5.17 และค่าความเฉพาะเจาะจงสำหรับค่า TF-IDF เท่ากับ เท่ากับ 95.67% 94.76% 97.04% และ 98.12% ตามลำดับ และค่าความถี่ให้ค่าความเฉพาะเจาะจงเท่ากับ 93.85% 95.44% 96.81% และ 95.90% ตามลำดับ แสดงดังภาพประกอบ 5.18



ภาพประกอบ 5.16 เปรียบเทียบค่าความถูกต้องของการทดลองแบบ A B C และ D ระหว่างค่าความถี่ กับ ค่า TF-IDF สำหรับชุดข้อมูล TIS+50



ภาพประกอบ 5.17 เปรียบเทียบค่าการตอบสนองไวของการทดลองแบบ A B C และ D ระหว่าง ค่าความถี่ กับ ค่า TF-IDF ของชุดข้อมูล TIS+50



ภาพประกอบ 5.18 เปรียบเทียบค่าความเฉพาะเจาะจงของการทดลองแบบ A B C และ D ระหว่างค่าความถี่ กับ ค่า TF-IDF สำหรับชุดข้อมูล TIS+50

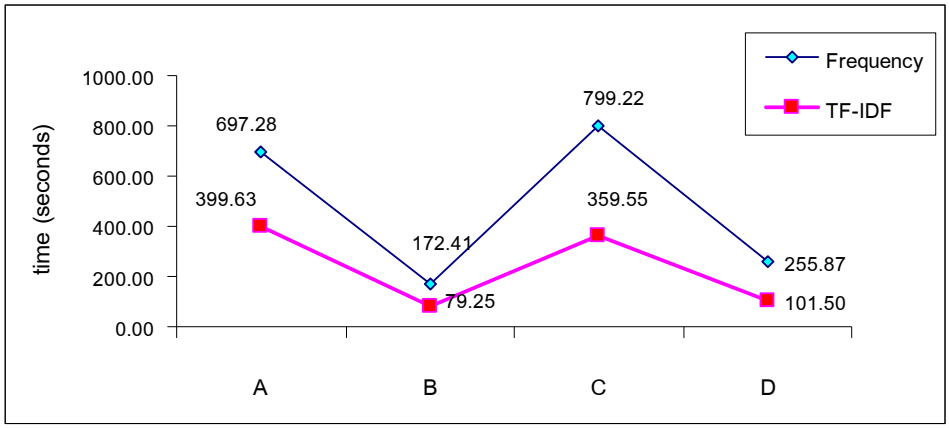
2) ประเด็นประสิทธิภาพของเวลาสำหรับค่า TF-IDF

พิจารณาระหว่างค่า TF-IDF และค่าความถี่ ผลการทดลองแสดงให้เห็นว่าโครงข่ายประสาทเทียมของค่า TF-IDF ใช้เวลาการสร้างแบบจำลองน้อยกว่าโครงข่ายประสาทเทียมของค่าความถี่สำหรับชุดข้อมูล Vertebrate ชุดข้อมูล A.thaliana และชุดข้อมูล TIS+50 ตัวอย่างเช่น แบ่งสายพันธุกรรมด้วยขนาดหน้าต่าง 303 นิวคลีโอไทด์ สร้างลักษณะเฉพาะ 1-แกรม และ 2-แกรม เลือกลักษณะเฉพาะด้วยเทคนิค CFS

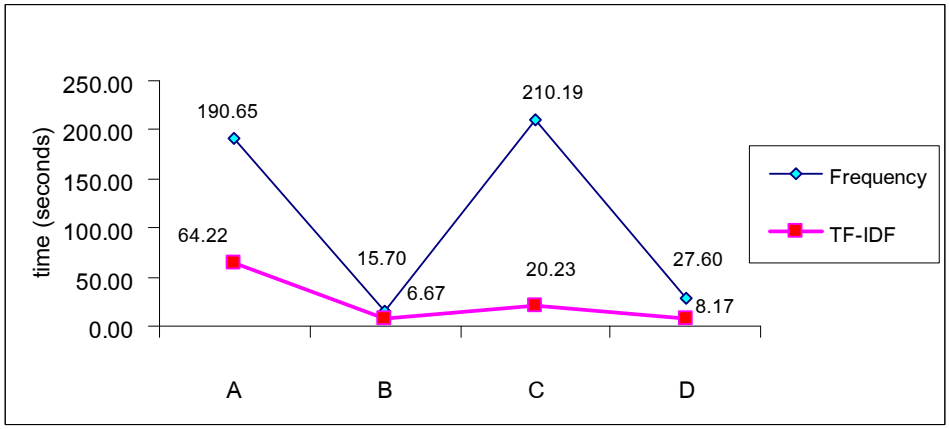
สำหรับชุดข้อมูล Vertebrate การทดลองแบบ A B C และ D ใช้เวลาการสร้างแบบจำลองสำหรับค่า TF-IDF เท่ากับ 399.63 79.25 359.55 และ 101.50 วินาที ตามลำดับ น้อยกว่าเวลาการสร้างแบบจำลองสำหรับค่าความถี่เท่ากับ 697.28 172.41 799.22 และ 255.87 วินาที ตามลำดับ แสดงดังภาพประกอบ 5.19

สำหรับชุดข้อมูล A.thaliana การทดลองแบบ A B C และ D ใช้เวลาการสร้างแบบจำลองสำหรับค่า TF-IDF เท่ากับ 64.22 6.67 20.23 และ 8.17 วินาที ตามลำดับ น้อยกว่าเวลาการสร้างแบบจำลองสำหรับค่าความถี่เท่ากับ 190.65 15.70 210.19 และ 27.60 วินาที ตามลำดับ แสดงดังภาพประกอบ 5.20

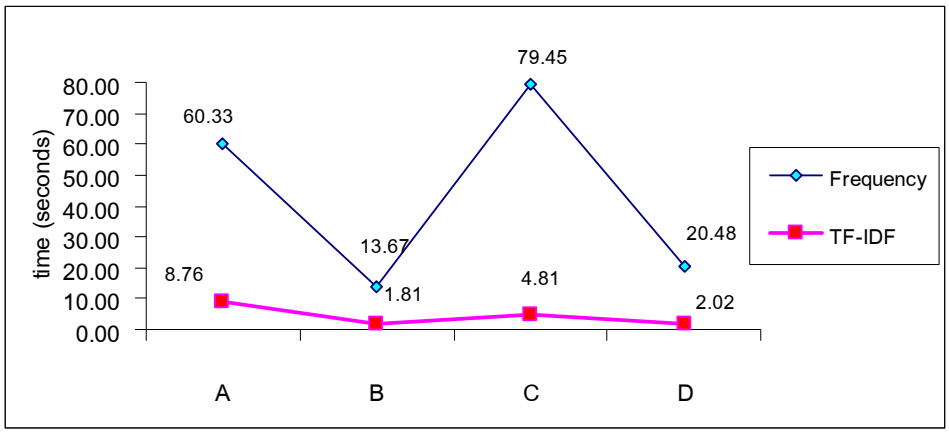
สำหรับชุดข้อมูล TIS+50 การทดลองแบบ A B C และ D ใช้การสร้างแบบจำลองสำหรับค่า TF-IDF เท่ากับ 8.76 1.81 4.81 และ 2.02 วินาที ตามลำดับ น้อยกว่าเวลาการสร้างแบบจำลองสำหรับค่าความถี่เท่ากับ 60.33 13.67 79.45 และ 20.48 วินาที ตามลำดับ แสดงดังภาพประกอบ 5.21



ภาพประกอบ 5.19 เปรียบเทียบเวลาการสร้างแบบจำลองระหว่างค่าความถี่กับค่า TF-IDF สำหรับชุดข้อมูล Vertebrate



ภาพประกอบ 5.20 เปรียบเทียบเวลาการสร้างแบบจำลองระหว่างค่าความถี่กับค่า TF-IDF สำหรับชุดข้อมูล A.thaliana



ภาพประกอบ 5.21 เปรียบเทียบเวลาการสร้างแบบจำลองระหว่าง
ค่าความถี่กับค่า TF-IDF สำหรับชุดข้อมูล TIS+50

3) ประเด็นประสิทธิภาพของเทคนิคการเลือกลักษณะเฉพาะ

การเลือกลักษณะเฉพาะสามารถเพิ่มประสิทธิภาพการทำนายผลลัพธ์ และลดเวลาการประมวลผลโดยสามารถสรุปเป็น 2 ประเด็นย่อย คือ 1) ประเด็นจำนวนลักษณะเฉพาะที่เหมาะสม และ 2) ประเด็นประสิทธิภาพเทคนิคการเลือกลักษณะเฉพาะ

ประเด็นย่อยที่ 1 ประสิทธิภาพของจำนวนลักษณะเฉพาะ

เนื่องจากเทคนิคไคสแควร์ เทคนิคอัตราส่วนเกน และเทคนิครีลีฟ-เอฟ เป็นเทคนิคการเลือกลักษณะเฉพาะแบบตัวกรอง จึงต้องระบุจำนวนลักษณะเฉพาะที่ใช้เป็นข้อมูลเข้าในโครงข่ายประสาทเทียม ผลการทดลองแสดงให้เห็นว่าจำนวนลักษณะเฉพาะ 5 ถึง 15 ลำดับแรกที่มีนัยสำคัญสูงสุดให้ค่าความถูกต้องสูงสำหรับชุดข้อมูล Vertebrate ชุดข้อมูล A.thaliana และชุดข้อมูล TIS+50 ตัวอย่างเช่น แบ่งสายพันธุกรรมด้วยขนาดหน้าต่าง 303 นิวคลีโอไทด์ สร้างลักษณะเฉพาะ 1-แกรม และ 2-แกรม กำหนดค่าลักษณะเฉพาะด้วยค่า TF-IDF

สำหรับชุดข้อมูล Vertebrate ค่าความถูกต้องของการทดลองแบบ B เมื่อเลือกลักษณะเฉพาะด้วยเทคนิคไคสแควร์จำนวนลักษณะเฉพาะ 15 ลำดับแรกให้ค่าความถูกต้องสูงสุดเท่ากับ 99.64% เทคนิคอัตราส่วนเกนจำนวนลักษณะเฉพาะ 15 ลำดับแรกให้ค่าความถูกต้องสูงสุดเท่ากับ 99.62% และเทคนิคเทคนิครีลีฟ-เอฟจำนวนลักษณะเฉพาะเท่ากับ 20 ลำดับแรกให้ค่าความถูกต้องสูงสุดเท่ากับ 99.65% แสดงดังตารางที่ 5.20

ตารางที่ 5.20 ค่าความถูกต้องของจำนวนลักษณะเฉพาะที่มีลำดับนัยสำคัญแตกต่างกันของเทคนิคการเลือกลักษณะเฉพาะแบบตัวกรองสำหรับข้อมูล Vertebrate

จำนวน ลักษณะเฉพาะ	ค่าความถูกต้อง (%)		
	เทคนิคไคสแควร์	เทคนิคอัตราส่วนเกน	เทคนิครีลีฟ-เอฟ
5	97.65	97.55	99.57
10	99.33	99.27	99.53
15	**99.64**	**99.62**	99.54
20	98.59	99.34	**99.65**
30	99.30	99.03	98.99

สำหรับชุดข้อมูล A.thaliana ค่าความถูกต้องของการทดลองแบบ B เมื่อเลือกลักษณะเฉพาะด้วยเทคนิคไคสแควร์จำนวนลักษณะเฉพาะ 5 ลำดับแรกให้ค่าความถูกต้องสูงสุดเท่ากับ 99.27% เทคนิคอัตราส่วนเกนจำนวนลักษณะเฉพาะ 5 ลำดับแรกให้ค่าความถูกต้องสูงสุดเท่ากับ 99.22% และเทคนิครีลีฟ-เอฟจำนวนลักษณะเฉพาะ 5 ลำดับแรกให้ค่าความถูกต้องสูงสุดเท่ากับ 99.27% แสดงดังตารางที่ 5.21

ตารางที่ 5.21 ค่าความถูกต้องของจำนวนลักษณะเฉพาะที่มีลำดับนัยสำคัญแตกต่างกันของเทคนิคการเลือกลักษณะเฉพาะแบบตัวกรองสำหรับข้อมูล A.thaliana

จำนวน ลักษณะเฉพาะ	ค่าความถูกต้อง (%)		
	เทคนิคไคสแควร์	เทคนิคอัตราส่วนเกน	เทคนิครีลีฟ-เอฟ
5	**99.27**	**99.22**	**99.27**
10	97.86	98.78	99.12
15	94.53	96.48	**99.27**
20	97.56	95.75	96.39
30	88.18	91.21	90.77

สำหรับชุดข้อมูล TIS+50 ค่าความถูกต้องของการทดลองแบบ B เมื่อเลือกลักษณะเฉพาะด้วยเทคนิคไคสแควร์จำนวนลักษณะเฉพาะเท่ากับ 10 ให้ค่าความถูกต้องสูงสุดเท่ากับ 90.18% เทคนิคอัตราส่วนเกนจำนวนลักษณะเฉพาะเท่ากับ 10 ให้ค่าความถูกต้องสูงสุดเท่ากับ 90.18% และเทคนิครีลีฟ-เอฟจำนวนลักษณะเฉพาะเท่ากับ 15 ให้ค่าความถูกต้องสูงสุดเท่ากับ 90.36% แสดงดังตารางที่ 5.22

ตารางที่ 5.22 ค่าความถูกต้องของจำนวนลักษณะเฉพาะที่มีลำดับนัยสำคัญแตกต่างกันของเทคนิคการเลือกลักษณะเฉพาะแบบตัวกรองสำหรับข้อมูล TIS+50

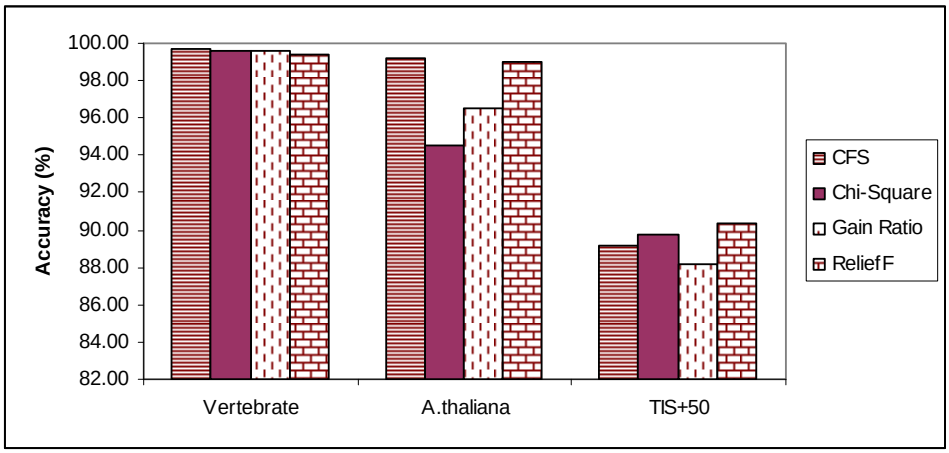
จำนวนลักษณะเฉพาะ	ค่าความถูกต้อง (%)		
	เทคนิคไคสแควร์	เทคนิคอัตราส่วนเกน	เทคนิครีลีฟ-เอฟ
5	90.39	87.73	89.37
10	**90.18**	**90.18**	89.78
15	89.78	88.14	**90.36**
20	89.57	89.98	89.78
30	88.96	89.57	89.57

ประเด็นย่อยที่ 2 การเปรียบเทียบประสิทธิภาพเทคนิคการเลือกลักษณะเฉพาะ ผลการทดลองแสดงให้เห็นว่าเทคนิค CFS ช่วยเพิ่มประสิทธิภาพการทำนายจุดเริ่มต้นการแปลรหัส สำหรับชุดข้อมูล Vertebrate ชุดข้อมูล A.thaliana และชุดข้อมูล TIS+50 ตัวอย่างเช่น การแบ่งสายพันธุกรรมด้วยขนาดหน้าต่าง 303 นิวคลีโอไทด์ สร้างลักษณะเฉพาะ 1-แกรม และ 2-แกรม กำหนดค่าลักษณะเฉพาะด้วยค่า TF-IDF เทคนิคไคสแควร์ เทคนิคอัตราส่วนเกน และเทคนิครีลีฟ-เอฟ เลือกจำนวนลักษณะเฉพาะ 15 ลำดับแรกที่มีนัยสำคัญสูงสุด

พิจารณาค่าความถูกต้องของการทดลองแบบ B ผลการทดลองแสดงให้เห็นว่า ชุดข้อมูล Vertebrate เทคนิค CFS ให้ค่าความถูกต้องสูงสุดเท่ากับ 99.70% ชุดข้อมูล A.thaliana เทคนิค CFS ให้ค่าความถูกต้องสูงสุดเท่ากับ 99.22% และชุดข้อมูล TIS+50 เทคนิครีลีฟ-เอฟให้ค่าความถูกต้องสูงสุดเท่ากับ 90.39% แสดงดังตารางที่ 5.23 และภาพประกอบ 5.22

ตารางที่ 5.23 ค่าความถูกต้องของเทคนิคการเลือกลักษณะเฉพาะ

ชื่อชุดข้อมูล	ค่าความถูกต้อง (%)			
	เทคนิค CFS	เทคนิคไคสแควร์	เทคนิคอัตราส่วนเกน	เทคนิครีลีฟ-เอฟ
Vertebrate	**99.70**	99.64	99.62	99.45
A.thaliana	**99.22**	94.53	96.48	99.02
TIS+50	89.16	89.78	88.14	**90.39**

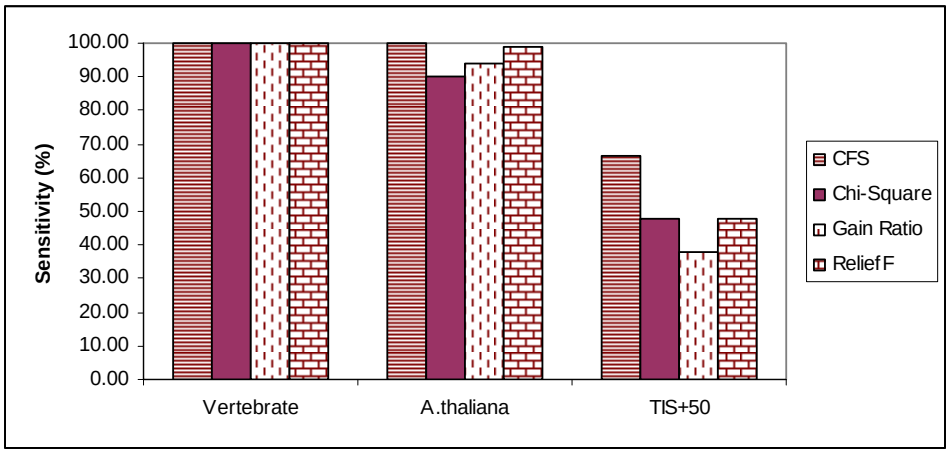


ภาพประกอบ 5.22 เปรียบเทียบค่าความถูกต้องของเทคนิคการเลือกลักษณะเฉพาะ

พิจารณาค่าการตอบสนองไวของการทดลองแบบ B ชุดข้อมูล Vertebrate เทคนิคไคสแควร์ เทคนิคอัตราส่วนเกน และเทคนิครีลีฟ-เอฟ ให้ค่าการตอบสนองไวสูงสุดเท่ากับ 100.00% ชุดข้อมูล A.thaliana เทคนิค CFS ให้ค่าการตอบสนองไวสูงสุดเท่ากับ 100.00% และชุดข้อมูล TIS+50 เทคนิค CFS ให้ค่าความถูกต้องสูงสุดเท่ากับ 66.67% แสดงดังตารางที่ 5.24 และภาพประกอบ 5.23

ตารางที่ 5.24 ค่าการตอบสนองไวของเทคนิคการเลือกลักษณะเฉพาะ

ชื่อชุดข้อมูล	ค่าการตอบสนองไว (%)			
	เทคนิค CFS	เทคนิคไคสแควร์	เทคนิคอัตราส่วนเกน	เทคนิครีลีฟ-เอฟ
Vertebrate	99.94	100.00	100.00	100.00
A.thaliana	**100.00**	90.23	93.69	99.04
TIS+50	66.67	48.00	38.00	48.00

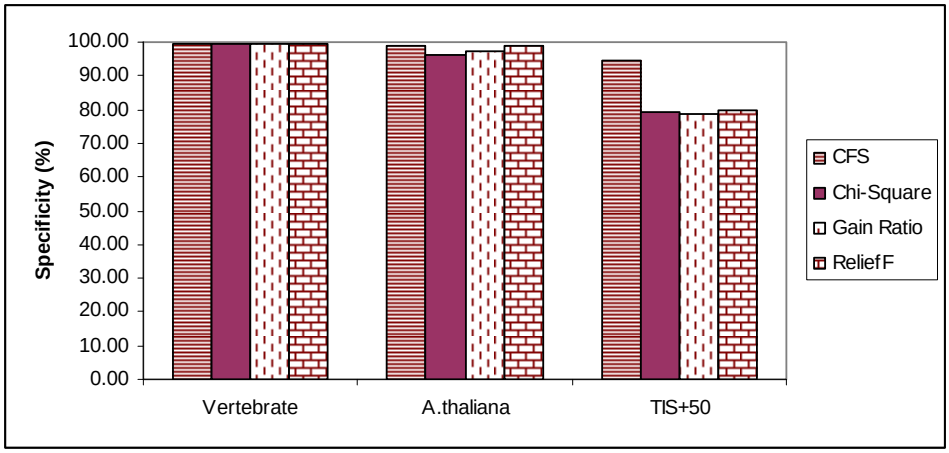


ภาพประกอบ 5.23 เปรียบเทียบค่าการตอบสนองไวของเทคนิคการเลือกลักษณะเฉพาะ

พิจารณาค่าความเฉพาะเจาะจงแสดงดังตารางที่ 5.25 และภาพประกอบ 5.24 สำหรับชุดข้อมูล TIS+50 การทดลองแบบ B เทคนิค CFS ให้ค่าความเฉพาะเจาะจงเท่ากับ 94.76% สูงกว่าเทคนิคไคสแควร์ เทคนิคอัตราส่วนเกน และเทคนิครีลีฟ-เอฟ ซึ่งให้ค่าความเฉพาะเจาะจงเท่ากับ 79.35% 78.78% และ 79.92% ตามลำดับ

ตารางที่ 5.25 ค่าความเฉพาะเจาะจงของเทคนิคการเลือกลักษณะเฉพาะ

ชื่อชุดข้อมูล	ค่าความเฉพาะเจาะจง (%)			
	เทคนิค CFS	เทคนิคไคสแควร์	เทคนิคอัตราส่วนเกน	เทคนิครีลีฟ-เอฟ
Vertebrate	99.63	99.53	99.50	99.27
A.thaliana	98.95	96.00	97.44	99.02
TIS+50	**94.76**	79.35	78.78	79.92



ภาพประกอบ 5.24 เปรียบเทียบค่าความเฉพาะเจาะจงของเทคนิคการเลือกลักษณะเฉพาะ

4) ประเด็นประสิทธิภาพของลักษณะเฉพาะที่สร้างจากรูปแบบคอนเซนซัส

ลักษณะเฉพาะที่สร้างจากรูปแบบคอนเซนซัสช่วยเพิ่มประสิทธิภาพการทำนายผลลัพธ์ของโครงข่ายประสาทเทียม ผลการทดลองแสดงให้เห็นว่าลักษณะเฉพาะที่สร้างจากรูปแบบคอนเซนซัสให้ประสิทธิภาพสูง โดยเฉพาะอย่างยิ่งสำหรับชุดข้อมูล TIS+50 ซึ่งแต่ละสายพันธุกรรมจะมีจุดเริ่มต้นการแปลรหัส และรหัสหยุด (โคดอน TAA TAG หรือ TGA) โดยแบ่งการพิจารณาเป็น 2 กรณี คือ กรณีที่ 1 พิจารณาระหว่างการทดลองแบบ C กับการทดลองแบบ A ซึ่งปราศจากการเลือกลักษณะเฉพาะ และ กรณีที่ 2 พิจารณาระหว่างการทดลองแบบ D กับ แบบ B ซึ่งมีการเลือกลักษณะเฉพาะ โดยการทดลองแบบ C และ D มีลักษณะเฉพาะที่สร้างจากรูปแบบคอนเซนซัส และการทดลองแบบ A และ B ไม่มีลักษณะเฉพาะที่สร้างจากรูปแบบคอนเซนซัส ผลการทดลองแสดงให้เห็นว่าการทดลองแบบ C

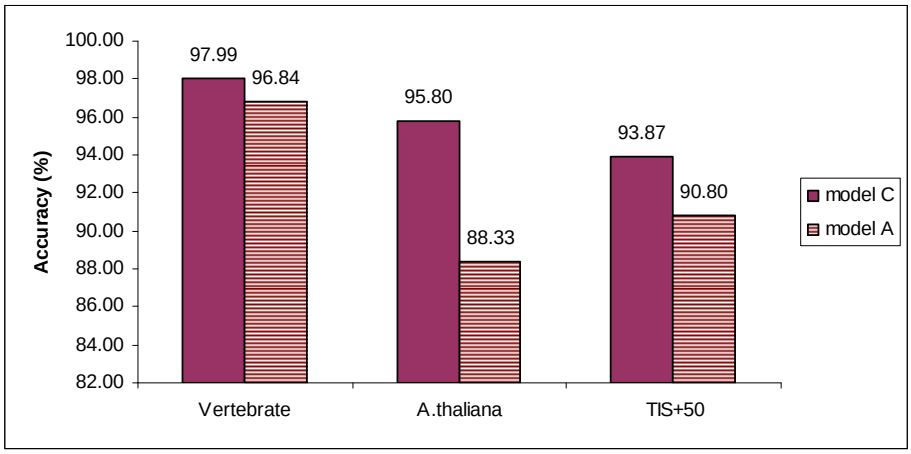
ให้ประสิทธิภาพสูงกว่าการทดลองแบบ A และการทดลองแบบ D ให้ประสิทธิภาพสูงกว่าการทดลองแบบ B

กรณีที่ 1 พิจารณาระหว่างการทดลองแบบ C กับการทดลองแบบ A ซึ่งปราศจากการเลือกลักษณะเฉพาะ ตัวอย่างเช่น แบ่งสายพันธุกรรมด้วยขนาดหน้าต่าง 303 นิวคลีโอไทด์ สร้างลักษณะเฉพาะ 1-แกรม และ 2-แกรม กำหนดค่าลักษณะเฉพาะด้วยค่า TF-IDF เลือกลักษณะเฉพาะด้วยเทคนิค CFS

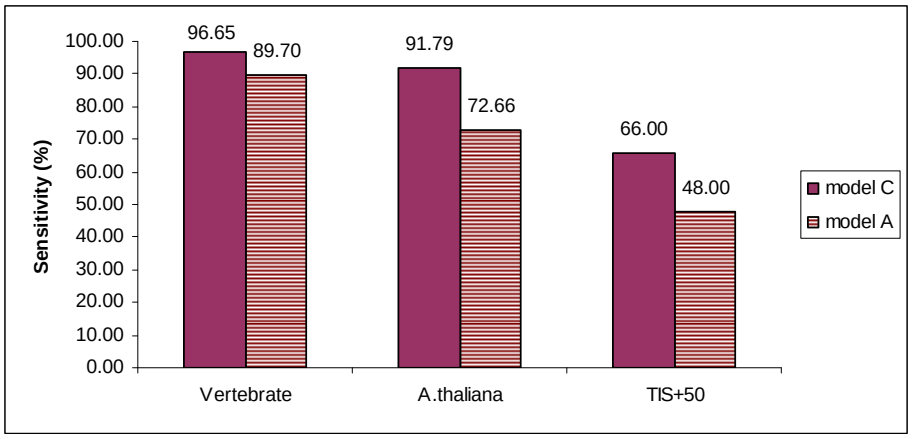
พิจารณาค่าความถูกต้อง สำหรับชุดข้อมูล Vertebrate ชุดข้อมูล A.thaliana และชุดข้อมูล TIS+50 พบว่า การทดลองแบบ C ให้ค่าความถูกต้องเท่ากับ 97.99% 95.80% และ 93.87% สูงกว่าการทดลองแบบ A ซึ่งให้ค่าความถูกต้องเท่ากับ 96.84% 88.33% และ 90.80% ตามลำดับ แสดงดังภาพประกอบ 5.25

พิจารณาค่าการตอบสนองไวสำหรับชุดข้อมูล Vertebrate ชุดข้อมูล A.thaliana และชุดข้อมูล TIS+50 พบว่า การทดลองแบบ C ให้ค่าการตอบสนองไวเท่ากับ 96.65% 91.79% และ 66.00% สูงกว่าการทดลองแบบ A ซึ่งให้ค่าการตอบสนองไวเท่ากับ 89.70% 72.66% และ 48.00% ตามลำดับ แสดงดังภาพประกอบ 5.26

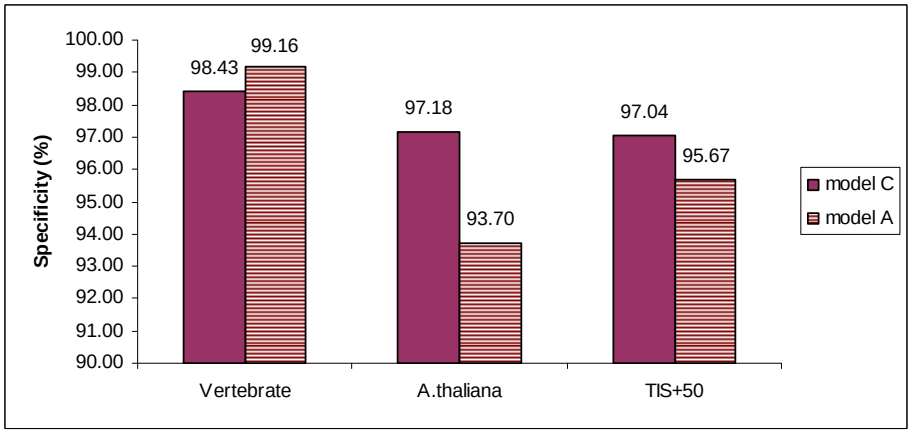
พิจารณาค่าความเฉพาะเจาะจงสำหรับชุดข้อมูล Vertebrate ชุดข้อมูล A.thaliana และชุดข้อมูล TIS+50 พบว่าการทดลองแบบ C ให้ค่าความเฉพาะเจาะจงเท่ากับ 99.16% 97.18% และ 97.04% และการทดลองแบบ A ให้ค่าความเฉพาะเจาะจงเท่ากับ 98.43% 93.70% และ 95.67% ตามลำดับ แสดงดังภาพประกอบ 5.27



ภาพประกอบ 5.25 เปรียบเทียบค่าความถูกต้องระหว่างการทดลองแบบ A กับการทดลองแบบ C



ภาพประกอบ 5.26 เปรียบเทียบค่าการตอบสนองไวระหว่าง
การทดลองแบบ A กับการทดลองแบบ C



ภาพประกอบ 5.27 เปรียบเทียบค่าความเฉพาะเจาะจง
ระหว่างการทดลองแบบ A กับการทดลองแบบ C

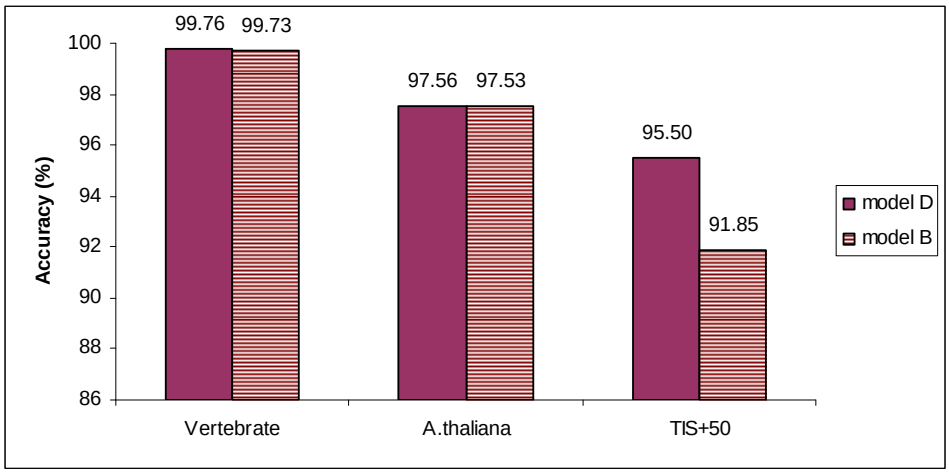
กรณีที่ 2 พิจารณาระหว่างการทดลองแบบ D กับการทดลองแบบ B มีการเลือกลักษณะเฉพาะ ตัวอย่างเช่น แบ่งสายพันธุกรรมด้วยขนาดหน้าต่าง 303 นิวคลีโอไทด์ สร้างลักษณะเฉพาะ 1-แกรม และ 2-แกรม กำหนดค่าลักษณะเฉพาะด้วยค่า TF-IDF เลือกลักษณะเฉพาะด้วยเทคนิค CFS

พิจารณาค่าความถูกต้อง สำหรับชุดข้อมูล Vertebrate ชุดข้อมูล A.thaliana และชุดข้อมูล TIS+50 พบว่า การทดลองแบบ D ให้ค่าความถูกต้องเท่ากับ 99.76% 97.56% และ 95.50% ตามลำดับ สูงกว่าการทดลองแบบ B ซึ่งให้ค่าความถูกต้องเท่ากับ 99.73% 97.53% และ 91.85% ตามลำดับ แสดงดังภาพประกอบ 5.28

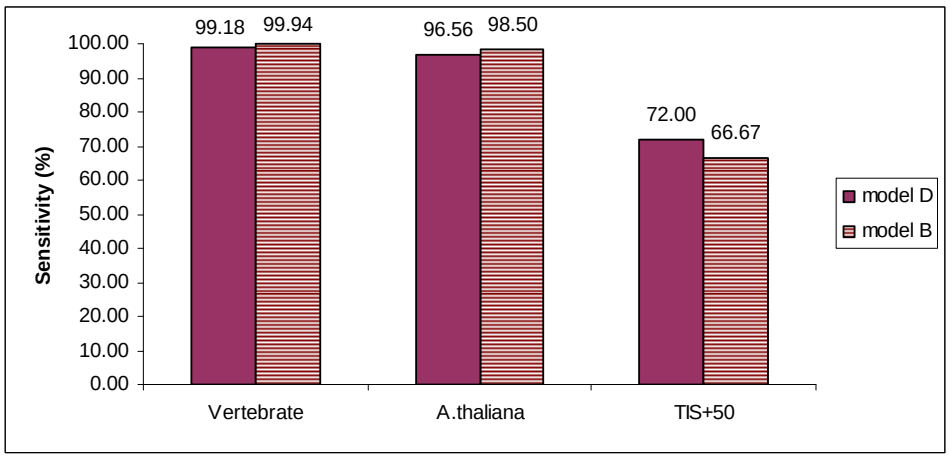
พิจารณาค่าการตอบสนองไว สำหรับชุดข้อมูล Vertebrate ชุดข้อมูล A.thaliana และชุดข้อมูล TIS+50 พบว่า การทดลองแบบ D ให้ค่าการตอบสนองไว

เท่ากับ 99.18% 96.54% และ 72.00% ตามลำดับ สูงกว่าการทดลองแบบ B ซึ่งให้ค่าการตอบสนองไวเท่ากับ 99.94% 98.50% และ 66.67% ตามลำดับ แสดงดังภาพประกอบ 5.29

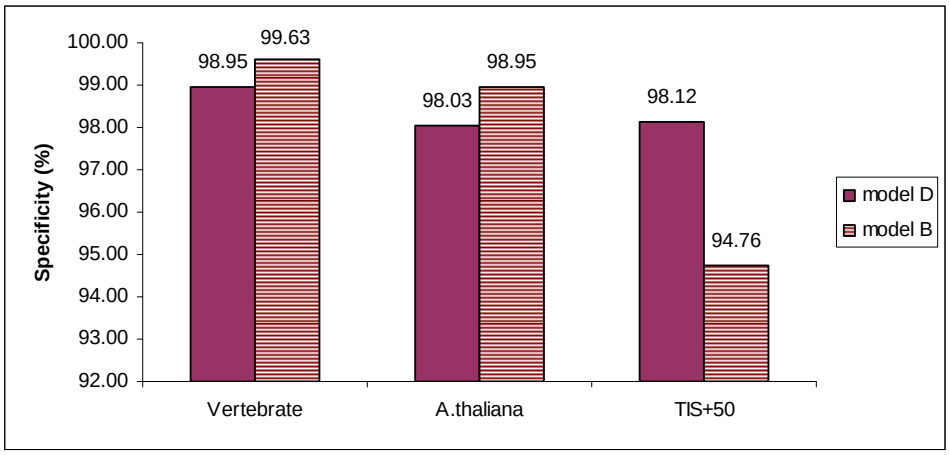
พิจารณาค่าความเฉพาะเจาะจง สำหรับชุดข้อมูล Vertebrate ชุดข้อมูล A.thaliana และชุดข้อมูล TIS+50 พบว่า การทดลองแบบ D ให้ค่าความเฉพาะเจาะจงเท่ากับ 98.95% 98.03% และ 98.12% ตามลำดับ สูงกว่าการทดลองแบบ B ซึ่งให้ค่าความเฉพาะเจาะจงเท่ากับ 99.63% 98.95% และ 94.76% ตามลำดับ แสดงดังภาพประกอบ 5.30



ภาพประกอบ 5.28 เปรียบเทียบค่าความถูกต้องระหว่างการทดลองแบบ B กับการทดลองแบบ D



ภาพประกอบ 5.29 เปรียบเทียบค่าการตอบสนองไวระหว่างการทดลองแบบ B กับการทดลองแบบ D



ภาพประกอบ 5.30 เปรียบเทียบค่าความเฉพาะเจาะจงระหว่าง
การทดลองแบบ B กับการทดลองแบบ D

5) ประเด็นประสิทธิภาพของแบบจำลอง TF-IDF-NN-TIS

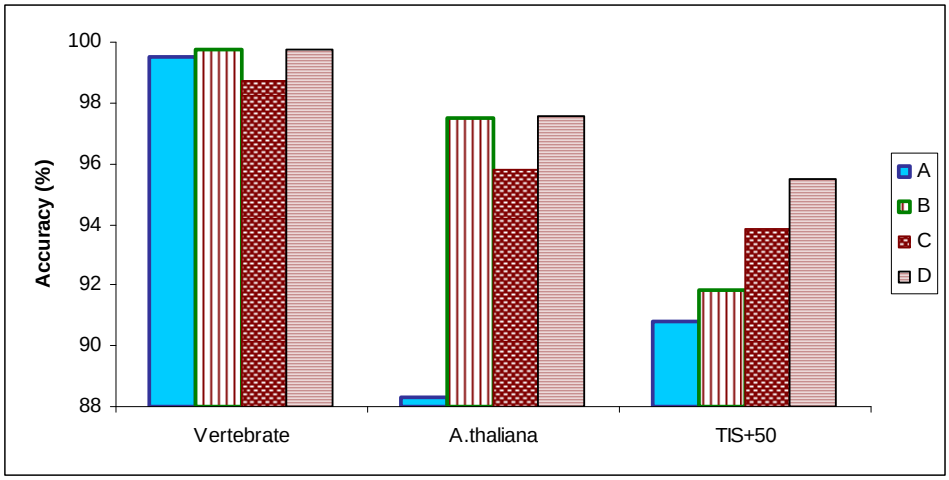
แบบจำลอง TF-IDF-NN-TIS ประกอบด้วยการทดลอง 4 แบบ คือ การทดลองแบบ A B C และ D แต่ละการทดลองจะมีข้อมูลเข้าในโครงข่ายประสาทเทียมแตกต่างกัน โดยการทดลองแบบ D เป็นการทดลองที่ใช้ลักษณะเฉพาะที่ผ่านการเลือกด้วยเทคนิคการเลือกลักษณะเฉพาะร่วมกับลักษณะเฉพาะรูปแบบคอนเซนซัส สำหรับการพิจารณาประเด็นประสิทธิภาพของแบบจำลอง TF-IDF-NN-TIS สรุปเป็น 2 ประเด็นย่อย คือ ประเด็นย่อยที่ 1 ประสิทธิภาพของแบบจำลอง TF-IDF-NN-TIS หรือการทดลองแบบ D และ ประเด็นย่อยที่ 2 เปรียบเทียบประสิทธิภาพของแบบจำลอง TF-IDF-NN-TIS กับงานวิจัยที่ศึกษาก่อนหน้า

ประเด็นย่อยที่ 1 ประสิทธิภาพของแบบจำลอง หรือการทดลองแบบ D ผลการทดลองแสดงให้เห็นว่าการทดลองแบบ D มีประสิทธิภาพการทำนายผลลัพธ์ของโครงข่ายประสาทเทียมสูงกว่าการทดลองแบบ A B และ C สำหรับชุดข้อมูล Vertebrate ชุดข้อมูล A.thaliana และชุดข้อมูล TIS+50 ตัวอย่างเช่น แบ่งสายพันธุกรรมด้วยขนาดหน้าต่าง 303 นิวคลีโอไทด์ สร้างลักษณะเฉพาะ 1-แกรม และ 2-แกรม กำหนดค่าลักษณะเฉพาะด้วยค่า TF-IDF เลือกลักษณะเฉพาะด้วยเทคนิค CFS

พิจารณาค่าความถูกต้องแสดงดังตารางที่ 5.26 และภาพประกอบ 5.31 สำหรับชุดข้อมูล TIS+50 ของการทดลองแบบ D ให้ค่าความถูกต้องเท่ากับ 95.50% สูงกว่าการทดลองแบบ A B และ C ซึ่งเท่ากับ 90.80% 89.16% และ 93.87% ตามลำดับ

ตารางที่ 5.26 ค่าความถูกต้องของการทดลองแบบ A B C และ D

ชื่อชุดข้อมูล	ค่าความถูกต้อง (%)			
	A	B	C	D
Vertebrate	99.53	99.73	98.73	99.76
A.thaliana	88.33	97.53	95.80	97.56
TIS+50	**90.80**	**89.16**	**93.87**	**95.50**

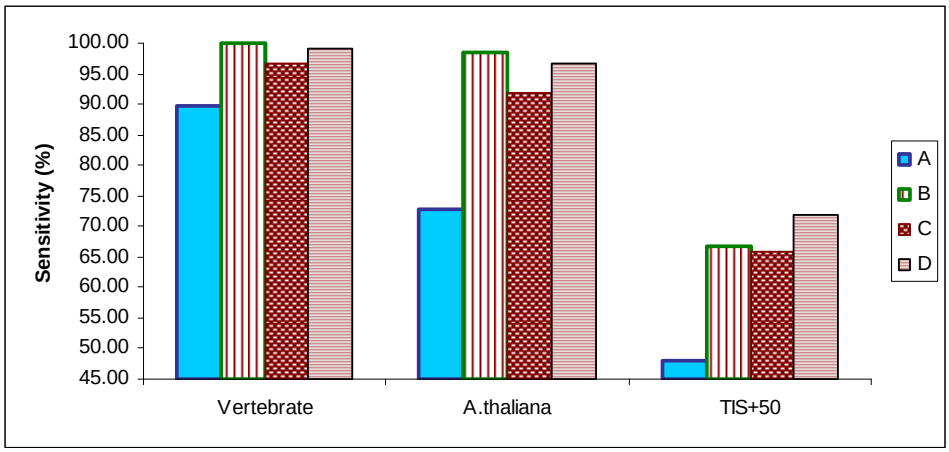


ภาพประกอบ 5.31 เปรียบเทียบค่าความถูกต้องระหว่างการทดลองแบบ A B C และ D

พิจารณาค่าตอบสนองไวแสดงดังตารางที่ 5.27 และภาพประกอบ 5.32 สำหรับชุดข้อมูล Vertebrate การทดลองแบบ D ให้ค่าการตอบสนองไวเท่ากับ 99.18% ซึ่งใกล้เคียงกับค่าการตอบสนองไวของการทดลองแบบ B ซึ่งเท่ากับ 99.94% สูงกว่าการทดลองแบบ A และ C ซึ่งให้ค่าการตอบสนองไวเท่ากับ 89.70% และ 96.65%

ตารางที่ 5.27 ค่าการตอบสนองไวของการทดลองแบบ A B C และ D

ชื่อชุดข้อมูล	ค่าการตอบสนองไว (%)			
	A	B	C	D
Vertebrate	**89.70**	**99.94**	**96.65**	**99.18**
A.thaliana	72.66	98.50	91.79	96.56
TIS+50	48.00	66.67	66.00	72.00

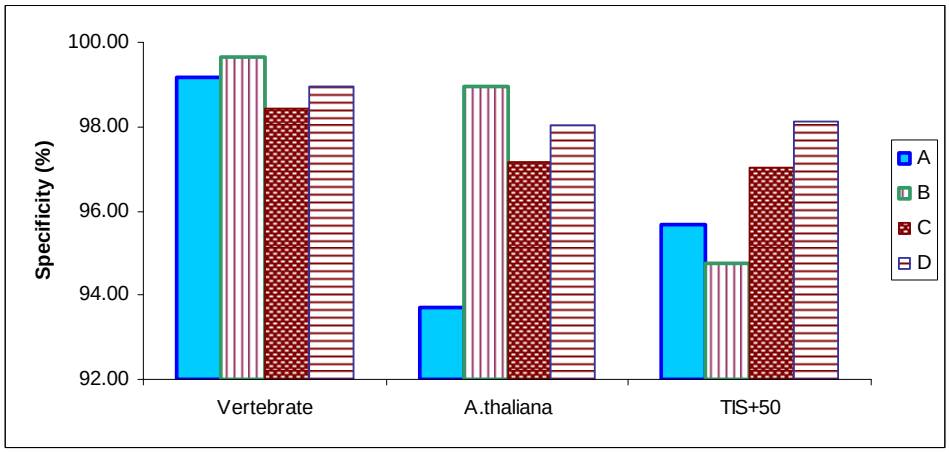


ภาพประกอบ 5.32 เปรียบเทียบค่าการตอบสนองไวระหว่างการทดลองแบบ A B C และ D

พิจารณาค่าความเฉพาะเจาะจงแสดงดังตารางที่ 5.28 และภาพประกอบ 5.33 สำหรับชุดข้อมูล TIS+50 การทดลองแบบ D ให้ค่าความเฉพาะเจาะจงเท่ากับ 98.12% สูงกว่าการทดลองแบบ A B และ C ซึ่งให้ค่าความเฉพาะเจาะจงเท่ากับ 95.67% 94.76% และ 97.04% ตามลำดับ

ตารางที่ 5.28 ค่าความเฉพาะเจาะจงของการทดลองแบบ A B C และ D

ชื่อชุดข้อมูล	ค่าความเฉพาะเจาะจง (%)			
	A	B	C	D
Vertebrate	99.16	99.63	98.43	98.95
A.thaliana	93.70	98.95	97.18	98.03
TIS+50	**95.67**	**94.76**	**97.04**	**98.12**



ภาพประกอบ 5.33 เปรียบเทียบค่าความเฉพาะเจาะจงระหว่างการทดลองแบบ A B C และ D

ตารางที่ 5.29 เปรียบเทียบค่าความถูกต้องระหว่างแบบจำลอง TF-IDF-NN-TIS กับ งานวิจัยที่ศึกษาก่อนหน้า

ชื่อนักวิจัย	ปีที่เผยแพร่	Vertebrate (%)	A.thaliana (%)	TIS+50 (%)
Pedersen และ Nielsen	1997	85.00	88.00	-
Zien และคณะ	2000	88.10	-	-
Liu และคณะ	2004	92.45	-	-
Rajapakse และ Ho	2005	96.10	-	-
Tzanis และคณะ	2007	97.26	97.07	-
Zeng และคณะ	2007	96.68	-	91.82
TF-IDF-NN-TIS	**2009**	**99.76**	**97.56**	**95.50**

ประเด็นย่อยที่ 2 เปรียบเทียบประสิทธิภาพการทำนายจุดเริ่มต้นการแปลรหัสระหว่างแบบจำลอง TF-IDF-NN-TIS กับงานวิจัยที่ศึกษาก่อนหน้า แสดงดังตารางที่ 5.29 เปรียบเทียบค่าความถูกต้องของแบบจำลอง TF-IDF-NN-TIS กับงานวิจัยที่ศึกษาก่อนหน้า (Pedersen and Neilsen, 1997; Zien et *al*., 2000; Liu et *al*., 2004; Rajapakse and Ho, 2005; Tzanis et *al*., 2007; Zeng and Alhajj, 2007) โดยแบบจำลอง TF-IDF-NN-TIS แบ่งสายพันธุกรรมด้วยขนาดหน้าต่าง 303 นิวคลีโอไทด์ สร้างลักษณะเฉพาะ 1-แกรม และ 2-แกรม กำหนดค่าลักษณะเฉพาะด้วยค่า TF-IDF ใช้ลักษณะเฉพาะ 1แกรม และ 2-แกรมที่ผ่านการเลือกด้วยเทคนิค CFS ร่วมกับลักษณะเฉพาะรูปแบบคอนเซนซัสเป็นข้อมูลเข้าสำหรับการทำนายจุดเริ่มต้นการแปลรหัสของโครงข่ายประสาทเทียม พบว่า สำหรับชุดข้อมูล Vertebrate แบบจำลองที่นำเสนอให้ค่าความถูกต้องเท่ากับ 99.76% สูงกว่างานวิจัยที่ศึกษาก่อนหน้าของ Tzanis และคณะ (Tzanis et *al*., 2007) ซึ่งเท่ากับ 97.26% สำหรับชุดข้อมูล A.thaliana ค่าความถูกต้องของแบบจำลองที่นำเสนอเท่ากับ 97.56% สูงกว่างานวิจัยที่ศึกษาก่อนหน้าของTzanis และคณะ (Tzanis et *al*., 2007) ซึ่งเท่ากับ 97.07% สำหรับชุดข้อมูล TIS+50 ค่าความถูกต้องของแบบจำลองเท่ากับ 95.50% สูงกว่างานวิจัยที่ศึกษาก่อนหน้าของนักวิจัย (Zeng et *al*., 2007) ซึ่งเท่ากับ 91.82% ดังนั้นแบบจำลอง TF-IDF-NN-TIS ให้ค่าความถูกต้องสูงกว่างานวิจัยที่ศึกษาก่อนหน้าสำหรับ ชุดข้อมูล Vertebrate ชุดข้อมูล A.thaliana และชุดข้อมูล TIS+50

บทที่ 6

บทสรุปและข้อเสนอแนะ

งานวิจัยนี้นำเสนอแบบจำลองการทำนายจุดเริ่มต้นการแปลรหัสโดยใช้วิธี TF-IDF และโครงข่ายประสาทเทียม (The TF-IDF and Neural Networks Approach for Translation Initiation Sites Prediction: TF-IDF-NN-TIS) ซึ่งใช้เทคนิคการสร้างลักษณะเฉพาะ n-แกรม กำหนดค่าลักษณะเฉพาะ TF-IDF ร่วมกับเทคนิคการเลือกลักษณะเฉพาะ จากผลการทดลอง จากผลการทดลองแสดงให้เห็นว่าแบบจำลองที่นำเสนอให้ค่าความถูกต้องสูง และใช้เวลาในการประมวลผลน้อย

6.1 สรุปผลงานวิจัย

งานวิจัยนี้ได้บรรลุตามวัตถุประสงค์โดยมีการออกแบบแบบจำลองการทำนายจุดเริ่มต้นการแปลรหัสโดยใช้วิธี TF-IDF และโครงข่ายประสาทเทียม ซึ่งเป็นแบบจำลองการวิเคราะห์สายจีโนมโดยใช้โครงข่ายประสาทเทียม ที่มุ่งเน้นเฉพาะส่วนของการระบุจุดเริ่มต้นการแปลรหัสในสายจีโนม แบบจำลองที่นำเสนอมีขั้นตอนทั้งหมด 5 ขั้นตอน คือ 1) ขั้นตอนการแบ่งสายพันธุกรรม 2) ขั้นตอนการสร้างลักษณะเฉพาะ n-แกรม 3) ขั้นตอนการเลือกลักษณะเฉพาะ 4) ขั้นตอนการสร้างลักษณะเฉพาะรูปแบบคอนเซนซัส และ 5) ขั้นตอนการทำนายผลลัพธ์ พัฒนาโปรแกรมตามแบบจำลองที่ได้ออกแบบไว้โดยใช้โปรแกรม MATLAB การทดลองแบ่ง 4 แบบ คือ การทดลองแบบ A B C และ D การทดลองแต่ละแบบมีชั้นข้อมูลเข้าของสถาปัตยกรรมโครงข่ายประสาทเทียมแบบไปข้างหน้าหลายชั้นที่แตกต่างกัน การทดลองแบบ A มีข้อมูลเข้าเป็นลักษณะเฉพาะทั้งหมดของ n-แกรม การทดลองแบบ B มีข้อมูลเข้าเป็นลักษณะเฉพาะ n-แกรม ที่ผ่านการเลือกด้วยเทคนิค CFS เทคนิคไคสแควร์ เทคนิคอัตราส่วนเกน และเทคนิครีลีฟ-เอฟ การทดลองแบบ C มีข้อมูลเข้าเป็นลักษณะเฉพาะ n-แกรม และลักษณะเฉพาะรูปแบบคอนเซนซัส และการทดลองแบบ D มีข้อมูลเข้าเป็นลักษณะเฉพาะ n-แกรมที่ผ่านการเลือกด้วยเทคนิคการเลือกลักษณะเฉพาะ และลักษณะเฉพาะรูปแบบคอนเซนซัส ทำการทดสอบประสิทธิภาพการทดลองแต่ละแบบด้วยชุดข้อมูล 3 ชุดข้อมูล ได้แก่ ชุดข้อมูล Vertebrate ชุดข้อมูล Arabidopsis thaliana และชุดข้อมูล TIS+50

โดยผลการทดลองเพื่อหาขนาดหน้าต่างที่เหมาะสมสำหรับชุดข้อมูล Vertebrate ด้วยเทคนิค CFS และโครงข่ายประสาทเทียมได้รับการตีพิมพ์ในงานประชุมวิชาการระดับชาติ The 12th National Computer Science and Engineering Conference (NCSEC 2008) ในระหว่างวันที่ 20-21 พฤศจิกายน 2551 ณ โรงแรมลองบีชการ์เด้น แอนด์ สปา จังหวัดชลบุรี เรื่อง การทำนายตำแหน่งเริ่มต้นการแปลรหัสพันธุกรรมของลำดับดีเอ็นเอสัตว์มีกระดูกสันหลังโดยเลือกลักษณะเฉพาะที่สัมพันธ์กันและโครงข่ายประสาทเทียม (Translation Initiation Sites Prediction of Vertebrate DNA Sequences Using Correlation-base Feature Selection and Neural Networks) แสดงดังภาคผนวก ก สำหรับแบบจำลอง TF-IDF-NN-TIS ได้รับการตีพิมพ์ในงานประชุมวิชาการระดับนานาชาติ The 2nd IEEE International Conference on Computer Science and Information Technology (IEEE ICCSIT 2009) ณ กรุงปักกิ่ง ประเทศจีน ระหว่างวันที่ 8-11 สิงหาคม 2552 แสดงดังภาคผนวก ข

ผลการทดลองตามแบบจำลองการทำนายจุดเริ่มต้นการแปลรหัสโดยใช้วิธี TF-IDF และโครงข่ายประสาทเทียม สามารถสรุปเป็นประเด็นต่างๆ 7 ประเด็น คือ 1) ประเด็นประสิทธิภาพของขนาดหน้าต่าง สรุปจากขั้นตอนที่ 1 การแบ่งสายพันธุกรรม 2) ประเด็นประสิทธิภาพ n-แกรม 3) ประเด็นประสิทธิภาพของค่า TF-IDF 4) ประเด็นประสิทธิภาพของเวลาสำหรับค่า TF-IDF ทั้งสามประเด็นนี้สรุปจากขั้นตอนที่ 2 การสร้างลักษณะเฉพาะ n-แกรม 5) ประเด็นประสิทธิภาพของเทคนิคการเลือกลักษณะเฉพาะ สรุปจากขั้นตอนที่ 3 การเลือกลักษณะเฉพาะ 6) ประเด็นประสิทธิภาพของรูปแบบคอนเซนซัส สรุปจากขั้นตอนที่ 4 การสร้างลักษณะเฉพาะรูปแบบคอนเซนซัส และ 7) ประเด็นประสิทธิภาพของแบบจำลอง TF-IDF-NN-TIS สรุปจากขั้นตอนที่ 1 ถึง ขั้นตอนที่ 5 โดยมีรายละเอียดดังนี้

6.1.1 ประเด็นประสิทธิภาพขนาดหน้าต่าง การแบ่งสายพันธุกรรมเป็นสายพันธุกรรมย่อยด้วยขนาดหน้าต่างที่เหมาะสมจะช่วยเพิ่มค่าความถูกต้องการทำนายจุดเริ่มต้นการแปล

6.1.2 ประเด็นประสิทธิภาพ n-แกรม การสร้างลักษณะเฉพาะสำหรับสายพันธุกรรมด้วยเทคนิค n-แกรมที่เหมาะสมจะช่วยเพิ่มค่าความถูกต้องการทำนายจุดเริ่มต้นการแปลรหัส

6.1.3 ประเด็นประสิทธิภาพของค่า TF-IDF การกำหนดค่าลักษณะเฉพา n-แกรมด้วยค่า TF-IDF จะให้ค่าความถูกต้อง ค่าการตอบสนองไว และค่าความเฉพาะเจาะจงการทำนายจุดเริ่มต้นการแปลรหัสสูงกว่าค่าลักษณะเฉพาะ n-แกรม ที่กำหนดด้วยค่าความถี่

6.1.4 ประเด็นประสิทธิภาพของเวลาสำหรับค่า TF-IDF การประมวลผลโครงข่ายประสาทเทียมซึ่งมีข้อมูลเข้าเป็นลักษณะเฉพาะที่กำหนดค่าลักษณะเฉพาะด้วยค่า TF-IDF ใช้เวลาการสร้างแบบจำลองน้อยกว่าค่าความถี่

6.1.5 ประเด็นประสิทธิภาพของเทคนิคการเลือกลักษณะเฉพาะ เทคนิคการเลือกลักษณะเฉพาะมี 4 เทคนิค คือ เทคนิค CFS เทคนิคไคสแควร์ เทคนิคอัตราส่วนเกน และเทคนิครีลีฟ-เอฟ สำหรับเทคนิคไคสแควร์ เทคนิคอัตราส่วนเกน และเทคนิครีลีฟ-เอฟเป็นเทคนิคการเลือกลักษณะเฉพาะแบบตัวกรองที่เรียงลำดับลักษณะเฉพาะตามนัยสำคัญทางสถิติจากมากไปน้อย จึงต้องระบุจำนวนลักษณะเฉพาะที่ต้องการใช้เป็นข้อมูลเข้าในโครงข่ายประสาทเทียม ผลการทดลองแสดงให้เห็นว่าจำนวนลักษณะเฉพาะตั้งแต่ 5 ถึง 15 ลำดับแรกที่มีนัยสำคัญสูงสุดให้ค่าความถูกต้องการทำนายจุดเริ่มต้นการแปลรหัสสูง สำหรับประสิทธิภาพของเทคนิคการเลือกลักษณะเฉพาะ ผลการทดลองแสดงให้เห็นว่าเทคนิค CFS ให้ค่าความถูกต้อง ค่าการตอบสนองไว และค่าความเฉพาะเจาะจงสูงในการทำนายผลลัพธ์ด้วยโครงข่ายประสาทเทียม

6.1.6 ประเด็นประสิทธิภาพของรูปแบบคอนเซนซัส ลักษณะเฉพาะที่สร้างจากรูปแบบคอนเซนซัสช่วยให้การระบุจุดเริ่มต้นการแปลรหัสถูกต้องมากขึ้น จากผลการทดลองแสดงให้เห็นว่าลักษณะเฉพาะที่สร้างจากรูปแบบคอนเซนซัสเพิ่มประสิทธิภาพการทำนายผลลัพธ์ของโครงข่ายประสาทเทียม โดยให้ค่าความถูกต้อง ค่าการตอบสนองไว และค่าความเฉพาะเจาะจงสูงกว่าการทำนายผลลัพธ์ของโครงข่ายประสาทเทียมที่ไม่มีลักษณะเฉพาะที่สร้างจากรูปแบบคอนเซนซัส

6.1.7 ประเด็นประสิทธิภาพของแบบจำลอง TF-IDF-NN-TIS จากการผลทดลองแสดงให้เห็นว่าการทดลองที่มีข้อมูลเข้าเป็นลักษณะเฉพาะ n-แกรม ที่ผ่านการเลือกด้วยเทคนิคการเลือกลักษณะเฉพาะร่วมกับลักษณะเฉพาะที่สร้างจากรูปแบบคอนเซนซัส ให้ค่าความถูกต้องการทำนายจุดเริ่มต้นการแปลรหัสสูง สำหรับการเปรียบเทียบค่าความถูกต้องระหว่างแบบจำลอง TF-IDF-NN-TIS กับงานวิจัยที่มีการศึกษาก่อนหน้าของชุดข้อมูล Vertebrate ชุดข้อมูล A.thaliana และชุดข้อมูล TIS+50 พบว่า แบบจำลองที่นำเสนอให้ค่าความถูกต้องเท่ากับ 99.76% 97.56% และ 95.50% ตามลำดับ ซึ่งสูงกว่างานวิจัยที่ศึกษาก่อนหน้า

6.2 ปัญหาและอุปสรรค

เนื่องจากข้อมูลที่ศึกษาส่วนใหญ่เป็นข้อมูลเกี่ยวกับชีววิทยา จึงต้องใช้เวลานานสำหรับการศึกษาเนื้อหาในส่วนของอณูชีววิทยา ชุดข้อมูล และชีวสารสนเทศศาสตร์

6.3 ข้อเสนอแนะ

เนื่องจากข้อมูลที่ใช้ในการทดลองมีจำนวนมาก จึงควรใช้เครื่องคอมพิวเตอร์ที่มีสมรรถนะสูง และอาจใช้เทคนิคการคำนวณแบบคลัสเตอร์ เพื่อลดเวลาในการทำงาน นอกจากนี้ยังสามารถนำแบบจำลองดังกล่าวไปใช้สำหรับการวิเคราะห์สายจีโนมเพื่อเป้าหมายอื่นๆ ตัวอย่างเช่น การทำนายตำแหน่งการเชื่อมต่อ (Splice Sites Junction Prediction) และการทำนายตำแหน่งโปรโมเตอร์ (Promoter Prediction)

บรรณานุกรม

บุรชัย สนธยานนท์ และคณะ. 2542. ชีวเคมี. ภาควิชาเคมี คณะวิทยาศาสตร์ มหาวิทยาลัยมหิดล. กรุงเทพฯ.

ประดิษฐ์ พงศ์ทองคำ. 2543. พันธุศาสตร์. คณะวิทยาศาสตร์, มหาวิทยาลัยเกษตรศาสตร์. กรุงเทพฯ.

อมรา คัมภิรานนท์. 2542. พันธุศาสตร์มนุษย์. บริษัท เท็กซ์ แอนด์ เจอ์นัล พับลิเคชั่น จำกัด. กรุงเทพฯ.

อุไรวรรณ วิจารณกุล. 2545. ดีเอ็นเอเทคโนโลยี, คณะวิทยาศาสตร์และเทคโนโลยี. สถาบันราชภัฏพิบูลสงคราม.

Cigan, A., Feng, L., and Donahue, T. 1988. tRNAi(met) Functions in Directing the Scanning Ribosome to the Start of Translation. Science, 242(4875): 93-97.

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

Computational Systems Biology Laboratory. 2004. Available: http://csbl.bmb.uga.edu/index.html. (accessed 01/10/08).

Conilione, P.C., and Wang, D. 2005. Effect of Non-Target Examples on E. coli Promoters Recognition Using Neural Networks. In Proceedings of International Joint Conference on Neural Networks. pp. 310-315.

Dana, J. 2006. The Era of Bioinformatics. IEEE Conferences Proceedings ICTTA 2nd. pp. 1860-1865.

Hall, M. A., and Smith, L.A. 1997. Feature Subset Selection: A Correlation Based Filter Approach. [Online] Available: http://www.cs.waikato.ac.nz/ml/publications/1997/HallSmith97.ps (accessed 12/12/08).

Hall, M. A. 2000. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. In Proceedings 17th Intelligence Conference on Machine Learning. pp. 359-366.

Hatzigeorgiou, A.G. 2002. Translation Initiation Start Prediction in Human cDNAs with High Accuracy. Bioinformatics. 18(2): 343-350.

Ho, L.S., and Rajapakse, J.C. 2004. High Sensitivity Techniques for Translation Initiation Site Detection. In Proceedings of the CIBCB'04 IEEE, pp. 1553-1559.

Joachims, T. 2003. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF fro Text Categorization. Ph.D. Thesis, Dortmund University, Dortmund, Germany.

Kantardzic, M. 2003. Data Mining: Concepts Models Methods and Algorithms. Wiley Interscience, New York.

Kozak, M. 1978. How Do Eucaryotic Ribosomes Select Initiation Regions in Messenger RNA. Cell, 15(4): 1109-1123.

Kozak, M. 1989. The Scanning Model for Translation: An Update. Journal of Cell Biology, 108(2): 229-241.

Kozak, M. 1987. An Analysis of 5'-Noncoding Sequences from 699 Vertebrate Messenger RNAs. Nucleic Acids Research, 15(20): 8125-8148.

Li, G. L., and Leong T. Y. 2005. Feature Selection for the Prediction of Translation Initiation Sites. Genomics, Proteomics and Bioinformatics. 3(2): 73-83.

Li, G. L., Leong, T. Y., and Zhang, L. 2005. Translation Initiation Sites Prediction with Mixture Gaussian Models in Human cDNA Sequences. IEEE Transactions on Knowledge and Data Engineering. August 2005, pp. 1152-1160.

Liu, H., Han, H., Li, J., and Wong, L. 2004. Using Amino Acid Patterns to Accurately Predict Translation Initiation Sites. In Silico Biology. 4(3): pp. 255-269.

Liu, D., Xiong, X., DasGupta, B., and Zhang, H. 2006. Motif Discoveries in Unaligned Molecular Sequences Using Self-Organizing Neural Networks. IEEE Transactions on Neural Networks. July 2006, pp. 919-928.

Marko, R. S., and Igor, K. 2003. Theoretical and Empirical Analysis of ReliefF and RReleifF. Machine Learning Journal. 53: 23-69.

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

National Center for Biotechnology Information. 1988. http://www.ncbi.nlm.nih.gov/ (accessed 01/11/08).

Nadershahi, A., Fahrenkrug, S. C., and Ellis, L. 2004. Comparison of computational methods for identifying translation initiation sites in EST data. BMC Bioinformatics. 5

Pedersen, A. G., and Nielsen, H. 1997. Neural Network Prediction of Translation Initiation Sites in Eukaryotes: Perspective for EST and Genome Analysis. Proceeding of 5th Intelligence Systems for Molecular Biology. pp. 226-233.

Roiger, R. J., and Geatz, M. W. 2003. Data Mining: A Tutorial-Based Primer, Pearson Education International Edition, New York.

Saidi, R., Maddouri, M. and Nguifo, E.M. 2007. Biological Sequences Encoding for Supervised Classification. Springer-Verlag Berlin Heidelberg 2007, pp. 224-238.

Stormo, G. D., Schneider, T. D., and Gold, L.M. 1982. Characterization of Translation Initiation Sites E. coli. Nucleic Acid Research. 10: 2971-2996.

Tan, T. Z., Ng, G. S., and Quck, C. 2006. Genetic Complementary Learning for Translation Initiation Sites Prediction. IEEE Congress on Evolutionary Computation. pp. 259-266.

Tzanis, G. Berberidis, C., Alexandridou, A., and Vlahavas, I. 2005. Improving the Accuracy of Classifiers for the Prediction of Translation Initiation Sites in Genomic Sequences. In PCI 2005 Springer-Verlag. Berlin Heidelberg, pp. 11-13.

Tzanis, G., Berberidis, C., and Vlahavas, I. 2006. A Novel Data Mining Approach for the Accurate Prediction of Translation Initiation Sites. In ISBMDA 2006 Springer-Verlag. Berlin Heidelberg, pp. 92-103.

Tzanis, G., and Vlahavas, I. 2006. Prediction of Translation Initiation Sites Using Classifier Selection. In SETN 2006 LNCS (LNAI) Springer. Heidelberg, pp. 367-377.

Tzanis, G., Berberidis, C., and Vlahavas, I. 2007 MANTIS: A Data Mining Methodology for Effective Translation Initiation Site Prediction. Proceedings of the 29th Annual International Conference of the IEEE EMBS. August 23-26, pp. 6343-6347.

Witten, I. H., Frank, E. 2005. Data Mining Practical Machine Learning Tools and Techniques. Elsevier, Inc., San Francisco.

Zeng, F., Yap, R., and Wong, L. 2002. Using Feature Generation and Feature Selection for Accurate Prediction of Translation Initiation Sites. In Proceedings of 13th International Conference on Genome Informatics. pp. 192-200.

Zeng, J., and Alhajj, R. 2007. Multi-Agent System for Translation Initiation Site Prediction. IEEE International Conference on Bioinformatics and Biomedicine. pp. 103-108.

Zien, A., Ratsch, G., Mika, S., Scholkopf, B., Lemmen, C., Smola, A., Lengauer, T., and Muller, K.-R. 2000. Engineering Support Vector Machine Kernels that recognize Translation Initiation Sites. Bioinformatics. 16: 799-807.

ภาคผนวก

ภาคผนวก ก

ผลงานวิจัยที่ได้รับการตีพิมพ์ในงานประชุมวิชาการ NCSEC 2008

เรื่อง การทำนายจุดเริ่มต้นการแปลรหัสพันธุกรรมของลำดับดีเอ็นเอสัตว์มีกระดูกสันหลังโดยเลือกลักษณะเฉพาะที่สัมพันธ์กันและโครงข่ายประสาทเทียม (Translation Initiation Sites Prediction of Vertebrate DNA Sequences Using Correlation-based Feature Selection and Neural Networks)

งานประชุมวิชาการ The 12th National Computer Science and Engineering Conference (NCSEC 2008)

วันที่ 20-21 พฤศจิกายน 2551

สถานที่ โรงแรมลองบีชการ์เด้น แอนด์ สปา จังหวัดชลบุรี

การทำนายตำแหน่งเริ่มต้นการแปลรหัสพันธุกรรมของลำดับดีเอ็นเอสัตว์มีกระดูกสันหลัง โดยการเลือกลักษณะเฉพาะที่สัมพันธ์กันและโครงข่ายประสาทเทียม

Translation Initiation Sites Prediction of Vertebrate DNA Sequences Using Correlation-based Feature Selection and Neural Networks

ฏรินทร คงมณี[1] ศิริรัตน์ วณิชโยบล[2] และ วิภาดา เวทย์ประสิทธิ์[1]
ห้องปฏิบัติการวิจัยปัญญาประดิษฐ์[1] ห้องปฏิบัติการวิจัยเทคโนโลยีระบบสารสนเทศและโปรแกรมประยุกต์[2] ภาควิชาวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ จังหวัดสงขลา 90112 โทรศัพท์: 0-7428-8581 โทรสาร: 0-7444-6917
E-mail: t_kongmanee@hotmail.com, sirirut.v@psu.ac.th, wwettayaprasit@yahoo.com

บทคัดย่อ

บทความนี้นำเสนอเทคนิคการทำนายตำแหน่งเริ่มต้นการแปลรหัสพันธุกรรม (Translation Initiation Sites) ของลำดับดีเอ็นสัตว์มีกระดูกสันหลังทำการเลือกลักษณะเฉพาะด้วยวิธีการเลือกลักษณะเฉพาะที่สัมพันธ์กัน (Correlation-based Feature Selection) และสอนด้วยโครงข่ายประสาทเทียม มีการทดลองเปรียบเทียบการเลือกขนาดหน้าต่างของลำดับดีเอ็นเอในขนาดต่างๆกัน พิจารณาทั้งกรณีอัพสตรีมและดาวน์สตรีม โดยสร้างลักษณะเฉพาะด้วยเทคนิค 1-แกรม และ 2-แกรม ทำการทดลองโดยใช้ชุดข้อมูลสัตว์มีกระดูกสันหลังจาก GenBank ผลการทดลองแสดงให้เห็นว่า เทคนิคที่นำเสนอให้ค่าความถูกต้องสูงสุด และใช้เวลาในการทำงานน้อย

คำสำคัญ: ตำแหน่งเริ่มต้นแปลรหัสพันธุกรรม, เอ็น-แกรม, โครงข่ายประสาทเทียม, การเลือกลักษณะเฉพาะที่สัมพันธ์กัน

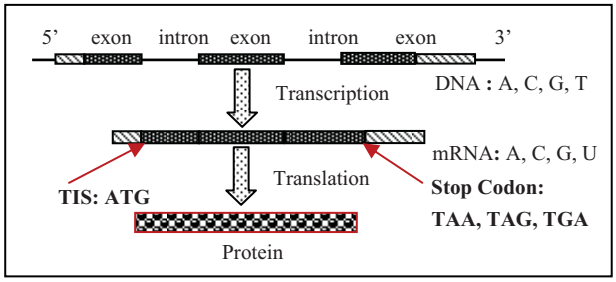
Abstract

This paper presents a technique of Translation Initiation Sites using Correlation-based Feature Selection and Neural Networks. There is comparison study for the selection of difference window sizes of DNA sequences. The study examines at both upstream and downstream nucleotides by using 1-gram and 2-gram techniques. The study uses data set of vertebrate DNA sequences from GenBank. The results of the study indicates that the proposed technique gives maximum accuracy with less time.

Keywords: Translation Initiation Sites, n-Gram, Neural Networks, Correlation-based Feature Selection

1. บทนำ

ในสิ่งมีชีวิตส่วนใหญ่ข้อมูลทางพันธุกรรมจะถูกเก็บไว้ในรูปของดีเอ็นเอ หน้าที่ของดีเอ็นเอคือการผ่านข้อมูลทางพันธุกรรมหรือลอกรหัสไปสู่อาร์เอ็นเอ (Transcription) จากนั้นทำการแปลรหัสจากอาร์เอ็นเอไปเป็นโปรตีน (Translation) กระบวนการดังกล่าวเรียกว่ากระบวนการผ่านข่าวสารทางพันธุกรรม หรือ Central Dogma [1] แสดงดังรูปที่ 1



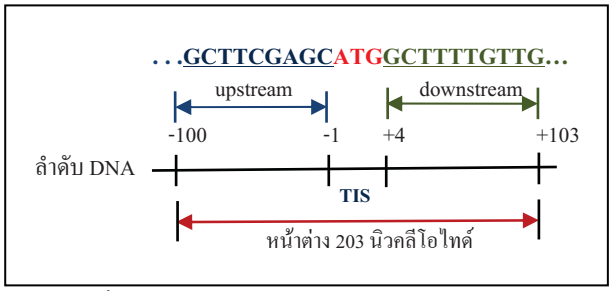
รูปที่ 1 โครงสร้างพื้นฐานการสร้างโปรตีนของเซลล์ยูคารีโอต

ดีเอ็นเอประกอบด้วยนิวคลีโอไทด์ (Nucleotide) หลายๆ นิวคลีโอไทด์มาต่อกันโดยที่นิวคลีโอไทด์มีทั้งหมด 4 ชนิดคือ A C G และ T ยีน คือดีเอ็นเอความยาวช่วงหนึ่งๆที่ก่อให้เกิดลักษณะต่างๆของสิ่งมีชีวิต และเป็นตัวกำหนดการเรียงตัวของกรดอะมิโนชนิดต่างๆของโปรตีน

รหัสพันธุกรรม (Genetic Code) หมายถึง ลำดับของนิวคลีโอไทด์บนอาร์เอ็นเอสื่อสาร (messenger RNA หรือ mRNA) ที่ถอดข้อความพันธุกรรมจากดีเอ็นเอไปเป็นลำดับของกรดอะมิโนของโปรตีนซึ่งรหัสพันธุกรรมสำหรับกรดอะมิโนหนึ่งตัวประกอบไปด้วย 3 นิวคลีโอไทด์ เรียกว่า โคดอน (Codon)

กระบวนการแปลรหัสพันธุกรรมบนลำดับเอ็มอาร์เอ็นเอไปเป็นโปรตีนจะเริ่มต้นที่ตำแหน่งเริ่มต้นการแปลรหัสพันธุกรรม

(Translation Initiation Sites: TIS) [2] ซึ่งมีรหัสพันธุกรรมเป็นโคดอน ATG สำหรับกระบวนการแปลรหัสพันธุกรรมตามสมมติฐานต้นแบบการตรวจสอบไรโบโซมนั้นสมมติให้ไรโบโซมเริ่มต้นตรวจสอบลำดับเอ็มอาร์เอ็นเอจากปลาย 5' ไปยังปลาย 3' จนกระทั่งเจอตำแหน่งเริ่มต้นการแปลรหัสพันธุกรรมจึงเริ่มต้นแปลรหัสพันธุกรรม และหยุดการแปลรหัสพันธุกรรมเมื่อเจอโคดอนหยุด (Stop Codon) ได้แก่ TAA TAG หรือ TGA [3-5] สำหรับนิวคลีโอไทด์ด้านซ้ายของ TIS เรียกว่า อัพสตรีม (Upstream) และ นิวคลีโอไทด์ด้านขวาของ TIS เรียกว่า ดาวน์สตรีม (Downstream) ดังรูปที่ 2 แสดงลำดับดีเอ็นเอย่อยของหน้าต่างขนาด 203 นิวคลีโอไทด์จะมีโคดอน ATG อยู่กึ่งกลางโดยที่นิวคลีโอไทด์ A ของ โคดอน ATG เป็นตำแหน่ง +1 สำหรับส่วนอัพสตรีมจะเริ่มต้นที่ตำแหน่ง -1 และลดลงเรื่อยๆไปด้านซ้าย สำหรับส่วนดาวน์สตรีมจะเริ่มต้นที่ตำแหน่ง +4 และเพิ่มขึ้นเรื่อยๆไปด้านขวา ข้อสังเกต สำหรับการทดลองจะเลือกพิจารณาลำดับของสายดีเอ็นเอที่มีโคดอน ATG เป็น TIS เพียงตำแหน่งเดียว แต่มีโคดอน ATG มากกว่า 1 ตำแหน่งได้



รูปที่ 2 ลำดับดีเอ็นเอย่อยของหน้าต่าง 203 นิวคลีโอไทด์

บทความนี้นำเสนอการทำนาย TIS โดยการเลือกลักษณะเฉพาะที่สัมพันธ์กันและสอนด้วยโครงข่ายประสาทเทียม ในส่วนที่ 2 กล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้อง ส่วนที่ 3 กล่าวถึงแบบจำลองการทำนายตำแหน่งเริ่มต้นการแปลรหัสพันธุกรรมโดยใช้การเลือกลักษณะเฉพาะที่สัมพันธ์กันและโครงข่ายประสาทเทียม (Translation Initiation Sites Prediction using Correlation-based Feature Selection and Neural Networks: TISP-CFS-NN) ส่วนที่ 4 กล่าวถึง ผลการทดลอง และส่วนที่ 5 คือบทสรุป

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 เทคนิคเอ็น-แกรม

เทคนิคเอ็น-แกรมเป็นวิธีสำหรับการสร้างลักษณะเฉพาะด้วยการกำหนดความยาว n คงที่ การสร้างเอ็น-แกรมทำได้โดยเลื่อนหน้าต่างของอักขระไปบนลำดับทั้งหมด n ตัว และในแต่ละครั้งที่ทำการเลื่อนลำดับย่อยขนาด n อักขระจะถูกสกัดออกมา เทคนิคเอ็น-แกรมมี 2 แบบ ได้แก่เทคนิคเอ็น-แกรมแบบทุกเฟรม (Any-frame) และเทคนิคเอ็นแกรมแบบอิน-เฟรม (In-frame)

2.1.1 เทคนิคเอ็น-แกรมแบบทุกเฟรม (Any-frame) มีการทำงานดังต่อไปนี้กำหนดให้ลำดับมีความยาว m ตัว แกรมหรือรูปแบบที่สกัดได้จะมีค่าเท่ากับ m-n+1 รูปแบบ [7] ตัวอย่างเช่นกำหนดให้ลำดับคือ "AAGGGCCTAG" ซึ่งมีความยาวเท่ากับ 10 ตัว (m = 10) ถ้าต้องการสร้าง 2-แกรม (n = 2) ดังนั้นจะมีชุดข้อมูลที่สกัดได้เท่ากับ 9 รูปแบบ (10-2+1 = 9) ได้แก่ AA AG GG GG GC CC CT TA และ AG เป็นต้น

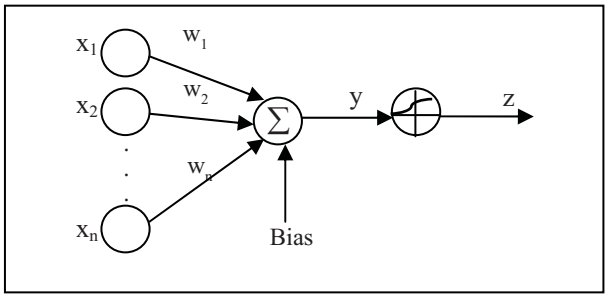
2.1.2 เทคนิคเอ็น-แกรมแบบอิน-เฟรม (In-frame) มีลักษณะการทำงานดังนี้ กำหนดให้ลำดับมีความยาว m ตัว แกรมหรือรูปแบบที่สกัดได้จะมีค่าเท่ากับ m/n รูปแบบ ตัวอย่างเช่นกำหนดให้ลำดับคือ "AA GG GC CT AG" ซึ่งมีความยาวเท่ากับ 10 ตัว (m = 10) ถ้าต้องการสร้าง 2-แกรม (n = 2) ดังนั้นจะมีชุดข้อมูลที่สกัดได้เท่ากับ 5 รูปแบบ (10/2 = 5) ได้แก่ AA GG GC CT และ AG สำหรับงานวิจัยได้นำเทคนิคเอ็น-แกรมมาสร้างลักษณะเฉพาะเพื่อทำนาย TIS ซึ่ง Zeng และคณะ [8] ใช้หน้าต่าง 203 นิวคลีโอไทด์

2.2 Correlation-based Feature Selection

ปัญหาที่เกิดขึ้นจากการนำวิธีการเรียนรู้ของเครื่องไปใช้ในการทำนาย TIS ของลำดับดีเอ็นเอคือ มีลักษณะเฉพาะจำนวนมาก แต่มีเพียงบางลักษณะเฉพาะเท่านั้นที่สัมพันธ์กับคลาส ทำให้ความถูกต้องของการเรียนรู้มีประสิทธิภาพลดลง ดังนั้นวิธีเลือกลักษณะเฉพาะจึงจำเป็นสำหรับการค้นหาลักษณะเฉพาะที่สำคัญสำหรับการทำนาย TIS เทคนิค Correlation-based Feature Selection หรือ CFS เป็นวิธีเลือกลักษณะเฉพาะที่อาศัยเทคนิคฮิวริสติกเพื่อค้นหาลักษณะเฉพาะที่มีความสัมพันธ์กับคลาสโดยทำการลดมิติของข้อมูล [10] ตัวอย่างเช่น Zeng และคณะ [8] ได้ใช้วิธี CFS เลือกลักษณะเฉพาะ ที่สร้างจากเทคนิคเอ็น-แกรมสำหรับทำนาย TIS ด้วยเทคนิคเหมืองข้อมูลที่หลายรูปแบบ

2.3 โครงข่ายประสาทเทียม

โครงข่ายประสาทเทียมคือเทคนิคการทำเหมืองข้อมูลเพื่อค้นหาความรู้ที่แฝงอยู่ในฐานข้อมูล มีรูปแบบการประมวลผลที่เลียนแบบการทำงานของเซลล์ประสาทของมนุษย์ประกอบด้วย หน่วยประมวลผลย่อยหลายหน่วยเชื่อมต่อกัน รูปแบบการเรียนรู้ของโครงข่ายประสาทเทียมเป็นการเรียนรู้แบบมีผู้สอน (Supervised Learning) ซึ่งจะต้องมีการสอนโครงข่ายประสาทเทียมก่อนนำไปใช้งานจริงรูปที่ 3 แสดงฟังก์ชันการทำงานในหน่วยประมวลผลย่อยของโครงข่ายประสาทเทียม ประกอบด้วยฟังก์ชันผลรวม (Summation Function) เพื่อคำนวณหาผลรวมของผลคูณระหว่างค่าน้ำหนักกับค่าข้อมูลเข้าและฟังก์ชันกระตุ้น (Activation Function)



รูปที่ 3 ฟังก์ชันการทำงานของหน่วยประมวลผลย่อย

โครงสร้างของโครงข่ายประสาทเทียมแบบหน่วยประมวลผลย่อยหลายชั้น (Multilayer Perceptron) มี 3 ระดับ คือ ชั้นข้อมูลเข้า (Input Layer) ชั้นซ่อน (Hidden Layer) และชั้นผลลัพธ์ (Output Layer) ใช้สำหรับงานที่มีความซับซ้อน เรียนรู้ด้วยขั้นตอนการส่งค่าย้อนกลับ (Backpropagation) ซึ่งกระบวนการส่งค่าย้อนกลับ ประกอบด้วย 2 ส่วนย่อยคือ การส่งผ่านไปข้างหน้า (Forward Pass) และการส่งผ่านย้อนกลับ (Backward Pass) สำหรับการส่งผ่านไปข้างหน้า ข้อมูลจะผ่านเข้าโครงข่ายประสาทเทียมที่ชั้นข้อมูลเข้า และจะส่งผ่าน จากอีกชั้นหนึ่งไปสู่อีกชั้นหนึ่งจนกระทั่งถึงชั้นข้อมูลออก ส่วนการส่งผ่านย้อนกลับค่าน้ำหนักการเชื่อมต่อจะถูกปรับเปลี่ยนให้สอดคล้องกับกฎการแก้ข้อผิดพลาด (Error-Correction) คือผลต่างของผลลัพธ์ที่ได้ (Actual Response) กับผลลัพธ์เป้าหมาย (Target Response) เกิดเป็นสัญญาณผิดพลาด (Error Signal) ซึ่งสัญญาณนี้จะถูกส่งย้อนกลับเข้าสู่โครงข่ายประสาทเทียมในทิศทางตรงกันข้ามกับการเชื่อมต่อ จากนั้นค่าน้ำหนักของการเชื่อมต่อจะถูกปรับจนกระทั่งผลลัพธ์ที่ได้เข้าใกล้ผลลัพธ์เป้าหมาย [11] ตัวอย่างงานวิจัยที่ใช้โครงข่ายประสาทเทียมเช่น Pedersen และ Nielsen [12] ใช้โครงข่ายประสาทเทียมทำนาย TIS ในเซลล์ยูคารีโอต ต่อมา Hatzigeorgiou [13] สร้างแบบจำลองการตรวจสอบไรโบโซมด้วยโครงข่ายประสาทเทียมเพื่อทำนาย TIS ใน cDNA ของมนุษย์ Ho และ Rajapakse [14] ใช้โครงข่ายประสาทเทียมเพื่อทำนาย TIS ของชุดข้อมูลลำดับดีเอ็นเอสัตว์มีกระดูกสันหลังซึ่งสร้างลักษณะเฉพาะด้วยวิธีลูกโซ่มาร์คอฟ เป็นต้น

2.4 การทดสอบไขว้เปลี่ยนแบบ K-กลุ่ม

การประเมินผลลัพธ์การทดสอบไขว้เปลี่ยนแบบ K กลุ่ม (K-Fold Cross Validation) เริ่มต้นการทำงานโดยแบ่งชุดข้อมูลขนาดเท่าๆกันจำนวน K กลุ่ม และไม่มีส่วนที่ซ้ำซ้อนกัน ในแต่ละครั้งของการประเมินผลลัพธ์จะทำการแบ่ง K-1 กลุ่มเป็นชุดสอน และ 1 กลุ่มเป็นชุดทดสอบ กระบวนการนี้ถูกทำซ้ำ K ครั้ง ด้วยชุดทดสอบที่แตกต่างกันในแต่ละครั้ง ดังนั้นข้อมูลทุกตัวจึงเป็นทั้งชุดสอน และชุดทดสอบ ตัวอย่างเช่น ข้อมูลจำนวน 600 ตัวอย่าง ทำการทดสอบไขว้เปลี่ยนแบบ 3-กลุ่ม เริ่มต้นแบ่งข้อมูลเป็น 3 กลุ่มๆละ 200 ตัวอย่าง รอบที่ 1 จะใช้ข้อมูลกลุ่มที่ 1 เป็นชุดทดสอบ กลุ่มที่ 2 และ 3 เป็นชุดสอน รอบที่ 2 ใช้ข้อมูลกลุ่มที่ 2 เป็นชุดทดสอบ กลุ่มที่ 1 และ 3 เป็นชุดสอน และรอบสุดท้ายใช้ข้อมูลกลุ่มที่ 3 เป็นชุดทดสอบ กลุ่มที่ 1 และ 2 เป็นชุดสอน เป็นต้น

ขั้นตอนที่ 1 การเตรียมข้อมูล

1.1 ค้นหาโคดอน ATG ที่ปรากฏในลำดับดีเอ็นเอทั้งหมด
1.2 แบ่งโคดอน ATG เป็นคลาส 1 และคลาส 0
1.3 แยกลำดับดีเอ็นเอเป็นลำดับดีเอ็นเอย่อยด้วยขนาดหน้าต่าง 63 103 203 และทั้งลำดับนิวคลีโอไทด์

↓

ขั้นตอนที่ 2 การสร้างลักษณะเฉพาะ

2.1 ใช้เทคนิค 1-แกรม และ 2-แกรม
2.1.1 นับความถี่ของ 1-แกรม และ 2-แกรมที่ปรากฏในอัพสตรีม (up_x และ up_xx)
2.1.2 นับความถี่ของ 1-แกรม และ 2-แกรมที่ปรากฏในดาวน์สตรีม (dn_x และ dn_xx)
2.2 ใช้ Local Feature (Optional)
2.2.1 พิจารณาตำแหน่ง -3 เป็นนิวคลีโอไทด์ A หรือ G (up_-3AG)
2.2.2 พิจารณาตำแหน่ง +4 เป็นนิวคลีโอไทด์ G (dn_+4G)
2.2.3 พิจารณาการปรากฏของโคดอนหยุด (TAA TAG หรือ TGA) ในอิน-เฟรมดาวน์สตรีม (dn_STOP)
2.2.4 พิจารณาการปรากฏของโคดอน ATG ในอิน-เฟรมอัพสตรีม (up_ATG)

↓

ขั้นตอนที่ 3 การเลือกลักษณะเฉพาะ

3.1 เลือกลักษณะเฉพาะที่สร้างจากเทคนิค 1-แกรม และ2-แกรม ด้วยวิธี Correlation-based Feature Selection

↓

ขั้นตอนที่ 4 การทำนาย TIS

4.1 กำหนดสถาปัตยกรรมของโครงข่ายประสาทเทียมโดยมีจำนวนโหนดเข้า (Input Node) เท่ากับจำนวนลักษณะเฉพาะที่เลือก
4.2 ทำนาย TIS ด้วยโครงข่ายประสาทเทียม
4.3 ประเมินผลการทำนายแบบ K-Fold Cross Validation

รูปที่ 4 แบบจำลอง TISP-CFS-NN

3. แบบจำลอง TISP-CFS-NN

แบบจำลองการทำนายตำแหน่งเริ่มต้นแปลรหัสพันธุกรรมโดยใช้การเลือกลักษณะเฉพาะที่สัมพันธ์กันและโครงข่ายประสาทเทียม (TISP-CFS-NN: Translation Initiation Sites Prediction Using

Correlation-based Feature Selection and Neural Networks) ใช้สำหรับการทำนาย TIS ของชุดข้อมูลลำดับดีเอ็นเอสัตว์มีกระดูกสันหลังมีการทำงาน 4 ขั้นตอน คือ 1) การเตรียมข้อมูล 2) การสร้างลักษณะเฉพาะ 3) การเลือกลักษณะเฉพาะ และ 4) การทำนาย TIS แสดงดังรูปที่ 4

3.1 ขั้นตอนที่ 1 การเตรียมข้อมูล

จากชุดข้อมูลที่มีอยู่ค้นหาโคดอน ATG ที่ปรากฏในลำดับดีเอ็นเอ กำหนดให้โคดอน ATG ที่มีตำแหน่งนิวคลีโอไทด์ A ตรงกับสัญลักษณ์ระบุ TIS เป็นคลาส 1 (ตรงกับตำแหน่ง TIS) และโคดอน ATG อื่นๆเป็นคลาส 0 จากนั้นแยกลำดับดีเอ็นเอเป็นลำดับดีเอ็นเอย่อยๆด้วยขนาดหน้าต่าง 63 103 203 และทั้งลำดับนิวคลีโอไทด์

3.2 ขั้นตอนที่ 2 การสร้างลักษณะเฉพาะ

การสร้างลักษณะเฉพาะแบ่งออกได้ 2 กรณี ได้แก่กรณีที่ 1 คือการใช้เทคนิค 1-แกรม และ 2-แกรม และกรณีที่ 2 คือการใช้เทคนิค Local Feature

กรณีที่ 1 การสร้างลักษณะเฉพาะโดยใช้เทคนิค 1-แกรม และ 2-แกรม (แทนค่าด้วย R) เนื่องจากลำดับดีเอ็นเอประกอบด้วยอักขระ 4 ตัวได้แก่ A G C และ T ดังนั้น 1-แกรมจึงมีรูปแบบทั้งหมดเท่ากับ 4 รูปแบบ ($4^1 = 4$) และ 2-แกรมมีรูปแบบทั้งหมดเท่ากับ 16 รูปแบบ ($4^2 = 16$) ได้แก่ AA AG AC AT GA GC GG GT CA CG CC CT TA TG TC และ TT จากนั้นนับความถี่ของรูปแบบทั้งหมดที่ปรากฏในหน้าต่างโดยพิจารณาแยกกันระหว่างอัพสตรีมและดาวน์สตรีมโดยมีจำนวนรูปแบบในอัพสตรีมทั้งหมดเท่ากับ 20 รูปแบบ (4+16 = 20) และดาวน์สตรีมอีก 20 รูปแบบ (4+16 = 20) ดังแสดงในตารางที่ 1 ตัวอย่างลักษณะเฉพาะ 1-แกรมในอัพสตรีมเช่น up_A และ up_G เป็นต้น และตัวอย่างลักษณะเฉพาะ 2-แกรมในดาวน์สตรีม เช่น dn_AA dn_AG และ dn_AT เป็นต้น

ตารางที่ 1 ลักษณะเฉพาะสร้างจาก 1-แกรม และ2-แกรม

รูปแบบ	คำอธิบาย	จำนวน
up_x	นับจำนวนแกรม x ในอัพสตรีม	4
dn_x	นับจำนวนแกรม x ในดาวน์สตรีม	4
up_xx	นับจำนวนแกรม xx ในอัพสตรีม	16
dn_xx	นับจำนวนแกรม xxในดาวน์สตรีม	16

กรณีที่ 2 การใช้ Local Feature (แทนด้วย LF) สร้างลักษณะเฉพาะซึ่ง LF เป็นการสร้างลักษณะเฉพาะจากสมมติฐานต้นแบบการตรวจสอบไรโบโซม [3-5] และรูปแบบลักษณะเด่นส่วนใหญ่ของ Koak [6] โดยลักษณะเฉพาะที่สร้างจากต้นแบบการตรวจสอบคือการตรวจสอบว่า โคดอน ATG ที่พิจารณาเป็นโคดอน ATG แรกหรือไม่ และมีโคดอนหยุดปรากฏในอิน-เฟรมอัพสตรีมหรือไม่ สำหรับลักษณะเฉพาะที่สร้างจากรูปแบบลักษณะเด่นส่วนใหญ่ของ Kozak คือการตรวจสอบว่าตำแหน่ง -3 ในอัพสตรีมเป็นนิวคลีโอไทด์ A หรือ G และตำแหน่ง +4 ในดาวน์สตรีมเป็นนิวคลีโอไทด์ G ดังนั้น LF จึงมี 4 รูปแบบแสดงดังตารางที่ 2 ซึ่งลักษณะเฉพาะที่สร้างจาก LF ช่วยให้ระบุโคดอน ATG ที่เป็น TIS ได้ถูกต้องมากยิ่งขึ้น

ตารางที่ 2 ลักษณะเฉพาะที่สร้างจาก Local Feature

รูปแบบ	คำอธิบาย	จำนวน
dn_+4G	ลักษณะเฉพาะบูลีนซึ่งเป็นจริงถ้ามีนิวคลีโอไทด์ G ที่ตำแหน่ง +4 [6]	1
up_-3AG	ลักษณะเฉพาะบูลีนซึ่งเป็นจริงถ้ามีนิวคลีโอไทด์ A หรือ G ที่ตำแหน่ง -3 [6]	1
up_ATG	ลักษณะเฉพาะบูลีนซึ่งเป็นจริงถ้ามีโคดอน ATG ในอิน-เฟรมอัพสตรีม [3-5]	1
dn_STOP	ลักษณะเฉพาะบูลีนซึ่งเป็นจริงถ้ามีโคดอนหยุด (TAA, TAG, หรือ TGA) ในอิน-เฟรมดาวน์สตรีม [3-5]	1

3.3 ขั้นตอนที่ 3 การเลือกลักษณะเฉพาะ

ลักษณะเฉพาะทั้งหมดที่ได้จาก 1-แกรมและ 2-แกรมมี 40 รูปแบบแต่มีเพียงบางลักษณะเท่านั้นที่สัมพันธ์กับคลาสจึงเลือกลักษณะเฉพาะที่สัมพันธ์กับคลาสด้วยวิธี CFS ซึ่งหลักการทำงานของวิธี CFS คือการใช้ฮิวริสติกสำหรับการประเมินค่ากลุ่มย่อยของลักษณะเฉพาะโดยพิจารณาระดับความสัมพันธ์ภายในที่เกี่ยวข้องระหว่างลักษณะเฉพาะกับคลาส และลักษณะเฉพาะกับลักษณะเฉพาะ โดยสูตรฮิวริสติกดังสมการที่ (1) ซึ่งจะให้ค่าคะแนนสูงสุดสำหรับกลุ่มของลักษณะเฉพาะที่มีความสัมพันธ์ระหว่างลักษณะเฉพาะกับคลาสสูง และมีความสัมพันธ์ภายในระหว่างลักษณะเฉพาะกับลักษณะเฉพาะต่ำ โดยจากสูตรจะให้ค่าที่บ่งชี้การทำนายของแต่ละคลาสของลักษณะเฉพาะ และจัดการกับลักษณะเฉพาะที่ไม่เกี่ยวข้อง

$$Merit_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad (1)$$

เมื่อ $Merit_S$ คือ ค่าฮิวริสติกของกลุ่มลักษณะเฉพาะ S ที่ประกอบด้วย k ลักษณะเฉพาะที่ถูกคัดเลือก $\overline{r_{cf}}$ คือค่าเฉลี่ยความสัมพันธ์ระหว่างลักษณะเฉพาะกับคลาส $\overline{r_{ff}}$ คือค่าเฉลี่ยความสัมพันธ์ภายในระหว่างลักษณะเฉพาะกับลักษณะเฉพาะ [10] โดยกำหนดให้ลักษณะเฉพาะที่ผ่านการเลือกด้วยวิธี CFS แทนด้วย M

3.4 ขั้นตอนที่ 4 การทำนาย TIS

การทำนาย TIS ใช้โครงข่ายประสาทเทียมแบบหน่วยประมวลผลย่อยหลายชั้น และฟังก์ชันกระตุ้นซิกมอยด์ กำหนดสถาปัตยกรรมของโครงข่ายประสาทเทียมให้ชั้นข้อมูลเข้ามีจำนวนโหนดเท่ากับจำนวนลักษณะเฉพาะที่ต้องการ จำนวนข้อมูลชั้นซ่อน (แทนด้วย H) มีจำนวน 5 โหนด และจำนวนข้อมูลชั้นแสดงผล (แทนด้วย O) มีจำนวน 2 โหนด ประเมินผลการทำนาย TIS แบบ K-Fold Cross Validation

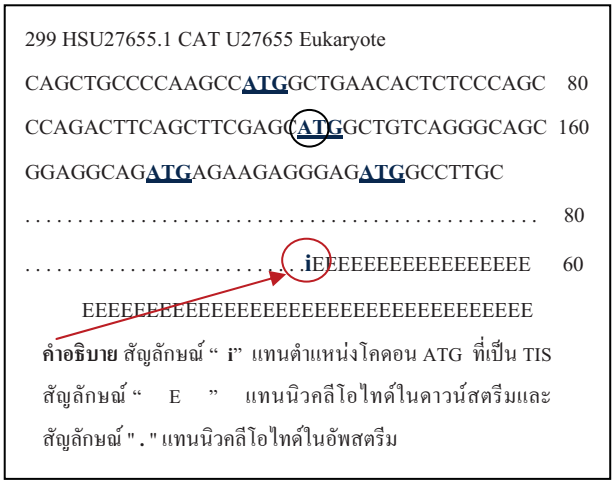
4. ผลการทดลอง

แบบจำลอง TISP-CFS-NN ได้ทำการทดลองบนเครื่องคอมพิวเตอร์ส่วนบุคคล หน่วยความจำ 2 กิกะไบต์ หน่วยประมวลผลกลางรุ่น Intel(R) Core(TM) 2 ความเร็วในการประมวลผล 1.83 กิกะเฮิร์ต มีการทดลองทั้งหมด 4 แบบคือการทดลองแบบ A B C และ D โดยการทดลอง A มีข้อมูลเข้าเป็นลักษณะเฉพาะที่ผ่านการเลือกด้วยวิธี CFS (M) การทดลอง B มีข้อมูลเข้าเป็นลักษณะเฉพาะที่ผ่านการเลือกด้วยวิธี CFS และ Local Feature (M และ LF) การทดลอง C มีข้อมูลเข้าเป็นลักษณะเฉพาะทั้งหมด และ Local Feature (R และ LF) และการทดลอง D มีข้อมูลเข้าเป็น Local Feature (LF) เท่านั้น ข้อสังเกตจะเห็นได้ว่าจำนวนข้อมูลเข้าเพื่อใช้ในการสอนโครงข่ายประสาทเทียมจะขึ้นอยู่กับการทดลองในแต่ละแบบส่วนชั้นข้อมูลซ่อนและชั้นแสดงผลเท่ากันทุกการทดลองดังตารางที่ 3 แสดงสถาปัตยกรรมโครงข่ายประสาทเทียมของการทดลอง แต่ละแบบตัวอย่างเช่น การทดลอง A มีสถาปัตยกรรมโครงข่ายประสาทเทียมเป็น M : H : O นั่นคือชั้นข้อมูลเข้ามีจำนวนโหนดเท่ากับ M ชั้นซ่อนมีข้อมูลเป็น H และชั้นแสดงผลมีข้อมูลเป็น O เป็นต้น

ตารางที่ 3 สถาปัตยกรรมโครงข่ายประสาทเทียมของการทดลอง

การทดลอง	ลักษณะเฉพาะ			สถาปัตยกรรมของโครงข่ายประสาทเทียม
	R	M	LF	
A	✗	✓	✗	M : H : O
B	✗	✓	✓	(M+LF) : H : O
C	✓	✗	✓	(R+LF) : H : O
D	✗	✗	✓	LF : H : O

ชุดข้อมูลในการทดลองประกอบด้วยลำดับดีเอ็นเอจำนวน 3,312 ลำดับรวบรวมจากสัตว์มีกระดูกสันหลังหลายชนิดจาก Kent Ridge Biomedical Data Set Repository สกัดจาก GenBank [15] แต่ละลำดับดีเอ็นเอมีโคดอน ATG เป็น TIS ที่ถูกต้องเพียงตำแหน่งเดียวแสดงตัวอย่างดังรูปที่ 5 ลำดับดีเอ็นเอมีจำนวนโคดอน ATG อยู่ทั้งหมด 4 ชุด โดยที่โคดอน ATG ชุดที่ 2 เป็น TIS เนื่องจากนิวคลีโอไทด์ A ในโคดอน ATG ชุดนี้มีตำแหน่งตรงกับสัญลักษณ์ " **i** " ซึ่งระบุ TIS พอดี



รูปที่ 5 ตัวอย่างลำดับดีเอ็นเอที่มีตำแหน่ง TIS

4.1 ขั้นตอนที่ 1 การเตรียมข้อมูล

จากชุดข้อมูลจำนวน 3,312 ลำดับพบว่าตำแหน่งนิวคลีโอไทด์ A ของโคดอน ATG ที่ตรงกับสัญลักษณ์ " **i** " มีจำนวน 3,263 ตำแหน่ง ดังนั้นจึงใช้ลำดับดีเอ็นเอสำหรับการทดลองเพียงจำนวน 3,263 ลำดับ และสกัดทิ้งไปจำนวน 49 ลำดับ เนื่องจากในแต่ละลำดับดีเอ็นเอมีชุดโคดอน ATG ที่ไม่ตรงกับสัญลักษณ์ระบุ TIS ดังนั้นจึงมีโคดอน ATG ทั้งหมด 13,308 ชุดประกอบด้วยโคดอน ATG ที่เป็นคลาส 1 จำนวน 3,263 (24.5%) ชุด และคลาส 0 จำนวน 10,045 (75.5%) ชุด ขั้นตอนต่อไปทำการแยกโคดอน ATG แต่ละชุดออกมาเป็นลำดับดีเอ็นเอย่อยตามขนาดหน้าต่าง 63 103 203 และทั้งลำดับนิวคลีโอไทด์

4.2 ขั้นตอนที่ 2 การสร้างลักษณะเฉพาะ

กำหนดให้ลักษณะเฉพาะที่สร้างด้วยเทคนิค 1-แกรม และ 2-แกรม มีทั้งหมด 40 รูปแบบ (R = 40) และกำหนดให้ลักษณะเฉพาะที่สร้างจาก Local Feature มีทั้งหมด 4 รูปแบบ (LF= 4) รวมทั้งหมด 44 รูปแบบ แสดงดังตารางที่ 4

ตารางที่ 4 รูปแบบลักษณะเฉพาะทั้งหมดที่ใช้ในการทดลอง

เทคนิคการสร้างลักษณะเฉพาะ	รูปแบบลักษณะเฉพาะ
1-แกรมอัพสตรีม	up_A up_C up_G up_T
1-แกรมดาวน์สตรีม	dn_A dn_C dn_G dn_T
2-แกรมอัพสตรีม	up_AA up_AG up_AC up_AT up_GA up_GG up_GC up_GT up_CA up_CG up_CC up_CT up_TA up_TG up_TC up_TT
2-แกรมดาวน์สตรีม	dn_AA dn_AG dn_AC dn_AT dn_GA dn_GG dn_GC dn_GT dn_CA dn_CG dn_CC dn_CT dn_TA dn_TG dn_TC dn_TT
Local Feature	up_-3AG dn_+4G up_ATG dn_STOP

4.3 ขั้นตอนที่ 3 การเลือกลักษณะเฉพาะ

ลักษณะเฉพาะที่สร้างจากเทคนิค 1-แกรม และ 2-แกรม ในแต่ละลำดับดีเอ็นเอย่อยมีจำนวนทั้งหมด 40 รูปแบบ ทำการทดลองเลือกลักษณะเฉพาะด้วยวิธี CFS โดยใช้โปรแกรม WEKA จำนวนลักษณะเฉพาะที่ผ่านการเลือกด้วยวิธี CFS แสดงดังตารางที่ 5 ข้อสังเกตจำนวนลักษณะเฉพาะที่ผ่านการเลือกด้วยวิธี CFS จะมีค่าน้อยกว่าหรือเท่ากับลักษณะเฉพาะที่สร้างขึ้นเสมอ (M ≤ R) ซึ่งหน้าต่าง 63 นิวคลีโอไทด์มีค่า M เท่ากับ 11 หน้าต่าง 103 นิวคลีโอไทด์มีค่า M เท่ากับ 12 หน้าต่าง 203 นิวคลีโอไทด์มีค่า M เท่ากับ 13 และหน้าต่างทั้งลำดับนิวคลีโอไทด์มีค่า M เท่ากับ 4

ตารางที่ 5 ลักษณะเฉพาะเทคนิคเอ็น-แกรมที่ผ่านการเลือกด้วยวิธี CFS

หน้าต่าง	ลักษณะเฉพาะที่ผ่านการเลือกด้วยวิธี CFS	จำนวน (M)
63	up_G up_T dn_C dn_G up_AT up_GC up_CG up_TG dn_GC dn_CG dn_CT	11
103	up_A up_T dn_C dn_G dn_T up_AT up_CG up_TG dn_GC dn_CG dn_CT dn_TC	12
203	up_A dn_A dn_C dn_G dn_T up_AT up_CG up_TG dn_GG dn_GC dn_CG dn_CT dn_TC	13
ทั้งลำดับ	dn_A dn_G dn_C dn_T	4

4.4 ขั้นตอนที่ 4 การทำนาย TIS

การทดลองแต่ละแบบมีสถาปัตยกรรมโครงข่ายประสาทเทียมแตกต่างกันและประเมินผลแบบ 3-Fold Cross Validation โดยพิจารณาประสิทธิภาพของขั้นตอนวิธีในประเด็นประสิทธิภาพของขนาดหน้าต่างทั้งลำดับนิวคลีโอไทด์ ประสิทธิภาพของการเลือกลักษณะเฉพาะด้วยวิธี CFS และประสิทธิภาพของเวลาการสร้างแบบจำลองดังรายละเอียดต่อไปนี้

4.4.1 ประสิทธิภาพของขนาดหน้าต่างทั้งลำดับนิวคลีโอไทด์

การทดลองของหน้าต่าง 63 103 203 และทั้งลำดับนิวคลีโอไทด์แสดงดังตารางที่ 6 ผลการทดลองแสดงให้เห็นว่าการทดลอง A ให้ค่าความถูกต้องเท่ากับ 77.05% 77.84% 80.17% และ 99.88% ตามลำดับ การทดลอง B ให้ค่าความถูกต้องเท่ากับ 80.21% 84.51% 92.79% และ 98.07% ตามลำดับ การทดลอง C ให้ค่าความถูกต้องเท่ากับ 80.91% 85.19% 92.31% และ 99.02% ตามลำดับ และการทดลอง D ให้ค่าความถูกต้องเท่ากับ 76.50% 78.92% 86.02% และ 89.06% ตามลำดับ จากการทดลองพบว่าหน้าต่างที่มีจำนวนนิวคลีโอไทด์มากค่าความถูกต้องสูง

ตารางที่ 6 ค่าความถูกต้องการทดลองของขนาดหน้าต่างแตกต่างกัน

การทดลอง	ค่าความถูกต้องของขนาดหน้าต่าง (%)			
	63	103	203	ทั้งลำดับ (MAX.=299)
A	77.05	77.84	80.17	**99.88**
B	80.21	84.51	92.79	98.07
C	80.91	85.19	92.31	99.02
D	76.50	78.92	86.02	89.06

4.4.2 ประสิทธิภาพของการเลือกลักษณะเฉพาะด้วยวิธี CFS

เมื่อเปรียบเทียบค่าความถูกต้องของหน้าต่างทั้งลำดับนิวคลีโอไทด์ของการทดลอง A B C และ D แสดงดังตารางที่ 6 ผลการทดลองให้ค่าความถูกต้องเท่ากับ 99.88% 98.07% 99.02% และ 89.06% ตามลำดับ ผลการทดลองแสดงให้เห็นว่าการเลือกลักษณะเฉพาะด้วยวิธี CFS ให้ผลการทดลองที่สูงสุดคือ 99.88% (การทดลอง A) แต่ถ้าใช้เฉพาะ LF โดยไม่ใช้เทคนิค CFS จะได้ผลการทดลองเพียง 89.06% (การทดลอง D) สำหรับประเด็นการเลือกใช้ LF ร่วมกับ CFS (การทดลอง B) ยังคงให้ผลการทดลองที่ดีคือ 98.07%

ตารางที่ 7 แสดงให้เห็นถึงจำนวนลักษณะเฉพาะที่สกัดได้จากขนาดหน้าต่างที่แตกต่างกัน ตัวอย่างเช่น เมื่อพิจารณาทั้งลำดับนิวคลีโอไทด์ของผลการทดลอง A B C และ D ได้จำนวนลักษณะเฉพาะเท่ากับ 4 8 44 และ 4 รูปแบบ ตามลำดับ จะเห็นได้ว่าการทดลอง A และ D มีจำนวนลักษณะเฉพาะน้อย

ตารางที่ 7 จำนวนลักษณะเฉพาะที่สกัดได้จากหน้าต่างที่แตกต่างกัน

การทดลอง	จำนวนลักษณะเฉพาะ			
	63	103	203	ทั้งลำดับ (MAX.=299)
A	11	12	13	**4**
B	15	16	17	8
C	44	44	44	44
D	4	4	4	**4**

4.4.3 ประสิทธิภาพของเวลาการสร้างแบบจำลอง

เปรียบเทียบเวลาการสร้างแบบจำลองของหน้าต่างทั้งลำดับนิวคลีโอไทด์ของการทดลอง A B C และ D แสดงดังตารางที่ 8 ผลการทดลองใช้เวลา 41.11 61.42 146.38 และ 46.73 วินาที ตามลำดับ แสดงให้เห็นว่าการทดลอง A ใช้เวลาน้อยที่สุดเนื่องจากได้ลดจำนวนลักษณะเฉพาะลง

ตารางที่ 8 เวลาการสร้างแบบจำลองของขนาดหน้าต่างแตกต่างกัน

การทดลอง	เวลาการสร้างแบบจำลองของขนาดหน้าต่าง (วินาที)			
	63	103	203	ทั้งลำดับ (MAX.=299)
A	76.06	57.39	67.25	**41.11**
B	88.53	75.50	77.66	61.42
C	157.78	153.72	159.23	146.38
D	41.50	39.64	41.42	46.73

5. บทสรุป

บทความนี้นำเสนอการทำนายการทำนายตำแหน่งเริ่มต้นการแปลรหัสพันธุกรรมของลำดับดีเอ็นเอสัตว์มีกระดูกสันหลังโดยการเลือกลักษณะเฉพาะที่สัมพันธ์กันและสอนด้วยโครงข่ายประสาทเทียม หรือแบบจำลอง TISP-CFS-NN ผลการทดลองสามารถสรุปได้ดังนี้ 1) หน้าต่างที่มีจำนวนนิวคลีโอไทด์มากจะมีค่าความถูกต้องที่สูงกว่า 2) การเลือกลักษณะเฉพาะด้วยวิธี CFS ทำให้สามารถลดมิติข้อมูลแต่ยังคงมีประสิทธิภาพการทำนายตำแหน่งเริ่มต้นแปลรหัสพันธุกรรมสูง 3) การเลือกลักษณะเฉพาะด้วยวิธี CFS ช่วยลดเวลาในการสร้างแบบจำลองลง

ดังนั้นการทำนายตำแหน่งเริ่มต้นการแปลรหัสพันธุกรรมของลำดับดีเอ็นเอสัตว์มีกระดูกสันหลังจึงควรใช้นิวคลีโอไทด์ทั้งลำดับ จากนั้นเลือกลักษณะเฉพาะด้วยวิธี CFS เพื่อลดมิติข้อมูลและลดเวลาการสร้างแบบจำลองในขณะที่ยังคงมีประสิทธิภาพการทำนาย TIS สูงโดยไม่ต้องใช้ Local Feature

6. กิตติกรรมประกาศ

งานวิจัยนี้ได้รับทุนสนับสนุนนักศึกษาระดับบัณฑิตศึกษาเป็นผู้ช่วยนักวิจัย (Research Assistant) ปีการศึกษา 2550 จากกองทุนวิจัย คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่

เอกสารอ้างอิง

[1] อัมรา คัมภิรานนท์, *พันธุศาสตร์มนุษย์*, บริษัท เท็กซ์ แอนด์ เจอร์นัล พับลิเคชั่น จำกัด, กรุงเทพฯ, 2542.

[2] อุไรวรรณ วิจารณกุล, *ดีเอ็นเอเทคโนโลยี*, คณะวิทยาศาสตร์และเทคโนโลยี, สถาบันราชภัฏพิบูลสงคราม, 2545.

[3] M. Kozak, "The Scanning Model for Translation: An Update," Cell Biology, vol. 108, no. 2, pp. 229-241, 1989.

[4] A. Cigan, L. Feng, and T. Donahue, "tRNAi(met) Functions in Directing the Scanning Ribosome to the Start of Translation," Science, vol. 242, no. 4875, pp. 93-97, 1988.

[5] M. Kozak, "How Do Eucaryotic Ribosomes Select Initiation Regions in Messenger RNA," Cell, vol. 15, no. 4, pp. 1109-1123, 1978.

[6] M. Kozak, "An Analysis of 5'-Noncoding Sequences from 699 Vertebrate Messenger RNAs," Nucleic Acids Research, vol. 15, no. 20, pp. 8125-8148, 1987.

[7] R. Saidi, M. Maddouri, and E. M. Nguifo, "Biological Sequences Encoding for Supervised Classification," Springer-Verlag Berlin Heidelberg 2007, pp. 224-238, 2007.

[8] F. Zeng, R. H. C. Yap, and L. Wong, "Using Feature Generation and Feature Selection for Accurate Prediction of Translation Initiation Sites," In Proceedings of the 13th International Conference on Genome Informatics, Tokyo, Japan, pp. 192-200, 2002.

[9] H. Liu, H. Han, J. Li, and L. Wong, "Using Amino Acid Patterns to Accurately Predict Translation Initiation Sites," In Silico Biology, vol. 4, no. 3, pp. 255-269, 2004.

[10] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," In Proceedings 17th Intelligence Conference on Machine Learning, pp. 359-366, 2000.

[11] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, A Wiley-Interscience Publication, USA, 2001.

[12] A. G. Pedersen, H. Nielsen, "Neural Network Prediction of Translation Initiation Sites in Eukaryotes: Perspective for EST and Genome Analysis," In Proceedings 5th International Conferences Intelligent Systems for Molecular Biology, pp. 226-233, 1997.

[13] A. G. Hatzigeorgiou, "Translation Initiation Start Prediction in Human cDNAs with High Accuracy," Bioinformatics, vol. 18, no. 2, pp. 343-350, 2002.

[14] Loi Sy Ho, and J. C. Rajapakse, "High Sensitivity Techniques for Translation Initiation Site Detection," In Proceedings of the 2004 Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB '04, pp. 1553 – 1559, 2004.

[15] Agency for Science Technology and Research, Singapore, Avialable: http://sdmc.i2r.a-star.edu.sg/rp [May/1/2008].

ประวัติผู้เขียนบทความ

ฏรินทร คงมณี นักศึกษาปริญญาโท ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ งานวิจัยที่สนใจได้แก่ Neural Networks, Bioinformatics, Data Mining

ผศ.ดร.วิภาดา เวทย์ประสิทธิ์ อาจารย์ประจำภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ งานวิจัยที่สนใจได้แก่ Knowledge Management, Neural Networks, Artificial Intelligent

ผศ.ดร.ศิริรัตน์ วณิชโยบล อาจารย์ประจำภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัย สงขลานครินทร์ วิทยาเขตหาดใหญ่ งานวิจัยที่สนใจได้แก่ Data Warehouse, Data Mining, GIS, Parallel Computing

ภาคผนวก ข

ผลงานวิจัยที่ได้รับการตีพิมพ์ในงานประชุมวิชาการ IEEE ICCSIT 2009

เรื่อง The TF-IDF and Neural Networks Approach for Translation Initiation Site Prediction

งานประชุมวิชาการ The 2nd IEEE International Conference on Computer Science and Information Technology (ICCSIT 2009)

วันที่ 8-11 สิงหาคม 2552

สถานที่ กรุงปักกิ่ง ประเทศจีน

The TF-IDF and Neural Networks Approach for Translation Initiation Site Prediction

Tarintorn Kongmanee
Artificial Intelligence Research Laboratory
Computer Science Department
Prince of Songkla University, Thailand
t_kongmanee@hotmail.com

Sirirut Vanichayobon
iSTAR Research Laboratory
Computer Science Department
Prince of Songkla University, Thailand
sirirut.v@psu.ac.th

Wiphada Wettayaprasit
Artificial Intelligence Research Laboratory
Computer Science Department
Prince of Songkla University, Thailand
wwettayaprasit@yahoo.com

***Abstract*—The precise prediction of translation initiation site is an important task for the analysis of genomic sequence. This study aims to increase the accuracy for the prediction of translation initiation site using a TF-IDF-NN-TIS model (TF-IDF and Neural Networks Approach for Translation Initiation Site Prediction). This study creates feature using 1-gram and 2-gram techniques for both upstream and downstream. Determining feature value uses TF-IDF approach and feature selection by correlation-based feature selection method. Evaluation prediction results use 10-fold cross validation. This study performed experiments on three different datasets that are Vertebrate, Arabidopsis thaliana, and TIS+50. The results of the study indicate that the proposed model gives highest accuracy with less processing time.**

***Keywords-translation initation sites; neural networks; correlation-based feature selection; TF-IDF*.**

I. INTRODUCTION

Gene is an important basis constituent of any living organisms. In general, genes are portions of the deoxyribonucleic acid, or DNA. DNA composes of several connecting nucleotides. The DNA has four bases that are adenine (A), cytosine (C), guanine (G), and thymine (T). DNA sequence has two ends called the 5' and 3' end. The DNA is transcribed to produce messenger RNA (mRNA), which is then translated to produce protein. For the translation process according to the ribosome scanning model that is ribosome scan mRNA sequence from the 5' end to the 3' end until it reads a translation initiation site (TIS) which is conserved ATG codon that has appropriate context. Then the translation process will begin and terminate when stopped codon is read which are TAA, TAG, or TGA [1]. The left-hand side nucleotide of TIS is called upstream region and the right-hand side nucleotide of TIS is called downstream region. The aim of TIS prediction is to correctly and efficiently identify the position of TIS in genome sequence.

This paper is outlined as follows. Section II is the related work and background. Section III proposes the TF-IDF and Neural Networks Approach for Translation Initiation Site (TF-IDF-NN-TIS) model. Section IV is the experimental result. Section V is the conclusion.

II. RELATED WORK AND BACKGROUND

Kozak [2] was the first researcher who proposed weight matrix to model the conserved motif around the TIS in cDNA sequence in 1987. The conserved motif derived from this matrix is GCC[**A/G**]CCATG**G**. Within this conserved motif, the nucleotide A or G in position -3 and nucleotide G in position +4 are the most highly conserved, assuming that +1 is the position of the nucleotide A of the conserved ATG codon. The first automatically system for TIS prediction is NetStart system that proposed by Pedersen and Nielsen [3]. The study trained feedforward multilayer neural networks with window size of 203 nucleotides. Whereas, Zien *et al.* [4] combined support vector machine with specially developed kernel function. This study carefully designed kernel functions for the purpose of achieving higher TIS prediction accuracy. Another method based on artificial neural networks was proposed by Hatzigeorgiou [5], who combined two neural networks that analyzed the conserved pattern and the coding or noncoding potential around the TIS, along with the ribosome scanning model. The study considered an input window size of 12 nucleotides around codon ATG.

Rajapakse and Ho [6] proposed a technique of encoding the input window of 103 nucleotides to neural networks. The encoding is based on lower-order Markov models. Zeng *et al.* [7] used feature generation with n-gram frequencies that the total number of generated feature quite large. Feature selection methods were used to find the most relevant feature that evaluated on a variety of standard machine learning methods. In later work, Liu *et al.* [8] used n-gram amino acid patterns instead of n-gram nucleotide. Tzanis *et al.* [9] improved accuracy with a novel TIS prediction based on component that mapped the biological problems identified. While Zeng and Alhajj [10] proposed approach which used multiple agents, each of which investigated some distinct biological perspective.

A. n-gram

Let sequence $S = s_1 s_2 \ldots s_N$ and alphabet $s_i \varepsilon \{A_1, A_2 \ldots, A_m\}$ where $i = 1,2,\ldots,N$, where N be a length of sequence S and m be a length of alphabet s_i. An n-gram of the sequence S is any subsequence of consecutive alphabet s_i which long n. The i^{th} n-gram of S is the subsequence $s_i s_{i+1} \ldots s_{i+n-1}$, There are total possible formats of n-gram equal to m^n formats.

For example, give a sequence $S = AACCAGT$ and alphabet $s_i \varepsilon \{A, C, G, T\}$ where N equaled to 7 and m equaled to 4. Then 1-gram are A, C, G, and T, 2-gram are AA, AC, CC, CA, AG, and GT. The total possible formats of 2-gram equaled to 16 ($4^2 = 16$) formats. An n-gram used

for a long time in a wide variety of problems. There are four good characteristics of n-gram. First, it is straightforwardness that is relatively insensitive to spelling errors. Second, it is domain independence which is independent from language and topic. Third, it is efficiency of one processing. Finally, it is simplicity which no linguistic knowledge is required [11].

B. TF-IDF weighting

TF-IDF is the most common weighting method used to describe documents in the vector space model. The TF-IDF function weights each vector component which each of them relates to a word of the vocabulary of each document on the following basis. First, it incorporates the word frequency in the document. Therefore, the more words TF (Term Frequency) appear in a document, the more significance of this document is estimated. In addition, IDF (Inverse Document Frequency) measures how infrequent a word is in the collection. This value is estimated using the whole training text collection at hand. Accordingly, if a word is very frequent in the text collection, it is not considered to be particularly representative of this document since it occurs in most documents, for example, stop words. In contrast, if the word is infrequent in the text collection, it is believed to be very relevant for the document [12].

C. Correlation-based Feature Selection (CFS)

One of the problems that has to be overcome in classifying tasks is high data dimensionality. High data dimensionality affects virtually all classifier neural networks as well as the others resulting in less accurate classification of high dimensional data. The CFS method performs when the classification is done by a multilayer perceptron. The CFS requires less computation time and provides very high statistical significance [13].

The CFS algorithm is a heuristic for evaluating the merit of a subset of features. This heuristic takes into account the usefulness of individual features for predicting the class label along with the level of intercorrelation among them. The hypothesis on which the heuristic is "good feature subsets containing features highly correlated with the class, yet uncorrelated with each other" by the heuristic formula as in (1).

$$Merit_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k+1)\overline{r_{ff}}}} \quad (1)$$

where $Merit_S$ is the heuristic "merit" of a feature subset S containing k features, $\overline{r_{cf}}$ is the average feature-class correlation, and $\overline{r_{ff}}$ is the average feature-feature intercorrelation. This will give maximum value for the subset of features with high relation between feature and class and low relation between feature and feature of each group [15].

III. TF-IDF-NN-TIS MODEL

A TF-IDF and Neural Networks Approach for Translation Inititation Site Prediction model, called TF-IDF-NN-TIS, composes of 5 steps that are 1) sequence segmentation, 2) n-gram feature generation, 3) feature selection, 4) consensus pattern, and 5) TIS prediction. The details of each step shows in Fig. 1.

A. Sequence Segmentation

For each sequence, select all ATGs codon then segment sequence into subsequence with window size of 303 nucleotides. This means that the selected subsequence should have the number of nucleotide in upstream and downstream less than or equal to 150 nucleotides as shows in Fig. 2. The subsequence will be divided into 2 groups that are positive group and negative group. The subsequence that has the position nucleotide A of target ATG codon matched with the position of TIS will be classified as positive group while other subsequence will be classified as negative group.

Step 1: Sequence Segmentation
1.1 Select all ATGs from each sequence. 1.2 Divide sequence into subsequence with window size of 303 nucleotides. 1.3 Divide subsequence into 2 groups that are positive group and negative group.
Step 2: n-gram Feature Generation
2.1 Let n equal to 1 and 2. The DNA sequence composes of A, C, G, and T. 2.2 Create all possible formats of 1-gram and 2-gram in upstream and downstream, separately. 2.3 For each subsequence, there are 40 features, determine value for all features as follows. (a) Frequency value: count the frequency of each feature that appears in subsequence. (b) TF-IDF value: calculate TF-IDF value of all features in subsequence.
Step 3: Feature Selection
3.1 Select feature that creates from n-gram technique by correlation-based feature selection (CFS) method.
Step 4: Consensus Pattern *(Optional)*
4.1 Count the frequencies of ATG codon that appear in the upstream (up_ATG) 4.2 Count the frequencies of stop codon that appear in the downstream (dn_stop) 4.3 Consider position -3 as whether to be nucleotide A or G (up_-3A/G) 4.4 Consider position +4 as whether to be nucleotide G (dn_+4G)
Step 5: TIS Prediction
5.1 Specify the number of input nodes of neural networks equals to the number of features needed. 5.2 Predict TIS using MLP neural networks 5.3 Evaluate prediction result using k-fold cross validation.

Figure 1. TF-IDF-NN-TIS model.

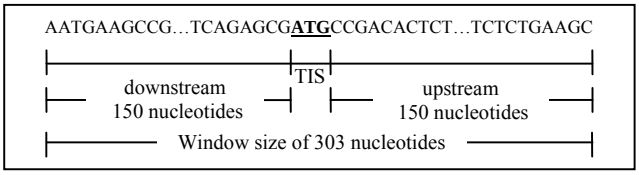


Figure 2. The sequence segmentation with window size of 303 nucleotides.

B. N-gram Feature Generation

We specify n equal to 1 and 2. The DNA sequence composes of 4 characters that are A, C, G, and T. Then 1-gram will have all possible formats equal to 4 ($4^1 = 4$) formats that are A, C, G, and T. The 2-gram will have all possible formats equal to 16 ($4^2 = 16$) formats that are AA, AC, AG, CA, CC, and etc. We considered on upstream and downstream, separately. This means that upstream will be 20 features (4+16=20) and downstream will be 20 features (4+16=20). For example, 1-gram in upstream terms denoted by up_A, and up_C, 1-gram in downstream terms denoted by dn_A and dn_C, 2-gram in upstream terms denoted by up_AA, up_AG, and etc, 2-gram in downstream denoted by dn_AA, dn_AG, and etc.

Let $T = (t_1, t_2, \ldots, t_M)$ be features created by 1-gram and 2-gram techniques with the total number of 40 features (M=40). Let $f_{t,s}$ be the frequency of feature t in sequence s. We used 2 ways for determining value of feature t in sequence s which were frequency value (denoted by $v_{t,s}$) and TF-IDF value (denoted by $v'_{t,s}$) [13].

1) Frequency value: This approach uses the frequency of the feature t in the sequence s which can be calculated by (2).

$$v_{t,s} = f_{t,s} \tag{2}$$

2) TF-IDF value: This approach takes into account the distribution of each 1-gram and 2-gram throughout all sequences in the training set which can be calculated by (3).

$$v'_{t,s} = \begin{cases} f_{t,s} \times log\dfrac{N}{n_t}; & if\ \ f_{t,s} \geq 1 \\ 0 \qquad\quad ; & otherwises \end{cases} \tag{3}$$

where N is the number of all sequences in the training set, and n_t is the total number of times feature t occurs in the training set.

C. Feature Selection

The total features received from 1-gram and 2-gram are 40 features, but there are only some features that are highly correlated with the class, but uncorrelated to each other. The CFS method was applied to keep the most significant features for the prediction of TIS (denoted by R) which can be calculated by (1). Note that the number of features selection from the CFS method will be less than or equal to the number of all features ($R \leq M$).

D. Consensus Pattern (optional)

Consensus pattern (denoted by CP) creates feature from the ribosome scanning model [1] and the conserved motif [2] from weight matrix model. The features created from the consensus patterns have the total number of 4 features (CP=4) as follows. 1) Count the frequencies of ATG codon that appear in the upstream (up_ATG) [3]. 2) Count the frequencies of stopped codon that appear in the downstream (dn_stop) [1]. The stopped codon is TAA, TAG, and TGA. 3) Consider position -3 as whether to be nucleotide A or G (up_-3A/G) [2] and 4) Consider position +4 as whether to be nucleotide G (dn_+4G) [2].

E. TIS Prediction

The prediction of TIS used feedforward multilayer perceptron neural networks and sigmoid activation function. To specify the architecture of neural networks, the input nodes are equal to the number of features needed. The hidden layer (denoted by H) composes of 10 nodes. The output layer (denoted by O) composes of 1 node. The evaluation of k-fold cross validation is used for the TIS classification.

Classification performance is measured by sensitivity (*Se*) and accuracy (*Acc*). Let *TP* be the number of the true positive ATGs classified as positive, *TN* be the number of the true negative ATGs classified as negative. *RP* is the number of the total true positive ATGs in the dataset, *R* is the number of the total dataset. *Se* is defined as *TP/RP*, the percentage of the correctly-predicted positives in the total true positives. *Acc* is defined as *(TP+TN)/R*, the percentage of the total correctly-predicted instances in all instances [17].

IV. Experiment and Result

The experiment was tested on personal computer with 2 gigabytes of memory, Intel(R) Core(TM) 2 Duo CPU T5550, and 1.83 gigahertz of processing speed. Experiment composes of 4 methods that are A, B, C, and D as shows in Table I. Method A uses all features of 1-gram and 2-gram. Method B uses feature of 1-gram and 2-gram that passes through the selection by CFS method. Method C combines all features of 1-gram and 2-gram and consensus pattern. Method D combines feature from n-gram that passes through the selection by CFS and consensus pattern which is the proposed (TF-IDF_NN_TIS) model. From Table I, M denoted by 1-gram and 2-gram features, R denoted by CFS feature, and CP denoted by consensus pattern feature.

TABLE I. The Experimental Design and Architecture of Neural Networks

Method	CFS (R)	Consensus Pattern (CP)	Architecture of Neural Networks
A	×	×	M:H:O
B	✓	×	R:H:O
C	×	✓	(M+CP):H:O
D	✓	✓	(R+CP):H:O

TABLE II. CHARACTERISTIC DATASET

Name	Min.	Max.	Mode	#Pos. ATGs	#Neg. ATGs	Positive/ Negative
Vertebrate	169	299	299	3312	10191	1/3
A.thaliana	169	299	299	523	1525	1/3
TIS+50	197	1112	469	50	469	1/9

TABLE III. COMPARISON THE ACCURACIES OF FREQUENCY VS. TF-IDF.

Method	Vertebrate (%)		A.thaliana (%)		TIS+50 (%)	
	Freq.	*TF-IDF*	*Freq.*	*TF-IDF*	*Freq.*	*TF-IDF*
A	93.97	99.53	89.40	**91.11**	88.95	**89.36**
B	92.18	**99.73**	93.52	**97.41**	90.39	90.80
C	96.81	98.73	94.78	95.17	92.02	**93.86**
D	98.05	**99.76**	97.41	**97.51**	93.26	**94.89**

TABLE IV. TIMES COMPARISON FOR MODELING OF FREQUENCY VS. TF-IDF VALUE ON ALL DATASETS.

Method	Vertebrate (seconds)		A.thaliana (seconds)		TIS+50 (seconds)	
	Freq.	*TF-IDF*	*Freq.*	*TF-IDF*	*Freq.*	*TF-IDF*
A	697	400	191	64	74	6
B	172	**79**	16	**7**	14	**2**
C	799	360	210	20	72	4
D	256	102	28	8	17	2

There are 3 datasets for this study that are Vertebrate, Arabidopsis thaliana, and TIS+50 as shows in Table II. The Vertebrate and Arabidopsis thaliana were constructed by Pedersen and Nielsen [3]. The datasets were originally extracted from Genbank and checked for suspicious annotations. The possible introns (noncoding region) are eliminated from all sequences. The upstream parts of the TIS are limited with at least 10 nucleotides and the downstream parts of the TIS are limited with at least 150 nucleotides. The TIS+50 datasets were constructed by Nadershahi *et al*. [16]. It is a standard EST dataset which contains 50 EST sequences. For each sequence in the dataset, there are one ATG as true TIS and other ATGs as false TIS. The Vertebrate composes of 13,503 ATGs which 3,312 (24.5%) of ATGs are true TISs. The Arabidopsis thaliana dataset composes of 2,048 ATGs which 523 (25.5%) of ATGs are true TISs. The TIS+50 composes of 519 ATGs which 50 (9.6%) of ATGs are true TISs.

The 10-folds cross validation is used for this study by dividing data into training set and testing set. The data will be divided into 10 equivalently parts. Then 9 parts will be used for training set and 1 part will be used for testing set. Data will be rotated 10 times for different testing set and different training set. The experimental results can be concluded into 5 issues that are 1) the efficiency of TF-IDF, 2) the efficiency of time for the TF-IDF, 3) the efficiency of CFS feature selection, 4) the efficiency of consensus pattern, and 5) the efficiency of TF-IDF-NN-TIS model.

1) Issue of the efficiency of TF-IDF value

Consider between TF-IDF value and frequency value, the experimental result shows that TF-IDF value received the accuracy higher than frequency value as shows in Table III. For example, method A of Vertebrate, A.thaliana, and TIS+50 dataset, TF-IDF value received the accuracy at 99.53%, 91.11%, and 89.36%, respectively, while the frequency value received the accuracy at 93.97%, 89.40%, and 88.95%, respectively.

2) Issue of the efficiecy of time for the TF-IDF

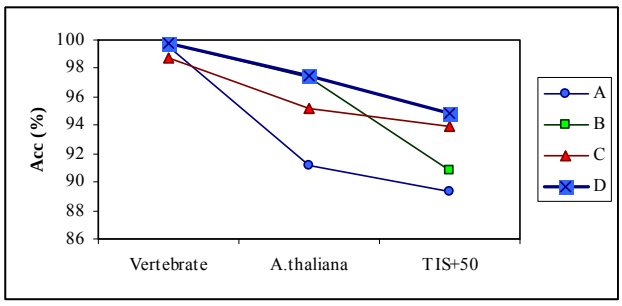
Consider between TF-IDF value and frequency value, the experimental result shows that the time for TF-IDF is less than the time for frequency as shows in Table IV. For example, method B of Vertebrate, A.thaliana, and TIS+50 dataset the time for TF-IDF are 79, 7, and 2 seconds, respectively which is less than the time for frequency which are 172, 16, and 14 seconds, respectively.

3) Issue of the efficiecy of CFS feature selection method

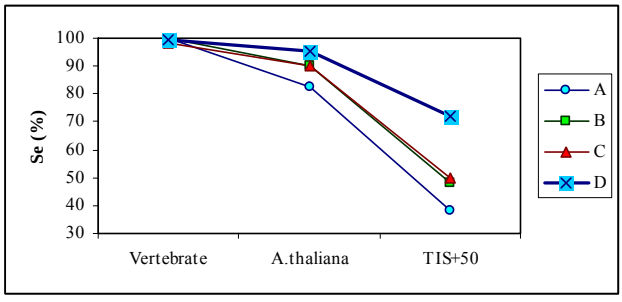
Consider between method B using CFS feature selection and method A without CFS feature selection, the experimental result shows that features through the selection of CFS method (method B) gives higher accuracy than 1-gram and 2-gram (method A) as shows in Table III. For example of A.thaliana, the TF-IDF of method B gave the accuracy at 97.41% which higher than method A at 91.11%. The time for method B is 7 seconds which is less than the time for method A at 64 seconds.

4) Issue of the efficiency of consensus pattern

Consider between method C with consensus pattern and method A without consensus pattern in the case of without CFS feature selection. The experimental result of the consensus pattern shows that method C gives higher accuracy than method A as shows in Table III. For example of TIS+50 dataset, TF-IDF of method C gave the accuracy at 93.86% which is higher than method A at 89.36%.



(a) The accuracy of TF-IDF-NN-TIS



(b) The sensitivity of TF-IDF-NN-TIS

Figure 3. The accuracy and the sensitivity of TF-IDF-NN-TIS model.

TABLE V. RESULTS OF PREVIOUS STUDIES.

Study	Year	Vertebrate	A.thaliana	TIS+50
Pedersen *et al.* [3]	1997	85.00	88.00	-
Zien *et al.* [4]	2000	88.10	-	-
Liu *et al.* [8]	2004	92.45	-	-
Rajapakse *et al.*[6]	2005	96.10	-	-
Tzanis et al. [9]	2007	97.26	97.07	-
Zeng *et al.* [10]	2007	96.68	-	91.82
Our approach	**2009**	**99.76**	**97.51**	**94.48**

Note that the consensus pattern is optional. This means that if there is no consensus pattern available from the previous researcher then the TF-IDF and CFS feature selection from method B also give high accuracy as mention in issue 3.

5) Issue of the efficiency of the purpose TF-IDF-NN-TIS model

The experimental result shows that the TF-IDF-NN-TIS (method D) gave highest accuracy and sensitivity than other methods on three datasets as shows in Table III and Fig. 3(a). For example, with TF-IDF of A.thaliana, method A, B, C, and D gave the accuracy at 91.11%, 97.41%, 95.17%, and 97.51%, respectively. The sensitivities are 82.79%, 90.06%, 90.25%, and 95.41%, respectively as shows in Fig. 3(b).

Table V compares the accuracy of the purposed TF-IDF-NN-TIS model with previous studies [3, 4, 6, 8, 9, 10] on three datasets of Vertebrate, A.thaliana, and TIS+50. For Vertebrate dataset, our approach gives the accuracy at 99.76% which is higher than Tzanis *et al.* [9] at 97.26%. For A.thaliana dataset, our approach gives the accuracy at 97.51% which is higher than Tzanis *et al.* [9] at 97.07%, For TIS+50 dataset, our approach gives the accuracy at 94.48% which is higher than Zeng *et al.* [10] at 91.82%. Therefore, the purposed TF-IDF-NN-TIS model gives higher accuracy than previous studies of the three datasets.

V. CONCLUSION

This paper proposes TF-IDF-NN-TIS model which aims to increase the accuracy for TIS prediction. The proposed model composes of the feature creation from 1-gram and 2-gram with TF-IDF value using CFS method for feature selection. The experimental result shows that proposed model received highest accuracy than the previous studies on all three datasets that are Vertebrate with the accuracy at 99.76%, A.thaliana with the accuracy at 97.51%, and TIS+50 with the accuracy at 94.48%. Moreover, the proposed model also uses less time.

For future work, we aim to apply TF-IDF-NN-TIS model for splice site and transcription start site prediction in genome sequence.

ACKNOWLEDGMENT

This paper receives financial support for graduate studies from Research Fund from the Faculty of Science at Prince of Songkla University, Hatyai Campus, Thailand for academic year 2007-2008.

REFERENCES

[1] M. Kozak, "The scanning model for translation: an update," J. Cell Biol.,vol. 108, pp. 229-241, 1989.

[2] M. Kozak, "An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs," Nucleic Acids Res., vol. 15, no. 20, pp. 8125-8148, 1987.

[3] A. G. Pedersen and H. Nielsen, "Neural Networks Prediction of Translation Initiation Sites in Eukaryotes: Perspective for EST and Genome analysis," in Proc. 5th Int. Conf. Intelligent Systems for Molecular Biology, pp. 226-233, 1997.

[4] A. Zien, G. Rätsch, S. Mika, B. Schölkopf, T. Lengauer, and K.R. Műller, "Engineering Support Vector Machine Kernels that Recognize Translation Initiation Sites," Bioinformatics, vol. 16, no. 9, pp. 799-807, 2000.

[5] A. Hatzigeorgiou, "Translation Initiation Start Prediction in Human cDNAs with High Accuracy," Bioinformatics, vol. 18, no. 2, pp. 343-350, 2002.

[6] J. C. Rajapakse and L. S. Ho, "Markov Encoding for Detecting Signals in Genomic Sequence," IEEE/ACM Trans. Comput. Biol. Bioinformatics, vol. 2, no. 2, pp. 131-142, 2005.

[7] F. Zeng, R. H.C. Yap, and L. Wong, "Using Feature Generation and Feature Selection for Accurate Prediction of Translation Initiation Sites," in Proc. 13th Inter. Conf. on Genome Informatics, pp. 192-200., 2002.

[8] H. Liu, H. Han, J. Li, and L. Wong, "Using Amino Acid Patterns to Accurately Predict Translation Initiation Sites," in Silico Biol., vol. 4, no. 3, pp. 255-269, 2004.

[9] G. Tzanis, C. Berberidis, an dI. Vlahavas, "MANTIS: A Data Mining Methodology for Effective Translation Initiation Site Prediction," in Proc. 29th Annual Inter. Conf. IEEE EMBS, pp. 6343-6347, 2007.

[10] J. Zeng and R. Alhajj, "Mluti-Agent System for Translation Initiation Site Prediction," 2007 IEEE Inter. Conf. on Bioinformatics and Biomedicine, pp. 103-108, 2007.

[11] A. Tomović, P. Janičić, V. Kešelj, "n-Gram-based Classification and Unspervised Hierarchical Clustering of Genome Sequence," J. Computer Methods and Programs in Biomedicine, vol. 81, pp. 137-153, 2006.

[12] T. Joachims, "A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization," Proc. 14th Inter. Conf. Machine Learning, pp. 143-151, 1997.

[13] Y. Yang and B.L. Lu, "Extracting features from protein sequences using chinese segmentation techniques for subcellular localization," Proc. 2005 IEEE Sym. on Computational Intelligence in Bioinformatics and Computational Biology, pp. 1-8, 2005.

[14] K. Michalak and H. Kwasnicka, "Correlation-based feature selection strategy in neural classification," in IEEE Proc. 6th Inter. Conf. Sys. Design and Application (ISDA'06), vol. 1, pp. 741-746 , 2006.

[15] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in Proc. 17th Inter. Conf. on Machine Learning, pp. 359-366, 2000.

[16] A. Nadershahi, S. C. Fahrenkrug, and L. Ellis, "Comparison of computational methods for identifying translation initiation sites in EST data," BMC Bioinformatics, vol. 5, 2004.

[17] R. J. Roiger and M. W. Geatz, "Data mining: A tutorial-based primer," Addison-Wisley, USA, 2003, pp.53-54.

ประวัติผู้เขียน

ชื่อ สกุล นางสาวฏรินทร คงมณี

รหัสประจำตัวนักศึกษา 5010220039

วุฒิการศึกษา

วุฒิ	ชื่อสถาบัน	ปีที่สำเร็จการศึกษา
วท.บ. (วิทยาการคอมพิวเตอร์)	มหาวิทยาลัยสงขลานครินทร์	2545

ทุนการศึกษา (ที่ได้รับในระหว่างการศึกษา)

ทุนสนับสนุนนักศึกษาระดับบัณฑิตศึกษาเป็นผู้ช่วยนักวิจัย (Research Assistant) จากกองทุนวิจัยคณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ ภาคการศึกษาที่ 2 ปีการศึกษา 2550

ตำแหน่งและสถานที่ทำงาน

ตำแหน่ง ครู

สถานที่ทำงาน โรงเรียนพรุพีพิทยาคม สำนักงานเขตพื้นที่การศึกษา สุราษฎร์ธานี เขต 3

การตีพิมพ์เผยแพร่ผลงาน

ฏรินทร คงมณี ศิริรัตน์ วณิชโยบล และวิภาดา เวทย์ประสิทธิ์. 2551. การทำนายตำแหน่งเริ่มต้นการแปลรหัสพันธุกรรมของลำดับดีเอ็นเอสัตว์มีกระดูกสันหลังโดยการเลือกลักษณะเฉพาะที่สัมพันธ์กันและโครงข่ายประสาทเทียม. The 12th Computer Science and Engineering Conference (NCSEC 2008). ชลบุรี ประเทศไทย. หน้า 209-216.

Kongmanee, T., Vanichayobon, S., and Wettayaprasit, W. 2009. The TF-IDF and Neural Networks Approach for Translation Initiation Site Prediction. The 2nd IEEE International Conference on Computer Science and Information Technology (IEEE ICCSIT 2009). August 8-11 2009, Bejing, China.