

บทที่ 4

การจำแนกชนิดข้อมูลของโมเลกุลเอชแอลเอโดยใช้สเปกตรัมเคอร์เนล

ในบทนี้จะเป็นการทดสอบกระบวนการวิธีของสเปกตรัมเคอร์เนลในการจำแนกข้อมูลสายลำดับโดยใช้วิธีซัพพอร์ตเวกเตอร์แมชชีน ที่ได้กล่าวไว้ในบทที่ 3 กรณีศึกษาคือ การจำแนกชนิดข้อมูลสายลำดับของโมเลกุลเอชแอลเอ โดยได้ แบ่งเนื้อหาเป็นส่วนๆดังนี้

- การเตรียมชุดข้อมูลของโมเลกุลเอชแอลเอ
- การวิเคราะห์กระบวนการวิธีดำเนินงาน
- การวัดประสิทธิภาพในการทำนาย
- ผลการทดลอง

4.1 การเตรียมชุดข้อมูลของโมเลกุลเอชแอลเอ

สำหรับกรณีศึกษาที่จะใช้ในงานวิจัยนี้คือ โมเลกุลของเอชแอลเอ ซึ่งสาเหตุในการศึกษาเนื่องจากโมเลกุลของเอชแอลเอ ดังกล่าวมีการแบ่งชนิดหรือกลุ่มออกเป็น 2 กลุ่มคือเอชแอลเอ กลุ่มที่ I และเอชแอลเอกลุ่มที่ II ตามที่บทที่ 2 ได้กล่าวไว้ โดยโครงสร้างหรือการจัดเรียงตัวของข้อมูลในแต่ละกลุ่มนั้นมีการจัดเรียงที่ความคล้ายคลึงกัน จึงต้องมีการสร้างผู้ช่วยในการตัดสินใจเพื่อจำแนกชนิดของโมเลกุลเอชแอลเอนั้นว่าอยู่ในกลุ่มใด ซึ่งต้องใช้การวิเคราะห์ที่ค่อนข้างละเอียด ดังนั้นจึงมีความจำเป็นในการนำสเปกตรัมเคอร์เนลมาใช้ในการจำแนกข้อมูลสายลำดับ เนื่องจากวิธีการนี้เหมาะสมสำหรับข้อมูลที่อยู่ในรูปของสายลำดับ

เนื่องจากการจัดเรียงโมเลกุลของเอชแอลเอนั้น จะอยู่ในรูปของดีเอ็นเอ ซึ่งจะมีองค์ประกอบของโมเลกุลอยู่ 4 เบสได้แก่ A, T, C, G ซึ่งจะพบภายในนิวคลีอไทด์ทั้งหมด โดยการจัดเรียงแต่ละโมเลกุลของเอชแอลเอนั้น จะมีการจัดเรียงและความยาวของข้อมูลสายลำดับที่แตกต่างกัน ไปตามชนิดและหน้าที่เฉพาะของโมเลกุลเอชแอลเอแต่ละตัว โดยรูปแบบการจัดเรียงของโมเลกุลเอชแอลเอนั้นสามารถแสดงได้ดังรูป 4.1

Copyright © by Chiang Mai University
All rights reserved

```

>HLA:HLA00001 A*01010101
ATGGCCGTCATGGCGCCCCGAACCTCCTCCTGCTACTCTCGGGGGCCCTGGCCCTGA
CATCCGTGTCCCGGCCCGGCCGCGGGGAGCCCCGCTTCATCGCCGTGGGCTACGTGGA
GAAGATGGAGCCCGCGGCCGCGTGGATAGAGCAGGAGGGGCGGGAGTATTGGGACCAG
AACCTGGGGACCCTGCGCGGCTACTACAACCAGAGCGAGGACGGTTCTCACACCATCC
TCCGCGGGTACCGGCAGGACGCCCTACGACGGCAAGGATTACATCGCCCTGAACGAGGA
CAAGCGCAAGTGGGAGGCGGCTCCATGCGGCGGAGCAGCGGAGAGTCTACCTGGAGGGC
GAGACGCTGCAGCGCACGGACCCCCCAAGACACATATGACCCACACCCCATCTCTG
CTGCGGAGATCACACTGACCTGGCAGCGGGATGGGAGGACCAGACCCAGGACACGGA
GTGGGCGGCTGTGGTGGTGCCTTCTGGAGAGGAGCAGAGATACACCTGCCATGTGCAG
TCTTCCCAGCCACCATCCCATCGTGGGCATCAITGCTGGCCTGGTTCTCCTTGGAG
GGAAGAGCTCAGATAGAAAAGGAGGGAGTTACACTCAGGCTGCAAGCAGTGACAGTGC
>HLA:HLA02169 A*01010102N
ATGGCCGTCATGGCGCCCCGAACCTCCTCCTGCTACTCTCGGGGGCCCTGGCCCTGA
CATCCGTGTCCCGGCCCGGCCGCGGGGAGCCCCGCTTCATCGCCGTGGGCTACGTGGA
GAAGATGGAGCCCGCGGCCGCGTGGATAGAGCAGGAGGGGCGGGAGTATTGGGACCAG
AACCTGGGGACCCTGCGCGGCTACTACAACCAGAGCGAGGACGGTTCTCACACCATCC
TCCGCGGGTACCGGCAGGACGCCCTACGACGGCAAGGATTACATCGCCCTGAACGAGGA
CAAGCGCAAGTGGGAGGCGGCTCCATGCGGCGGAGCAGCGGAGAGTCTACCTGGAGGGC
GAGACGCTGCAGCGCACGGACCCCCCAAGACACATATGACCCACACCCCATCTCTG
CTGCGGAGATCACACTGACCTGGCAGCGGGATGGGAGGACCAGACCCAGGACACGGA
GTGGGCGGCTGTGGTGGTGCCTTCTGGAGAGGAGCAGAGATACACCTGCCATGTGCAG
TCTTCCCAGCCACCATCCCATCGTGGGCATCAITGCTGGCCTGGTTCTCCTTGGAG
GGAAGAGCTCAGATAGAAAAGGAGGGAGTTACACTCAGGCTGCAAGCAGTGACAGTGC

```

รูป 4.1 ลักษณะของข้อมูลสายลำดับของโมเลกุลเอชแอลเอที่นำมาวิเคราะห์
(ที่มา: <http://www.ebi.ac.uk/imgt/hla/>)

ชุดข้อมูลที่ใช้ในการวิเคราะห์จะเป็นข้อมูลเข้าในการเรียนรู้ด้วยเครื่อง เพื่อใช้ในการจดจำรูปแบบของข้อมูลเข้า ซึ่งนำไปสู่ตัวแบบจำลองที่ทำการศึกษา โดยชุดข้อมูลเหล่านี้จะเป็นตัวแบบของกลุ่มยีน ซึ่งข้อมูลที่ศึกษานั้นได้ใช้ข้อมูลจากฐานข้อมูลไอเอ็มจีที/เอชแอลเอ (IMGT/HLA) โดยสามารถดาวน์โหลดชุดข้อมูลในการเรียนรู้ได้จาก <http://www.ebi.ac.uk/imgt/hla> เนื่องจากโมเลกุลของเอชแอลเออยู่ในรูปแบบสายลำดับ (Sequence) และแบ่งออกเป็น 2 กลุ่มคือเอชแอลเอ กลุ่มที่ I และเอชแอลเอกลุ่มที่ II ซึ่งโมเลกุลของเอชแอลเอสามารถแบ่งออกเป็นกลุ่มย่อยได้ และมีหน้าที่การทำงานที่แตกต่างกันออกไปในแต่ละชนิด โดยกลุ่มย่อยของเอชแอลเอที่ I จะประกอบด้วย A, B, Cw, E, F, G ส่วนกลุ่มย่อยของเอชแอลเอกลุ่มที่ II จะประกอบด้วย DMA, DMB, DOA, DOB, DPA1, DPB1, DQA1, DQB1, DRA, DRB โดยรายละเอียดข้อมูลจะแสดงตามตาราง 4.1

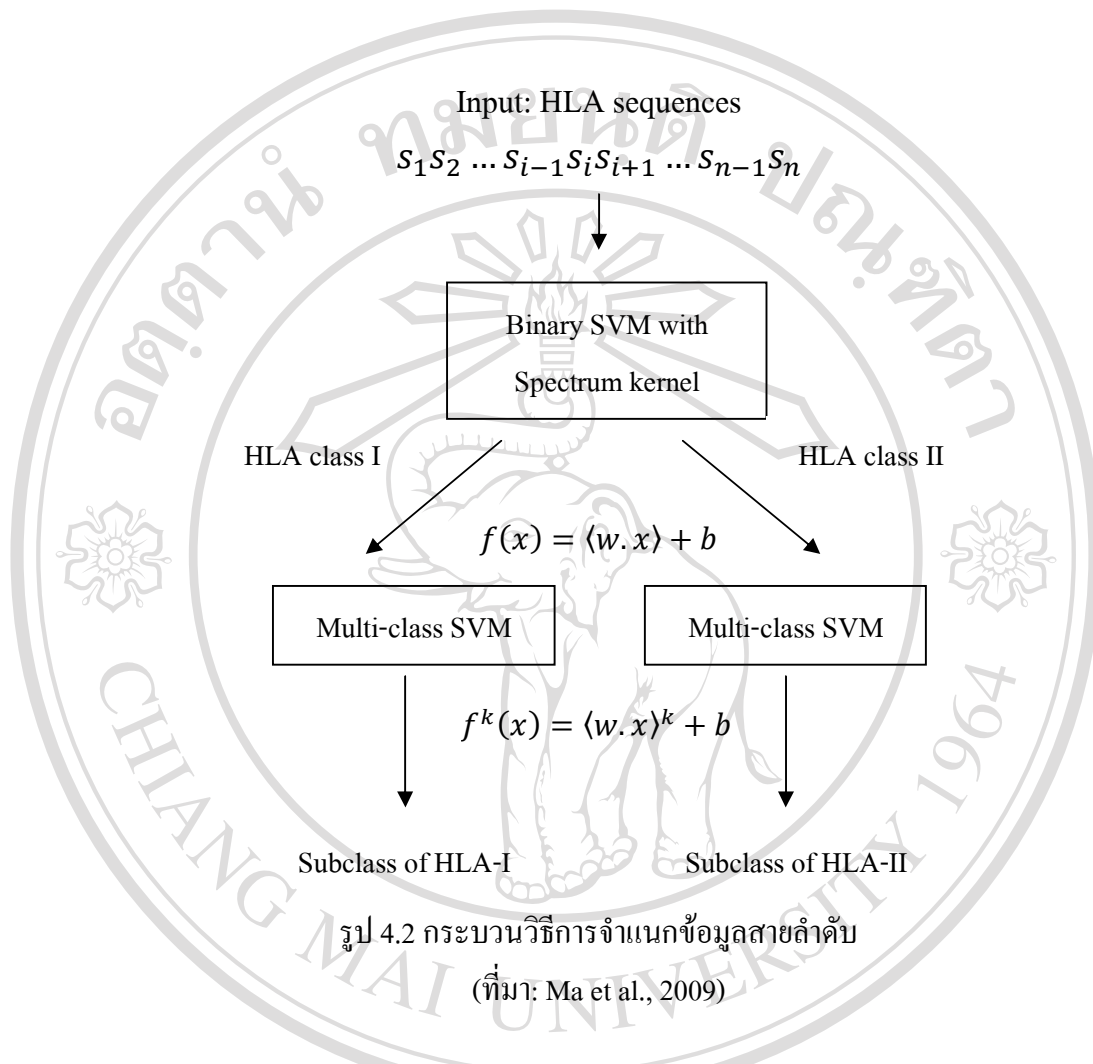
ตาราง 4.1 รายละเอียดของชุดข้อมูลที่นำมาวิเคราะห์

กลุ่ม	กลุ่มย่อย	จำนวนของข้อมูล สายลำดับ
HLA-I	HLA-A	965
	HLA-B	1,540
	HLA-C	626
	HLA-E	9
	HLA-F	21
	HLA-G	45
	รวม	3,206
HLA-II	HLA-DMA	4
	HLA-DMB	7
	HLA-DOA	12
	HLA-DOB	9
	HLA-DPA1	27
	HLA-DPB1	138
	HLA-DQA1	35
	HLA-DQB1	107
	HLA-DRA	3
	HLA-DRB	855
	รวม	1,197
รวม		4,403

4.2 การวิเคราะห์กระบวนการวิธีดำเนินงาน

ในการวิเคราะห์กระบวนการวิธีดำเนินงานนั้น จะเริ่มตั้งแต่การนำข้อมูลเข้าจากฐานข้อมูลซึ่งข้อมูลเข้าแต่ละตัวจะอยู่ในรูปแบบของสายลำดับ โดยป้อนข้อมูลของโมเลกุลของเอชแอลเอทั้งหมดให้กับเครื่องเพื่อทำการฝึกสอน โดยมีการปรับคุณลักษณะให้อยู่ในรูปของเวกเตอร์ ซึ่งเครื่องสามารถจะทำการรู้จำรูปแบบและคุณลักษณะต่างๆ ของข้อมูลได้ เมื่อเครื่องทำการรู้จำคุณลักษณะต่างๆ จากนั้นเราจะใช้ซอฟต์แวร์เวกเตอร์แมชชีนเป็นตัวแบบจำลองในการจำแนกข้อมูล ซึ่งตัวแบบจำลองที่ใช้นั้นจะอยู่ในรูปแบบเคอร์เนลซอฟต์แวร์เวกเตอร์แมชชีน (เคเอสวีเอ็ม) ได้แก่ เคอร์เนลฟังก์ชันแบบเชิงเส้น โพลีโนเมียล แกสเซียน และวิธีการที่เรานำเสนอคือ สเปกตรัมเคอร์เนล เป็นตัวจำแนกข้อมูลออกเป็น 2 กลุ่มคือโมเลกุลของเอชแอลเอกลุ่มที่ I และ เอชแอลเอกลุ่มที่

II จากนั้นเราจึงทำการจำแนกข้อมูลสายลำดับของโมเลกุลเอชแอลเอในกลุ่มย่อยของแต่ละกลุ่มหลัก ด้วยวิธีการเดียวกันต่อไป โดยจะแสดงในรูป 4.2



จากรูปข้างต้น จะเห็นได้ว่าเมื่อทำการจำแนกชนิดข้อมูลสายลำดับของโมเลกุลเอชแอลเอได้แล้ว จากนั้นจะทำการวัดประสิทธิภาพในการทำนายของการจำแนกชนิดข้อมูลสายลำดับของโมเลกุลเอชแอลเอตามหัวข้อที่ 4.3 ต่อไป

ก่อนอื่นเราจะแสดงให้เห็นถึงลักษณะของโมเลกุลเอชแอลเอทั้ง 2 กลุ่ม โดยจะแสดงให้เห็นว่าการจัดเรียงของโมเลกุลเอชแอลเอทั้ง 2 กลุ่มนั้นมีความแตกต่างกัน แต่จะมีลักษณะเฉพาะของในแต่ละโมเลกุลแต่ละกลุ่มดังรูปที่ 4.3

HLA	Type	Sequence
HLA:HLA00132	B*	GCTCCCACTCCATGAGGTATTTCTACACCTCCGTGTCCCGGCCCGGCCGCGG
HLA:HLA00133	B*	GCTCCCACTCCATGAGGTATTTCTACACCTCCGTGTCCCGGCCCGGCCGCGG
HLA:HLA00134	B*	GCTCCCACTCCATGAGGTATTTCTACACCTCCGTGTCCCGGCCCGGCCGCGG
HLA:HLA01763	B*	GCTCCCACTCCATGAGGTATTTCTACACCTCCGTGTCCCGGCCCGGCCGCGG
HLA:HLA02949	B*	GCTCCCACTCCATGAGGTATTTCTACACCTCCGTGTCCCGGCCCGGCCGCGG
HLA:HLA02990	B*	GCTCCCACTCCATGAGGTATTTCTACACCTCCGTGTCCCGGCCCGGCCGCGG
HLA:HLA03089	B*	GCTCCCACTCCATGAGGTATTTCTACACCTCCGTGTCCCGGCCCGGCCGCGG
HLA:HLA00514	DE	CTGTGGTCCAGGGCAGGGCCACTCCAGAGAAATTACGTGTACCAGGGACGGCA
HLA:HLA00515	DE	CTGTGGTCCAGGGCAGGGCCACTCCAGAGAAATTACGTGTACCAGGGACGGCA
HLA:HLA00517	DE	CTGTGGTCCAGGGCAGGGCCACTCCAGAGAAATTACCTTTCCAGGGACGGCA
HLA:HLA00519	DE	CTGTGGTCCAGGGCAGGGCCACTCCAGAGAAATTACCTTTCCAGGGACGGCA

รูป 4.3 ลักษณะของโมเลกุลเอชแอลเอกลุ่มที่ I และกลุ่มที่ II

4.3 การวัดประสิทธิภาพในการทำนาย

ในหัวข้อนี้จะเป็นการนำเสนอวิธีการในการประเมินค่าของประสิทธิภาพในการทำนาย การจำแนกชนิดข้อมูลสายลำดับของโมเลกุลเอชแอลเอ โดยจะใช้ค่าการวัดที่เป็นมาตรฐาน เพื่อใช้ในการเปรียบเทียบกับเทคนิคทันสมัยอื่น ต่อไป ได้แก่ ค่าความถูกต้อง (ACC), ค่าความแม่นยำ (Prec), ค่าความไว (Sens), ค่าสัมประสิทธิ์ความสัมพันธ์แมทริว (MCC)

ค่าความถูกต้อง เป็นค่าสัดส่วนของข้อมูลสายลำดับของ x ทุกกรณีที่น่ามาวิเคราะห์ ในการทำนายที่ถูกต้องทั้งหมด ซึ่งเทียบด้วยอัตราส่วนของหน่วยวัดทุกกรณี สำหรับการจำแนกชนิดข้อมูลสายลำดับของโมเลกุลเอชแอลเอ จะกำหนดได้ดังนี้:

$$\text{ความถูกต้อง} = \frac{TP_x + TN_x}{TP_x + TN_x + FP_x + FN_x} \quad (4.1)$$

โดยที่ x คือ ข้อมูลสายลำดับของโมเลกุลเอชแอลเอกลุ่มที่ I หรือเอชแอลเอกลุ่มที่ II

ค่าความแม่นยำ เป็นค่าสัดส่วนของข้อมูลสายลำดับของ x ทุกกรณีที่น่ามาวิเคราะห์แล้ว ตรงกับการวิเคราะห์โดยแบบจำลองส่วนด้วยข้อมูลที่ถูกต้องทำนายโดยแบบจำลองทั้งหมด ในการจำแนกกลุ่มข้อมูลของเอชแอลเอโดยจะกำหนดได้ดังนี้:

$$\text{ความแม่นยำ} = \frac{TP_x}{TP_x + FP_x} \quad (4.2)$$

โดยที่ FP_x คือค่าของผลรวมที่ผิดแต่ทำนายเป็นบวกสำหรับ x ทุกกรณี

TP_x คือค่าของผลรวมที่ถูกแต่ทำนายเป็นบวกสำหรับ x ทุกกรณี

ค่าความไว เป็นค่าสัดส่วนของข้อมูลสายลำดับของ x ทุกกรณีที่นำมาวิเคราะห์แล้ว ตรงกับการวิเคราะห์โดยแบบจำลองส่วนด้วยข้อมูลที่ถูกทำนายโดยแบบจำลองทั้งหมด ในการ จำแนกกลุ่มข้อมูลของเอชแอลเอโดยจะกำหนดได้ดังนี้:

$$\text{ความไว} = \frac{TP_x}{TP_x + FN_x} \quad (4.3)$$

โดยที่ FN_x คือค่าของผลรวมที่ผิดแต่ทำนายเป็นลบสำหรับ x ทุกกรณี

TN_x คือค่าของผลรวมที่ถูกแต่ทำนายเป็นลบสำหรับ x ทุกกรณี

ค่าสัมประสิทธิ์ความสัมพันธ์แมทริว เป็นค่าสัดส่วนของข้อมูลสายลำดับของ x ทุกกรณี ที่ นำมาวิเคราะห์ทั้งหมด ซึ่งเทียบด้วยอัตราส่วนของหน่วยวัดทุกกรณี ซึ่งครอบคลุมการพิจารณาของ ค่าวัดประสิทธิภาพทั้งหมด สำหรับการจำแนกชนิดข้อมูลสายลำดับของ โมเลกุลเอชแอลเอ จะกำหนดได้ดังนี้:

$$\text{ค่าสัมประสิทธิ์ความสัมพันธ์แมทริว} = \frac{N_i}{N_j} \quad (4.4)$$

โดย $N_i = TP_x TN_x - FP_x FN_x$

$$N_j = \sqrt{(TP_x + FN_x)(TP_x + FP_x)(TN_x + FP_x)(TN_x + FN_x)}$$

4.4 ผลการทดลอง

สตริงเคอร์เนลเป็นเครื่องมือทางคณิตศาสตร์ที่ใช้กันอย่างกว้างขวาง ส่วนใหญ่จะใช้ในการ วิเคราะห์ข้อมูลที่อยู่ในรูปแบบของข้อมูลสายลำดับ เพื่อจะจัดกลุ่มหรือจำแนกชนิดของข้อมูล โดยเฉพาะอย่างยิ่ง งานวิจัยทางด้านข้อความและการวิเคราะห์กลุ่มยีน ซึ่งวิธีการทำเรานำเสนอนั้น คือ สเปกตรัมเคอร์เนล ซึ่งเป็นส่วนหนึ่งของสตริงเคอร์เนล เนื่องจากวิธีการนี้จะเหมาะสมกับข้อมูล ที่มีลักษณะเป็นสายลำดับ โดยจะทำการเปรียบเทียบความคล้ายหรือความเหมือนของข้อมูลสาย ลำดับทั้งหมด เป็นต้น

ซึ่งในการวัดประสิทธิภาพในการจำแนกชนิดข้อมูลสายลำดับของ โมเลกุลเอชแอลเอ นั้น วิธีการที่นำเสนอคือ สตริงเคอร์เนลซึ่งจะใช้สเปกตรัมเคอร์เนลเป็นฟังก์ชันในการจำแนกชนิด

ข้อมูลสายลำดับของโมเลกุลเอชแอลเอ โดยเราจะทำการแบ่งขนาดหรือความถี่ของช่วงการพิจารณาข้อมูล โดยใช้หลักการของ เก-สเปกตรัม ซึ่งจะแบ่งระดับการพิจารณาออกเป็น 7 ระดับ คือระดับ 3 ถึงระดับ 9 โดยมีการวัดประสิทธิภาพของการจำแนกชนิดข้อมูลสายลำดับของโมเลกุลเอชแอลเอ ด้วยค่าความถูกต้อง ค่าความแม่นยำ ค่าความไว และค่าสัมประสิทธิ์ความสัมพันธ์-แมทริวของโมเลกุลเอชแอลเอทั้ง 2 กลุ่มได้ดังตาราง 4.2

ตาราง 4.2 การวัดประสิทธิภาพในการจำแนกชนิดข้อมูลสายลำดับของโมเลกุลเอชแอลเอ โดยใช้สเปกตรัมเคอร์เนลในความถี่ที่แตกต่างกัน

เก- สเปกตรัม	เอชแอลเอ-I		เอชแอลเอ-II	
	<i>ACC</i>	<i>MCC</i>	<i>Prec</i>	<i>Sens</i>
3	98.6	98.6	99.6	98.5
4	98.8	97.0	99.8	98.6
5	98.5	96.4	99.8	98.3
6	99.6	99.0	99.8	99.7
7	99.2	98.0	99.4	99.5
8	98.6	96.6	99.5	98.6
9	99.1	97.7	99.8	99.0

จากตารางข้างต้นจะเห็นได้ว่า 3-สเปกตรัมหรือโคดอน เป็นคุณลักษณะที่มีความสำคัญในการศึกษาระดับโมเลกุล ซึ่งได้ศึกษาถึงความเหมือนกันของข้อมูลในรูปแบบของโมเลกุล โดยโมเลกุลจะมีรูปแบบการจัดเรียงที่แตกต่างกันออกไปตามแต่ละชนิดของสิ่งมีชีวิต ซึ่งในสิ่งมีชีวิตแต่ละชนิดก็จะมีคุณลักษณะเด่นที่จำเพาะ ซึ่งบ่งชี้ถึงตัวตนของสิ่งมีชีวิตนั้น โดยจะแสดงออกในระดับโมเลกุล (Milhon et al., 1995. และ Mitreva et al, 2007.) ซึ่งผลลัพธ์ที่ได้คือค่าความถูกต้องเท่ากับ 98.6% ส่วนค่าความถูกต้องในการทำนายของเก-สเปกตรัม เท่ากับ 4 และ 5 นั้นจะได้ผลลัพธ์เท่ากับ 98.8% และ 98.5% ตามลำดับ ส่วนการใช้สเปกตรัมเคอร์เนลที่มีความถี่เท่ากับ 6 นั้นจะส่งผลให้ประสิทธิภาพของการทำนาย ได้ผลลัพธ์เป็นที่น่าสนใจ คือได้ค่าความถูกต้องและความแม่นยำเท่ากับ 99.6% และ 99.0% ตามลำดับ โดยในส่วนของคุณค่าความแม่นยำในข้อมูลสายลำดับของโมเลกุลเอชแอลเอกลุ่มที่ I จะได้ค่าความถูกต้องเท่ากับ 99.8% และโมเลกุลเอชแอลเอกลุ่มที่ II จะได้เท่ากับ 99.1% จากผลการทดลองนี้สามารถสรุปได้ว่า สเปกตรัมเคอร์เนลที่มีความถี่เท่ากับ 6 หรือเรียกว่าได-โคดอน (Di – codon) นั้นเป็นคุณสมบัติหนึ่งที่สำคัญสำหรับการจำแนกชนิดของโมเลกุล

เอชแอลเอ ซึ่งสามารถสร้างคุณลักษณะที่มีผลต่อการจำแนกกลุ่มของโมเลกุลเอชแอลเอในแต่ละกลุ่ม ดังนั้น 6- สเปกตรัม จึงเป็นตัวบ่งชี้ที่ในการแสดงออกของยีนในการศึกษาวิวัฒนาการของสิ่งมีชีวิตในระดับโมเลกุล ดังแสดงได้ในตาราง 4.2

ในการจำแนกชนิดของโมเลกุลในสิ่งมีชีวิตนั้น (Nguyen et al., 2009) เราจะทำการศึกษาผลลัพ์ที่ดีที่สุดในการตรวจสอบชนิดของข้อมูลสายลำดับของโมเลกุลเอชแอลเอ โดยทำการเพิ่มความยาวของสเปกตรัมที่มีความถี่เท่ากับ 7, 8 และ 9 ซึ่งจะเห็นได้ว่าผลลัพ์ที่ได้มีค่าต่ำกว่าผลลัพ์ที่ได้จากการใช้สเปกตรัมที่มีความถี่เท่ากับ 6 โดยผลลัพ์ของค่าความถูกต้องในเค-สเปกตรัมเท่ากับ 7 และ 8 คือ 99.2% และ 98.6% ตามลำดับ นอกจากนี้ในวิธีการทดลองที่น่าเสนอนี้ได้ทดสอบ เค-สเปกตรัมที่มีความถี่เท่ากับ 9 หรือที่เรียกว่าไตร-โคดอน (Tri-codon) ซึ่งจะเห็นได้ว่าผลลัพ์ของค่าความถูกต้องเป็น 99.1%

โดยจากผลลัพ์ทั้งหมดนี้ สามารถสรุปได้ว่าเมื่อมีการเพิ่มความยาวในการพิจารณาข้อมูลสายลำดับแต่ละโมเลกุลมากขึ้นนั้น ไม่ได้หมายความว่าผลลัพ์ที่มีประสิทธิภาพมากขึ้น แต่ในทางกลับกันเมื่อเพิ่มความยาวในการพิจารณาที่มากขึ้น จะทำให้ใช้เวลาในการพิจารณาโมเลกุลแต่ตำแหน่งมากขึ้น แต่ค่าความถูกต้องไม่ได้มากขึ้นตามไปด้วย ส่วนการพิจารณาความยาวของข้อมูลน้อยเกินไปนั้น จะทำให้การพิจารณาข้อมูลสายลำดับทั้งหมดนี้ไม่ครอบคลุมทุกกรณี อาจทำให้ขาดคุณสมบัติต่างๆ ที่ทำให้ผลลัพ์น้อยลงหรือทำให้ขาดความจำเพาะหรือลักษณะเด่นของข้อมูล เป็นต้น

4.4.1 สเปกตรัมเคอร์เนลเปรียบเทียบกับวิธีการทันสมัยอื่นในการจำแนกชนิดข้อมูลของเอชแอลเอกลุ่มหลัก

ในการเปรียบเทียบกับวิธีการทันสมัยอื่น ๆ สำหรับการจำแนกชนิดข้อมูลสายลำดับของโมเลกุลเอชแอลเอ ซึ่งใช้หลักเกณฑ์เดียวกันในการวิเคราะห์ข้อมูล โดยทำการ 10 ครอสวาริเดชั่น (10-Cross validation) กับชุดข้อมูลเข้าทั้งหมดของโมเลกุลเอชแอลเอ ดังตาราง 4.3 ซึ่งในการเปรียบเทียบกับวิธีการทันสมัยอื่นๆ กับวิธีที่เรานำเสนอนั้น เราได้เลือกวิธีการต่างๆ ที่มีรูปแบบในการตรวจสอบที่เป็นที่รู้จักและใช้กันอย่างแพร่หลาย ในงานด้านการจำแนกประเภทหรือชนิดของข้อมูล ได้แก่ โครงข่ายประสาทเทียมและเคอร์เนลซัพพอร์ตเวกเตอร์แมชชีน (เคเอสวีเอ็ม) (Ripley et al., 2009) ซึ่งในการตรวจสอบของเคเอสวีเอ็มนั้น จะใช้เคอร์เนลฟังก์ชันต่างๆ ได้แก่ เคอร์เนลฟังก์ชันแบบเชิงเส้นและแบบโพลิโนเมียลที่มีมิติ (d) = 2, 3, 4 และแบบเกาส์เซียน ซึ่งการจำแนกชนิดข้อมูลด้วยวิธีการทันสมัยนี้ จะต้องทำการแปลงข้อมูลนำเข้าของโมเลกุลเอชแอลเอให้อยู่ในรูปของเวกเตอร์ ที่มีคุณลักษณะทั้งหมดที่เป็นไปได้เท่ากับ 59 มิติ (Ma et al., 2009) ด้วยวิธี

อาเอสซียู ตามบทที่ 2 ที่ได้กล่าวไว้ ซึ่งวิธีการทันสมัยเหล่านี้จะต้องใช้เวลาในการดำเนินการนาน และยากต่อการแปลงคุณลักษณะของข้อมูล

ตาราง 4.3 การเปรียบเทียบประสิทธิภาพของวิธีการสเปกตรัมเคอร์เนล
กับวิธีการทันสมัยในการจำแนกชนิดข้อมูลของโมเลกุลเอชแอลเอกลุ่มหลัก

วิธีการ และ พารามิเตอร์		เอชแอลเอ-I				เอชแอลเอ-II	
		Acc	MCC	Prec	Sens	Prec	Sens
โครงข่ายประสาท	size=2	98.5	96.6	87.6	99	99.2	95.4
เชิงเส้น	c=20	91.4	98.3	97.7	99.1	99.2	98.9
โพลีโนเมียล	d=2	99	98.9	98.8	99.2	99.3	99.3
โพลีโนเมียล	d=3	99.3	99.2	99.1	99.4	99.7	99.5
โพลีโนเมียล	d=4	99	98.8	98.8	99	99.3	99.4
เกาส์เซียน	sig=0.01	99.5	99.5	99.5	99.8	99.8	99.7
สเปกตรัมเคอร์เนล	k=6	99.6	99.0	99.8	99.7	99.1	99.4

จากตารางข้างต้น จะเห็นได้ว่า ผลลัพธ์ในการเปรียบเทียบกับวิธีการทันสมัย ในการจำแนกชนิดข้อมูลสายลำดับของโมเลกุลเอชแอลเอ ที่ได้นำเสนอในตาราง 4.3 นั้นวิธีการที่เรานำเสนอคือการใช้สเปกตรัมเคอร์เนลที่ความถี่เท่ากับ 6 โดยผลลัพธ์ที่ได้ นั่นคือค่าความถูกต้องเท่ากับ 99.6% และค่าความแม่นยำของโมเลกุลเอชแอลเอกลุ่มที่ I เท่ากับ 99.8% ซึ่งเป็นวิธีการที่ได้ ผลลัพธ์ดีกว่าวิธีการคำนวณแบบอื่น ถึงแม้ว่าผลลัพธ์ในการจำแนกชนิดข้อมูลของโมเลกุลเอชแอลเอกลุ่มที่ II นั้นจะได้ค่าความแม่นยำและค่าความไว เท่ากับ 99.1% และ 99.4% ตามลำดับ แต่ในการทดลองโดยใช้เคอร์เนลแบบเกาส์เซียนนั้น จะได้ผลลัพธ์ที่ดีกว่าวิธีการที่เรานำเสนอเล็กน้อยคือค่าสัมประสิทธิ์ความสัมพันธ์แมทริวและค่าความไวในการจำแนกชนิดข้อมูลของโมเลกุลเอชแอลเอกลุ่มที่ I และค่าความไวกับค่าความแม่นยำในการจำแนกชนิดข้อมูลของโมเลกุลเอชแอลเอกลุ่มที่ II แต่ก็ได้ผลลัพธ์ที่ใกล้เคียงกับฟังก์ชันเคอร์เนลแบบเกาส์เซียน เป็นต้น

4.4.2 สเปกตรัมเคอร์เนลเปรียบเทียบกับวิธีการทันสมัยอื่นในการจำแนกชนิดข้อมูลของ เอชแอลเอกลุ่มย่อย

ในการจำแนกชนิดข้อมูลสายลำดับของโมเลกุลเอชแอลเอกลุ่มย่อยของเอชแอลเอกลุ่มที่ I และโมเลกุลเอชแอลเอกลุ่มที่ II นั้น วิธีการที่เรานำเสนอคือสเปกตรัมเคอร์เนลที่มีความถี่เท่ากับ 6 โดยรายละเอียดของโมเลกุลเอชแอลเอ ในแต่ละกลุ่มสามารถอธิบายได้ ดังตาราง 4.1 ซึ่งผลลัพธ์ของการจำแนกชนิดข้อมูลสายลำดับของโมเลกุลเอชแอลเอกลุ่มย่อย จะแสดงตามตาราง 4.4 โดยจะเห็นว่าฟังก์ชันเคอร์เนลแบบเกาส์เซียน ให้ผลลัพธ์ที่มีประสิทธิภาพดีที่สุดของความถูกต้องในกลุ่มย่อยของเอชแอลเอกลุ่มที่ I และกลุ่มย่อยของเอชแอลเอกลุ่มที่ II เท่ากับ 99.6% และ 99.4% ตามลำดับ ในการจำแนกชนิดของข้อมูลสายลำดับนี้ อย่างไรก็ตามการใช้สเปกตรัมเคอร์เนลที่เราเสนอนั้น ได้ผลลัพธ์ของความถูกต้องในกลุ่มย่อยของเอชแอลเอกลุ่มที่ I และกลุ่มย่อยของเอชแอลเอกลุ่มที่ II เท่ากับ 99.4% และ 99.0% ตามลำดับ ซึ่งใกล้เคียงกับวิธีการคำนวณของฟังก์ชันเคอร์เนลแบบเกาส์เซียน นอกจากนี้การวัดประสิทธิภาพในการทำนายของวิธีการที่เราเสนอนั้น ได้ผลลัพธ์ที่ดีกว่าวิธีการของโครงข่ายประสาทเทียมและเคเอสวีเอ็มที่ใช้ฟังก์ชันเคอร์เนลแบบเชิงเส้นและแบบโพลิโนเมียล เป็นต้น

ตาราง 4.4 การเปรียบเทียบประสิทธิภาพของวิธีการสเปกตรัมเคอร์เนล
กับวิธีการทันสมัยในการจำแนกชนิดข้อมูลของโมเลกุลเอชแอลเอกลุ่มย่อย

การจำแนกชนิด ของข้อมูล	วิธีการและตัวแปรต่างๆ						สเปกตรัมเคอร์เนล $k=6$
	<i>NN</i>	<i>Linear</i>	<i>Poly</i>	<i>Poly</i>	<i>Poly</i>	<i>Gaussian</i>	
			$d=2$	$d=3$	$d=4$		
กลุ่มย่อยของ เอชแอล เอกลุ่มที่ I	65.9	98.1	98.6	98.4	98.1	99.6	99.4
กลุ่มย่อยของ เอชแอล เอกลุ่มที่ II	71.4	98.8	98.5	98.9	98.5	99.3	99.0