ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

วิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)

ปริญญา

วิศวกรรมคอมพิวเตอร์ วิศวกรรมคอมพิวเตอร์

สาขา **ภาควิชา**

เรื่อง การจำแนกซับแฟมมิลี่ของเอมไซม์โดยใช้กฎความสัมพันธ์ที่มีลำดับเหตุการณ์

Sequential Associative Classification for Enzyme Subfamily Prediction

นามผู้วิจัย นายบัลลังก์ นิยมศักดิ์

ได้พิจารณาเห็นชอบโดย

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(รองศาสตราจารย์กฤษณะ ไวยมัย, Ph.D.)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม

(ผู้ช่วยศาสตราจารย์พีรวัฒน์ วัฒนพงศ์, Ph.D.)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม

(รองศาสตราจารย์อรินทิพย์ ธรรมชัยพิเนต, Ph.D.)

หัวหน้าภาควิชา

(ผู้ช่วยศาสตราจารย์ภุชงค์ อุทโยภาศ, Ph.D.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

(รองศาสตราจารย์กัญจนา ธีระกุล, D.Agr.)

คณบดีบัณฑิตวิทยาลัย

วันที่ ..... **เดือน** ..... **พ.ศ.** .....

วิทยานิพนธ์

เรื่อง

การจำแนกซับแฟมมิลี่ของเอมไซม์โดยใช้กฎความสัมพันธ์ที่มีลำดับเหตุการณ์

Sequential Associative Classification for Enzyme Subfamily Prediction

โดย

นายบัลลังก์ นิยมศักดิ์

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์
เพื่อความสมบูรณ์แห่งปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)
พ.ศ. 2553

บัลลังก์ นิยมศักดิ์ 2553: การจำแนกซับแฟมมิลี่ของเอมไซม์โดยใช้กฎความสัมพันธ์ที่มีลำดับเหตุการณ์ ปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์) สาขาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก: รองศาสตราจารย์กฤษณะ ไวยมัย, Ph.D. 69 หน้า

เทคนิคจำแนกซับแฟมมิลี่ของเอมไซม์เป็นงานวิจัยด้านชีวสารสนเทศศาสตร์ที่มีอย่างแพร่หลายในปัจจุบัน วิธีส่วนใหญ่ใช้ความรู้พื้นฐานทางสถิติซึ่งให้ความถูกต้องได้ดีในระดับหนึ่ง อย่างไรก็ตาม อย่างไรก็ตามวิธีเหล่านี้ไม่ได้สนใจในเรื่องการสร้างโมเดลในการทำนายที่สามารถอธิบายได้และลำดับความสัมพันธ์ของแอททิบิวต์ ว่ามีผลอย่างไรต่อการสร้างโมเดลในการทำนาย

ในงานวิจัยนี้จึงได้เสนอเทคนิคการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์แบบมีลำดับในการจำแนกซับแฟมมิลี่ของเอมไซม์(SAC : Sequential Associative Classification for Enzyme Subfamily Prediction) โดยรวมเทคนิคการหากฎความสัมพันธ์แบบมีลำดับเหตุการณ์ (Sequential Pattern Mining)และเทคนิคการจัดกลุ่มข้อมูลโดยใช้กฎความสัมพันธ์ (Associative Classification) เข้าด้วยกันเพื่อการทำนาย โดยใช้ความรู้พื้นฐานเกี่ยวกับ “โมทีฟเชิงปฏิกิริยาชีวเคมี” มารวมเข้าด้วยกันกับขั้นตอนการหากฎความสัมพันธ์แบบมีลำดับเหตุการณ์ นอกจากนี้ในการจัดกลุ่มข้อมูลโดยใช้กฎความสัมพันธ์ได้ถูกขยายขอบเขตออกไปโดยการพิจารณาลำดับความสัมพันธ์ของแอททิบิวต์ในขั้นตอนการสร้างโมเดลในการทำนาย จากผลการทดลองนอกจากจะได้โมเดลจำแนกประเภทที่ง่ายต่อการตีความแล้วเทคนิค SAC นี้สามารถสรุปให้เห็นได้ถึงความสัมพันธ์ระหว่างลำดับแอททิบิวต์ ในสายโปรตีนมีผลต่อความแม่นยำในการทำนาย โดยให้ความแม่นยำที่ดีที่สุดอยู่ที่ 70.95% โดยใช้ชุดข้อมูลโปรตีนเอมไซม์จากฐานข้อมูล SWISS-PROT

_____ _____ ___ / ___ / ___

ลายมือชื่อนิสิต ลายมือชื่ออาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

Banlang Niyomsak 2010: Sequential Associative Classification for Enzyme Subfamily Prediction. Master of Engineering (Computer Engineering), Major Field: Computer Engineering, Department of Computer Engineering. Thesis Advisor: Associate Professor Kitsana Waiyamai, Ph.D. 69 pages.

Enzyme subfamily prediction is nowadays a very active bioinformatics research topic. Using statistics to build predictor, enzyme subfamily prediction methods are very accurate. However those methods generate non-explainable predictor models and order between features is not taken into account for prediction model construction.

In this research work, we propose a Sequential Associative Classification (SAC) for enzyme subfamily prediction. SAC integrates sequential pattern mining with associative classification to perform prediction task. Background knowledge about biding and active site motifs is integrated in the sequential pattern mining step. Associative classification is extended by taking into account the order between features during the prediction model construction step. Experimental results using SWISS-PROT sequence dataset show that SAC is able to generate a very easy to understand descriptive predictor in the forms of class sequential pattern rules, and a very high accuracy of 70.95%.

_____ _____ ___ / ___ / ___

Student's signature Thesis Advisor's signature

กิตติกรรมประกาศ

ข้าพเจ้าขอกราบขอบพระคุณรองศาสตราจารย์ ดร. กฤษณะ ไวยมัย ประธานกรรมการที่ปรึกษาที่ได้สอนกระบวนการคิด ให้คิดอย่างเป็นระบบ มีเหตุมีผล มีจุดประสงค์ที่แน่นอน รู้จักแยกแยะ ระหว่าง อารมณ์ และเหตุผลให้ข้าพเจ้านำไปใช้ในเรื่องชีวิตประจำวัน รวมไปถึงการทำวิทยานิพนธ์ฉบับนี้

ขอกราบขอบพระคุณ ผู้ช่วยศาสตราจารย์พีรวัฒน์ วัฒนพงศ์ และ รองศาสตราจารย์อรินทิพย์ ธรรมชัยพิเนต กรรมการที่ปรึกษาวิทยานิพนธ์ร่วม ที่กรุณาให้ความรู้ ข้อเสนอแนะ คำปรึกษา คำแนะนำ เปิดโลกทัศน์และมุมมองต่าง ๆ เกี่ยวกับงานวิจัย

ขอขอบคุณ คุณธนภัทร ฆังคะจิตร อาจารย์ธนาวินท์ รักธรรมานนท์ คุณสาวิณี แสงสุริยันต์ คุณพีรพล เวทีกูล ที่ให้คำปรึกษาที่ดีไม่ว่าจะเป็นเรื่องงานวิจัยและเรื่องอื่นๆมากมาย ขอขอบคุณเอกสิทธิ์ พัชรวงศ์ศักดา เพื่อนผู้คอยช่วยเหลือให้คำปรึกษาในงานวิจัยและแผนการเรียน รวมทั้งขอขอบคุณสมาชิกห้องปฏิบัติการ DAKDL ที่สละเวลามาประชุมแลกเปลี่ยนประสบการณ์งานวิจัย ความรู้ใหม่ๆและให้คำแนะนำต่างๆมากมาย สุดท้ายขอบพระคุณ ทุกท่านที่มีส่วนในงานวิจัยนี้ทั้งที่ได้กล่าวถึง และไม่ได้กล่าวถึงไว้ ณ ที่นี้ด้วย

ขอขอบคุณเจ้าหน้าที่โครงการบัณฑิตศึกษา และเจ้าหน้าที่ธุรการภาควิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเกษตรศาสตร์ที่ช่วยเหลือในการประสานงาน และดำเนินงานด้านเอกสารต่างๆ ให้เป็นไปอย่างสะดวกลุล่วงไปด้วยดี

คุณงามความดี หรือประโยชน์อันใดเนื่องมาจากวิทยานิพนธ์ฉบับนี้ ขออุทิศแด่มารดา บิดา บุพการี และผู้มีพระคุณทุกท่าน

บัลลังก์ นิยมศักดิ์

สิงหาคม 2553

สารบัญ

หน้า

สารบัญ	(1)
สารบัญตาราง	(2)
สารบัญภาพ	(3)
คำนำ	1
วัตถุประสงค์	2
การตรวจเอกสาร	3
อุปกรณ์และวิธีการ	32
อุปกรณ์	32
วิธีการ	32
ผลและวิจารณ์	46
ผล	46
วิจารณ์	51
สรุปและข้อเสนอแนะ	53
สรุป	53
ข้อเสนอแนะ	53
เอกสารและสิ่งอ้างอิง	55
ภาคผนวก	59
ประวัติการศึกษา และการทำงาน	69

สารบัญตาราง

ตารางที่		หน้า
1	แสดงชื่อและสัญลักษณ์อักษรย่อ 3 ตัว และ 1 ตัว ของกรดอะมิโนมาตรฐาน 2 ชนิดที่พบในธรรมชาติ	10
2	แสดงสายของโปรตีนที่ถูกแทนที่ด้วยโมทีฟและจำนวนครั้งที่พบ *sp* ในสาย	40
3	แสดงการเกิด Reoccur ของ *sp* ในสายโปรตีนเส้นที่ 2	40
4	แสดงการคำนวนค่า *LFSS* ของ *sp*	41

ตารางผนวกที่

1	แสดงข้อมูลเอนไซม์ (Assession number) ที่ใช้ในงานวิจัยนี้	61

สารบัญภาพ

ภาพที่		หน้า
1	ขั้นตอนของเทคนิคการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์	5
2	ตัวอย่าง Sequence Database	7
3	รูปแบบลำดับเหตุการณ์ที่น่าสนใจแยกตาม k-sequence พร้อมแสดงค่า Support	8
4	แสดงสายโปรตีนที่เกิดจากการเรียงตัวของกรดอะมิโนหลายชนิดต่อกัน	8
5	แสดงการเชื่อมต่อระหว่างกรดอะมิโน 2 ชนิด ด้วยพันธะเปปไทด์	9
6	โครงสร้างปฐมภูมิของโปรตีน Immunoglobulin-binding protein G	12
7	โครงสร้างทุติยภูมิ จากภาพคือ Alpha-helix (ซ้าย) และ Beta sheet หรือ Beta-pleated sheet (ขวา)	12
8	โครงสร้างตติยภูมิของโปรตีนแคไนน์ จีดีพี-แรน (CANINE GDP-RAN, PDB id: 1BYU)	13
9	โครงสร้างจตุรภูมิของโปรตีนฮีโมโกลบิน (Hemoglobin)	13
10	กลไกในบริเวณจับและบริเวณเร่งของเอนไซม์ซูเครส (sucrase)	14
11	ตาราง BLOSUM62	16
12	แสดงไดอะแกรมวิธีการสร้างโมเดลทำนายฟังก์ชั่นด้วยวิธีการ homology-based ที่มีการปรับเปลี่ยนวิธีการบางส่วน	20
13	CBS-RG	25
14	CBS-CB	26
15	CBS_ALL (RG) อัลกอริทึม	27
16	CBS_ALL (CB) อัลกอริทึม	28
17	CBS_CLASS (RG) อัลกอริทึม	29
18	CBS_CLASS (CB) อัลกอริทึม	30
19	ภาพรวมของ SAC	33
20	ตารางแสดงคุณสมบัติโมทีฟ	34
21	การคำนวนคะแนนของโมทีฟที่พบในสายโปรตีน	35
22	การคำนวนคะแนนของโมทีฟที่พบในสายโปรตีน	36

สารบัญภาพ (ต่อ)

ภาพที่		หน้า
23	แสดงการเลือกโมทีฟแบบที่ดีที่สุด	37
24	รหัสเทียมของอัลกอริทึม SAC	38
25	กฎที่ได้จากการหา Class Sequential Pattern Mining	42
26	รหัสเทียมของ SAC Classifier	44
27	แสดงความแม่นยำ (แยกตามคลาส) แบบ Reoccurrence และ No Reoccurrence	48
28	แสดงความแม่นยำ (แยกตามคลาส) ของระบบจำแนกเอนไซม์ซับแฟมิลี สำหรับข้อมูลชุดที่เมื่อเปรียบเทียบกับ CBS	49
29	แสดงความแม่นยำ (แยกตามคลาส) ของระบบจำแนกเอนไซม์ซับแฟมิลี สำหรับข้อมูลชุดที่ 2 (unknown enzymes) จำนวน 2124 เอนไซม์ (independent test)	50
30	แสดงปัญหาของ CBS ที่เกิดจากหารใช้คะแนนจากความยาวกฎ	51
31	แสดงการเลือกโมทีฟแบบกลุ่ม	54

การจำแนกซับแฟมมิลี่ของเอมไซม์โดยใช้กฎความสัมพันธ์ที่มีลำดับเหตุการณ์

Sequential Associative Classification for Enzyme Subfamily Prediction

คำนำ

การทำนายเอนไซม์ซับแฟมมิลี่นั้นมีความสำคัญในการช่วยให้เราเข้าใจกลไกในการทำงานของเอมไซม์ในสิ่งมีชีวิต ซึ่งมีความสำคัญต่อการพัฒนาและวิจัยยารักษาโรคเป็นต้น โดยเอนไซม์ทำหน้าที่ในการเร่งปฏิกิริยาเคมีภายในเซลล์สิ่งมีชีวิต หากขาดเอนไซม์อาจทำให้ผลิตภัณฑ์จากปฏิกิริยาเคมีกลายเป็นสารเคมีชนิดอื่น ส่งผลให้การทำงานของเซลล์ผิดปรกติได้ ปัจจุบันมีการค้นพบเอนไซม์ชนิดใหม่เพิ่มขึ้นอย่างรวดเร็วและเป็นปริมาณมาก การจำแนกประเภทเอนไซม์โดยวิธีการวิเคราะห์และทดสอบในห้องปฏิบัติการอาจใช้เวลานานและเสียค่าใช้จ่ายสูง ดังนั้นจึงมีการพัฒนาระบบอัตโนมัติเพื่อทำการจำแนกเอนไซม์ซับแฟมมิลี่ จากอดีตจนถึงปัจจุบันมีงานวิจัยมากมายเกี่ยวกับการจำแนกเอนไซม์ซับแฟมมิลี่ แต่ยังไม่มีงานวิจัยใดที่ให้ความสนใจในการพัฒนาระบบในการจำแนกข้อมูลเอนไซม์ซับแฟมิลีที่สามารถทำความเข้าใจการทำงานเอนไซม์โดยสามารถบอกได้ว่าส่วนใดของโปรตีนที่ก่อให้เกิดกลไกการทำงานของเอมไซม์แบบใด

ในงานวิจัยนี้มุ่งเน้นพัฒนาระบบจำแนกประเภทเอนไซม์ซับแฟมิลีที่ให้ความถูกต้องสูง โดยเน้นไปในเรื่องการทำความเข้าใจกลไกการทำงานของเอมไซม์และลักษณะการเกิดโมทีฟซ้ำๆ ว่ามีความสัมพันธ์อย่างไรกับการทำงานของเอมไซม์ โดยโมทีฟที่ใช้ในงานวิจัยนี้จะเป็นส่วนย่อยๆ ของสายโปรตีนที่มีความสัมพันธ์โดยตรงกับการทำงานของเอมไซม์ โดยเสนอเทคนิคที่ชื่อว่า Sequential Associative Classification for Enzyme Subfamily Prediction (SAC) ที่ได้จากการนำเทคนิคการหารูปแบบลำดับเหตุการณ์ที่น่าสนใจ (Sequential Pattern Mining) และการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ (Associative Classification) มาผสมผสานกันนอกจากนี้ยังใช้ความรู้พื้นฐานของตาราง BLOSUM62 ในการคัดสรรโมทีฟที่ดีที่สุดสำหรับสายโปรตีนนั้น

วัตถุประสงค์

1. พัฒนาระบบในการจำแนกข้อมูลเอนไซม์ซับแฟมิลีโดยการรวมเทคนิค Sequential Pattern Mining และ Associative Classification เข้าด้วยกัน

2. ศึกษาความสัมพันธ์ระหว่างลำดับของฟีเจอร์ในสายโปรตีนและโครงสร้างปฐมภูมิที่พบจากตาราง BLOSUM62 Matrix

3. ศึกษาความสัมพันธ์ของรูปแบบฟีเจอร์ที่เกิดขึ้นซ้ำๆกันในสายโปรตีนว่ามีผลอย่างไรต่อการจำแนกข้อมูลเอนไซม์ซับแฟมิลี

4. พัฒนาระบบในการจำแนกข้อมูลเอนไซม์ซับแฟมิลีโดยใช้ค่าระยะทางระหว่างฟีเจอร์ในการบอกความใกล้เคียงของกฎและ Unseen Data

5. พัฒนาระบบในการจำแนกข้อมูลเอนไซม์ซับแฟมิลีได้อย่างมีประสิทธิภาพเหมาะสมและให้ความถูกต้องแม่นยำสูง โดยใช้ข้อมูลที่มีอยู่จำกัดเพียง 3.34% ของข้อมูลฝึกสอนระบบมาใช้ในการสร้างฟีเจอร์ตัวแทนสายโปรตีน

6. พัฒนาระบบในการจำแนกข้อมูลเอนไซม์ซับแฟมิลีที่สามารถทำความเข้าใจการทำงานเอนไซม์ได้ดีขึ้น โดยสามารถบอกได้ว่าส่วนใดของโปรตีนที่ก่อให้เกิดกลไกการทำงานของ Enzyme Function

การตรวจเอกสาร

การตรวจเอกสารของวิทยานิพนธ์นี้ประกอบไปด้วย ความรู้พื้นฐานทั่วไปเกี่ยวกับการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ (Associative Classification) การหารูปแบบลำดับเหตุการณ์ที่น่าสนใจ (Sequential Pattern Mining) ความรู้พื้นฐานด้านชีววิทยา แนวทางในการจำแนกเอนไซม์ซับแฟมิลี่ และ งานวิจัยที่เกี่ยวข้อง

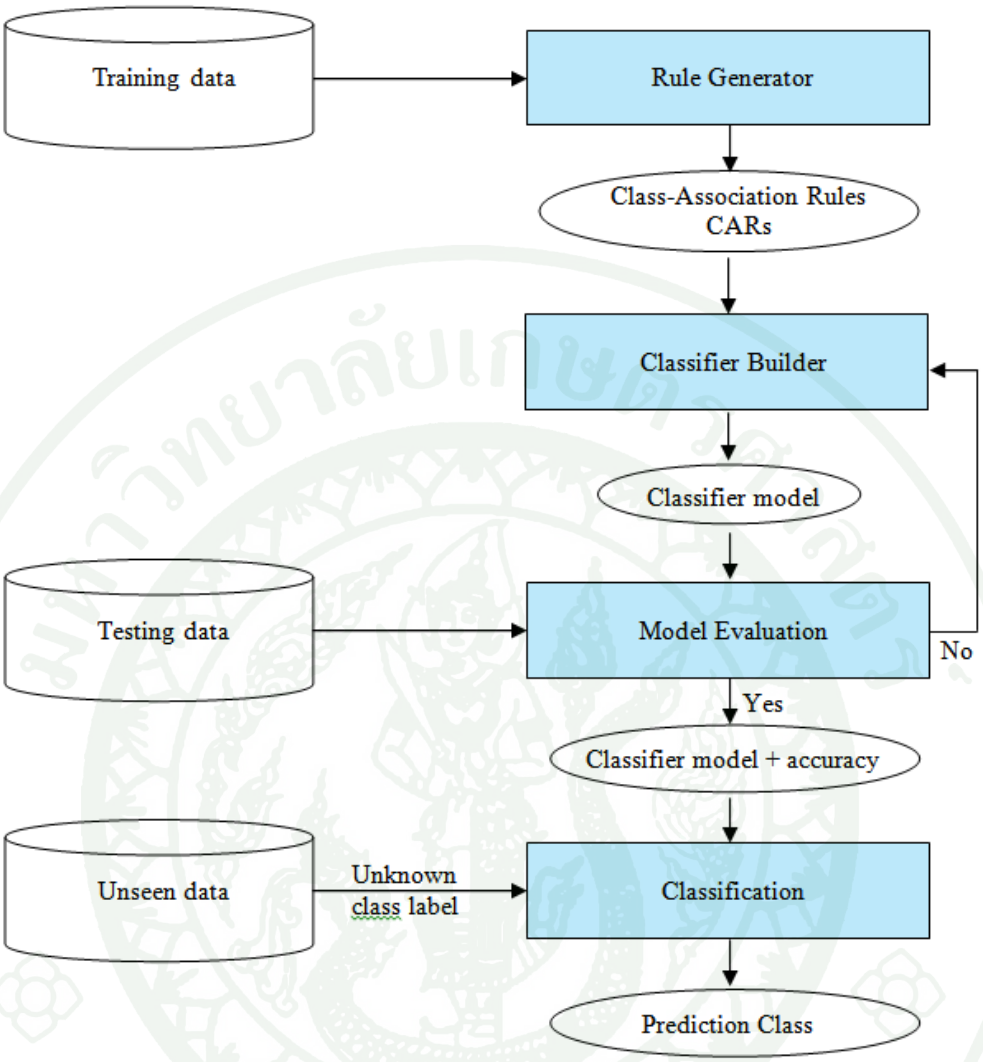
ความรู้พื้นฐานของเทคนิค Data mining

1. การจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ (Associative Classification)

การจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ (Associative classification) เป็นเทคนิคที่เกิดจากการรวมกันระหว่าง 2 เทคนิค (Liu *et al.,* 1998) ที่ได้กล่าวในหัวข้อข้างต้น นั่นคือ การจำแนกประเภทข้อมูล (Data classification) และ การสืบค้นกฎความสัมพันธ์ (Association rule discovery) โดยที่จุดประสงค์ของเทคนิคการจำแนกประเภทข้อมูลคือ เพื่อค้นหาโมเดลหรือเซตที่เล็กที่สุดของกฎในฐานข้อมูลเพื่อสร้างโมเดลจำแนกประเภทข้อมูลที่มีความถูกต้องแม่นยำมากที่สุด และจุดประสงค์ของเทคนิคการสืบค้นกฎความสัมพันธ์คือ เพื่อค้นหากฎความสัมพันธ์ทั้งหมดที่มีความสำคัญและบ่งบอกถึงคุณลักษณะของฐานข้อมูล โดยที่กฎเหล่านั้นจะต้องผ่านค่าสนับสนุนและค่าความมั่นใจขั้นต่ำด้วย โดยเทคนิคการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ (Associative classification) นี้ ได้แบ่งออกเป็น 2 ส่วนหลักๆ คือ ส่วนที่ใช้ในการสร้างกฎความสัมพันธ์ (Rule generator phase) และส่วนที่นำกฎความสัมพันธ์ไปสร้างโมเดลเพื่อใช้ทำนายข้อมูล (Classifier builder phase)

โดยในส่วนของการสร้างกฎความสัมพันธ์ (Rule generator phase) นั้นจะใช้หลักการหรือวิธีการเดียวกันกับเทคนิค Association rule discovery เกือบทั้งหมด ยกเว้นกฎที่ถูกสร้างจากกระบวนการสร้างกฎความสัมพันธ์นั้นจะต้องเป็นกฎเฉพาะที่เรียกว่า Class-Association Rules (CARs) นั่นคือกฎความสัมพันธ์ที่สับเซตของกฎทางด้านขวามือจะต้องเป็นแอตทริบิวต์ Class เท่านั้น เช่น {A, B, C → Class} โดยอัลกอริทึมการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ที่มีอยู่ (Dong *et al.,* 1999; Liu *et al.,* 1998) จะถูกดัดแปลงเพื่อค้นหา CARs ทั้งหมดที่ผ่านค่าสนับสนุนขั้นต่ำ (Minimum support) และค่าความมั่นใจขั้นต่ำ (Minimum confidence)

ในส่วนของการสร้างโมเดลในการทำนาย (Classifier builder phase) จะนำกฎความสัมพันธ์ที่ได้จากส่วนการสร้างกฎมาใช้เพื่อสร้างโมเดลในการทำนายข้อมูล โดยในการทำนายข้อมูลนั้น จะมีการพิจารณานแบ่งออกเป็น 2 วิธี วิธีที่1 จะทำการพิจารณากฎความสัมพันธ์ทีละกฎ (Single rule) โดยวิธีการพิจารณาแบบนี้ จะต้องทำการเรียงลำดับกฎความสัมพันธ์ก่อน โดยทั่วไปแล้วจะเรียงลำดับกฎความสัมพันธ์ตามค่าความมั่นใจ (Confidence) ก่อน แต่ถ้าค่าความมั่นใจของกฎความสัมพันธ์เท่ากัน ก็จะเรียงลำดับของกฎความสัมพันธ์ตามค่าสนับสนุน (Support) แต่ถ้าทั้งค่าความมั่นใจ และ ค่าสนับสนุนของกฎเกิดเท่ากันอีก ก็จะเรียงลำดับกฎโดยดูจาก กฎไหนถูกสร้างมาก่อน ก็จะเรียงกฎนั้นก่อนตามลำดับ หลังจากเรียงลำดับกฎความสัมพันธ์เป็นที่เรียบร้อยแล้ว ก็พร้อมที่จะทำนายข้อมูล โดยการทำนายข้อมูลนั้นจะทำนายตาม class ของกฎที่มีศักย์ (Precedence) สูงที่สุด ส่วนวิธีที่2 จะทำการพิจารณากฎความสัมพันธ์ทีละหลายๆกฎพร้อมกัน (Multiple rules) โดยในการทำนายข้อมูลนั้นจะนำกลุ่มของกฎที่มีอยู่ในคลาสเดียวกัน มาคำนวณผ่านสูตรที่ได้กำหนดเอาไว้แล้วดูว่าคลาสไหนที่ให้ค่ามากที่สุดคลาสนั้นก็จะเป็นคำตอบ โดยที่ภาพรวมขั้นตอนทั้งหมดของเทคนิคการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์ สามารถดูได้จากภาพที่ 1



ภาพที่ 1 ขั้นตอนของเทคนิคการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์

ที่มา: วีระพล, 2549

2. การหารูปแบบลำดับเหตุการณ์ที่น่าสนใจ(Sequential Pattern Mining)

เทคนิคการสืบค้นรูปแบบลำดับเหตุการณ์ คือการหารูปแบบของข้อมูลที่เกิดขึ้นบ่อย และมีเวลาหรือลำดับการเกิดเข้ามาเกี่ยวข้อง การหารูปแบบลำดับเหตุการณ์สามารถอธิบายในรูปของโมเดลทางคณิตศาสตร์ได้ดังนี้

ให้ M เป็นเซ็ตของโมทีฟทั้งหมดที่ปรากฏอยู่ในฐานข้อมูลโปรตีนคือ $M = \{m_1, m_2, ...m_l\}$ และสายโปรตีน P สามารถเขียนแทนด้วยลำดับของเหตุการณ์ E ที่ต่อเนื่องกันโดย $P =< E_1 E_2 ... E_k >$ เมื่อเหตุการณ์ E_i แทนเซตของโมทีฟทั้งหมดที่เกิดขึ้นซ้อนทับกันบนสายโปรตีน P โดยนับเป็นชุดที่ i เมื่อนับจากต้นสายโปรตีน P และเขียนแทนด้วยสัญลักษณ์ $E = (m_{k1}, m_{k2}, ..., m_{kn})$ เมื่อ $0 \le k1 < k1 < ... < kn$ ดังนั้นสำหรับสายโปรตีนที่มีเพียงโมทีฟ m_x และโมทีฟ m_y โดยที่ m_x เกิดขึ้นก่อนโมทีฟ m_y และไม่มีส่วนใดซ้อนทับกันจะสามารถเขียนสายโปรตีนนี้ได้เป็น $P =< (m_x)(m_y) >$

k-sequence หมายถึงรูปแบบลำดับเหตุการณ์ที่มีขนาด k โดย $k = \sum_i |E_i|$ เช่น $< (m_1)(m_4, m_2) >$ เป็น 3-sequence

สายโปรตีน $A =< a_1 a_2 ... a_n >$ ประกอบด้วยเหตุการณ์ $a_1, a_2, ..., a_n$ ตามลำดับดับ จะเรียกสายโปรตีนนี้ว่าเป็นรูปแบบลำดับเหตุการณ์ A และจะเป็นรูปแบบลำดับเหตุการณ์ย่อย (subsequence) ของรูปแบบลำดับเหตุการณ์ $B =< b_1 b_2 ... b_n >$ ก็ต่อเมื่อมีจำนวนเต็ม $j_1 < j_2 < ... < j_n$ ที่ทำให้ $a_1 \subseteq b_{j1}, a_2 \subseteq b_{j2}, ..., a_n \subseteq b_{jn}$ เช่น $< (m_1)(m_6, m_8)(m_3) >$ เป็นรูปแบบลำดับเหตุการณ์ย่อยของ $< (m_7)(m_1, m_3)(m_{12})(m_6, m_8, m_5)(m_3) >$ เพราะ $(m_1) \subseteq (m_1, m_3), (m_6, m_8) \subseteq (m_6, m_8, m_5)$ และ $(m_3) \subseteq (m_3)$ ส่วน $< (m_1)(m_8) >$ ไม่เป็นรูปแบบลำดับเหตุการณ์ย่อยของ $< (m_1, m_8) >$ และในทางกลับกัน

รูปแบบลำดับเหตุการณ์ย่อยที่น่าสนใจ (Frequent SubSequences) คือ รูปแบบลำดับเหตุการณ์ที่มีค่า Support มากกว่าหรือเท่ากับค่า Minimum Support ที่กำหนดไว้ เช่น จากฐานข้อมูลโปรตีนเอมไซม์ในภาพที่ 2 กำหนดให้มีค่า Minimum Support เท่ากับ 3 รูปแบบลำดับเหตุการณ์ที่น่าสนใจจะเป็นดังภาพที่ 3

รูปแบบลำดับเหตุการณ์ที่น่าสนใจและมีขนาดใหญ่ที่สุด (Maximal Frequent Sequences) คือ รูปแบบลำดับเหตุการณ์ที่น่าสนใจที่ไม่เป็นรูปแบบลำดับเหตุการณ์ย่อยของรูปแบบลำดับเหตุการณ์ที่น่าสนใจอื่นๆแสดงด้วยอักษรตัวหนาในภาพที่ 3

Protein-Sequence	Motif Sequence
1	<(M_3,M_4) (M_1,M_2,M_3) (M_1,M_2,M_5)>
2	<(M_1,M_2,M_3) (M_5) (M_1,M_3) (M_7) (M_6)>
3	<(M_1,M_2,M_5) (M_5) (M_8) (M_7) (M_6)>
4	<(M_2,M_3,M_4) (M_2,M_5) (M_7) (M_6)>
5	<(M_1,M_2,M_5)>
6	<(M_1,M_3) (M_5)>

ภาพที่ 2 ตัวอย่าง Sequence Database

อัลกอริทึมนี้เป็นอัลกอริทึมพื้นฐานชื่อ AprioriAll(Agrawal and Srikant, 1995) ที่พัฒนาจากอัลกอริทึม Apriori มีขั้นตอนในการทำงานคือขั้นแรกจะหาชุดข้อมูลที่เกิดพร้อมกันหรือเป็นรูปแบบลำดับ (sequence) ที่มี 1 เหตุการณ์ โดยมีการทำงานเหมือนการหา Frequent Itemsets ของอัลกอริทึม Apriori แต่นับค่า Support ตามจำนวน Sequence-id ที่เกิดชุดข้อมูลนั้น จากนั้นชุดข้อมูลเหล่านั้นซึ่งผ่านค่า Support จะถูกเรียกว่า Frequent Itemsets (L_1)มาสร้างเป็นกลุ่มของรูป แบบลำดับ 2-sequence (C_2) ที่เป็นไปได้ แล้วตรวจสอบค่า Support จากนั้นนำ 2-sequence ที่ผ่านค่า Support (L_2) มาสร้างเป็นกลุ่มของรูปแบบลำดับ 3-sequence (C_3) ที่เป็นไปได้ โดยกระบวนการนี้จะทำซ้ำไปเรื่อยๆจนกระทั่งไม่พบรูปแบบลำดับที่ผ่านค่า Support อีก

ลำดับเหตุการณ์ที่ใหญ่ที่สุด (Maximal Sequence) ลำดับเหตุการณ์ใด ๆ จะถูกเรียกว่าเป็นลำดับเหตุการณ์ที่ใหญ่ที่สุดก็ต่อเมื่อลำดับเหตุการณ์นั้นไม่เป็นลำดับเหตุการณ์ย่อยของลำดับเหตุการณ์ใด ๆเลยจากภาพที่ 3 ลำดับเหตุการณ์ที่ใหญ่ที่สุดจะถูกแสดงตัวตัวหนา

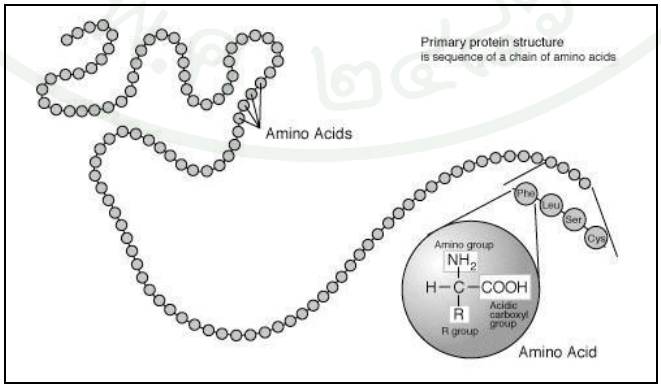
1-freq	2-freq	3-freq	4-freq
$<(M_1)>$: 5	$<(M_1,M_2)>$: 4	$<(M_1,M_2,M_5)>$: 3	**$<(M_2)\ (M_5)\ (M_7)\ (M_6)>$: 3**
$<(M_2)>$: 5	$<(M_1,M_3)>$: 3	**$<(M_1,M_2)\ (M_5)>$: 3**	
$<(M_3)>$: 4	<(M1,M5)> : 3	**$<(M_1,M_3)\ (M_5)>$: 3**	
$<(M_5)>$: 6	$<(M_2,M_3)>$: 3	**$<(M_2,M_3)\ (M_5)>$: 3**	
$<(M_6)>$: 3	$<(M_2,M_5)>$: 4	$<(M_2)\ (M_7)\ (M_6)>$: 3	
$<(M_7)>$: 3	$<(M_1\ M_5)>$: 4	$<(M_2)\ (M_5)\ (M_7)>$: 3	
	$<(M_2)\ (M_5)>$: 4	$<(M_5)\ (M_7)\ (M_6)>$: 3	
	$<(M_2)\ (M_6)>$: 3		
	$<(M_2)\ (M_7)>$: 3		
	$<(M_3)\ (M_5)>$: 3		
	$<(M_5)\ (M_6)>$: 3		
	$<(M_5)\ (M_7)>$: 3		
	$<(M_7)\ (M_6)>$: 3		

ภาพที่ 3 รูปแบบลำดับเหตุการณ์ที่น่าสนใจแยกตาม k-sequence พร้อมแสดงค่า Support

ความรู้พื้นฐานเกี่ยวกับชีววิทยา

ความรู้เบื้องต้นเกี่ยวกับโปรตีนและเอนไซม์

โปรตีน คือ สารประกอบอินทรีย์ที่เกิดจากการเรียงตัวของกรดอะมิโนซึ่งเชื่อมต่อกันด้วยพันธะเปปไทด์ เรียกว่าสายโพลีเปปไทด์ (polypeptide) ดังภาพที่ 4 โปรตีนถูกพบครั้งแรกโดย Jöns Jakob Berzelius ในปี ค.ศ. 1838 โปรตีนจัดเป็นสารโมเลกุลขนาดใหญ่ เป็นส่วนประกอบสำคัญของโครงสร้างและกิจกรรมภายในเซลล์สิ่งมีชีวิตทุกชนิด

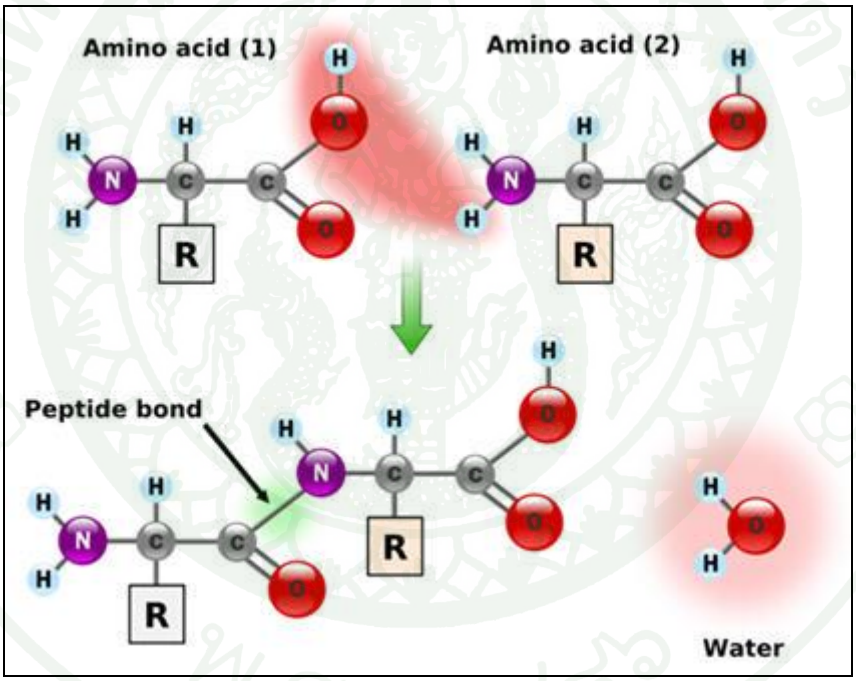


ภาพที่ 4 แสดงสายโปรตีนที่เกิดจากการเรียงตัวของกรดอะมิโนหลายชนิดต่อกัน

ภาพที่ 4 (ต่อ)

ที่มา: Wikipedia (2009)

ส่วนประกอบย่อยของสายโพลีเปปไทด์ คือกรดอะมิโน (amino acid) มีสูตรโครงสร้างเป็น NH2-CHR-COOH ดังภาพที่ 5 ในสูตรโครงสร้างนี้อะตอมคาร์บอนในตำแหน่งอัลฟา (α -Carbon) จะเป็นอะตอมคาร์บอนไม่สมมาตร (asymmetric carbon atom) ดังนั้นกรดอะมิโนทุกตัว (นอกจากไกลซีน ซึ่งมี R เป็นไฮโดรเจน) จะมีสเตริโอไอโซเมอร์ได้ สองชนิด คือ D- และ L- โดยที่กรดอะมิโนที่พบในธรรมชาติส่วนใหญ่เป็นชนิด L-



ภาพที่ 5 แสดงการเชื่อมต่อระหว่างกรดอะมิโน 2 ชนิด ด้วยพันธะเปปไทด์

ที่มา: Wikipedia (2009)

กรดอะมิโนในธรรมชาติมีอยู่มากกว่า 80 ชนิด แต่กรดอะมิโนที่พบมากและถือว่ามีความสำคัญนั้นมีจำนวน 20 ชนิด แต่ละชนิดมีชื่อเต็มและชื่อย่อ โดยชื่อย่อแบ่งออกเป็น 2 รูปแบบ คือ รูปแบบสามตัวอักษร และรูปแบบหนึ่งตัวอักษร แสดงดังตารางที่ 1

ตารางที่ 1 แสดงชื่อและสัญลักษณ์อักษรย่อ 3 ตัว และ 1 ตัว ของกรดอะมิโนมาตรฐาน 2 ชนิดที่พบในธรรมชาติ

ชื่อเต็ม	ชื่อย่อ (3 ตัวอักษร)	ชื่อย่อ (1 ตัวอักษร)
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

ที่มา: Wikipedia (2009)

ข้อมูลโครงสร้างโปรตีนที่นักชีวเคมีให้ความสนใจมีอยู่ด้วยกัน 4 รูปแบบได้แก่ โครงสร้างปฐมภูมิ (primary structure) โครงสร้างทุติยภูมิ (secondary structure) โครงสร้างตติยภูมิ (tertiary structure) และโครงสร้างจตุรภูมิ (quaternary structure) แสดงดังภาพที่ 6, 7, 8 และ 9 ตามลำดับ สำหรับงานวิจัยนี้สนใจเฉพาะข้อมูลโครงสร้างปฐมภูมิหรือสายลำดับกรดอะมิโน (Protein Sequence) เท่านั้น

โครงสร้างปฐมภูมิหรือสายลำดับกรดอะมิโน แสดงลำดับการเรียงตัวของกรดอะมิโนในสายโปรตีน (เป็นข้อมูลที่พบมากที่สุดในฐานข้อมูลโปรตีน)

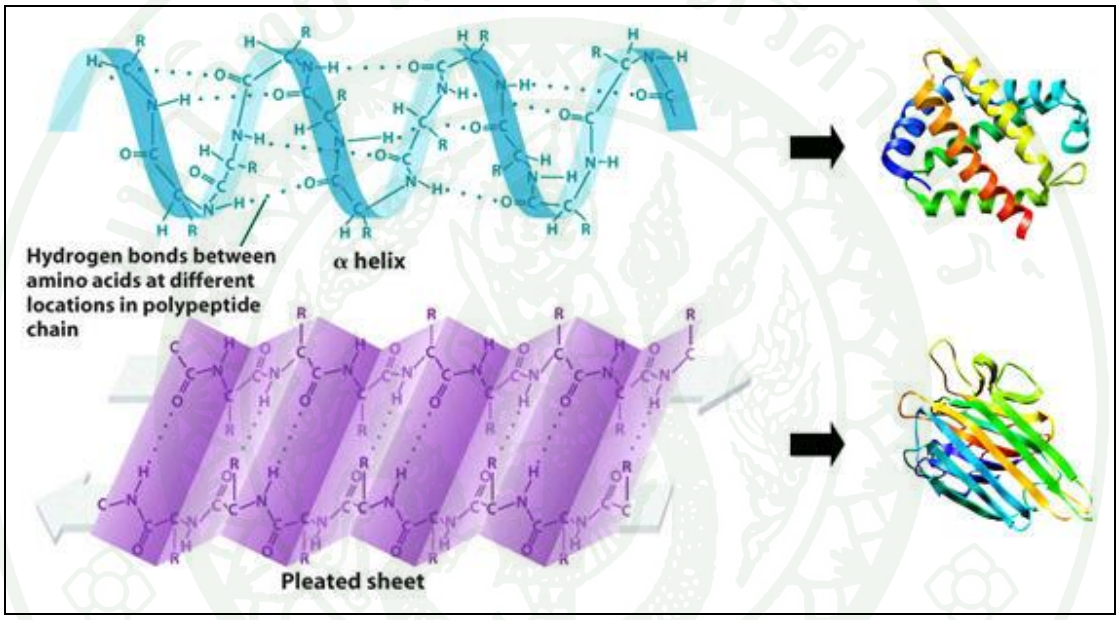
โครงสร้างทุติยภูมิหรือโครงสร้าง 2 มิติ แสดงการพันหรือม้วนตัว (fold) ของกรดอะมิโนที่อยู่ใกล้เคียงกันเพียงบางส่วนของสายโพลีเปปไทด์ เป็นรูปแบบเฉพาะอยู่ 2 ลักษณะคือ อัลฟาเฮลิกซ์ (alpha helix) สายเปปไทด์มีลักษณะขดเป็นเกลียว และเบต้า ชีท (beta sheet) หรือ เบต้า-พลีท ชีท (beta-pleated sheet) สายเปปไทด์มีลักษณะซิกแซกหรือพับเป็นแผ่น

โครงสร้างตติยภูมิหรือโครงสร้าง 3 มิติ แสดงการพันหรือม้วนตัว (fold) ของกรดอะมิโนตลอดทั้งสายโพลีเปปไทด์ โดยแสดงตำแหน่งพิกัดของทุกอะตอมของกรดอะมิโนในสายโพลีเปปไทด์ ข้อมูลดังกล่าวได้จากกระบวนการทางฟิสิกส์คือ X-Ray Crystallographic และ Nuclear Magnetic Resonance (NMR) ข้อมูลรูปร่างโครงสร้าง 3 มิตินี้มีประโยชน์อย่างมาก เนื่องจากสามารถบ่งบอกหน้าที่พื้นฐานของโปรตีนได้ ฐานข้อมูลที่สำคัญที่เก็บข้อมูลโครงสร้าง 3 มิติของโปรตีนคือ ฐานข้อมูลพีดีบี (Protein Data Bank, PDB) สามารถเข้าใช้งานได้ที่เว็บไซด์ www.rcsb.org

โครงสร้างจตุรภูมิ แสดงภาพรวมของโครงสร้าง 3 มิติ ของโปรตีน เนื่องจากโปรตีนบางชนิดประกอบด้วยสายโพลีเปปไทด์มากกว่า 1 สายขึ้นไป เช่น ฮีโมโกลบิน (hemoglobin) ของคนประกอบด้วยสายโพลีเปปไทด์ 4 สาย เราเรียกสายโพลีเปปไทด์แต่ละสายนี้ว่า “หน่วยย่อย” (subunit) (www.vcharkarn.com, 2009)

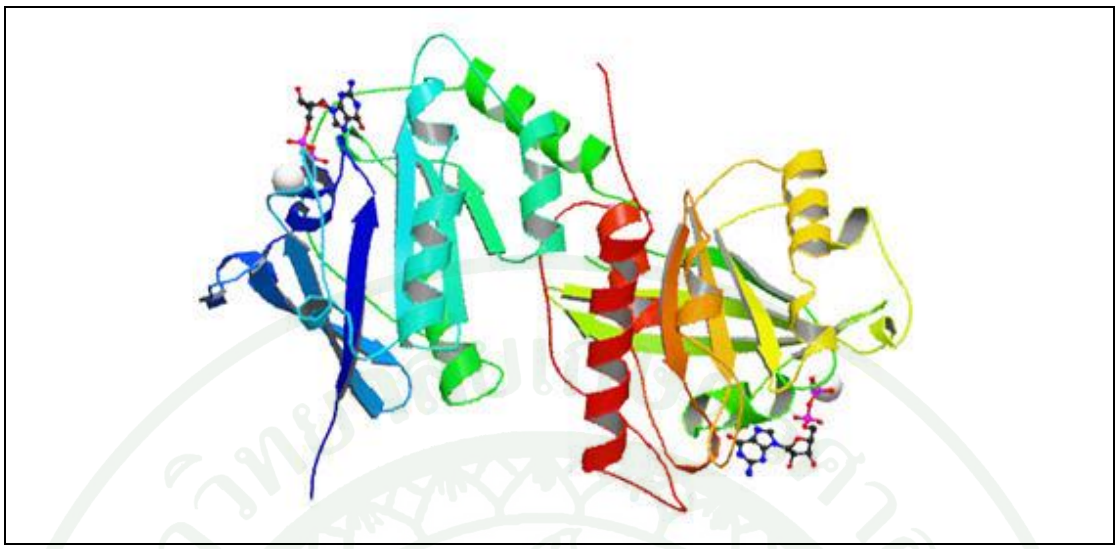
SFTLTNKNVIFVAGLGGIGLDTSKELLKRDLKNLVILDRIENPAAIAELKAINPKVTVTF
YPYDVTVPIAETTKLLKTIFAQLKTVDVLINGAGILDDHQIERTIAVNYTGLVNTTTAIL
DFWDKRKGGPGGIICNIGSVTGFNAIYQVPVYSGTKAAVVNFTSSLAKLAPITGVTAYTV
NPGITRTTLVHKFNSWLDVEPQVAEKLLAHPTQPSLACAENFVKAIELNQNGAIWKLDLG
TLEAIQWTKHWDSGI

ภาพที่ 6 โครงสร้างปฐมภูมิของโปรตีน Immunoglobulin-binding protein G



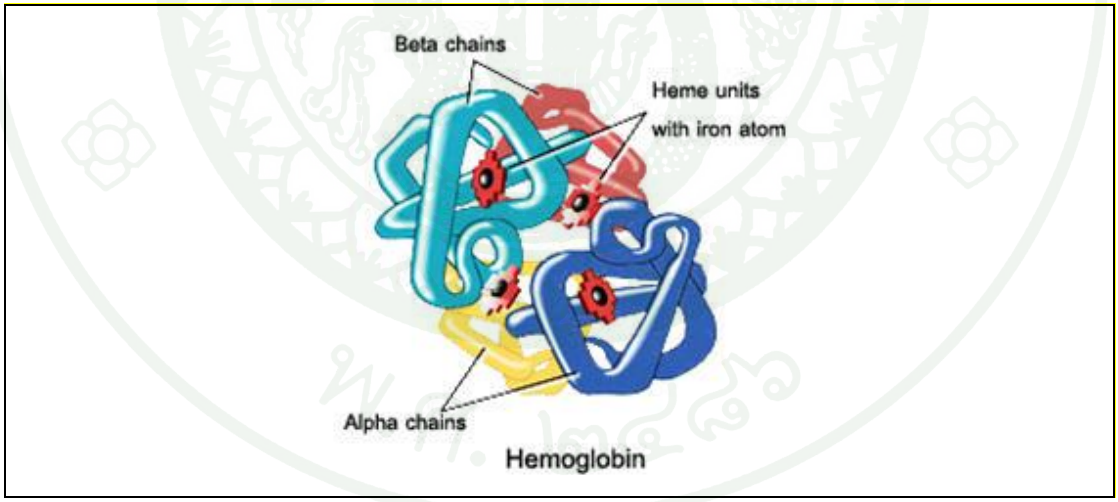
ภาพที่ 7 โครงสร้างทุติยภูมิ จากภาพคือ Alpha-helix (ซ้าย) และ Beta sheet หรือ Beta-pleated sheet (ขวา)

ที่มา: barleyworld.org (2009) และ Protein data bank (2009)



ภาพที่ 8 โครงสร้างตติยภูมิของโปรตีนแคไนน์ จีดีพี-แรน (CANINE GDP-RAN, PDB id: 1BYU)

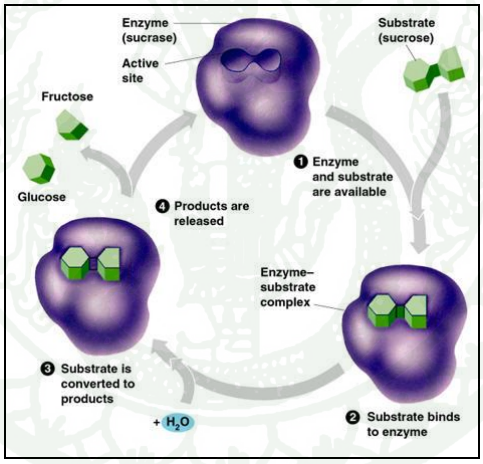
ที่มา: Protein data bank (2009)



ภาพที่ 9 โครงสร้างจตุรภูมิของโปรตีนฮีโมโกลบิน (Hemoglobin)

ที่มา: biomed.brown.edu (2009)

ฟังก์ชั่นเอนไซม์เป็นฟังก์ชั่นโปรตีนประเภทหนึ่งที่สำคัญอย่างยิ่งในสิ่งมีชีวิตทุกชนิด ลักษณะพิเศษของฟังก์ชั่นเอนไซม์คือความสามารถในการเร่งปฏิกิริยาที่สิ่งมีชีวิตต้องการได้อย่างรวดเร็ว เช่น การให้พลังงาน การย่อยอาหาร ฯลฯ การทำงานฟังก์ชั่นเอนไซม์ใช้บางส่วนของเอนไซม์ที่เรียกว่าบริเวณจับ (binding site) และบริเวณเร่ง (catalytic site) โดยบริเวณจับคือบริเวณที่เอนไซม์ใช้เข้าจับกับสารตั้งต้น (substrates) ให้อยู่ในสภาพที่พร้อมเกิดฟังก์ชั่น ส่วนบริเวณเร่งคือบริเวณที่เอนไซม์ใช้เหนี่ยวนำให้เกิดปฏิกิริยาเคมีกับสารตั้งต้นให้กลายเป็นผลิตภัณฑ์ (products) ดังนั้นโปรตีนที่มีโครงสร้างหรือลำดับกรดอะมิโนที่คล้ายคลึงกัน แต่บริเวณจับและบริเวณเร่งทำงานต่างกัน จึงไม่จำเป็นที่จะต้องมีฟังก์ชั่นเอนไซม์เหมือนกัน ในขณะที่โปรตีนที่มีโครงสร้างหรือลำดับกรดอะมิโนต่างกันมาก แต่มีลำดับกรดอะมิโนที่ทำหน้าที่บริเวณจับและบริเวณเร่งเหมือนกัน ย่อมสามารถมีฟังก์ชั่นเอนไซม์ที่เหมือนกันได้ (พีระ ลิ่วลม, 2551) รายละเอียดการทำงานของเอนไซม์แสดงดังภาพที่ 10



ภาพที่ 10 กลไกในบริเวณจับและบริเวณเร่งของเอนไซม์ซูเครส (sucrase)

ที่มา: porpax.bio.miami.edu (2009)

จากภาพที่ 8 แสดงการเข้าจับและเร่งปฏิกิริยาระหว่างเอนไซม์ซูเครส (sucrase) และสารตั้งต้นซูโครส (sucrose) จากภาพ หมายเลข 1 แสดงบริเวณเร่ง (active site) ของเอนไซม์ซูเครส และสารตั้งต้นซูโครส หมายเลข 2 แสดงการจับกัน (binding) ระหว่างเอนไซม์ซูเครสและสารตั้งต้นซูโครส ณ ตำแหน่งบริเวณเร่ง หมายเลข 3 แสดงการทำปฏิกิริยาของสารตั้งต้นซูโครสและน้ำ โดยมี

ตัวเร่งปฏิกิริยาคือเอนไซม์ซูเครส หมายเลข 4 แสดงผลลัพธ์จากปฏิกิริยาเคมีดังกล่าว ได้เป็นสารกลูโคส (glucose) และสารฟรัคโทส (fructose)

ดังนั้นงานด้านชีวสารสนเทศที่มุ่งเน้นศึกษาด้านเอนไซม์จึงไม่ได้มีเพียงความต้องการทำนายฟังก์ชั่นเอนไซม์ที่ถูกต้องแม่นยำเท่านั้น แต่หมายรวมไปถึงการใช้ข้อมูลจำนวนมากที่มีอยู่ในการศึกษาทำความเข้าใจการทำงานเอนไซม์ได้ดีขึ้น นั่นคือความสามารถในการระบุว่าส่วนใดของโปรตีนที่สามารถทำงานเป็นบริเวณจับหรือบริเวณเร่งที่ก่อให้เกิดกลไกการทำงานของฟังก์ชั่นเอนไซม์ได้อย่างสมเหตุผลน่าเชื่อถือในเชิงวิทยาศาสตร์ ซึ่งมีความสำคัญอย่างยิ่งต่อการนำไปใช้งานในระดับห้องปฏิบัติการ (หรือเรียกย่อว่า wet lab.) ในการทำความเข้าใจกลไกเอนไซม์และการออกแบบเอนไซม์ตัวใหม่

จากความสำคัญของเอนไซม์ ความซับซ้อนของฟังก์ชั่นเอนไซม์ และความต้องการในระดับห้องปฏิบัติการดังกล่าว ในงานวิจัยนี้จึงเน้นกลุ่มเป้าหมายคือเอนไซม์ และการพัฒนาตัวแทนสายโปรตีนที่ใช้ทำนายฟังก์ชั่นเอนไซม์ได้อย่างมีประสิทธิภาพ รวมทั้งยังสามารถหาความสัมพันธ์ระหว่างเอนไซม์โดยใช้โมทีฟที่ได้จากบริเวณจับและบริเวณเร่งของเอนไซม์เป็นตัวแทนสายโปรตีน

BLOSUM62 Matrix

ลำดับของกรดอะมิโนในสายโปรตีนมีการพัฒนาตลอดเวลาและวิวัฒนาการได้ ดังนั้นจึงทำให้เกิดการเปลี่ยนแปลงในลำดับของกรดอะมิโนในสายโปรตีนที่เกิดขึ้น กรดอะมิโนตัวนั้นๆ บางครั้งสามารถแทนที่ด้วยกรดอะมิโนตัวอื่นได้ การกลายพันธุ์ของกรดอะมิโนหนึ่งไปยังอีกกรดอะมิโนหนึ่ง

วิวัฒนาการของสายโปรตีนเหล่านี้สามารถจะนำมาสร้างแบบจำลองโดยการให้คะแนนของเมทริกซ์หรือที่เรียกว่าตาราง BLOSUM62 (Henikoff, 1992) ดังภาพที่ 11 ตัวอย่างเช่นคะแนนจากการแทนกรดอะมิโน K ด้วย R คือ 2 คะแนน โดยช่วงของคะแนนจะอยู่ระหว่าง -4 ถึง 11 โดยคะแนนที่มากจะแสดงให้เห็นถึงความคล้ายคลึงกันของแทนที่กรดอะมิโน

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W

ภาพที่ 11 ตาราง BLOSUM62

โมทีฟและรีแอคทีฟโมทีฟ

โมทีฟ คือตัวแทนของสายโปรตีน เนื่องจากฐานข้อมูลที่มีข้อมูลโปรตีนเอนไซม์ส่วนมากอยู่ในรูปของสายลำดับโปรตีน (protein sequences) ดังนั้นตัวแทนสายโปรตีนที่เหมาะสมกับข้อมูลลักษณะนี้ก็คือรูปแบบลำดับสายโปรตีน (protein sequence pattern) ที่เรียกว่าโมทีฟ (motif) ซึ่งรู้จักกันดีมากว่า 25 ปีแล้ว (Weber *et al.,* 1982) โดยโมทีฟทำหน้าที่เป็นตัวแทนสายโปรตีนของกลุ่มโปรตีนหนึ่ง ยกตัวอย่างเช่นในโปรตีนกลุ่มหนึ่งที่มีคุณสมบัติ phosphorylation site มีรูปแบบสายลำดับ [ST]-x-[RK] เป็นโมทีฟ หมายถึงในโปรตีนกลุ่มนี้ทุกตัวจะมีสายลำดับโปรตีนที่ตำแหน่งใดใดเริ่มต้นด้วยกรดอะมิโนรหัส S หรือ T ตามด้วยกรดอะมิโนรหัสใดใดอีก 1 ตัว (แทนด้วยรหัส x) แล้วปิดท้ายด้วยกรดอะมิโนรหัส R หรือ K เป็นโมทีฟแสดงคุณสมบัติฟังก์ชั่น phosphorylation site ของกลุ่มโปรตีนนี้ ส่วนองค์ประกอบของโมทีฟที่แสดงด้วยสัญลักษณ์ [] หมายถึงกลุ่มแทนที่กรดอะมิโน (substitution group) ซึ่งสามารถเกิดขึ้นได้จากการกลายพันธุ์ในสายวิวัฒนาการของโปรตีนกลุ่มนั้น โดยในกรณีที่กลุ่มแทนที่กรดอะมิโนมีกรดอะมิโนเป็นสมาชิกเพียง 1 ชนิด เรียกบริเวณนั้นว่าบริเวณอนุรักษ์ (conserve region) เช่น [ST]-x-R กรดอะมิโน R ก็คือบริเวณอนุรักษ์

นอกจากนี้ โมทีฟยังสามารถมีรูปแบบสายลำดับที่ยืดหยุ่นได้ที่เรียกว่า gap ยกตัวอย่างเช่น โมทีฟ [ST]-x(1,2)-[RK] มีรูปแบบสายลำดับที่ยืดหยุ่นได้โดยตรงกลางของโมทีฟมีรูปแบบ x(1,2)

หมายถึงการมีกรดอะมิโนใดใดจำนวน 1 หรือ 2 ตัวก็ได้ เป็นต้น โดยการยืดหยุ่นดังกล่าวนี้ในทางชีววิทยาหมายถึงการแทรก (insertion) และการขาดหาย (deletion) ในสายวิวัฒนาการของโปรตีน

ดังนั้นองค์ประกอบพื้นฐานของโมทีฟจึงประกอบด้วยสิ่งที่เรียกว่า กลุ่มแทนที่กรดอะมิโน บริเวณอนุรักษ์ และ gap โดยแต่ละโมทีฟก็คือการเรียงลำดับกันของกลุ่มแทนที่ บริเวณอนุรักษ์ และ gap นั่นเอง เมื่อกำหนดให้โมทีฟเท่ากับ G[GA]KVLG การแทนที่สายโปรตีนด้วยโมทีฟจะได้ผลลัพธ์ดังตัวอย่างข้างล่าง โดยบริเวณที่ถูกแทนที่จะแสดงด้วยตัวอักษรที่ขีดเส้นใต้

MNDLSGKTVIITGGAKVLGAEAARQAVAAGARAKSADVLDE

รีแอคทีฟโมทีฟ (โมทีฟเชิงปฏิกิริยาเคมี) คือ โมทีฟจากข้อมูลบริเวณจับและบริเวณเร่งในการค้นพบรีแอคทีฟโมทีฟจากบริเวณจับและบริเวณเร่ง โดยทั่วไปใช้โครงสร้างข้อมูลที่เรียกว่า “บลอก” (block) ในการค้นพบรีแอคทีฟโมทีฟ โดยขั้นตอนพื้นฐานสามารถให้นิยามเป็นขั้นตอนได้ดังนี้

นิยามที่ 4 *บลอกจากบริเวณจับหรือบริเวณเร่ง* (หรือ block) B_{mxn}: เป็นเมทริกซ์ของกรดอะมิโน x_{ij} จำนวน m แถว n คอลัมน์ โดย i = 1 ถึง m และ j = 1 ถึง n ทั้งนี้ $x_{ij} \in \sum, \sum =$ {20 ชนิดของกรดอะมิโน}

บลอกเป็นโครงสร้างข้อมูลประกอบด้วยเซ็ทของลำดับกรดอะมิโนในบริเวณจับหรือบริเวณเร่งจำนวน m สาย ขนาดความยาว 15 กรดอะมิโน โดยตำแหน่งที่ 8 ซึ่งเป็นตรงกลางของบริเวณดังกล่าวเป็นกรดอะมิโนที่ถูกระบุว่าเป็นกลไกในการทำงานบริเวณจับหรือบริเวณเร่งดังกล่าวที่ได้จากฐานข้อมูล SWISSPROT ดังนั้น “บลอก” จึงสามารถจัดโครงสร้างข้อมูลให้อยู่ในรูปของเมทริกซ์ B_{mxn} ของกรดอะมิโน x_{ij} โดย m หมายถึงจำนวนสายของบริเวณจับหรือบริเวณเร่งในบลอก และ n หมายถึงขนาดความยาวของ บลอก (n=15) โดย i หมายถึงตำแหน่งสายลำดับกรดอะมิโนในแถวของเมทริกซ์บลอก และ j หมายถึง ตำแหน่งของกรดอะมิโนในคอลัมน์ของเมทริกซ์บลอก ทั้งนี้ $x_{ij} \in \sum, \sum =$ {20 ชนิดของกรดอะมิโน} ซึ่งจากโครงสร้างบลอกดังกล่าวนี้สามารถนำไปค้นพบกลุ่มแทนที่ซึ่งเป็นองค์ประกอบของโมทีฟได้ตามนิยามที่ 5

นิยามที่ 5 *ลำดับกรดอะมิโนในบริเวณจับหรือบริเวณเร่ง* (site sequence) s_i: คือสายลำดับกรดอะมิโน (string) $s_i = x_{i1}, x_{i2}, x_{i3},.., x_{in}$ ในบลอก B โดย $\forall\ s_i \in B_{mxn}$, $x_{ij} \in \sum$, และ $\sum$= {20 ชนิดของกรดอะมิโน}

นิยามที่ 6 *กลุ่มแทนที่กรดอะมิโน* (substitution group) $\hat{A}_j$: หมายถึงเซ็ทของกรดอะมิโนที่ได้จากบลอก B_{mxn} ที่ตำแหน่ง j ซึ่งก็คือกรดอะมิโน x_{ij} ทุกชนิดจากตำแหน่ง j ของทุกลำดับกรดอะมิโน s_i ใน B สามารถเขียนให้อยู่ในรูปฟอร์มของ $\hat{A}_j = \bigcup_{i=1}^{m} x_{ij}$ เมื่อ กรดอะมิโน $x_{ij} \in \sum$, และ $\sum$= {20 ชนิดของกรดอะมิโน}

ในกรณีที่ $\hat{A}_j$ มีแค่กรดอะมิโนเพียง 1 ชนิด เป็นกลุ่มของกรดอะมิโนชนิดพิเศษที่เราเรียกว่า “บริเวณอนุรักษ์” (conserved region)

นิยามที่ 7 *โมทีฟที่ค้นพบจากบลอก* (รีแอคทีฟโมทีฟ หรือ reactive motif) M: ก็คือกลุ่มแทนที่ทั้งหมดที่ได้จากทุกตำแหน่ง j ในบลอก B_{mxn} โดยสามารถเขียนให้อยู่ในรูปฟอร์มของ M = $\hat{A}_1, \hat{A}_2, \hat{A}_3, \ldots, \hat{A}_n$ เมื่อ กลุ่มแทนที่ $\hat{A}_j \subseteq \sum$ และ $\sum$= {20 ชนิดของกรดอะมิโน} หรือเขียนโมทีฟให้อยู่ในรูปฟอร์ม M = $\bigcup_{i=1}^{m} x_{i1}, \bigcup_{i=1}^{m} x_{i2}, \bigcup_{i=1}^{m} x_{i3}, .., \bigcup_{i=1}^{m} x_{in}$ เมื่อ กรดอะมิโน $x_{ij} \in \sum$, และ $\sum$= {20 ชนิดของกรดอะมิโน}

จากสูตรการค้นพบรีแอคทีฟโมทีฟจากบลอกในบริเวณจับหรือบริเวณเร่ง M = $\bigcup_{i=1}^{m} x_{i1}, \bigcup_{i=1}^{m} x_{i2}, \bigcup_{i=1}^{m} x_{i3}, .., \bigcup_{i=1}^{m} x_{in}$ สามารถวิเคราะห์ปัญหาเบื้องต้นของโมทีฟที่ได้จากบลอกก็คือ ที่ขนาดของโมทีฟ n = 15 คงที่ พบว่าคุณภาพของกลุ่มแทนที่ $\hat{A}_j$ หรือ $\bigcup_{i=1}^{m} x_{ij}$ ขึ้นกับจำนวนสายลำดับกรดอะมิโนในบลอกหรือค่า m นั่นเอง ซึ่งในกรณีที่ค่า m = 1 และแต่ละตำแหน่งของโมทีฟสามารถเป็นกรดอะมิโนได้ 20 ชนิด จะได้ค่าความน่าจะเป็นในการค้นพบโมทีฟ M ดังกล่าวในสายโปรตีนขนาดความยาว y อยู่ที่ประมาณ $y\left(\frac{1}{20}\right)^{15}$ ซึ่งโอกาสที่จะพบโมทีฟมีน้อยมาก (แม้ในฐานข้อมูลสายโปรตีนขนาดใหญ่) ทำให้นำโมทีฟนั้นไปใช้ประโยชน์ได้น้อย ในทางตรงกันข้ามในบลอกขนาดใหญ่ที่มีค่า m มาก จะทำให้กลุ่มแทนที่ในทุกตำแหน่ง j มีค่าใกล้เคียง $\sum$ ซึ่งก็คือมีความน่าจะเป็นในการค้นพบโมทีฟนั้นในทุกสายโปรตีน

ดังนั้นในการได้มาซึ่งรีแอคทีฟโมทีฟที่มีคุณภาพจึงขึ้นอยู่กับการพัฒนากลุ่มแทนที่ที่มีคุณภาพที่เราเรียกว่า “กลุ่มแทนที่ที่สมบูรณ์” (Maximal Substitution Group) และ “การพัฒนาคุณภาพบลอก” ซึ่งทำให้ได้บลอกที่มีขนาดเหมาะสมในการได้มาซึ่งโมทีฟที่มีคุณภาพ ซึ่งความสัมพันธ์ระหว่างขั้นตอนต่างๆ ในการค้นพบและพัฒนารีแอคทีฟโมทีฟเพื่อทำนายประเภทฟังก์ชั่นเอนไซม์จะสามารถศึกษาเพิ่มเติมโดยละเอียดได้จากงานวิจัย โมทีฟเชิงปฏิกิริยาเคมี (พีระ, 2551)

การจำแนกซับแฟมมิลี่ของเอมไซม์ (Enzyme Subfamily Prediction)

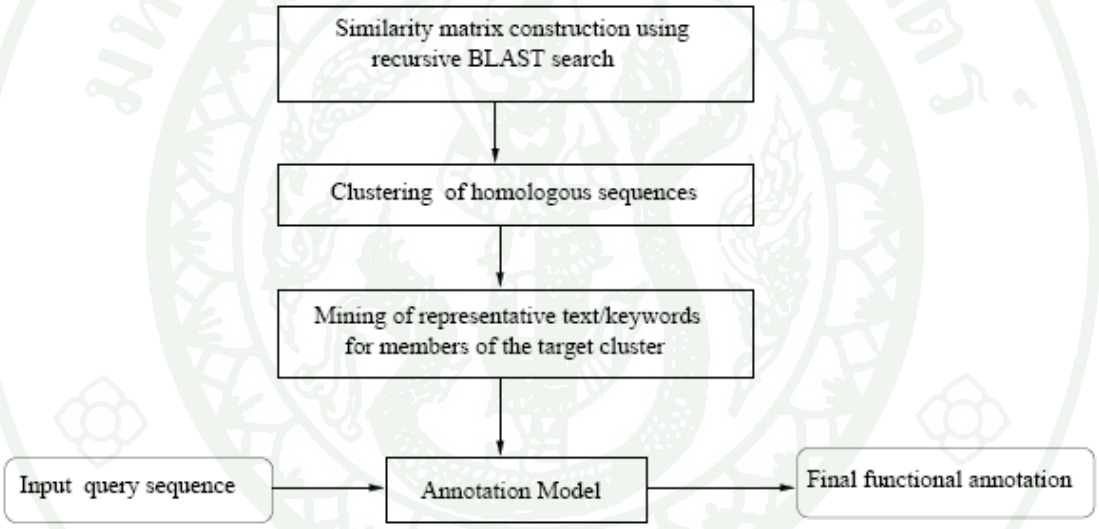
ปัญหาของการจำแนกซับแฟมมิลี่ของเอมไซม์คือการสร้างกระบวนการกำหนดประเภทให้กับข้อมูลที่ยังระบุประเภทไม่ได้ ตัวอย่างของ Application ที่ทำหน้าที่นี้เช่น ในกรณีของสาขาชีวสารสนเทศ ที่มีฐานข้อมูลเป็นลำดับนิวคลีโอไทด์ (DNA/RNA) หรือ ลำดับกรดอะมิโน

โปรตีน ประกอบไปด้วยลำดับของกรดอะมิโนเมื่อนักชีววิทยาต้องการทราบ Family หรือ Function ของโปรตีนใหม่ที่ค้นพบ เมื่อหลายปีที่ผ่านมาการขยายตัวของข้อมูลทางสาขานี้มีอัตราการฌติบโตที่สูงมาก สิ่งนี้เองจึงเป็นที่มาของความต้องการระบบอัตโนมัติในการจำแนกข้อมูลแบบมีลำดับโดยแบ่งได้เป็น 3 กลุ่มตามวิธีการที่ใช้ในงานวิจัยได้แก่ (1) การเปรียบเทียบความเหมือนของลำดับกรดอะมิโนในสายโปรตีน (homology-based approach) (2) การพิจารณาส่วนย่อยของสายโปรตีน (subsequence-based approach) และ (3) การสร้างตัวแทนสายโปรตีนหรือฟีเจอร์ (feature-based approach) ซึ่งมีรายละเอียดดังต่อไปนี้

ระบบจำแนกประเภทเอนไซม์โดยเปรียบเทียบความเหมือนของลำดับกรดอะมิโนในสายโปรตีน (Homology-based approach)

สมมติฐานของวิธีการนี้คือ เอนไซม์ที่มีลำดับของกรดอะมิโนหรือโครงสร้างปฐมภูมิที่คล้ายคลึงกัน (homology) จะมีฟังก์ชั่นที่เหมือนกัน วิธีการทำนายฟังก์ชั่นลักษณะนี้ เริ่มจากมีเอนไซม์ต้นแบบที่ทราบฟังก์ชั่นแล้ว จากนั้นทำการพิจารณาเอนไซม์ใหม่ที่ต้องการทราบฟังก์ชั่นว่ามีโครงสร้างปฐมภูมิคล้ายคลึงกับเอนไซม์ต้นแบบตัวใดมากที่สุด ก็จะทำนายได้ว่ามีฟังก์ชั่นการทำงานเหมือนกับเอนไซม์ต้นแบบตัวนั้นนั่นเอง สำหรับวิธีการพิจารณาความคล้ายคลึงของ

โครงสร้างปฐมภูมินั้นใช้วิธีการที่เป็นที่นิยมเช่น BLAST (Altschul *et al.,* 1990) PSI-BLAST (Altschul *et al.*, 1997) FASTA (Lipman and Pearson, 1985) และ HMM (Eddy, 1998) เป็นต้น ตัวอย่างงานวิจัยที่ใช้วิธีการนี้ได้แก่ GeneQuiz (Andrade, 1999) PEDANT (Riley *et al.*, 2005) และ Auto-FACT (Koski *et al.*, 2005) เป็นต้น นอกจากนี้ได้มีการพัฒนาวิธีการข้างต้น (Xie *et al*., 2002; Abascal and Valencia, 2003; Sasson *et al.,* 2006) โดยพยายามจัดกลุ่มเอนไซม์ต้นแบบที่มีค่าความเหมือนใกล้เคียงกัน จากนั้นการสืบค้นตัวแทนกลุ่ม และนำไปสร้างเป็นโมเดลทำนายฟังก์ชั่น สำหรับในการทำนายฟังก์ชั่นเอนไซม์ทำได้โดย นำเอนไซม์ที่ไม่ทราบฟังก์ชั่นนั้นไปสืบค้นในโมเดลและถ้ามีบางส่วนตรงกับตัวแทนกลุ่มใดของเอนไซม์ต้นแบบ จะได้ว่ามีฟังก์ชั่นการทำงานตามเอนไซม์กลุ่มต้นแบบนั้น ดังภาพที่ 12



ภาพที่ 12 แสดงไดอะแกรมวิธีการสร้างโมเดลทำนายฟังก์ชั่นด้วยวิธีการ homology-based ที่มีการปรับเปลี่ยนวิธีการบางส่วน

ที่มา: Computational Approaches for Protein Function Prediction: A Survey (Pandy, Kumar and Steinbach, 2006)

ข้อดีของการทำนายฟังก์ชั่นของเอนไซม์ด้วยวิธีการนี้คือเป็นวิธีการที่ง่ายและให้ความถูกต้องสูงสำหรับสายโปรตีนที่มีค่าความเหมือนระหว่างสายโปรตีนสูงๆ (pairwise sequence similarity) ยกตัวอย่างเช่นในงานวิจัยของ Tian et *al.* (2004) กล่าวว่าสายโปรตีนที่ไม่ทราบฟังก์ชั่น

ที่มีค่าความเหมือนกันระหว่างสายโปรตีนต้นแบบ เฉลี่ยประมาณ 60% นั้น จะให้ค่าความถูกต้องของการทำนายฟังก์ชั่นได้สูงถึง 90% เป็นต้น

ข้อเสียของการทำนายฟังก์ชั่นของเอนไซม์ด้วยวิธีการนี้คือค่าความถูกต้องของการทำนายจะลดลงมากเมื่อค่าความเหมือนระหว่างสายโปรตีนต้นแบบกับสายโปรตีนที่ต้องการทราบฟังก์ชั่นนั้นมีค่าน้อยกว่า 30% (Emanuelsson *et al.*, 2007; Chou and Shen, 2008)

ระบบจำแนกประเภทเอนไซม์โดยพิจารณาส่วนย่อยของสายโปรตีน (Subsequence-based approach)

เนื่องจากการทำงานของเอนไซม์นั้นเกิดขึ้นจากส่วนย่อยบางส่วนของลำดับกรดอะมิโนในสายโปรตีนเท่านั้น ดังนั้นวิธีการนี้จึงพยายามหาส่วนย่อยของสายโปรตีนที่มีความสำคัญต่อการเกิดฟังก์ชั่น โดยส่วนย่อยที่สำคัญนั้นได้แก่ส่วนที่เรียกว่า (1.) โมทีฟ (motif) เป็นบริเวณส่วนย่อยของสายโปรตีนที่มีตำแหน่งบางตำแหน่งร่วมกัน (conserved) เพื่อบ่งชี้ว่าเป็นเอนไซม์ในกลุ่มตระกูลหรือแฟมิลีเดียวกัน (Bork and Koonin, 1996) นอกจากนี้บริเวณส่วนย่อยดังกล่าวอาจเป็นบริเวณจับ (binding site) ในการเกิดปฏิกิริยาเคมีกับสารตั้งต้น หรืออาจเป็นบริเวณที่มีการติดกัน (interaction) ระหว่างโปรตีน ซึ่งบริเวณดังกล่าวเป็นข้อมูลที่มีประโยชน์มากต่อการทำนายฟังก์ชั่นของเอนไซม์ (Bork and Koonin, 1996; Huang and Brutlag, 2001) (2.) โดเมน (Domain) สืบเนื่องจากสมมติฐานที่ว่าฟังก์ชั่นหลายๆ ฟังก์ชั่น เกิดจากบริเวณส่วนย่อยของเอนไซม์มีโครงสร้างและฟังก์ชั่นที่แตกต่างกัน (Servant *et al.*, 2002) บริเวณดังกล่าวนั้นเรียกว่า ฟังก์ชั่นโดเมน (function domain) กล่าวคือฟังก์ชั่นของเอนไซม์นั้นเกิดจากการรวมตัวของฟังก์ชั่นย่อยๆ ของแต่ละโดเมนนั่นเอง

จากข้างต้นจะเห็นได้ว่า โมทีฟ หรือโดเมน นั้นเป็นส่วนที่สำคัญที่สามารถบ่งชี้ฟังก์ชั่นการทำงานของเอนไซม์ได้ งานวิจัยที่ใช้วิธีการนี้ได้แก่งานวิจัยของ Hannenhalli and Russell (2000) ได้พัฒนาวิธีการสืบค้นและระบุบริเวณส่วนย่อยของสายโปรตีนที่สามารถแบ่งแยกคลาสของเอนไซม์ได้ดีที่สุด โดยเริ่มจากทำการระบุตำแหน่งของบริเวณอนุรักษ์ ซึ่งได้จากการเทียบเรียงสายโปรตีนทุกสาย (multiple sequence alignment) ในกลุ่มแฟมิลีเดียวกัน จากนั้นทำการคำนวณค่าความสัมพันธ์ของเอ็นโทรปี (relative entropy) ในแต่ละตำแหน่งของกรดอะมิโนที่อยู่ในแฟมิลีเดียวกัน บริเวณที่มีค่าความสัมพันธ์ของเอ็นโทรปีรวมกันมากที่สุด จะเป็นบริเวณที่สามารถแบ่งแยกคลาสของเอนไซม์ได้ดีที่สุด จากนั้นนำมาเป็นฟีเจอร์ให้กับฮิดเด้นมาร์คอฟโมเดลเพื่อสร้าง

เป็นโมเดลจำแนกประเภทเอนไซม์ งานวิจัยของ Blekas *et al.* (2005) ใช้โมทีฟจากฐานข้อมูล PROSITE (Hulo, 2006) ซึ่งได้จากการทดสอบโดยผู้เชี่ยวชาญ และโมทีฟที่ได้จากการสืบค้นด้วยอัลกอริทึม MEME (Bailey *et al.*, 1999) เป็นฟีเจอร์ในการสร้างโมเดลจำแนกประเภทเอนไซม์ด้วยนิวรอน เน็ทเวิร์ค (neuron network) งานวิจัยของ Ben-Hur and Brutlag (2005) ใช้การพิจารณาความถี่ของการปรากฏโมทีฟแต่ละตัวในสายโปรตีนเป็นตัวแทนสายโปรตีนและสร้างเป็นฟีเจอร์เวคเตอร์นำเข้าอัลกอริทึมวิธีเรียนรู้ ซัพพอร์ตเวคเตอร์แมชชีน งานวิจัยของ Schug *et al.* (2002) เป็นงานวิจัยแรกที่เริ่มมีการนำโดเมนมาใช้ในการทำนายฟังก์ชั่น ในงานวิจัยนี้ทำการสกัดโดเมนจากฐานข้อมูลโปรดอม (ProDom) (Servant *et al.*, 2002) และฐานข้อมูลซีดีดี (Conserved Domain Database, CDD) (Marchler-Bauer *et al.* 2005) จากนั้นสร้างโมเดลโดยใช้การสืบค้นด้วยอัลกอริทึม BLAST งานวิจัย WILMA (Prlic *et al.*, 2004) ใช้โดเมนจากฐานข้อมูล PROSITE และ Pfam (Sonnhammer *et al.*, 1997) และ PRINTS (Attwood *et al.*, 2003) ส่วนวิธีการทำนายฟังก์ชั่นใช้วิธีการสืบค้นร่วมกันระหว่างหลายอัลกอริทึม (ensemble) ได้แก่ RPS-BLAST (Altschul *et al.* 1997) PROSITE scans และ Finger-PRINTScan (Scordis *et al.*, 1999) เป็นต้น

ข้อดีของการทำนายฟังก์ชั่นของเอนไซม์ด้วยวิธีการนี้คือส่วนย่อยของลำดับกรดอะมิโนที่ได้จากโมทีฟหรือโดเมนนั้น มีความสัมพันธ์กับโครงสร้างโปรตีนที่ทำให้เกิดฟังก์ชั่นการทำงานเฉพาะอย่าง ซึ่งเป็นบริเวณที่สามารถบ่งบอกฟังก์ชั่นของเอนไซม์ได้เป็นอย่างดี

ข้อเสียของการทำนายฟังก์ชั่นของเอนไซม์ด้วยวิธีการนี้คือการนิยามส่วยย่อยของลำดับกรดอะมิโนว่าเป็นโมทีฟหรือโดเมนนั้นยังมีความคลาดเคลื่อนอยู่ และการใช้โมทีฟหรือโดเมนเป็นตัวแทนสายโปรตีนนั้น ในบางกรณีอาจยังไม่เพียงพอที่จะใช้เป็นตัวแทนสายโปรตีนทั้งเส้นได้

ระบบจำแนกประเภทเอนไซม์โดยการสร้างตัวแทนสายโปรตีนหรือฟีเจอร์ (Feature-based approach)

เป็นการแปลงข้อมูลปฐมภูมิให้อยู่ในรูปแบบฟีเจอร์เวคเตอร์ซึ่งสามารถใช้เป็นตัวแทนสายโปรตีนได้ โดยฟีเจอร์แต่ละตัวอาจเป็นข้อมูลเกี่ยวกับชีววิทยาหรือชีวเคมี หรือเป็นชุดของกรดอะมิโนที่เกิดขึ้นบ่อย (frequent item set) จากนั้นจึงนำฟีเจอร์เวคเตอร์ดังกล่าวมาสร้างโมเดลจำแนกประเภทโดยใช้อัลกอริทึมเรียนรู้ได้แก่ ซัพพอร์เวคเตอร์แมชชีน (SVM) นิวรอนเน็ตเวิร์ค (Neuron Network) เนอีฟเบย์ (Naïve Bayesian) หรือซี 4.5 (C4.5) เป็นต้น

งานวิจัยที่ใช้วิธีการสร้างตัวแทนสายโปรตีนหรือฟีเจอร์ได้แก่ งานวิจัย SVM-Prot (Cai *et al.*, 2003) ทำการสร้างตัวแทนสายโปรตีนโดยพิจารณาคุณสมบัติของกรดอะมิโน เช่นแรงวัลเดอร์วาวล์ การมีขั้วไม่มีขั้ว การมีประจุไม่มีประจุ และแรงตึงผิวของกรดอะมิโน งานวิจัย CD (Chou and Elrod, 2003) ทำการพิจารณาความถี่ของการปรากฏของลำดับกรดอะมิโนในสายโปรตีน งานวิจัย CDA (Chou, 2005) เป็นการพิจารณาคุณสมบัติทางเคมีร่วมกับพิจารณาความถี่ของการปรากฏของกรดอะมิโน และสร้างโมเดลจำแนกประเภทด้วยโคแวเรียนดิสคริมิแนนท์ งานวิจัยของ Zhou *et al.* (2007) ใช้วิธีการสกัดฟีเจอร์เช่นเดียวกับงานวิจัย CDA แต่ต่างกันตรงที่สร้างโมเดลจำแนกประเภทด้วยอัลกอริทึมเรียนรู้ ซัพพอร์ตเวคเตอร์แมชชีน งานวิจัยของ King *et al.* (2000) สร้างฟีเจอร์เวคเตอร์จากชุดของกรดอะมิโนที่เกิดขึ้นบ่อย โดยสร้างโมเดลจำแนกประเภทด้วยอัลกอริทึม C4.5 เป็นต้น

งานวิจัยที่เกี่ยวข้อง

ในวิทยานิพนธ์เล่มนี้จะนำเสนองานวิจัยที่เกี่ยวข้องกับเทคนิคการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์แบบมีลำดับ (Sequential Associative Classification) โดยจะนำเสนออัลกอริทึม CBS (Tseng and Lee, 2005) ซึ่งเป็นต้นแบบของเทคนิคดังกล่าว

CBS อัลกอริทึม

Classification Method byUsing Sequential Patterns (CBS) เป็นงานแรกในการเสนอวิธีการรวมเทคนิคการหารูปแบบลำดับเหตุการณ์ที่น่าสนใจ (Sequential Pattern Mining) เข้ากับการคำนวณเชิงสถิติเพื่อหาความเป็นไปได้ (Probabilistic Induction) ซึ่งอัลกอริทึม CBS ประกอบด้วย 2 ส่วนนั่นคือ ในคือในส่วนการสร้างกฎความสัมพันธ์ เรียกว่า CBS-RG ซึ่งอัลกอริทึมในส่วนแรกนี้ก็จะอ้างอิงกับอัลกอริทึม Apriori-like (Agrawal, 1995) และในส่วนของการสร้างโมเดลในการทำนาย เรียกว่า CBS-CB และแบ่งออกเป็น 2 วิธีด้วยกันคือ CBS_ALL และ CBS_CLASS

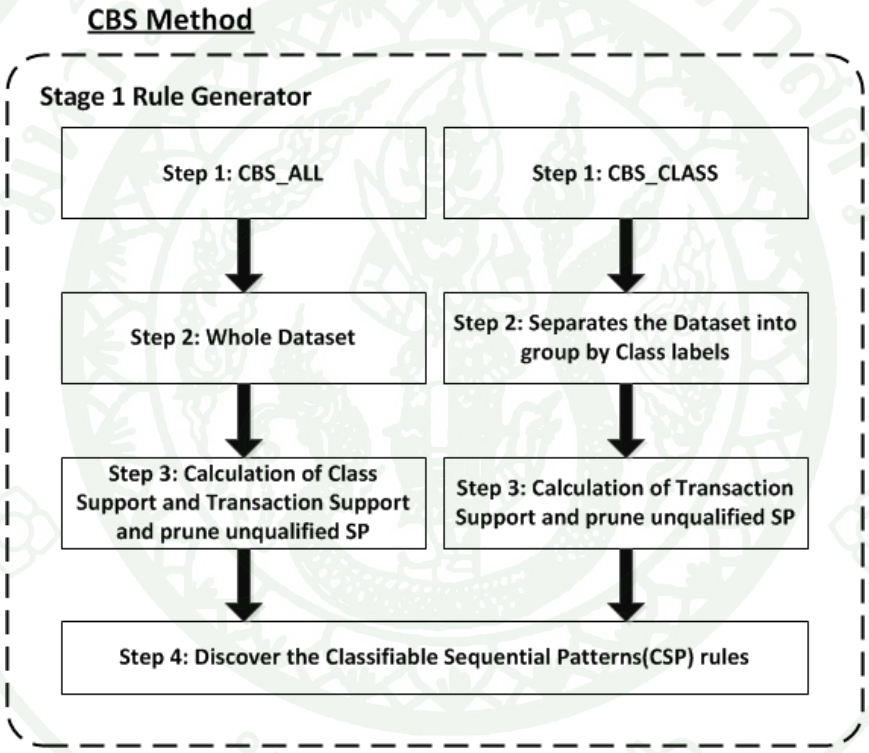
ก่อนที่จะทำความเข้าใจอัลกอริทึม CBS_ALL และ CBS_CLASS นั้น จะต้องทราบเกี่ยวกับแนวคิดเบื้องต้นที่ใช้ในอัลกอริทึมในส่วนของ CBS-RG เสียก่อน โดยจุดมุ่งหมายในส่วนของ CBS-RG นั่นก็คือการค้นหากฎความสัมพันธ์ทั้งหมดที่มีค่าสนับสนุนมากกว่าค่าสนับสนุนขั้นต่ำ โดยที่กฎความสัมพันธ์เหล่านั้นจะอยู่ในรูปของ (อธิบายเพิ่มเติมในส่วนถัดไป)

$$SP_i \rightarrow C_m$$

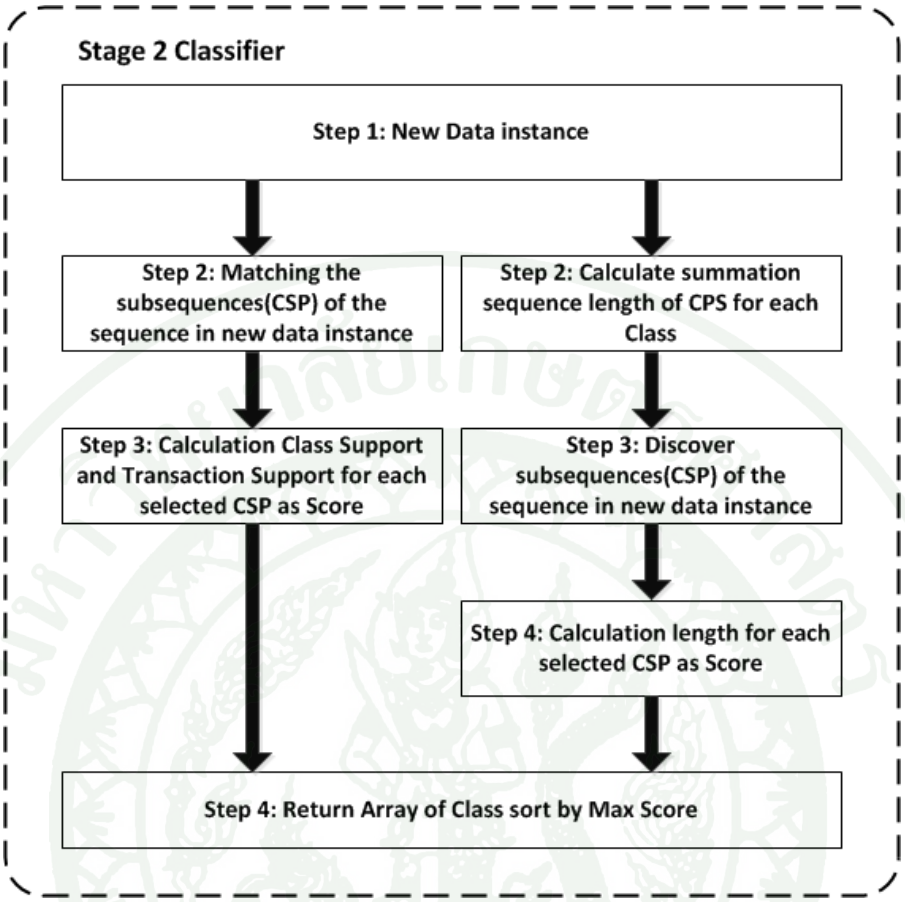
เมื่อ D_i แทนข้อมูลแบบมีลำดับของคลาสที่ i ดังนั้นฐานข้อมูลทั้งหมดเท่ากับ $D = \{D_1, D_2, D_3, ...D_N\}$ เมื่อกำหนดให้มี Class ทั้งหมด N คลาสในฐานข้อมูล ในแต่ละ D_i จะประกอบไปด้วยข้อมูบแบบมีลำดับในรูปของ $\{a_1, a_2, a_3, ..., a_n\}$ เมื่อ a_n เป็นค่า ณ ตำแหน่งที่ n ในลำดับของข้อมูล เพื่อให้ง่ายต่อความเข้าใจ สมมุติให้ค่า ณ ตำแน่ง n ใดๆถูกแบ่งช่วงและแปลงไปอยู่ในรูปค่าที่จำแนกประเภทได้ (Categorical Values) ไว้ล่วงหน้าแล้ว

Classifiable Sequential Pattern (CSP) จะถูกใช้เป็นกฎในการแยกประเภทสำหรับสร้างโมเดลในการทำนาย กฎ CSP จะอยู่ในรูปของ $SP_i \rightarrow C_m$ เมื่อ SP_i เป็นลำดับของค่าที่จำแนกประเภทได้ $SP_i = a_2 \rightarrow a_3 \rightarrow a_7 ...,$ และ C_m คือคลาสที่ m

จากความเข้าใจพื้นฐานที่ได้กล่าวมาแล้วในขั้นต้น CBS จะแบ่งได้เป็น 2 แนวทางคือ CBS_ALL และ CBS_CLASS และจะมีขั้นตอนย่อย 2 ขั้นตอนด้วยกันคือ CBS-RG และ CBS-CB ดังภาพที่ 13 และภาพที่ 14



ภาพที่ 13 CBS-RG



ภาพที่ 14 CBS-CB

1. CBS_ALL โดยหลักการของวิธีนี้จะหา CSP จากทั้งฐานข้อมูลและคำนวน Classify-Score จาก**แต่ละท**รานเซ็คชั่น (Transaction) โดยอัลกอริทึมจะคำนวนหาค่า Transaction Support และ Class Support ไปพร้อมๆกันโดยนำอัลกอริทึม Apriori-like (Agrawal, 1995) มาปรับปรุงโดยเพิ่มการนับ Class Support เข้าไปดังรูปที่ 15

```
CBS _ ALL (Dataset D, min_seq_su p, min_rule_s up)
{
1  CSP_1 = {large 1-items}
2     for(i = 2;CSP_i ≠ φ;i++) do
3        SP_i = gen _ candidateS P(CSP_{i-1});
4        for each data d ∈ D do
5           SP_s = SP_i ∩ subseq(d);
6           for each sp ∈ SP_s do
7              sp.seq_ sup++;
8              sp.class_ sup[d.class]++;
9           end
10       end
11       CSP_i = {sp|sp ∈ SP_i, sp.seq_ sup ≥ min_seq_su p
 and ∃w let sp.class_ sup[w] / sp.seq_ sup ≥ min_rule_s up}
12    end
13 CSP = ∪_i CSP_i
}
```

ภาพที่ 15 CBS_ALL (RG) อัลกอริทึม

จากพื้นฐานดาต้าไมนิ่งที่ได้กล่าวไปแล้วเพื่อให้ความเข้าใจตรงกัน จะกำหนดให้ $CSP_i = L_i$ เมื่อ L_i = large $i-items$ และ $SP_i = C_i$ เมื่อ C_i คือ แคนดิเดท (Candidate) และ i คือ ความยาวของ items จากรูปที่ 8 จะได้ CSP_1 จาก SP_1 ที่ผ่านค่า Min.Support แล้วเพื่อมาสร้าง SP_2 (ขั้นตอนที่ 1-3) ต่อมาจะนำ SP_i ไปหาว่าเป็น Subsequence ของแต่ละ Transaction d ไหมเมื่อ $d \in D$ แล้วนำไปเก็บไว้ที่ตัวแปร SP_s เมื่อ $sp \in SP_s$ (ขั้นตอนที่ 4-5) หลังจากนั้นจะทำการเพิ่มค่า Support ของ sp และเพิ่ม Class Support โดยเช็คว่า sp ที่เป็น Subsequence ของ d นั้นๆมีคลาสเป็นอะไรและเพิ่มค่า Support ของแต่ละคลาสลงใน $class_$sup อาเรย์ (ขั้นตอนที่ 6-9) หลังจากนั้นจะทำการตัดกฎ (Prune) ที่ไม่ผ่านค่า Min.Support และ Class Support ก็จะได้ CSP_2 (ขั้นตอนที่ 10-11) และกลับไปทำซ้ำในขั้นตอนที่ 2 จนกว่า CSP_i จะเท่ากับเซตว่างก็จะได้กฎทั้งหมดมาเก็บไว้ใน CSP (ขั้นตอนที่ 12-13) และจะนำไปใช้สร้างโมเดลในการทำนายต่อไป

ในขั้นตอนของโมเดลการทำนาย CBS_ALL (CB) นั้นจะใช้วิธีการให้คะแนนในการกำหนดคลาสให้กับข้อมูลทำสอบ *sequence* x ดังภาพที่ 16

```
Class _of_ sequence (sequence x)
{
1  M = φ;
2  for each csp_i ∈ CSP do
3     if csp_i.sp ∈ subseq (x)
4        M.add(csp_i);
5  end
6  score_arra y[] = new array[clas s_set(D).c ount];
7  for each csp_m ∈ M do
8     for each c_n ∈ class _set (D) do
9        score_arra y[n] + =
csp_m. support / csp_m.class_sup [n];
10    end
11 end
12 k = indexof (Max{score_ array[]});
13 return k;
}
```

ภาพที่ 16 CBS_ALL (CB) อัลกอริทึม

ในขั้นตอนแรกจะทำการหาว่ามีกฎ (csp_i) ใดบ้างที่เป็น Subsequence ของ x โดยจะเก็บกฎเหล่านั้นไว้ในตัวแปร M (ขั้นตอนที่ 1-5) หลังจากนั้นจะประกาศตัวแปรอาเรย์ของคะแนน (score_array[ ]) มีจำนวนช่องเท่ากับจำนวนคลาส (ขั้นตอนที่ 6) หลังจากนั้นจะใช้ Class Support และ Sequence Support ที่เก็บมาในตอนแรกมาคำนวนคะแนนของแต่ละ csp_m และเก็บคะแนนลงใน score_array ตามคลาสของกฎ (ขั้นตอนที่ 7-11) ในตอนสุดท้ายจะทำการตอบคลาสที่มีคะแนนสูงสุด (ขั้นตอนที่ 12-13)

2. CBS_CLASS หลักการของวิธีนี้จะแบ่งฐานข้อมูลออกเป็นแต่ละคลาสและคำนวนหา CSPของแต่ละคลาส โดยจะต่างจาก CBS_ALL ตรงที่ใช้แค่ค่า Min.Support โดยขั้นตอนการทำงานดูได้จากภาพที่ 17 โดยยังใช้อัลกอริทึม Apriori-like (Agrawal, 1995) เป็นพื้นฐาน

```
CBS_CLASS( Dataset D , min_sup)
{
1  for each c_i ∈ class_set( D ) do
2      D_i = class _ dataset ( D , c_i );
3      CSP_i = FindSP( D_i , min _ sup )
}
FindSP(Dat aset D, min_sup)
{
4  SP_1 = {large 1 - items}
5  for( i = 2; SP_{i-1} ≠ φ ; i + + ) do
6      SP_C_i = gen_candid ateSP( SP_{i-1} );
7      for each data d ∈ D do
8          SP_s = SP_C_i ∩ subseq(d);
9          for each sp ∈ SP_s do
10             sp.sup + + ;
11         end
12     end
13     SP_i = {sp | sp ∈ SP_s , sp.sup ≥ min_sup}
14 end
15 return ⋃_i SP_i
}
```

ภาพที่ 17 CBS_CLASS (RG) อัลกอริทึม

กำหนดให้ $SP\_C_k = C_k$ (Candidate itemset of size k) และ $SP_i = L_i$ (Frequent Itemsets of size i) จากรูปที่ 9 ในขั้นตอนแรกจะทำการแบ่งฐานข้อมูลออกเป็นแต่ละคลาส (D_i) แล้วเรียก Method FindSP โดยส่ง Input เป็น D_i และค่า Min.Support (ขั้นตอนที่ 1-3) หลังจากนั้นใน Method FindSP จะทำการสร้าง Candidate จาก SP_1 โดยการ Join กันเป็น SP_C_2 (ขั้นตอนที่ 4-6) หลังจากนั้นจะนำ Candidate ที่ได้ไปตรวจดูว่าเป็น Subsequence ของ d หรือไม่และเก็บตัวที่

เป็นเข้า SP_s (ขั้นตอนที่ 7-8) หลังจากนั้นจะนับ Support ของแต่ละ sp (ขั้นตอนที่ 9-12) เช็คว่า sp ตัวไหนผ่าน Min.Support นำเข้ามาเก็บใน SP_i (ขั้นตอนที่ 13) หลังจากนั้นก็จะกลับไปทำซ้ำในขั้นตอนที่ 2-13 ใหม่จนกว่า SP_{i-1} จะเท่ากับเซตว่างแล้วจะคืนผลลัพธ์ให้กับ CSP_i เมื่อจับขั้นตอนเหล่านี้จะได้ CSP ของแต่ละคลาสเพื่อนำกฎเหล่าไปใช้ทำนายกฎต่อไป

ขั้นตอนของโมเดลการทำนาย CBS_CLASS (CB) จะใช้ความยาวกฎเป็นคะแนนของ CSP หลังจากนั้นทำการการนอร์มัลไลซ์คะแนนรวมของแต่ละคลาสให้เป็นมาตรฐานเดียวกันโดยมีค่าสูงสุดเท่ากับ 1 โดยมีขั้นตอนการทำงานดังภาพที่ 18

```
class_of_s   equence(se   quence   x)
{
1 total_   score[]   = new   array[clas   s_count(D)   ];
2 for   each   csp  i  ∈ CSP   do
3   to   tal_score[   csp  i .class]   + = csp  i .sp.length   ;
4 end
5 score_arra   y[] = new   array[clas   s_count(D)   ];
6 for   each   csp  i  ∈ CSP   do
7   if   csp  i  ∈ subseq(x)
8     for   each   c  m  ∈ belong_cla   ss_set(   csp  i )
9       score_arra   y[m] + =
csp  i .sp.length   / total_s   core[m];
10     end
11   end   if
12 end
13 k = index_of(M   ax{score_a   rray[]});
14 return   c  k
}
```

ภาพที่ 18 CBS_CLASS (CB) อัลกอริทึม

ในขั้นตอนแรกจะสร้างอาเรย์ผลรวมของคะแนน (total_score[ ]) มีจำนวนช่องเท่ากับคลาสหลังจากนั้นจะนำความยาวกฎของแต่ละคลาสมาบวกกันและเก็บลง total_score[ ] (ขั้นตอนที่ 1-4) ในขั้นต่อมาสร้างอาเรย์ของคะแนน (score_array[ ]) และนำกฎ csp_i เมื่อ $csp_i \in CSP$ ไปเช็คว่าเป็น Subsequence ของ *sequence* x หลังจากนั้นจะคำนวนคะแนนโดยนำความยาวของกฎ ($csp_i.sp.length$) หารด้วยผลรวมของคะแนนที่ตรงกับคลาสของมัน (C_m) แล้วใส่ผลลัพธ์ลงในอาเรย์ของคะแนนช่องที่ m (ขั้นตอนที่ 5-12) และในตอนสุดท้ายจะคืนค่าของคลาสจะช่องของคะแนน score_array[ ] ที่มีค่ามากที่สุด

จากงานวิจัย CBS ได้นำเสนอเทคนิคใหม่โดยการผสมผสานกันระหว่าง การหารูปแบบลำดับเหตุการณ์ที่น่าสนใจ (Sequential Pattern Mining) และการคำนวณเชิงสถิติเพื่อหาความเป็นไปได้ (Probabilistic Induction) เป็น CBS (Classification Method by Using Sequential Patterns) (Tseng and Lee, 2005) โดยใช้ Subsequence-based approach เป็นพื้นฐาน ในการจำแนกประเภท ซึ่งจากการทดลองในการจำแนกซับแฟมมิลี่ของเอมไซม์พบว่าให้ความถูกต้องในการทำนายได้ไม่ดีนัก

ข้อดีของการทำนายฟังก์ชั่นของเอนไซม์ด้วยวิธีการนี้คือจะได้กฎที่มีความเกี่ยวข้องกับคลาสนั้นๆโดยตรง

ข้อเสียของการทำนายฟังก์ชั่นของเอนไซม์ด้วยวิธีการนี้คือปัญหาของระบบเกิดขึ้นทั้งในส่วนการสร้างกฎและในส่วนการสร้างโมเดลในการทำนาย โดยปัญหาในส่วนของการสร้างกฎความสัมพันธ์จากงานทดลองแรกๆ คือการสร้างกฎความสัมพันธ์โดยใช้ MEME (Multiple Em for Motif Elicitation) (Timothy and Charles, 1994) ในการหาโมทีฟ พบว่ากฎความสัมพันธ์ยึดติดกับคลาสมากเกินไปอีกทั้งในการหาโมทีฟไม่ได้ใช้ความรู้ทางด้านชีววิทยามาช่วย ทำให้โมทีฟที่ได้ไม่สามารถอธิบายคุณสมบัติด้วยตัวมันเองได้ ทั้งด้านความแม่นยำของโมเดลที่จะต้องใช้กฎเหล่านั้นในการทำนายโดยใช้เพียงแค่ คำนวณเชิงสถิติเพื่อหาความเป็นไปได้ (Probabilistic Induction)

อุปกรณ์และวิธีการ

อุปกรณ์

ฮาร์ดแวร์

1. เครื่องคอมพิวเตอร์ 1 เครื่อง ประกอบด้วยอุปกรณ์ดังต่อไปนี้
2. ซีพียู (CPU) Intel Q8200 ความเร็ว 2.3 GHz
3. หน่วยความจำหลัก 4 GB
4. ฮาร์ดดิกส์ขนาด 1 TB

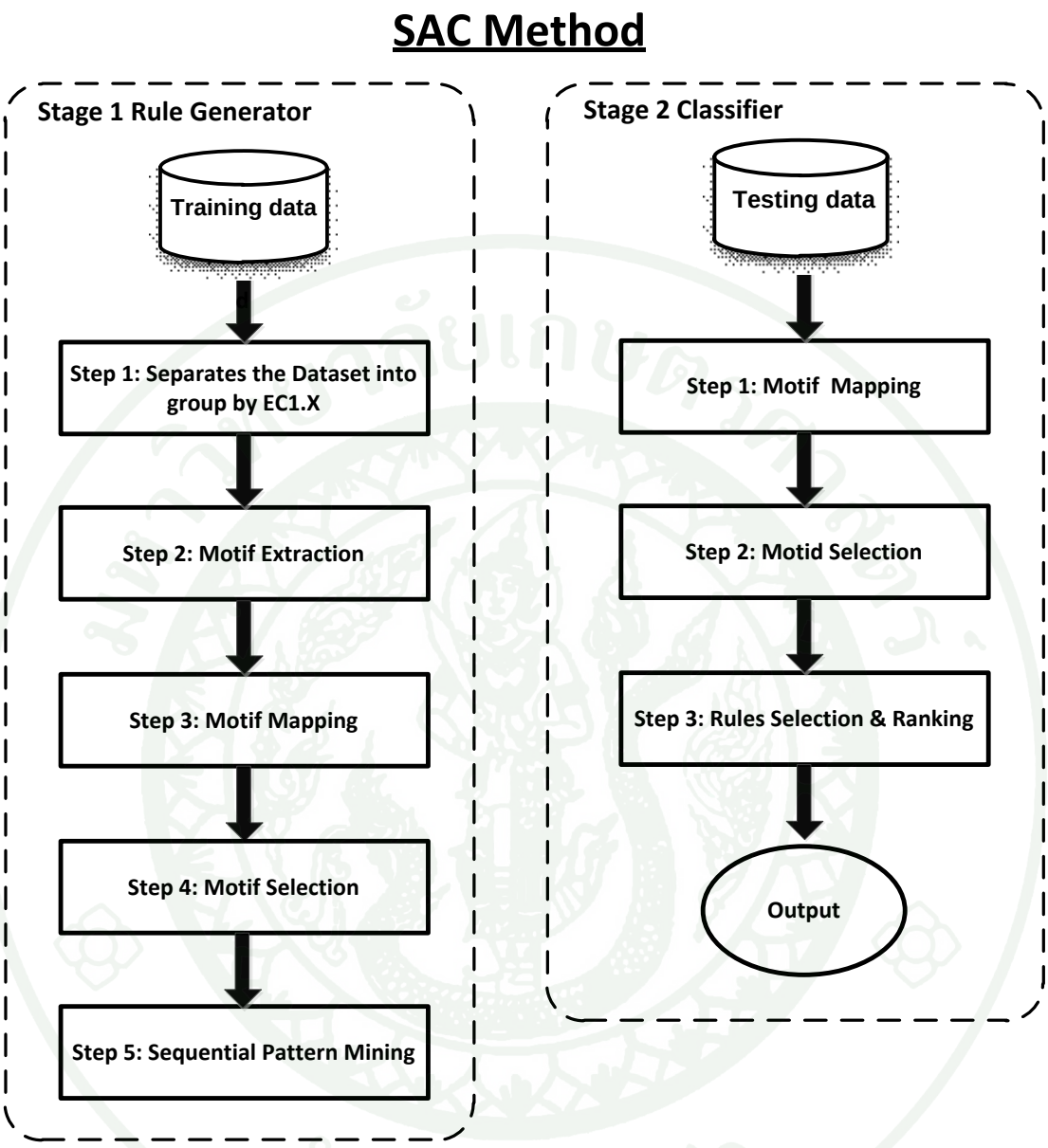
ซอฟต์แวร์

1. ระบบปฏิบัติการ Windows Server 2003 R2
2. SQL Server 2003 R2
3. Microsoft Visual Studio 2008

วิธีการ

ภาพรวมของระบบ

ดังที่กล่าวมาแล้วในส่วนของงานวิจัยที่เกี่ยวข้อง ดังนั้นในการเพิ่มประสิทธิภาพในส่วนของการสร้างกฎความสัมพันธ์และการสร้างโมเดลในการทำนาย ในวิทยานิพนธ์เล่มนี้จะนำเสนอนำเสนออัลกอริทึม SAC (Sequential Associative Classification) ซึ่งเป็นอัลกอริทึมที่ใช้ในการจำแนกซับแฟมมิลี่ของเอมไซม์ มีวัตถุประสงค์คือ ค้นหากฎความสัมพันธ์ที่ให้ความแม่นยำสูงในการนำไปใช้จำแนกซับแฟมมิลี่ของเอมไซม์ และสามารถนำกฎเหล่านั้นไปให้ผู้เชี่ยวชาญในด้านชีวสารสนเทศศาสตร์วิเคราะห์และตีความต่อไปได้ก่อนที่จะทำความเข้าใจในส่วนรายละเอียดนั้น สามารถที่จะดูภาพรวมของระบบได้จากภาพที่ 19 โดยจะแบ่งออกเป็น 2 ส่วนด้วยกันใน Stage 1 จะเป็นในส่วนของการสร้างกฎความสัมพันธ์และ Stage 2 จะเป็นส่วนของตัวจำแนกประเภทที่จะใช้ทำนาย Unseen Data



ภาพที่ 19 ภาพรวมของ SAC

การสร้างกฎความสัมพันธ์

การเตรียมข้อมูล (Data-Preprocessing)

1. การสกัดคุณสมบัติ (Motif Extraction)

ในงานวิจัยนี้ได้นำประโยชน์จาก โมทีฟเชิงปฏิกิริยาเคมี (พีระ, 2551) ในการสกัดคุณสมบัติวิธีการนี้ทำให้ได้โมทีฟจากส่วนที่เกี่ยวข้องกับฟังก์ชั่นเอมไซม์โดยตรงทำให้ได้โมทีฟที่มีคุณสมบัติที่ขึ้นตรงกลับคลาสนั้นโดยเฉพาะและโมทีฟที่เป็นตัวร่วมระหว่างคลาสนำไปใช้เป็นตัวแทนของโมทีฟในการสร้างกฎความสัมพันธ์ต่อไป โดยจะได้โมทีฟออกมาในรูปแบบของ กลุ่มของสัญลักษณ์ที่ใช้ในการเปรียบเทียบข้อมูล (Regular Expression) โดยจะจัดกลุ่มโมทีฟตามฟังก์ชั่นจากภาพที่ 20 จะเห็นว่าฟังก์ชั่น ion-binding site สามารถมีได้หลาย Regular Expression คือ โมทีฟตัวที่ 1 และ 2 แต่มีฟังก์ชั่นการทำงานเดียวกัน

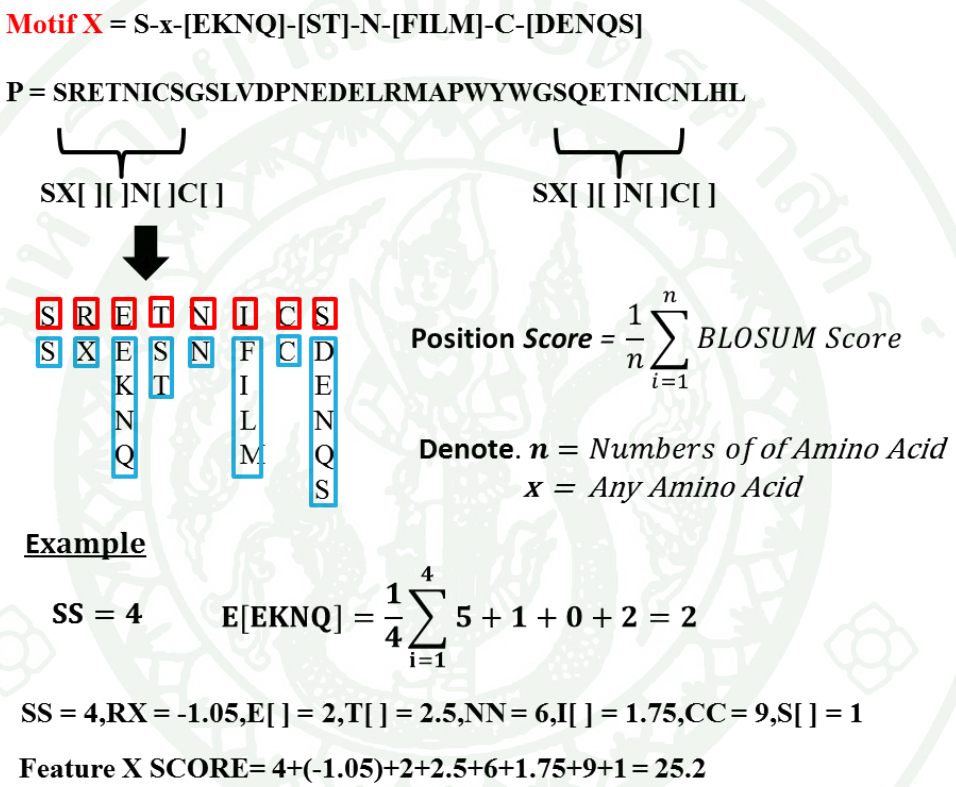
IndexMotif	ClassID	Function	Length	Regular expression
1	1	ion-binding site	8	S-x-[EKNQ]-[ST]-N-[FILM]-C-[DENQS]
2	1	ion-binding site	5	C-x-x-C-G
3	1	active site	9	S-x-[EKNQ]-[ST]-N-[FILM]-C-[DENQS]-C
4	2	Copper B	7	C-x-x-C-x-x-C

ภาพที่ 20 ตารางแสดงคุณสมบัติโมทีฟ

2. การหาโมทีฟในสายโปรตีน (Motif Mapping)

ในขั้นตอนนี้จะนำโมทีฟที่ได้จากภาพที่ 20 มาทำการค้นหาว่าในแต่ละสายของโปรตีนพบโมทีฟอะไรบ้าง แต่ไม่สามารถนำ Regular Expression มาใช้โดยตรงได้เพราะเมื่อ คุณลักษณะสำคัญ (Feature) ที่ค้นพบในสายมีการซ้อนทับกันในบางตำแหน่งจะทำให้ไม่สามารถรู้ได้ว่า Feature ตัวไหนจะเป็นตัวแทนที่ดีที่สุดของช่วงสายโปรตีนนั้นจึงใช้ตาราง BLOSUM62 (Henikoff,

1992) ซึ่งเป็นตารางคะแนนความเหมือนมาใช้ในการคำนวณในการเลือกโมทีฟที่ดีที่สุดโดยจะนำโมทีฟทุกตัวที่ได้ไปค้นหาตั้งแต่ตำแหน่งแรกสุดของสายจนจบสายโปรตีนจากภาพที่ 21 ให้ P เป็นสายโปรตีนเมื่อนำโมทีฟ X ไปค้นหาปรากฏว่าพบ 2 ครั้งในสายของโปรตีนโดยนำช่วงที่เหมือนไปหาคะแนนจากตาราง BLOSUM62 ในภาพที่ 11 จากรูปที่ 21 ตำแหน่งแรกที่พบโมทีฟได้คะแนนเท่ากับ 25.2



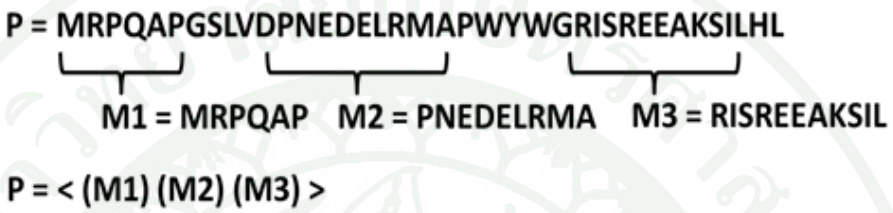
ภาพที่ 21 การคำนวนคะแนนของโมทีฟที่พบในสายโปรตีน

3. การเลือก (Motif Selection)

ในขั้นตอนนี้จะเป็นการเลือกตัวแทนโมทีฟที่พบในสายโปรตีนโดยแปลงจากสายของตัวอักษรเป็นโมทีฟ จากภาพที่ 22 ให้ P แทนสายโปรตีนและ M_n เป็นแอททริบิวต์โมทีฟตัวที่ n

เมื่อนำโมทีฟที่ได้มาเลือกจะสามารถแทนที่สายโปรตีน P ได้ด้วย M1, M9, M5 โดยถ้ามอง P เป็นทรานเซ็คชั่นในเรื่องของปัญหา Sequential Pattern Mining และโมทีฟเป็นไอเท็มแล้ว

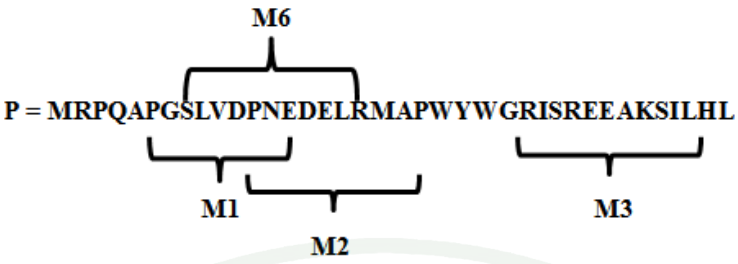
จะได้ Transaction P = < (M1) (M2) (M3) > โดยเกิด M1, M2, M3 ตามลำดับในขั้นตอนนี้ได้เสนอการเลือกโมทีฟแบบ การเลือกโมทีฟที่ดีที่สุด (Best Motif Selection)



ภาพที่ 22 การแทนที่สายโปรตีนด้วยโมทีฟ

4.1 การเลือกโมทีฟที่ดีที่สุด (Best Motif Selection)

จากขั้นตอนที่ 3 จะได้ตำแหน่งของโมทีฟทั้งหมดเก็บไว้ในฐานข้อมูลโดยจะทำการ Query (การเรียกดูสารสนเทศ) โดยมีเงื่อนไขคือเรียงตามตำแหน่งของโมทีฟจากน้อยไปมาก จากภาพที่ 23 ผลลัพธ์ที่ได้จากการ Query นั้นพบว่าในสายโปรตีน P พบโมทีฟ M1, M2, M3 และ M6 ตามลำดับตำแหน่งบนสายโปรตีนจากน้อยไปมาก และโมทีฟสามตัวแรกนั้นมีตำแหน่งที่เหลื่อมกันอยู่โดยจะเลือกตัวที่มีคะแนนสูงสุดเป็นตัวแทนกลุ่มโดยจะได้ตัวแทนสายโปรตีนคือ P = < (M6) (M3) >



Motif	Start	End	Score
M6	9	18	42
M1	7	13	-5
M2	12	20	37
M3	26	37	28

P = < (M6) (M3) >

ภาพที่ 23 แสดงการเลือกโมทีฟแบบที่ดีที่สุด

4. SAC-RG

ในขั้นตอนนี้จะทำการหากฎความสัมพันธ์ทีละคลาส โดยจะเรียกกฎที่ได้ว่าที่ CSP (Classifiable Sequential Pattern) โดยใช้อัลกอริทึม CBS Class มาปรับปรุงเป็น SAC โดยเพิ่มขั้นตอนการหาคะแนนของลำดับเหตุการณ์ที่น่าสนใจหรือเกิดขึ้นบ่อย (FSS : Frequent Sequence Score) และ การหาคะแนนของลำดับเหตุการณ์ที่น่าสนใจที่เกิดขึ้นซ้ำๆภายในสายโปรตีน (LFSS : Local Frequent Sequence Score) สำหรับรูปแบบลำดับเหตุการณ์ที่น่าสนใจหรือเกิดขึ้นบ่อย (Frequent Sequence Pattern) หรือ CSP นั้นเอง โดยมีรหัสเทียม (psudocode) ของอัลกอริทึม SAC ดังภาพที่ 24

```
Input: Dataset D, min_sup
Output: CSP
SAC Class (DatasetD, min_sup)
{
1   foreach c_i ∈ class_set(D) do
2       D_i = class_dataset(D, c_i);
3       CSP_i = FindSP(D_i, min_sup)
4   end
}

FindSP (DatasetD, min_sup)
{
5       SP_1 = {large1 − items}
6       foreach (i = 2; SP_{i−1} ≠ ∅; i++) do
7           SP_C_i = gencandidateSP(SP_{i−1});
8           foreach (data d ∈ D) do
9            SP_s = SP_C_i ∩ subseq(d);
10           SP_s.newid;
11           foreach (sp ∈ SP_s) do
12            sp.sup++;
13            sp.localscore+ = LFSS(d, sp);
14            sp_score[sp.id]+ = sp.localscore;
15            sp_count[sp.id] ++;
16           end
17          end
18          foreach (sp.id in sp_score[ ]) do
19              sp.score =
                FSS(sp_score[sp.id], sp_count[sp.id]);
20          end
21          SP_i = {sp | sp ∈ SP_s, sp.sup ≥ min_sup}

22      end
23 return ∪_i SP_i
}
```

ภาพที่ 24 รหัสเทียมของอัลกอริทึม SAC

อัลกอริทึมนี้จะเริ่มต้นด้วย SAC Class เป็นการแบ่งชุดข้อมูลในส่วนของการสอนระบบออกเป็นกลุ่มย่อยๆตามคลาส (บรรทัดที่ 1-4) หลังจากนั้นจะเริ่มทำการหา Frequent Itemsets (FindSP) โดยกำหนดให้ $SP\_C_k = C_k$ (Candidate itemset of size k) และ $SP_i = L_i$ (Frequent Itemsets of size i) จากบรรทัดที่ 3 จะเรียก Method FindSP โดยมี Input 2 ตัวคือ Data ที่ถูกแบ่ง

ข้อมูลตามกลุ่มของ Class และค่า Min.Support ในบรรทัดที่ 5-7 จะทำการ Generate Candidate จากความยาว k-sequence เท่ากับ 1 โดยจะเก็บ Candidate ไว้ในตัวแปร SP_C_i ในบรรทัดที่ 8-9 จะนำสายของโปรตีนเข้ามาทีละเส้นเพื่อดูว่ามี Candidate ตัวไหนเป็น Subsequence ของสายโปรตีนนั้นบ้างและเก็บไว้ในตัวแปร SP_s ต่อทำการกำหนด ID ให้สมาชิกแต่ละตัว (sp)ใน SP_s (บรรทัดที่10) ในบรรทัดที่ 11-16 ทำการนับ Support ของแต่ละ sp และคำนวนค่า $sp.localscore$ โดยใช้ Method LFSS (จะอธิบายในส่วนถัดไป) นำ Score ที่ได้ไปเก็บใน Score Array และเพิ่มการนับ Support ของแต่ละ $sp.id$ ในบรรทัดที่ 18-20 ทำการคำนวนค่าของ $sp.score$ แต่ละตัวโดยใช้ Method FSS (จะอธิบายในส่วนถัดไป) ในบรรทัดที่ 21 ทำการกรองเฉพาะ sp ที่ผ่านค่า Min.support มาเก็บไว้ใน SP_i สุดท้ายในบรรทัด 23 จะนำกฎทั้งหมดของ Class นั้นๆไปเก็บไว้ที่ CSP_i

4.1 การคำนวนหาค่า LFSS

ในการทำคำนวนหาค่า $LFSS$ ของแต่ละ sp เมื่อกำหนดให้ $LFSS$ เป็นค่าที่คำนวนได้จาก 1 สายโปรตีน(1 PID) และ $sp_i \in SP_s$ เมื่อ sp เกิดจากการ Generate Candidate ที่ความยาว k-sequence $sp = M_1, M_3, ..., M_k$ และ reoccur เป็นค่าการเกิดซ้ำของ sp ใน 1 PID จะได้สูตรการคำนวนดังสมการที่ 1

$$LFSS\_sp_i = \left[\frac{\sum_{n=1}^{reoccur} ScoreM_1}{reoccur}, ..., \frac{\sum_{n=1}^{reoccur} ScoreM_k}{reoccur} \right] \quad (1)$$

โดย

$LFSS\_sp_i$ = ค่า $LFSS$ ของ Pattern sp ตัวที่ i

$reoccur$ = จำนวนการเกิดซ้ำของ Pattern sp ใน

$ScoreM_k$ = คะแนนของโมทีฟตัวที่ k

สามารถดูได้จากตัวอย่างการคำนวนต่อไปนี้กำหนดให้มี PID ทั้งหมด 3 เส้น Min.Support = 66.67% และ $sp = C_3 = M1, M3, M1$ ดังตารางที่ 2

ตารางที่ 2 แสดงสายของโปรตีนที่ถูกแทนที่ด้วยโมทีฟและจำนวนครั้งที่พบ *sp* ในสาย

PID No.	Feature of Seq. (Motif)	Global Sup	Reoccur
1	M1,M3,M3,M3,M1,M3,M1	1	10
2	M1,M7,M3,M9,M1,M3,M1	1	3
3	M1,M3,M79,M16,M4	0	0
	TOTAL	2	13

ตารางที่ 3 แสดงการเกิด Reoccur ของ *sp* ในสายโปรตีนเส้นที่ 2

	Score of Feature	4	4	7	-6	-2	6	27
PID	Round	M1	M7	M3	M9	M1	M3	M1
2	1	α		β		Ω		
2	2	α					β	Ω
2	3					α	β	Ω

จากตารางที่ 3 *sp* ที่เกิดซ้ำกันบนสาย PID 2 จะสามารถคำนวนค่า *LFSS* ของ $sp = C_3 = M1, M3, M1$ เมื่อค่า Reoccur ของ Pattern *sp* มีค่าเท่ากับ 3 ได้ดังตารางที่ 4

ตารางที่ 4 แสดงการคำนวนค่า *LFSS* ของ *sp*

Feature	M1	M3	M1
PID 2	$\frac{\sum_{n=1}^{3}\alpha}{3}$	$\frac{\sum_{n=1}^{3}\beta}{3}$	$\frac{\sum_{n=1}^{3}\Omega}{3}$

ค่า *LFSS* ของ *sp* ใน PID 2 = $[2, 6.33, 17.33]$ สามารถมองในรูปของ Array ได้ดังนี้ $sp = \begin{bmatrix} M1 \\ M3 \\ M1 \end{bmatrix} = \begin{bmatrix} 2 \\ 6.33 \\ 17.33 \end{bmatrix}$

4.2 การคำนวนหาค่า FSS

จากในขั้นตอนการคำนวน *LFSS* จะนำคะแนนมาคำนวนค่า FSS เมื่อกำหนดให้ *FSS* คือค่าที่คำนวนได้จากการนำค่า *LFSS* ของ Pattern *sp* จากทุก PID ที่พบ Pattern นี้มาคำนวนเมื่อกำหนดให้ $sp = C_3 = M1, M3, M1$ จากตารางที่ 1 จะคำนวนค่า *LFSS* ได้ทั้งหมด 3 ค่า (จาก 3 PID ในตาราง) ดังสมการที่ 2

$$\mathrm{FSS}\_sp_i = \left[\frac{\sum_{n=1}^{\sup port} LFSS\_sp_{i_n}}{\sup port} \right] \quad (2)$$

โดย

FSS_sp_i = ค่า *FSS* ของ Pattern *sp* ตัวที่ i

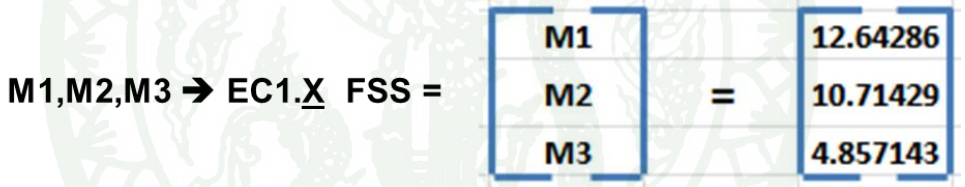
$\sup port$ = จำนวนการเกิดของ Pattern *sp* ในฐานข้อมูล

$LFSS\_sp_{i_n}$ = คะแนน *LFSS* ของ Pattern sp_i ตัวที่ n

ค่า FSS ของ sp ตัวที่ i มีค่าเท่ากับ $\mathrm{FSS}\_sp_i = \left[\frac{\sum_{n=1}^{\sup port=3} L\mathrm{FSS}\_\mathrm{sp}_{i_3}}{3} \right] = \begin{bmatrix} 3.20 \\ 4.18 \\ 5.81 \end{bmatrix}$

ตัวจำแนกประเภท

จากขั้นตอน SAC-RG จะทำการหากฎที่ละคลาส (Class Sequential Pattern Mining) ดังนั้น SAC จึงสามารถให้ความสัมพันธ์ของกฎที่มีความเฉพาะเจาะจงกับชุดของสายโปรตีนในแต่ละคลาส หลังจากเสร็จสิ้นการหากฎในขั้นตอนที่ผ่านมาจะได้กฎที่มีคะแนน FSS เก็บไว้ใน DB เพื่อจะนำไปใช้ในการคำนวนเพื่อหากฎที่เหมาะสมที่สุดในการทำนาย Unseen Data โดยมีตัวอย่างของกฎดังภาพที่ 25 จากตัวอย่างกฎนี้จะได้ว่ามีโมทีฟ 1 2 และ 3 ตามลำดับโดยจะเป็นกฎของคลาสที่ X (X = จำนวนเต็มที่เป็นคลาสใดๆ) โดยมีคะแนน FSS ของแต่ละโมทีฟอยู่ในรูปของเวกเตอร์



ภาพที่ 25 กฎที่ได้จากการหา Class Sequential Pattern Mining

เมื่อมี Unseen Data เข้ามาในระบบก่อนขั้นตอนการทำนาย Class ของข้อมูลจะต้องผ่านขั้นตอนการทำ Motif Mapping และ Motif Selection หลังจากนั้นจะเข้าสู่ขั้นตอนการเลือกกฎและการเรียงกฎ (Rules Selection and Rules Ranking) โดย SAC Classifier มีรหัสทียมดังภาพที่ 26

การเลือกกฎและการเรียงกฎ (Rules Selection and Rules Ranking)

จากข้างต้นที่ได้กล่าวไปแล้วว่าโมทีฟบางตัวสามารถเป็นตัวร่วมในหลายๆคลาสและโมทีฟบางตัวก็เป็นคุณสมบัติเฉพาะของคลาสนั้นเพียงคลาสเดียวนอกจากนี้โมทีฟ 1ตัว สามารถมีได้หลาย Regular Expression ทำให้คุณสมบัติเดียวกันมีคะแนน BLOSUM ที่ต่างกันจึงไม่สามารถใช้การคำนวณค่าความเชื่อมั่น (Confidence) ได้เมื่อต้องตัดสินใจในการเลือกกฎที่มีลักษณะดังตัวอย่าง

ต่อไปนี้ ให้ P เป็นสายโปรตีนเส้นหนึ่งในชุดทดสอบที่ผ่านขั้นตอนการเลือกโมทีฟแล้วได้เป็น P = < (M6) (M3) > และ C เป็นคลาสของโปรตีนเอมไซม์เมื่อมีการเลือกกฎออกมาเพื่อทำนายปรากฏว่าได้สองกฎที่มความสัมพันธ์กับ P คือ

$$(M6)(M3) \rightarrow C1$$
$$(M6)(M3) \rightarrow C2$$

จากตัวอย่างนี้ทำให้ไม่สามารถใช้ค่าความเชื่อมั่นในการเลือกกฎที่เหมาะสมได้ ในงานวิจัยนี้จึงได้เสนอวิธีการเรียงกฎแบบใหม่โดยใช้ การวัดระยะแบบยูคลิด (Euclidean distance) และ การวัดระยะแบบแมนฮัตตัน (Manhattan distance) เข้ามาช่วยในการแก้ปัญหาดังกล่าวในการเลือกกฎที่เหมาะสมที่สุดเมื่อกำหนดให้ csp เป็นกฎความสัมพันธ์กฎหนึ่งที่เก็บอยู่ในฐานข้อมูล (CSP) $csp \in CSP$ โดยมีสูตรการคำนวณดังนี้

1.1 การคำนวน Euclidean Distance

$$d(csp, input) = \sqrt{\sum_{i=1}^{n} \left(F_{i_{csp}} - F_{i_{input}} \right)^2} \qquad (3)$$

โดย

csp = Pattern ที่เป็น subsequence ของ input

$F_{i_{csp}}$ = Feature หรือ Motif ตัวที่ i ของ Pattern csp

$F_{i_{input}}$ = Feature หรือ Motif ตัวที่ i ของ Pattern input

1.2 การคำนวน Manhattan Distance

$$d(csp, input) = \sum_{i=1}^{n} \left| F_{i_{csp}} - F_{i_{input}} \right| \qquad (4)$$

โดย

csp = Pattern ที่เป็น subsequence ของ input

$F_{i_{csp}}$ = Feature หรือ Motif ตัวที่ i ของ Pattern csp

$F_{i_{input}}$ = Feature หรือ Motif ตัวที่ i ของ Pattern input

```
Input: sequence X
Output: csp_i
Class Sequence sequence(X)
{
1   foreach (csp_i ∈ CSP) do
2       if csp_i ∈ subseq(X) then
3           csp_i.distance =
            distance_function(csp_i, seqx);
4           tmp.add(csp_i);
5       end
6   end
7   if tmp.count = 0 then
8       return default class;
9   else
10      tmp =
        orderby(tmp, length_desc, distance_asc);
11  end
12  if max length.count(tmp) = 1 then
13      return csp_i;
14  else if lowsid tance.count(tmp) > 1 then
15      return vote(class);
16  else
17      return csp_i(low_distance);
18  end
}
```

ภาพที่ 26 รหัสเทียมของ SAC Classifier

SAC Classifier จะเริ่มต้นโดยตรวจดูว่ากฎไหนบ้างที่เป็น Subsequence ของ Sequence X เมื่อมี csp_i ใดที่เป็น Subsequence ก็จะคำนวนค่าระยะทาง (Distance Function) และเก็บค่าเข้าตัว

แปร tmp (ขั้นตอนที่ 1-5) จากนั้นจะเช็คดูว่าถ้าใน *tmp* ไม่มีกฎใดเลยที่เป็น Subsequence ของ X จะตอบ Default Class (ขั้นตอนที่ 7-8) แต่ถ้าไม่เท่ากับ 0 จะทำการ Sort ข้อมูลใน tmp ตามความยาวกฎจากมากไปน้อยและค่า Distance จากน้อยไปมาก (ขั้นตอนที่ 9-11) หลังจากนั้นจะเช็คว่ากฎที่ยาวที่สุดมีกฎเดียวหรือเปล่าถ้าใช้จะตอบกฎนั้นพร้อม Class ของกฎ (ขั้นตอนที่ 12-13) แต่ถ้ามีมากกว่า 1 กฎที่มีความยาวกฎเท่ากันจะพิจารณาคะแนนของ Distance ถ้าทั้งความยาวกฎและค่า Distance เท่ากันจะใช้หลักการตัดสินใจเสียงข้างมาก (Majority vote) ในการเลือก Class จากกฎที่เป็น Subsequence ทั้งหมดของ X (ขั้นตอนที่ 14-15) แต่ถ้าไม่ตอบกฎที่มีค่า Distance น้อยที่สุดจากกฎทั้งหมดที่มีความยาวเท่ากัน (ขั้นตอนที่ 17-18)

ผลและวิจารณ์

ผล

1. วิธีวัดผลการทดลอง

ในหัวข้อนี้จะกล่าวถึงผลการวิจัยของงานวิจัยนี้เปรียบเทียบกับงานวิจัย Nakashima (1986) Chou and Elrod (2003) Kuo-Chen Chou (2005) และงานวิจัย Tseng and Lee (2005) โดยในงานวิจัยของ Chou and Elrod (2003) และ Kuo-Chen Chou (2005) ใช้ความน่าจะเป็นของการเกิดกรดอะมิโนที่อยู่ติดกัน หรือเรียกว่าอะมิโนแอซิดคอมโพสิชัน (amino acid composition) เป็นตัวแทนสายโปรตีนหรือฟีเจอร์ แต่ต่างกันสำหรับงานวิจัย Chu (2005) นั้น ไม่ได้ใช้ข้อมูลอะมิโนแอซิดคอมโพสิชันโดยตรง แต่จะใช้ซูโดอะมิโนแอซิด คอมโพสิชัน (pseudo amino acid composition) ซึ่งเป็นการพิจารณคู่กรดอะมิโนโดยที่มีระยะห่างหรือแก๊ปเท่ากับ 1, 2 หรือ 3 เป็นต้น และมีการคำนวณค่าไฮโดรโฟบิกซิตี้และไฮโดรฟิลิกซิตี้ระหว่างกรดอะมิโนร่วมด้วย จากนั้นนำฟีเจอร์ที่ได้ไปสร้างตารางตัวแทนสายข้อมูลโปรตีนทั้งเส้น เพื่อเป็นข้อมูลนำเข้าอัลกอริทึมวิธีการเรียนรู้ และทำการสร้างแบบจำลองเพื่อใช้ในการจำแนกเอนไซม์ซับแฟมิลี่คลาส ส่วนของ Tseng and Lee (2005) จะใช้วิธีการรวมเทคนิคการหารูปแบบลำดับเหตุการณ์ที่น่าสนใจ (Sequential Pattern Mining) เข้ากับการคำนวณเชิงสถิติเพื่อหาความเป็นไปได้ (Probabilistic Induction) ในการจำแนกเอนไซม์ซับแฟมิลี่คลาส

ข้อมูลที่ใช้ในงานวิจัยนี้ใช้ข้อมูลชุดเดียวกันกับงานวิจัย Chu and Elrod (2003) โดยแบ่งชุดข้อมูลออกเป็น 2 กลุ่มคือ (1) ข้อมูลชุดที่ 1 เป็นชุดข้อมูลสอนระบบ จำนวน 2640 โปรตีน โดยเป็นข้อมูลที่ใช้สกัดฟีเจอร์และสร้างแบบจำลอง (2) ข้อมูลชุดที่ 2 (unknown enzymes) เป็นชุดข้อมูลที่ไม่มีความสัมพันธ์กับข้อมูลชุดแรก ใช้เพื่อทดสอบแบบจำลองที่สร้างได้จากข้อมูลชุดที่ 1 เท่านั้น มีจำนวน 2124 โปรตีน

2. ผลการทดลอง

1.1 สำหรับผลการทดลองของวิธีวิจัยในงานวิจัยนี้ในส่วนแรกจำทำการพิสูจน์ให้เห็นว่าการพิจารณาการหาคะแนนของลำดับเหตุการณ์ที่น่าสนใจที่เกิดขึ้นซ้ำๆภายในสายโปรตีน (LFSS :

1.2 Local Frequent Sequence Score) มีผลต่อความแม่นยำในการจำแนกประเภทเอนไซม์ซับแฟมิลี่หรือไม่ จากสมการที่ 2 เมื่อเราไม่ต้องการพิจารณาค่า LFSS ที่เป็นการคำนวนค่าการเกิดซ้ำของ sp ใน 1 PID จะเขียนสมาการใหม่ได้ดังสมการที่ 5 โดยในผลการทดลองแบบไม่พิจารณา LFSS นั้นจะทำให้มีสูตรคำนวนแค่สมการเดียวดังสมการที่ 5 โดยจะแบ่งผลการทดลองออกเป็นแบบ Reoccurrence และ No Reoccurrence ของ sp โดยแต่ละแบบจะมี 2 แบบย่อยคือ การวัดระยะแบบยูคลิด (Euclidean distance) และ การวัดระยะแบบแมนฮัตตัน (Manhattan distance) ดังภาพที่ 27

$$FSS\_sp_i = \left[\frac{\sum_{n=1}^{\sup port} sp_{i_n}}{\sup port} \right] \quad (5)$$

โดย

FSS_sp_i = ค่า FSS ของ Pattern sp ตัวที่ i

$\sup port$ = จำนวนการเกิดของ Pattern sp ในฐานข้อมูล

sp_{i_n} = คะแนนของ Pattern sp_i ตัวที่ n ในฐานข้อมูล

Subfamily class	Number of samplesb	SAC (%)			
		No Reoccurrence		Reoccurrence	
		Euclidean	Manhattan	Euclidean	Manhattan
1	626	69.01	68.21	71.88	71.57
2	216	67.13	64.81	71.30	70.83
3	25	52.00	48.00	52.00	48.00
4	17	76.47	47.06	76.47	64.71
5	14	50.00	71.43	78.57	71.43
6	608	59.21	58.72	64.97	58.72
7	7	57.14	42.86	57.14	42.86
8	6	100.00	83.33	100.00	83.33
9	253	79.45	78.26	81.82	80.63
10	12	75.00	66.67	75.00	66.67
11	20	90.00	85.00	90.00	85.00
13	12	66.67	66.67	66.67	66.67
14	257	69.26	60.31	69.26	65.37
15	20	95.00	85.00	95.00	85.00
17	11	63.64	54.55	63.64	54.55
18	20	75.00	65.00	75.00	65.00
Overall	2124	1435/2124 = 67.56%	1384/2124 = 65.16%	1507/2124 = 70.95%	1440/2124 = 67.80%

ภาพที่ 27 แสดงความแม่นยำ (แยกตามคลาส) แบบ Reoccurrence และ No Reoccurrence

1.3 สำหรับผลการทดลองในส่วนที่ 2 จะแสดงผลการทดลองแบบพิจารณา Reoccurrence ของ *sp* โดยแยกเป็น 2 แบบคือการวัดระยะแบบยูคลิด (Euclidean distance) และ การวัดระยะแบบแมนฮัตตัน (Manhattan distance) โดยเปรียบเทียบกับงานวิจัยอื่นๆด้วยการวัดค่าความแม่นยำ (Accuracy) ของระบบจำแนกประเภทเอนไซม์ซับแฟมิลี่ โดยมีผลการวิจัยโดยละเอียดดังนี้

Subfamily class	Number of samplesb	CBS (Tseng and Lee, 2005) (%)	SAC (%)	
			Euclidean	Manhattan
1	626	22.68	71.88	71.57
2	216	28.70	71.30	70.83
3	25	20.00	52.00	48.00
4	17	5.88	76.47	64.71
5	14	35.71	78.57	71.43
6	608	25.49	64.97	58.72
7	7	0.00	57.14	42.86
8	6	0.00	100.00	83.33
9	253	62.06	81.82	80.63
10	12	8.33	75.00	66.67
11	20	10.00	90.00	85.00
13	12	8.33	66.67	66.67
14	257	11.67	69.26	65.37
15	20	90.00	95.00	85.00
17	11	0.00	63.64	54.55
18	20	35.00	75.00	65.00
Overall	2124	586/2124=27.59%	1507/2124 = 70.95%	1440/2124 = 67.80%

ภาพที่ 28 แสดงความแม่นยำ (แยกตามคลาส) ของระบบจำแนกเอนไซม์ซับแฟมิลี สำหรับข้อมูลชุดที่เมื่อเปรียบเทียบกับ CBS

Subfamily class	Number of samplesb	Least Euclidean predictor (Nakashima et al., 1986) (%)	Covariant- discriminant predictor (Chou and Elrod, 2003) (%)	Am-Pse-AA composition and Covariant- discriminant predictor (Kuo-Chen Chou,2005) (%)	SAC (%)	
					Euclidean	Manhattan
1	626	26.68	49.68	73	71.88	71.57
2	216	47.22	57.41	70.37	71.30	70.83
3	25	36	48	56	52.00	48.00
4	17	17.65	52.94	70.59	76.47	64.71
5	14	7.14	50	50	78.57	71.43
6	608	71.38	72.37	77.3	64.97	58.72
7	7	28.57	57.14	42.86	57.14	42.86
8	6	33.33	50	50	100.00	83.33
9	253	73.91	86.17	84.58	81.82	80.63
10	12	41.67	58.33	83.33	75.00	66.67
11	20	50	75	90	90.00	85.00
13	12	25	75	66.67	66.67	66.67
14	257	68.87	77.82	84.05	69.26	65.37
15	20	70	90	95	95.00	85.00
17	11	72.73	81.82	63.64	63.64	54.55
18	20	50	60	80	75.00	65.00
Overall	2124	1134/2124 = 53.39%	1398/2124 = 65.82%	1626/2124 = 76.55%	1507/2124 = 70.95%	1440/2124 = 67.80%

ภาพที่ 29 แสดงความแม่นยำ (แยกตามคลาส) ของระบบจำแนกเอนไซม์ซับแฟมิลี สำหรับข้อมูลชุดที่ 2 (unknown enzymes) จำนวน 2124 เอนไซม์ (independent test)

วิจารณ์

จากผลการทดลองที่ได้นั้นในด้านประสิทธิภาพความถูกต้องแม่นยำของระบบจำแนกประเภทเอนไซม์ซับแฟมิลีนั้น สามารถวิเคราะห์และวิจารณ์ได้ดังนี้

จากภาพที่ 27 จะเห็นได้ว่าสมมุติฐานในการพิจารณาการเกิดซ้ำๆของ *sp* มีผลต่อความแม่นยำในการการจำแนกเอนไซม์ซับแฟมิลีของคลาสนั้นๆโดยเฉพาะในขั้นตอนการทำ Best Motif Selection เมื่อพบ *sp* ที่เกิดซ้ำเป็นจำนวนมากในแต่ละคลาสการคำนวน LFSS จะช่วยให้การจำแนกเอนไซม์ซับแฟมิลีนั้นทำได้ดีขึ้น

จากภาพที่ 28 จะเห็นได้ว่า CBS ที่ใช้การคำนวณเชิงสถิติเพื่อหาความเป็นไปได้ (Probabilistic Induction) ในการจำแนกเอนไซม์ซับแฟมิลีนั้นให้ความแม่นยำที่ 27.59% ในขณะที่ SAC ที่ใช้ความรู้พื้นฐานของตาราง BLOSUM62 ในการคำนวนคะแนนของโมทีฟเพื่อนำมาคำนวนค่าระยะทางระหว่างโมทีฟของข้อมูลทดลองกับกฎหรือรูปแบบลำดับเหตุการณ์ย่อยที่น่าสนใจ (Frequent SubSequences) ในการจำแนกเอนไซม์ซับแฟมิลีนั้นให้ความแม่นยำที่ 70.95% ผลจากการทดลองของ CBS นั้นสามารถบ่งชี้ได้ว่าการใช้วิธีการทางสถิติไม่สามารถรองรับคุณสมบัติเฉพาะตัวของโมทีฟ ที่มีลักษณะที่โมทีฟตัวหนึ่งๆเป็นตัวร่วมของหลายๆซับแฟมิลีได้จากการวิเคราะห์ข้อมูลดังภาพที่ 30 โดยกำหนดให้สายโปรตีนทดสอบเท่ากับ P และอยู่ในรูปของลำดับโมทีฟ $P = M3, M3, M130, M6, M11$ และมีผลเฉลยเป็นคลาสที่ 1

Class	Motif	Class	RuleLength
C1	M3,M6	1	2
C1	M6,M11	1	2
C3	M3,M3	3	2
C1	M3,M11	1	2
C1	M3,M3	1	2
C3	M,M11	3	2
C3	M6	3	1
C3	M11	3	1
C1	M6	1	1
C3	M3	3	1
C1	M11	1	1
C1	M3	1	1

Frequent Subsequence

Class 1	Total Rules	86
	Total Frequent Subsequence	7
	Sum Rule length	143
Class3	Total Rules	63
	Total Frequent Subsequence	5
	Sum Rule length	84

	C1	C2	C3	Cn
ScoreArray	0.0769	-	0.0833	-

ภาพที่ 30 แสดงปัญหาของ CBS ที่เกิดจากหารใช้คะแนนจากความยาวกฎ

จากภาพที่ 30 เมื่อข้อมูลทดสอบเข้ามาในระบบ CBS จะทำการหากฎที่เป็นรูปแบบย่อยของข้อมูลทดสอบ โดยในขั้นตอนของจะเห็นว่า CBS CLASS Classifier จะคำนวนหาผลรวมของความยาวกฎในแต่ละคลาสโดยคลาสที่ 1 มีค่าเท่ากับ 143 และคลาสที่ 3 มีค่าเท่ากับ 84 และมีกฎทั้งหมดในคลาสที่ 1 เท่ากับ 86 และมีกฎที่เป็น SubSequences ของข้อมูลทดสอบเท่ากับ 7 กฎในขณะที่คลาสที่ 3 มีกฎทั้งหมด 63 กฎและมีกฎที่เป็น SubSequences เท่ากับ 5 กฎ เมื่อมีโมทีฟ M1 ที่เป็นตัวรวมที่เกิดในเกือบทุกคลาสยกเว้นคลาสที่ 3 และมีค่า Support สูงมากผ่านที่จะผ่านค่า Minimum Support ได้ในแต่ละคลาสจะทำให้จำนวนกฎและและผลรวมความยาวกฎที่ใช้เป็นน้ำหนักในการหารความยาวกฎที่เป็น SubSequences เพิ่มขึ้นทำให้ได้ค่าคะแนนต่อกฎของคลาสที่ 1น้อยกว่ากฎของคลาสที่ 3 ถึงแม้ปริมาณกฎที่เป็น SubSequences ของคลาสที่ 1 จะมีมากกว่า แต่ก็ไม่ชนะคลาสที่ 3 ที่มีน้ำหนักในการหารความยาวกฎที่ต่ำกว่าทำให้ทำนายผิดคลาส จากการวิเคราะห์ฐานข้อมูลพอโมทีฟที่เป็นตัวรวมเป็นจำนวนมากซึ่งส่งผลกับการทำนายของ CBS

โมทีฟที่เป็นตัวร่วมระหว่างคลาสนั้นจะทำให้กฎมีความยืดหยุ่นมากกว่าอีกทั้งในขั้นตอนการสร้างกฎจะทำให้กฎมีความยาวมากขึ้นกว่าปกติ ถ้าตัดโมทีฟที่เป็นตัวร่วมระหว่างคลาสออกจะทำให้ได้กฎที่มีความยาวสั้นเกินไปที่จะใช้จำแนกเอนไซม์ซับแฟมิลี ในแง่ของสายโปรตีนนั้นกฎที่มีความยาวควรจะเป็นกฎที่เฉพาะเจาะจงและเหมาะสมกว่าในการแสดงถึงคุณสมบัติของโมทีฟในสายโปรตีนนั้นๆมากกว่ากฎที่มีความยาวสั้นเพราะโมทีฟเพียงตัวเดียวไม่สามารถบ่งบอกถึงคุณสมบัติในการทำงานของมันได้ต้องใช้โมทีฟหลายๆตัวประกอบกันในการทำงานหนึ่งๆ

จากภาพที่ 29 จะเห็นได้อัลกอริธึม SAC ได้ผลความถูกต้องในการทำนายน้อยกว่าอัลกอริธึม Am-Pse-AA อยู่ 5.6% ในแบบ Euclidean Distance และ 8.75% ในแบบ Manhattan Distance แต่ถ้าดูในรายละเอียดของแต่ละคลาสแล้วการใช้วิธี Euclidean Distance จะให้ผลดีกว่าในการจำแนกคลาสบางคลาส โดยจำแนกได้ดีกว่า 5 คลาส, เสมอ 4 คลาส, แพ้ 7 คลาส จากทั้งหมด 16 คลาส โดยปัญหาความถูกต้องในการทำนายที่พบในบ้างคลาสนั้นเกิดจากการขาดแคลนข้อมูลที่มีความสัมพันธ์โดยตรงกับการเกิดฟังก์ชั่นเอนไซม์ในบริเวณจับและบริเวณเร่งในการนำมาสร้างรีแอคทีฟโมทีฟ (reactive motif) ซึ่ง มีเพียงประมาณ 3.34% เท่านั้นจากจำนวนเอนไซม์ทั้งหมดในการนำมาสร้างรีแอคทีฟโมทีฟ ทำให้ไม่สามารถหาตัวแทนของโมทีฟได้เพียงพอ

สรุปและข้อเสนอแนะ

สรุป

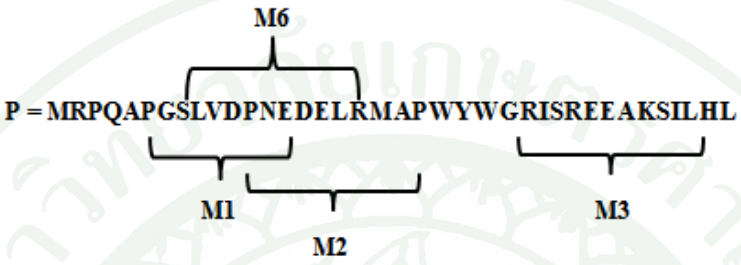
งานวิจัยนี้จึงได้เสนอเทคนิคการจำแนกประเภทข้อมูลโดยใช้กฎความสัมพันธ์แบบมีลำดับในการจำแนกซับแฟมมิลี่ของเอมไซม์(SAC : Sequential Associative Classification for Enzyme Subfamily Prediction) โดยรวมเทคนิคการหากฎความสัมพันธ์แบบมีลำดับเหตุการณ์ (Sequential Pattern Mining) และเทคนิคการจัดกลุ่มข้อมูลโดยใช้กฎความสัมพันธ์ (Associative Classification) เข้าด้วยกันเพื่อการทำนาย โดยใช้ความรู้พื้นฐานเกี่ยวกับ “โมทีฟเชิงปฏิกิริยาชีวเคมี” มารวมเข้าด้วยกันกับขั้นตอนการหากฎความสัมพันธ์แบบมีลำดับเหตุการณ์ นอกจากนี้ในการจัดกลุ่มข้อมูลโดยใช้กฎความสัมพันธ์ได้ถูกขยายขอบเขตออกไปโดยการพิจารณาลำดับความสัมพันธ์ของแอททิบิวต์ในขั้นตอนการสร้างโมเดลในการทำนาย จากผลการทดลองนอกจากจะได้โมเดลจำแนกประเภทที่ง่ายต่อการตีความแล้วในด้านชีวสารสนเทศนอกจากจะต้องการจำแนกเอนไซม์ซับแฟมิลียังรวมไปถึงการทำความเข้าใจการทำงานเอนไซม์ได้ดีขึ้น ว่าส่วนใดของโปรตีนที่ก่อให้เกิดกลไกการทำงานของ Enzyme Function โดยเทคนิค SAC นี้สามารถสรุปให้เห็นได้ถึงความสัมพันธ์ระหว่างลำดับแอททิบิวต์ในสายโปรตีนและยังให้ผลความแม่นยำในการทำนายที่อยู่ในเกณฑ์ที่สามารถนำไปใช้ได้จริงโดยมีความถูกต้องสูงสุดอยู่ที่ 70.95%

อย่างไรก็ดีปัญหาความถูกต้องในการทำนายที่พบในบ้างคลาสนั้นเกิดจากการขาดแคลนข้อมูลที่มีความสัมพันธ์โดยตรงกับการเกิดฟังก์ชั่นเอนไซม์ในบริเวณจับและบริเวณเร่งในการนำมาสร้างรีแอคทีฟโมทีฟ (reactive motif) ซึ่ง มีเพียงประมาณ 3.34% เท่านั้นจากจำนวนเอนไซม์ทั้งหมดทำให้ไม่สามารถหาตัวแทนของโมทีฟได้เพียงพอในบ้างคลาส ในอนาคตเมื่อมีการวิจัยข้อมูลในส่วนดังกล่าวเพิ่มมากขึ้นเทคนิค SAC ก็จะมีความถูกต้องที่สูงขึ้นตามไปด้วย

ข้อเสนอแนะ

งานวิจัยนี้ยังสามารถทำการปรับปรุงเพื่อเพิ่มประสิทธิภาพได้ในส่วนของการบูตค่า Support ของกฎได้โดยใช้แนวคิดการรวมกลุ่มโมทีฟ (Group of Overlap Motif) ในขั้นตอนการเลือก โมทีฟดังภาพที่ 24 วิธีนี้จะต่างจากแบบแรกตรงที่จะนำกลุ่มของโมทีฟที่

ซ้อนทับกันมารวมเข้าเป็นกลุ่มเดียวกันจากภาพที่ 25 ผลลัพธ์ที่ได้จากการ Query นั้นพบว่าในสายโปรตีน P พบโมทีฟ M1, M2, M3 และ M6 ตามลำดับ ตำแหน่งบนสายโปรตีนเรียงจากน้อยไปมาก และโมทีฟสามตัวแรกนั้นมีตำแหน่งที่เหลื่อมกันอยู่ โดยจะได้ตัวแทนสายโปรตีนคือ P = < (M6 | M1 | M2) (M3) >



Motif	Start	End	Score
M6	9	18	42
M1	7	13	-5
M2	12	20	37
M3	26	37	28

P = < (M6 | M1 | M2) (M3) >

ภาพที่ 31 แสดงการเลือกโมทีฟแบบกลุ่ม

โดยปกตินั้นการหากฎความสัมพันธ์แบบมีลำดับเหตุการณ์ (Sequential Pattern Mining) โอกาสที่จะเกิดกฎที่มีความยาวมากๆนั้นค้อนข้างต่ำทำให้ต้องใช้ค่า Support ที่ค้อนข้างต่ำเพื่อให้ได้ปริมาณกฎที่เพียงพอต่อการใช้งานและมีคุณภาพที่เหมาะสมต่อการเป็นตัวแทนกฎของคลาสนั้นๆ จากแนวคิดนี้ทำให้รูปแบบของ 1 สายโปรตีนสามารถนับ Support ของ Candidate ได้เพิ่มขึ้นทำให้กฎที่ไม่ผ่านค่า Support ขั้นต่ำมีโอกาสเกิดและทำให้กฎยาวขึ้น

เอกสารและสิ่งอ้างอิง

พีระ ลิ่วลม. 2551. **การพัฒนาตัวแทนสายโปรตีนสำหรับการทำนายประเภทฟังก์ชั่นเอนไซม์.** วิทยานิพนธ์ปริญญาเอก, มหาวิทยาลัยเกษตรศาสตร์.

Agrawal, R., T. Imieli ski, and A. Swami. 1993. Mining association rules between sets of items in large databases. **ACM SIGMOD Record** 22(2): 207-216.

Agrawal, R. and R. Srikant. 1994. **Fast algorithms for mining association rules**, Citeseer.

Agrawal, R. and R. Srikant. 1995. **Mining sequential patterns**.

Altschul, S., T. Madden, and J. Zhang. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Research** 25(17): 3389.

Bailey, T. and C. Elkan. 1994. **Fitting a mixture model by expectation maximization to discover motifs in biopolymers**, Citeseer.

Bairoch, A. and R. Apweiler. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. **Nucleic Acids Research** 28(1): 45.

Bork, P. and E. Koonin. 1996. Protein sequence motifs. **Current Opinion in Structural Biology** 6(3): 366-376.

Burges, C. 1998. A tutorial on support vector machines for pattern recognition. **Data mining and knowledge discovery** 2(2): 121-167.

Chou, K. 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. **Bioinformatics** 21(1):

Chou, K. and D. Elrod. 2003. Prediction of enzyme family classes. **Journal of Proteome Research** 2(2): 183-190.

Chou, K. and H. Shen. 2008. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. **Nature protocols** 3(2): 153-162.

Eddy, S. 2004. Where did the BLOSUM62 alignment score matrix come from? **Nature Biotechnology** 22(8): 1035-1036.

Emanuelsson, O., S. Brunak, and H. Nielsen. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. **Nature protocols** 2(4): 953-971.

Exarchos, T., M. Tsipouras, and C. Papaloukas. 2008. A two-stage methodology for sequence classification based on sequential pattern mining and optimization. **Data & Knowledge Engineering** 66(3): 467-487.

Han, X. 2003. **CPAR: Classification based on predictive association rules**, Society for Industrial & Applied.

Henikoff, S. and J. Henikoff. 1992. Amino acid substitution matrices from protein blocks. **Proceedings of the National Academy of Sciences of the United States of America** 89(22): 10915.

Hopp, T. and K. Woods. 1981. Prediction of protein antigenic determinants from amino acid sequences. **Proceedings of the National Academy of Sciences of the United States of America** 78(6): 3824.

Jaillet, S., A. Laurent, and M. Teisseire. 2006. Sequential patterns for text categorization. **Intelligent Data Analysis** 10(3): 199-214.

Kudenko, D. and H. Hirsh. 1998. **Feature generation for sequence categorization**, JOHN WILEY & SONS LTD.

V. Tseng and C. Lee. 2005. CBS: a new classification method by using sequential patterns. **Proceedings of the SIAM International Data Mining Conference**, California, USA,

Lent, B., R. Agrawal, and R. Srikant. 1997. **Discovering trends in text databases**.

Lesh, N., M. Zaki, and M. Ogihara. 1999. **Mining features for sequence classification**, ACM.

Liewlom, P., T. Rakthanmanon, and K. Waiyamai. 2007. Prediction of Enzyme Class by Using Reactive Motifs Generated from Binding and Catalytic Sites. **Advanced Data Mining and Applications** : 442-453.

Liu, B., W. Hsu, and Y. Ma. 1998. Integrating classification and association rule mining. **Knowledge Discovery and Data Mining** : 80–86.

Nakashima, H., K. Nishikawa, and T. Ooi. 1986. The folding type of a protein is relevant to the amino acid composition. **Journal of biochemistry** 99(1): 153.

Pearson, W. and D. Lipman. 1988. Improved tools for biological sequence comparison. **Proceedings of the National Academy of Sciences** 85(8): 2444.

Pei, W. **CMAR: Accurate and efficient classification based on multiple class-association rules**.

Rak, R., W. Stach, and M. Antonie. 2005. Considering re-occurring features in associative classifiers. **Advances in Knowledge Discovery and Data Mining** : 240-248.

Sun, Y., M. Kamel, and Y. Wang. 2007. Cost-sensitive boosting for classification of imbalanced data. **Pattern Recognition** 40(12): 3358-3378.

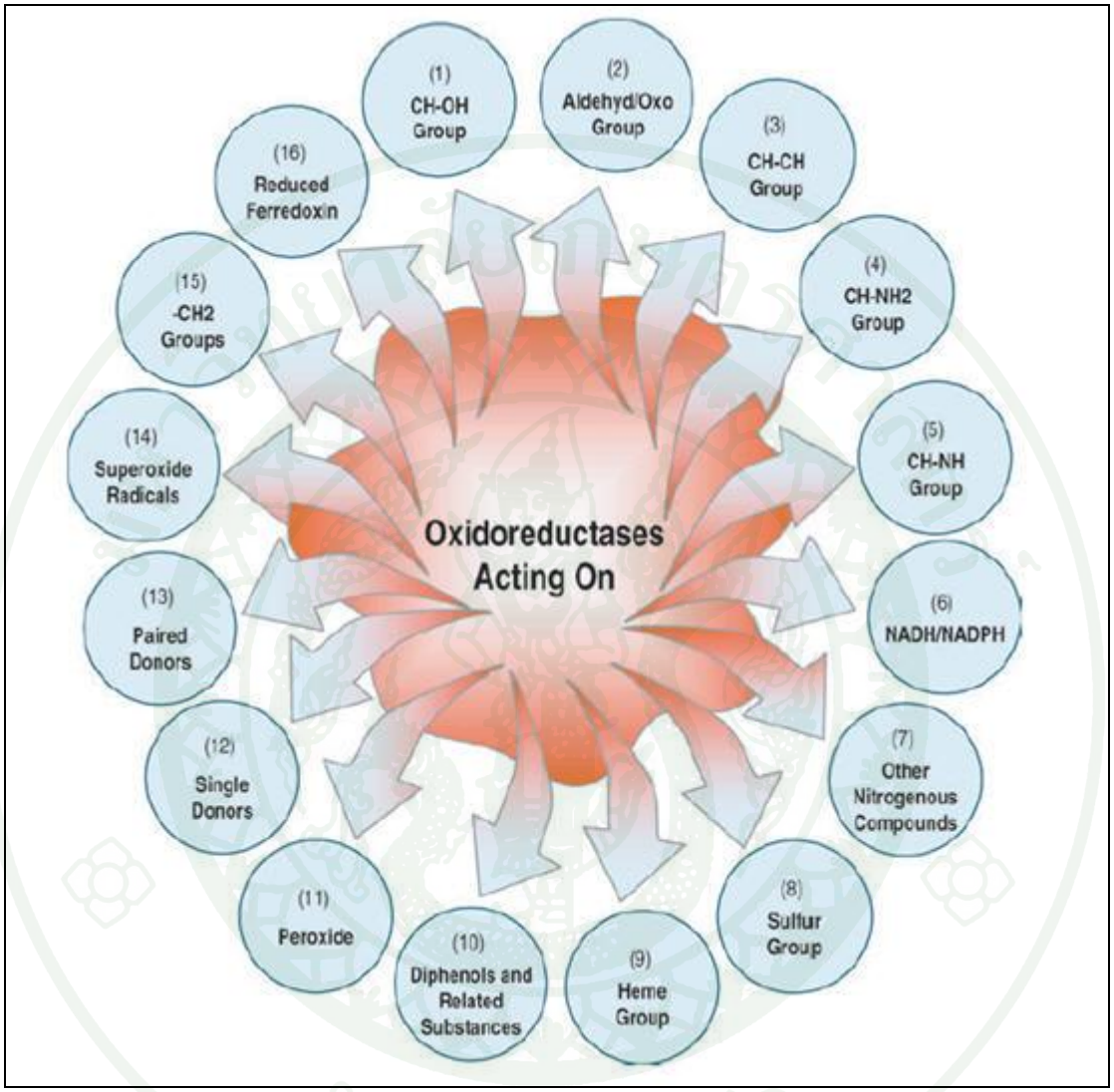
Tanford, C. 1962. Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. **Journal of the American Chemical Society** 84(22): 4240-4247.

Waiyamai, K., P. Liewlom, T. Kangkachait, and T. Rakthanmanon. 2008. **Concept lattice-based mutation control for reactive motifs discovery, Springer-Verlag**.

Yan, X., J. Han, and R. Afshar. 2003. **CloSpan: Mining closed sequential patterns in large datasets**.

ภาคผนวก

ข้อมูลเอนไซม์และข้อมูลซับแฟมิลีคลาสที่ใช้ในงานวิจัยนี้



ภาพผนวกที่ 1 ไดอะแกรมแสดงซับแฟมิลีคลาสของเอนไซม์อ็อกซิโดรีดักเทส (Oxidoreductases) ที่ใช้ในงานวิจัยนี้ จำนวนทั้งสิ้น 16 คลาส

ที่มา: Chou (2005)

ตารางผนวกที่ 1 แสดงข้อมูลเอนไซม์ (Assession number) ที่ใช้ในงานวิจัยนี้

Class 1: 314 Sequences; Average Sequence Identity: 13.16%									
O00097	O45687	P00325	P00328	P00331	P00334	P06525	P07158	P07161	P07246
P08319	P09370	P10847	P12854	P14139	P14674	P17648	P20306	P21518	P22246
P23237	P23361	P25139	P25406	P25988	P27581	P28469	P33010	P38113	P40394
P41682	P42328	P48585	P48814	P49383	P49645	P51550	P54202	P80338	Q00669
Q00672	Q05114	Q09009	Q17334	Q64413	O07399	O85141	P33207	P50941	P70720
P73826	O34268	P50169	Q27979	P50842	P11759	P29781	P55463	P27867	Q00796
Q59787	P15428	P16232	P50172	P51975	Q29608	P35270	P45856	P77851	O51544
O84838	P37417	P95837	P19337	O28578	O67619	O65992	P42957	Q45421	O75828
P47844	Q29529	O26337	P95872	O24562	P30360	P31657	P42734	Q02971	Q40976
O57380	P25984	P50578	P28475	P08793	P91711	O50316	P06981	P20839	P24547
P39567	P47996	P50095	P50098	Q07152	Q50715	P07943	P21300	P45377	P14720
P51103	P51106	P51110	O05973	O60701	O86422	P76373	Q07172	Q57346	O26327
O34651	P10370	P28736	Q02136	P50163	P25415	P49365	P15770	P46240	Q44607
P50166	Q12634	O33734	P00337	P00340	P00343	P04034	P06151	P10655	P13715
P16115	P19858	P20619	P29038	P42119	P42123	P50934	P78007	Q07251	Q27888
Q60009	Q95028	P26298	P51011	P13443	P08499	P31116	P46806	P56429	P29147
P29266	O26662	P04035	P12684	P16237	P29057	P48020	P51639	Q01559	Q12577
Q58116	O08756	P34439	Q61425	P23238	P50204	O08349	O59028	P06994	P11708
P17783	P25077	P37228	P44427	P48364	P80038	Q04820	Q42972	Q59202	P26616
P45868	O30807	P37225	P12628	P22178	P36444	P43279	P51615	O43837	P28834
P50213	Q93353	O14254	O75874	P16100	P39126	P41562	P50214	P50217	P54071
P80046	Q58991	O13287	P00349	P14062	P37754	P41570	P41576	P52208	P70718
P96789	P12310	P39483	P40288	O00091	O24357	O83491	P11410	P11413	P22992
P37986	P44311	P48848	P54996	P77809	Q27464	Q43727	Q04520	P19871	P39160
P32816	P38945	P14061	P51656	P51659	P70385	Q62904	P50199	O57656	P08507
P21695	P37606	P52425	Q00055	Q27567	Q44472	P21528	P52426	O27441	O59930
O94114	P05644	P08791	P18869	P24098	P29696	P34733	P41019	P43860	P54354
P87186	P94631	P96197	Q02143	Q56268	O27491	O33114	O67289	P05989	P37253
P78827	Q02138	Q57179	Q59818	P39849	O34296	P76251	O67555	O08651	O33116
P08328	P40510	P87228	P09437	P32891	Q00922	P47195	P81156	P22637	Q01745
P54223	O05807	P33940	P13650	Q07982	P13032	P18158	P43304	P52111	P90795
O05542	P15279	P28036	Q44002						
Class 2: 216 Sequences; Average Sequence Identity: 17.27%									
P45382	P78870	P77580	O26890	O30706	O67716	P10539	P23247	P30903	P41399
P41404	P47730	P97049	Q51344	Q55512	Q56734	Q59291	O06822	O13507	O27090
O34425	O43026	O59494	O67161	P00354	P00356	P00358	P00360	P00362	P04796
P04970	P07486	P08439	P09124	P09316	P10097	P12858	P12860	P16858	P17329

ตารางผนวกที่ 1 (ต่อ)

Class 2: 216 Sequences; Average Sequence Identity: 17.27%									
P17331	P17729	P17819	P19089	P19315	P20286	P20445	P22513	P25856	P25858
P26517	P26519	P26521	P27726	P29272	P30724	P32636	P32638	P32810	P34783
P34917	P34919	P34921	P34923	P35143	P44304	P46713	P47543	P49433	P50321
P50362	P51009	P52987	P53430	P54118	P54270	P56649	P78958	P87197	Q00584
Q01077	Q01597	Q01982	Q07234	Q12552	Q27890	Q41595	Q42977	Q46450	Q58546
Q59800	Q64467	Q92243	O32507	P38947	Q55585	P06131	P24183	P33160	P46448
Q07103	Q50570	Q60316	P37685	P51650	P42412	Q02252	Q07536	P00352	P08157
P12693	P13601	P20000	P23240	P24549	P30837	P30840	P33008	P40047	P41751
42236	P46329	P46368	P47738	P48644	P51648	P54115	Q25417	Q28399	P07702
P50113	O08318	P23715	P54895	P54899	P96136	Q59279	O67166	P07004	P39821
P45638	P54885	P54903	P74935	P96489	P11883	P43353	P47771	P48448	P08639
P29236	P80505	O24174	P17445	P42757	P56533	P77674	P81406	Q43272	P07003
Q06278	Q54970	P45851	O66112	P06958	P11177	P21873	P21881	P26267	P26284
P32473	P35487	P45119	P47516	P51266	P52899	P52902	P52904	P75391	Q09171
Q10504	Q59097	P07015	P20967	P45303	Q02218	P09060	P11178	P21839	P37940
P50136	O05651	O27772	O58415	O73986	P56815	Q51803	Q51805	Q56317	Q57715
Q57717	O27113	O29779	O29782	Q57956	O27743	P26693	P27989	Q49161	Q49163
Q50538	Q57617	O27002	O31112	P95294	Q58571				
Class 3: 194 Sequences; Average Sequence Identity: 12.89%									
P43901	P20049	P08088	Q12882	Q18164	Q28007	Q28943	P42330	Q04828	P46844
P53004	O26891	O29353	O67061	O84369	O86836	P24703	P38103	P40110	P42976
P45153	P46829	P72024	P72642	Q52419	Q57865	P15047	P39071	Q56632	O30847
O84992	O87612	P27137	P94135	Q45072	P13653	P17652	P21218	P26156	P26163
P26180	P26237	P26238	P28372	P29683	P36208	P36437	P37846	P48099	P48100
P51188	P51278	P54208	P56302	P56303	Q00864	Q04607	Q95666	P42593	Q16698
Q64591	O27083	P21920	P72711	Q10680	Q53139	P07772	P23102	O07400	O24990
O67505	P16657	P42829	P44432	P46533	P54616	P73016	P80030	Q05069	O06236
O27281	O29513	O66461	P05021	P25468	P25996	P28272	P28294	P32747	P32748
P45477	P46539	P46727	P54321	P54322	P74782	Q47741	Q58070	Q63707	O75845
O88822	P11353	P33771	P35055	P36552	P36553	P43898	P72848	Q42840	Q42946
O24163	O24164	O53230	P27863	P32397	P40012	P50336	P51175	P55826	P56601
Q12737	P05335	P06598	P07872	P08790	P11356	P13711	P34355	Q15067	O42772
P21801	P21911	P21912	P21914	P31039	P31040	P47052	P48932	P48933	P80477
P80480	Q00711	Q09508	Q09545	O06913	O06914	P00363	P00364	P07014	P08065
P08066	P10444	P17596	P20921	P20922	P31038	P44893	P44894	P51053	P51054
Q10760	Q10761	Q59661	Q59662	P12007	P26440	P34275	P07670	P15650	P28330
P51174	P79274	Q51697	Q51698	P15651	P16219	Q06319	Q07417	P08503	P11310

ตารางผนวกที่ 1 (ต่อ)

Class 3: 194 Sequences; Average Sequence Identity: 12.89%									
P41367	P45952	Q22347	Q04616	P18405	P24008	P31213	P31214	Q28891	Q28892
P31210	P51857	Q60759	Q92947						
Class 4: 130 Sequences; Average Sequence Identity: 13.94%									
P17557	P30234	Q08352	P09831	P09832	P39812	Q05755	Q05756	Q03460	P04964
P24295	P27346	P28997	P33327	P39633	P41755	Q03578	P22823	P23307	O52310
O59650	O74024	P00366	P00367	P10860	P26443	P49448	P52596	P54385	P80053
P80319	P96110	Q47951	Q53199	Q56304	P00369	P00370	P07262	P14657	P15111
P28724	P29051	P29507	P31026	P39708	P43793	P54386	P54387	P55990	P94316
P94598	P95544	P13154	P54531	P31228	Q99489	P16636	P28300	P28301	Q05063
O06595	P10902	P38032	P74562	Q51363	O93364	P23623	O35078	P00371	P14920
P18894	P22942	P80324	Q19564	Q99042	P19643	P21396	P21397	P21398	P27338
P49253	O06207	O33065	P21159	P28225	P38075	P44909	P74211	O46406	O70423
O75106	P12807	P19801	P36633	P46881	P46883	Q07121	Q07123	Q12556	Q16853
Q29437	Q43077	Q59118	O49850	O49954	P15505	P23378	P26969	P33195	P49095
P49361	P49362	P54377	Q09785	Q50601	P23225	P51375	P55037	P55038	Q06434
P29011	P00372	P22619	P22641	P23006	P29894	Q49124	Q50420	Q59542	Q59543
Class 5: 112 Sequences; Average Sequence Identity: 13.47%									
P07275	P30038	P39634	P78568	P55818	Q02046	P00386	Q44524	O25773	O66553
P00373	P17817	P22008	P22350	P27771	P32263	P43869	P46540	P46725	P52053
P54893	P54904	P74572	Q04708	Q12641	Q12740	Q20848	O74927	O80585	P42898
P46151	P53128	Q17693	P15244	Q44297	O62583	P00374	P00375	P00376	P00377
P00379	P00380	P00381	P00382	P00383	P00384	P04174	P04382	P05794	P07807
P09503	P10167	P11045	P11731	P12833	P13955	P16184	P17719	P22573	P22906
P27421	P27422	P27498	P28019	P31074	P31500	P36591	P43791	P47470	P78028
P78218	P95524	Q07801	Q54277	Q54801	Q57452	Q59397	Q59408	Q59487	Q59908
Q93341	O75891	P28037	P38997	P38998	P43065	Q09694	O87386	O87388	P23342
P40854	P40859	P40873	P40874	P40875	Q46336	Q46338	O64411	P08159	P30986
P87111	P94132	Q08822	Q11190	Q48303	Q63342	P16099	O29544	P55300	P94951
Q50501	Q58441								
Class 6: 305 Sequences; Average Sequence Identity: 11.71%									
P07001	P41077	P51995	P00387	P36060	P00389	P16603	P37040	P50126	P00390
P27456	P42770	P48638	P48641	Q43621	O62768	O84101	P38816	P43788	P50971
P52214	P75531	P94284	Q17745	Q92375	P39040	O03060	O03172	O03175	O03203
O03206	O03850	O21000	O21070	O21233	O21325	O21333	O21336	O21405	O21408
O21514	O21798	O47430	O47492	O47498	O53307	O63850	O66842	O68853	O78680
O78688	O78694	O78697	O78701	O78704	O78707	O78710	O78714	O78748	O78755

ตารางผนวกที่ 1 (ต่อ)

Class 6: 305 Sequences; Average Sequence									
O79408	O79411	O79421	O79427	O79435	O79438	O79677	O79874	O79881	O84970
O85274	O99823	O99826	O99978	P03888	P03891	P03894	P03897	P03900	P03903
P03909	P03912	P03916	P03919	P03922	P03925	P04540	P05507	P05510	P06253
P06256	P06259	P06262	P06265	P06410	P07706	P07709	P08740	P09045	P11628
P11631	P11658	P11991	P12099	P12126	P12129	P12132	P12199	P12771	P12774
P12777	P15550	P15553	P15577	P15580	P15584	P15956	P15959	P16673	P18903
P18931	P18934	P18938	P18941	P19044	P19050	P20113	P20686	P21301	P24873
P24877	P24884	P24887	P24895	P24969	P24972	P24975	P24978	P24982	P24997
P25707	P26289	P26522	P26525	P26847	P26850	P27572	P29801	P29915	P29918
P29921	P29924	P30826	P31978	P32421	P33509	P33512	P33598	P33602	P33605
P33608	P33903	P34192	P34195	P34847	P34850	P34853	P34856	P34859	P38599
P38602	P41296	P41299	P41304	P41307	P41315	P42032	P43191	P43194	P43197
P43200	P43203	P43206	P46619	P46722	P48176	P48653	P48656	P48897	P48900
P48903	P48907	P48910	P48913	P48916	P48919	P48922	P48925	P48928	P48931
P50367	P50940	P50975	P51097	P51100	P52765	P55780	P55783	P56752	P56755
P56896	P56908	P56911	P56914	P92475	P92483	P92486	P92659	P92667	P92670
P92697	P92700	P93401	P95174	P95177	P95180	Q00236	Q00244	Q00540	Q00543
Q00570	Q01562	Q04050	Q31849	Q32238	Q33635	Q33821	Q34050	Q34573	Q34947
Q34950	Q35100	Q35535	Q35542	Q35585	Q35813	Q36346	Q36424	Q36428	Q36457
Q36460	Q36836	Q37312	Q37371	Q37375	Q37381	Q37546	Q37603	Q37626	Q37680
Q37710	Q37714	Q37809	Q44241	Q56218	Q56221	Q56224	Q56227	Q60010	Q95704
Q95710	Q95891	Q95915	Q95918	Q96007	Q96067	Q96070	Q96186	P28304	P43903
Q28452	P11605	P17569	P27967	P39865	P39868	P43101	P27783	P22945	P39863
P49050	P22944	P42435	P15344	O87948	O85762	P41816	Q03558	P05982	Q64669
P37061	P75389	Q60049	P24232	Q03331					
Class 7: 64 Sequences; Average Sequence Identity: 13.69%									
O04420	O32141	O74409	P04670	P09118	P11645	P16163	P16164	P22673	P23194
P25689	P33282	P34798	P34799	P53763	P78609	Q00511	Q45697	Q50925	P05314
P39661	Q51879	P25006	P38501	Q01537	Q06006	Q53239	Q60214	P09152	P11349
P11351	P19316	P19317	P19318	P19319	P33937	P39185	P39458	P42175	P42176
P42178	P42434	P73448	P81186	Q06457	Q53176	Q56350	O54235	O67422	P00394
P45208	P71319	P19573	P94127	Q01710	Q51705	Q59105	Q59746	O06844	O50651
Q51662	Q52527	Q59646	Q59647						
Class 8: 59 Sequences; Average Sequence Identity: 19.70%									
P17846	P38038	P38039	P39692	P52673	P52674	Q09878	O00087	O08749	O18480
O50286	O50311	O84561	P00391	P09063	P09622	P09623	P09624	P11959	P14218
P21880	P31023	P31046	P31052	P43784	P47513	P49819	P50970	P52992	P54533

ตารางผนวกที่ 1 (ต่อ)

Class 8: 59 Sequences; Average Sequence Identity: 19.70%									
P75393	P90597	P95596	Q04829	Q04933	Q59822	P07850	P51687	Q07116	P30008
O33998	P45573	P45574	P45575	Q59109	Q59110	O05927	O06737	P17853	P17854
P52672	P56859	P56860	P56891	P71752	P72794	P94498	Q10270	Q55309	
Class 9: 254 Sequences; Average Sequence Identity: 21.32%									
O03167	O03198	O03848	O13082	O21327	O21399	O21403	O47425	O47491	O47667
O47669	O47671	O47673	O47675	O47677	O47679	O47681	O47686	O47688	O47690
O47692	O47694	O47696	O47698	O47700	O47702	O47705	O47708	O47710	O48316
O48374	O54069	O74471	O78682	O78750	O79404	O79417	O79433	O79673	O79876
O99255	O99819	P00395	P00397	P00399	P00401	P00403	P00405	P00407	P00409
P00411	P00413	P00415	P00417	P00419	P00421	P00423	P00425	P00427	P00429
P03945	P04038	P04371	P04373	P05490	P05502	P05505	P06030	P07255	P07471
P07657	P08306	P08741	P08743	P08745	P08749	P09669	P10175	P10606	P10888
P11947	P11950	P12074	P12700	P12702	P12787	P13182	P13184	P14058	P14544
P14546	P14574	P14578	P14852	P14854	P15545	P15952	P15954	P16262	P18943
P18945	P19536	P20374	P20386	P20609	P20674	P20682	P20684	P24010	P24012
P24310	P24794	P24881	P24891	P24894	P24985	P24987	P24989	P25002	P25312
P26455	P26457	P26857	P27168	P29505	P29856	P29860	P29864	P29870	P29872
P29874	P29876	P29878	P29880	P30815	P32799	P33504	P33508	P33518	P34189
P34838	P34842	P35171	P38596	P41293	P41295	P41311	P41775	P43024	P43370
P43372	P43374	P43376	P47918	P48171	P48659	P48661	P48772	P48867	P48869
P48871	P48873	P48887	P48889	P48891	P50253	P50268	P50666	P50672	P50674
P50676	P50678	P50680	P50684	P50686	P50688	P50690	P50692	P55777	P56392
P79010	P80439	P80441	P92478	P92514	P92662	P92692	P92696	P98000	P98002
P98012	P98020	P98023	P98025	P98027	P98031	P98033	P98035	P98037	P98039
P98042	P98044	P98047	P98049	P98053	P98055	P98057	Q00527	Q00529	Q01555
Q02211	Q02221	Q02766	Q03227	Q03439	Q03736	Q04441	Q04452	Q05572	Q06474
Q07063	Q08855	Q10375	Q20779	Q33824	Q34941	Q35101	Q35539	Q36309	Q36452
Q36675	Q36837	Q36952	Q37369	Q37374	Q37416	Q37430	Q37472	Q37548	Q37604
Q37620	Q37677	Q37684	Q37705	Q37718	Q42841	Q94514	Q95840	Q95914	Q96065
Q96133	Q96190	P24474	Q51700						
Class 10: 94 Sequences; Average Sequence Identity: 12.71%									
O01374	O14949	O14957	O31214	O60044	P00126	P00127	P00128	P00130	P05417
P07056	P07256	P07257	P07552	P07919	P08067	P08525	P13271	P13272	P14927
P16536	P22289	P22695	P23004	P31800	P31930	P32551	P37299	P37841	P43264
P43265	P43266	P46269	P47985	P48502	P48503	P48504	P48505	P49345	P49346
P50523	P51130	P51132	P51133	P51134	P51135	P78761	P81380	Q02762	Q09154
P43309	P43310	P43311	Q00024	Q06215	Q08296	Q08304	Q08305	Q08306	Q08307

ตารางผนวกที่ 1 (ต่อ)

Class 10: 94 Sequences; Average Sequence Identity: 12.71%									
P06811	P10574	P17489	P56193	Q01679	Q02075	Q02079	Q02081	Q02497	Q03966
Q12541	Q12542	Q12718	Q12719	Q12729	Q12739	Q99044	Q99046	Q99049	Q99055
P14133	P24792	P37064	Q00624	Q40588	P08980	P14698	P26290	P26292	P30361
P49728	P70758	Q02585	Q46136						
Class 11: 154 Sequences; Average Sequence Identity: 19.01%									
O31158	O31168	P04963	P25026	P49053	P49323	Q55921	P48534	P19136	P00431
P14532	P37197	O13289	O52762	O61235	O68146	P04040	P04762	P06115	P07145
P07820	P11934	P12365	P13029	P15202	P17336	P17598	P17750	P18122	P18123
P21179	P24168	P25819	P25890	P26901	P29422	P29611	P29756	P30263	P30264
P30567	P32290	P37743	P42234	P42321	P44390	P45737	P45739	P46817	P48062
P48350	P48351	P48352	P49284	P49316	P49317	P49319	P50979	P55303	P55304
P55305	P55306	P55307	P55308	P55310	P55311	P55312	P55313	P77872	P78574
P78619	P81138	P95539	P95631	Q01297	Q04657	Q08129	Q27710	Q42547	Q43206
Q59296	Q59337	Q59602	Q59635	Q59714	Q64405	Q92405	Q96528	O35244	O77834
P00433	P00434	P05164	P11247	P11678	P11965	P15004	P15232	P17179	P17180
P22079	P22195	P22196	P24101	P27337	P28313	P28314	P30041	P37834	P37835
P49290	P80025	Q01603	Q02200	Q05855	P07202	P09933	P14650	P35419	O02621
O18994	O23968	O23970	O32770	O46607	O59858	O62327	O75715	P04041	P07203
P11352	P11909	P12079	P18283	P21765	P22352	P28714	P30708	P30710	P35665
P35666	P36014	P36968	P36969	P37141	P38143	P40581	P46412	P52032	P52033
P74250	Q00277	Q64625	Q95003						
Class 12: 94 Sequences; Average Sequence Identity: 13.66%									
O33950	O67987	P05403	P07773	P11451	P27098	P31019	P20351	P48775	P48776
Q09474	P08170	P09186	P09439	P09918	P14856	P27480	P29114	P29250	P37831
P38414	P38415	P38416	P38417	P38419	Q06327	Q05353	P06622	P08127	P17262
P17295	P17296	P31003	Q04285	Q53034	P21816	Q16878	O23920	O42764	O48604
P49429	P80064	P93836	Q00415	Q02110	Q22633	Q27203	Q53407	P00436	P00437
P15109	P15110	P20371	P20372	P16469	P18054	P39654	P39655	P55249	Q02759
Q01284	Q12723	P12530	P16050	P12527	P48999	P51399	P08695	P11122	P17297
P47228	P47231	P47233	P14902	P28776	O09173	Q00667	Q93099	P46952	P46953
P22635	P22636	P11295	P04029	P06617	P25017	Q04564	Q09109	P21795	P27652
P17554	P08659	P13129	Q01158						
Class 13: 257 Sequences; Average Sequence Identity: 15.71%									
O75936	Q19000	Q12797	P13674	P54001	Q60716	O60568	P24802	Q20679	P28038
Q05963	Q06942	P07770	P23094	Q51494	P08084	Q07944	Q05182	P23262	O24312
P37114	P48522	Q04468	Q43033	Q43240	P17549	P18125	P46634	Q64505	P20586

ตารางผนวกที่ 1 (ต่อ)

Class 13: 257 Sequences; Average Sequence Identity:*a* 15.71%									
P27138	P11987	P22868	P27353	P27355	Q08477	O54705	O61309	P29473	P29475
P29477	P70313	Q26240	Q28969	P42535	P19729	P19731	P19733	P31020	P16549
P17636	P31513	P36365	P36367	P49326	P97501	Q01740	Q28505	O09158	O16805
O18809	O18992	O35293	O42231	O42457	O46420	O54749	O55071	O62671	O73686
O93299	P00178	P00181	P00184	P00186	P04167	P04799	P05176	P05178	P05180
P05182	P05184	P08683	P10610	P10613	P10615	P10633	P10635	P11509	P11511
P11711	P11713	P12789	P12791	P12939	P13108	P13584	P14762	P15128	P15149
P16141	P17666	P19225	P20812	P20814	P20852	P21595	P24453	P24455	P24457
P24460	P24462	P24464	P24903	P29981	P30608	P30610	P33260	P33262	P33264
P33268	P33270	P33274	P43083	P49602	P51538	P51589	P51869	P51871	P56590
P56592	P56654	P56656	P70091	P79304	P79401	P79690	P79760	P93846	Q00557
Q05047	Q05555	Q06367	Q09736	Q12586	Q12589	Q16678	Q16850	Q27593	Q27606
Q27712	Q27902	Q29488	Q29510	Q29624	Q64391	Q64406	Q64417	Q64458	Q64462
Q64481	Q64654	Q64680	Q92087	Q92090	Q92100	Q92110	Q92112	Q92148	P07739
P09140	P12744	P19839	P19907	P23146	P24113	P29239	P08516	P14579	P14581
P20817	P15150	P15538	P19099	P30100	P97720	Q29552	Q64658	P00189	P10612
P79153	Q07217	P04176	P30967	P90925	O42091	P07101	P17289	P24529	P09810
P17532	P70080	P09172	Q05754	P08478	P12890	P19021	O42713	P06845	P11344
P33180	P55023	P55025	Q04604	O02768	O62698	P05979	P23219	P35354	P70682
Q05769	P00191	P08686	Q02390	O19998	O70453	O78497	P09601	P14901	P30519
P43242	P71119	P74133	P07308	P13516	Q64420	P22243	P28645	P32061	P32063
Q01753	Q40731	Q42770	Q43593	O48651	O65403	O65726	P32476	P52020	Q92206
O73853	P05185	P12394	P27786	P70085	Q29497	Q92113			
Class 14: 155 Sequences; Average Sequence Identity: 26.92%									
O04997	O09164	O12933	O13401	O15905	O22373	O30563	O30826	O42724	O46412
O49044	O49066	O51917	O54086	O59924	O67470	O86165	O93724	P00441	P00442
P00443	P00444	P00445	P00446	P00448	P00449	P03946	P04178	P04179	P07505
P07509	P07632	P08228	P08294	P09157	P09212	P09213	P09214	P09223	P09224
P09670	P09671	P09678	P09737	P09738	P10791	P10792	P11418	P11796	P11964
P13367	P13926	P14830	P14831	P15107	P15453	P17550	P17670	P18655	P18868
P19665	P19666	P19685	P20379	P23345	P23346	P23744	P24669	P24702	P24704
P24705	P24706	P25842	P27082	P27084	P28755	P28756	P28757	P28758	P28759
P28763	P28764	P31108	P31161	P33431	P34107	P34461	P34697	P36214	P37369
P41962	P41963	P41973	P41974	P41975	P41976	P41978	P41979	P41980	P41981
P43019	P43312	P43725	P47201	P50059	P50061	P50911	P51547	P53635	P53636
P53637	P53638	P53640	P53641	P53642	P53649	P53651	P53652	P53653	P53654

ตารางผนวกที่ 1 (ต่อ)

Class 14: 155 Sequences; Average Sequence Identity: 26.92%									
P54375	P77928	P77929	P77968	P80174	P80293	P80566	P80734	P80857	P81926
P93258	P93407	Q00637	Q01137	Q02610	Q03299	Q03301	Q03302	Q03303	Q07182
Q07449	Q07796	Q08420	Q08713	Q42684	Q43779	Q59094	Q59448	Q59452	Q59519
Q59623	Q59679	Q60036	Q92429	Q92450					
Class 15: 84 Sequences; Average Sequence Identity: 18.55%									
O15910	O46310	O61065	O66503	O83092	O83972	O84834	P00452	P03174	P03175
P03190	P06474	P07201	P07742	P08543	P09247	P09853	P09938	P10224	P11156
P11157	P11158	P12848	P16782	P20493	P20503	P21524	P21672	P23921	P26685
P26713	P28846	P29883	P31350	P32209	P32282	P32984	P33799	P36602	P36603
P37426	P37427	P39452	P42170	P42491	P42492	P42521	P43754	P47471	P47473
P48591	P48592	P49723	P49730	P50620	P50621	P50641	P50642	P50643	P50644
P50645	P50646	P50647	P50648	P50650	P50651	P52343	P55982	P55983	P74240
P75461	P78027	P79733	Q01037	Q01038	Q01319	Q03604	Q08698	Q10840	Q60561
P07071	P28903	P43752	Q05262						
Class 16: 154 Sequences; Average Sequence Identity: 16.99%									
P42454	O04397	O04977	O23877	P00454	P00455	P08165	P10933	P22570	P28861
P31973	P41343	P41344	P41345	P41346	P53991	Q00598	Q10547	Q41014	Q44532
Q44549	Q55318	Q61578	P07771	P13452	P21394	P23101	P37337	P77650	Q03304
Q07946	Q52126	O07643	O26739	O27605	O27606	O68940	O68943	O68946	O68951
P00457	P00458	P00459	P00460	P00461	P00462	P00463	P00464	P00467	P00468
P06117	P06118	P06119	P06120	P06121	P06122	P06662	P06769	P07328	P07329
P08624	P08625	P08717	P08718	P09552	P09553	P09554	P09555	P09772	P11347
P15052	P15332	P15334	P15335	P16266	P16267	P16268	P16269	P16855	P16856
P17303	P19066	P19067	P19068	P20620	P20621	P22548	P22921	P25314	P25767
P26248	P26250	P26251	P26252	P33178	P46034	P51754	P54799	P54800	P55170
P71526	P71527	P77874	P95296	Q00240	Q02452	Q07933	Q07934	Q07935	Q07942
Q44044	Q44045	Q46083	Q46244	Q50218	Q50785	Q50788	Q55029	Q55030	Q57118
Q58289	Q59270	P07598	P07603	P12635	P12636	P12943	P12944	P13063	P13065
P13628	P13629	P15283	P15284	P17632	P17633	P18188	P18190	P18191	P18636
P18637	P19927	P19928	P21852	P21949	P21950	P29166	P31891	P31892	P33374
P33375	P37181	Q46046	Q46847						

ประวัติการศึกษา และการทำงาน

ชื่อ	นายบัลลังก์ นิยมศักดิ์
เกิดวันที่	27 มีนาคม 2525
สถานที่เกิด	กรุงเทพมหานครฯ
ประวัติการศึกษา	วท.บ. (วิทยาการคอมพิวเตอร์) คณะวิทยาศาสตร์ มหาวิทยาลัยกรุงเทพ (พ.ศ. 2546)
ตำแหน่งปัจจุบัน	อาจารย์พิเศษ
สถานที่ทำงานปัจจุบัน	มหาวิทยาลัยกรุงเทพ
ผลงานดีเด่นและ/หรือรางวัลทางวิชาการ	
ทุนการศึกษาที่ได้รับ	