วิทยานิพนธ์

การพัฒนาตัวแทนสายโปรตีนสำหรับการทำนายประเภทฟังก์ชั่นเอนไซม์

DEVELOPMENT OF PROTEIN SEQUENCE REPRESENTATION FOR ENZYME FUNCTION CLASSIFICATION

นายพีระ ลิ่วลม

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

พ.ศ. 2551

ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

วิศวกรรมศาสตรดุษฎีบัณฑิต (วิศวกรรมคอมพิวเตอร์)

ปริญญา

วิศวกรรมคอมพิวเตอร์	วิศวกรรมคอมพิวเตอร์
สาขา	**ภาควิชา**

เรื่อง การพัฒนาตัวแทนสายโปรตีนสำหรับการทำนายประเภทฟังก์ชั่นเอนไซม์

Development of Protein Sequence Representation for Enzyme Function Classification

นามผู้วิจัย นายพีระ ลิ่วลม

ได้พิจารณาเห็นชอบโดย

ประธานกรรมการ .....

(รองศาสตราจารย์กฤษณะ ไวยมัย, Doctorat d'Universite)

กรรมการ .....

(รองศาสตราจารย์พันธุ์ปิติ เปี่ยมสง่า, D.Sc.)

กรรมการ .....

(ผู้ช่วยศาสตราจารย์ยอดเยี่ยม ทิพย์สุวรรณ, Ph.D.)

หัวหน้าภาควิชา .....

(ผู้ช่วยศาสตราจารย์เขมะฑัต วิภาตะวนิช, Ph.D.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

.....

(รองศาสตราจารย์กัญจนา ธีระกุล, D.Agr.)

คณบดีบัณฑิตวิทยาลัย

วันที่ 4 **เดือน** มิถุนายน **พ.ศ.** 2551

วิทยานิพนธ์

เรื่อง

การพัฒนาตัวแทนสายโปรตีนสำหรับการทำนายประเภทฟังก์ชั่นเอนไซม์

Development of Protein Sequence Representation for Enzyme Function Classification

โดย

นายพีระ ลิ่วลม

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์
เพื่อความสมบูรณ์แห่งปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต (วิศวกรรมคอมพิวเตอร์)
พ.ศ. 2551

พีระ ลิ่วลม 2551: การพัฒนาตัวแทนสายโปรตีนสำหรับการทำนายประเภทฟังก์ชั่นเอนไซม์ ปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต (วิศวกรรมคอมพิวเตอร์) สาขาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ ประธานกรรมการที่ปรึกษา: รองศาสตราจารย์กฤษณะ ไวยมัย, Doctorat d'Universite 102 หน้า

วิทยานิพนธ์นี้มุ่งพัฒนาตัวแทนสายโปรตีนประเภทโมทีฟสำหรับการทำนายประเภทฟังก์ชั่นเอนไซม์ โดยนำเสนอตัวแทนสายโปรตีนชนิดใหม่ที่เรียกว่า รีแอคทีฟโมทีฟ (reactive motif) ซึ่งมีความสัมพันธ์โดยตรงกับการเกิดฟังก์ชั่นเอนไซม์ในบริเวณจับและบริเวณแร่ง โดยปัญหาวิจัยที่สำคัญก็คือ การขาดแคลนข้อมูล (มีเพียงประมาณ 3.34% เท่านั้นจากจำนวนเอนไซม์ทั้งหมด) ในการนำมาสร้างรีแอคทีฟโมทีฟ ดังนั้นในวิทยานิพนธ์นี้จึงนำเสนอวิธีการด้านสถิติร่วมกับการใช้ความรู้พื้นหลังด้านชีวเคมีในการสร้างและพัฒนาคุณภาพของรีแอคทีฟโมทีฟ สำหรับวิธีการทางสถิติใช้ในการขยายกลุ่มข้อมูลบริเวณจับและบริเวณแร่งให้ได้ปริมาณที่เหมาะสมและมีคุณภาพ ส่วนการใช้ความรู้พื้นหลังเช่น แผนภาพของ Taylor และ BLOSUM62 ถูกนำมาใช้สำหรับเพิ่มคุณภาพองค์ประกอบย่อยของรีแอคทีฟโมทีฟ โดยใช้คอนเซ็ปต์ของ การควบคุมการกลายพันธุ์ (mutation control) ในการพัฒนากลุ่มแทนที่กรดอะมิโนเป็น กลุ่มแทนที่กรดอะมิโนที่สมบูรณ์ โดยคอนเซ็ปต์การควบคุมการกลายพันธุ์ดังกล่าวถูกอธิบายและจัดให้อยู่ในระเบียบแบบแผนของทฤษฎีคอนเซ็ปต์แลททิส ทั้งนี้เครื่องมือวัดประเภท sensitivity precision specificity และ coverage ถูกนำมาใช้วัดคุณภาพของรีแอคทีฟโมทีฟที่ได้จากงานวิจัย ผลการวัดพบว่ารีแอคทีฟโมทีฟมีคุณภาพในเชิง sensitivity และ coverage ดีกว่าโมทีฟ PROSITE ที่พัฒนาโดยผู้เชี่ยวชาญ ในขณะที่ค่าคุณภาพที่เหลือมีค่าใกล้เคียงกัน สำหรับเครื่องมือวัดประเภท accuracy ที่ได้จากการทำนายประเภทฟังก์ชั่นเอนไซม์ ใช้วัดคุณภาพของรีแอคทีฟโมทีฟเมื่อนำมาใช้เป็นคุณลักษณะเด่นในการเรียนรู้ด้วยขั้นตอนวิธี C4.5 สำหรับทำนายประเภทฟังก์ชั่นเอนไซม์ ผลวิจัยพบว่าการใช้รีแอคทีฟโมทีฟดังกล่าวให้ค่า accuracy ถึง 72% เมื่อเปรียบเทียบกับ 67% ที่ใช้ PROSITE

พีระ ลิ่วลม 3 / 6 / 2551

ลายมือชื่อนิสิต ลายมือชื่อประธานกรรมการ

Peera Liewlom 2008: Development of Protein Sequence Representation for Enzyme Function Classification. Doctor of Engineering (Computer Engineering), Major Field: Computer Engineering, Department of Computer Engineering. Thesis Advisor: Associate Professor Kitsana Waiyamai, Doctorat d'Universite ¡102 pages.

This thesis is concentrated in developing a motif-based protein sequence representation for enzyme function classification. The main objective of this thesis is to propose a new protein sequence representation called reactive motifs. Reactive motifs are motifs that are related directly to enzyme functions and are generated from binding and catalytic sites. Main challenge is a lack of data (only 3.34% data available of all enzymes) at binding and catalytic sites to generate reactive motifs. Therefore, a method that combines statistics and bio-chemistry background knowledge is proposed to generate and improve reactive motifs. Statistics are used to extend binding and catalytic sites data. Bio-chemistry background knowledge such as Taylor – Venn's diagram and BLOSUM62 are used to improve the quality of the reactive motif elements. The concept of mutation control is introduced, which uses amino acid substitution groups to generate maximal amino acid substitution groups. Mutation control operations are described and formalized using concept lattice. Sensitivity, precision, specificity, and coverage measures are used to assess the quality of the discovered reactive motifs. Experimental results show that the discovered reactive motifs provide better quality in terms of sensitivity and coverage compared to PROSITE expert-based motifs. To assess the accuracy in enzyme function classification, the reactive motifs are used as input to the C4.5 learning algorithm. Experiments using reactive motifs as feature predict enzyme function with 72% accuracy compared with 67% accuracy using PROSITE motifs.

พีร เลิศลม

Student's signature

Thesis Advisor's signature

3 / 6 / 2008

กิตติกรรมประกาศ

ขอกราบขอบพระคุณ รศ.ดร.กฤษณะ ไวยมัย ประธานกรรมการที่ปรึกษาวิทยานิพนธ์ ที่ให้โอกาสในการศึกษาและความรู้ในการศึกษาวิจัย ขอกราบขอบพระคุณ รศ.ดร.พันธุ์ปิติ เปี่ยมสง่า กรรมการที่ปรึกษาวิชาเอก และ ผศ.ดร.ยอดเยี่ยม ทิพย์สุวรรณ กรรมการที่ปรึกษาวิชารอง ที่ให้คำปรึกษาในการค้นคว้าวิจัย ตลอดจนการตรวจทานแก้ไขวิทยานิพนธ์จนเสร็จสมบูรณ์ และขอกราบขอบพระคุณ ผศ.ดร.อรรถสิทธิ์ สุรฤกษ์ ผู้ทรงคุณวุฒิภายนอก ที่ได้กรุณาให้คำแนะนำที่เป็นประโยชน์อย่างยิ่งต่อวิทยานิพนธ์ฉบับนี้

ขอกราบขอบพระคุณคุณปู่วีระ คุณพ่อ วัลลภ และคุณแม่เสริมสุข ในความอดทน สนับสนุนส่งเสริมทั้งด้านการเงินและกำลังใจข้าพเจ้าจนเรียนสำเร็จ รวมถึง รศ.ดร.พงษ์ศักดิ์ สุริยวนากุล ผู้บังคับบัญชาที่ให้โอกาสในการศึกษาต่อครั้งนี้

ขอขอบคุณพี่ศิวะเป็นอย่างยิ่งที่ทุ่มเทแนะนำในเรื่องงานเขียนเป็นระยะเวลายาวนาน ขอบคุณน้องปียะและภีมที่คอยสนับสนุนอย่างสม่ำเสมอ รวมถึงน้องวรรณทิชาที่เป็นกำลังใจและช่วยทำรูปเล่มวิทยานิพนธ์นี้จนสำเร็จเรียบร้อยด้วยดี

สุดท้ายขอบคุณเพื่อนๆ และน้องๆ ในห้องแลบ DAKDL ตั้งแต่รุ่นปี พ.ศ. 2547 จนถึงปี พ.ศ. 2550 ที่ได้มีส่วนช่วยกันแนะนำให้งานมีความก้าวหน้า รวมถึงขอบคุณเจ้าหน้าที่ธุรการที่ให้การประสานงานแก้ไขปัญหาต่างๆ จนลุล่วงราบรื่นมาตลอด

พีระ ลิ่วลม

เมษายน 2551

สารบัญ

หน้า

สารบัญตาราง

สารบัญตาราง (ต่อ)

สารบัญภาพ

สารบัญภาพ (ต่อ)

คำอธิบายสัญลักษณ์และคำย่อ

CL	=	Concept Lattice
MSA	=	Multiple Sequence Alignment
แลททิส	=	คอนเซ็ปต์แลททิส หรือ Concept Lattice

การพัฒนาตัวแทนสายโปรตีนสำหรับการทำนายประเภทฟังก์ชั่นเอนไซม์

Development of Protein Sequence Representation for Enzyme Function Classification

คำนำ

โปรตีนเอนไซม์เป็นสารพื้นฐานของสิ่งมีชีวิต การเข้าใจฟังก์ชั่นการทำงานของโปรตีนเอนไซม์ย่อมหมายถึงโอกาสในการประยุกต์ใช้งานเอนไซม์ในสิ่งมีชีวิตทั้งหลายรวมทั้งในร่างกายของมนุษย์ เช่น การทำงานของยารักษาโรคที่มีประสิทธิภาพมากขึ้น การเพิ่มผลผลิตของพืชและสัตว์ เป็นต้น ทั้งนี้ปัจจุบันมีโปรตีนเอนไซม์ในฐานข้อมูล ENZYME nomenclature database ทั้งหมด 116,330 ชนิด จากโปรตีนทั้งหมดที่ค้นพบจริงและมีการตรวจสอบคุณสมบัติแล้วเพียง 333,445 ชนิด (UNIPROT, 2007) โดยโปรตีนที่ค้นพบจากจีโนม (genome) มีทั้งหมด 5,139,891 ชนิด และยังคงมีการปรับปรุงข้อมูลอยู่เรื่อยๆ ซึ่งจากข้อมูลที่มีพอสมควรนี้ทำให้การศึกษาฟังก์ชั่นเอนไซม์ด้วยเทคนิคด้านชีวสารสนเทศ (Bioinformatics) เป็นอีกหนทางหนึ่งในการศึกษาวิจัยโปรตีนเอนไซม์ที่น่าสนใจ โดยการพัฒนาเทคนิคการทำนายฟังก์ชั่นเอนไซม์ที่มีความแม่นยำสูงเป็นหนทางหนึ่งในการใช้ประโยชน์จากข้อมูลที่มีอยู่ดังกล่าว

อย่างไรก็ตามความแม่นยำในการทำนายไม่ใช่สิ่งที่สร้างความเชื่อมั่นในการใช้งานระดับห้องปฏิบัติการ ที่มีความต้องการมากกว่าการระบุว่าโปรตีนที่สงสัยทำงานฟังก์ชั่นใดเพียงเท่านั้น แต่ต้องการทราบว่าโปรตีนนั้นมีกลไกฟังก์ชั่นเอนไซม์ใดอยู่ที่ส่วนใดของโปรตีน ซึ่งมีผลต่อการทำความเข้าใจการทำงานของเอนไซม์ชนิดนั้น และสามารถประเมินความเชื่อมั่นในผลการทำนายได้จากความสมเหตุผล ดังนั้นการพัฒนาตัวแทนสายโปรตีนจากส่วนที่เกิดกลไกเอนไซม์โดยตรงเพื่อทำนายฟังก์ชั่นเอนไซม์ จึงเป็นหัวข้อที่น่าสนใจอย่างยิ่ง เนื่องจากตัวแบบการทำนายฟังก์ชั่นที่ได้จากส่วนที่เป็นกลไกการทำงานของเอนไซม์โดยตรงสมควรมีความแม่นยำสูง อีกทั้งสามารถระบุได้ว่าส่วนใดของโปรตีนที่เป็นกลไกการทำงานเอนไซม์

ตัวแทนสายโปรตีนประเภทหนึ่งที่นิยมใช้กันมากในศาสตร์ด้านชีวสารสนเทศก็คือ โมทีฟ (motif) โดยเป็นรูปแบบสายลำดับ (sequence pattern) ที่ใช้ระบุเอกลักษณ์ (identity) ของกลุ่มสายโปรตีนนั้น โดยโมทีฟที่ใช้เป็นตัวแทนสายโปรตีนในการแสดงคุณสมบัติฟังก์ชั่นเอนไซม์ควรมี

ลักษณะที่สอดคล้องกับการทำงานของเอนไซม์นั้นคือการทำงานเข้าจับพอดีกับสารตั้งต้น (substrate) แล้วก่อฟังก์ชั่น หรือ “fit and function” ดังนั้นงานวิจัยนี้จึงเน้นการใช้งานโมทีฟประเภทที่ไม่มีการยืดหยุ่นของรูปแบบสายลำดับที่เรียกว่า “no gap” ทั้งนี้การค้นพบโมทีฟประเภทนี้มีพื้นฐานเริ่มต้นจากกลุ่มข้อมูลสายโปรตีนที่จัดเรียงโครงสร้างข้อมูลแบบเดียวกับกลุ่มสายโปรตีนในฐานข้อมูล BLOCKS (Henikoff and Henikoff, 1991) โดยในงานศึกษานี้เรียกโครงสร้างข้อมูลชนิดนี้ว่า บลอก (block) โดยแต่ละบลอกเป็นการรวบรวมส่วนของสายโปรตีนในบริเวณจับหรือบริเวณเร่งชนิดเดียวกันเข้ามาอยู่ในบลอกเดียวกัน ทั้งนี้การทำงานฟังก์ชั่นเอนไซม์ของโปรตีนเกิดขึ้นในส่วนที่เรียกว่าบริเวณจับและบริเวณเร่ง ซึ่งปัจจุบันมีข้อมูลบริเวณดังกล่าวในฐานข้อมูลเพียงประมาณ 3% เท่านั้น นอกจากนี้การเกิดฟังก์ชั่นเอนไซม์แต่ละประเภทเกิดขึ้นจากการประสานงานระหว่างบริเวณจับและบริเวณเร่ง ทำให้ 1 ฟังก์ชั่นเอนไซม์ประกอบด้วยหลายบริเวณจับและบริเวณเร่ง ในทางกลับกันบริเวณจับหรือบริเวณเร่งประเภทหนึ่งๆ สามารถทำงานได้ในฟังก์ชั่นเอนไซม์หลายประเภท ซึ่งการมีข้อมูลน้อยและลักษณะเฉพาะที่ซับซ้อนของบริเวณจับและบริเวณเร่งดังกล่าว เป็นปัญหาอย่างยิ่งต่อการค้นพบโมทีฟจากบริเวณเหล่านี้เพื่อใช้งานอย่างจริงจังในการทำนายฟังก์ชั่นเอนไซม์

ปัญหาการมีข้อมูลน้อยและความซับซ้อนของโมทีฟจากบริเวณจับและบริเวณเร่งดังกล่าวสามารถศึกษาผ่านโมทีฟที่รู้จักกันดีในปัจจุบัน เช่น PROSITE (Bairoch, 1991) และ eMotif (Huang and Brutlag, 2001) โดย PROSITE เป็นฐานข้อมูลโมทีฟที่พัฒนาขึ้นโดยผู้เชี่ยวชาญด้านโปรตีนมีการพัฒนามาอย่างยาวนานตั้งแต่ปี 1988 แต่มีโมทีฟที่ค้นพบจากบริเวณจับและบริเวณเร่งเพียง 152 โมทีฟ ซึ่งครอบคลุมฟังก์ชั่นเอนไซม์ 396 ฟังก์ชั่นจากทั้งหมด 3,845 ฟังก์ชั่น การมีโมทีฟประเภทนี้อยู่น้อยเนื่องจากการมีข้อมูลที่ใช้ค้นพบโมทีฟมีน้อย อีกทั้งการพัฒนาโดยผู้เชี่ยวชาญเป็นกระบวนการที่ค่อนข้างช้า นอกจากนี้โมทีฟของ PROSITE ยังสามารถระบุความซับซ้อนของบริเวณจับและบริเวณเร่งได้อย่างชัดเจน โดยบางโมทีฟเป็นส่วนทำงานของฟังก์ชั่นเอนไซม์ถึง 46 ฟังก์ชั่น และมีถึง 139 ฟังก์ชั่นที่ประกอบด้วยโมทีฟจากบริเวณจับและบริเวณเร่งมากกว่า 1 โมทีฟ ในขณะที่ eMotif เป็นฐานข้อมูลที่เก็บโมทีฟที่ได้จากการค้นพบอัตโนมัติจากฐานข้อมูล BLOCKS มีโมทีฟอยู่เป็นจำนวนมาก แต่เนื่องจากฐานข้อมูล BLOCKS มีข้อมูลบริเวณจับและบริเวณเร่งอยู่เพียงประมาณ 30 บลอกเท่านั้น จึงมีโมทีฟที่ได้จากบริเวณจับและบริเวณเร่งในจำนวนที่ไม่มากนัก

ดังนั้นในงานวิจัยนี้จึงออกแบบเนื้องานการทำนายฟังก์ชั่นเอนไซม์จากโมทีฟที่ค้นพบจากข้อมูลบริเวณจับและบริเวณเร่งออกเป็น 2 ส่วนงานหลักคือ 1.การพัฒนาคุณภาพของข้อมูลที่มีอยู่

น้อยมากเพื่อใช้สร้างตัวแทนสายโปรตีนที่มีคุณภาพ และ 2.การสร้างตัวแบบทำนายฟังก์ชั่นเอนไซม์ที่มีการผสมผสานบริเวณจับและบริเวณเร่งที่ซับซ้อน

ในส่วนงานหลักแรกคือ การพัฒนาคุณภาพของข้อมูลที่มีอยู่น้อยมากเพื่อสร้างโมทีฟตัวแทนสายโปรตีนในการแสดงคุณสมบัติฟังก์ชั่นเอนไซม์ที่มีคุณภาพนั้น สามารถดำเนินการได้ใน 2 แนวทางควบคู่กัน คือการหาขั้นตอนวิธีในการพัฒนาคุณภาพองค์ประกอบย่อยของโมทีฟที่เรียกว่ากลุ่มแทนที่กรดอะมิโน (substitution group) และการพัฒนาคุณภาพของชุดข้อมูลบลอกที่ใช้ค้นพบกลุ่มแทนที่

ในการพัฒนาคุณภาพของกลุ่มแทนที่นั้นประยุกต์ใช้เทคนิคคอนเซ็ปต์แลททิส หรือ Concept Lattice (CL) (Wille, 1982) ในการเชื่อมโยงความรู้พื้นหลัง (background knowledge) ที่ได้จากงานวิจัยอื่น เช่น BLOSUM62 (Henikoff and Henikoff, 1992) หรือ แผนภาพ Taylor-Venn's diagram (Taylor, 1986) เป็นต้น และเชื่อมโยงความรู้เชิงวิทยาศาสตร์ (scientific knowledge) ด้านกลไกการเกิดฟังก์ชั่นเอนไซม์ที่เรียกว่า การควบคุมการกลายพันธุ์ (mutation control) โดยเชื่อมประสานความรู้เหล่านี้เข้ากับขั้นตอนการค้นพบกลุ่มแทนที่หรือองค์ประกอบย่อยของโมทีฟจากข้อมูลบลอก ทำให้ได้องค์ประกอบย่อยโมทีฟที่ได้มีคุณภาพที่ดีขึ้น และเนื่องจากการควบคุมการกลายพันธุ์เป็นหัวใจหลักของการพัฒนาคุณภาพกลุ่มแทนที่ ดังนั้นจึงเรียกส่วนงานย่อยนี้ว่า งานควบคุมการกลายพันธุ์ (mutation control task)

อย่างไรก็ตาม การเพิ่มคุณภาพของกลุ่มแทนที่เพียงอย่างเดียวไม่เพียงพอต่อการประยุกต์ใช้กับข้อมูลที่มีน้อยมากๆ ได้ เช่น บลอกที่มีข้อมูลบริเวณจับหรือบริเวณเร่งอยู่เพียง 1 ระเบียน ดังนั้นจึงได้พัฒนาเทคนิคการเพิ่มคุณภาพบลอกให้สามารถใช้ร่วมกับงานควบคุมการกลายพันธุ์ได้อย่างมีประสิทธิภาพ ประกอบด้วย งานคัดกรองบลอกที่มีคุณภาพ (block scan filtering) เป็นขั้นตอนการเตรียมการก่อนงานควบคุมการกลายพันธุ์ และงานจัดกลุ่มโมทีฟเชิงปฏิกิริยา (reactive motif – group definition) ที่เป็นขั้นตอนจัดการหลังส่วนงานควบคุมการกลายพันธุ์ โดยผลของการพัฒนาคุณภาพทั้งในส่วนบลอกและกลุ่มแทนที่กรดอะมิโนดังกล่าวนี้ ได้ออกมาเป็นโมทีฟชนิดใหม่ที่เรียกว่า รีแอคทีฟโมทีฟ (reactive motif)

สำหรับส่วนงานหลักที่ 2 คือ การสร้างตัวแบบทำนายฟังก์ชั่นเอนไซม์ที่มีการผสมผสานบริเวณจับและบริเวณเร่งที่ซับซ้อน โดยใช้เครื่องมือที่มีอยู่แล้วและนิยมใช้กันในปัจจุบันนั่นคือ

เครื่องจักรเรียนรู้ (machine learning) ประเภท C4.5 (Quinlan, 1993) ซึ่งมีจุดเด่นเป็นตัวแบบระบบทำนายที่ใช้โครงสร้างแบบต้นไม้ (tree) ที่ค่อนข้างรวดเร็วทั้งในขั้นตอนการสร้างตัวแบบและการใช้ทำนายประเภทของฟังก์ชั่นเอนไซม์ โดยใช้รีแอคทีฟโมทีฟที่ได้จากส่วนงานหลักแรกเป็นคุณลักษณะเด่นของกลุ่มสายโปรตีนที่เป็นข้อมูลนำเข้าในการสร้างตัวแบบทำนายฟังก์ชั่นเอนไซม์และทดสอบความแม่นยำในการทำนายฟังก์ชั่นเอนไซม์ ซึ่งนอกจากการใช้ประโยชน์ในเชิงทำนายฟังก์ชั่นเอนไซม์แล้ว ในอีกทางหนึ่งก็คือการประเมินคุณภาพของรีแอคทีฟโมทีฟที่ได้จากงานวิจัยนี้

จากความสำคัญดังที่กล่าวมานี้ จึงเป็นที่มาของวิทยานิพนธ์หัวข้อ “การพัฒนาตัวแทนสายโปรตีนสำหรับการทำนายประเภทฟังก์ชั่นเอนไซม์” โดยเป้าหมายหลักคือการประยุกต์ความรู้พื้นหลังและความรู้ในเชิงกลไกการเกิดฟังก์ชั่นเอนไซม์เข้าไปในเทคนิคการสร้างตัวแทนสายโปรตีนสำหรับทำนายประเภทฟังก์ชั่นโปรตีน

วัตถุประสงค์

เพื่อศึกษารูปแบบลำดับของกรดอะมิโนในสายโปรตีนหรือตัวแทนสายโปรตีนที่เป็นคุณลักษณะเด่นของฟังก์ชั่นเอนไซม์ โดยมีวัตถุประสงค์หลักแยกเป็นหัวข้อดังนี้

1. เพื่อพัฒนาตัวแทนสายโปรตีนโดยประยุกต์ใช้ “ความรู้” ลักษณะต่างๆ เช่น ทฤษฎีทางชีววิทยา ปฏิกิริยาชีวเคมี (biochemical reaction) และความรู้พื้นหลัง (background knowledge) ที่สัมพันธ์กับฟังก์ชั่นเอนไซม์

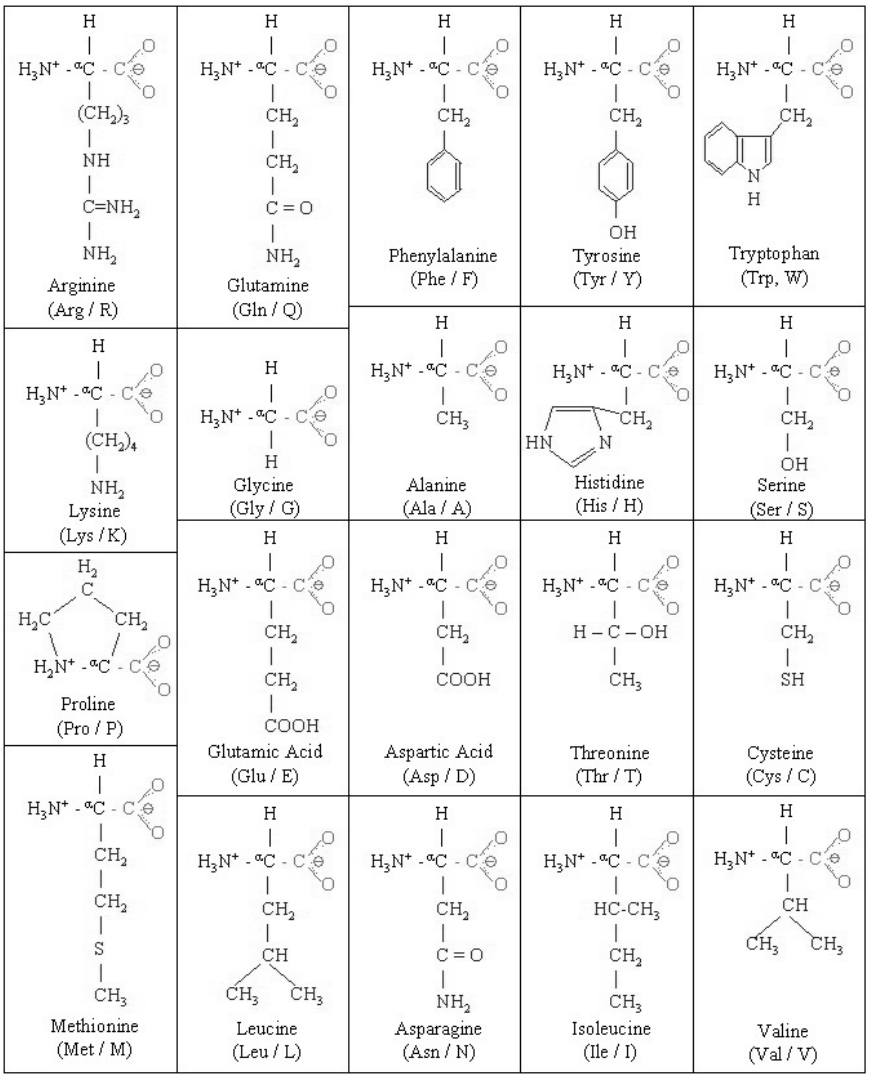
2. ประเมินคุณภาพของตัวแทนสายโปรตีนที่พัฒนาขึ้นนี้ รวมถึงประเมินคุณภาพการนำตัวแทนสายโปรตีนดังกล่าวเมื่อนำมาเป็นคุณลักษณะเด่นกลุ่มข้อมูลนำเข้าสำหรับสร้างโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์

3. พัฒนากรอบงานเชิงทฤษฎีในการพัฒนาตัวแทนสายโปรตีน

การตรวจเอกสาร

1. ข้อมูลโปรตีนและฟังก์ชั่นเอนไซม์

1.1 ข้อมูลสายโปรตีน



ภาพที่ 1 รายละเอียดของกรดอะมิโนแต่ละชนิดและอักษรย่อ

ที่มา: Wikipedia (2004)

โปรตีนประกอบขึ้นจากกรดอะมิโน 20 ชนิดที่เรียงต่อกันเป็นสายยาว (Lesk, 2004) โดยกรดอะมิโนแต่ละชนิดแทนด้วยอักษรย่อขนาด 3 ตัวอักษร และสามารถเขียนรหัสเป็นอักษร 1 ตัวคือ A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, และ Y แสดงรายละเอียดดังภาพที่ 1

โปรตีนแต่ละชนิดจะมีลำดับกรดอะมิโนไม่เหมือนกันเป็นเอกลักษณ์ ดังนั้นโปรตีนทุกชนิดทุกตัวจะต้องมีข้อมูลลำดับกรดอะมิโน โดยข้อมูลเพิ่มเติมอื่นๆ เกี่ยวกับโปรตีนเช่น แหล่งที่มาของโปรตีน ฟังก์ชั่นการทำงานของโปรตีน โครงสร้างโปรตีน ก็จะมีการอ้างอิงมาที่ลำดับกรดอะมิโนนี้ ซึ่งข้อมูลเหล่านี้มีการเก็บรวบรวมอยู่ในลักษณะที่ใช้งานได้ง่ายเช่น เป็นฐานข้อมูลบนเครือข่ายอินเตอร์เน็ท ทั้งนี้การเก็บข้อมูลโปรตีนที่สมบูรณ์ที่สุดเป็นความร่วมมือระหว่างหลายสถาบันที่เก็บรวบรวมข้อมูลโปรตีนที่เรียกกลุ่มตัวเองว่า UNIPROT (Apweiler, 2004) สามารถใช้งานฐานข้อมูลผ่านอินเตอร์เน็ทได้ที่ uniprot.org

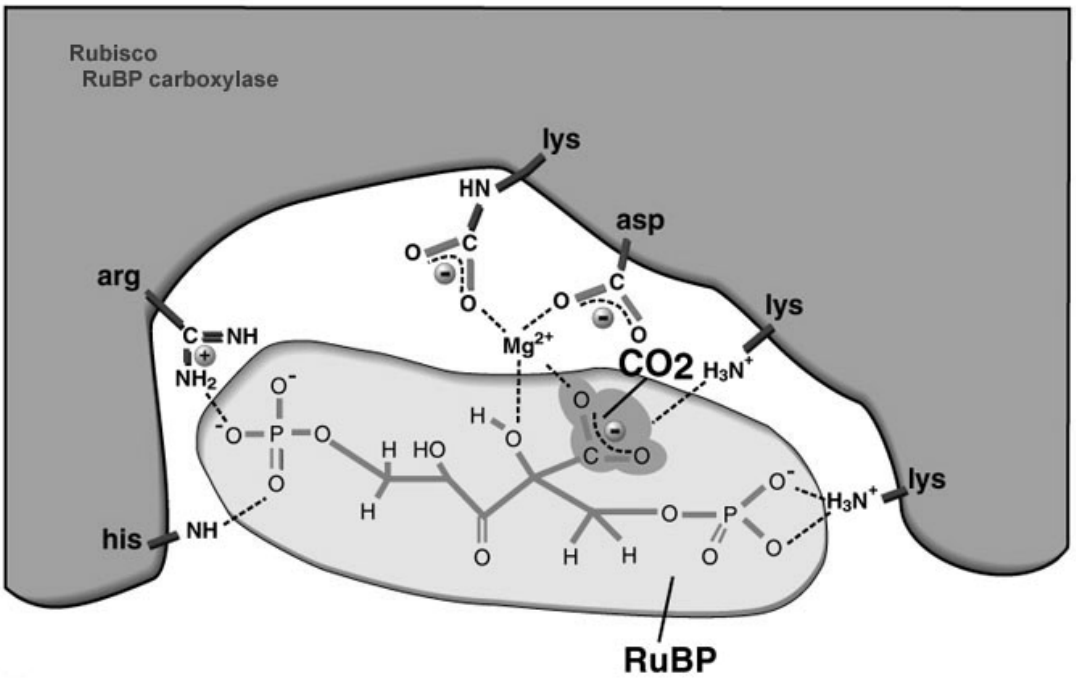
1.2 ฐานข้อมูลโปรตีนและฟังก์ชั่นโปรตีน

ปัจจุบันฐานข้อมูลโปรตีนที่สมบูรณ์ที่สุดคือฐานข้อมูลที่เกิดจากความร่วมมือจากหลายสถาบันที่เรียกว่า UNIPROT โดยการเก็บข้อมูลเกี่ยวกับฟังก์ชั่นโปรตีนมีพื้นฐานเริ่มต้นจากการนำผลการวิจัยของโปรตีนนั้นมาใส่ลงไปในฐานข้อมูล และมีการระบุคุณสมบัติสำคัญของโปรตีนเป็นคำสำคัญ (keywords) รวมทั้งรายละเอียดการทำงานของโปรตีนนั้น ซึ่งทำให้ผู้สนใจโดยทั่วไปสามารถคัดเลือกข้อมูลโปรตีนเฉพาะตัวที่สนใจออกมาใช้งานได้สะดวกได้ นอกจากนี้ยังมีฐานข้อมูลอื่นที่น่าสนใจ เช่น ฐานข้อมูลฟังก์ชั่นโปรตีนที่มีการจัดรวบรวมเป็นหมวดหมู่เป็นการเฉพาะ เช่น ฐานข้อมูลโปรตีนเอนไซม์ที่เวบไซต์ http://www.expasy.org/enzyme/ (Appel, 1994) ฐานข้อมูล Amino Acid-Nucleotide Interaction ที่เวบไซต์ http://aant.icmb.utexas.edu/ (Hoffman *et al.*, 2004) หรือ ระบบ ontology ของฟังก์ชั่นโปรตีนขึ้นมา เช่นที่ Gene Ontology ที่เวบไซต์ http://geneontology.org/ (The Gene Ontology Consortium, 2000) ฯลฯ

1.3 ฟังก์ชั่นเอนไซม์

ฟังก์ชั่นเอนไซม์เป็นฟังก์ชั่นโปรตีนประเภทหนึ่งที่สำคัญอย่างยิ่งในสิ่งมีชีวิตทุกชนิด ลักษณะพิเศษของฟังก์ชั่นเอนไซม์คือความสามารถในการเร่งปฏิกิริยาที่สิ่งมีชีวิตต้องการได้อย่างรวดเร็ว เช่น การให้พลังงาน การย่อยอาหาร ฯลฯ การทำงานฟังก์ชั่นเอนไซม์ใช้บางส่วนของ

โปรตีนเอนไซม์ที่เรียกว่าบริเวณจับ (binding site) และบริเวณเร่ง (catalytic site) โดยบริเวณจับคือบริเวณที่โปรตีนเอนไซม์ใช้เข้าจับกับสารตั้งต้น (substrates) ให้อยู่ในสภาพที่พร้อมเกิดฟังก์ชั่น ส่วนบริเวณเร่งคือบริเวณที่เอนไซม์ใช้เหนี่ยวนำให้เกิดปฏิกิริยาเคมีกับสารตั้งต้นให้กลายเป็นผลิตภัณฑ์ (products) (Patrick, 1995) ดังนั้นโปรตีนที่มีโครงสร้างหรือลำดับกรดอะมิโนที่คล้ายคลึงกัน แต่บริเวณจับและบริเวณเร่งทำงานต่างกัน จึงไม่จำเป็นที่จะต้องมีฟังก์ชั่นเอนไซม์เหมือนกัน ในขณะที่โปรตีนที่มีโครงสร้างหรือลำดับกรดอะมิโนต่างกันมาก แต่มีลำดับกรดอะมิโนที่ทำหน้าที่บริเวณจับและบริเวณเร่งเหมือนกัน ย่อมสามารถมีฟังก์ชั่นเอนไซม์ที่เหมือนกันได้ รายละเอียดการทำงานของเอนไซม์แสดงดังภาพที่ 2



ภาพที่ 2 กลไกในบริเวณจับและบริเวณเร่งของเอนไซม์ RuBP Carboxylase
ที่มา: Mallery (2007)

จากภาพที่ 2 แสดงการเข้าจับและเร่งปฏิกิริยาปลดปล่อย CO_2 จากสาร RuBP โดยมีบริเวณจับประกอบด้วยกรดอะมิโน arg และ his ด้านซ้ายของภาพ และ lys 2 ตัวทางขวาของภาพทำหน้าที่เป็นบริเวณจับกับสาร RuBP ด้วยแรงไอออนิกจากคุณสมบัติประจุบวก จากนั้นกรดอะมิโน asp จึงทำหน้าที่ร่วมกับโคเอนไซม์คือแมกนีเซียมในการถ่ายเทประจุเพื่อสลายพันธะของกลุ่มคาร์บอนไดออกไซด์ออกมาจากสาร RuBP

ดังนั้นงานด้านชีวสารสนเทศที่มุ่งเน้นศึกษาด้านเอนไซม์จึงไม่ได้มีเพียงความต้องการทำนายฟังก์ชั่นเอนไซม์ที่ถูกต้องแม่นยำเท่านั้น แต่หมายรวมไปถึงการใช้ข้อมูลจำนวนมากที่มีอยู่ในการศึกษาทำความเข้าใจการทำงานเอนไซม์ได้ดีขึ้น นั่นคือความสามารถในการระบุว่าส่วนใดของโปรตีนที่สามารถทำงานเป็นบริเวณจับหรือบริเวณเร่งที่ก่อให้เกิดกลไกการทำงานของฟังก์ชั่นเอนไซม์ได้อย่างสมเหตุผลน่าเชื่อถือในเชิงวิทยาศาสตร์ ซึ่งมีความสำคัญอย่างยิ่งต่อการนำไปใช้งานในระดับห้องปฏิบัติการ (หรือเรียกย่อว่า wet lab.) ในการทำความเข้าใจกลไกเอนไซม์และการออกแบบเอนไซม์ตัวใหม่

จากความสำคัญของเอนไซม์ ความซับซ้อนของฟังก์ชั่นเอนไซม์ และความต้องการในระดับห้องปฏิบัติการดังกล่าวในงานศึกษานี้จึงเน้นกลุ่มเป้าหมายคือโปรตีนเอนไซม์ และการพัฒนาตัวแทนสายโปรตีนที่ใช้ทำนายฟังก์ชั่นเอนไซม์ได้อย่างมีประสิทธิภาพ

2. ตัวแทนสายโปรตีน

2.1 ตัวแทนสายโปรตีนประเภทโมทีฟ

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 ยกตัวอย่าง เช่น โมทีฟ [ST]-x(1,2)-[RK] มีรูปแบบสายลำดับที่ยืดหยุ่นได้โดยตรงกลางของโมทีฟมีรูปแบบ x(1,2) หมายถึงการมีกรดอะมิโนใดใดจำนวน 1 หรือ 2 ตัวก็ได้ เป็นต้น โดยการยืดหยุ่นดังกล่าวนี้ ในทางชีววิทยาหมายถึงการแทรก (insertion) และการขาดหาย (deletion) ในสายวิวัฒนาการของโปรตีน

ดังนั้นองค์ประกอบพื้นฐานของโมทีฟจึงประกอบด้วยสิ่งที่เรียกว่า กลุ่มแทนที่กรดอะมิโน บริเวณอนุรักษ์ และ gap โดยแต่ละโมทีฟก็คือการเรียงลำดับกันของกลุ่มแทนที่ บริเวณอนุรักษ์ และ gap นั่นเอง

สำหรับโมทีฟที่มีอยู่ในปัจจุบันอาจแบ่งออกเป็น 2 ประเภทกว้างๆ คือ โมทีฟที่ค้นพบโดยผู้เชี่ยวชาญ และโมทีฟที่ค้นพบจากขั้นตอนวิธีอัตโนมัติ โดยจะได้ทำการตรวจเอกสารเทคนิคเหล่านี้เฉพาะเทคนิคที่สำคัญและมีเนื้อหาครอบคลุมต่อการกำหนดทิศทางงานพัฒนาโมทีฟในวิทยานิพนธ์นี้ ดังนี้

2.2 เทคนิคการพัฒนาตัวแทนสายโปรตีน

ในหัวข้อนี้กล่าวถึงเทคนิคการพัฒนาตัวแทนสายโปรตีน ประกอบด้วย โมทีฟที่ค้นพบและพัฒนาโดยผู้เชี่ยวชาญ คือ PROSITE (Bairoch, 1988) และโมทีฟที่ค้นพบจากขั้นตอนวิธีอัตโนมัติ ประกอบด้วย เทคนิค Multiple Sequence Alignment (MSA) (Waterman *et al.*, 1976) เทคนิค eMotif (Huang and Brutlag, 2001) และเทคนิค 3MOTIF (Bennett *et al.*, 2003) เป็นต้น

2.2.1 PROSITE (Bairoch, 1988) PROSITE (Bairoch, 1988)

โมทีฟที่ค้นพบโดยผู้เชี่ยวชาญที่รู้จักกันดีและมีคุณภาพดีที่สุดคือ PROSITE เป็นโมทีฟที่สั้นที่สร้างโดยผู้เชี่ยวชาญ เก็บไว้ในฐานข้อมูลที่สร้างขึ้นครั้งแรกโดย Amos Bairoch ในปี 1988 ซึ่งมีการจัดการปรับปรุงจนถึงปัจจุบันโดยสถาบัน expasy ปัจจุบันสามารถใช้งานข้อมูล PROSITE ได้ที่เวบไซต์ www.expasy.org

จุดเด่นของ PROSITE คือการใช้ความรู้ด้านชีวเคมีในระดับลึกในการค้นพบโมทีฟทำให้โมทีฟที่ได้มีความน่าเชื่อถือในระดับห้องปฏิบัติการสูงมาก แต่ข้อด้อยก็คือการค้นพบโมทีฟจะต้องใช้ “ผู้เชี่ยวชาญ” เท่านั้นซึ่งทำได้ค่อนข้างช้า ดังนั้นในปัจจุบันโมทีฟที่สร้างโดย PROSITE จึงมีเพียง 1,639 รูปแบบ เท่านั้น โดยเป็นฟังก์ชั่นทางปฏิกิริยาชีวเคมีเพียง 188 รูปแบบ เมื่อเทียบกับเอนไซม์ที่มี 4,361 ฟังก์ชั่น

2.2.2 Multiple Sequence Alignment

สำหรับขั้นตอนวิธีอัตโนมัติที่เป็นเทคนิคพื้นฐานและนิยมใช้ค้นพบโมทีฟจากกลุ่มโปรตีนเรียกว่า Multiple Sequence Alignment (MSA) ซึ่งเป็นเทคนิคที่พัฒนามาอย่างต่อเนื่องยาวนาน ยกตัวอย่างเช่น งานของ Waterman *et al.* (1976) งานของ Feng and Doolittle (1987) และงานของ Barton (1990) เป็นต้น

ในการทำ MSA มีหลักการที่สำคัญคือการจัดเรียงลำดับสายโปรตีนของกลุ่มโปรตีนให้อยู่ในตำแหน่งที่มีค่าคะแนนความเหมือน (similarity) มากที่สุด โดยใช้ตารางคะแนนความเหมือน (similarity score table) ประเภทต่างๆ เช่น PAM (Dayhoff *et al.*, 1978) และ BLOSUM ดังแสดงตัวอย่าง PAM ในภาพที่ 3 เข้ามาช่วยในการคำนวณ โดยองค์ประกอบของตารางประเภทนี้ประกอบด้วยคะแนนความเหมือนที่ได้จากการจับคู่กันของกรดอะมิโนแต่ละประเภทรวมทั้ง gap (ใช้สัญลักษณ์ *) โดยจากภาพที่ 3 ได้ค่าคะแนนความเหมือนของกรดอะมิโน Y และ F เท่ากับ 7 คะแนน

ในการจัดเรียงลำดับสายโปรตีนของกลุ่มโปรตีนให้อยู่ในตำแหน่งที่มีค่าคะแนนความเหมือนมากที่สุดอธิบายได้จากตัวอย่างในภาพที่ 4 แสดงการจัดเรียงสายโปรตีนสายสั้น 4 สายคือ NYLS, NKYLS, NFS, และ NFLS

จากภาพที่ 4 สัญลักษณ์ ★ หมายถึงบริเวณอนุรักษ์ โดยผลในการทำ MSA ได้โมทีฟของกลุ่มโปรตีนคือ N-x(0,1)-[YF]-x(0,1)-S เป็นลักษณะเฉพาะร่วมกันของโปรตีนในกลุ่มนี้ โดยรูปแบบสายลำดับที่ได้มีชื่อเรียกหลากหลายตามวิธีได้มาที่แตกต่างกันเช่น consensus sequence (Taylor, 1986) ที่ได้จาก global alignment หรือ motif ที่ได้จาก local alignment เป็นต้น

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0	0	0	0	-8
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2	-1	0	-1	-8
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2	2	1	0	-8
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2	3	3	-1	-8
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2	-4	-5	-3	-8
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2	1	3	-1	-8
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2	3	3	-1	-8
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1	0	0	-1	-8
H	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2	1	2	-1	-8
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4	-2	-2	-1	-8
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2	-3	-3	-1	-8
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2	1	0	-1	-8
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2	-2	-2	-1	-8
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1	-4	-5	-2	-8
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1	-1	0	-1	-8
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1	0	0	0	-8
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0	0	-1	0	-8
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6	-5	-6	-4	-8
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2	-3	-4	-2	-8
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4	-2	-2	-1	-8
B	0	-1	2	3	-4	1	3	0	1	-2	-3	1	-2	-4	-1	0	0	-5	-3	-2	3	2	-1	-8
Z	0	0	1	3	-5	3	3	0	2	-2	-3	0	-2	-5	0	0	-1	-6	-4	-2	2	3	-1	-8
X	0	-1	0	-1	-3	-1	-1	-1	-1	-1	-1	-1	-1	-2	-1	0	0	-4	-2	-1	-1	-1	-1	-8
*	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	1

ภาพที่ 3 ค่าคะแนนความยากง่ายการแทนที่กรดอะมิโนในสายวิวัฒนาการของ PAM250 ที่ได้จากโปรแกรม BioEdit version 5.0.9 (Hall, 1999)

```
N-YLS
NKYLS
N-F-S
N-FLS
★   ★
```

ภาพที่ 4 การทำ Multiple Sequence Alignment ของสายโปรตีนสายสั้น 4 สาย

2.2.3 eMOTIF (Huang and Brutlag, 2001)

สำหรับ eMOTIF เป็นเทคนิคการค้นพบโมทีฟจากฐานข้อมูล BLOCKS (Henikoff and Henikoff, 1991) ซึ่งฐานข้อมูล BLOCKS เป็นชุดโครงสร้างข้อมูลที่ได้จากเทคนิค MSA ดังจะได้กล่าวถึงรายละเอียดในภายหลัง โดยหลักการพื้นฐานของ eMOTIF อธิบายดังในภาพที่ 5

ภาพที่ 5 ขั้นตอนวิธีในการค้นพบ eMOTIF

จากภาพที่ 5 แบ่งภาพออกเป็นสองส่วนคือ ในภาพด้านซ้ายเป็นความรู้พื้นหลังที่ใช้ประมวลผลกลุ่มแทนที่ซึ่งเป็นองค์ประกอบย่อยที่สำคัญของโมทีฟ ในภาพด้านขวาในส่วน a คือเมทริกซ์ของสายลำดับกรดอะมิโนที่มีความสัมพันธ์กันสูงที่ประมวลผลจากเทคนิค MSA และเก็บไว้ในฐานข้อมูล BLOCKS โดยในภาพด้านขวาในส่วน b และ c คือโมทีฟที่ไม่ขัดแย้งกับความรู้พื้นหลังที่แสดงไว้ในภาพด้านซ้าย

พิจารณาจากข้อมูล BLOCKS ในภาพด้านขวาส่วน a ในคอลัมน์ที่ 2 ของข้อมูลทุกแถว พบเฉพาะกรดอะมิโน F และ Y ในตำแหน่งนี้ ซึ่งเมื่อพิจารณาประกอบความรู้พื้นหลังในภาพด้านซ้าย พบว่ากลุ่มแทนที่ที่เป็นไปได้ก็คือ {F} {Y} {F,Y} {F,Y,W} {F,L,Y} {F,I,L,V,Y} และ {F,I,L,M,V,Y} โดยการประมวลผล eMOTIF จะประมวลกลุ่มแทนที่ที่เป็นไปได้ทั้งหมดในทุกคอลัมน์ของ BLOCKS แล้วนำมาประกอบเข้าด้วยกันเป็นโมทีฟในทุกกรณีที่เป็นไปได้ จากนั้น

นำมาตรวจสอบความแม่นยำในการใช้งานโมทีฟผ่านฐานข้อมูลโปรตีน ถ้าโมทีฟใดมีความแม่นยำในการทำนายประเภทโปรตีนผ่านเกณฑ์ที่กำหนดไว้ก็จะถูกจัดเก็บไว้ในฐานข้อมูล eMOTIF ส่วนโมทีฟที่ไม่ผ่านเกณฑ์ความแม่นยำการใช้งานก็จะถูกคัดทิ้งออกไปจากระบบ ทั้งนี้ในภาพด้านขวาส่วน b และ c เป็นตัวอย่างของโมทีฟที่ผ่านเกณฑ์ความแม่นยำทั้ง 2 โมทีฟ

จากขั้นตอนวิธีดังกล่าว ทำให้แต่ละ BLOCKS สามารถประมวลผลได้โมทีฟออกมาจำนวนมากเก็บไว้ในฐานข้อมูล eMOTIF ถึงประมาณ 170,000 โมทีฟ

2.2.4 3MOTIF (Bennett *et al.*, 2003)

สำหรับเทคนิค 3MOTIF เป็นการพัฒนาเทคนิคต่อเนื่องจาก eMOTIF ทั้งนี้เนื่องจาก eMOTIF มีโมทีฟถึงประมาณ 170,000 โมทีฟ ซึ่งมากเกินความจำเป็นในการใช้งานทำนายฟังก์ชั่นโปรตีนเอนไซม์ที่มีเพียง 70,000 ชนิด ดังนั้นเพื่อคัดกรองโมทีฟที่ไม่สัมพันธ์กับฟังก์ชั่นโปรตีนออกไป เทคนิค 3MOTIF จึงเลือกใช้ข้อมูลเฉพาะในบริเวณที่เป็นโครงสร้าง 3 มิติ (3D-Structure) ที่มี eMOTIF ปรากฏอยู่เท่านั้น โดยมีสมมติฐานเบื้องต้นว่าฟังก์ชั่นโปรตีนเกี่ยวข้องกับรูปโครงสร้าง 3 มิติของโปรตีนอย่างแนบแน่น

อย่างไรก็ตาม พื้นฐานเทคนิคของ 3MOTIF ยังคงอิงฐานข้อมูล BLOCKS ซึ่งจะได้กล่าวถึงข้อจำกัดของการใช้งาน BLOCKS อย่างละเอียดอีกครั้งในภายหลัง

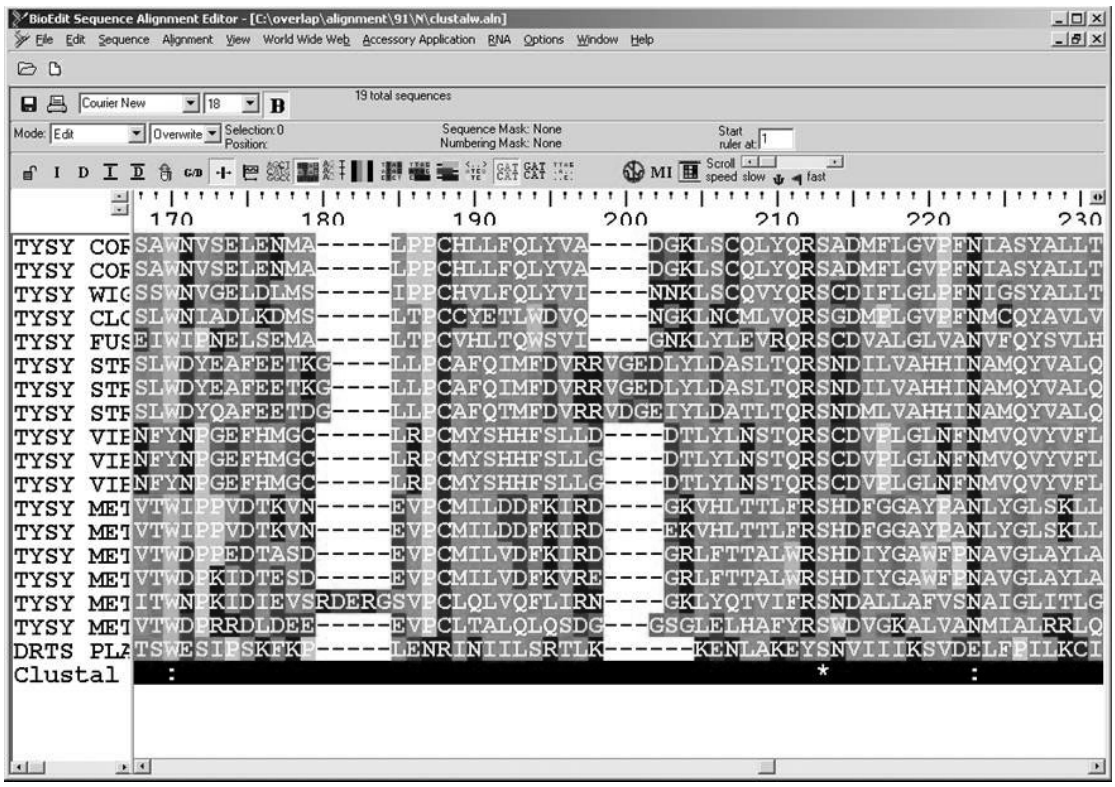
นอกจากเทคนิคค้นพบโมทีฟโดยผู้เชี่ยวชาญ คือ PROSITE และเทคนิคค้นพบโมทีฟจากขั้นตอนวิธีอัตโนมัติ คือ MSA eMOTIF และ 3MOTIF ดังที่ยกตัวอย่างมานี้แล้ว ยังมีเทคนิคอื่นที่ค้นพบโมทีฟด้วยขั้นตอนวิธีอัตโนมัติ เช่น การค้นพบโมทีฟแบบไม่มี gap หรือในงานของ Attasena and Waiyamai (2007) Rattanakornkul *et al.* (2003) และ ธนพัฒน์ และคณะ (2548) ในการค้นพบโมทีฟแบบมี gap ดังนั้นโมทีฟจึงมีมากมายหลายประเภท ซึ่งจะได้ตรวจสอบว่าโมทีฟประเภทใดที่เหมาะสมใช้เป็นตัวแทนสายโปรตีนสำหรับพัฒนาระบบทำนายประเภทฟังก์ชั่นเอนไซม์ในหัวข้อต่อไป

3. ลักษณะเฉพาะของโมทีฟและโครงสร้างข้อมูลที่สัมพันธ์กับฟังก์ชั่นเอนไซม์

3.1 การเลือกใช้โมทีฟเป็นตัวแทนสายโปรตีนสำหรับทำนายประเภทฟังก์ชั่นเอนไซม์

โมทีฟเป็นรูปแบบลำดับสายโปรตีนสั้นๆ ที่สามารถแสดงเอกลักษณ์ของกลุ่มโปรตีนแต่ละชุด โดยเอกลักษณ์ดังกล่าวอาจเป็นฟังก์ชั่น โครงสร้าง หรือบริเวณอนุรักษ์ในสายวิวัฒนาการ (conserved region) และเนื่องจากโปรตีนแต่ละชนิดล้วนมีคุณสมบัติที่หลากหลายผสมผสานกัน ดังนั้นโมทีฟที่ได้จากโปรตีนแต่ละชุดจึงไม่สามารถระบุได้ว่าโมทีฟของโปรตีนชุดนั้นเป็นการแสดงออกของเอกลักษณ์ใด และมีแนวโน้มที่จะได้โมทีฟแสดงคุณสมบัติที่ไม่ซับซ้อน เช่น โมทีฟแสดงลำดับวิวัฒนาการ เป็นต้น รวมทั้งอาจรบกวนกันเองจนมีผลต่อการค้นพบโมทีฟดังแสดงในภาพที่ 6

จากภาพที่ 6 นี้ ใช้โปรแกรม BioEdit (Hall, 1999) เป็นเครื่องมือในการค้นพบโมทีฟของกลุ่มโปรตีนที่มีฟังก์ชั่น Thymidylate synthase ด้วยเทคนิค multiple sequence alignment โดยมีชื่อของโปรตีนปรากฏในเฟรมทางซ้ายมือและผลการ alignment ทางขวามือ และมี consensus sequence ที่เป็นโมทีฟชนิดหนึ่งปรากฏในบรรทัดสุดท้าย ทั้งนี้โปรตีนที่มีชื่อขึ้นต้นเหมือนกันหมายถึงโปรตีนที่ได้จากสายวิวัฒนาการเดียวกัน โดยในตัวอย่างนี้มีกลุ่มโปรตีนที่ชื่อขึ้นต้นด้วย TYSY ถึง 17 ชนิด และมีกลุ่มโปรตีนที่ชื่อขึ้นต้นด้วย DRTS เพียง 1 ชนิด ซึ่งจากภาพจะพิจารณาได้อย่างชัดเจนว่าโปรตีนในกลุ่ม TYSY มีความคล้ายคลึงกันสูงทำให้มีแนวโน้มการค้นพบโมทีฟที่เป็นตัวแทนด้านสายวิวัฒนาการของโปรตีน โดยความคล้ายคลึงกันนี้สามารถสังเกตได้จากการมีตำแหน่งตรงกันของรหัสกรดอะมิโนและโซนสีที่แสดงความคล้ายคลึงกันของกรดอะมิโน อย่างไรก็ตามเมื่อได้พยายามค้นพบโมทีฟที่แสดงคุณสมบัติฟังก์ชั่นเอนไซม์โดยเพิ่มข้อมูลโปรตีนตัวที่ 18 ที่มีฟังก์ชั่นเดียวกันแต่ได้จากสายวิวัฒนาการอื่นคือโปรตีนที่ชื่อขึ้นต้นด้วย DRTS สิ่งที่เกิดขึ้นก็คือ consensus sequence หรือโมทีฟที่ได้ไม่สามารถนำไปใช้งานได้ เนื่องจากการรบกวนกันของคุณสมบัติด้านสายวิวัฒนาการที่แตกต่างกัน



ภาพที่ 6 ปัญหาการรบกวนจากข้อมูลบริเวณอนุรักษ์ในสายวิวัฒนาการที่มีต่อการค้นพบโมทีฟที่เป็นตัวแทนสายโปรตีนของกลุ่มโปรตีนที่มีฟังก์ชั่นเหมือนกัน

หมายเหตุ ใช้โปรแกรม BioEdit ประมวลผลข้อมูลที่เตรียมจาก UNIPROT และ Prosite

★ หมายถึงในตำแหน่งนั้นมีกรดอะมิโนชนิดที่ตรงกันในโปรตีนทุกตัว

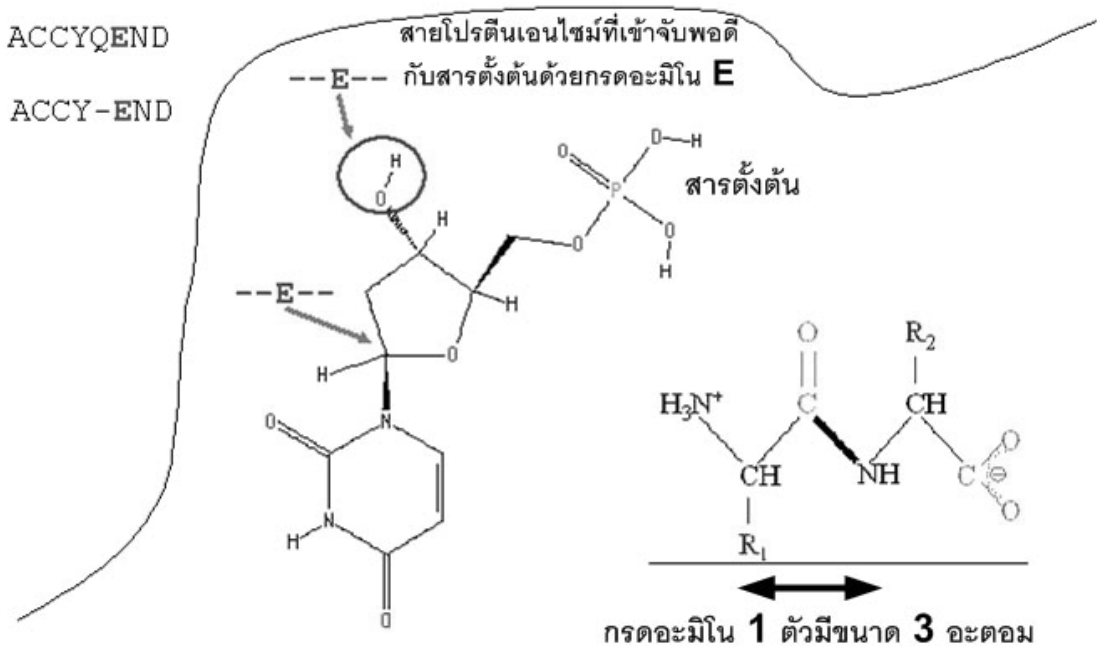
: หมายถึงในตำแหน่งนั้นมีกรดอะมิโนชนิดที่ตรงกันเป็นส่วนใหญ่

ดังนั้นการค้นพบโมทีฟที่แสดงคุณสมบัติฟังก์ชั่นเอนไซม์จึงควรได้มาจากการจัดกลุ่มชุดข้อมูลสายโปรตีนที่เกี่ยวข้องกับการทำงานฟังก์ชั่นเอนไซม์โดยตรง นั้นคือข้อมูลในส่วนบริเวณจับและบริเวณเร่ง รวมทั้งประเภทของโมทีฟที่ใช้ควรมีความเหมาะสมสอดคล้องกับลักษณะเฉพาะการเกิดฟังก์ชั่นเอนไซม์

จากเหตุผลที่โปรตีนต่างสายวิวัฒนาการสามารถมีรูปร่างโครงสร้างโปรตีนที่แตกต่างกันแต่สามารถทำงานฟังก์ชั่นเอนไซม์เหมือนกันได้เมื่อมีบริเวณจับและบริเวณเร่งที่ทำงานเหมือนกัน ดังนั้นงานวิจัยนี้จึงสนใจโมทีฟจากบริเวณจับและบริเวณเร่งที่มีขนาดความยาวไม่มากนักแต่ควรมีขนาดที่สอดคล้องกับสารตั้งต้นส่วนใหญ่ของฟังก์ชั่นเอนไซม์ ซึ่งจากการประมาณ

ขนาดของสารตั้งต้นในฐานข้อมูล BRENDA (Schomburg *et al.*, 2004) ได้ค่าเฉลี่ยของขนาดสารตั้งต้นที่ประมาณ 21 อะตอม ซึ่งเมื่อคำนวณเทียบกับกรดอะมิโน 1 ตัวที่มีขนาด 3 อะตอม โมทีฟจากบริเวณจับและบริเวณเร่งที่สนใจจึงมีขนาดความยาวบวกลบ 7 ลำดับกรดอะมิโน ได้เป็นขนาด 15 ลำดับกรดอะมิโน

นอกจากนี้การทำงานของเอนไซม์คือการเข้าจับพอดีกับสารตั้งต้นและเหนี่ยวนำให้เกิดฟังก์ชั่นหรือ “fit and function” ดังนั้นโมทีฟที่เหมาะสมกับฟังก์ชั่นเอนไซม์จึงควรมีลักษณะที่พอดีหรือไม่ยืดหยุ่น (no gap) เนื่องจากการมี gap จะทำให้บริเวณจับหรือบริเวณเร่งเกือบทั้งหมดมีความไม่พอดีกับสารตั้งต้นและไม่สามารถทำงานกลไกฟังก์ชั่นเอนไซม์ได้ โดยสามารถทำความเข้าใจเหตุผลดังกล่าวได้จากภาพที่ 7



ภาพที่ 7 ผลกระทบจาก gap ที่มีต่อการทำงานฟังก์ชั่นเอนไซม์

จากภาพที่ 7 เส้นยาวโค้งไปมาที่เห็นนี้สมมติให้เป็นสายโปรตีนที่เข้าจับกับสารเป้าหมาย โดยสมมติให้ส่วนที่โปรตีนเอนไซม์เข้าจับสารเป้าหมายคือส่วนสายลำดับ ACCYQEND โดยให้กรดอะมิโนที่มีรหัส E เป็นกรดอะมิโนตัวสำคัญที่เข้าจับกับหมู่ OH ของสารตั้งต้นดังแสดงในวงกลมสีแดง ซึ่งถ้าหากใช้ตัวแทนสายโปรตีนในบริเวณนั้นด้วยโมทีฟที่มี gap แทนที่ลงไปในตำแหน่งของกรดอะมิโนรหัส Q สิ่งที่เกิดขึ้นก็คือกรดอะมิโน E สามารถเลื่อนเข้ามาได้ 1 ตำแหน่ง

ซึ่งกรดอะมิโน 1 ตัวมีโครงสร้างแกนกลางของโมเลกุลขนาด 3 อะตอม (N-C-C หรือ C-C-N) หมายความว่าการเลื่อนเข้ามา 1 ตำแหน่งของกรดอะมิโน มีผลทำให้ E เปลี่ยนตำแหน่งไปถึง 3 อะตอม ซึ่งเป็นตำแหน่งที่ไม่สามารถเข้าจับกับสารตั้งต้นเพื่อเกิดฟังก์ชั่นเอนไซม์ได้

การอธิบายดังกล่าวได้รับการพิสูจน์อย่างชัดเจนในงานสร้างระบบทำนายฟังก์ชั่นเอนไซม์จากโมทีฟที่ได้จากบริเวณจับและบริเวณเร่ง (พีระ และ กฤษณะ, 2549) ที่โมทีฟประเภทไม่มี gap ให้ผลการทำนายที่แม่นยำเฉลี่ยถึง 76.6% ดีกว่าโมทีฟประเภทมี gap ที่ให้ผลเพียง 50.8% ในกลุ่มฟังก์ชั่นเอนไซม์ที่มีข้อมูลบริเวณจับและบริเวณเร่งที่สมบูรณ์มากที่สุดจำนวน 5 ฟังก์ชั่น

ดังนั้นโมทีฟที่เหมาะสมเป็นตัวแทนสายโปรตีนจากข้อมูลบริเวณจับและบริเวณเร่งในวิทยานิพนธ์นี้จึงเลือกใช้โมทีฟที่ขนาดความยาว 15 กรดอะมิโนและเป็นโมทีฟประเภทไม่มี gap ซึ่งโครงสร้างข้อมูลที่ใช้ในการค้นพบโมทีฟประเภทนี้ก็คือโครงสร้างแบบ BLOCKS หรือที่เรียกขึ้นใหม่เพื่อใช้กับงานศึกษานี้โดยเฉพาะว่าบลอก (block)

3.2 บลอก: โครงสร้างชุดข้อมูลสำหรับค้นพบโมทีฟ

BLOCKS (Henikoff and Henikoff, 1991) เป็นฐานข้อมูลที่รวบรวมกลุ่มสายลำดับโปรตีนขนาดสั้น (substring) ที่บ่งชี้ถึงการมีเอกลักษณ์เดียวกัน เช่น ส่วนที่มีฟังก์ชั่นเดียวกัน ส่วนที่เป็นสายวิวัฒนาการเดียวกัน เป็นต้น โดย BLOCKS เป็นชุด substring ที่ได้จากเทคนิค MSA แล้วเก็บรวบรวมไว้เป็นชุดข้อมูล ตัวอย่างของชุดข้อมูลใน BLOCKS แสดงให้เห็นดังในภาพที่ 8

```
GUB_CLOTM|P29716     ( 272) GDVNGDGHVNSSDYSLFKRYLL  18
GUNA_CLOTM|P04955    ( 416) GDVNGDGNVNSTDLTMLKRYLL  10
GUNB_CLOTM|P04956    ( 501) GDVNGDGRVNSSDVALLKRYLL  12
GUND_CLOTM|P04954    ( 584) GDVNDDGKVNSTDLTLLKRYVL  10
GUNE_CLOTM|P10477    ( 414) GDVNGDGKINSTDCTMLKRYIL  15
GUNS_CLOTM|P38686    ( 678) GDVNDDGKVNSTDAVALKRYVL  12
GUNX_CLOTM|P15329    ( 167) GDVNLDGQVNSTDFSLLKRYIL  13
XYNY_CLOTM|P51584    ( 733) GDVNGDGTINSTDLTMLKRSVL  18
```

ภาพที่ 8 ชุดข้อมูลบางส่วนของ dockerin type I รหัส IPB002105B ใน BLOCKS

ที่มา: ฐานข้อมูล BLOCKS จากเวบไซต์ http://blocks.fhcrc.org/ (Henikoff and Henikoff, 1991)

จากภาพที่ 8 แต่ละบรรทัดประกอบด้วยข้อมูล 4 ส่วนคือ ชื่อและรหัสโปรตีน, ตำแหน่งของโปรตีนที่เริ่มต้นสายลำดับกรดอะมิโนที่เป็นเอกลักษณ์ของกลุ่ม, สายลำดับกรดอะมิโนที่เป็นเอกลักษณ์ของกลุ่ม, และ ค่าระยะทาง (distance) ที่บ่งชี้ความแตกต่างของสายลำดับโปรตีนในกลุ่มโปรตีนนั้น

อย่างไรก็ตาม ชุดข้อมูลเอนไซม์ที่มีใน BLOCKS มีน้อยมาก โดยข้อมูลบริเวณจับและบริเวณเร่งมีอยู่เพียง 6 ชนิด รวมทั้งหมด 39 ชุดข้อมูลในฐานข้อมูล BLOCKS ดังนั้นงานศึกษาในวิทยานิพนธ์นี้จึงนำข้อมูลบริเวณจับและบริเวณเร่งจากฐานข้อมูลของ UNIPROT มาสร้างชุดข้อมูลขึ้นใหม่ เรียกชุดข้อมูลเหล่านี้ว่า บลอก หรือ block

โครงสร้างข้อมูลของ BLOCKS หรือบลอกนี้ เป็นประโยชน์อย่างยิ่งในงานค้นพบองค์ประกอบย่อยของโมทีฟคือ กลุ่มแทนที่กรดอะมิโน เพื่อค้นพบโมทีฟประเภทไม่มี gap รายละเอียดดังแสดงในหัวข้อต่อไป

3.3 การค้นพบกลุ่มแทนที่กรดอะมิโนจากบลอก และความรู้พื้นหลัง

จากชุดข้อมูลที่มีโครงสร้างแบบบลอกสามารถค้นพบโมทีฟที่เป็นตัวแทนสายโปรตีนของบลอกได้จากแต่ละตำแหน่งคอลัมน์ของบลอกนั้น โดยชนิดของกรดอะมิโนที่ปรากฏอยู่ในคอลัมน์เดียวกันหมายถึงความสามารถในการแทนที่กันได้ จึงสามารถจัดให้อยู่ในกลุ่มแทนที่กรดอะมิโน (substitution group) เดียวกัน ซึ่งแต่ละคอลัมน์ของบลอกจะได้กลุ่มแทนที่กรดอะมิโนที่เป็นองค์ประกอบย่อยของโมทีฟเรียงลำดับกันออกมาเป็นโมทีฟตัวแทนสายโปรตีนของบลอกนั้น ตัวอย่างของโมทีฟที่ได้จากบลอกแสดงดังภาพที่ 9

คอลัมน์ที่	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	H	T	S	P	H	R	P	R	F	S	P	A	T	H	P
	H	E	T	P	H	K	P	R	F	S	P	D	E	N	P
	T	S	S	P	H	T	D	R	V	E	Q	S	P	V	I
โมทีฟ	HT	TES	ST	P	H	RKT	PD	R	FV	SE	PQ	ADS	TEP	HNV	PI

ภาพที่ 9 โมทีฟที่ได้จากบลอก

จากภาพที่ 9 แสดงการค้นพบกลุ่มแทนที่กรดอะมิโนจากบลอกออกมาเป็นโมทีฟ อย่างไรก็ตามกลุ่มแทนที่ที่ได้จากบลอกที่มีสมาชิกของสายลำดับกรดอะมิโนอยู่น้อยย่อมได้กลุ่มแทนที่ที่ไม่สมบูรณ์ ดังเช่นในตัวอย่างที่มีสายลำดับกรดอะมิโนเพียง 3 เส้นในบลอก ย่อมได้กลุ่มแทนที่กรดอะมิโนที่มีสมาชิกเป็นกรดอะมิโนได้ไม่เกิน 3 ชนิด ในขณะที่อาจมีกรดอะมิโนชนิดอื่นที่สามารถเป็นสมาชิกในกลุ่มแทนที่ได้อีกทำให้โมทีฟที่ได้ไม่สมบูรณ์เพียงพอที่จะนำไปใช้งานได้จริง ดังนั้นเพื่อให้ได้กลุ่มแทนที่กรดอะมิโนที่มีคุณภาพที่ดีขึ้น จึงมีความพยายามใช้ความรู้ในลักษณะต่างๆ เข้ามาช่วยค้นพบกลุ่มแทนที่กรดอะมิโนซึ่งสามารถแบ่งความรู้ดังกล่าวออกเป็น 2 ลักษณะคือ ความรู้จากผู้เชี่ยวชาญ และความรู้พื้นหลัง (background knowledge) ที่ได้จากงานวิจัยอื่น

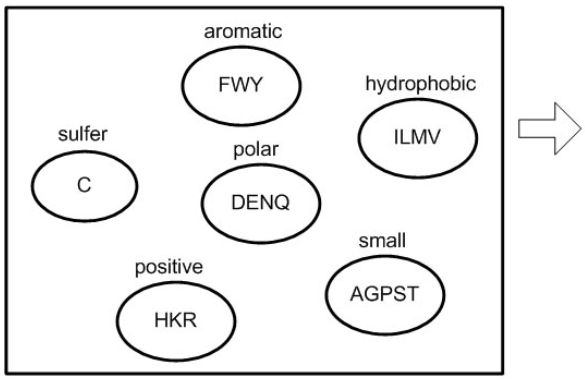
ตัวอย่างการใช้ความรู้จากผู้เชี่ยวชาญในการค้นพบกลุ่มแทนที่และโมทีฟที่รู้จักกันดีก็คือ PROSITE ซึ่งทำให้โมทีฟที่ได้จาก PROSITE มีความน่าเชื่อถือในการใช้งานสูง และกลุ่มแทนที่กรดอะมิโนที่ใช้ใน PROSITE มีคุณภาพที่ดี (สาวิณี และคณะ, 2549) อย่างไรก็ตามกระบวนการที่ดำเนินการโดยผู้เชี่ยวชาญเป็นกระบวนการที่ค่อนข้างช้าดังนั้นจึงมีความพยายามในการใช้ความรู้พื้นหลังประเภทอื่นๆ เข้ามาประยุกต์ใช้กับการค้นพบกลุ่มแทนที่ ซึ่งในงานศึกษานี้แบ่งความรู้พื้นหลังออกเป็น 2 ประเภทหลักๆ คือความรู้พื้นหลังแบบตารางคะแนนความเหมือน และความรู้พื้นหลังในลักษณะคอนเท็กซ์ (context)

สำหรับความรู้พื้นหลังแบบตารางแสดงคะแนนความเหมือนนั้นได้กล่าวถึงไปแล้วในหัวข้อตัวแทนสายโปรตีนประเภทโมทีฟดังแสดงตัวอย่างตาราง PAM ในภาพที่ 3 โดยการค้นพบกลุ่มแทนที่ที่มีคุณภาพดีขึ้นจากความรู้พื้นหลังประเภทนี้ มีหลักการพื้นฐานที่สมาชิกกรดอะมิโนในกลุ่มควรมีความคล้ายคลึงกันสูง ซึ่งสามารถหาความคล้ายคลึงกันได้จากตารางแสดงความเหมือน ดังนั้นในขั้นตอนแรกจึงจำเป็นต้องทราบกรดอะมิโนที่เป็นแกนกลางของกลุ่ม เช่น จาก บลอกตำแหน่งคอลัมน์ที่ 3 ในภาพที่ 9 คือ [ST] กรดอะมิโนที่เป็นแกนกลางของกลุ่มคือ S จากนั้นกำหนดค่าคะแนนวิกฤต (threshold) ขึ้นมา เช่น ให้ค่าคะแนนวิกฤตเป็น 1 จะได้กลุ่มแทนที่ที่ประกอบด้วยกรดอะมิโนชนิดที่มีความเหมือนกับ S ตั้งแต่คะแนน 1 ขึ้นไปเป็น [SANGPT] เป็นต้น

อย่างไรก็ตาม การเป็นสมาชิกของกลุ่มแทนที่กรดอะมิโนเดียวกันสามารถเกิดจากบางเงื่อนไขหรือบางคุณสมบัติที่กรดอะมิโนเหล่านั้นมีตรงกัน ซึ่งทำให้กรดอะมิโนบางชนิดมีค่าคะแนนความเหมือนกันน้อยเมื่อเทียบกับกรดอะมิโนอื่นๆ ในกลุ่ม ดังนั้นจึงมีหลายงานวิจัยที่

นำเสนอความรู้พื้นหลังที่สามารถปรับแต่งให้อยู่ในลักษณะคอนเท็กซ์สำหรับค้นพบกลุ่มแทนที่จากการมีเงื่อนไขหรือคุณสมบัติบางประการตรงกัน

คอนเท็กซ์คือตารางระบุความสัมพันธ์ระหว่างสิ่งของกับคุณสมบัติที่สิ่งของนั้นมี ดังนั้นความรู้พื้นหลังที่สามารถปรับแต่งเข้าสู่รูปแบบของคอนเท็กซ์ได้จึงมักอยู่ในรูปของเซ็ทหรือรากต้นไม้แสดงสับเซ็ทของกรดอะมิโน เช่น เซ็ทของกรดอะมิโน 6 กลุ่มในงานของ (Dayhoff *et al.*, 1978), แผนภาพคุณสมบัติเคมีฟิสิกส์ของ Taylor (Taylor-Venn's diagram) (Taylor, 1986), หรือรากต้นไม้แสดงสับเซ็ทของกรดอะมิโนในงานของ (Wu and Brutlag, 1996) ซึ่งความรู้พื้นหลังของกรดอะมิโนในลักษณะนี้มีการรวบรวมข้อมูลไว้ในงานของ (Wu and Brutlag, 1996) โดยตัวอย่างการแปลงความรู้พื้นหลังเป็นคอนเท็กซ์แสดงในภาพที่ 10



Amino Acid	small	polar	positive	aromatic	hydrophobic	sulfer
A	X					
C						X
D		X				
E		X				
F				X		
G	X					
H			X			
I					X	
K			X			
L					X	
M					X	
N		X				
P	X					
Q		X				
R			X			
S	X					
T	X					
V					X	
W				X		
Y				X		

ภาพที่ 10 การแปลงความรู้พื้นหลังเซ็ทกรดอะมิโนของ Dayhoff เป็นคอนเท็กซ์

สำหรับคอนเท็กซ์ของกรดอะมิโนและคุณสมบัติของกรดอะมิโนดังแสดงในภาพที่ 10 นี้ สามารถอธิบายในลักษณะเดียวกับคอนเท็กซ์ในทฤษฎีคอนเซ็ปต์แลททิส หรือ Concept Lattice (CL) ซึ่งจะได้กล่าวถึงในหัวข้อถัดๆ ไปในบทตรวจเอกสารนี้

จากตัวอย่างคอนเท็กซ์ในภาพที่ 10 เมื่อใช้กับกลุ่มแทนที่จากบลอกตำแหน่งคอลัมน์ที่ 3 ในภาพที่ 9 คือ [ST] จะพบการมีคุณสมบัติตรงกันคือ "ขนาดเล็ก" หรือ small ดังนั้นกลุ่มแทนที่กรดอะมิโนที่มีคุณสมบัติดังกล่าวก็คือ [AGPST]

อย่างไรก็ตาม ฟังก์ชั่นโปรตีนเป็นสิ่งที่ค่อนข้างซับซ้อน การใช้เพียงความรู้พื้นหลังในการค้นพบกลุ่มแทนที่โดยอัตโนมัติจึงได้กลุ่มแทนที่ที่ค่อนข้างหยาบ และต้องการเชื่อมโยงองค์ความรู้ของผู้เชี่ยวชาญเข้าไปในขั้นตอนวิธีค้นพบกลุ่มแทนที่ให้มีคุณภาพที่ใกล้เคียงกับผลที่ได้จากกระบวนการของผู้เชี่ยวชาญมากขึ้น และกลายเป็นวัตถุประสงค์สำคัญหนึ่งในวิทยานิพนธ์นี้ การนำ “ความรู้” ของปฏิกิริยาชีวเคมี (biochemical reaction) ประยุกต์เป็นกรอบประมวลผลเชิงคอนเซ็ปต์ในการพัฒนาตัวแทนสายโปรตีน

ในการพัฒนาเทคนิคตามวัตถุประสงค์ดังกล่าว จำเป็นต้องเข้าใจลักษณะปัญหาและข้อจำกัดของบลอกซึ่งส่งผลเป็นข้อจำกัดของการค้นพบกลุ่มแทนที่และโมทีฟที่ได้จากบริเวณจับและบริเวณเร่ง ดังจะได้นำเสนอในหัวข้อถัดไป

3.4 ปัญหาการใช้งานบลอกและกลุ่มแทนที่กรดอะมิโนที่ได้จากบริเวณจับและบริเวณเร่ง

จากลักษณะเฉพาะของฟังก์ชั่นเอนไซม์ทำให้โมทีฟที่เหมาะสมเป็นตัวแทนสายโปรตีนสำหรับระบบทำนายประเภทฟังก์ชั่นเอนไซม์ ก็คือโมทีฟที่ได้จากบลอกในบริเวณจับและบริเวณเร่ง ซึ่งฐานข้อมูล BLOCKS (Henikoff and Henikoff, 1991) ที่สร้างจากเทคนิค MSA มีข้อมูลบริเวณดังกล่าวอยู่น้อยมากคือประมาณ 50 บลอกเท่านั้น ดังนั้นเทคนิคการค้นพบโมทีฟจากบลอกด้วยขั้นตอนวิธีอัตโนมัติทุกวิธี เช่น eMOTIF 3MOTIF เป็นต้น จึงได้ผลส่วนใหญ่เป็นโมทีฟที่ไม่สัมพันธ์กับฟังก์ชั่นเอนไซม์ แต่เป็นโมทีฟจากบริเวณอื่น เช่น บริเวณอนุรักษ์ในสายวิวัฒนาการ เป็นต้น ซึ่งการใช้งานโมทีฟจากบริเวณเหล่านี้ในการทำนายประเภทฟังก์ชั่นเอนไซม์ มีความแม่นยำขึ้นอยู่กับลักษณะกลุ่มของข้อมูลที่มี ณ เวลานั้น เป็นหลัก และได้โมทีฟที่ไม่สอดคล้องกับความต้องการใช้งานในระดับห้องปฏิบัติการ ซึ่งต้องการโมทีฟที่มีความสัมพันธ์โดยตรงกับฟังก์ชั่นเอนไซม์ นั่นคือ การได้มาซึ่งโมทีฟจากบริเวณจับและบริเวณเร่งที่เป็นกลไกฟังก์ชั่นเอนไซม์โดยตรง

อย่างไรก็ตาม ปัญหาที่สำคัญที่สุดของการค้นพบโมทีฟหรือกลุ่มแทนที่กรดอะมิโน จากบลอกที่ได้จากบริเวณจับและบริเวณเร่งก็คือ การมีข้อมูลในบริเวณดังกล่าวอยู่น้อยมาก โดยบลอกที่มีข้อมูลบริเวณจับหรือบริเวณเร่งน้อยกว่า 5 ระเบียนมีถึง 85% และบลอกที่มีเพียง 1 ระเบียนมีถึง 57% ซึ่งการค้นพบกลุ่มแทนที่จำเป็นต้องมีอย่างน้อย 2 ระเบียน และการค้นพบกลุ่มแทนที่ที่มีความครอบคลุมสมบูรณ์ขึ้นกับการมีข้อมูลที่มากพอในระดับหนึ่ง

นอกจากนี้ การทำงานในบริเวณจับหรือบริเวณเร่งที่ทำงานเหมือนกันอาจมีลักษณะเฉพาะที่แตกต่างกันได้ เช่น การเข้าจับสารตั้งต้นในทิศทางที่แตกต่างกันเป็นต้น ซึ่งหมายความว่าในแต่ละบลอกอาจมีคลัสเตอร์ (cluster) ย่อยๆ ได้ ซึ่งการค้นพบโมทีฟจากบลอกที่มีคลัสเตอร์ย่อยจะประสบกับปัญหาการรบกวนกันแบบเดียวกับการรบกวนของกลุ่มโปรตีนต่างสายวิวัฒนาการที่แสดงไว้ในภาพที่ 6 ในขณะที่การแบ่งคลัสเตอร์จากบลอกที่มีปริมาณน้อยนั้นทำได้ยาก อีกทั้งเมื่อทำสำเร็จย่อมเป็นการเพิ่มปัญหาบลอกที่มี 1 ระเบียนหรือน้อยกว่า 5 ระเบียนให้มีมากยิ่งขึ้น

ดังนั้นในงานศึกษาวิทยานิพนธ์นี้ จึงมุ่งเน้นไปที่การเชื่อมโยงองค์ความรู้ลักษณะต่างๆ เข้ากับบลอกเพื่อชดเชยข้อด้อยของการมีข้อมูลน้อยในการพัฒนาคุณภาพของกลุ่มแทนที่ และใช้เทคนิคทางสถิติประกอบความรู้พื้นหลังในการพัฒนาคุณภาพของบลอกที่มีสมาชิกเพียง 1 ระเบียนได้

ในการแก้ปัญหาโดยใช้ความรู้พื้นหลังในการพัฒนาตัวแทนสายโปรตีนจากบริเวณที่มีข้อมูลน้อยนั้น ทฤษฎีพื้นฐานหนึ่งที่ใช้ในการเชื่อมโยงองค์ความรู้ลักษณะต่างๆ เข้ากับบลอกก็คือทฤษฎีคอนเซ็ปต์แลททิส มีรายละเอียดดังจะได้กล่าวถึงในส่วนทฤษฎีและเทคนิคพื้นฐานในหัวข้อถัดไป

4. ทฤษฎีและเทคนิคพื้นฐาน

4.1 ทฤษฎีคอนเซ็ปต์แลททิส: กรอบทำงานคอนเซ็ปต์ความรู้ของกลุ่มข้อมูล

เทคนิคคอนเซ็ปต์แลททิสหรือ Concept Lattice (CL) ถูกนำเสนอขึ้นเป็นครั้งแรกโดย (Wille, 1982) เป็นโมเดลของโครงสร้างข้อมูลแบบหนึ่งเพื่อใช้อธิบายความสัมพันธ์ขององค์ประกอบทั้งหมดในคอนเท็กซ์ให้อยู่ในรูปของคอนเซ็ปต์และความสัมพันธ์ระหว่างคอนเซ็ปต์นั้นเรียกโครงสร้างข้อมูลนี้ว่าคอนเซ็ปต์แลททิส (เรียกย่อว่าแลททิส)

สำหรับหลักการของ CL ยึดรูปแบบจากงานวิจัยของ Wille (1982, 1989); Waiyamai *et al.* (1997); Waiyamai *et al.* (2008) ดังนี้

ตารางที่ 1 คอนเท็กซ์ของการดำรงอยู่ของสิ่งมีชีวิตและน้ำ

Entities(abbreviation)	A	B	C	D	E	F	G	H	I
Leech(le)	X	X					X		
Bream(br)	X	X					X	X	
Frog(fr)	X	X	X				X	X	
Dog(do)	X		X				X	X	X
Spike-weed(sw)	X	X		X		X			
Reed(re)	X	X	X	X		X			
Bean(be)	X		X	X	X				
Maize(ma)	X		X	X		X			

ที่มา: Wille (1989)

หมายเหตุ คุณสมบัติที่ระบุในแต่ละคอลัมน์มีความหมายดังนี้ A: ขาดน้ำไม่ได้, B: อาศัยอยู่ในน้ำ, C: อาศัยอยู่บนบก, D: ต้องใช้คลอโรฟิลด์, E: ใบเลี้ยงคู่, F: ใบเลี้ยงเดี่ยว, G: เคลื่อนไหวได้, H: มีระยาง, I: ออกลูกเป็นตัว

นิยามที่ 1 คอนเท็กซ์ (context): เป็นความสัมพันธ์แบบไตรภาคี (Σ, P, R) โดย Σ และ P เป็นเซ็ทของสิ่งของ (Entities) และคุณสมบัติของสิ่งของ (Properties) มีความสัมพันธ์ทวิภาค R $\subseteq \Sigma$ x P หรือ eRp เมื่อสิ่งของ e $\in \Sigma$ มีความสัมพันธ์ R ไปที่ p $\in$ P เมื่อตรวจสอบแล้วว่าสิ่งของ e มีคุณสมบัติ p ตัวอย่างคอนเท็กซ์ของ การดำรงอยู่ของสิ่งมีชีวิตและน้ำ (living beings and water) (Wille, 1989) แสดงดังในตารางที่ 1

จากคอนเท็กซ์การดำรงอยู่ของสิ่งมีชีวิตและน้ำในตารางที่ 1 สามารถค้นพบคอนเซ็ปต์และแลททิสตามขั้นตอนของคอนเซ็ปต์แลททิสได้จากนิยามต่างๆ ดังนี้

นิยามที่ 2 คอนเซ็ปต์ (formal concept): อยู่ในรูปฟอร์มของคู่ลำดับ (Extent, Intent) ที่ได้จากคอนเท็กซ์ (Σ, P, R) เมื่อกำหนดให้ Extent $\subseteq \Sigma$ และ Intent $\subseteq$ P และกำหนดให้คอนเซ็ปต์เป็นคู่ลำดับที่มีฟังก์ชั่นการเชื่อมต่อ Galois (Galois Connection) ซึ่งกันและกันคือ f(Extent) = Intent และ g(Intent) = Extent โดยจะกล่าวถึงรายละเอียดอีกครั้งเมื่อประยุกต์คอนเซ็ปต์กับกรดอะมิโนในบทอุปกรณ์และวิธีการ

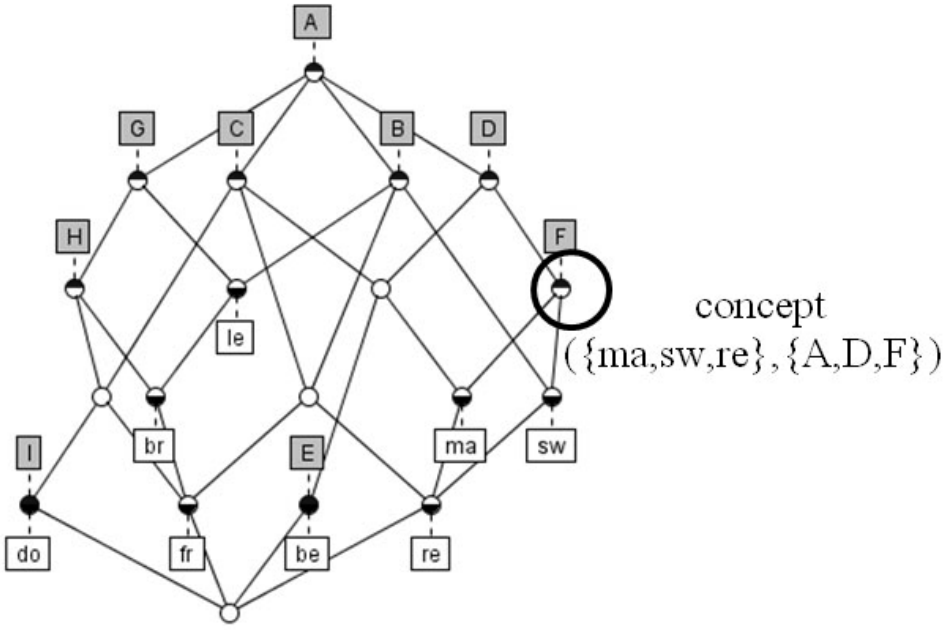
พิจารณาตัวอย่างคอนเซ็ปต์ ({re,fr}, {A,B,C}) เป็นคอนเซ็ปต์ที่ได้จากคอนเท็กซ์การดำรงอยู่ของสิ่งมีชีวิตและน้ำ เนื่องจากสิ่งมีชีวิต {re,fr} มีคุณสมบัติร่วมที่มากที่สุดคือ {A,B,C} หรือ f(Extent) = Intent ในขณะเดียวกันสิ่งมีชีวิตในกลุ่ม {re,fr} เป็นกลุ่มสิ่งมีชีวิตที่ใหญ่ที่สุดของการมีคุณสมบัติ {A,B,C} หรือ g(Intent) = Extent

นิยามที่ 3 คอนเซ็ปต์แลททิส (concept lattice): คือโครงสร้างข้อมูลของกลุ่มคอนเซ็ปต์ทั้งหมดที่ได้จากคอนเท็กซ์ (Σ, P, R) โดยแต่ละคอนเซ็ปต์ c $\in$ L มีความสัมพันธ์แบบ (L,$\leq$) ที่เรียงตามลำดับขนาดของคอนเซ็ปต์ (partial ordering) โดยเริ่มจากคอนเซ็ปต์ที่มีขนาด Extent ใหญ่ที่สุดไว้ข้างบนสุด ตามด้วยคอนเซ็ปต์ที่ขนาดใหญ่รองลงมาที่ Extent เป็นสับเซ็ทของ Extent ในคอนเซ็ปต์ด้านบน จนกระทั่งได้คอนเซ็ปต์ที่เล็กที่สุดด้านล่าง

ลักษณะหนึ่งที่สำคัญของโครงสร้างข้อมูลในคอนเซ็ปต์แลททิสก็คือ ความสัมพันธ์แบบ super-concept และ sub-concept โดยในกลุ่มคอนเซ็ปต์ที่มีความสัมพันธ์แบบ (L,$\leq$) คอนเซ็ปต์ A ใดใด ที่มี extent เป็น subset ของ extent ในคอนเซ็ปต์ B ใดใด กล่าวได้ว่าคอนเซ็ปต์ A นั้นเป็น sub-concept ของคอนเซ็ปต์ B และ คอนเซ็ปต์ B เป็น super-concept ของคอนเซ็ปต์ A

ทั้งนี้ตัวอย่างของคอนเซ็ปต์แลททิสที่ได้จากคอนเท็กซ์การดำรงอยู่ของสิ่งมีชีวิตและน้ำ ในตัวอย่างตารางที่ 1 แสดงดังในภาพที่ 11 โดยการดำเนินการที่สัมพันธ์กันระหว่างคอนเซ็ปต์ที่ประยุกต์ใช้กับเรื่องกลุ่มแทนที่กรดอะมิโนจะได้กล่าวถึงในบทอุปกรณ์และวิธีการต่อไป

ทฤษฎีคอนเซ็ปต์แลททิสกล่าวถึงเทคนิคที่ประมวลคอนเซ็ปต์และความสัมพันธ์ระหว่างคอนเซ็ปต์ที่ได้จากคอนเท็กซ์ความรู้พื้นหลัง ดังนั้นเทคนิคคอนเซ็ปต์แลททิสจึงเหมาะสมอย่างยิ่งในการนำมาใช้เชื่อมโยงความรู้พื้นหลังใดใดที่อยู่ในรูปคอนเท็กซ์อย่างเป็นระบบ



ภาพที่ 11 คอนเซ็ปต์แลททิสของคอนเท็กซ์การดำรงอยู่ของสิ่งมีชีวิตและน้ำ

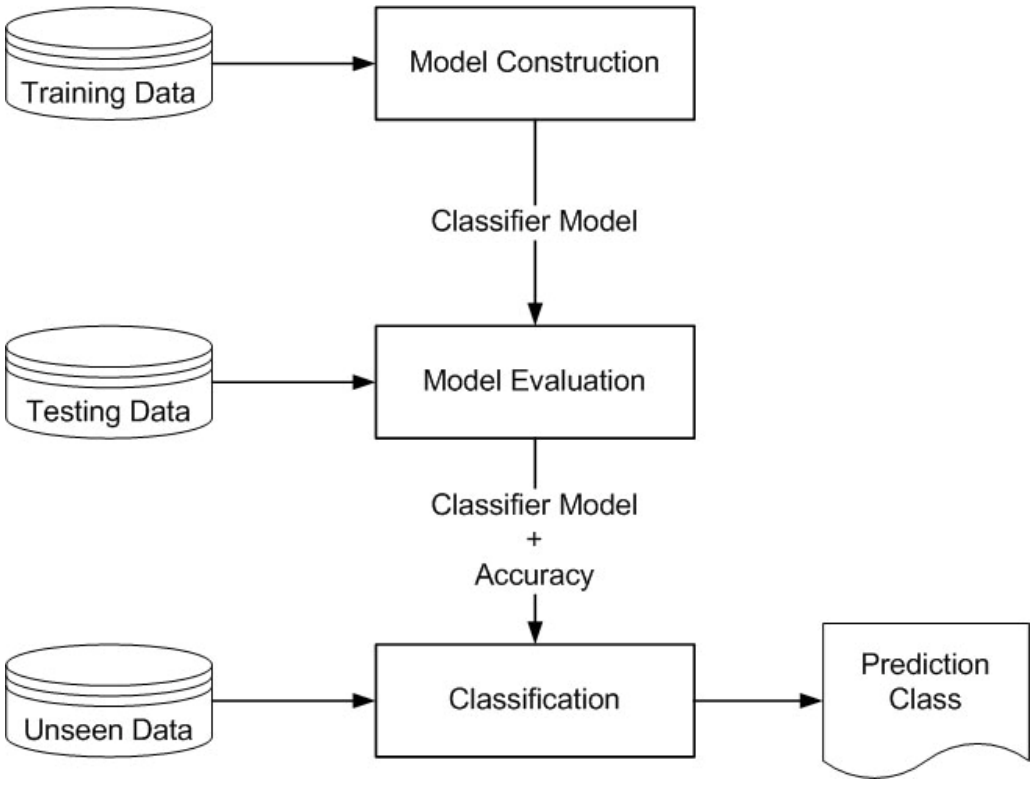
หมายเหตุ สร้างคอนเซ็ปต์แลททิสด้วยโปรแกรม concept explorer 1.3 โดยแต่ละครเซ็ปต์แสดงภาพในลักษณะย่อ โดยรายละเอียดเต็มของ Extent คือเซ็ทของสิ่งของ (enyities) ทั้งหมดที่อยู่ด้านล่างของคอนเซ็ปต์นั้น และมี Intent คือเซ็ทของ คุณสมบัติ (properties) ทั้งหมดที่อยู่ด้านบนของคอนเซ็ปต์นั้น ซึ่งจากตัวอย่างคอนเซ็ปต์ในวงกลมใหญ่ก็คือ ({ma,sw,re},{A,D,F})

ในหัวข้อต่อไปจะได้กล่าวถึงเทคนิคพื้นฐานของเทคนิคเหมืองข้อมูล (data mining) ที่เรียกว่าระบบทำนายประเภทข้อมูล ซึ่งใช้ในการทำนายประเภทฟังก์ชั่นเอนไซม์

4.2 เทคนิคเหมืองข้อมูล (data mining) และการพัฒนาระบบทำนายประเภทข้อมูล (data classification)

เทคนิคเหมืองข้อมูล (data mining) เป็นเทคนิคที่ดึงข้อมูลสำคัญหรือรูปแบบข้อมูลที่น่าสนใจ (interesting patterns) จากกลุ่มข้อมูลขนาดใหญ่ (Han and Kamber, 2001; กฤษณะ, 2549) สามารถแบ่งออกเป็น 2 กลุ่มเทคนิคคือ 1) กลุ่มเทคนิคที่ใช้ในการทำนายข้อมูล (predictive mining tasks) เช่น การทำนายผลลัพธ์เชิงปริมาณ (prediction) และการทำนายประเภทข้อมูล (data

classification) เป็นต้น และ 2) กลุ่มเทคนิคเชิงพรรณนาข้อมูล (descriptive mining tasks) เช่น การแบ่งกลุ่มข้อมูล (data clustering) และการค้นพบกฎความสัมพันธ์ (association rule discovery) เป็นต้น โดยในงานวิจัยนี้ต้องการเครื่องมือในกลุ่มเทคนิคที่ใช้ในการทำนายข้อมูลที่เรียกว่าระบบทำนายประเภทข้อมูล (data classification) ในการทำนายฟังก์ชั่นเอนไซม์ที่มีลักษณะเป็นกลุ่มประเภทข้อมูล (category)



ภาพที่ 12 ขั้นตอนการพัฒนาระบบทำนายประเภทข้อมูล

ในการพัฒนาระบบทำนายประเภทข้อมูลมีหลักการพื้นฐานคือการวิเคราะห์ข้อมูลเพื่อให้ได้มาซึ่งโมเดลหรือแบบจำลองอธิบายลักษณะสำคัญของข้อมูลแต่ละกลุ่มประเภท ซึ่งเราสามารถใช้โมเดลนั้นในการทำนายประเภทข้อมูลของข้อมูลใหม่ได้ รายละเอียดนำเสนอดังในภาพที่ 12

ในภาพที่ 12 นำเสนอขั้นตอนการพัฒนาระบบทำนายประเภทข้อมูลโดยเริ่มต้นจากการ นำข้อมูลเรียนรู้ระบบ (Training Data) มาผ่านกระบวนการเรียนรู้แบบมีผู้สอน (Supervise Learning) ในขั้นตอนแรกคือการสร้างโมเดล (Model Construction) ได้ผลจากขั้นตอนนี้เป็นโมเดล

จำแนกประเภท (Classifier Model) จากนั้นนำข้อมูลทดสอบ (Testing Data) มาทดสอบความถูกต้องแม่นยำของโมเดลจำแนกประเภทเพื่อพัฒนาจนได้ระบบทำนายประเภทที่มีประสิทธิภาพในขั้นตอนพัฒนาโมเดล (Model Evaluation) ซึ่งจากขั้นตอนนี้จะได้ระบบทำนายประเภทข้อมูล (Classifier Model) ที่พร้อมใช้ทำนายประเภทของข้อมูลตัวใหม่ที่ยังไม่ทราบประเภทของข้อมูล (Unseen Data) ในขั้นตอนการทำนายประเภทข้อมูล (Classification) ได้

ในการเลือกเครื่องมือเหมืองข้อมูลที่ใช้ในงานวิจัย เลือกจากการเปิดโอกาสให้ใช้งานฟรี มีความสามารถด้านเหมืองข้อมูลที่ครอบคลุม มีความนิยมใช้งานสูง และมีความน่าเชื่อถือจากการพัฒนาที่ต่อเนื่องยาวนาน โดยในงานวิจัยนี้เลือกใช้ชุดโปรแกรม WEKA หรือ Waikato Environment for Knowledge Analysis (Witten and Frank, 2005) ที่เริ่มต้นพัฒนามาจากชุดโปรแกรมสำหรับเครื่องจักรเรียนรู้ (machine learning) ตั้งแต่ปี 1993 โดยใช้ภาษา C จากนั้นเปลี่ยนมาใช้ภาษา Java ในปี 1997 และได้รับการยอมรับเป็นเครื่องมือยอดนิยมด้านเหมืองข้อมูลในปี 2005 โดยชุดเครื่องมือ WEKA สามารถใช้งานเทคนิคเหมืองข้อมูลที่ครอบคลุมตั้งแต่ขั้นตอนการเตรียมข้อมูล การทำนายประเภทข้อมูล การแบ่งกลุ่มข้อมูล การค้นพบกฎความสัมพันธ์ รวมไปถึงการสร้างภาพข้อมูล (visualization) โดยแต่ละกลุ่มเทคนิคมีขั้นตอนวิธี (algorithms) ที่หลากหลายให้เลือกใช้เป็นจำนวนมาก

สำหรับการทดสอบความถูกต้องแม่นยำของโมเดลการทำนายประเภทข้อมูลนั้น มีเทคนิคที่สำคัญคือการจัดกลุ่มข้อมูลฝึกสอนในการพัฒนาโมเดลการทำนายประเภทข้อมูล และการจัดกลุ่มข้อมูลทดสอบในการวัดความถูกต้องแม่นยำของโมเดลดังกล่าวด้วยเครื่องมือวัดต่างๆ โดยเทคนิคพื้นฐานที่ใช้ในวิทยานิพนธ์มีดังนี้

4.2.1 การจัดกลุ่มข้อมูลฝึกสอนและกลุ่มข้อมูลทดสอบแบบ n-fold cross-validation สำหรับระบบทำนายประเภทข้อมูล

โดยทั่วไปการแบ่งกลุ่มข้อมูลฝึกสอนและกลุ่มข้อมูลทดสอบเพื่อพัฒนาและวัดประสิทธิภาพของระบบทำนายประเภทข้อมูลใช้วิธีสุ่มในการแบ่งกลุ่มข้อมูลทั้งหมดกลุ่มหนึ่งออกเป็น 2 ส่วน โดยเป็นกลุ่มข้อมูลฝึกสอนมีขนาด 2 ใน 3 ของข้อมูลทั้งหมด และกลุ่มข้อมูลที่เหลือ 1 ใน 3 เป็นกลุ่มข้อมูลทดสอบ ซึ่งการแบ่งข้อมูลแบบสุ่มเพื่อให้ได้กลุ่มข้อมูลฝึกสอนและกลุ่มข้อมูลทดสอบทำให้ค่าความแม่นยำที่วัดได้จากโมเดลการทำนายประเภทข้อมูลมีความไม่

แน่นอนขึ้นอยู่กับผลการสุ่ม ดังนั้นในการวัดค่าความแม่นยำของโมเดลการทำนายประเภทข้อมูลจึงมักนิยมใช้วิธีทางสถิติเข้าช่วยซึ่งให้ผลที่น่าเชื่อถือมากขึ้น

เทคนิค n-fold cross-validation เป็นการวิเคราะห์ข้อมูลด้วยสถิติที่รู้จักกันดีมากว่า 40 ปี (Mosteller and Tukey, 1968) มีหลักการเรียบง่ายแต่มีประสิทธิภาพในการใช้งาน (Han and Kamber, 2001) โดยเริ่มต้นจากการแบ่งกลุ่มแบบสุ่มจากกลุ่มข้อมูลทั้งหมดเป็นกลุ่มย่อยหรือ “fold” จำนวน n กลุ่ม คือ S_1, S_2,.., S_nที่มีขนาดใกล้เคียงกันทุกกลุ่ม ข้อมูล fold เหล่านี้จะถูกนำมาสร้างเป็นกลุ่มข้อมูลฝึกสอนและกลุ่มข้อมูลทดสอบจำนวน n รอบ โดยแต่ละรอบที่ i = 1 ถึง n กำหนดให้ S_i เป็นกลุ่มข้อมูลทดสอบ ที่เหลือเป็นกลุ่มข้อมูลฝึกสอนเพื่อสร้างโมเดลการทำนายประเภทข้อมูล (classifier) โดยค่าความแม่นยำของโมเดลการทำนายประเภทข้อมูลวัดจากจำนวนที่ทำนายประเภทถูกต้องทั้งหมดจาก n รอบนั้น เทียบกับจำนวนข้อมูลที่มีในกลุ่มข้อมูลทั้งหมด

สำหรับเครื่องมือวัดประสิทธิภาพของระบบทำนายประเภทข้อมูลในกลุ่มข้อมูลทดสอบนั้น สามารถแบ่งเครื่องมือวัดออกเป็น 2 กลุ่มคือ การวัดประสิทธิภาพของระบบทำนายประเภทข้อมูลที่มี 2 ประเภทข้อมูล และ การวัดประสิทธิภาพของระบบทำนายประเภทข้อมูลที่มีมากกว่า 2 ประเภทข้อมูล รายละเอียดดังนี้

4.2.2 การวัดประสิทธิภาพของระบบทำนายประเภทข้อมูลที่มี 2 ประเภทข้อมูล

สำหรับการวัดประสิทธิภาพของระบบทำนายข้อมูลที่มี 2 ประเภท (binary classes) นิยมใช้มาตรวัดที่พัฒนาจากงานของ van Rijsbergen (1979) โดยใช้ Contingency Table หรือ Confusion Matrix ในการระบุค่า true positive (TP) true negative (TN) false positive (FP) และ false negative (FN) (Kohavi and Provost, 1998) เพื่อใช้ในการคำนวณค่าคุณภาพต่างๆ เช่น accuracy precision sensitivity และ specificity เป็นต้น โดยตัวอย่างของ Contingency Table แสดงดังในตารางที่ 2 และ Confusion Matrix ดังในตารางที่ 3

ตารางที่ 2 องค์ประกอบของ Contingency Table

Contingency Table		**Predicted Class**		**ASum**
		Retrieved	**Not Retrieved**	
Actual Class	**Relevant**	$A \cap B$	$A \cap B^{-}$	A
	Non-Relevant	$A^{-} \cap B$	$A^{-} \cap B^{-}$	A^{-}
PSum		B	B^{-}	N

ตารางที่ 3 องค์ประกอบของ Confusion Matrix สำหรับผลการทำนายประเภทข้อมูล 2 ประเภท

Confusion Matrix		**Predicted Class**	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

จากตารางที่ 2 แสดงองค์ประกอบของ Contingency Table แสดงผลการทำนายประเภทของกลุ่มข้อมูลชุดหนึ่งประกอบด้วยรายละเอียดคือ

N หมายถึง เซ็ทของข้อมูลทั้งหมด

A หมายถึง เซ็ทของข้อมูลที่สัมพันธ์กับประเภทข้อมูลนั้นจริง

A^{-} หมายถึง เซ็ทของข้อมูลที่ไม่สัมพันธ์กับประเภทข้อมูลนั้น

B หมายถึง เซ็ทของข้อมูลที่ทำนายว่าเกี่ยวข้องกับประเภทข้อมูลนั้น

B^{-} หมายถึง เซ็ทของข้อมูลที่ทำนายว่าไม่เกี่ยวข้องกับประเภทข้อมูลนั้น

สำหรับตารางที่ 3 แสดงองค์ประกอบของ Confusion Matrix แสดงผลการทำนายประเภทของกลุ่มข้อมูลชุดหนึ่งประกอบด้วยรายละเอียดคือ

ค่า TP ใน Confusion Matrix แสดงจำนวนที่ระบบทำนายประเภทของข้อมูล (Predicted Class = Yes) ตรงกับความเป็นจริง (Actual Class = Yes) ตรงกับ $|A \cap B|$ ใน Contingency Table

ค่า FP ใน Confusion Matrix แสดงจำนวนที่ระบบทำนายประเภทของข้อมูล (Predicted Class = Yes) แต่ผิดไปจากความเป็นจริงโดยข้อมูลนั้นมิใช่ประเภทดังกล่าว (Actual Class = No) ตรงกับ $|A^{-} \cap B|$ ใน Contingency Table

ค่า FN ใน Confusion Matrix แสดงจำนวนที่ระบบทำนายผลว่าไม่ใช่ประเภทของข้อมูลนั้น (Predicted Class = No) แต่ผิดไปจากความเป็นจริงที่ข้อมูลนั้นเป็นประเภทดังกล่าว (Actual Class = Yes) ตรงกับ $|A \cap B^{-}|$ ใน Contingency Table

ค่า TN ใน Confusion Matrix แสดงจำนวนที่ระบบทำนายผลว่าไม่ใช่ประเภทของข้อมูลนั้น (Predicted Class = No) ตรงกับความเป็นจริงที่ข้อมูลนั้นมิใช่ประเภทดังกล่าว (Actual Class = No) ตรงกับ $|A^{-} \cap B^{-}|$ ใน Contingency Table

โดยค่าเหล่านี้ถูกนำมาคำนวณเป็นมาตรวัดคุณภาพของระบบทำนายประเภทข้อมูลได้หลายลักษณะด้วยกันคือ

Accuracy = $\frac{|A \cap B| + |A^{-} \cap B^{-}|}{|N|} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$, หมายถึงค่าความแม่นยำในการทำนายผลโดยรวมทั้งที่ใช่และไม่ใช่ประเภทของข้อมูลนั้นได้อย่างถูกต้อง

Precision = $\frac{|A \cap B|}{|B|} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$, หมายถึงค่าความถูกต้องในการทำนายประเภทของข้อมูลจากจำนวนที่ทำนายประเภทข้อมูลทั้งหมด

Sensitivity = $\frac{|A \cap B|}{|A|} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$, หมายถึงค่าความครอบคลุมในการทำนายประเภทของข้อมูลได้อย่างถูกต้องจากที่มีอยู่ทั้งหมด

Specificity = $\frac{|\bar{A} \cap \bar{B}|}{|\bar{A}|} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$, หมายถึงความเฉพาะเจาะจงในการระบุได้อย่างไม่ผิดพลาดว่าข้อมูลนั้นมิใช่ประเภทข้อมูลใด

เทคนิคการวัดประสิทธิภาพของระบบทำนายประเภทข้อมูลที่มี 2 ประเภทข้อมูลดังกล่าวนี้ มีความเหมาะสมอย่างยิ่งในการนำมาวัดคุณภาพของโมทีฟ หรือวัดความสามารถในการระบุว่าสายโปรตีนนั้นควรมีโมทีฟใดหรือไม่ควรมีโมทีฟใด ในกรณีของเอนไซม์ก็คือการใช้โมทีฟในการระบุประเภทของสายโปรตีนว่า “มี” หรือ “ไม่มี” บริเวณจับหรือบริเวณเร่งใด ซึ่งก็คือลักษณะของการทำนายข้อมูลที่มี 2 ประเภท (binary classes) ของแต่ละโมทีฟที่พัฒนาขึ้นมานั่นเอง

สำหรับในงานวิจัยนี้ เพิ่มวิธีการวัดคุณภาพโมทีฟอีกลักษณะหนึ่งคือ “ค่าความครอบคลุม” (หรือ % Coverage Value) (Cleverdon *et al.*, 1966) หมายถึง ความสามารถในการนำชุดโมทีฟไปใช้ประโยชน์ในการเป็นตัวแทนของชุดข้อมูลที่มีอยู่ ซึ่งก็คือจำนวนเปอร์เซ็นต์ของข้อมูลที่มีโมทีฟต่อจำนวนข้อมูลทั้งหมด (Liewlom *et al.*, 2007) แสดงการคำนวณดังสูตรด้านล่าง

% Coverage Value = $100\left(\frac{S_{motif}}{S_{all}}\right)$, โดยกำหนดให้ S_{motif} เป็นจำนวนข้อมูลตัวอย่างที่มีโมทีฟอย่างน้อย 1 โมทีฟจากชุดโมทีฟนั้น และ S_{all} เป็นจำนวนข้อมูลตัวอย่างทั้งหมดของชุดข้อมูลนั้น

4.2.3 การวัดประสิทธิภาพของระบบทำนายประเภทข้อมูลที่มีมากกว่า 2 ประเภทข้อมูล

ในการพัฒนาระบบทำนายประเภทของข้อมูลโดยทั่วไปที่มีมากกว่า 2 ประเภทข้อมูลนั้น จำเป็นต้องมีมาตรวัดคุณภาพว่าระบบที่พัฒนาขึ้นนั้นให้ผลการทำงานที่ดีเพียงใด เพื่อหาวิธีปรับแต่งจนได้ระบบทำนายประเภทข้อมูลที่มีคุณภาพดี โดยวิธีหนึ่งที่นิยมใช้ก็คือ เปอร์เซ็นต์การทำนายผลที่ถูกต้องจากข้อมูลทั้งหมด (หรือ % Correctly Classified Instance) (Witten and Frank, 2005) โดยคำนวณจาก Confusion Matrix ที่มีประเภทข้อมูลมากกว่า 2 ประเภท ดังตารางที่ 4

ตารางที่ 4 แสดงองค์ประกอบของ Confusion Matrix สำหรับผลการทำนายประเภทข้อมูลที่มากกว่า 2 ประเภท

Confusion Matrix		Predicted Class			ASum
		A	**B**	**C**	
Actual Class	**A**	10	0	0	10
	B	0	9	1	10
	C	0	2	8	10
PSum		10	11	9	Sum = 30

จากตารางที่ 4 เป็นผลการทำนายประเภทข้อมูล A B และ C จากจำนวนข้อมูลทั้งหมด Sum = 30 ตัวอย่าง (Instances) โดยคอลัมน์ของตารางแสดงผลการใช้ระบบทำนายประเภทข้อมูลในการทำนายประเภทของข้อมูล ส่วนแถวของตารางแสดงประเภทของข้อมูลที่มีอยู่จริง โดยสามารถแปลผลในตารางที่ 4 ได้ว่า ในจำนวนข้อมูล 10 ตัวอย่างที่เป็นประเภท A สามารถใช้ระบบทำนายประเภทข้อมูลทำนายถูกต้อง $PSum_A$ =10 ตัว สำหรับข้อมูลประเภท B นั้น ระบบทำนายประเภทข้อมูลทำนายเป็นประเภท B จำนวน $PSum_B$ = 11 ตัวอย่าง แต่ถูกต้องเพียง 9 ตัวอย่าง โดย 2 ตัวอย่างที่เหลือทำนายข้อมูลประเภท C เป็น B สำหรับข้อมูลประเภท C นั้นใช้ระบบทำนายประเภทข้อมูลทำนายได้ประเภท C จำนวน $PSum_C$ = 9 ตัวอย่าง แต่ถูกต้องเพียง 8 ตัวอย่าง โดยที่เหลือ 1 ตัวอย่างทำนายข้อมูลประเภท B เป็น C ดังนั้นจึงสามารถคำนวณ “เปอร์เซ็นต์การทำนายผลที่ถูกต้อง” จากจำนวนตัวอย่างที่ทายถูกในเซลสีเทาที่เป็น diagonal ทั้งหมดของตาราง หารด้วยจำนวนตัวอย่างทั้งหมด ดังแสดงในสมการด้านล่าง

$$\% \text{ Correctly Classified Instance} = 100 \left(\frac{\sum_{i=1}^{n} M_{ii}}{Sum} \right) \%$$, โดยกำหนดให้ M_{ii} คือ จำนวนข้อมูลของเซลใน Confusion Matrix ในโซนที่เป็น diagonal แทยงมุมจากบนซ้ายลงมาล่างขวาของเมทริกซ์ โดย Sum คือจำนวนตัวอย่างทั้งหมด และ n คือจำนวนประเภทของข้อมูลโดยจากตัวอย่างในตารางที่ 4 ได้ค่าความแม่นยำ = 100 ((10+9+8)/30) = 90%

ทั้งนี้สามารถคำนวณค่าความแม่นยำได้อีกหลายลักษณะคือ precision sensitivity และ specificity ที่ได้จาก Confusion Matrix M ดังนี้

$$\text{precision} = \left(\frac{\sum_{i=1}^{n} M_{ii}}{\sum_{i=1}^{n}\sum_{k=1}^{n} M_{ki}} \right) = \left(\frac{\sum_{i=1}^{n} M_{ii}}{\sum_{i=1}^{n} PSum_i} \right)$$

$$\text{sensitivity} = \left(\frac{\sum_{i=1}^{n} M_{ii}}{Sum} \right)$$

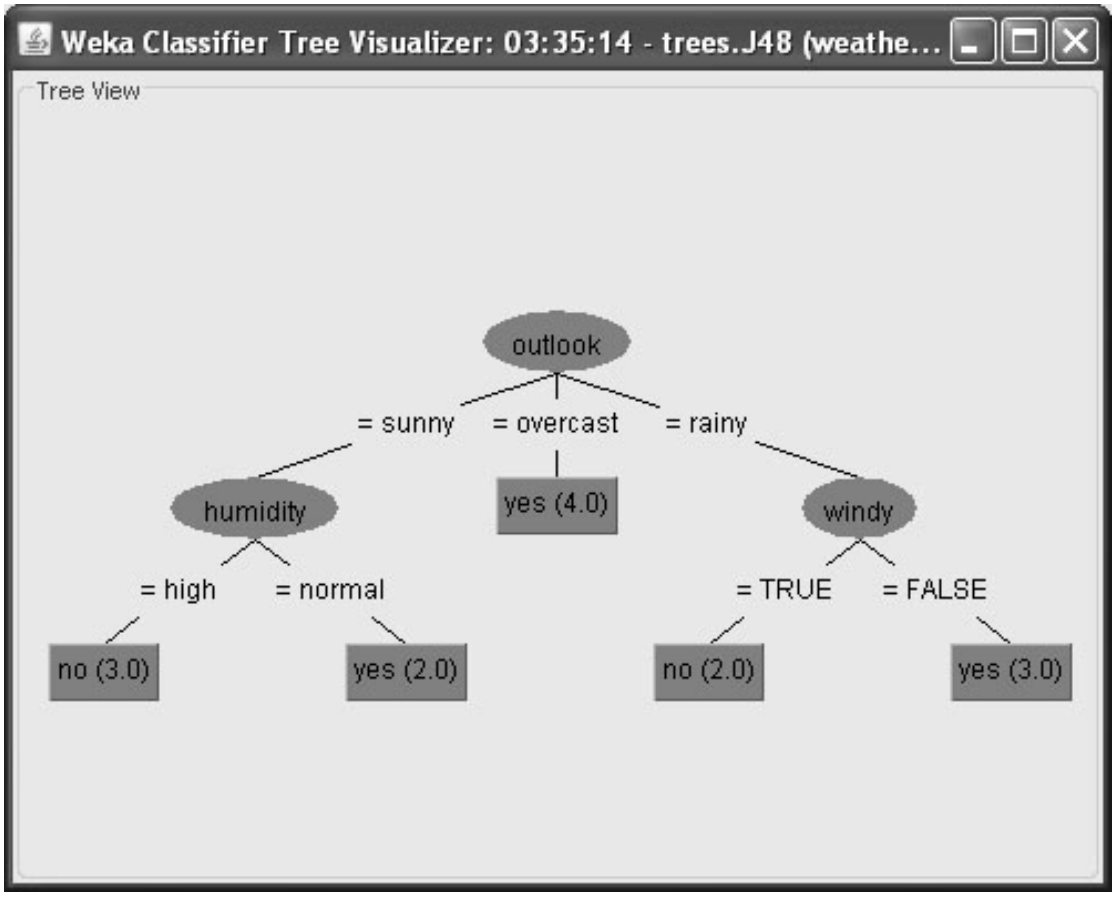
$$\text{specificity} = \left(\sum_{i=1}^{n} \left(\frac{Sum - ASum_i - PSum_i + M_{ii}}{Sum - ASum_i} \right) \right)$$

สำหรับเทคนิคพื้นฐานที่ใช้พัฒนาโมเดลการทำนายประเภทข้อมูลที่ใช้ในวิทยานิพนธ์นี้คือ เทคนิคต้นไม้ช่วยการตัดสินใจ (Decision Tree) ที่เรียกว่า C4.5 (Quinlan, 1993)

4.3 เทคนิคต้นไม้ช่วยการตัดสินใจ Decision Tree: C4.5

การเรียนรู้แบบมีผู้สอนเพื่อสร้างโมเดลทำนายประเภทข้อมูล (Model Construction) ที่มีให้เลือกใช้ใน WEKA นั้นมีขั้นตอนวิธีที่หลากหลาย เช่น Decision Tree, Bayesian Classification, Neural Network, Genetic Algorithm และ Association based Classification เป็นต้น โดยในงานวิทยานิพนธ์นี้ใช้เทคนิคประเภท Decision Tree ที่ชื่อว่า C4.5 (Quinlan, 1993)

เทคนิคต้นไม้ช่วยการตัดสินใจ (Decision Tree) เป็นโครงสร้างข้อมูลหรือโมเดลข้อมูลที่ใช้จำแนกประเภทของข้อมูล มีลักษณะทั่วไปคล้ายโครงสร้างต้นไม้ที่เริ่มต้นจากโหนดราก (root node) ระบุคุณสมบัติที่ใช้เริ่มต้นในการพิจารณา และเส้นกิ่ง (edge) แสดงเงื่อนไขของคุณสมบัตินั้นที่เชื่อมโยงไปพิจารณาคุณสมบัติในโหนดลูก (children node) อื่นๆ จนได้คำตอบที่โหนดปลายหรือลีฟโหนด (leaf node) ตัวอย่างของต้นไม้ช่วยการตัดสินใจนำเสนอดังในภาพที่ 13



ภาพที่ 13 ต้นไม้ช่วยการตัดสินใจการเล่นเทนนิสในสภาพอากาศต่างๆ

หมายเหตุ สร้างภาพจากโปรแกรม WEKA

จากภาพที่ 13 เป็นต้นไม้ช่วยการตัดสินใจที่เป็นโมเดลการทำนายการเลือกเล่นเทนนิสของชายผู้หนึ่ง ซึ่งโมเดลดังกล่าวสามารถนำมาใช้ทำนายได้ว่าชายผู้นี้น่าจะเล่นหรือไม่เล่นเทนนิสในสภาพอากาศของวันใหม่ได้ เช่น ถ้าในวันใหม่มีสภาพอากาศคือฝนตก (rainy) ความชื้น (humidity) มีสูง และลมแรง (windy) สามารถใช้ต้นไม้ช่วยการตัดสินใจทำนายได้ว่าชายผู้นี้น่าจะไม่เล่นเทนนิส โดยพิจารณาเริ่มต้นจากโหนดรากแทนค่าแอททริบิวต์สภาพอากาศ (outlook) พิจารณาไล่ไปตามกิ่งที่มีค่าสภาพอากาศฝนตก (rainy) พบโหนดลูกคือลมแรง windy พิจารณาจากกิ่งที่ให้ค่าลมแรงเป็น true ได้ลีฟโหนดเป็นการทำนายว่าชายผู้นี้ไม่น่าจะเล่นเทนนิสในสภาพอากาศแบบนี้

การสร้างต้นไม้ช่วยการตัดสินใจโดยทั่วไปประกอบด้วย 2 ขั้นตอนหลักคือ ขั้นตอนสร้างต้นไม้ (tree construction) เพื่อให้ได้โครงสร้างข้อมูลของทั้งชุดข้อมูลฝึกสอน และขั้นตอนตัดแต่งกิ่ง (tree pruning) เพื่อให้ได้โมเดลจำแนกประเภทข้อมูลที่ไม่เป็นการใช้งานที่เฉพาะเจาะจง (over fit) กับกลุ่มข้อมูลใดข้อมูลหนึ่งมากเกินไป

ในขั้นตอนการสร้างต้นไม้จากชุดข้อมูลฝึกสอน เป็นการวิเคราะห์ แอททริบิวต์ (attributes) ที่ใช้ระบุข้อมูลแต่ละระเบียนในชุดข้อมูลฝึกสอนนั้นโดยใช้ค่าฟังก์ชั่นความดี (goodness function) ในการเลือกแอททริบิวต์ที่มีค่าฟังก์ชั่นความดีดีที่สุดเป็นโหนดเริ่มต้น (root node) และเลือกจำนวนการแตกกิ่งจากโหนดนั้นให้ได้จำนวนกิ่งที่เหมาะสมที่สุด และวนลูปในการใช้ค่าฟังก์ชั่นความดีในการเลือกแอททริบิวต์เป็นโหนดลูกและจำนวนกิ่งที่แตกจากโหนดลูกไปเรื่อยๆ จนได้ลีฟโหนดที่เป็นการระบุประเภทข้อมูล

ขั้นตอนต่อมาคือการตัดแต่งกิ่งให้ได้โครงสร้างข้อมูลที่ไม่เป็นการใช้งานเฉพาะเจาะจงกับกลุ่มข้อมูลใดข้อมูลหนึ่งมากเกินไป โดยเครื่องมือทางคณิตศาสตร์ที่ใช้ก็คือการลดอัตราความผิดพลาด (error rate reduce) โดยทำการยุบลีฟโหนดที่เป็นการลดอัตราความผิดพลาดออกไป

สำหรับเทคนิคต้นไม้ช่วยการตัดสินใจ Decision Tree C4.5 (Quinlan, 1993) มีความสามารถในการรองรับโหนดที่มีคุณสมบัติที่เป็นค่าตัวเลขต่อเนื่องได้ เช่น อุณหภูมิ ความสูง เป็นต้น โดยในขั้นตอนการสร้างต้นไม้ C4.5 จะแบ่งช่วงของตัวเลขให้ได้จำนวนเงื่อนไขหรือเส้นกิ่งที่เหมาะสมที่สุดกับคำตอบปลายทางโดยใช้ค่า information gain และ gain ratio เป็นฟังก์ชั่นความดี (goodness function) และใช้เทคนิค estimated true error rate reduce ในการคำนวณอัตราความผิดพลาดในขั้นตอนการตัดแต่งกิ่งทำให้เทคนิค C4.5 เป็นวิธีที่ได้รับความนิยมสูงในการนำไปใช้สร้างโมเดลจำแนกประเภท เนื่องจากมีความเร็วในการเรียนรู้เพื่อสร้างโมเดลทำนายประเภทข้อมูลสูงกว่าวิธีอื่นๆ โดยที่ให้เปอร์เซ็นต์ความถูกต้องของผลลัพธ์พอ ๆ กัน

สำหรับบทนี้ได้ทำการตรวจเอกสารในเนื้อหาที่เกี่ยวข้องและที่เป็นเทคนิคพื้นฐาน โดยในบทต่อไปจะได้แสดงรายละเอียดเกี่ยวกับขั้นตอนวิธีที่พัฒนาขึ้นใช้ในงานวิจัยนี้

อุปกรณ์และวิธีการ

อุปกรณ์

1. ฮาร์ดแวร์
    1.1 คอมพิวเตอร์ Intel Centrino ความเร็ว CPU clock rate 1.3 GHz
    1.2 หน่วยความจำหลัก 768 MB
    1.3 ฮาร์ดดิสก์ขนาด 40 GB

2. ซอฟต์แวร์
    2.1 โปรแกรมตัวแปลภาษา Delphi 7.0
    2.2 โปรแกรมเหมืองข้อมูลและเครื่องจักรเรียนรู้ Waikato Environment for Knowledge Analysis (WEKA) เวอร์ชั่น 3.5.3 (Frank *et al.*, 2004)
    2.3 เวบแอพพลิเคชั่นที่ให้บริการฐานข้อมูลด้านโปรตีนและเอนไซม์บนอินเตอร์เน็ท

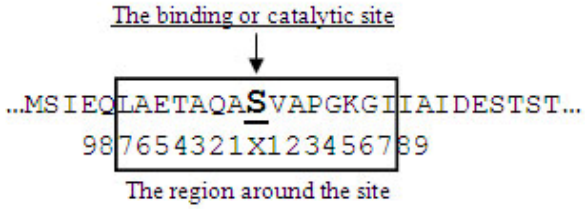
วิธีการ

การพัฒนาตัวแทนสายโปรตีนประเภทโมทีฟเพื่อทำนายประเภทฟังก์ชั่นเอนไซม์ มีเนื้องานหลักคือการค้นพบโมทีฟจากข้อมูลบริเวณจับและบริเวณเร่งที่เราเรียกว่า รีแอคทีฟโมทีฟ (reactive motif) ดังนั้นในรายละเอียดของวิธีการจึงเริ่มกล่าวจาก *1) การเตรียมข้อมูลพื้นฐานจากบริเวณจับและบริเวณเร่ง* จากนั้นจึงอธิบาย *2) หลักการพื้นฐานในการค้นพบรีแอคทีฟโมทีฟจากบริเวณจับและบริเวณเร่ง* โดยในภาพรวมของงานวิจัยนี้แบ่งออกเป็นสองส่วนงานหลักคือ *3) ส่วนงานที่ 1 การค้นพบและพัฒนารีแอคทีฟโมทีฟ* และ*4) ส่วนงานที่ 2 ระบบทำนายประเภทฟังก์ชั่นเอนไซม์ที่สร้างจากรีแอคทีฟโมทีฟ* ซึ่งหลังจากอธิบายภาพรวมแล้ว จะได้อธิบายในเนื้องานสำคัญที่นำไปสู่การค้นพบโมทีฟคือ *5) การค้นพบกลุ่มแทนที่ที่สมบูรณ์จากการประมวลผลการควบคุมการกลายพันธุ์บนฐานคอนเซ็ปต์แลททิส* ซึ่งเป็นหัวใจของการค้นพบโมทีฟจากข้อมูลบลอกในบริเวณจับและบริเวณเร่งที่มีข้อมูลน้อย โดยมี *6) การพัฒนาคุณภาพบลอก* เป็นเทคนิคที่ทำให้ข้อมูลบลอกในบริเวณจับและบริเวณเร่งที่มีน้อยมากๆ ให้สามารถใช้งานสำหรับค้นพบกลุ่มแทนที่และโมทีฟได้

โดยการขยายกลุ่มข้อมูลด้วยวิธีทางสถิติ และ *7) การแปลงรูปฟอร์มความรู้พื้นหลัง* ซึ่งทำให้ความรู้พื้นหลังที่มีอยู่ในปัจจุบันในรูปฟอร์มของคอนเท็กซ์และตารางคะแนนความเหมือนสามารถนำมาประยุกต์ใช้กับงานวิจัยนี้ได้ โดยเนื้องานทั้งหมดจะได้กล่าวถึงตามลำดับดังนี้

1. การเตรียมข้อมูลพื้นฐาน

ข้อมูลสายลำดับโปรตีนที่เป็นข้อมูลพื้นฐานของงานวิจัยจัดเตรียมจากส่วนข้อมูล SWISSPROT (Bairoch and Apweiler, 2000) ในฐานข้อมูล UNIPROT เวอร์ชั่น 9.2 (UNIPROT, 2005) และจัดเตรียมข้อมูลฟังก์ชั่นเอนไซม์ตามมาตรฐานของ ENZYME NOMENCLATURE (Nomenclature Committee of the International Union of Biochemistry and Molecular Biology, 1992) ในฐานข้อมูล ENZYME NOMENCLATURE เวอร์ชั่น 37.0 ขององค์กร SWISS Institution of Bioinformatics (SIB) จากนั้นคัดเลือกเฉพาะโปรตีนที่มีคุณสมบัติเอนไซม์เป็นชุดข้อมูลสายลำดับโปรตีนเอนไซม์ (Enzyme Sequence Dataset) ซึ่งมีข้อมูลบางระเบียนที่ระบุตำแหน่งการทำงานของบริเวณจับและบริเวณเร่งของเอนไซม์ โดยแยกออกมาเป็นอีกฐานข้อมูลหนึ่งคือ ฐานข้อมูลบริเวณจับและบริเวณเร่ง (Binding and Catalytic Site Database) ประกอบด้วยข้อมูลสายลำดับกรดอะมิโนในบริเวณจับและบริเวณเร่ง ที่เตรียมขึ้นจากตำแหน่งของบริเวณจับ (หรือบริเวณเร่ง) และบริเวณรอบๆ รวมทั้งหมด 15 ตำแหน่งดังแสดงในภาพที่ 14



ภาพที่ 14 การเตรียมข้อมูลสายลำดับโปรตีนบริเวณจับและบริเวณเร่ง

แต่ละบริเวณจับและบริเวณเร่งที่มีความหมายกลไกการทำงานฟังก์ชั่นเอนไซม์เหมือนกันจะจัดให้อยู่ในกลุ่มย่อยเดียวกันในฐานข้อมูลบริเวณจับและบริเวณเร่ง โดยความหมายเหล่านี้ได้จากการระบุคุณลักษณะสายโปรตีนในฐานข้อมูล SWISSPROT ซึ่งนำมาจัดเก็บเป็นข้อมูลประกอบสายลำดับกรดอะมิโนในฐานข้อมูลบริเวณจับและบริเวณเร่ง โดยในกรณีของบริเวณจับประกอบด้วย 4 แอททริบิวต์คือ คำจำกัดความการทำงาน, สารตั้งต้น (substrate), วิธีการเข้าจับ (เช่น via amide

nitrogen), และชนิดของกรดอะมิโนที่ใช้เข้าจับ ส่วนในกรณีของบริเวณเร่งประกอบด้วย 3 แอททริบิวต์คือ คำจำกัดความการทำงาน, ลักษณะกลไกทำงาน (เช่น proton acceptor), และชนิดของกรดอะมิโนที่ทำงาน รวมเป็นข้อมูลในฐานข้อมูลบริเวณจับและบริเวณเร่งทั้งหมด 291 กลุ่มกลไกการทำงานฟังก์ชั่นเอนไซม์ ครอบคลุมข้อมูลบริเวณจับและบริเวณเร่งทั้งหมด 3,084 ระเบียน

ในการคัดเลือกสายลำดับโปรตีนเอนไซม์เพื่อใช้ในการสร้างโมเดลการทำนายประเภทฟังก์ชั่นจากชุดข้อมูลสายลำดับโปรตีนเอนไซม์นั้น เริ่มจากการกรองเอาโปรตีนที่มีข้อมูลบริเวณจับหรือบริเวณเร่งเพียง 1 ชนิดออกไป เนื่องจากการระบุว่าโปรตีนตัวใดทำงานฟังก์ชั่นเอนไซม์ใดจะต้องมีการระบุอย่างน้อย 2 บริเวณจับหรือบริเวณเร่ง นอกจากนี้การคัดเลือกกลุ่มฟังก์ชั่นเอนไซม์ที่มีข้อมูลสายลำดับโปรตีนอยู่ในช่วงข้อมูล 10 ถึง 1,000 ระเบียน ได้เป็นชุดข้อมูลสายลำดับโปรตีนเอนไซม์ที่คัดเลือกแล้วทั้งหมด 19,258 ระเบียนในเอนไซม์ 235 ฟังก์ชั่น

2. หลักการพื้นฐานในการค้นพบรีแอคทีฟโมทีฟจากบริเวณจับและบริเวณเร่ง

ในการค้นพบรีแอคทีฟโมทีฟจากบริเวณจับและบริเวณเร่ง 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 $x_{ij} \in \Sigma$, Σ = {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 ตำแหน่งของกรดอะมิโนในคอลัมน์ของเมท

ริกซ์บลอก ทั้งนี้ $x_{ij} \in \Sigma, \Sigma$ = {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 $\forall\ s_i \in B_{mxn}$, $x_{ij} \in \Sigma$, และ Σ = {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 $x_{ij} \in \Sigma$, และ Σ = {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 $\hat{A}_j \subseteq \Sigma$, และ Σ = {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 $x_{ij} \in \Sigma$, และ Σ = {20 ชนิดของกรดอะมิโน}

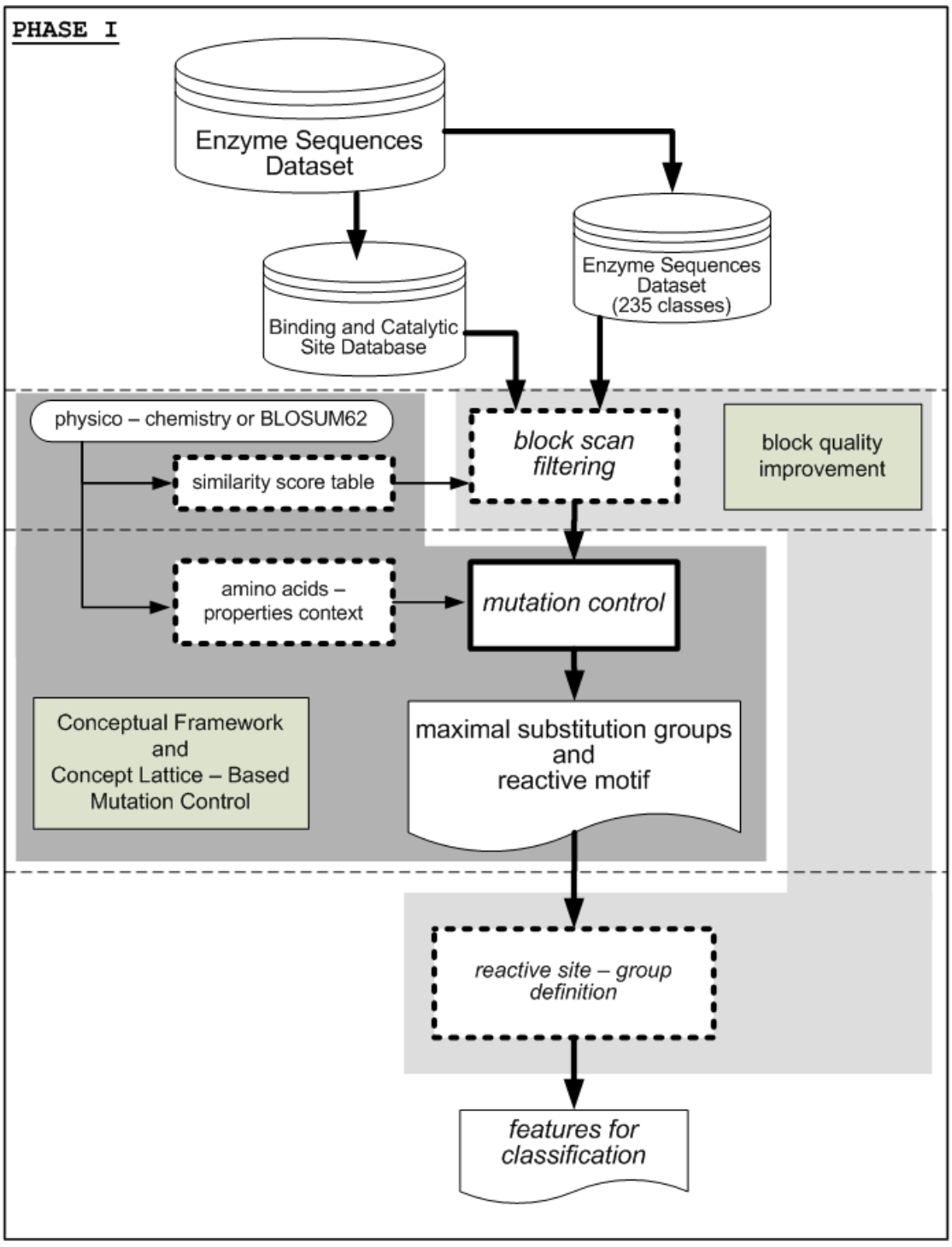
จากสูตรการค้นพบรีแอคทีฟโมทีฟจากบลอกในบริเวณจับหรือบริเวณเร่ง M = $\bigcup_{i=1}^{m} x_{i1}, \bigcup_{i=1}^{m} x_{i2}, \bigcup_{i=1}^{m} x_{i3}, .., \bigcup_{i=1}^{m} x_{in}$ สามารถวิเคราะห์ปัญหาเบื้องต้นของโมทีฟที่ได้จากบลอกก็คือ ที่ขนาดของโมทีฟ n = 15 คงที่ พบว่าคุณภาพของกลุ่มแทนที่ $\hat{A}_j$ หรือ $\bigcup_{i=1}^{m} x_{ij}$ ขึ้นกับจำนวนสายลำดับกรดอะมิโนในบลอกหรือค่า m นั่นเอง ซึ่งในกรณีที่ค่า m = 1 และแต่ละตำแหน่งของโมทีฟสามารถเป็นกรดอะมิโนได้ 20 ชนิด จะได้ค่าความน่าจะเป็นในการค้นพบโมทีฟ M ดังกล่าวในสายโปรตีนขนาดความยาว y อยู่ที่ประมาณ y $(\frac{1}{20})^{15}$ ซึ่งโอกาสที่จะพบโมทีฟมีน้อยมาก (แม้ในฐานข้อมูลสาย

โปรตีนขนาดใหญ่) ทำให้นำโมทีฟนั้นไปใช้ประโยชน์ได้น้อย ในทางตรงกันข้ามในบลอกขนาดใหญ่ที่มีค่า m มาก จะทำให้กลุ่มแทนที่ในทุกตำแหน่ง j มีค่าใกล้เคียง Σ ซึ่งก็คือมีความน่าจะเป็นในการค้นพบโมทีฟนั้นในทุกสายโปรตีน

ดังนั้นเนื้องานหลักในการได้มาซึ่งรีแอคทีฟโมทีฟที่มีคุณภาพจึงขึ้นอยู่กับการพัฒนากลุ่มแทนที่ที่มีคุณภาพที่เราเรียกว่า “กลุ่มแทนที่ที่สมบูรณ์” (Maximal Substitution Group) และ “การพัฒนาคุณภาพบลอก” ซึ่งทำให้ได้บลอกที่มีขนาดเหมาะสมในการได้มาซึ่งโมทีฟที่มีคุณภาพ ซึ่งความสัมพันธ์ระหว่างขั้นตอนต่างๆ ในการค้นพบและพัฒนารีแอคทีฟโมทีฟเพื่อทำนายประเภทฟังก์ชั่นเอนไซม์จะได้อธิบายให้เข้าใจภาพรวมใน 2 ส่วนงานหลักก่อน คือ ส่วนงานที่ 1 การค้นพบและพัฒนาคุณภาพรีแอคทีฟโมทีฟ และ ส่วนงานที่ 2 ระบบทำนายประเภทฟังก์ชั่นเอนไซม์ที่สร้างจากรีแอคทีฟโมทีฟ ก่อนอธิบายเจาะลึกลงไปในรายละเอียดที่มีมากของเนื้องานหลัก คือ “การค้นพบกลุ่มแทนที่ที่สมบูรณ์” และ “การพัฒนาคุณภาพบลอก” ในภายหลัง

3. ส่วนงานหลักที่ 1 การค้นพบและพัฒนาคุณภาพรีแอคทีฟโมทีฟ

จากภาพที่ 15 แสดงภาพรวมในการค้นพบและพัฒนาคุณภาพรีแอคทีฟโมทีฟ เริ่มจากการใช้ข้อมูลพื้นฐานที่จัดเตรียมไว้คือชุดข้อมูลสายลำดับโปรตีนเอนไซม์และฐานข้อมูลบริเวณจับและบริเวณเร่ง โดยหัวใจของส่วนงานหลักที่ 1 นี้ เป็นขั้นตอนการค้นพบองค์ประกอบย่อยของโมทีฟจากข้อมูลบลอกในบริเวณจับและบริเวณเร่ง คือ *กลุ่มแทนที่ที่สมบูรณ์* (Maximal Substitution Group) ที่อิง *แนวคิดการควบคุมการกลายพันธุ์* (mutation control) ที่ใช้ *“ความสัมพันธ์เชิงวิทยาศาสตร์ระหว่างรีแอคทีฟโมทีฟกับประเภทฟังก์ชั่นเอนไซม์”* ในการประมวลความรู้พื้นหลังลักษณะต่างๆ คือ *ตารางแสดงความเหมือน* (similarity score table) และ *คอนเท็กซ์แสดงความสัมพันธ์กรดอะมิโนกับคุณสมบัติกรดอะมิโน* (amino acids-properties context) ให้อยู่ในรูป *คอนเซ็ปต์* โดยกรอบงานนี้แสดงในโซนสีเทาแก่เรียกว่า *กรอบงานเชิงคอนเซ็ปต์* (Conceptual Framework) ที่นำ *ความรู้ในรูปของคอนเซ็ปต์* เหล่านี้มาประมวลผล *การควบคุมการกลายพันธุ์บนฐานคอนเซ็ปต์แลททิส* (Concept Lattice – Based Mutation Control) ได้ผลออกมาเป็น กลุ่มแทนที่ที่สมบูรณ์ ซึ่งนำมาประกอบรวมกันเป็น *รีแอคทีฟโมทีฟ*



ภาพที่ 15 ภาพรวมส่วนงานหลักที่ 1 การค้นพบและพัฒนาคุณภาพรีแอคทีฟโมทีฟ

ทั้งนี้ ประสิทธิภาพของการประมวลผลเพื่อให้ได้กลุ่มแทนที่ที่สมบูรณ์ ขึ้นอยู่กับคุณภาพของบลอก จึงมีขั้นตอน *การพัฒนาคุณภาพของบลอก* ที่เป็นกระบวนการก่อนและหลังขั้นตอนการ

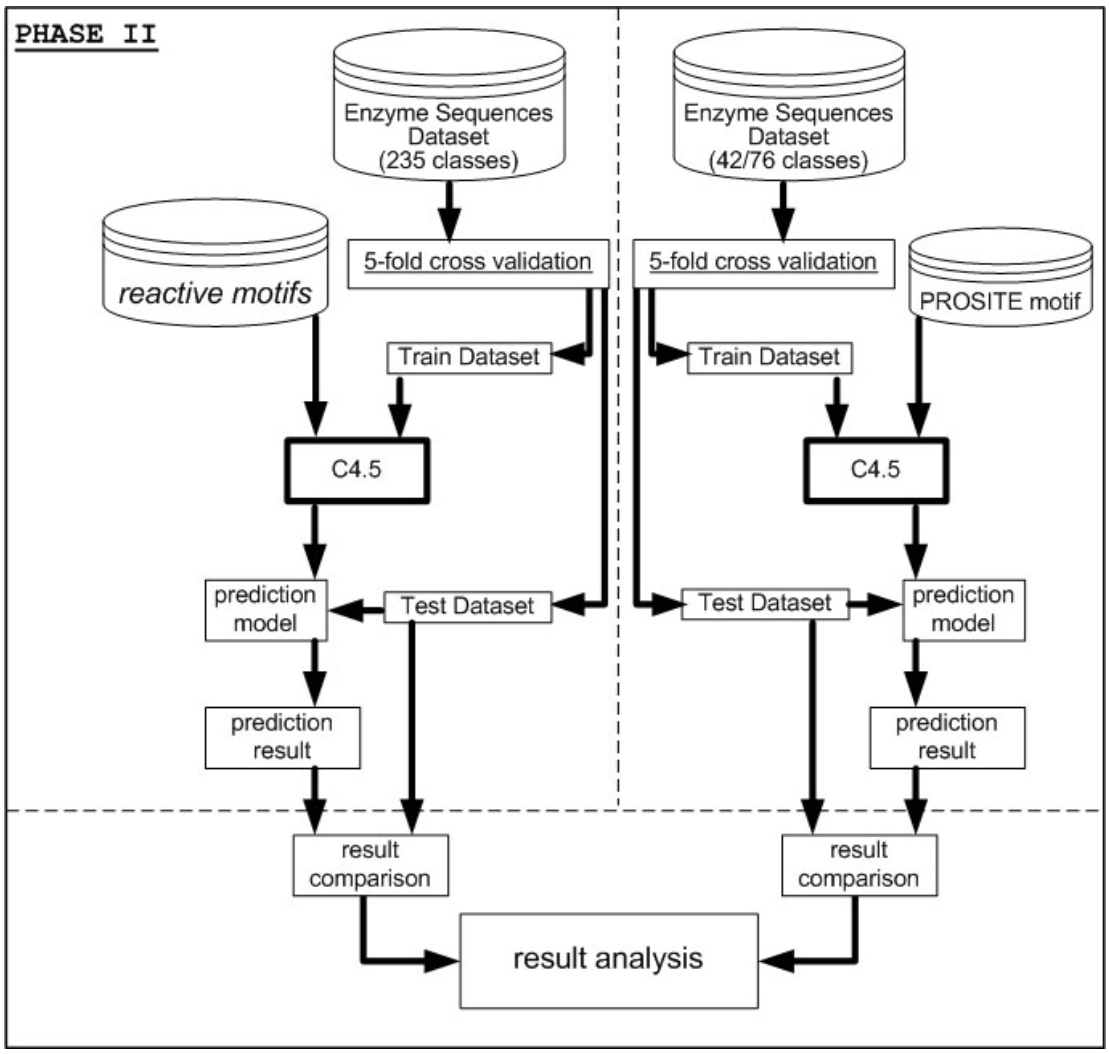
ควบคุมการกลายพันธุ์ (โซนสีเทาอ่อน) คือ *การคัดสรรบลอกที่มีคุณภาพ* (block scan filtering) และ *การจัดกลุ่มบริเวณปฏิกิริยาชีวเคมี* (reactive site – group definition) ได้ผลสุดท้ายเป็นกลุ่มรีแอคทีฟโมทีฟ เป็นคุณลักษณะเด่นของกลุ่มสายโปรตีนข้อมูลนำเข้าสำหรับใช้พัฒนาระบบทำนายประเภทฟังก์ชั่นเอนไซม์ โดยภาพรวมของการพัฒนาระบบทำนายประเภทฟังก์ชั่นเอนไซม์จะได้กล่าวถึงในลำดับถัดไป ก่อนที่จะอธิบายรายละเอียดในเนื้องานส่วนต่างๆ ที่กล่าวมาคือ *การค้นพบกลุ่มแทนที่ที่สมบูรณ์จากประมวลผลการควบคุมการกลายพันธุ์บนฐานคอนเซ็ปต์แลททิส* และ *การพัฒนาคุณภาพบลอก* ในลำดับต่อจากนั้น

4. ส่วนงานหลักที่ 2 ระบบทำนายประเภทฟังก์ชั่นเอนไซม์ที่สร้างจากรีแอคทีฟโมทีฟ

เนื่องจากแต่ละฟังก์ชั่นเอนไซม์เกิดจากการทำงานร่วมกันของบริเวณจับและบริเวณเร่ง การใช้รีแอคทีฟโมทีฟจากบริเวณเดียวจึงไม่สามารถใช้ทำนายฟังก์ชั่นเอนไซม์ได้ อีกทั้งในงานวิจัยนี้ยังใช้รีแอคทีฟโมทีฟจากบริเวณจับและบริเวณเร่งจำนวนมากที่เป็นผลจากส่วนงานหลักที่ 1 ดังนั้นจึงจำเป็นต้องการเครื่องมืออัตโนมัติในการแก้ปัญหาความซับซ้อนของการทำงานร่วมกันระหว่างรีแอคทีฟโมทีฟในการทำนายประเภทฟังก์ชั่นเอนไซม์ โดยในวิทยานิพนธ์นี้เลือกหนึ่งในเทคนิคที่รู้จักกันดีก็คือเครื่องจักรเรียนรู้ที่เรียกว่า C4.5 (Quinlan, 1993) ดังแสดงภาพรวมของการพัฒนาระบบดังกล่าวในภาพที่ 16

จากภาพที่ 16 เริ่มต้นจากกลุ่มข้อมูลสายโปรตีนเอนไซม์ 2 กลุ่มที่ใช้กับรีแอคทีฟโมทีฟกับโมทีฟ PROSITE จากนั้นแต่ละกลุ่มโปรตีนถูกนำไปจัดกลุ่มฝึกสอนและทดสอบด้วยวิธี 5-fold cross validation โดยกลุ่มฝึกสอนนำไปสร้างโมเดลการทำนายประเภทเอนไซม์ด้วย C4.5 ก่อนที่จะนำโมเดลนั้นมาทดสอบความแม่นยำด้วยกลุ่มข้อมูลทดสอบ โดยวิเคราะห์ผลการทำนายประเภทเอนไซม์ของกลุ่มโปรตีนที่ใช้รีแอคทีฟโมทีฟเปรียบเทียบกับผลการทำนายประเภทเอนไซม์ของกลุ่มโปรตีนที่ใช้โมทีฟ PROSITE ในท้ายสุด

ในส่วนงานการพัฒนาระบบทำนายประเภทฟังก์ชั่นเอนไซม์นี้จะได้กล่าวถึงรายละเอียดตามลำดับคือ การเตรียมข้อมูลสำหรับเปรียบเทียบโมเดลการทำนายฟังก์ชั่นเอนไซม์, การเตรียมข้อมูลฝึกสอนและทดสอบโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์, และเครื่องมือที่ใช้สร้างโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์



ภาพที่ 16 ภาพรวมส่วนงานหลักที่ 2 ระบบทำนายประเภทฟังก์ชั่นเอนไซม์ที่สร้างจากรีแอคทีฟโมทีฟ

4.1 การเตรียมข้อมูลสำหรับเปรียบเทียบโมเดลทำนายประเภทฟังก์ชั่นเอนไซม์

ในการเตรียมข้อมูลสำหรับเปรียบเทียบโมเดลทำนายประเภทฟังก์ชั่นเอนไซม์ เป็นการเปรียบเทียบระหว่างกลุ่มข้อมูลที่ใช้รีแอคทีฟโมทีฟกับ PROSITE ซึ่งกลุ่มข้อมูลสายโปรตีนที่ใช้กับรีแอคทีฟเป็นกลุ่มข้อมูลเดียวกับชุดข้อมูลสายลำดับโปรตีนเอนไซม์ที่จัดเตรียมไว้ตั้งแต่ก่อนเริ่มส่วนงานหลักที่ 1 แต่สำหรับกลุ่มข้อมูลสายโปรตีนที่ใช้กับ PROSITE เพื่อเปรียบเทียบกันนี้ รวบรวมขึ้นใหม่โดยจาก 152 โมทีฟของ PROSITE ที่เป็นรูปแบบสายลำดับในบริเวณจับและบริเวณเร่ง ได้ถูกสำรวจว่าโมทีฟเหล่านี้มีอยู่ในฟังก์ชั่นใดบ้าง จากนั้นใช้เงื่อนไขเดียวกับชุดข้อมูล

สายลำดับโปรตีนเอนไซม์คือ เลือกเฉพาะฟังก์ชั่นที่พบโมทีฟมากกว่า 1 ชนิด และแต่ละกลุ่มฟังก์ชั่นเอนไซม์มีโปรตีนเอนไซม์เป็นสมาชิกอยู่ระหว่าง 10 ถึง 1,000 ชนิด ได้กลุ่มข้อมูลสายโปรตีนที่ใช้กับ PROSITE เพียง 42 ฟังก์ชั่น ครอบคลุมโปรตีนเพียง 2,579 ชนิด และโมทีฟ 36 โมทีฟ ซึ่งเป็นกลุ่มข้อมูลขนาดเล็ก อีกทั้งทดสอบแล้วมีความแม่นยำที่ต่ำเกินไป จึงได้ปรับแต่งกลุ่มข้อมูลนี้ใหม่โดยให้แต่ละกลุ่มฟังก์ชั่นมีโปรตีนเป็นสมาชิกอยู่ระหว่าง 5 ถึง 1,000 ชนิด ครอบคลุม 65 โมทีฟ ใน 76 ฟังก์ชั่น รวมเป็นโปรตีนทั้งหมด 2,815 ชนิด

4.2 การเตรียมข้อมูลฝึกสอนและทดสอบโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์

ในการเตรียมข้อมูลฝึกสอนและข้อมูลทดสอบของโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์ในแต่ละโมเดลนั้น ใช้วิธี n-fold cross validation (Mosteller and Tukey, 1968) ในการได้มาซึ่งผลความแม่นยำการทำนายของแต่ละโมเดล โดยค่า n ที่เลือกใช้ก็คือ n = 5 เนื่องจากสมาชิกสายโปรตีนในกลุ่มฟังก์ชั่นเอนไซม์ที่มีขนาดต่ำสุดมี 5 ชนิด

ในการฝึกสอนและทดสอบโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์ จะต้องรวมข้อมูลโมทีฟกับสายโปรตีนเอนไซม์เข้าด้วยกัน โดยในวิทยานิพนธ์นี้ใช้วิธีการ (Diplaris *et al.*, 2005) โดยการนำเสนอข้อมูลสายโปรตีนทุกเส้นจะถูกแปลงให้อยู่ในฐานการนำเสนอแอททริบิวต์ในรูปของตารางที่มีขนาดเซ็ทของแอททริบิวต์เท่ากัน ซึ่งในที่นี้แต่ละแอททริบิวต์ก็คือรีแอคทีฟโมทีฟในกลุ่มที่เป็นบริเวณจับและบริเวณเร่งเดียวกัน (same site description) สายโปรตีนใดที่ตรวจสอบแล้วพบสายลำดับกรดอะมิโนที่สอดคล้องกับรีแอคทีฟโมทีฟที่อยู่ในบริเวณจับหรือบริเวณเร่ง (site description) ใด ให้กำกับค่าเซลจุดตัดระหว่างสายโปรตีนกับรีแอคทีฟโมทีฟนั้นในตารางด้วยค่า 1 นอกนั้นใส่เป็นค่า 0 หรือไม่พบรีแอคโมทีฟเหล่านั้น จากนั้นปิดท้ายแอททริบิวต์ในตำแหน่งสุดท้ายด้วยประเภทของฟังก์ชั่นเอนไซม์ของสายโปรตีนนั้น ตัวอย่างการจัดเตรียมข้อมูลแสดงในภาพที่ 17

จากภาพที่ 17 แสดงสายโปรตีน ID:1 ที่มีฟังก์ชั่น EC 3.2.1.1 พบลำดับกรดอะมิโนที่สอดคล้องกับรีแอคทีฟโมทีฟ M1 และ M2 โดย M1 เป็นบริเวณจับ:1 (site description:1) ซึ่งสามารถนำเสนอข้อมูลดังกล่าวในรูปของเมทริกซ์ โดยกำหนดให้เซลในแถวของสายโปรตีน ID:1 และในคอลัมน์ของบริเวณจับ:1 มีค่าเป็น 1 เป็นต้น

Protein ID	Reactive Motif / PROSITE						Enzyme Function (EC number)
	Site Description 1	Site Description 2	Site Description 3	Site Description 4	....	Site Description N	
1	1	1	0	0	....	0	3.2.1.1
2	1	0	1	0	....	0	3.2.1.1
3	0	0	0	1	....	1	4.1.2.13
....	....	....	....	....	....	....	....
M	0	1	0	0	....	1	2.6.1.1

ภาพที่ 17 การนำเสนอข้อมูลสายโปรตีนในรูปฐานนำเสนอแอททริบิวต์เป็นรีแอคทีฟโมทีฟที่จัดกลุ่มตามบริเวณจับและบริเวณเร่งเดียวกันและฟังก์ชั่นเอนไซม์

4.3 เครื่องมือที่ใช้สร้างโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์

ในวิทยานิพนธ์นี้ใช้ชุดเครื่องมือเครื่องจักรเรียนรู้ที่เรียกว่า WEKA (Frank *et al.*, 2004) ซึ่งมีโปรแกรม C4.5 (ให้บริการในโมดูล J4.8 ของโปรแกรม WEKA) ในการสร้างโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์แต่ละโมเดลที่ได้จากข้อมูลฝึกสอนและทดสอบแบบ 5-fold cross-validation ดังอธิบายในหัวข้อก่อนหน้านี้เพื่อเปรียบเทียบผลความแม่นยำของโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์จากรีแอคทีฟโมทีฟและโมทีฟ PROSITE ในลักษณะต่างๆ

5. การค้นพบกลุ่มแทนที่ที่สมบูรณ์จากประมวลผลการควบคุมการกลายพันธุ์บนฐานคอนเซ็ปต์แลททิส

ในหัวข้อนี้แสดงรายละเอียดขั้นตอนวิธีในการค้นพบ “กลุ่มแทนที่ที่สมบูรณ์” (Maximal Substitution Group) ซึ่งเป็นองค์ประกอบย่อยของรีแอคทีฟโมทีฟและนำไปสู่การค้นพบรีแอคทีฟโมทีฟที่เป็นเนื้อหาที่สำคัญที่สุดของงานวิจัยนี้ โดยมุ่งเน้นไปที่การแก้ปัญหาการมีข้อมูลบริเวณจับและบริเวณเร่งน้อย ทำให้หัวใจของงานคือการนำ “ความรู้” ในงานวิจัยอื่นเข้ามาช่วยในการประมวลผลเพื่อให้ได้กลุ่มแทนที่ที่สมบูรณ์

ความรู้ที่สำคัญที่สุดคือ *"ความสัมพันธ์เชิงวิทยาศาสตร์ระหว่างคุณลักษณะเด่นของกลุ่มข้อมูลนำเข้ากับประเภทข้อมูลที่ใช้สร้างโมเดลการทำนายประเภทข้อมูล"* ทั้งนี้เนื่องจากสามารถใช้การอธิบายเชิง "วิทยาศาสตร์" ดังกล่าวมาใช้ในงานพัฒนาคุณภาพของคุณลักษณะเด่นที่ใช้กับกลุ่มข้อมูลนำเข้าได้อย่างแน่นอน โดยไม่ต้องสร้างเกณฑ์วัดคุณภาพของสิ่งที่นำมาใช้พัฒนาคุณลักษณะเด่นของกลุ่มข้อมูลนำเข้าเหมือนกับการใช้ข้อมูลลักษณะอื่น ซึ่งยากต่อการหาข้อสรุปที่ชัดเจนเมื่อกลุ่มข้อมูลมีการเปลี่ยนแปลงไป

ในกรณีของฟังก์ชั่นเอนไซม์ ความสัมพันธ์ระหว่างคุณลักษณะเด่นที่ได้จากบริเวณจับและบริเวณเร่งกับประเภทฟังก์ชั่นเอนไซม์ ก็คือ *แนวคิดการควบคุมการกลายพันธุ์* (mutation control) และ *กลไกการเกิดฟังก์ชั่นเอนไซม์* (enzyme mechanism) ซึ่งอธิบายได้ด้วย *ความรู้พื้นหลัง* ที่อยู่ในรูปฟอร์ม *คอนเท็กซ์* (context) และแปลงความรู้ดังกล่าวให้อยู่ในรูปฟอร์มของ *คอนเซ็ปต์* (concepts) ก่อนที่จะนำเสนอการประมวลผลด้วย *ทฤษฎีคอนเซ็ปต์แลททิส* (Concept Lattice Theory) ได้ผลเป็นกลุ่มแทนที่ที่สมบูรณ์ในท้ายสุด (Liewlom, 2008)

ขั้นตอนในการแปลงความรู้ให้อยู่ในรูปของคอนเซ็ปต์ เรียกว่า *กรอบงานเชิงคอนเซ็ปต์* (Conceptual Framework) และการใช้คอนเซ็ปต์แลททิสในการประมวลผลกลุ่มแทนที่ที่สมบูรณ์ที่อิงจากแนวคิดการควบคุมการกลายพันธุ์ เรียกว่า *การประมวลผลการควบคุมการกลายพันธุ์บนฐานคอนเซ็ปต์แลททิส* (Concept Lattice – Based Mutation Control) โดยอธิบายตามลำดับดังนี้

5.1 กรอบงานเชิงคอนเซ็ปต์ (conceptual framework)

ขั้นตอนนี้มีจุดประสงค์ในการแปลงความสัมพันธ์เชิงวิทยาศาสตร์ระหว่างคุณลักษณะเด่นของกลุ่มข้อมูลนำเข้า (รีแอคทีฟโมทีฟ) กับประเภทของข้อมูล (ฟังก์ชั่นเอนไซม์) ซึ่งก็คือ แนวคิดการควบคุมการกลายพันธุ์ และ กลไกการเกิดฟังก์ชั่นเอนไซม์ ให้อยู่ในรูปของ คอนเซ็ปต์ ดังนั้นในส่วนนี้จะได้กล่าวถึงความรู้ดังกล่าวตามด้วยขั้นตอนการแปลงความรู้ให้เป็นคอนเซ็ปต์ดังนี้

5.1.1 แนวคิดการควบคุมการกลายพันธุ์ และกลไกการเกิดฟังก์ชั่นเอนไซม์

จากข้อเท็จจริงที่แต่ละสายลำดับกรดอะมิโนภายในบลอกของบริเวณจับหรือบริเวณเร่งเดียวกันสามารถมีลำดับที่แตกต่างกันได้ แต่ยังคงความสามารถในการเกิดฟังก์ชั่นเอนไซม์ได้เหมือนกัน ดังนั้นการกลายพันธุ์ในบริเวณจับและบริเวณเร่งจึงไม่ได้เกิดขึ้นอย่างสุ่ม (random) แต่มีการควบคุม (control) ทิศทางการกลายพันธุ์ให้ยังคงเกิดกลไกฟังก์ชั่นเอนไซม์เดียวกันได้ เราเรียกแนวคิดนี้ว่า การควบคุมการกลายพันธุ์ (mutation control) โดยการค้นพบกลุ่มแทนที่ที่สมบูรณ์จากแนวคิดดังกล่าวเป็นไปตามสมมติฐานที่ 1

<u>สมมติฐานที่ 1</u>: ถ้าเราสามารถแจกแจงลักษณะการควบคุมการกลายพันธุ์ที่ซ่อนอยู่ในแต่ละตำแหน่งของบลอกได้ เราย่อมสามารถหากลุ่มแทนที่ที่สมบูรณ์ที่เป็นองค์ประกอบของรีแอคทีฟโมทีฟได้

การควบคุมการกลายพันธุ์มีเป้าหมายที่การรักษาความสามารถในการเกิดกลไกฟังก์ชั่นเอนไซม์ซึ่งมีอยู่ด้วยกัน 2 ลักษณะ (Patrick, 1995) คือ *กลไกที่ทำให้เกิดฟังก์ชั่นเอนไซม์* (supporter) และ*กลไกที่ไม่ยับยั้งการเกิดฟังก์ชั่นเอนไซม์* นั้น (not inhibitor) ดังนั้นถ้าเราสามารถแจกแจงลักษณะของกลไกดังกล่าวได้ ย่อมสามารถหากลุ่มแทนที่ที่สมบูรณ์ตามสมมติฐานที่ 1 ได้

การแจกแจงความรู้เหล่านี้ จำเป็นต้องมีเครื่องมือทางทฤษฎีการประมวลผลสนับสนุน โดยการแจกแจงความรู้เหล่านี้สามารถทำให้อยู่ในรูปของ *คอนเซ็ปต์* ได้ ซึ่งทำให้สามารถใช้งานทฤษฎีคอนเซ็ปต์แลททิสได้ ดังนั้น เป้าหมายในการทำงานของกรอบงานเชิงคอนเซ็ปต์ก็คือ การแปลงแนวคิดการควบคุมกลายพันธุ์ กลไกที่ทำให้เกิดฟังก์ชั่นเอนไซม์ และกลไกไม่ยับยั้งการเกิดฟังก์ชั่นเอนไซม์ ให้อยู่ในรูปฟอร์มของ *คอนเซ็ปต์การควบคุมการกลายพันธุ์* (mutation control concept) *คอนเซ็ปต์สนับสนุน* (supporter concept) และ *คอนเซ็ปต์ไม่ยับยั้ง* (not inhibitor concept หรือ (~inhibitor concept) ดังที่จะได้กล่าวถึงต่อไป

5.1.2 ขั้นตอนการแปลงความสัมพันธ์ระหว่างคุณลักษณะเด่นของกลุ่มข้อมูลกับประเภทข้อมูลให้เป็นคอนเซ็ปต์

การแปลงความสัมพันธ์ระหว่างคุณลักษณะเด่นของกลุ่มสายโปรตีน (reactive motif) กับประเภทข้อมูล (ฟังก์ชั่นเอนไซม์) ให้เป็นคอนเซ็ปต์ ประกอบด้วยขั้นตอนย่อย 2 ขั้นตอน คือ 1) การกำหนดรูปฟอร์มของคอนเซ็ปต์ และ 2) การกำหนดลักษณะคอนเท็กซ์ที่ใช้อธิบายคอนเซ็ปต์ ซึ่งสามารถนำไปประมวลผลด้วยทฤษฎีคอนเซ็ปต์แลททิสต่อไปได้ รายละเอียดแต่ละขั้นตอนมีดังนี้

ก. การกำหนดรูปฟอร์มของคอนเซ็ปต์

ในการแปลงความรู้เป็นคอนเซ็ปต์ กำหนดให้ใช้คอนเซ็ปต์ที่อยู่ในรูป formal concept ตามนิยามที่ 2 คือเป็นคู่ลำดับของสิ่งของ (extent) และรายละเอียดของสิ่งของ (intent) หรือเขียนในรูปฟอร์ม (extent, intent) ซึ่งในกรณีของคอนเซ็ปต์การควบคุมการกลายพันธุ์ คอนเซ็ปต์สนับสนุน และคอนเซ็ปต์ไม่ยับยั้ง สามารถเขียนให้อยู่ในรูปฟอร์มเบื้องต้นของคอนเซ็ปต์ก่อนกำหนดรายละเอียดที่ถูกต้องต่อไปดังนี้

นิยามที่ 8 คอนเซ็ปต์สนับสนุน (supporter concept): สามารถเขียนให้อยู่ในรูปฟอร์มของ (extent, intent) ตามนิยามของคอนเซ็ปต์ในนิยามที่ 2 ได้ในเบื้องต้นคือ คอนเซ็ปต์สนับสนุน = (เซ็ทกรดอะมิโน, การแจกแจงปัจจัยกลไกเอนไซม์ที่มีร่วมกันของเซ็ทกรดอะมิโน) ซึ่งมีลักษณะเฉพาะในแต่ละตำแหน่ง j ของ บลอก B ที่ได้จากบริเวณจับและบริเวณเร่ง

นิยามที่ 9 คอนเซ็ปต์ไม่ยับยั้ง (not inhibitor concept): สามารถเขียนให้อยู่ในรูปฟอร์มของ (extent, intent) ตามนิยามของคอนเซ็ปต์ในนิยามที่ 2 ได้ในเบื้องต้นคือคอนเซ็ปต์ไม่ยับยั้ง = (เซ็ทกรดอะมิโน, การแจกแจงปัจจัยการไม่ยับยั้งกลไกเอนไซม์ที่มีร่วมกันของเซ็ทกรดอะมิโน) ซึ่งมีลักษณะเฉพาะในแต่ละตำแหน่ง j ของ บลอก B ที่ได้จากบริเวณจับและบริเวณเร่ง

นิยามที่ 10 คอนเซ็ปต์การควบคุมการกลายพันธุ์ (mutation control concept): สามารถเขียนให้อยู่ในรูปฟอร์มของ (extent, intent) ตามนิยามของคอนเซ็ปต์ในนิยามที่ 2 ได้ในเบื้องต้นคือคอนเซ็ปต์การควบคุมการกลายพันธุ์ = (กลุ่มแทนที่ที่สมบูรณ์, การแจกแจงปัจจัยกลไก

การเกิดฟังก์ชั่นเอนไซม์และไม่ยับยั้งฟังก์ชั่นเอนไซม์ที่มีร่วมกันของกลุ่มแทนที่ที่สมบูรณ์) ซึ่งมีลักษณะเฉพาะในแต่ละตำแหน่ง j ของ บลอก B ที่ได้จากบริเวณจับและบริเวณเร่ง โดยมีลักษณะเฉพาะของทั้งคอนเซ็ปต์สนับสนุนและคอนเซ็ปต์ไม่ยับยั้งปรากฏอยู่

ในการแจกแจงกลไกเอนไซม์ต่างๆ ในทั้ง 3 คอนเซ็ปต์นี้ จำเป็นต้องระบุลักษณะของคอนเท็กซ์ที่ใช้ในการแจกแจงดังกล่าว ดังนี้

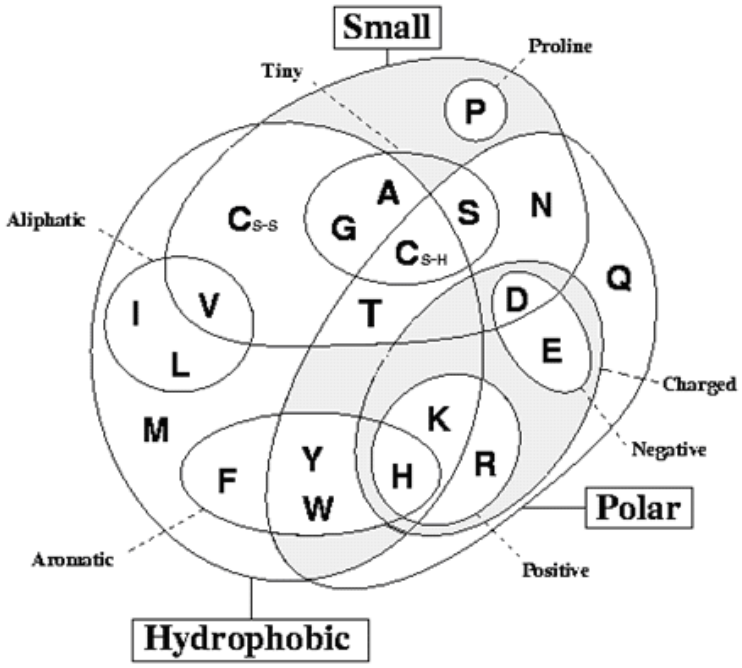
ข. การกำหนดลักษณะของคอนเท็กซ์ที่ใช้อธิบายคอนเซ็ปต์

ในกรอบงานเชิงคอนเซ็ปต์จำเป็นต้องมีความรู้พื้นหลังในรูปฟอร์มคอนเท็กซ์สำหรับการแจกแจงรายละเอียดของแต่ละคอนเซ็ปต์ โดยในกรณีของเอนไซม์นั้นความสามารถในการเกิดฟังก์ชั่นเอนไซม์ตามคอนเซ็ปต์สนับสนุนและคอนเซ็ปต์ไม่ยับยั้งมีความเกี่ยวข้องกับแรงในระดับโมเลกุล 3 ประเภท (Patrick, 1995) คือ แรงขั้วไฟฟ้า (ionic force) แรงเสตอริก (steric force) และแรงแวนเดอร์วาลล์ (Van der Waals force) ยกเว้นแรงสุดท้ายที่เป็นแรงอ่อนแล้ว แรงขั้วไฟฟ้าและแรงเสตอริกเป็นแรงที่มีลักษณะเฉพาะตามคุณสมบัติของอะตอมหรือโมเลกุลในตำแหน่งพิเศษพอดีที่ก่อปฏิกิริยาเอนไซม์ เช่น ขนาด, ประจุไฟฟ้า เป็นต้น ดังนั้นจึงสามารถใช้คุณสมบัติของกรดอะมิโนในบริเวณจับและบริเวณเร่งเป็นคอนเท็กซ์ความรู้พื้นหลังในการอธิบายแรงพื้นฐานที่ก่อปฏิกิริยาเอนไซม์เหล่านี้ได้ โดยคอนเท็กซ์จะต้องอยู่ในรูปความสัมพันธ์ระหว่างชนิดของกรดอะมิโนและคุณสมบัติที่กรดอะมิโนชนิดนั้นมีอยู่ตามนิยามที่ 11

นิยามที่ 11 คอนเท็กซ์กรดอะมิโนและคุณสมบัติ (amino acid – properties context): เป็นคอนเท็กซ์ความสัมพันธ์ระหว่างชนิดของกรดอะมิโนและคุณสมบัติของกรดอะมิโนที่สามารถเขียนอยู่ในรูปของไตรภาคี (Σ, P, R) เมื่อ Σ และ P เป็นเซ็ทของชนิดและคุณสมบัติของกรดอะมิโน และมีความสัมพันธ์ทวิภาค $R \subseteq \Sigma \times P$ ที่แสดงรายละเอียดแต่ละ eRp ที่หมายถึงความสัมพันธ์ระหว่างกรดอะมิโนแต่ละชนิด $e \in \Sigma$ ที่ถูกกำหนดให้มีความสัมพันธ์ทวิภาค R กับคุณสมบัติ $p \in P$ เมื่อ e ถูกตรวจสอบแล้วว่ามีคุณสมบัติ p ดังกล่าว

ในภาพที่ 18 แสดงความรู้พื้นหลังคุณสมบัติเชิงเคมีฟิสิกส์ของกรดอะมิโนในแผนภาพของ Taylor ที่ถูกจัดรูปให้อยู่ในลักษณะของคอนเท็กซ์กรดอะมิโนและคุณสมบัติตาม

นิยามที่ 11 โดยในคอนเท็กซ์แสดงความสัมพันธ์แต่ละ eRp เมื่อกรดอะมิโนชนิดใดมีคุณสมบัติกรดอะมิโนใดให้มาร์คเซลจุดตัดความสัมพันธ์ด้วยสัญลักษณ์ “X”



	Small	Tiny	Proline	Polar	Charge	Positive	Negative	Hydrophobic	Aromatic	Aliphatic
A	X	X						X		
C	X	X		X				X		
D	X			X	X		X			
E				X	X		X			
F								X	X	
G	X	X						X		
H				X	X	X		X	X	
I								X		X
K				X	X	X		X		
L								X		X
M								X		
N	X			X						
P	X		X							
Q				X						
R				X	X	X				
S	X	X		X						
T	X			X				X		
V	X							X		X
W				X				X	X	
Y				X				X	X	

ภาพที่ 18 ความรู้พื้นหลังคุณสมบัติเชิงเคมีฟิสิกส์ในแผนภาพของ Taylor ที่ถูกจัดรูปให้อยู่ในลักษณะของคอนเท็กซ์กรดอะมิโนและคุณสมบัติ

หมายเหตุ ภาพที่แสดงนี้ปรับปรุงมาจากภาพในเวบไซต์ wikipedia

ในหลายงานวิจัยเช่น EMOTIF, 3MOTIF, และ Structure Pattern (Eidhammer *et al.*, 2000) ล้วนใช้ความรู้พื้นหลังแบบคอนเท็กซ์จากแผนภาพของ Taylor (Taylor-Venn's diagram) ในการค้นพบกลุ่มแทนที่ในองค์ประกอบย่อยของโมทีฟ แต่การใช้คอนเท็กซ์ในงานเหล่านี้ใช้เพื่อสนับสนุนการพัฒนาตัวแทนสายโปรตีนโดยตรงเท่านั้น ไม่ได้นำมาแจกแจงความสัมพันธ์ระหว่างคุณลักษณะเด่นของกลุ่มข้อมูลกับประเภทข้อมูลเพื่อพัฒนาเป็นตัวแทนสายโปรตีน ดังเช่นที่นำเสนอในงานวิทยานิพนธ์นี้

จากลักษณะคอนเท็กซ์ที่ได้นี้ ทำให้สามารถนิยามลักษณะของคอนเซ็ปต์ที่ต้องการทุกคอนเซ็ปต์คือ คอนเซ็ปต์สนับสนุน คอนเซ็ปต์ไม่ยับยั้ง และคอนเซ็ปต์การควบคุมการกลายพันธุ์ให้อยู่ในรูปฟอร์มคอนเซ็ปต์ของกลุ่มกรดอะมิโนได้ตามนิยามที่ 12

นิยามที่ 12 คอนเซ็ปต์ (concept) ของกลุ่มกรดอะมิโนและคุณสมบัติ: อยู่ในรูปฟอร์มของคู่ลำดับ (Extent, Intent) ที่ได้จากคอนเท็กซ์ (Σ, P, R) ที่มี Σ เป็นเซ็ทของกรดอะมิโน 20 ชนิด และ P เป็นเซ็ทคุณสมบัติของกรดอะมิโน โดยกำหนดให้ Extent $\subseteq \Sigma$ และ Intent $\subseteq$ P และกำหนดให้คอนเซ็ปต์เป็นคู่ลำดับที่มีฟังก์ชั่นการเชื่อมต่อ Galois (Galois Connection) ซึ่งกันและกันคือ f(Extent) = Intent และ g(Intent) = Extent หรือเขียนได้ดังนี้

$$\text{Intent} = f(\text{Extent}) \text{ เมื่อ } f(\text{Extent}) = \{p \in \text{Intent} \mid \forall a \in \Sigma, aRp\}$$
$$\text{Extent} = g(\text{Intent}) \text{ เมื่อ } g(\text{Intent}) = \{a \in \text{Extent} \mid \forall p \in P, aRp\}$$

ยกตัวอย่างคอนเซ็ปต์ ({A, C, G, S}, {ขนาดเล็ก, ขนาดเล็กพิเศษ}) ที่ได้จากคอนเท็กซ์ที่แปลงมาจากแผนภาพของ Taylor เป็นคอนเซ็ปต์เนื่องจากมี f({A, C, G, S}) = {ขนาดเล็ก, ขนาดเล็กพิเศษ} หรือเซ็ทคุณสมบัติที่เป็น Intent ร่วมที่ใหญ่ที่สุดของกรดอะมิโนกลุ่มนี้ก็คือ {ขนาดเล็ก, ขนาดเล็กพิเศษ} ในขณะที่มี g({ขนาดเล็ก, ขนาดเล็กพิเศษ}) = {A, C, G, S} หรือเซ็ทกลุ่มกรดอะมิโนที่เป็น Extent ร่วมที่ใหญ่ที่สุดของเซ็ทคุณสมบัตินี้ก็คือ {A, C, G, S} เป็นต้น

สำหรับการประมวลผลคอนเซ็ปต์สนับสนุน คอนเซ็ปต์ไม่ยับยั้ง และคอนเซ็ปต์การควบคุมการกลายพันธุ์ที่เดิมอยู่ในรูปฟอร์มคอนเซ็ปต์เบื้องต้นในนิยามที่ 8 9 และ 10 ให้อยู่ในรูปฟอร์มคอนเซ็ปต์ของกลุ่มกรดอะมิโนตามนิยามที่ 12 นั้น ใช้ขั้นตอนวิธีที่เรียกว่า *การประมวลผลการควบคุม*

การกลายพันธุ์บนฐานคอนเซ็ปต์แลททิส (Concept Lattice – Based Mutation Control) มีรายละเอียดดังนี้

5.2 การประมวลผลการควบคุมการกลายพันธุ์บนฐานคอนเซ็ปต์แลททิส (Concept Lattice – Based Mutation Control)

การประมวลผลด้วยคอนเซ็ปต์แลททิสเพื่อแจกแจงคอนเซ็ปต์สนับสนุน คอนเซ็ปต์ไม่ยับยั้ง และคอนเซ็ปต์การควบคุมการกลายพันธุ์นั้น มีขั้นตอนวิธีที่เหมือนกันจำนวน 4 ขั้นตอนคือ 1) การกำหนดคอนเซ็ปต์ที่เป็น input และ output 2) เลือกคอนเท็กซ์เพื่อก่อรูปคอนเซ็ปต์แลททิส 3) การใส่คอนเซ็ปต์ที่เป็น input ลงไปในคอนเซ็ปต์แลททิส และ 4) ดำเนินการบนคอนเซ็ปต์แลททิสเพื่อค้นพบคอนเซ็ปต์ที่เป็น output

การค้นพบคอนเซ็ปต์สนับสนุน คอนเซ็ปต์ไม่ยับยั้ง และคอนเซ็ปต์การควบคุมการกลายพันธุ์ สามารถแจกแจงตามขั้นตอนวิธีดังกล่าวได้ตามลำดับดังนี้

5.2.1 การค้นพบคอนเซ็ปต์สนับสนุน

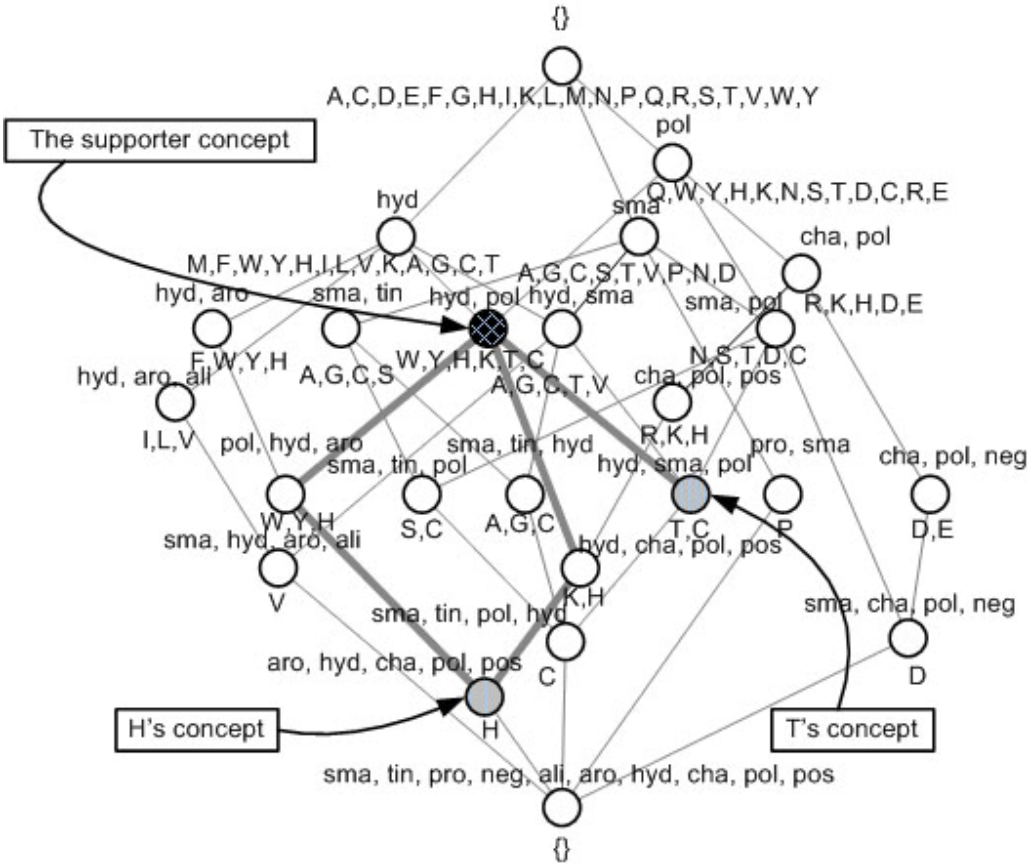
ขั้นตอนที่ 1 การกำหนดคอนเซ็ปต์ที่เป็น input และ output

จากนิยามที่ 8 คอนเซ็ปต์สนับสนุนหมายถึง คอนเซ็ปต์ (เซ็ทกรดอะมิโน, การแจกแจงปัจจัยกลไกเอนไซม์ที่มีร่วมกันของเซ็ทกรดอะมิโน) ซึ่งมีลักษณะเฉพาะในแต่ละตำแหน่ง j ของ บลอก B ที่ได้จากบริเวณจับและบริเวณเร่ง

ดังนั้นจึงสามารถระบุคอนเซ็ปต์ที่เป็น input คือ กลุ่มคอนเซ็ปต์ของกรดอะมิโนแต่ละชนิดที่ปรากฏในตำแหน่ง j ของ บลอก B หรือกลุ่มคอนเซ็ปต์ของกลุ่มแทนที่ $\hat{A}_j$ ที่ได้จากบลอกตามนิยามที่ 6 โดยคอนเซ็ปต์ที่เป็น output ก็คือ คอนเซ็ปต์สนับสนุนที่เป็น super-concept ของกลุ่มคอนเซ็ปต์เหล่านั้น โดยพิจารณาจากลักษณะ intent ที่ระบุการมี “คุณสมบัติร่วม”

ขั้นตอนที่ 2 เลือกคอนเท็กซ์เพื่อก่อรูปคอนเซ็ปต์แลททิส

การเลือกคอนเท็กซ์เพื่อประมวลผลส่วน intent ของคอนเซ็ปต์สนับสนุนคือ “ปัจจัยกลไกเอนไซม์ที่มีร่วมกัน” สามารถใช้คอนเท็กซ์ของกรดอะมิโนและคุณสมบัติได้ตามนิยามที่ 11 ดังนั้นคอนเท็กซ์จากนิยามที่ 11 นี้จึงสามารถนำมาก่อรูปแลททิสกรดอะมิโนกับคุณสมบัติได้ ดังแสดงในภาพที่ 19 ซึ่งก่อรูปแลททิสตามกำหนดไว้ในนิยามที่ 13



ภาพที่ 19 โครงสร้างคอนเซ็ปต์แลททิสที่ได้จากคอนเท็กซ์กรดอะมิโนและคุณสมบัติ รวมทั้งการค้นพบคอนเซ็ปต์สนับสนุนจากคอนเซ็ปต์ของกรดอะมิโน H และ T ที่ได้จากบลอกตำแหน่งที่ 1 ในภาพที่ 9

หมายเหตุ แต่ละโหนดวงกลมหมายถึงแต่ละคอนเซ็ปต์ โดยด้านบนและด้านล่างของแต่ละโหนดคอนเซ็ปต์แสดงข้อมูล Intent และ Extent แทนด้วยตัวย่อ 3 ตัวอักษร

นิยามที่ 13 คอนเซ็ปต์แลททิสกรดอะมิโนกับคุณสมบัติ (amino acid – properties concept lattice): กำหนดสัญญลักษณ์ย่อเป็น $L = (L,\leq)$ เป็นโครงสร้างข้อมูลเชิงคอนเซ็ปต์ของคอนเท็กซ์กรดอะมิโนและคุณสมบัติ (Σ, P, R) ที่เรียกว่าโครงสร้างข้อมูลแลททิส ประกอบด้วยโหนดของแต่ละคอนเซ็ปต์และการเชื่อมต่อคอนเซ็ปต์แบบเรียงตามลำดับ (partial ordering) โดยเริ่มต้นจากโหนดคอนเซ็ปต์ที่มีขนาดของ Extent ใหญ่ที่สุด ตามด้วยคอนเซ็ปต์ที่มีขนาดของ Extent เป็นสับเซ็ทที่มีขนาดรองลงมาตามลำดับในทุกเส้นทางเชื่อมต่อจนสิ้นสุดที่โหนดคอนเซ็ปต์ที่มีขนาด Extent เป็นสับเซ็ทที่เล็กที่สุดของ Extent ของโหนดเริ่มต้น โดยเมื่อกำหนดให้ c1 และ c2 เป็นคอนเซ็ปต์ใน L และ c1 เป็นคอนเซ็ปต์ที่มี Extent เป็นสับเซ็ทของ Extent ของ c2 ($Extent(c1) \subseteq Extent(c2)$ หรือ $Intent(c2) \subseteq Intent(c1)$) สามารถแทนความสัมพันธ์ดังกล่าวด้วยสัญลักษณ์ $c1 \leq c2$ หรือ c1 เป็น sub-concept ของ c2 และ c2 เป็น super-concept ของ c1 ทั้งนี้ ตัวอย่างคอนเซ็ปต์แลททิสที่ได้จากคอนเท็กซ์กรดอะมิโนและคุณสมบัตินำเสนอดังในภาพที่ 19

จากแลททิสที่มีอยู่ สามารถใส่กลุ่มคอนเซ็ปต์ของกรดอะมิโนในกลุ่มแทนที่ $\hat{A}_j$ ที่ได้จากบลอก B ตำแหน่งที่ j ลงไปบนคอนเซ็ปต์แลททิสได้ตามขั้นตอนต่อไป

ขั้นตอนที่ 3 การใส่คอนเซ็ปต์ที่เป็น input ลงไปในคอนเซ็ปต์แลททิส

กรดอะมิโนแต่ละชนิดที่พบในตำแหน่ง j เดียวกันของบลอก B สามารถค้นพบคอนเซ็ปต์บนคอนเซ็ปต์แลททิสได้จากวิธีการในนิยามที่ 14

นิยามที่ 14 คอนเซ็ปต์กรดอะมิโน (amino acid concept): เป็นคอนเซ็ปต์ที่อิงนิยามคอนเซ็ปต์ออบเจคต์ (object concept) ของ (Gabriela, 2007) โดยกรดอะมิโนทุกชนิดจะมีคอนเซ็ปต์ของตนบนแลททิสเรียกว่าคอนเซ็ปต์กรดอะมิโน เป็นคอนเซ็ปต์ที่ประกอบด้วยกลุ่มกรดอะมิโนที่เล็กที่สุดที่มีคุณสมบัติของกรดอะมิโนชนิดนั้นครบถ้วน ยกตัวอย่างเช่น กรดอะมิโน T จะมีคอนเซ็ปต์กรดอะมิโนของตนบนแลททิสกรดอะมิโนและคุณสมบัติเป็นคอนเซ็ปต์ ({T, C}, {ขนาดเล็ก, มีขั้ว, ไม่ชอบน้ำ}) เป็นคอนเซ็ปต์ที่เล็กที่สุดที่มีกรดอะมิโน T และคุณสมบัติของ T อยู่ครบ

เมื่อพิจารณาจากบลอกในภาพที่ 9 ตำแหน่งที่ 1 จะได้ $\hat{A}_j = \{H,T\}$ ซึ่งสามารถค้นพบคอนเซ็ปต์ของกรดอะมิโน H และ T ได้ตามนิยามที่ 14 นี้ โดยแสดงดังในภาพที่ 19 โดยกลุ่มคอนเซ็ปต์ที่ได้กรดอะมิโนแต่ละชนิดใน $\hat{A}_j$ สามารถนิยามเป็น $C(\hat{A}_j)$ ได้ตามนิยามที่ 15

นิยามที่ 15 กลุ่มคอนเซ็ปต์ของ $\hat{A}_j$ หรือ $C(\hat{A}_j)$ บนคอนเซ็ปต์แลททิสกรดอะมิโนและคุณสมบัติ: หมายถึงกลุ่มคอนเซ็ปต์บนคอนเซ็ปต์แลททิสกรดอะมิโนและคุณสมบัติตามนิยามที่ 6 โดยแต่ละคอนเซ็ปต์ได้จากกรดอะมิโนแต่ละชนิดใน $\hat{A}_j$ จากบลอก B ที่ตำแหน่ง j ด้วยวิธีการดังที่นิยามไว้ในนิยามที่ 14

ขั้นตอนที่ 4 ดำเนินการบนคอนเซ็ปต์แลททิสเพื่อค้นพบคอนเซ็ปต์ที่เป็น output

จากการวิเคราะห์ในขั้นตอนที่ 1 ได้ความสัมพันธ์คอนเซ็ปต์สนับสนุนเป็น super-concept ของกลุ่มคอนเซ็ปต์กรดอะมิโน $\hat{A}_j$ หรือ $C(\hat{A}_j)$ ที่ได้จากบลอกตำแหน่ง j ซึ่งสามารถใช้การดำเนินการ Join เป็นวิธีการหา super-concept ของกลุ่มคอนเซ็ปต์ที่มีขนาด extent เล็กที่สุดบนคอนเซ็ปต์แลททิสได้ตามนิยามที่ 16

นิยามที่ 16 การดำเนินการ Join(C): เป็นความสัมพันธ์ระหว่างกลุ่มคอนเซ็ปต์ใน C ที่ให้ผลเป็นคอนเซ็ปต์พิเศษที่มี Intent เป็นเซ็ทคุณสมบัติที่ใหญ่ที่สุดที่กลุ่มคอนเซ็ปต์นั้นมีร่วมกัน และ Extent ของคอนเซ็ปต์นั้นเป็น Extent ที่เล็กที่สุดที่ประกอบด้วยทุกกรดอะมิโนในกลุ่มคอนเซ็ปต์ C โดยการค้นพบคอนเซ็ปต์ Join ใช้การเชื่อมต่อ Galois คือ f และ g ในทฤษฎีคอนเซ็ปต์แลททิส ดังแสดงในสมการที่ 1

$$\text{Join (C)} = (g(f(\bigcup_{c\in C} Extent(c))), \bigcap_{c\in C} Intent(c)) \quad \text{...(สมการที่ 1)}$$

ในการค้นพบคอนเซ็ปต์สนับสนุน ได้จากการ Join ของกลุ่มคอนเซ็ปต์ที่ได้จากกลุ่มแทนที่ $\hat{A}_j$ ที่ได้จากบลอก B ตำแหน่งที่ j โดยทำการ Join บนคอนเซ็ปต์แลททิสกรดอะมิโนและคุณสมบัติ โดยคอนเซ็ปต์สนับสนุนเป็นองค์ประกอบสำคัญของการค้นพบกลุ่มแทนที่ที่สมบูรณ์ จึงนิยามคอนเซ็ปต์สนับสนุนในเลมมา (lemma) ที่ 1 ดังนี้

เลมมาที่ 1 คอนเซ็ปต์สนับสนุน: เป็นคอนเซ็ปต์ที่ได้จากการดำเนินการ Join (นิยามที่ 16) ของกลุ่มคอนเซ็ปต์ $C(\hat{A}_j)$ (นิยามที่ 15) ที่ได้จากบลอก B ตำแหน่ง j โดยดำเนินการบนแลททิสกรดอะมิโนและคุณสมบัติ (นิยามที่ 13) โดยกลุ่มคอนเซ็ปต์กรดอะมิโน $C(\hat{A}_j)$ ประกอบด้วย

คอนเซ็ปต์ $c \in C(\hat{A}_j)$ และ $C(\hat{A}_j) \subseteq L$ บนแลททิสกรดอะมิโนและคุณสมบัติ ดังนั้นคอนเซ็ปต์สนับสนุนสามารถค้นพบได้ด้วย Join($C(\hat{A}_j)$) สรุปเป็นสมการเลมมาที่ 1 ดังนี้

คอนเซ็ปต์สนับสนุน = Join($C(\hat{A}_j)$)

$$= (g(f(\bigcup_{c \in C(\hat{A}j)} Extent(c))), \bigcap_{c \in C(\hat{A}j)} Intent(c)) \text{ สมการเลมมาที่ 1}$$

จากตัวอย่างกลุ่มแทนที่ {H,T} ที่ได้จากตำแหน่งคอลัมน์ที่ 1 จากบลอกในภาพที่ 9 สามารถนำมาค้นพบคอนเซ็ปต์ของ H และ T บนแลททิสได้ดังแสดงในภาพที่ 19 โดยในขั้นตอนนี้เป็นการค้นพบคอนเซ็ปต์สนับสนุนด้วยการดำเนินการ Join ของกลุ่มคอนเซ็ปต์ C({H,T}) เป็นการดำเนินการ Join(C({H,T}))บนแลททิสกรดอะมิโนและคุณสมบัติได้ ได้เป็นคอนเซ็ปต์สนับสนุน ({W,H,Y,K,T,C}, {มีขั้ว, ไม่ชอบน้ำ}) แสดงดังในภาพที่ 19 เช่นเดียวกัน

5.2.2 การค้นพบคอนเซ็ปต์ไม่ยับยั้ง

ขั้นตอนที่ 1 การกำหนดคอนเซ็ปต์ที่เป็น input และ output

จากนิยามที่ 9 คอนเซ็ปต์ไม่ยับยั้งหมายถึง คอนเซ็ปต์ (เซ็ทกรดอะมิโน, การแจกแจงปัจจัยการไม่ยับยั้งกลไกเอนไซม์ที่มีร่วมกันของเซ็ทกรดอะมิโน) ซึ่งมีลักษณะเฉพาะในแต่ละตำแหน่ง j ของ บลอก B ที่ได้จากบริเวณจับและบริเวณเร่ง

ดังนั้นจึงสามารถระบุคอนเซ็ปต์ที่เป็น input คือ กลุ่มคอนเซ็ปต์ของกรดอะมิโน $C(\hat{A}_j)$ ที่ได้จากตำแหน่ง j ของ บลอก B ตามนิยามที่ 15 โดยคอนเซ็ปต์ที่เป็น output ก็คือ คอนเซ็ปต์ไม่ยับยั้งที่เป็น super-concept ของกลุ่มคอนเซ็ปต์เหล่านั้น โดยพิจารณาจากลักษณะ intent ที่ระบุการมี “คุณสมบัติร่วม”

ขั้นตอนที่ 2 เลือกคอนเท็กซ์เพื่อก่อรูปคอนเซ็ปต์แลททิส

การเลือกคอนเท็กซ์เพื่อประมวลผลส่วน intent ของคอนเซ็ปต์ไม่ยับยั้งคือ คอนเท็กซ์ที่สามารถอธิบายการ “ไม่มี” คุณสมบัติที่ขัดขวางการเกิดฟังก์ชั่นเอนไซม์ หรือ คุณสมบัติ

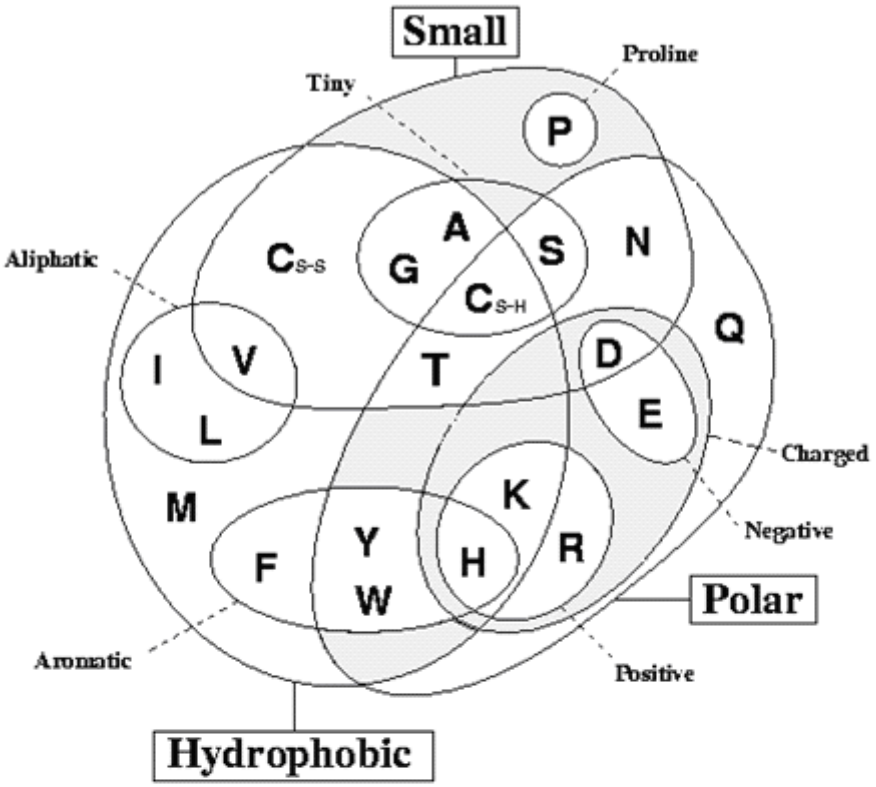
ต้องห้าม (prohibited properties) ที่กรดอะมิโนทุกชนิดในกลุ่มแทนที่นั้นจะต้อง “ไม่มี” คุณสมบัตินั้นเหมือนกัน ดังนั้นจึงต้องการคอนเท็กซ์พิเศษที่รองรับการแจกแจงคุณสมบัติ “ไม่ยับยั้ง” ดังกล่าว

เมื่อพิจารณาจากความหมายของ “คุณสมบัติไม่ยับยั้ง” สามารถเขียนให้อยู่ในรูปของ “คุณสมบัติขอบเขต” ที่เป็นรูปฟอร์มที่สามารถนำไปใช้ในทฤษฎีคอนเซ็ปต์แลททิสได้ดังแสดงในนิยามที่ 17

นิยามที่ 17 คุณสมบัติขอบเขต: เป็นการนำเสนอความหมายในอีกรูปหนึ่งของ “คุณสมบัติไม่ยับยั้ง” โดยพิจารณาจากบทบาทของมันที่สามารถรวมกรดอะมิโนที่แตกต่างกันให้สามารถอยู่ในขอบเขตของเซ็ทเดียวกันได้ ยกตัวอย่างเช่น กำหนดให้คุณสมบัติไฮโดรโฟบิค (ไม่ชอบน้ำ) เป็นคุณสมบัติต้องห้าม จะได้กลุ่มแทนที่กรดอะมิโนที่ใหญ่ที่สุดที่มี “คุณสมบัติไม่ไฮโดรโฟบิค” คือ {D, E, N, P, Q, R, S} โดยเมื่อพิจารณากรดอะมิโน E กับ P จะพบว่า E มีคุณสมบัติ {มีขั้ว, มีประจุ, ประจุลบ} ซึ่งไม่มีคุณสมบัติใดที่เหมือนกับกรดอะมิโน P ที่มีคุณสมบัติ {ขนาดเล็ก, วงแหวนโปรไลน์} เลย แต่ E กับ P ถูกจัดให้อยู่ในขอบเขตเซ็ทเดียวกันได้ด้วยคุณสมบัติไม่ไฮโดรโฟบิค ดังนั้นงานวิจัยนี้จึงนำเสนอคุณสมบัติไม่ต้องห้ามในรูปของ “คุณสมบัติขอบเขต” ตามบทบาทดังกล่าว ซึ่งสามารถใช้งานในทฤษฎีคอนเซ็ปต์แลททิสได้สะดวกขึ้น

ดังนั้น งานวิจัยนี้จึงใช้คอนเท็กซ์พิเศษที่แสดงความสัมพันธ์ระหว่างชนิดและคุณสมบัติขอบเขตของกรดอะมิโนในการสนับสนุนคอนเซ็ปต์ไม่ยับยั้งที่เรียกว่า คอนเท็กซ์กรดอะมิโนและคุณสมบัติขอบเขต โดยคอนเท็กซ์ดังกล่าวจะต้องแปลงมาจากความรู้พื้นหลังเดียวกับคอนเท็กซ์กรดอะมิโนและคุณสมบัติ รายละเอียดนำเสนอดังในนิยามที่ 18

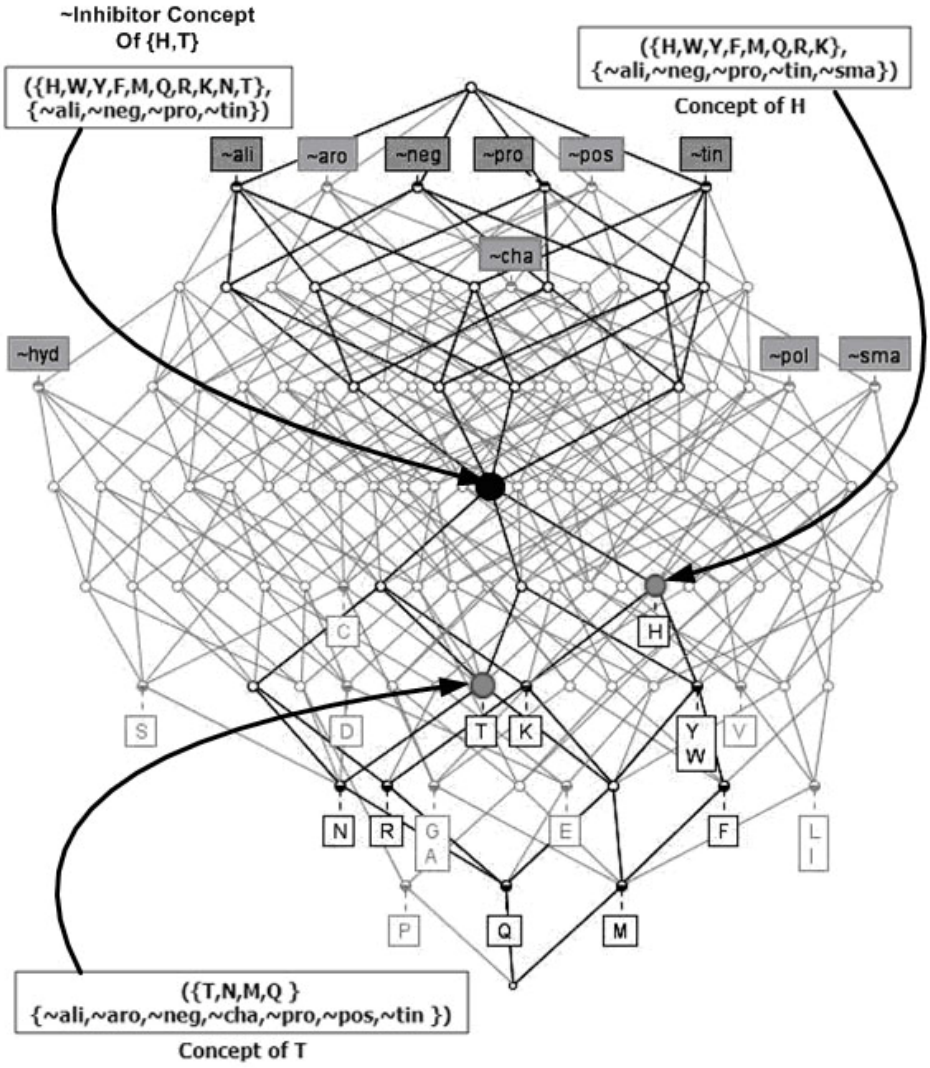
นิยามที่ 18 คอนเท็กซ์กรดอะมิโนและคุณสมบัติขอบเขต: เป็นคอนเท็กซ์ที่เน้นคุณสมบัติคู่ตรงข้าม (inverse properties) หรือการ “ไม่มีคุณสมบัติ” หรือการมี “คุณสมบัติขอบเขต” ดังนั้นจึงสามารถแปลงคอนเท็กซ์กรดอะมิโนและคุณสมบัติในนิยามที่ 11 ให้อยู่ในรูปของไตรภาคี ($\sum$, P', R') โดยกำหนดให้มีความสัมพันธ์ R' $\subseteq \sum$ x P' เมื่อ $\sum$ เป็นเซ็ทของกรดอะมิโน 20 ชนิด และ P' เป็นเซ็ทคุณสมบัติขอบเขต โดยกรดอะมิโน e $\in \sum$ ใดมีคุณสมบัติขอบเขต p' $\in$ P' ใด จะแสดงความสัมพันธ์ eRp' ในคอนเท็กซ์กรดอะมิโนและคุณสมบัติขอบเขตด้วยมาร์ค “X” เมื่อตรวจสอบแล้วว่า e มีคุณสมบัติขอบเขต (p') นั้น



Inverse Context	not small	not tiny	not proline	not polar	not charge	not positive	not negative	not hydrophobic	not aromatic	not aliphatic
A			X	X	X	X	X		X	X
C			X		X	X	X		X	X
D		X	X			X		X	X	X
E	X	X	X			X		X	X	X
F	X	X	X	X	X	X	X			X
G			X	X	X	X	X		X	X
H	X	X	X				X			X
I	X	X	X	X	X	X	X		X	
K	X	X	X				X		X	X
L	X	X	X	X	X	X	X		X	
M	X	X	X	X	X	X	X		X	X
N		X	X		X	X	X	X	X	X
P		X		X	X	X	X	X	X	X
Q	X	X	X		X	X	X	X	X	X
R	X	X	X				X	X	X	X
S			X		X	X	X	X	X	X
T		X	X		X	X	X		X	X
V		X	X	X	X	X	X		X	
W	X	X	X		X	X	X			X
Y	X	X	X		X	X	X			X

ภาพที่ 20 คอนเท็กซ์กรดอะมิโนและคุณสมบัติขอบเขตที่ได้จากความรู้พื้นหลังแผนภาพของ Taylor

ตัวอย่างของคอนเท็กซ์กรดอะมิโนและคุณสมบัติขอบเขตที่แปลงมาจากความรู้พื้นหลังเดียวกับตัวอย่างของคอนเท็กซ์กรดอะมิโนและคุณสมบัติในภาพที่ 18 นำเสนอดังในภาพที่ 20



ภาพที่ 21 แลททิสกรดอะมิโนและคุณสมบัติขอบเขตที่ได้จากคอนเท็กซ์กรดอะมิโนและคุณสมบัติขอบเขต รวมทั้งแสดงคอนเซ็ปต์ไม่ยับยั้งจากคอนเซ็ปต์ของกรดอะมิโน H และ T

หมายเหตุ ภาพนี้สร้างโดยใช้โปรแกรม Concept Explorer 1.3 โดยสัญลักษณ์ “~” แทนความหมายของคำว่า “ไม่มีคุณสมบัติ”

จากคอนเท็กซ์กรดอะมิโนและคุณสมบัติขอบเขตที่ได้ในภาพที่ 20 สามารถก่อรูปเป็นคอนเซ็ปต์แลททิสดังแสดงในภาพที่ 21 ที่ก่อรูปแลททิสตามนิยามที่ 19

นิยามที่ 19 คอนเซ็ปต์แลททิสกรดอะมิโนและคุณสมบัติขอบเขต (amino acid – boundary properties concept lattice): เป็นแลททิสที่สร้างขึ้นมาในลักษณะเดียวกับแลททิสกรดอะมิโนและคุณสมบัติในนิยามที่ 13 แต่เป็นโครงสร้างแลททิสที่สร้างจากคอนเท็กซ์กรดอะมิโนและคุณสมบัติขอบเขต (Σ, P', R') โดยตัวอย่างของแลททิสกรดอะมิโนและคุณสมบัติขอบเขตแสดงดังในภาพที่ 21

ขั้นตอนที่ 3 การใส่คอนเซ็ปต์ที่เป็น input ลงไปในคอนเซ็ปต์แลททิส

กรดอะมิโนแต่ละชนิดที่พบในตำแหน่ง j เดียวกันหรือ $\hat{A}_j$ ของบลอก B สามารถค้นพบคอนเซ็ปต์บนคอนเซ็ปต์แลททิสกรดอะมิโนและคุณสมบัติขอบเขตได้ด้วยวิธีการในลักษณะเดียวกับการค้นพบคอนเซ็ปต์ของกรดอะมิโนในนิยามที่ 14 โดยกลุ่มคอนเซ็ปต์ของ $\hat{A}_j$ ที่ค้นพบบนคอนเซ็ปต์แลททิสกรดอะมิโนกับคุณสมบัติขอบเขต นิยามได้ดังนี้

นิยามที่ 20 กลุ่มคอนเซ็ปต์ $C'(\hat{A}_j)$ บนคอนเซ็ปต์แลททิสกรดอะมิโนและคุณสมบัติขอบเขต: หมายถึงกลุ่มคอนเซ็ปต์ที่ได้จากกรดอะมิโนแต่ละชนิดใน $\hat{A}_j$ ที่ได้จาก บลอกที่ตำแหน่ง j บนคอนเซ็ปต์แลททิสกรดอะมิโนและคุณสมบัติขอบเขตตามนิยามที่ 19 ด้วยวิธีการดังที่นิยามไว้ในนิยามที่ 14

ทั้งนี้ ตัวอย่างการใส่คอนเซ็ปต์ของ H และ T หรือ C'({H,T}) ที่ได้จากบลอกในภาพที่ 9 ตำแหน่งที่ 1 ลงไปบนคอนเซ็ปต์แลททิสกรดอะมิโนและคุณสมบัติขอบเขต แสดงตัวอย่างให้เห็นดังในภาพที่ 21

ขั้นตอนที่ 4 ดำเนินการบนคอนเซ็ปต์แลททิสเพื่อค้นพบคอนเซ็ปต์ที่เป็น output

จากการวิเคราะห์ในขั้นตอนที่ 1 ได้ความสัมพันธ์คอนเซ็ปต์ไม่ยับยั้งเป็น super-concept ของกลุ่มคอนเซ็ปต์กรดอะมิโน $\hat{A}_j$ ในลักษณะเดียวกับคอนเซ็ปต์สนับสนุน โดยเมื่อมีกลุ่มคอนเซ็ปต์ $C'(\hat{A}_j)$ ย่อมสามารถนำมาใช้ค้นพบคอนเซ็ปต์ไม่ยับยั้งบนคอนเซ็ปต์แลททิสกรดอะมิโน

และคุณสมบัติขอบเขตด้วยการดำเนินการ Join ในลักษณะเดียวกับคอนเซ็ปต์สนับสนุนได้ ดังแสดงในเลมมา 2

<u>เลมมาที่ 2 คอนเซ็ปต์ไม่ยับยั้ง</u>: เป็นคอนเซ็ปต์ที่ได้จากการดำเนินการแลททิส Join (นิยามที่ 16) ของกลุ่มคอนเซ็ปต์ $C'(\hat{A}_j)$ (นิยามที่ 20) ที่ได้จากแต่ละตำแหน่งในบลอกจากบริเวณจับหรือบริเวณเร่ง โดยดำเนินการบนแลททิสกรดอะมิโนและคุณสมบัติขอบเขต (นิยามที่ 19) โดยกำหนดให้ $A_j \subseteq \Sigma$ เป็นกลุ่มกรดอะมิโนที่ได้จากแต่ละตำแหน่งคอลัมน์ในบลอกจากบริเวณจับหรือบริเวณเร่ง โดยกำหนดให้คอนเซ็ปต์ $c' \in C'(\hat{A}_j)$ บนแลททิสกรดอะมิโนและคุณสมบัติขอบเขต จากนั้นค้นพบคอนเซ็ปต์ไม่ยับยั้งจากการดำเนินการกลุ่มคอนเซ็ปต์กรดอะมิโนนั้นด้วย Join($C'(\hat{A}_j)$) แสดงดังในสมการเลมมาที่ 2 ดังนี้

คอนเซ็ปต์ไม่ยับยั้ง $= \text{Join}(C'(\hat{A}_j))$

$$= \left(g\left(f\left(\bigcup_{c' \in C''(\hat{A}j)} Extent(c')\right)\right), \bigcap_{c' \in C''(\hat{A}j)} Intent(c')\right) \qquad \text{สมการเลมมาที่ 2}$$

ตัวอย่างของการ Join($C'(\{H,T\})$) เพื่อค้นพบคอนเซ็ปต์ไม่ยับยั้งบนคอนเซ็ปต์แลททิสกรดอะมิโนและคุณสมบัติขอบเขต นำเสนอดังในภาพที่ 21 ได้ผลเป็นคอนเซ็ปต์ไม่ยับยั้ง ({F,H,K,M,N,Q,R,T,W,Y},{ไม่เล็กพิเศษ, ไม่เป็นวงแหวนโปรไลน์, ไม่ใช่กิ่งอะลิฟาติค, ไม่ใช่ประจุลบ})

5.2.3 การค้นพบคอนเซ็ปต์การควบคุมการกลายพันธุ์

ขั้นตอนที่ 1 การกำหนดคอนเซ็ปต์ที่เป็น input และ output

จากนิยามที่ 10 คอนเซ็ปต์การควบคุมการกลายพันธุ์ คือ คอนเซ็ปต์ (กลุ่มแทนที่ที่สมบูรณ์, การแจกแจงปัจจัยกลไกการเกิดฟังก์ชั่นเอนไซม์และไม่ยับยั้งฟังก์ชั่นเอนไซม์ที่มีร่วมกันของกลุ่มแทนที่ที่สมบูรณ์) ซึ่งมีลักษณะเฉพาะในแต่ละตำแหน่ง j ของ บลอก B ที่ได้จากบริเวณจับและบริเวณเร่ง โดยมีลักษณะเฉพาะของทั้งคอนเซ็ปต์สนับสนุนและคอนเซ็ปต์ไม่ยับยั้งปรากฏอยู่

จากนิยามดังกล่าวนี้สามารถระบุคอนเซ็ปต์ที่เป็น output คือ คอนเซ็ปต์ควบคุมการกลายพันธุ์ และคอนเซ็ปต์ที่เป็น input คือ คอนเซ็ปต์สนับสนุน และคอนเซ็ปต์ไม่ยับยั้ง ซึ่งได้กล่าวถึงวิธีค้นพบคอนเซ็ปต์สนับสนุน และคอนเซ็ปต์ไม่ยับยั้งดังกล่าวในขั้นตอนก่อนหน้านี้ พร้อมที่จะดำเนินการในขั้นต่อไป

ขั้นตอนที่ 2 เลือกคอนเท็กซ์เพื่อก่อรูปคอนเซ็ปต์แลททิส

ในการเลือกคอนเท็กซ์เพื่อก่อรูปคอนเซ็ปต์แลททิสในกรณีของคอนเซ็ปต์การควบคุมการกลายพันธุ์มีความซับซ้อนกว่า คอนเซ็ปต์สนับสนุน และคอนเซ็ปต์ไม่ยับยั้ง เนื่องจากคอนเซ็ปต์ควบคุมการกลายพันธุ์ต้องการคอนเซ็ปต์แลททิสที่สามารถรองรับทั้งคอนเซ็ปต์สนับสนุน และคอนเซ็ปต์ไม่ยับยั้ง ซึ่งไม่มีอยู่เดิม ดังนั้นจึงมีความจำเป็นต้องสร้างแลททิสขึ้นใหม่อีก 1 แลททิสที่สามารถรองรับทั้ง 2 คอนเซ็ปต์ได้ และนำไปสู่การค้นพบคอนเซ็ปต์การควบคุมการกลายพันธุ์ ดังนั้นจึงเรียกแลททิสที่จะสร้างขึ้นใหม่นี้ว่าแลททิสการควบคุมการกลายพันธุ์ ซึ่งก่อรูปมาจากคอนเท็กซ์การควบคุมการกลายพันธุ์

วิธีที่เรียบง่ายในการสร้างคอนเท็กซ์และแลททิสใหม่ที่สามารถรองรับคอนเซ็ปต์ที่อยู่ต่างแลททิสกันก็คือ การยุบรวมทั้ง 2 คอนเซ็ปต์เข้าด้วยกัน ยุบ Extent เข้าด้วยกันเป็นกรดอะมิโนชุดหนึ่ง และยุบ Intent เข้าด้วยกันเป็นคุณสมบัติชุดใหม่อีกชุดหนึ่ง ซึ่งกรดอะมิโนชุดดังกล่าวจะก่อรูปความสัมพันธ์กับชุดคุณสมบัติใหม่ได้เป็นคอนเท็กซ์ใหม่ที่เรียกว่าคอนเท็กซ์การควบคุมการกลายพันธุ์ดังแสดงในนิยามที่ 21

<u>นิยามที่ 21 คอนเท็กซ์การควบคุมการกลายพันธุ์ (mutation control context)</u>: คือความสัมพันธ์ R" $\subseteq$ A" x P" เมื่อ A" $\subseteq \sum$ โดย $\sum$ = {กรดอะมิโน 20 ชนิด} และกลุ่มกรดอะมิโน A" ได้มาจากการยุบรวมกันของ Extent จากคอนเซ็ปต์สนับสนุนในเลมมาที่ 1 และ Extent จากคอนเซ็ปต์ไม่ยับยั้งในเลมมาที่ 2 หรือ A" = g(f($\bigcup_{c\in\hat{A}'} Extent(c)$)) $\cup$ g(f($\bigcup_{c'\in\hat{A}''} Extent(c')$)) และได้เซ็ทคุณสมบัติกรดอะมิโนชุดใหม่จากการยุบรวมกันของ Intent จากคอนเซ็ปต์สนับสนุนในเลมมาที่ 1 และ Intent จากคอนเซ็ปต์ไม่ยับยั้งในเลมมาที่ 2 หรือ P" = $\bigcap_{c\in\hat{A}'} Intent(c) \cup \bigcap_{c'\in\hat{A}''} Intent(c')$ โดย P" $\subseteq$ P $\cup$ P' เมื่อ P = {เซ็ทคุณสมบัติกรดอะมิโน} จากคอนเท็กซ์กรดอะมิโนและคุณสมบัติ และ P' = {เซ็ทคุณสมบัติขอบเขตของกรดอะมิโน} จากคอนเท็กซ์กรดอะมิโนและคุณสมบัติขอบเขต ทั้งนี้

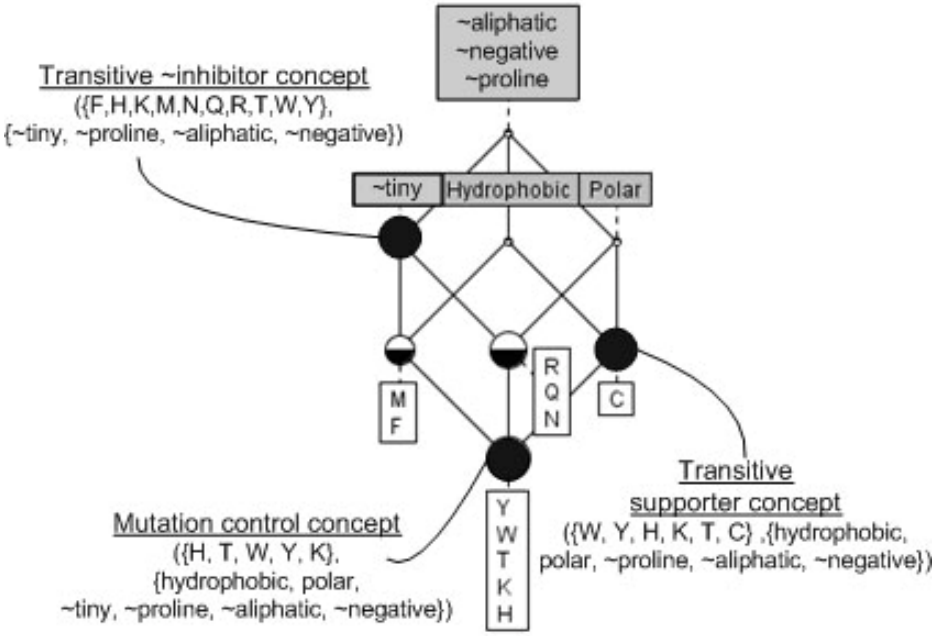
การยุบรวมทั้ง 2 คอนเซ็ปต์เข้าด้วยกันจะต้องเป็นคอนเซ็ปต์ที่ได้มาจากกลุ่มแทนที่ที่ได้จากแต่ละตำแหน่งคอลัมน์ในบลอกชุดเดียวกัน

ตัวอย่างของคอนเท็กซ์การควบคุมการกลายพันธุ์ที่สร้างจากการยุบรวมคอนเซ็ปต์สนับสนุนและคอนเซ็ปต์ไม่ยับยั้งของกลุ่มแทนที่ {H,T} แสดงดังในตารางที่ 5 โดยจากคอนเท็กซ์ดังกล่าวนี้ สามารถใช้หลักทฤษฎีของคอนเซ็ปต์แลททิสในการสร้างคอนเซ็ปต์แลททิสการควบคุมการกลายพันธุ์ดังแสดงในนิยามที่ 22

ตารางที่ 5 คอนเท็กซ์การควบคุมการกลายพันธุ์ของกลุ่มแทนที่ {H,T}

	ไม่มีขั้ว	ไม่ชอบน้ำ	ไม่เล็กพิเศษ	ไม่เป็นวงแหวนโปรไลน์	ไม่เป็นประจุลบ	ไม่เป็นกิ่งอะลิฟาติค
C	X	X		X	X	X
F		X	X	X	X	X
H	X	X	X	X	X	X
K	X	X	X	X	X	X
M		X	X	X	X	X
N	X		X	X	X	X
Q	X		X	X	X	X
R	X		X	X	X	X
T	X	X	X	X	X	X
W	X	X	X	X	X	X
Y	X	X	X	X	X	X

นิยามที่ 22 คอนเซ็ปต์แลททิสการควบคุมการกลายพันธุ์ (mutation control – concept lattice): เป็นแลททิสที่สร้างจากคอนเท็กซ์คล้ายกับแลททิสในนิยามที่ 13 แต่ใช้คอนเท็กซ์การควบคุมการกลายพันธุ์ในการค้นพบทุกคอนเซ็ปต์ เพื่อก่อรูปเป็นแลททิสการควบคุมการกลายพันธุ์ แสดงดังในภาพที่ 22



ภาพที่ 22 แลททิสการควบคุมการกลายพันธุ์และคอนเซ็ปต์ที่เกี่ยวข้อง

ขั้นตอนที่ 3 การใส่คอนเซ็ปต์ที่เป็น input ลงไปในคอนเซ็ปต์แลททิส

ในการใส่คอนเซ็ปต์ที่เป็น input ลงไปในคอนเซ็ปต์แลททิสนั้น เดิมใช้วิธีการหาคอนเซ็ปต์ของกรดอะมิโนดังแสดงในนิยามที่ 15 และ 20 แต่เนื่องจากในกรณีคอนเซ็ปต์การกลายพันธุ์มีคอนเซ็ปต์ที่เป็น input อยู่แล้ว แต่เป็นคอนเซ็ปต์ที่อยู่ต่างคอนเซ็ปต์แลททิสกัน ดังนั้นเราจึงต้องการวิธีในการนำคอนเซ็ปต์เหล่านี้มาอยู่ด้วยกันบนแลททิสเดียวกัน ด้วยการระบุคอนเซ็ปต์ที่มี extent และ intent ที่สอดคล้องกันมากที่สุด กับคอนเซ็ปต์ที่อยู่ต่างแลททิสกัน โดยเราเรียกคอนเซ็ปต์ที่สอดคล้องกันดังกล่าวว่าคอนเซ็ปต์ส่งผ่าน (transitive concept) นิยามดังในเลมมาที่ 3

<u>เลมมาที่ 3 คอนเซ็ปต์ส่งผ่าน (transitive concept)</u>: คอนเซ็ปต์ U เป็นคอนเซ็ปต์ส่งผ่านของคอนเซ็ปต์ V เขียนให้อยู่ในฟอร์ม $U = Tr(V)$ ก็ต่อเมื่อ U และ V เป็นคอนเซ็ปต์ที่อยู่ต่างแลททิสกัน โดย $Extent(V) \subseteq Extent(U)$ และ $Intent(V) \subseteq Intent(U)$

จากเลมมาที่ 3 สามารถเขียนคอนเซ็ปต์ส่งผ่านของคอนเซ็ปต์สนับสนุนให้อยู่ในรูป $Tr(Join(C(\hat{A}_j)))$ และคอนเซ็ปต์ส่งผ่านของคอนเซ็ปต์ไม่ยับยั้งสามารถเขียนให้อยู่ในรูป $Tr(Join(C'(\hat{A}_j)))$ แสดงตัวอย่างดังในภาพที่ 22

ขั้นตอนที่ 4 ดำเนินการบนคอนเซ็ปต์แลททิสเพื่อค้นพบคอนเซ็ปต์ที่เป็น output

จากคอนเซ็ปต์ส่งผ่านของคอนเซ็ปต์สนับสนุนและคอนเซ็ปต์ไม่ยับยั้งที่อยู่บนแลททิสเดียวกัน ดังนั้นจึงสามารถดำเนินการแลททิสแบบพิเศษในการหาคอนเซ็ปต์ภาพรวมของคอนเซ็ปต์แยกย่อยทั้ง 2 คอนเซ็ปต์ได้เป็นคอนเซ็ปต์การควบคุมการกลายพันธุ์

โดยจากนิยามที่ 10 ของคอนเซ็ปต์การกลายพันธุ์ คือ *การมีลักษณะเฉพาะของทั้งคอนเซ็ปต์สนับสนุนและคอนเซ็ปต์ไม่ยับยั้งปรากฏอยู่* ดังนั้นการดำเนินการแลททิสที่ครอบคลุมการอธิบายทุกคุณสมบัติที่เกี่ยวข้องกับกลไกการเกิดเอนไซม์ก็คือ Meet ในทฤษฎีคอนเซ็ปต์แลททิส ดังแสดงในนิยามที่ 23

นิยามที่ 23 การดำเนินการ Meet(C): เป็นความสัมพันธ์ระหว่างกลุ่มคอนเซ็ปต์ใน C ที่ให้ผลเป็นคอนเซ็ปต์พิเศษ โดยคอนเซ็ปต์ดังกล่าวมี Extent เป็นเซ็ทกรดอะมิโนที่ใหญ่ที่สุดที่กลุ่มคอนเซ็ปต์นั้นมีร่วมกัน และ Intent ของคอนเซ็ปต์นั้นประกอบด้วยทุกคุณสมบัติของกรดอะมิโนในกลุ่มคอนเซ็ปต์นั้น โดยการค้นพบคอนเซ็ปต์ Meet ใช้การเชื่อมต่อ Galois คือ f และ g ในทฤษฎีคอนเซ็ปต์แลททิส ดังแสดงในสมการที่ 2

$$\text{Meet (C)} = \left(\bigcap_{c\in \hat{A}} Extent(c)\right), f(g(\bigcup_{c\in \hat{A}} Intent(c))) \quad \ldots(\text{สมการที่ 2})$$

จากคอนเซ็ปต์สนับสนุนในเลมมาที่ 1 และคอนเซ็ปต์ไม่ยับยั้งในเลมมาที่ 2 สามารถใช้งานบนแลททิสการควบคุมการกลายพันธุ์เดียวกันในนิยามที่ 22 ได้ผ่านคอนเซ็ปต์ส่งผ่านในเลมมาที่ 3 ทำให้สามารถใช้การดำเนินการ Meet ในนิยามที่ 23 เพื่อค้นพบคอนเซ็ปต์การควบคุมการกลายพันธุ์ตามแบบแผนของทฤษฎีคอนเซ็ปต์แลททิสได้ในท้ายสุด ดังนำเสนอเป็นทฤษฎีหลักในวิทยานิพนธ์ดังนี้

ทฤษฎีหลัก คอนเซ็ปต์การควบคุมการกลายพันธุ์ (mutation control concept): เป็นคอนเซ็ปต์ที่ได้จากการดำเนินการ Meet (ในนิยามที่ 23) ของคอนเซ็ปต์ส่งผ่าน (เลมมา 3) ของคอนเซ็ปต์สนับสนุน (เลมมา 1) และคอนเซ็ปต์ไม่ยับยั้ง (เลมมา 2) บนแลททิสการควบคุมการกลายพันธุ์ (ในนิยามที่ 22) รายละเอียดดังนี้

กำหนดให้คอนเซ็ปต์ส่งผ่านของคอนเซ็ปต์สนับสนุน Tr(Join(C($\hat{A}_j$))) เป็นรูปฟอร์มคอนเซ็ปต์ที่ประมวลจากเลมมาที่ 1 และ 3 โดย C($\hat{A}_j$) เป็นเซ็ทของคอนเซ็ปต์กรดอะมิโนบนแลททิสกรดอะมิโนและคุณสมบัติ ของกลุ่มแทนที่ $\hat{A}_j$ ที่ได้จากบลอก B ที่ตำแหน่ง j และกำหนดให้มีคอนเซ็ปต์ส่งผ่านของคอนเซ็ปต์ไม่ยับยั้ง Tr(Join(C'($\hat{A}_j$))) เป็นรูปฟอร์มคอนเซ็ปต์ที่ประมวลจากเลมมาที่ 2 และ 3 โดย C'($\hat{A}_j$) เป็นเซ็ทคอนเซ็ปต์กรดอะมิโนบนแลททิสกรดอะมิโนและคุณสมบัติขอบเขตของกลุ่มแทนที่ $\hat{A}_j$ ที่ได้จากบลอก B ที่ตำแหน่ง j เช่นเดียวกัน โดยทั้ง 2 คอนเซ็ปต์ส่งผ่านเป็นคอนเซ็ปต์ c" ในเซ็ทของคอนเซ็ปต์ C"($\hat{A}_j$) หรือ c" $\in$ C"($\hat{A}_j$)

ดังนั้นคอนเซ็ปต์การควบคุมการกลายพันธุ์ จึงสามารถอธิบายได้ด้วยการดำเนินการ Meet บนกลุ่มคอนเซ็ปต์ C"($\hat{A}_j$) บนแลททิสการควบคุมการกลายพันธุ์เดียวกัน ดังนี้

คอนเซ็ปต์การควบคุมการกลายพันธุ์

= Meet (C"($\hat{A}_j$))

= Meet ({Tr(Join (C($\hat{A}_j$))), Tr(Join (C'($\hat{A}_j$)))})

$$= \left(\bigcap_{c''\in\{Tr(\text{Join }(C(\hat{A}j))),Tr(\text{Join }(C'(\hat{A}j)))\}} Extent(c'') \right), f\left(g\left(\bigcup_{c''\in\{Tr(\text{Join }(C(\hat{A}j))),Tr(\text{Join }(C'(\hat{A}j)))\}} Intent(c'') \right)\right) \text{ ...(สมการทฤษฎีหลัก)}$$

จากคอนเซ็ปต์การควบคุมการกลายพันธุ์ในทฤษฎีหลัก สามารถค้นพบกลุ่มแทนที่กรดอะมิโนที่สมบูรณ์ได้จาก Extent คือ $\bigcap_{c''\in\{Tr(\text{Join }(C(\hat{A}j))),Tr(\text{Join }(C'(\hat{A}j)))\}} Extent(c'')$ เนื่องจากกลุ่ม Extent นี้มีคุณสมบัติของ Intent คือ $f(g(\bigcup_{c''\in\{Tr(\text{Join }(C(\hat{A}j))),Tr(\text{Join }(C'(\hat{A}j)))\}} Intent(c'')))$ ที่เป็นกลไกการเกิดฟังก์ชั่นเอนไซม์ทั้งหมด และดังนั้นคอนเซ็ปต์การควบคุมการกลายพันธุ์จึงเป็นการพัฒนาคุณภาพของกลุ่มแทนที่กรดอะมิโน $\hat{A}_j$ ที่ได้จากบลอกในบริเวณจับหรือบริเวณเร่งในตำแหน่ง j โดยเมื่อดำเนินการกับทุกตำแหน่ง j บนบลอก B ก็จะได้กลุ่มแทนที่ที่สมบูรณ์เป็นองค์ประกอบของรีแอคทีฟโมทีฟ M ที่มีคุณภาพ

ตัวอย่างคอนเซ็ปต์การกลายพันธุ์ของกลุ่มแทนที่กรดอะมิโน {H,T} ที่ได้จากการประมวลตามขั้นตอนในทฤษฎีหลักถูกแสดงในภาพที่ 22 ได้ผลเป็นคอนเซ็ปต์ ({H,T,W,Y,K}, {ไม่ชอบน้ำ, มีขั้ว, ไม่ใช่ขนาดเล็กพิเศษ, ไม่เป็นวงแหวนโปรไลน์, ไม่เป็นกิ่งอะลิฟาติก, ไม่มีประจุ

ลบ}) และกลุ่มแทนที่กรดอะมิโนที่สมบูรณ์หรือมีประสิทธิภาพดีที่ค้นพบจากคอนเซ็ปต์การควบคุมการกลายพันธุ์ก็คือ {H,T,W,Y,K}

ในหัวข้อที่ 5.2.4 จะได้นำเสนอขั้นตอนวิธีคำนวณกลุ่มแทนที่ที่สมบูรณ์ โดยอิงทฤษฎีหลักที่นำเสนอในวิทยานิพนธ์นี้ ซึ่งทำให้เชื่อมั่นได้ว่าหลักความสัมพันธ์ระหว่างคุณลักษณะเด่นของกลุ่มสายโปรตีน (reactive motif) กับประเภทข้อมูล (ฟังก์ชั่นเอนไซม์) เหล่านี้มีอยู่ในรีแอคทีฟโมทีฟที่เป็นผลผลิตของงานวิจัยนี้ เป็นขั้นตอนสุดท้ายของการพัฒนาคุณภาพกลุ่มแทนที่กรดอะมิโนด้วยองค์ความรู้

5.2.4 ขั้นตอนวิธีประมวลการค้นพบกลุ่มแทนที่กรดอะมิโนที่สมบูรณ์ที่อิงคอนเซ็ปต์การควบคุมการกลายพันธุ์

จากทฤษฎีหลัก มีการใช้ 3 แลททิสในการควบรวมองค์ความรู้พื้นหลังและความรู้เชิงวิทยาศาสตร์เพื่อค้นพบองค์ประกอบย่อยโมทีฟที่มีคุณภาพและมีความหมาย (meaningful motif) ที่สัมพันธ์กับประเภทข้อมูลที่ใช้ในการทำนายประเภทข้อมูล (ฟังก์ชั่นเอนไซม์) อย่างไรก็ตาม การค้นพบแต่ละกลุ่มแทนที่จาก 3 แลททิส ไม่ใช่ขั้นตอนที่มีประสิทธิภาพนัก

ดังนั้นในการประยุกต์โปรแกรมเพื่อค้นพบกลุ่มแทนที่กรดอะมิโนที่สมบูรณ์จึงใช้เพียงบางส่วนของทฤษฎีหลักในการประมวลผลใช้งานเพียง 2 คอนเท็กซ์เพื่อให้สามารถเขียนโปรแกรมได้ง่ายและทำงานได้เร็ว ทั้งนี้การประยุกต์ใช้จากทฤษฎีหลักยังเป็นการสร้างความเชื่อมั่นให้กับขั้นตอนวิธีดังกล่าวว่าจะยังคงให้ผลเป็นรีแอคทีฟโมทีฟที่มีความสัมพันธ์แนบแน่นกับประเภทฟังก์ชั่นเอนไซม์และใช้เป็นคุณลักษณะเด่นของกลุ่มข้อมูลนำเข้าสำหรับสร้างโมเดลในการทำนายประเภทฟังก์ชั่นเอนไซม์ได้ดี

```
1)  Procedure Aminos-to-Common-Property(A[1..n], C_m)
2)  Start
3)  I ← {all properties P in C_m};
4)  for k ← 1 to n
5)  {check properties p of A_k
6)  I ← I – (P-p);
7)  }
8)  return I;
9)  End;
10)
11) Procedure Property-to-Common-Aminos(I, C_m)
12) Start
13) G ← {};
14) for each amino acid a in C_m
15) {if a has I
16) then G ← a;
17) }
18) return G;
19) End;
20)
21) Discovering Maximal Substitution Group Algorithm
22) Main(C_AP, C_BP, A[1..n])
23) Start
24) I_C ← Aminos-to-Common-Property(A[1..n], C_AP);
25) G_S ← Property-to-Common-Aminos(I_C, C_AP);
26) I_B ← Aminos-to-Common-Property(A[1..n], C_BP);
27) G_I ← Property-to-Common-Aminos(I_B, C_BP);
28) SG ← G_S ∩ G_I;
29) End.
```

ภาพที่ 23 ขั้นตอนวิธีในการค้นพบกลุ่มแทนที่กรดอะมิโนที่สมบูรณ์ที่อิงคอนเซ็ปต์การควบคุมการกลายพันธุ์

จากทฤษฎีหลัก เน้นพิจารณาไปที่ส่วน Extent ของคอนเซ็ปต์การควบคุมการกลายพันธุ์ $\bigcap_{c'' \in \{Tr(\text{Join }(C(\hat{A}j))), Tr(\text{Join }(C'(\hat{A}j)))\}} Extent(c'')$ พบว่ากลุ่มแทนที่กรดอะมิโนนี้เกิดจากการอินเตอร์เส็คกันระหว่าง Extent ของคอนเซ็ปต์สนับสนุน และ Extent ของคอนเซ็ปต์ไม่ยับยั้ง ดังนั้นจึงสามารถจำกัดการคำนวณให้อยู่ในขอบเขตเพียง 2 คอนเซ็ปต์ และเพื่อหลีกเลี่ยงความซับซ้อนในการประมวลแลททิส จึงดำเนินการค้นพบคอนเซ็ปต์สนับสนุนและคอนเซ็ปต์ไม่ยับยั้งจากการประมวลบนคอนเท็กซ์โดยตรง ซึ่งมีเพียงแค่ 2 คอนเท็กซ์ รายละเอียดการคำนวณนำเสนอดังในภาพที่ 23

จากภาพที่ 23 ขั้นตอนวิธีในการค้นพบกลุ่มแทนที่กรดอะมิโนที่สมบูรณ์เริ่มจากบรรทัดที่ 22 เริ่มจากการกำหนดข้อมูลนำเข้า 3 ชนิดด้วยกันคือ คอนเท็กซ์กรดอะมิโนและคุณสมบัติที่ได้จากความรู้พื้นหลัง (แทนด้วยสัญลักษณ์ C_{AP}), คอนเท็กซ์กรดอะมิโนและคุณสมบัติขอบเขตที่ได้จากความรู้พื้นหลังเดียวกัน (แทนด้วยสัญลักษณ์ C_{BP}), และกลุ่มแทนที่กรดอะมิโนที่ได้จากบลอก (แทนด้วยตัวแปรอาเรย์ A[1..n])

จากนั้นสืบค้น I_C ที่เป็น Intent ของคอนเซ็ปต์สนับสนุนได้จากการทำอินเตอร์เส็คคุณสมบัติของกรดอะมิโนทุกตัวที่อยู่ใน A[1..n] ด้วย procedure ที่ชื่อว่า Aminos-to-Common-Property(A[1..n], C_{AP}) ในบรรทัดที่ 24 หลังจากนั้นใช้ I_C ใน procedure ที่ชื่อว่า Property-to-Common-Aminos(I_C, C_{AP}) เพื่อค้นพบ Extent ของคอนเซ็ปต์สนับสนุน G_S ในบรรทัดที่ 25 ซึ่งก็คือกลุ่มของกรดอะมิโนที่ใหญ่ที่สุดที่มีคุณสมบัติ I_C

ในลักษณะการดำเนินการแบบเดียวกัน การค้นพบ Extent ของคอนเซ็ปต์ไม่ยับยั้ง G_I ใช้ทั้ง 2 procedure ในบรรทัดที่ 26 และ 27 โดยเปลี่ยนข้อมูลนำเข้าจาก C_{AP} เป็น C_{BP} สุดท้ายค้นพบกลุ่มแทนที่กรดอะมิโนที่สมบูรณ์ SG ได้จากการอินเตอร์เส็กระหว่าง Extent ของคอนเซ็ปต์สนับสนุน G_S และ Extent ของคอนเซ็ปต์ไม่ยับยั้ง G_I ซึ่งให้ผลที่เหมือนกับ Extent ที่นำเสนอในทฤษฎีหลัก

6. การพัฒนาคุณภาพบลอกข้อมูลจากบริเวณจับและบริเวณเร่ง

จากข้อเท็จจริงที่ข้อมูลบลอกที่ได้จากบริเวณจับและบริเวณเร่งกว่าครึ่งหนึ่งมีข้อมูลลำดับกรดอะมิโนเพียง 1 สายเท่านั้น ดังนั้นการใช้งานรีแอคทีฟโมทีฟในกลุ่มข้อมูลเอนไซม์ขนาดใหญ่จะใช้งานได้ก็ต่อเมื่อแก้ปัญหาบลอกที่มีขนาดเท่ากับ 1 นี้ได้ โดยมีสมมติฐานการแก้ปัญหาดังนี้

<u>สมมติฐานที่ 2:</u> แม้ว่าบลอกจะมีสมาชิกลำดับกรดอะมิโนบริเวณจับหรือบริเวณเร่งอยู่น้อยหรือมีเพียง 1 สาย แต่ถ้ามีเครื่องมือในการค้นพบบริเวณที่ **"คล้ายคลึง"** กับข้อมูลที่มีอยู่ ย่อมสามารถพัฒนาบลอกที่มีคุณภาพในการค้นพบกลุ่มแทนที่กรดอะมิโนได้

องค์ความรู้ที่ระบุความ "คล้ายคลึง" ของกรดอะมิโน ก็คือองค์ความรู้ในลักษณะตารางคะแนนความเหมือนซึ่งสามารถใช้คำนวณค่าคะแนนที่ระบุความคล้ายคลึงกันระหว่างสายโปรตีน 2 สายได้ ดังนั้นจึงใช้ตารางคะแนนความเหมือนในการพัฒนาคุณภาพของบลอกได้

ในงานวิจัยนี้การพัฒนาคุณภาพบลอกประกอบไปด้วยขั้นตอนก่อนและหลังขั้นตอนวิธีการควบคุมการกลายพันธุ์ที่เรียกว่า "การคัดสรรบลอกที่มีคุณภาพ" (block scan filtering) และ "การจัดกลุ่มบริเวณปฏิกิริยาชีวเคมี" (reactive site – group definition) โดยรายละเอียดจะได้นำเสนอตามลำดับดังนี้

6.1 การคัดสรรบลอกที่มีคุณภาพ (block scan filtering)

เป้าหมายของขั้นตอนนี้ก็เพื่อให้ข้อมูลบริเวณจับหรือบริเวณเร่งเพียง 1 ระเบียนสามารถพัฒนาเป็นบล็อกข้อมูลที่มีคุณภาพ ซึ่งเป็นสิ่งจำเป็นต่อการนำไปใช้ค้นพบโมทีฟจากขั้นตอนการกลายพันธุ์ต่อไป โดยเทคนิคพื้นฐานของการคัดสรรบลอกที่มีคุณภาพประกอบด้วยขั้นตอนย่อย 2 ขั้นตอนคือ การแสกนสร้างบลอกอิงความเหมือน (similarity block scanning) และการคัดกรองด้วยกรอบคุณภาพ (constraint filter) ในขั้นตอนย่อยแรก ใช้หลัก "ความคล้ายคลึง" เพื่อเลือกสรรลำดับกรดอะมิโนจากกลุ่มโปรตีนที่เกี่ยวข้องเพื่อใช้สร้างบลอก จากนั้นในขั้นตอนย่อยต่อมาจึงมี "กรอบคุณภาพ" ในการคัดกรองให้ได้บลอกที่มีคุณภาพ

6.1.1 การแสกนสร้างบลอกอิงความเหมือน (similarity block scanning)

ในขั้นตอนย่อยแรกนั้น แต่ละบริเวณจับหรือบริเวณเร่ง 1 ระเบียนจากฐานข้อมูลบริเวณจับและบริเวณเร่ง สามารถนำมาหากลุ่มสายโปรตีนที่เกี่ยวข้องกับบริเวณดังกล่าวได้จากข้อมูลระบุลักษณะการทำงานของระเบียนนั้นในฐานข้อมูลบริเวณจับหรือบริเวณเร่ง โดยทุกฟังก์ชั่นเอนไซม์ที่มีการระบุลักษณะการทำงานของบริเวณจับหรือบริเวณเร่งที่เหมือนกันกับ

ระเบียนที่มีอยู่ก็คือฟังก์ชั่นเอนไซม์ที่เกี่ยวข้อง และทุกสายโปรตีนในฟังก์ชั่นเหล่านั้นก็คือกลุ่มสายโปรตีนที่ควรมีบางส่วนของสายโปรตีนทำงานเป็นบริเวณจับหรือบริเวณเร่งในลักษณะเช่นเดียวกัน

จากนั้นนำทุกสายโปรตีนที่เกี่ยวข้องนั้นมาค้นพบส่วนของสายโปรตีนที่คาดว่าน่าจะเป็นบริเวณทำงานเหมือนกันกับข้อมูลบริเวณจับหรือบริเวณเร่งที่มีอยู่ โดยมีสมมติฐานว่าส่วนของโปรตีนที่มีความ “คล้ายคลึง” กับบริเวณจับหรือบริเวณเร่งที่มีอยู่ น่าจะเป็นส่วนที่เกิดกลไกการทำงานในลักษณะเดียวกัน ดังนั้นจึงจำเป็นต้องมีเครื่องมือในการคำนวณความเหมือนเช่น ตารางคะแนนความเหมือน BLOSUM62 เป็นต้น เพื่อใช้ในการค้นพบส่วนของโปรตีนที่มีความคล้ายคลึงสูงสุดกับบริเวณจับหรือบริเวณเร่ง 1 ระเบียนที่มีอยู่

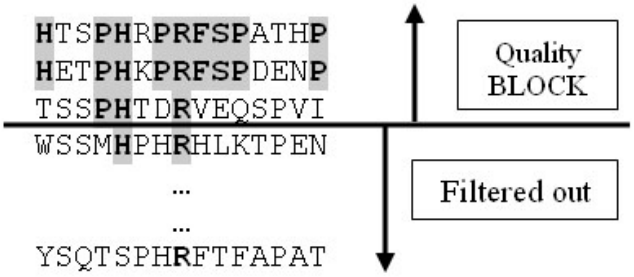
ในการนี้เริ่มจากการใช้บริเวณจับหรือบริเวณเร่ง 1 ระเบียนขนาด 15 กรดอะมิโนที่มีอยู่ดำเนินการแสกนเทียบเรียงส่วนย่อยในแต่ละสายโปรตีนตั้งแต่ต้นสายจนหมดสาย โดยคำนวณความคล้ายคลึงกันของกรดอะมิโนทั้ง 15 ตำแหน่งที่เทียบเรียงกันจากตารางคะแนนความเหมือน จากนั้นเลือกเฉพาะส่วนที่มีคะแนนสูงสุดเก็บไว้เป็นสมาชิกในบลอกที่เรียกว่าบลอกอิงความเหมือน (similarity block) ซึ่งจากข้อมูลของบลอกบริเวณจับหรือบริเวณเร่ง 1 ระเบียนที่มีอยู่ ก็จะกลายเป็นบลอกอิงความเหมือนที่มีสมาชิกเป็นสายลำดับกรดอะมิโนของบริเวณที่คาดว่าเป็นบริเวณจับหรือบริเวณเร่งนั้นเป็นจำนวนมาก

อย่างไรก็ตาม บริเวณจับหรือบริเวณเร่งที่ทำงานในลักษณะเดียวกัน อาจมีรายละเอียดที่แตกต่างกันได้ เช่น มีโครงสร้างต่างกัน ทิศทางในการเข้าจับและทำปฏิกิริยาต่างกัน เป็นต้น ดังนั้นจึงต้องคัดกรองบลอกอิงความเหมือนนั้นให้เหลือแต่เฉพาะส่วนข้อมูลที่มีคุณภาพเหมาะสมกับข้อมูลบริเวณจับหรือบริเวณเร่ง 1 ระเบียนที่ใช้เริ่มต้นสร้างบลอกนั้น รายละเอียดนำเสนอในขั้นตอนย่อยที่ 2 ของการคัดสรรบลอกที่มีคุณภาพ

6.1.2 การคัดกรองด้วยกรอบคุณภาพ

ในขั้นตอนย่อยที่สองนั้น แต่ละบล็อกอิงความเหมือนที่ได้มาเบื้องต้นนี้มีขนาดที่ใหญ่และไม่มีคุณภาพสำหรับค้นพบโมทีฟ จึงต้องมีการคัดกรองให้เหลือเฉพาะส่วนที่มีคุณภาพเหมาะสมกับข้อมูล 1 ระเบียนที่ใช้เริ่มต้นสร้างบลอกอิงความเหมือนนั้น โดยเริ่มต้นจากการจัดเรียง

สายลำดับกรดอะมิโนในบล็อกที่มีคะแนนความเหมือนสูงสุดไปต่ำสุด ก่อนที่จะใช้แนวคิดกรอบคุณภาพในการคัดกรองให้เหลือเฉพาะส่วนบลอกที่มีคุณภาพ



ภาพที่ 24 การคัดกรองบลอกอิงความเหมือนเป็นบลอกที่มีคุณภาพด้วยกรอบคุณภาพของ Smith และคณะวิจัย (Smith *et al.*, 1990)

กรอบคุณภาพที่ใช้จะต้องมีความเสถียรในแต่ละบลอกที่มีช่วงคะแนนความเหมือนที่แตกต่างกัน ซึ่งยากที่จะตัดสินใจได้ว่าควรใช้คะแนนความเหมือนเท่าใดเป็นเกณฑ์แยกบลอกที่มีคุณภาพออกจากบลอกที่ไม่มีคุณภาพ ดังนั้นจึงเลือกกรอบคุณภาพจากข้อสรุปในงานวิจัยของ Smith และคณะวิจัย (Smith *et al.*, 1990) ที่สรุปว่าโมทีฟที่มี “คุณภาพสูง” จะต้องประกอบด้วย “บริเวณอนุรักษ์” อย่างน้อย 3 ตำแหน่ง ซึ่งใช้เป็นกรอบคุณภาพ (constraint filter) ในการคัดกรองบลอกที่มีคุณภาพจากบลอกอิงความเหมือนดังตัวอย่างที่แสดงในภาพที่ 24

บลอกที่มีคุณภาพที่ได้นี้นำไปใช้ค้นพบโมทีฟจาก “แนวคิดการกลายพันธุ์” ดังที่ได้นำเสนอไปแล้ว อย่างไรก็ตามโมทีฟที่ได้ไม่สามารถนำไปใช้งานได้ในทันที ต้องมีการจัดกลุ่มโมทีฟอันเนื่องมาจากผลของการใช้ข้อมูลบริเวณจับและบริเวณเร่ง 1 ระเบียนในการคัดสรรบลอกที่มีคุณภาพในขั้นตอนนี้ ดังจะได้นำเสนอรายละเอียดในหัวข้อต่อไป

6.2 การจัดกลุ่มบริเวณปฏิกิริยาชีวเคมี (reactive site – group definition)

เพื่อแก้ปัญหาการมีข้อมูลบริเวณจับหรือบริเวณเร่งอยู่เพียง 1 ระเบียน ทำให้ในงานวิจัยนี้นำเสนอวิธีในการแก้ปัญหาดังกล่าวโดยการให้ข้อมูล 1 ระเบียนสร้างเป็นบลอกที่มีคุณภาพ 1 บลอก เพื่อค้นพบ 1 โมทีฟจากขั้นตอนแนวคิดการกลายพันธุ์ได้

อย่างไรก็ตาม ในกรณีที่บางบริเวณจับหรือบริเวณเร่งหลายบริเวณมีคำอธิบายความหมายการทำงานกลไกเอนไซม์ (site description) ที่เหมือนกัน แม้ว่าบริเวณจับหรือบริเวณเร่งเหล่านั้นจะนำมาพัฒนาได้ผลสุดท้ายเป็นหลายรีแอคทีฟโมทีฟ แต่รีแอคทีฟเหล่านั้นล้วนเป็นตัวแทนของบริเวณที่มีความหมายการทำงานกลไกเอนไซม์เดียวกัน ดังนั้นจึงควรจัดกลุ่มรีแอคทีฟโมทีฟที่มีความหมายการทำงานกลไกเอนไซม์ (site description) เหมือนกันเข้าด้วยกัน ก่อนนำไปพัฒนาเป็นโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์ ซึ่งในเบื้องต้นจัดกลุ่มรีแอคทีฟโมทีฟตามความหมายเชิงกลไกปฏิกิริยาชีวเคมีดังกล่าวที่ได้จากฐานข้อมูลบริเวณจับและบริเวณเร่ง รวมทั้งหมด 291 โมทีฟ

อย่างไรก็ตามพบว่าในกลุ่มรีแอคทีฟเดียวกัน แต่ละรีแอคทีฟโมทีฟอาจมีแพทเทิร์น (sequence pattern) ที่แตกต่างกันมาก อันเป็นผลมาจากขั้นตอนการคัดกรองบลอกที่มีคุณภาพด้วยกรอบคุณกรอบคุณภาพของ Smith และคณะวิจัย (Smith *et al.*, 1990) ทำให้รีแอคทีฟโมทีฟที่ได้จากบริเวณจับหรือบริเวณเร่งเดียวกันอาจมีตำแหน่งแกนของแพทเทิร์นที่เป็นบริเวณอนุรักษ์ไม่ตรงกันได้ ดังนั้นจึงใช้ผลความแตกต่างดังกล่าวในการแบ่งกลุ่มย่อยของรีแอคททีฟโมทีฟที่ได้จากบริเวณจับและบริเวณเร่งเดียวกัน โดยรีแอคทีฟโมทีฟที่มีบริเวณอนุรักษ์เหมือนกันจะถูกจัดเป็นกลุ่มย่อยเดียวกัน เรียกวิธีจัดแบ่งกลุ่มโมทีฟลักษณะนี้ว่า การจัดกลุ่มบริเวณปฏิกิริยาชีวเคมีด้วยบริเวณอนุรักษ์ (conserved region-group definition)

ด้วยวิธีดังกล่าว จึงสามารถจัดกลุ่มรีแอคทีฟโมทีฟที่มีประสิทธิภาพได้มากกว่าเดิม โดยในรีแอคทีฟโมทีฟที่ได้จากตาราง BLOSUM จัดกลุ่มย่อยได้ถึง 1,328 กลุ่ม และจัดกลุ่มรีแอคทีฟโมทีฟจากคอนเท็กซ์คุณสมบัติเคมีฟิสิกส์ของกรดอะมิโน (แผนภาพของ Taylor) ได้กลุ่มย่อย 1,390 กลุ่ม

จากส่วนงานหลักที่ 1 นี้ได้ผลออกมาเป็นรีแอคทีฟโมทีฟที่ได้จากแต่ละบริเวณจับและบริเวณเร่ง อย่างไรก็ตามเนื่องจากแต่ละฟังก์ชั่นเอนไซม์เกิดขึ้นจากการผสมผสานกันระหว่างบริเวณจับและบริเวณเร่ง ดังนั้นการใช้แต่ละรีแอคโมทีฟจากบริเวณจับหรือบริเวณเร่งในการทำนายประเภทฟังก์ชั่นเอนไซม์จึงไม่เพียงพอ และต้องการเครื่องมือในการสร้างโมเดลจากการผสมผสานรีแอคทีฟโมทีฟสำหรับทำนายประเภทฟังก์ชั่นเอนไซม์ เป็นเนื้อหาที่จะอธิบายในส่วนงานหลักที่ 2

7. การแปลงรูปฟอร์มความรู้พื้นหลัง

ในหัวข้อ 5 และหัวข้อ 6 เป็นการใช้องค์ความรู้ลักษณะต่างๆ ในการพัฒนาคุณภาพกลุ่มแทนที่กรดอะมิโนและบลอก อย่างไรก็ตามองค์ความรู้ที่ใช้ในการพัฒนาคุณภาพกลุ่มแทนที่อยู่ในรูปฟอร์มของคอนเท็กซ์ ในขณะที่การพัฒนาบลอกใช้องค์ความรู้ที่อยู่ในรูปฟอร์มของตารางคะแนนความเหมือน ซึ่งการพัฒนาคุณภาพกลุ่มแทนที่และบลอกเป็นกระบวนการต่อเนื่องกันให้ผลสุดท้ายเป็นรีแอคทีฟโมทีฟ ดังนั้นการใช้องค์ความรู้ที่ใช้ในการพัฒนาคุณภาพกลุ่มแทนที่และบลอกจึงควรมาจากความรู้พื้นหลังเดียวกัน ซึ่งจำเป็นต้องมีเครื่องมือในการแปลงความรู้พื้นหลังที่อยู่ในรูปฟอร์มหนึ่งไปเป็นอีกรูปฟอร์มหนึ่งได้ โดยเครื่องมือดังกล่าวได้มีการนำเสนอในงานวิจัย (Liewlom *et al.*, 2007) ที่เป็นเนื้อหาหลักของวิทยานิพนธ์นี้ ซึ่งทำให้มีทางเลือกในการใช้คอนเท็กซ์จากความรู้พื้นหลังในงานวิจัยอื่นที่เปิดกว้างมากขึ้น โดยรายละเอียดจะได้แสดงตามลำดับ

7.1 การแปลงความรู้พื้นหลังจากรูปฟอร์มคอนเท็กซ์เป็นตารางคะแนนความเหมือน

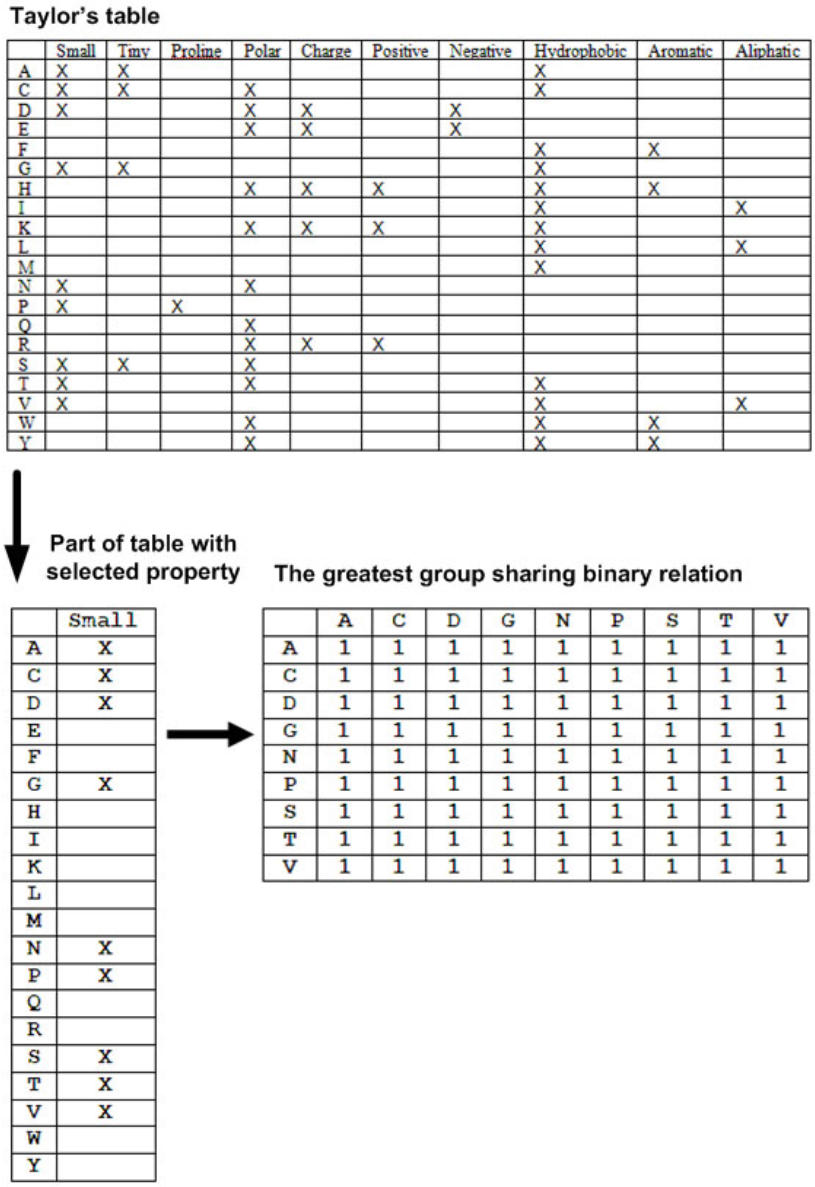
การแปลงคะแนนความเหมือนจากคอนเท็กซ์กรดอะมิโนและคุณสมบัติใช้หลักการเรียบง่ายโดยเปรียบเทียบคู่กรดอะมิโนใดใด ถ้ากรดอะมิโนคู่นั้นมีคุณสมบัติใดเหมือนกันให้มีค่าคะแนนความเหมือนเพิ่มขึ้น 1 คะแนน ยกตัวอย่างเช่น กรดอะมิโน A มีคุณสมบัติ {ขนาดเล็ก, ขนาดเล็กพิเศษ, ไม่ชอบน้ำ} ในขณะที่กรดอะมิโน C มีคุณสมบัติ {ขนาดเล็ก, ขนาดเล็กพิเศษ, ไม่ชอบน้ำ, มีขั้ว} ดังนั้นคะแนนความเหมือนของคู่กรดอะมิโน A และ C ก็คือ 3 ซึ่งได้จากจำนวนคุณสมบัติที่ A กับ C มีเหมือนกันคือ |{ขนาดเล็ก, ขนาดเล็กพิเศษ, ไม่ชอบน้ำ}|

อย่างไรก็ตาม ในกรณีที่จับคู่เพื่อคิดคะแนนความเหมือนระหว่างกรดอะมิโนชนิดเดียวกัน คะแนนความเหมือนจะมีน้ำหนักมากเป็นพิเศษ โดยเทียบได้กับกรณีของบริเวณอนุรักษ์ซึ่งต้องการกรดอะมิโนที่มีลักษณะเฉพาะพิเศษ ซึ่งจากการทดสอบพบว่าในงานศึกษานี้ใช้ค่าน้ำหนัก 4 เป็นค่าที่ดีที่สุด ยกตัวอย่างเช่น จับคู่กรดอะมิโน A กับ A จะมีค่าคะแนนความเหมือนเท่ากับ 4 x |{ขนาดเล็ก, ขนาดเล็กพิเศษ, ไม่ชอบน้ำ}| = 12 โดยตัวอย่างตารางคะแนนความเหมือนที่แปลงมาจากคอนเท็กซ์กรดอะมิโนและคุณสมบัติที่ได้จากแผนภาพของ Taylor นำเสนอดังในตารางที่ 6

ตารางที่ 6 ตารางคะแนนความเหมือนที่แปลงมาจากคอนเท็กซ์กรดอะมิโนและคุณสมบัติที่ได้จากแผนภาพคุณสมบัติเคมีฟิสิกส์ของกรดอะมิโนของ Taylor

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	12	3	1	0	1	3	1	1	1	1	1	1	1	0	0	2	2	2	1	1
C	3	16	2	1	1	3	2	1	2	1	1	2	1	1	1	3	3	2	2	2
D	1	2	16	3	0	1	2	0	2	0	0	2	1	1	2	2	2	1	1	1
E	0	1	3	12	0	0	2	0	2	0	0	1	0	1	2	1	1	0	1	1
F	1	1	0	0	8	1	2	1	1	1	1	0	0	0	0	0	1	1	2	2
G	3	3	1	0	1	12	1	1	1	1	1	1	1	0	0	2	2	2	1	1
H	1	2	2	2	2	1	20	1	4	1	1	1	0	1	3	1	2	1	3	3
I	1	1	0	0	1	1	1	8	1	2	1	0	0	0	0	0	1	2	1	1
K	1	2	2	2	1	1	4	1	16	1	1	1	0	1	3	1	2	1	2	2
L	1	1	0	0	1	1	1	2	1	8	1	0	0	0	0	0	1	2	1	1
M	1	1	0	0	1	1	1	1	1	1	4	0	0	0	0	0	1	1	1	1
N	1	2	2	1	0	1	1	0	1	0	0	8	1	1	1	2	2	1	1	1
P	1	1	1	0	0	1	0	0	0	0	0	1	8	0	0	1	1	1	0	0
Q	0	1	1	1	0	0	1	0	1	0	0	1	0	4	1	1	1	0	1	1
R	0	1	2	2	0	0	3	0	3	0	0	1	0	1	12	1	1	0	1	1
S	2	3	2	1	0	2	1	0	1	0	0	2	1	1	1	12	2	1	1	1
T	2	3	2	1	1	2	2	1	2	1	1	2	1	1	1	2	12	2	2	2
V	2	2	1	0	1	2	1	2	1	2	1	1	1	0	0	1	2	12	0	0
W	1	2	1	1	2	1	3	1	2	1	1	1	0	1	1	1	2	0	12	3
Y	1	2	1	1	2	1	3	1	2	1	1	1	0	1	1	1	2	0	3	12

7.2 การแปลงความรู้พื้นหลังจากรูปฟอร์มตารางคะแนนความเหมือนเป็นคอนเท็กซ์



	Small	Tiny	Proline	Polar	Charge	Positive	Negative	Hydrophobic	Aromatic	Aliphatic
A	X	X						X		
C	X	X		X				X		
D	X			X	X		X			
E				X	X		X			
F								X	X	
G	X	X						X		
H				X	X	X		X	X	
I								X		X
K				X	X	X		X		
L								X		X
M								X		
N	X			X						
P	X		X							
Q				X						
R				X	X	X				
S	X	X		X						
T	X			X				X		
V	X							X		X
W				X				X	X	
Y				X				X	X	

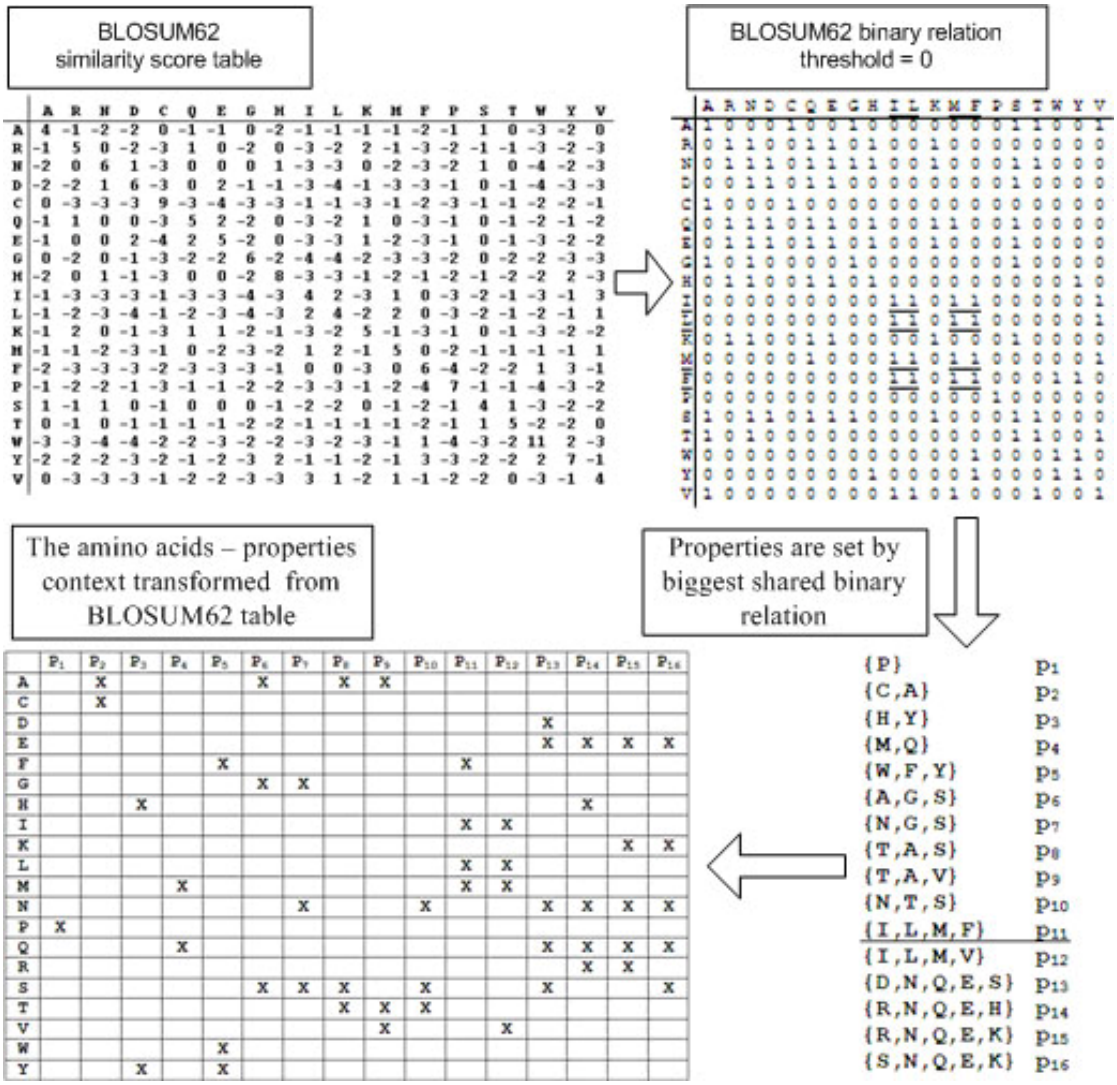
	Small
A	X
C	X
D	X
E	
F	
G	X
H	
I	
K	
L	
M	
N	X
P	X
Q	
R	
S	X
T	X
V	X
W	
Y	

	A	C	D	G	N	P	S	T	V
A	1	1	1	1	1	1	1	1	1
C	1	1	1	1	1	1	1	1	1
D	1	1	1	1	1	1	1	1	1
G	1	1	1	1	1	1	1	1	1
N	1	1	1	1	1	1	1	1	1
P	1	1	1	1	1	1	1	1	1
S	1	1	1	1	1	1	1	1	1
T	1	1	1	1	1	1	1	1	1
V	1	1	1	1	1	1	1	1	1

ภาพที่ 25 ความสัมพันธ์ระหว่างคุณสมบัติของกรดอะมิโนที่แฝงอยู่รูปคะแนนในตารางคะแนนความเหมือน

ในการแปลงความรู้พื้นหลังจากรูปฟอร์มตารางคะแนนความเหมือน เช่น ตาราง BLOSUM62 ให้เป็นคอนเท็กซ์ มีขั้นตอนที่ค่อนข้างซับซ้อน โดยในลำดับแรกจำเป็นต้องระบุให้ได้เสียก่อนว่า คุณสมบัติของกรดอะมิโนแบบคอนเท็กซ์แฝงอยู่ในรูปคะแนนในตารางคะแนนความเหมือนได้อย่างไร ปัญหานี้สามารถพิจารณาจากภาพที่ 25

พิจารณาจากภาพที่ 25 จะเห็นได้อย่างชัดเจนว่าแต่ละคุณสมบัติกรดอะมิโนสามารถแปลงเป็นความสัมพันธ์ทวิภาคแบบทั้งกลุ่มของกลุ่มกรดอะมิโนที่มีคุณสมบัติเดียวกันที่ใหญ่ที่สุดได้ ดังนั้นจึงใช้หลักการนี้ในการแปลงตารางคะแนนความเหมือนเป็นคอนเท็กซ์ ประกอบด้วย 3 ขั้นตอนดังนี้



ภาพที่ 26 ขั้นตอนการแปลงตารางคะแนนความเหมือน BLOSUM62 เป็นคอนเท็กซ์

ขั้นตอนแรก ให้กำหนดค่าวิกฤต (threshold) ของคะแนนความเหมือนที่ชี้ว่ากรดอะมิโนคู่นั้นมีคุณสมบัติที่ตรงกันอย่างน้อย 1 คุณสมบัติ (ในงานศึกษานี้ให้ค่าวิกฤตเท่ากับ 0) จากนั้นแปลงค่าคะแนนทั้งหมดในตารางให้กลายเป็นความสัมพันธ์ทวิภาคให้หมด โดยคะแนนความเหมือนของกรดอะมิโนคู่ใดมีค่ามากกว่าหรือเท่ากับค่าวิกฤตให้แทนค่าด้วย 1 ส่วนคะแนนความเหมือนที่ต่ำกว่าค่าวิกฤตให้แทนค่าด้วย 0 จากนั้นในขั้นตอนที่ 2 ทำการค้นหากลุ่มของกรดอะมิโนที่ใหญ่ที่สุดทุกกลุ่มที่มีความสัมพันธ์ทวิภาคแบบทั้งกลุ่ม โดยแต่ละกลุ่มหมายถึงการมีคุณสมบัติที่ตรงกันของกรดอะมิโนในกลุ่มนั้น จำนวนกลุ่มที่ได้ย่อมหมายถึงจำนวนคุณสมบัติที่แฝงอยู่ในตารางคะแนนความเหมือนนั้น ซึ่งในขั้นตอนสุดท้ายก็คือการนำทุกกลุ่มคุณสมบัติมารวมกันเป็นตารางคอนเท็กซ์ ทั้งนี้สามารถศึกษาขั้นตอนอย่างละเอียดได้จากภาพที่ 26

ในบทต่อไปเป็นการแสดงผลการวิเคราะห์คุณภาพของรีแอคทีฟโมทีฟและประสิทธิภาพเมื่อใช้ทำนายประเภทฟังก์ชั่นเอนไซม์ ซึ่งเป็นการพิสูจน์ความเป็นไปได้ของแนวทางวิจัยที่พัฒนาคุณลักษณะเด่นของกลุ่มข้อมูลที่สร้างจากส่วนที่มีข้อมูลน้อยในการทำนายประเภทข้อมูลขนาดใหญ่ได้ดี

ผลและวิจารณ์

ในบทนี้นำเสนอผลการทดลองเปรียบเทียบประสิทธิภาพของรีแอคทีฟโมทีฟชนิดต่างๆ กับโมทีฟ PROSITE ในการทำนายประเภทฟังก์ชั่นเอนไซม์ โดยการนำเสนอแยกออกเป็น 3 ส่วน คือ 1) เปรียบเทียบคุณภาพและความเสถียรของรีแอคทีฟโมทีฟ 2) เปรียบเทียบความแม่นยำโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์ที่สร้างจากรีแอคทีฟโมทีฟชนิดต่างๆ ที่ไม่ใช้เทคนิคการพัฒนาคุณภาพบลอก 3) เปรียบเทียบความแม่นยำโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์ที่สร้างจากรีแอคทีฟโมทีฟชนิดต่างๆ ที่ใช้ร่วมกับเทคนิคการพัฒนาคุณภาพบลอก

1. เปรียบเทียบคุณภาพของรีแอคทีฟโมทีฟ

งานวิจัยนี้มีวัตถุประสงค์หลักในการพัฒนาคุณลักษณะเด่นของกลุ่มข้อมูลสายโปรตีนสำหรับทำนายประเภทฟังก์ชั่นเอนไซม์ที่เรียกว่า “รีแอคทีฟโมทีฟ” ดังนั้นในการนำคุณลักษณะเด่นของกลุ่มข้อมูลสายโปรตีนชนิดนี้ไปใช้งานทำนายประเภทฟังก์ชั่นเอนไซม์จึงต้องมีการวิเคราะห์เปรียบเทียบคุณภาพและความเสถียรของโมทีฟประเภทต่างๆ ประกอบด้วย โมทีฟ PROSITE และรีแอคทีฟโมทีฟที่ได้จากความรู้พื้นหลังที่แตกต่างกัน โดยรีแอคทีฟโมทีฟที่ได้จากความรู้พื้นหลังตาราง BLOSUM62 เราเรียกว่า รีแอคทีฟโมทีฟแสดงคุณสมบัติ BLOSUM (BLOSUM – reactive motif) รีแอคทีฟโมทีฟที่ได้จากความรู้พื้นหลังแผนภาพของ Taylor เรียกว่า รีแอคทีฟโมทีฟแสดงคุณสมบัติเคมีฟิกส์ (physicochemistry – reactive motif) โดยมีรีแอคทีฟโมทีฟอีกชนิดหนึ่งที่ไม่มีกลุ่มแทนที่กรดอะมิโนเป็นองค์ประกอบของโมทีฟ ไม่มีการใช้ความรู้พื้นหลังใดใด โดยใช้แต่เพียงบริเวณอนุรักษ์เป็นองค์ประกอบย่อยของรีแอคทีฟโมทีฟ เราเรียกรีแอคทีฟโมทีฟชนิดนี้ว่า รีแอคทีฟโมทีฟที่ไม่มีกลุ่มแทนที่ (without substitution group – reactive motif)

คุณภาพของโมทีฟที่พัฒนามาจากบริเวณจับและบริเวณเร่งที่ดีก็คือ ความสามารถในการระบุว่าสายโปรตีนใด “มี” หรือ “ไม่มี” บริเวณจับหรือบริเวณเร่งที่สัมพันธ์ต่อการเกิดฟังก์ชั่นเอนไซม์นั้นได้อย่างถูกต้อง โดยในการวัดค่าคุณภาพของโมทีฟ ใช้ค่า true positive (TP) false positive (FP) true negative (TN) และ false negative (FN) ในการคำนวณค่าคุณภาพ คือ ค่า accuracy precision sensitivity และ specificity โดยคำนวณค่า %coverage เพิ่มเติมเข้าไปเพื่อเปรียบเทียบคุณภาพการใช้งานโมทีฟที่ขนาดของกลุ่มข้อมูลตั้งแต่ 3 ฟังก์ชั่น 25 ฟังก์ชั่น ไป

จนกระทั่ง 235 ฟังก์ชั่น โดยผลการเปรียบเทียบคุณภาพของโมทีฟประเภทต่างๆ นำเสนอดังในตารางที่ 7 และ 8

ตารางที่ 7 เปรียบเทียบค่า TP FP TN FN ของแต่ละรีแอคทีฟโมทีฟและ PROSITE เมื่อใช้กับกลุ่มโปรตีนระหว่าง 3 ถึง 235 ฟังก์ชั่น

Motif	#fn	all positive (P)	all negative (N)	TP	FP	TN	FN
PROSITE	3	288	576	133	0	576	155
	25	3,114	71,730	1,479	8	71,722	1,635
	76	6,219	477,961	3,619	27	477,934	2,600
รีแอคทีฟโมทีฟแสดงคุณสมบัติ BLOSUM	3	288	576	224	194	382	64
	25	5,095	127,961	2,969	1,876	126,085	2,126
	235	79,533	13,015,907	41,665	81,497	12934410	37868
รีแอคทีฟโมทีฟแสดงคุณสมบัติเคมีฟิสิกส์	3	288	576	143	5	571	145
	25	5,095	127,961	2,850	1,214	126,747	2,245
	235	79,533	13,015,907	39,747	17,402	12,998,505	39,786
รีแอคทีฟโมทีฟที่ไม่มีกลุ่มแทนที่	3	N/A	N/A	N/A	N/A	N/A	N/A
	25	5,095	127,961	3,564	10,310	117,651	1,531
	235	79,533	13,015,907	47,766	613,836	1,2402,071	31,767

หมายเหตุ ค่า #fn หมายถึงจำนวนฟังก์ชั่นที่มีในแต่ละกลุ่มโปรตีนเอนไซม์ ค่า P หมายถึงจำนวนบริเวณจับและบริเวณเร่งที่มีทั้งหมดในกลุ่มโปรตีนนั้น และ ค่า N หมายถึงจำนวนทั้งหมดของบริเวณจับหรือบริเวณเร่งที่โมทีฟแต่ละชนิดไม่ควรระบุผิดพลาดว่ามีโมทีฟนั้น

ตารางที่ 8 เปรียบเทียบค่าคุณภาพของแต่ละรีแอคทีฟโมทีฟและ PROSITE เมื่อใช้กับกลุ่มโปรตีนระหว่าง 3 ถึง 235 ฟังก์ชั่น

Motif	#fn	accuracy	precision	sensitivity	specificity	%coverage
PROSITE	3	0.8206	1.0000	0.4618	1.0000	46.18
	25	0.9780	0.9946	0.4750	0.9999	70.42
	76	0.9946	0.9926	0.5819	0.9999	79.04
รีแอคทีฟโมทีฟแสดงคุณสมบัติ BLOSUM	3	0.7014	0.5359	0.7778	0.6632	92.71
	25	0.9699	0.6128	0.5827	0.9853	96.32
	235	0.9909	0.3383	0.5239	0.9937	96.55
รีแอคทีฟโมทีฟแสดงคุณสมบัติเคมีฟิสิกส์	3	0.8264	0.9662	0.4965	0.9913	50.35
	25	0.9740	0.7013	0.5594	0.9905	88.74
	235	0.9956	0.6955	0.4998	0.9987	85.61
รีแอคทีฟโมทีฟที่ไม่มีกลุ่มแทนที่	3	N/A	N/A	N/A	N/A	N/A
	25	0.9110	0.2569	0.6995	0.9194	99.28
	235	0.9507	0.0722	0.6006	0.9528	99.69

จากตารางที่ 7 และ 8 ในส่วนสีเทาคือการวัดคุณภาพของโมทีฟที่ใช้กับข้อมูลขนาด 3 ฟังก์ชั่น ซึ่งมีลักษณะพิเศษคือเป็นการใช้งานรีแอคทีฟโมทีฟที่สร้างจากข้อมูลเฉพาะบริเวณจับหรือบริเวณเร่งโดยตรงเท่านั้น โดยไม่ได้ใช้ร่วมกับเทคนิคการพัฒนาคุณภาพบลอก ซึ่งพบว่ารีแอคทีฟโมทีฟแสดงคุณสมบัติเคมีฟิสิกส์ให้ค่าคุณภาพในทุกค่าที่ใกล้เคียงกับ PROSITE มากที่สุด

สำหรับคุณภาพโดยทั่วไปของโมทีฟ PROSITE เมื่อใช้กับทุกระดับขนาดกลุ่มข้อมูล คือ 3 ฟังก์ชั่น 25 ฟังก์ชั่น และ 76 ฟังก์ชั่น มีลักษณะที่เหมือนกันคือมีค่า precision สูงมากที่ประมาณ 1 แต่มีค่า sensitivity และ %coverage ที่ต่ำมากที่สุด โดยรีแอคทีฟโมทีฟที่มีลักษณะใกล้เคียงกับ PROSITE เมื่อใช้กับทุกระดับขนาดกลุ่มข้อมูล คือ 3 ฟังก์ชั่น 25 ฟังก์ชั่น และ 235 ฟังก์ชั่น คือ รีแอคทีฟโมทีฟแสดงคุณสมบัติเคมีฟิสิกส์ โดยที่กลุ่มข้อมูลขนาด 25 ฟังก์ชั่นและ 235 ฟังก์ชั่น รีแอคทีฟโมทีฟแสดงคุณสมบัติเคมีฟิสิกส์มีค่า precision ที่ประมาณ 0.7 ตกลงจากค่า precision ที่ใช้กับข้อมูลขนาด 3 ฟังก์ชั่นที่มีค่าประมาณ 0.97 ทั้งนี้การลดลงของค่า precision นี้มีสมมติฐานเนื่องจากการไม่มีข้อมูลบริเวณจับและบริเวณเร่งให้ใช้งานโดยตรง จึงต้องใช้ร่วมกับเทคนิคการพัฒนาข้อมูล

ซึ่งมีประสิทธิภาพดีในระดับหนึ่ง แต่เทคนิคดังกล่าวยังคงไม่สามารถทดแทนการขาดแคลนข้อมูลได้อย่างสมบูรณ์

2. เปรียบเทียบความแม่นยำในโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์ที่สร้างจากรีแอคทีฟโมทีฟชนิดต่างๆ ที่ได้จากเทคนิคการพัฒนากลุ่มแทนที่

เนื่องจากการพัฒนาคุณภาพของรีแอคทีฟโมทีฟมี 2 งานย่อยคือ การพัฒนาคุณภาพกลุ่มแทนที่ด้วยแนวคิดการควบคุมการกลายพันธุ์ และการพัฒนาคุณภาพบลอก โดยกล่าวได้ว่าการพัฒนาคุณภาพกลุ่มแทนที่เป็นส่วนงานสำคัญที่ทำให้ได้มาซึ่งองค์ประกอบของรีแอคทีฟโมทีฟที่เรียกว่ากลุ่มแทนที่ที่สมบูรณ์ โดยการพัฒนาคุณภาพบลอกเป็นส่วนงานที่เข้ามาช่วยทำให้สามารถใช้ข้อมูลบริเวณจับและบริเวณเร่งที่มีปริมาณน้อยมากๆ ขนาด 1 ระเบียนได้ และทำให้การใช้งานรีแอคทีฟโมทีฟสามารถขยายขนาดกลุ่มข้อมูลเอนไซม์ได้ถึง 235 ฟังก์ชั่น 291 รีแอคทีฟโมทีฟ (ทำได้สูงสุด 720 รีแอคทีฟโมทีฟ) จากเดิมที่สามารถค้นพบรีแอคทีฟโมทีฟได้เพียงประมาณ 50 รีแอคทีฟโมทีฟเท่านั้น

ดังนั้นจึงเป็นสิ่งที่น่าสนใจในการวิเคราะห์คุณภาพของรีแอคทีฟโมทีฟเมื่อใช้แต่เทคนิคการพัฒนาคุณภาพกลุ่มแทนที่โดยไม่ได้ใช้ร่วมกับเทคนิคการพัฒนาคุณภาพบลอก ในการทำนายประเภทฟังก์ชั่นเอนไซม์ที่กลุ่มข้อมูลขนาดเล็กจำนวนหนึ่ง ที่มีข้อมูลบริเวณจับและบริเวณเร่งที่เพียงพอต่อการนำมาสร้างรีแอคทีฟโมทีฟ

สำหรับกลุ่มข้อมูลที่ใช้เป็นกลุ่มฝึกสอนและทดสอบด้วยโมเดลการทำนายประเภทเอนไซม์ C4.5 ขนาด 5 ฟังก์ชั่น ประกอบด้วยสายโปรตีนทั้งหมด 439 สาย โดยเป็นสายโปรตีนที่มีข้อมูลบริเวณจับและบริเวณเร่งจำนวน 90 สาย มีข้อมูลบริเวณจับและบริเวณเร่งทั้งหมด 221 ระเบียน โดยบริเวณจับและบริเวณเร่งที่มีข้อมูลน้อยที่สุดคือ 1 ระเบียน มากที่สุดคือ 32 ระเบียน รวมทั้งสิ้น 19 บริเวณ ได้เป็นรีแอคทีฟโมทีฟทั้งหมด 19 รีแอคทีฟโมทีฟ โดยการจัดกลุ่มฝึกสอนและทดสอบมีทั้งหมด 2 ลักษณะคือ 1) โปรตีนเอนไซม์ทั้ง 439 สายแบ่งออกเป็นกลุ่มฝึกสอน 67% (293) ส่วนที่เหลือ 146 สายคือกลุ่มข้อมูลทดสอบ และ 2) โปรตีนเอนไซม์ทั้ง 439 จัดกลุ่มฝึกสอนและทดสอบด้วยวิธี 5-fold cross validation ผลการทดสอบแสดงดังตารางที่ 9

ตารางที่ 9 เปรียบเทียบผลการทำนายฟังก์ชั่นเอนไซม์จากรีแอคทีฟโมทีฟประเภทต่างๆ ด้วย C4.5 เมื่อไม่ได้ใช้เทคนิคการพัฒนาคุณภาพบลอก จำนวน 5 ฟังก์ชั่น ข้อมูลโปรตีน 439 สาย

รีแอคทีฟโมทีฟ	**ความแม่นยำ (%Correctly Classified Instance)**	
	กลุ่มฝึกสอน(67%) / ทดสอบ (33%) (293 สาย/ 146 สาย)	**5-fold cross validation**
รีแอคทีฟโมทีฟที่ไม่มีกลุ่มแทนที่	74.67	76.77
รีแอคทีฟโมทีฟแสดงคุณสมบัติ BLOSUM	76.67	75.17
รีแอคทีฟโมทีฟแสดงคุณสมบัติเคมีฟิสิกส์	76.00	77.68

จากตารางที่ 9 แสดงถึงผลความแม่นยำในการทำนายประเภทฟังก์ชั่นเอนไซม์โดยใช้รีแอคทีฟโมทีฟประเภทต่างๆ ที่ 5 ฟังก์ชั่นโดยพบว่าทุกประเภทให้ผลใกล้เคียงกัน ดังนั้นจึงมีการวิเคราะห์เพิ่มเติมในระดับคุณภาพของโมทีฟในการระบุการมีบริเวณจับและบริเวณเร่งได้อย่างถูกต้อง ซึ่งแสดงผลดังกล่าวรวมไว้ในตารางที่ 8 ของส่วนที่ 1 ในบทนี้แล้ว

ในการวิเคราะห์คุณภาพของรีแอคทีฟโมทีฟเพื่อใช้ในการทำนายประเภทฟังก์ชั่นเอนไซม์ให้ชัดเจนขึ้นนั้น เราเปรียบเทียบกับโมทีฟ PROSITE ที่พัฒนาโดยผู้เชี่ยวชาญ โดยใน 5 ฟังก์ชั่นมีเพียง 3 ฟังก์ชั่น (288 สายโปรตีน) และ 3 โมทีฟที่มีคำอธิบายการทำงาน (site description) ที่ตรงกันหรือเป็นบริเวณจับและบริเวณเร่งเดียวกัน โดยผลเปรียบเทียบความแม่นยำในการทำนายประเภทฟังก์ชั่นเอนไซม์ด้วยรีแอคทีฟโมทีฟและ PROSITE แสดงดังตารางที่ 10

ตารางที่ 10 เปรียบเทียบผลความแม่นยำการทำนายประเภทฟังก์ชั่นเอนไซม์ระหว่างระบบการทำนายประเภทฟังก์ชั่นเอนไซม์ C4.5 ที่ได้จากรีแอคทีฟโมทีฟประเภทต่างๆ กับ PROSITE ในชุดข้อมูลขนาด 3 ฟังก์ชั่น จำนวนโปรตีน 288 สาย ที่มีโมทีฟตรงกัน จำนวน 3 โมทีฟ เมื่อไม่ใช้เทคนิคการพัฒนาคุณภาพบลอก

ประเภทโมทีฟ	% Correctly Classified Instance (C4.5)	
	กลุ่มฝึกสอน(67%) / ทดสอบ (33%) (192 สาย/ 96 สาย)	5-fold cross validation
PROSITE	76.5306	75.6944
รีแอคทีฟโมทีฟแสดงคุณสมบัติ BLOSUM	71.4286	70.1389
รีแอคทีฟโมทีฟแสดงคุณสมบัติเคมีฟิสิกส์	76.5306	77.4306
รีแอคทีฟโมทีฟที่ไม่มีกลุ่มแทนที่	N/A	N/A

จากตารางที่ 10 ในฟังก์ชั่นเอนไซม์ที่ตรงกันจำนวน 3 ฟังก์ชั่นที่ทุกระบบทำนายประเภทฟังก์ชั่นเอนไซม์ใช้โมทีฟจากบริเวณจับและบริเวณเร่งเดียวกัน พบว่ารีแอคทีฟโมทีฟแสดงคุณสมบัติเคมีฟิสิกส์ให้ผลที่ดีที่สุด โดยดีกว่า PROSITE เพียงเล็กน้อย ในขณะที่รีแอคทีฟโมทีฟที่ไม่มีกลุ่มแทนที่ไม่สามารถหาโมทีฟจากบริเวณจับและบริเวณเร่งเดียวกันมาเปรียบเทียบได้

สำหรับการวิเคราะห์ผลคุณภาพของโมทีฟที่มีต่อความแม่นยำในการทำนายประเภทเอนไซม์ที่ 3 ฟังก์ชั่น พบว่าค่า precision ของโมทีฟมีผลต่อความแม่นยำมากกว่าค่า sensitivity และ %coverage โดยเมื่อเปรียบเทียบกับรีแอคทีฟโมทีฟแสดงคุณสมบัติ BLOSUM ที่มีค่า sensitivity และ ค่า %coverage ที่มากกว่า แต่กลับให้เปอร์เซ็นต์ความแม่นยำในการทำนายที่ต่ำกว่าทั้งในระดับโมทีฟที่ใช้ทำนายบริเวณจับและบริเวณเร่ง และในระดับทำนายประเภทฟังก์ชั่นเอนไซม์ โดยแนวโน้มนี้มีปรากฏในการวัดระดับ 25 และ 235 ฟังก์ชั่นด้วย

ทั้งนี้ผลการเปรียบเทียบจะสรุปผลได้ชัดเจนก็ต่อเมื่อขยายขนาดกลุ่มทดสอบให้มีจำนวนฟังก์ชั่นและจำนวนโปรตีนมากกว่านี้ ซึ่งทำได้เมื่อใช้ประกอบกับเทคนิคการพัฒนาคุณภาพบลอก ทำให้เพิ่มกลุ่มฟังก์ชั่นเอนไซม์ที่ใช้พัฒนาโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์ได้ถึง 235 ฟังก์ชั่น ครอบคลุมโปรตีนถึง 19,258 สาย โดยผลวิเคราะห์ดังกล่าวแสดงในส่วนที่ 3 ในบทนี้

3. เปรียบเทียบความแม่นยำในโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์ที่สร้างจากรีแอคทีฟโมทีฟชนิดต่างๆ ที่ได้จากเทคนิคการพัฒนากลุ่มแทนที่ร่วมกับเทคนิคการพัฒนาคุณภาพบลอก

ส่วนนี้นำเสนอผลเปรียบเทียบความแม่นยำโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์ที่ได้จากรีแอคทีฟโมทีฟเมื่อใช้ร่วมกับเทคนิคการพัฒนาคุณภาพบลอก โดยเปรียบเทียบกับโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์ที่ได้จาก PROSITE พบว่ากลุ่มข้อมูลที่ใช้รีแอคทีฟโมทีฟมีถึง 19,258 สายโปรตีนใน 235 ฟังก์ชั่น เมื่อกำหนดให้แต่ละฟังก์ชั่นมีโปรตีนอยู่ระหว่าง 10 ถึง 1,000 สาย และมีข้อมูลบริเวณจับและบริเวณเร่งไม่น้อยกว่า 2 บริเวณ (site descriptions) ต่อ 1 ฟังก์ชั่น ในขณะที่กลุ่มข้อมูลที่ใช้ PROSITE มีเพียง 2,815 สายใน 76 ฟังก์ชั่น เมื่อกำหนดให้แต่ละฟังก์ชั่นมีโปรตีนอยู่ระหว่าง 5 ถึง 1,000 สายและมีข้อมูลโมทีฟไม่น้อยกว่า 2 โมทีฟต่อ 1 ฟังก์ชั่น

ทั้งนี้พบว่ากลุ่มโปรตีนที่ใช้รีแอคทีฟโมทีฟและ PROSITE มีกลุ่มข้อมูลที่ตรงกันที่ 25 ฟังก์ชั่น ครอบคลุมโปรตีน 2,579 สาย จึงนำข้อมูลกลุ่มเดียวกันนี้มาเปรียบเทียบผลความแม่นยำของการใช้โมทีฟชนิดต่างๆ ในการสร้างโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์ แสดงผลดังตารางที่ 11

จากตารางที่ 11 เมื่อเปรียบเทียบระบบทำนายประเภทฟังก์ชั่นเอนไซม์ที่ได้จากรีแอคทีฟโมทีฟประเภทต่างๆ และที่ได้จาก PROSITE พบว่ารีแอคทีฟโมทีฟทุกประเภทสามารถนำมาใช้ประโยชน์ในการทำนายประเภทฟังก์ชั่นเอนไซม์ด้วย C4.5 ได้แม่นยำกว่า PROSITE ที่เอนไซม์ 25 ฟังก์ชั่นเดียวกัน

ตารางที่ 11 เปรียบเทียบคุณภาพรีแอคทีฟโมทีฟและ PROSITE ในการทำนายบริเวณจับและบริเวณเร่ง และความแม่นยำเมื่อใช้ทำนายประเภทฟังก์ชั่นเอนไซม์ที่ 25 ฟังก์ชั่น 2,579 สายโปรตีนด้วย C4.5 (5-fold cross validation)

ประเภทโมทีฟ	% Correctly Classified Instance
PROSITE	65.8009
รีแอคทีฟโมทีฟแสดงคุณสมบัติ BLOSUM	74.5310
รีแอคทีฟโมทีฟแสดงคุณสมบัติเคมีฟิสิกส์	73.0880
รีแอคทีฟโมทีฟที่ไม่มีกลุ่มแทนที่	71.2843

สำหรับการเปรียบเทียบความแม่นยำของโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์ที่ได้จากรีแอคทีฟโมทีฟที่ 235 ฟังก์ชั่น และที่ได้จาก PROSITE ที่ 76 ฟังก์ชั่น เปรียบเทียบดังแสดงในตารางที่ 12 และ 13 โดยในตารางที่ 12 ที่เป็นการเปรียบเทียบเฉพาะโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์ที่ได้จากรีแอคทีฟโมทีฟประเภทต่างๆ นั้น เปรียบเทียบในหลายลักษณะด้วยกันคือ โมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์ที่ใช้รีแอคทีฟแบบไม่มีกลุ่มแทนที่ และที่ได้กลุ่มแทนที่จากความรู้พื้นหลังที่แตกต่างกันคือระหว่าง BLOSUM กับ เซ็ทกรดอะมิโนและคุณสมบัติเชิงเคมีฟิสิกส์ของ Taylor นอกจากนี้ยังเปรียบเทียบการใช้รีแอคทีฟโมทีฟเป็นคุณลักษณะเด่นของกลุ่มข้อมูลสายโปรตีนสำหรับสร้างระบบทำนายประเภทฟังก์ชั่นเอนไซม์เมื่อมีการจัดกลุ่มรีแอคทีฟโมทีฟตามคำอธิบายบริเวณจับและบริเวณเร่งที่มีตรงกันในฐานข้อมูล (site description) (เรียกย่อว่าการจัดกลุ่มด้วยฐานข้อมูล) เปรียบเทียบกับเมื่อมีการพัฒนาวิธีจัดกลุ่มพิเศษตามบริเวณอนุรักษ์ที่รีแอคทีฟโมทีฟในกลุ่มนั้นมีเหมือนกัน (เรียกย่อว่าการจัดกลุ่มด้วยบริเวณอนุรักษ์)

ทุกโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์ดังที่กล่าวมา สร้างโดยขั้นตอนวิธี C4.5 แบบ 5-fold cross validation ผลการเปรียบเทียบแสดงดังตารางที่ 12 และ 13 ตามลำดับ

ตารางที่ 12 ผลการเปรียบเทียบความแม่นยำระหว่างโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์ที่ใช้รีแอคทีฟโมทีฟประเภทต่างๆ ที่ 235 ฟังก์ชั่น โปรตีน 19,258 สาย

การจัดกลุ่มบริเวณปฏิกิริยาชีวเคมี	รีแอคทีฟโมทีฟ					
	ไม่มีกลุ่มแทนที่		แสดงคุณสมบัติ BLOSUM		แสดงคุณสมบัติเคมีฟิสิกส์	
	จำนวนโมทีฟ	C4.5 (%)	จำนวนโมทีฟ	C4.5 (%)	จำนวนโมทีฟ	C4.5 (%)
จัดกลุ่มด้วยฐานข้อมูล	291	60.84	291	**68.69**	291	64.38
จัดกลุ่มด้วยบริเวณอนุรักษ์	1324	70.57	1328	71.66	1390	**72.58**

ตารางที่ 13 ความแม่นยำของโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์ที่ใช้โมทีฟ PROSITE

เงื่อนไขจำนวนสมาชิกโปรตีนในแต่ละฟังก์ชั่นเอนไซม์	จำนวนฟังก์ชั่น	จำนวนโมทีฟ	จำนวนสายโปรตีน	C4.5 (%)
ระหว่าง 10 และ 1000	42	36	2579	37.15
ระหว่าง 5 และ 1000	76	65	2815	**67.25**

ผลการเปรียบเทียบเฉพาะโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์ที่ใช้รีแอคทีฟโมทีฟ พบว่าในกรณีของการจัดกลุ่มบริเวณปฏิกิริยาชีวเคมีด้วยฐานข้อมูลบริเวณจับและบริเวณเร่ง โมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์ที่ใช้รีแอคทีฟแสดงคุณสมบัติ BLOSUM ให้ความแม่นยำดีที่สุดที่ 68.9% ในขณะที่รีแอคทีฟโมทีฟแสดงคุณสมบัติเคมีฟิสิกส์ให้ผลแม่นยำดีที่สุดที่ 72.58% ในกรณีของการจัดกลุ่มรีแอคทีฟโมทีฟด้วยบริเวณอนุรักษ์ของโมทีฟนั้น อย่างไรก็ตามค่าความแม่นยำของทุกโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์มีค่าที่ใกล้เคียงกันมาก

ในการพิจารณาความแม่นยำของโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์เมื่อใช้โมทีฟ PROSITE เมื่อใช้เงื่อนไขเดียวกับการใช้รีแอคทีฟโมทีฟคือแต่ละกลุ่มฟังก์ชั่นเอนไซม์มีสมาชิกโปรตีนตั้งแต่ 10 ถึง 1,000 ชนิด พบว่ามีความแม่นยำค่อนข้างต่ำมากที่ 37.15% อย่างไรก็ตาม เมื่อ

ได้ปรับแก้ให้เงื่อนไขละเอียดที่สุดที่สามารถใช้ 5-fold cross-validation ได้อย่างมีประสิทธิภาพคือที่แต่ละกลุ่มฟังก์ชั่นเอนไซม์มีสมาชิกโปรตีนตั้งแต่ 5 ถึง 1,000 ชนิด พบว่ามีความแม่นยำเพิ่มขึ้นใกล้เคียงกับการใช้รีแอคทีฟโมทีฟที่ 67.25% อย่างไรก็ตามเมื่อเปรียบเทียบเป็นขนาดกลุ่มข้อมูลที่โมทีฟแต่ละชนิดสามารถใช้งานได้สูงสุด พบว่ากลุ่มข้อมูลที่ใช้รีแอคทีฟโมทีฟมีขนาดใหญ่กว่าที่ใช้โมทีฟ PROSITE ถึง 7 เท่า อีกนัยหนึ่งรีแอคทีฟโมทีฟสามารถประยุกต์ใช้กับกลุ่มข้อมูลโปรตีนที่มีอยู่ในปัจจุบันได้มากกว่า PROSITE หลายเท่า

จากตารางที่ 8 และ 11 เมื่อวิเคราะห์ผลของคุณภาพโมทีฟที่มีต่อความแม่นยำในการทำนายประเภทฟังก์ชั่นเอนไซม์ ที่กลุ่มข้อมูลขนาด 25 ฟังก์ชั่นแสดงอย่างชัดเจนว่า PROSITE เป็นโมทีฟที่มีคุณภาพดีที่สุดในแง่ของ accuracy precision และ specificity แต่เนื่องจากค่า % coverage และ sensitivity ที่ต่ำทำให้โมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์ให้ผลความแม่นยำ (% Correctly Classified Instances) ต่ำที่สุดเมื่อเทียบกับรีแอคทีฟโมทีฟประเภทอื่น อย่างไรก็ตามแม้ว่ารีแอคทีฟโมทีฟที่ไม่มีกลุ่มแทนที่จะมีค่า %coverage และ sensitivity ที่สูง แต่จากการที่มีค่า precision ที่ต่ำมาก ทำให้การใช้งานรีแอคทีฟประเภทนี้ในการสร้างโมเดลทำนายประเภทฟังก์ชั่นเอนไซม์ให้ผลที่แม่นยำสู้รีแอคทีฟโมทีฟประเภทอื่นไม่ได้

เมื่อพิจารณาเปรียบเทียบคุณภาพของโมทีฟที่มีต่อความแม่นยำการทำนายประเภทฟังก์ชั่นเอนไซม์ที่กลุ่มข้อมูลขนาด 235 ฟังก์ชั่นในตารางที่ 8 และ 12 พบว่า ค่า specificity มีผลต่อค่า accuracy ของโมทีฟโดยรวม โดยค่า sensitivity ที่สูงของรีแอคทีฟโมทีฟที่ไม่มีกลุ่มแทนที่พบว่ามิใช่ปัจจัยที่ส่งผลให้ค่าความแม่นยำในการทำนายฟังก์ชั่นเอนไซม์สูงตามไปด้วย อันเนื่องมาจากการมีค่า precision ที่ต่ำมากอันเป็นเหตุผลเดียวกับการเปรียบเทียบที่ 3 และ 25 ฟังก์ชั่น อย่างไรก็ตาม ในกรณีที่ค่า precision ที่ต่ำมากในกรณีของรีแอคทีฟโมทีฟแสดงคุณสมบัติ BLOSUM ที่ 235 ฟังก์ชั่น ที่มีค่าเพียง 0.3383 แต่ให้ค่าความแม่นยำเมื่อใช้งานทำนายฟังก์ชั่นเอนไซม์ได้สูงถึง 68.69% ก็เนื่องจากการมีค่า %coverage ที่สูงถึง 96.55% ทำให้สามารถใช้ศักยภาพของ C4.5 ในการสร้างระบบทำนายประเภทฟังก์ชั่นเอนไซม์ได้อย่างแม่นยำ แม้ว่าโมทีฟชนิดนี้จะมีค่า precision ในการระบุบริเวณจับและบริเวณเร่งได้ต่ำกว่าตาม

สรุปและข้อเสนอแนะ

สรุปผลการวิจัย

การศึกษาวิจัยในวิทยานิพนธ์นี้มุ่งเน้นไปที่การพัฒนาตัวแทนสายโปรตีนสำหรับการทำนายประเภทฟังก์ชั่นของโปรตีน โดยกำหนดขอบเขตการศึกษาไปที่ฟังก์ชั่นเอนไซม์และพัฒนาตัวแทนสายโปรตีนประเภทโมทีฟจากบริเวณจับและบริเวณเร่งซึ่งเป็นบริเวณกลไกก่อฟังก์ชั่นเอนไซม์โดยตรง เพื่อใช้เป็นคุณลักษณะเด่นของกลุ่มข้อมูลนำเข้าสำหรับสร้างโมเดลการทำนายประเภทฟังก์ชั่นเอนไซม์ที่มีประสิทธิภาพดี

หัวข้อการวิจัยดังกล่าวนี้เป็นงานวิจัยด้านชีวสารสนเทศ จึงมีการวิจัยและนำเสนอเนื้อหางานวิจัยในทั้งด้านชีววิทยาและด้านวิศวกรรมคอมพิวเตอร์ โดยในเนื้อหาด้านชีววิทยามีดังนี้

1) การนำข้อมูลบริเวณจับและบริเวณเร่งมาใช้สร้างโมทีฟที่เรียกว่า รีแอคทีฟโมทีฟ

2) การนำความรู้พื้นหลังและทฤษฎีทางเอนไซม์สรุปเรียกเป็น “แนวคิดการควบคุมการกลายพันธุ์” มาใช้หากลุ่มแทนที่ที่สมบูรณ์ ซึ่งเป็นองค์ประกอบที่สำคัญของรีแอคทีฟโมทีฟ

3) การนำคำอธิบายบริเวณจับและบริเวณเร่ง (site description) มาใช้จัดกลุ่มรีแอคทีฟโมทีฟเพื่อใช้ในระบบทำนายประเภทฟังก์ชั่นเอนไซม์

สำหรับการนำเสนอเนื้อหาด้านวิศวกรรมคอมพิวเตอร์มีดังนี้

1) พัฒนาคุณลักษณะเด่นกลุ่มข้อมูลนำที่สร้างจากข้อมูลปริมาณน้อยสำหรับทำนายประเภทข้อมูลปริมาณมาก

- การพัฒนาตัวแทนสายโปรตีนที่เรียกว่า รีแอคทีฟโมทีฟจากข้อมูลบริเวณจับและบริเวณเร่ง เพื่อใช้ทำนายประเภทฟังก์ชั่นเอนไซม์ของสายโปรตีน

2) นำเสนอขั้นตอนวิธีในการขยายข้อมูลปริมาณน้อยด้วยวิธีการทางสถิติ

- การพัฒนาบลอกที่มีคุณภาพด้วยการนำเสนอขั้นตอนวิธีที่เรียกว่าการคัด สรรบลอกที่มีคุณภาพ และการจัดกลุ่มบริเวณปฏิกิริยาชีวเคมี

- สามารถใช้ขยายปริมาณของคุณลักษณะเด่นกลุ่มข้อมูลที่มีอยู่น้อยเพื่อใช้งานทำนายประเภทข้อมูลขนาดใหญ่ได้

3) ปรับปรุงขั้นตอนวิธีการได้มาซึ่งคุณลักษณะเด่นของกลุ่มข้อมูลที่มีคุณภาพ ที่สร้างจากข้อมูลปริมาณน้อย โดยใช้ความรู้พื้นหลังลักษณะต่างๆ

- นำเสนอการแปลง “ความสัมพันธ์เชิงวิทยาศาสตร์ระหว่างคุณลักษณะเด่นของกลุ่มข้อมูลกับประเภทของข้อมูล” ให้อยู่ในรูปฟอร์มของคอนเซ็ปต์ด้วย กรอบงานเชิงคอนเซ็ปต์

- นำเสนอการใช้ทฤษฎีคอนเซ็ปต์แลททิสในการนำความรู้พื้นหลังลักษณะต่างๆ เข้ามาประมวลผลเชิงคอนเซ็ปต์เพื่อให้ได้มาซึ่งคุณลักษณะเด่นของกลุ่มข้อมูลที่มีคุณภาพ

4) ขยายขอบเขตการใช้งานความรู้พื้นหลังที่มีอยู่ให้สามารถใช้งานได้กว้างขวางขึ้น

- การแปลงความรู้พื้นหลังจากรูปฟอร์มคอนเท็กซ์เป็นตารางคะแนนความเหมือน และการแปลงจากรูปฟอร์มตารางคะแนนความเหมือนเป็นคอนเท็กซ์

การพัฒนารีแอคทีฟโมทีฟจากบริเวณจับและบริเวณเร่งมีลักษณะปัญหาที่พิเศษคือข้อมูลบริเวณจับและบริเวณเร่งมีน้อยมากคือมีเพียงประมาณ 3.34% และแต่ละฟังก์ชั่นเอนไซม์มีการผสมผสานกันระหว่างบริเวณจับและบริเวณเร่งที่ซับซ้อนเพื่อทำงานฟังก์ชั่นเอนไซม์ในลักษณะ 1 ฟังก์ชั่นมีหลายบริเวณจับและบริเวณเร่ง และ 1 บริเวณจับหรือบริเวณเร่งมีได้ในหลายฟังก์ชั่น

เพื่อแก้ปัญหาดังกล่าว ในวิทยานิพนธ์นี้จึงได้นำเสนอการใช้เทคนิคคอนเซ็ปต์แลททิสที่มีการประยุกต์กรอบงานเชิงคอนเซ็ปต์เพื่อเชื่อมโยงความรู้ในหลายลักษณะเข้ามาพัฒนาคุณลักษณะเด่นของกลุ่มข้อมูลที่เรียกว่ารีแอคทีฟโมทีฟให้มีคุณภาพมากขึ้น นอกจากนี้ยังพัฒนาคุณภาพของบลอกที่ใช้ค้นพบรีแอคทีฟโมทีฟด้วยขั้นตอนวิธีใหม่ที่เรียกว่าการคัดสรรบลอกที่มีคุณภาพ และการ

จัดกลุ่มบริเวณปฏิกิริยาชีวเคมี ทำให้รีแอคทีฟโมทีฟสามารถใช้งานทำนายประเภทฟังก์ชั่นเอนไซม์ที่กลุ่มข้อมูลขนาดใหญ่ขึ้นมาก

ในการใช้งานรีแอคทีฟโมทีฟเพื่อทำนายประเภทฟังก์ชั่นเอนไซม์ด้วยเครื่องจักรเรียนรู้ C4.5 แบบ 5-fold cross validation ได้ผลความแม่นยำดีที่ประมาณ 70% ดีกว่าโมทีฟ PROSITE อีกทั้งสามารถประยุกต์ใช้กับกลุ่มข้อมูลที่ใหญ่กว่าหลายเท่า จากกลุ่มข้อมูลฝึกสอนที่มีข้อมูลบริเวณจับและบริเวณเร่งเพียง 5.8% ดังนั้นจึงมีความเป็นไปได้ที่จะนำขั้นตอนวิธีในงานวิทยานิพนธ์นี้ในการประยุกต์ใช้กับข้อมูลชีวสารสนเทศที่มีจำนวนน้อยอื่นๆ

รีแอคทีฟโมทีฟให้ผลแม่นยำดีที่สุดคือรีแอคทีฟโมทีฟที่ได้จากความรู้พื้นหลังประเภทคอนเท็กซ์คุณสมบัติเคมีฟิสิกส์ของกรดอะมิโนแม้ว่ารีแอคทีฟโมทีฟชนิดนี้จะให้ค่าความครอบคลุมหรือ sensitivity ที่ด้อยที่สุด แต่ให้ค่าเฉลี่ยโมทีฟต่อโปรตีนหนึ่งสายที่ค่อนข้างดี ซึ่งหมายถึงความเฉพาะเจาะจง (specific) ในการนำไปใช้งานโมทีฟเป็นตัวแทนบริเวณกลไกย่อยของฟังก์ชั่นเอนไซม์ที่ดีที่สุดเมื่อเทียบกับรีแอคทีฟชนิดอื่นที่ได้จากวิธีอัตโนมัตินี้

นอกจากนี้การใช้งานรีแอคทีฟโมทีฟในการทำนายฟังก์ชั่นเอนไซม์มีการเพิ่มค่าความแม่นยำสูงขึ้นมากเมื่อใช้วิธีการจัดกลุ่มบริเวณปฏิกิริยาชีวเคมีด้วยบริเวณอนุรักษ์ แสดงให้เห็นถึงการให้รายละเอียดบริเวณจับและบริเวณเร่งในฐานข้อมูลที่มีอยู่ในปัจจุบันมีความไม่สมบูรณ์เพียงพอและต้องการปรับปรุงคุณภาพให้ดีขึ้น

ข้อเสนอแนะแนวทางการพัฒนางานวิจัย

สำหรับงานวิจัยต่อเนื่องในอนาคตสามารถนำแต่ละส่วนของวิทยานิพนธ์มาพัฒนาต่อให้มีประสิทธิภาพมากขึ้นได้ดังนี้

1) พิจารณาจากศักยภาพของรีแอคทีฟโมทีฟที่สามารถประยุกต์ใช้กับข้อมูลบริเวณจับและบริเวณเร่งซึ่งมีน้อยมากได้ ดังนั้นจึงมีความเป็นไปได้ที่จะประยุกต์ใช้รีแอคทีฟโมทีฟกับข้อมูลสายโปรตีนในบริเวณอื่นที่มีน้อยได้ ทั้งนี้อาจเน้นไปที่บริเวณโคแฟคเตอร์ของฟังก์ชั่นเอนไซม์หรือบริเวณที่ก่อฟังก์ชั่นโปรตีนในลักษณะอื่น เป็นต้น

2) การพัฒนากรอบทฤษฎีคอนเซ็ปต์แลททิสให้ครอบคลุมหลากหลายกรณีมากกว่านี้ เช่น การพัฒนากรอบงานเชิงคอนเซ็ปต์ ซึ่งทำให้มองปัญหาด้านชีวสารสนเทศให้อยู่ในรูปของคอนเซ็ปต์และสามารถใช้งานคอนเซ็ปต์แลททิสได้ หรือพัฒนาแนวทางการประยุกต์ใช้คอนเซ็ปต์แลททิสด้วยขั้นตอนวิธีที่แตกต่างจากที่นำเสนอในงานวิจัยนี้ เป็นต้น

3) การพัฒนาความรู้พื้นหลังและการแปลงความรู้พื้นหลังให้มีประสิทธิภาพมากขึ้น เช่น วิธีแปลงความรู้พื้นหลังจากรูปฟอร์มหนึ่งไปยังรูปฟอร์มที่หลากหลายมากขึ้น นอกเหนือไปจากคอนเท็กซ์และตารางคะแนนความเหมือนที่มีอยู่ในงานวิจัยนี้ รวมถึงการพัฒนาขั้นตอนวิธีให้มีประสิทธิภาพในการนำไปประยุกต์ใช้งานมากขึ้น

4) การพัฒนากรอบคุณภาพในการได้มาซึ่งบลอกที่มีคุณภาพ เช่น การพัฒนาเกณฑ์ชี้วัดคุณภาพของบลอกขึ้นใหม่ เป็นต้น

5) อาจสามารถจัดข้อมูลให้อยู่ในโครงสร้างอื่นที่ไม่ใช่บลอก ซึ่งจะทำให้ประยุกต์ใช้งานขั้นตอนวิธีที่มีอยู่ได้กว้างขึ้น ทั้งนี้เนื่องจากบลอกเหมาะสมกับบริเวณกลุ่มสายข้อมูลที่ไม่มี gap ซึ่งเหมาะสมกับปัญหาประเภทฟังก์ชั่นเอนไซม์ในบริเวณจับและบริเวณเร่งเท่านั้น ดังนั้นถ้าหากต้องการประยุกต์ใช้งานขั้นตอนวิธีในงานวิจัยนี้ให้กว้างขวางขึ้น ย่อมจำเป็นต้องมีการพัฒนาโครงสร้างข้อมูลให้สามารถรองรับปัญหาด้านชีวสารสนเทศได้กว้างขวางกว่าโครงสร้างบลอกที่มีอยู่

เอกสารและสิ่งอ้างอิง

Appel, R.D., A. Bairoch and D.F. Hochstrasser. 1994. A New Generation of Information Retrieval Tools for Biologists : The Example of The EXPASY WWW Server. **Trends in biochemical sciences** 19 (6): 258-260.

Apweiler, R., A. Bairoch, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi และ L.L. Yeh. 2004. UniProt: the Universal Protein knowledgebase. **Nucleic Acids Research** 32: Database issue D115-D119.

Attasena, V. and K. Waiyamai. 2007. Discovering Motifs from Frequent Sequence Patterns in Protein Sequences, *In* Supan Tungjitkusolmun, ed. **The International Conference on Engineering, Applied Sciences, and Technology (ICEAST 2007)** . King Mongkut's Institute of Technology Ladkrabang (KMITL), Thailand.

Bairoch, A. 1991. PROSITE: a dictionary of sites and patterns in proteins. **Nucleic Acids Research** 19: 2241-2245.

_____ and R. Apweiler. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. **Nucleic Acids Research** 28: 45-48.

Barton, G.J. 1990. Protein multiple sequence alignment and flexible pattern matching. **Methods Enzymol.** (183): 403-428.

Bennett, S.P., L. Lu and D.L. Brutlag. 2003. 3MATRIX and 3MOTIF: a protein structure visualization system for conserved sequence. **Nucleic Acids Research** 31: 3328–3332.

Cleverdon, C.J., J. Mills and M. Keen. 1966. **Factors Determining the Performance of Indexing Systems, Volume I - Design, Volume II - Test Results.** ASLIB Cranfield Project, Cranfield.

Dayhoff, M.O., R.M. Schwartz and B.C. Orcutt. 1978. A model of evolutionary change in proteins, pp. 345-352. *In* M.O. Dayhoff, ed. **Atlas of Protein Sequence and Structure Vol. 5. Suppl. 3.** National Biomedical Research Foundation, Washington.

Diplaris, S., G. Tsoumakas, P.A. Mitkas and I. Vlahavas. 2005. Protein Classification with Multiple Algorithms, pp. 448-456. *In* P. Bozanis and E.N. Houstis, eds. **10th Panhellenic Conference on Informatics, PCI 2005,** Volas, Greece, November 11-13, 2005. Springer-Verlag, Berlin / Heidelberg.

Eidhammer, I., I. Jonassen and W.R. Taylor. 2000. Protein structure comparison and structure patterns. **Journal of Computational Biology** 7 (5): 685-716.

Feng, DF and RF Doolittle. 1987. Progressive Sequence Alignment as a Prerequisite to Correct Phylogenetic Trees. **Journal of Molecular Evolution** (25): 351-360.

Frank, E., M. Hall, L. Trigg, G Holmes and I.H. Witten. 2004. Data mining in bioinformatics using Weka. **Bioinformatics** 20 (15): 2479-2481.

Gabriela, A., B. Anne, H. Marianne, P. Guillaume and S. Alain. 2007. Performances of Galois Sub-hierarchy-building Algorithms, pp. 166-180. *In* S.O. Kuznetsov and S. Schmidt, eds. **5th International Conference, ICFCA 2007** . Springer-Verlag, Berlin / Heidelberg.

Hall, T.A. 1999. BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. **Nucleic Acids Symp. Ser.** (41): 95–98.

Han, J. and M. Kamber. 2001. **Data Mining Concepts and Techniques.** Morgan Kaufmann, USA.

Henikoff, S. and J.G. Henikoff. 1991. Automated assembly of protein blocks for database searching. **Nucleic Acids Research** 19: 6565–6572.

_____ and _____. 1992. Amino acid substitution matrices from protein blocks. **Proc. Natl. Acad. Sci USA** (89): 10915-10919.

Hoffman, M.M., M.A. Khrapov, J.C. Cox, J. Yao, L. Tong and A.D. Ellington. 2004. AANT: the Amino Acid-Nucleotide Interaction Database. **Nucleic Acids Research** 32: Database issue:D174-81.

Huang, J.Y. and D.L. Brutlag. 2001. The EMOTIF database. **Nucleic Acids Research** 29: 202–204.

Kohavi, R. and F. Provost. 1998. Glossary of Terms. **Machine Learning** (30): 271-274.

Lesk, A.M. 2004. **Introduction to Protein science.** University of Cambridge, OXFORD University Press Inc, Newyork.

Liewlom, P. 2008. **Using Concept Lattice – Based Mutation Control to Embed Enzyme Mechanism Representation to Reactive Motif for Enzyme Function Prediction.** (submitted to **Journal of Computational Biology**). Kasetsart University, Department of Computer Engineering, Thailand.

_____., T. Rakthanmanon and K. Waiyamai. 2007. Prediction of Enzyme Class using Reactive Motifs generated from Binding and Catalytic Sites, *In* R. Alhajj, H. Gao, S. Li, J. Li and O.R. Zaiane, eds. **The 3rd International Conference on Advanced Data Mining and Applications (ADMA 2007)** . Springer-Verlag, Berlin / Heidelberg.

Mallery, C. 2007. **Enzymology.** Enzymology. Available Source: http://fig.cox.miami.edu/~cmallery/255/255enz/enzymology.htm, February 15, 2008.

Mosteller, F. and J.W. Tukey. 1968. Data analysis, including statistics, *In* G. Lindzey and E. Aronson, eds. **Handbook of Social Psychology, Vol. 2.** Addison–Wesley,

Nomenclature Committee of the International Union of Biochemistry and Molecular Biology. 1992. **Enzyme Nomenclature. Recommendations 1992.** Academic Press,

Patrick, G.L. 1995. **An Introduction to Medical Chemistry.** Oxford University Press Inc., Newyork.

Quinlan, J.R. 1993. **C4.5: Programs for Machine Learning.** Morgan Kaufman,

Rattanakronkul, N., T. Wattarujeekrit and K. Waiyamai. 2003. Predicting Protein Structure Class from Closed Protein Sequences, *In* K.Y. Whang, J. Jeon, K. Shim and J Srivastava, eds. **Advances in Knowledge Discovery and Data Mining (PAKDD) 7. ed.** Springer-Verlag, Berlin Heidelberg.

Schomburg, I., A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn and D. Schomburg. 2004. Brenda, the enzyme database: updates and major new developments. **Nucleic Acids Research** 32: D431–D433.

Smith, H.O., T.M. Annau and S. Chandrasegaran. 1990. Finding sequence motifs in groups of functionally related proteins. **Proceedings of the National Academy of Sciences** 87 (2): 826–830.

Taylor, W.R. 1986. Identification of protein sequence homology by consensus template alignment. **J. Mol. Biol** (188): 233-258.

The Gene Ontology Consortium. 2000. Gene Ontology: tool for the unification of biology. **Nature Genet** (25): 25-29.

UNIPROT. 2005. **UniProt Release 4.0 Note 1-Feb-2005.** http://www.ebi.uniprot.org/support/docs/rel_notes/. Available Source: UNIPROT, February 15, 2005.

_____. 2007. **UniProt Release 12.8.** UNIPROT. Available Source: http://uniprot.org, February 10, 2007.

van Rijsbergen, C.J. 1979. **Information Retrieval.** 2 ed. Butterworths, London.

Waiyamai, K., R. Taouil and L. Lakhal. 1997. Towards an object database approach for managing concept lattices, *In* W.E. David and C.G. Robert, eds. **The 16th International Conference on Conceptual Modeling** . Springer-Verlag, Berlin / Heidelberg.

_____, P. Liewlom, T. Kangkachit and T. Rakthanmanon. 2008. Concept Lattice – Based Mutation Control for Reactive Motifs Discovery, *In* **The Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2008)** . Springer-Verlag,

Waterman, M.S., T.F. Smith and W.A. Beyer. 1976. Some biological sequence metrics. **Advances in Mathematics** (20): 367-387.

Weber, I.T., D.B. McKay and T.A. Steitz. 1982. Two helix DNA binding motif of CAP found in lac repressor and gal repressor. **Nucleic Acids Research** 10: 5085–5102.

Wikipedia. 2004. **Amino Acid.** Amino acid - Wikipedia, the free encyclopedia. Available Source: http://en.wikipedia.org/wiki/Amino_acid, December 2, 2004.

Wille, R. 1982. Restructuring lattice theory: an approach based on hierarchies of concepts, pp. 445-470. *In* I. Rival, ed. **Ordered sets.** Dordrecht–Boston,

_____. 1989. Knowledge acquisition by methods of formal concept analysis, pp. 365-380. *In* R. Diday, ed. **Data Analysis, Learning Symbolic and Numeric Knowledge.** Nova Science Publishers, Inc. Commack, Newyork, USA.

Witten, I. H. and E. Frank. 2005. **Data Mining: Practical machine learning tools and techniques.** 2 ed. Morgan Kaufmann, San Francisco.

Wu, T.D. and D.L. Brutlag. 1996. Discovering Empirically Conserved Amino Acid Substitution Groups in Databases of Protein Families. **Proc Int Conf Intell Syst Mol Biol.** (4): 230-240.

กฤษณะ ไวยมัย. 2549. **คลังข้อมูลและการทำเหมืองข้อมูล.** ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์, กรุงเทพ.

ธนภัทร ฆังคะจิตร, พีระ ลิ่วลม, ปนัดดา ปันสุวรรณ์ และ กฤษณะ ไวยมัย. 2548. ระบบทำนายประเภทฟังก์ชั่นโปรตีนจากลำดับกรดอะมิโน, น. 277-286. *ใน* จีรเดช อู่สวัสดิ์, บรรณาธิการ. **การประชุมวิชาการวิทยาการคอมพิวเตอร์และวิศวกรรมคอมพิวเตอร์แห่งชาติ ครั้งที่ 9.** ภาควิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยหอการค้าไทย, กรุงเทพ.

พีระ ลิ่วลม และ กฤษณะ ไวยมัย. 2549. การประยุกต์ใช้โมทีฟที่มีการกลายพันธุ์จากบริเวณจับและบริเวณเร่งสำหรับทำนายฟังก์ชั่นเอนไซม์, น. 561-571. *ใน* บุญส่ง วัฒนกิจ, บรรณาธิการ. **การประชุมวิชาการวิทยาการคอมพิวเตอร์และวิศวกรรมคอมพิวเตอร์แห่งชาติ ครั้งที่ 10.** ภาควิชาวิศวกรรมคอมพิวเตอร์ และภาควิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยขอนแก่น, ขอนแก่น.

สาวิณี แสงสุริยันต์, วรัญญา อรรถเสนา และ กฤษณะ ไวยมัย. 2549. การเพิ่มประสิทธิภาพโมทีฟโดยใช้โมทีฟคอนเซพท์, น. 651-658. *ใน* บุญส่ง วัฒนกิจ, บรรณาธิการ. **การประชุมวิชาการวิทยาการคอมพิวเตอร์และวิศวกรรมคอมพิวเตอร์แห่งชาติ ครั้งที่ 10.** ภาควิชาวิศวกรรมคอมพิวเตอร์ และภาควิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยขอนแก่น, ขอนแก่น.

ภาคผนวก

ประวัติการศึกษา และการทำงาน

ชื่อ	นายพีระ ลิ่วลม
เกิดวันที่	วันที่ 9 ธันวาคม 2513
สถานที่เกิด	กรุงเทพมหานคร
ประวัติการศึกษา	วิทยาศาสตรบัณฑิต (เคมี), มหาวิทยาลัยมหิดล
	วิทยาศาสตรมหาบัณฑิต (เทคโนโลยีการจัดการสารสนเทศ), มหาวิทยาลัยมหิดล
ตำแหน่งปัจจุบัน	อาจารย์
สถานที่ทำงานปัจจุบัน	คณะวิทยาศาสตร์และวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตเฉลิมพระเกียรติจังหวัดสกลนคร