ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

วิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)

ปริญญา

วิศวกรรมคอมพิวเตอร์ วิศวกรรมคอมพิวเตอร์

สาขา **ภาควิชา**

เรื่อง การจำแนกซับแฟมิลีและวิเคราะห์โครงสร้างของเอนไซม์ ด้วยวิธีการพิจารณาการติดกันของคู่ลำดับกรดอะมิโนและโพรไฟล์ฮิดเด้นมาร์คอฟโมเดล

Enzyme Subfamily Classification and Structure Analysis Using Residue Contact and Profile HMMs

นามผู้วิจัย นางสาวสิริวรรณ แต้วิจิตร

ได้พิจารณาเห็นชอบโดย

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(รองศาสตราจารย์กฤษณะ ไวยมัย, Ph.D.)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม

(รองศาสตราจารย์พันธุ์ปิติ เปี่ยมสง่า, Ph.D.)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม

(อาจารย์สุภาวดี อิงศรีสว่าง, Ph.D.)

หัวหน้าภาควิชา

(ผู้ช่วยศาสตราจารย์เขมะฑัต วิภาตะวนิช, Ph.D.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

(รองศาสตราจารย์กัญจนา ธีระกุล, D.Agr.)

คณบดีบัณฑิตวิทยาลัย

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

วิทยานิพนธ์

เรื่อง

การจำแนกซับแฟมิลีและวิเคราะห์โครงสร้างของเอนไซม์ ด้วยวิธีการพิจารณาการติดกันของคู่ลำดับกรดอะมิโนและโพรไฟล์ฮิดเด้นมาร์คอฟโมเดล

Enzyme Subfamily Classification and Structure Analysis Using Residue Contact and Profile HMMs

โดย

นางสาวสิริวรรณ แต้วิจิตร

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์
เพื่อความสมบูรณ์แห่งปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)
พ.ศ. 2552

สิริวรรณ แต้วิจิตร 2552: การจำแนกซับแฟมิลีและวิเคราะห์โครงสร้างของเอนไซม์ด้วยวิธีการพิจารณาการติดกันของคู่ลำดับกรดอะมิโนและโพรไฟล์ฮิดเด้นมาร์คอฟโมเดล ปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์) สาขาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก: รองศาสตราจารย์กฤษณะ ไวยมัย, Ph.D. 111 หน้า

การค้นพบเอนไซม์ใหม่มีอัตราการค้นพบที่สูงขึ้นอย่างรวดเร็ว เพื่อลดเวลาและค่าใช้จ่ายในการจำแนกประเภทเอนไซม์ จึงมีการพัฒนางานวิจัยเพื่อสร้างระบบอัตโนมัติด้วยวิธีการต่างๆ ได้แก่ (1) วิธีการเปรียบเทียบเอนไซม์ใหม่และเอนไซม์ต้นแบบที่ทราบฟังก์ชั่นแล้ว วิธีดังกล่าวให้ความถูกต้องสูงเมื่อเอนไซม์ใหม่มีค่าความเหมือนกับเอนไซม์ต้นแบบมากกว่า 60% (2) วิธีการพิจารณาส่วนย่อยของลำดับกรดอะมิโน ซึ่งสามารถบ่งบอกฟังก์ชั่นได้ (3) วิธีการสร้างฟีเจอร์เพื่อเป็นตัวแทนเอนไซม์ เช่นพิจารณาความถี่ในการเกิดกรดอะมิโน วิธีดังกล่าวให้ความถูกต้องได้ดี แต่กระนั้นผลลัพธ์ที่ได้ไม่สามารถแสดงความสัมพันธ์เกี่ยวกับโครงสร้างของเอนไซม์ใหม่ได้

ในงานวิจัยนี้ทำการพัฒนาระบบจำแนกประเภทเอนไซม์ซับแฟมิลีที่ให้ความถูกต้องสูงและให้ผลลัพธ์ที่สามารถนำไปวิเคราะห์โครงสร้างเอนไซม์ใหม่ ในรูปแบบการติดกันของโครงสร้างทุติยภูมิ โดยมีสมมติฐานที่ว่าส่วนย่อยของโครงสร้างเอนไซม์สามารถบ่งบอกหน้าที่การทำงานของเอนไซม์ได้ สำหรับส่วนย่อยของโครงสร้างนั้นสืบค้นจากตารางเมตริกซ์คอนแทคแมป ซึ่งเป็นการแปลงข้อมูลโครงสร้าง 3 มิติให้อยู่ในรูปตารางเมตริกซ์ที่เข้าใจง่าย และนำแต่ละส่วนย่อยมาประกอบกันเป็นตัวแทนสายโปรตีนด้วยโพรไฟล์ฮิดเด้นมาร์คอฟโมเดล ดังนั้นข้อมูลที่ใช้ในการสกัดฟีเจอร์จำเป็นต้องทราบโครงสร้าง 3 มิติแล้วเท่านั้น คิดเป็นสัดส่วนเพียง 11.21% ของข้อมูลทั้งหมดในงานวิจัยนี้ และสร้างแบบจำลองด้วยซัพพอร์ตเวคเตอร์แมชชีน จากผลการทดลองพบว่าฟีเจอร์จากตารางเมตริกซ์คอนแทคแมป มีความสัมพันธ์และมีนัยสำคัญเพียงพอในการจำแนกประเภทเอนไซม์ โดยแบบจำลองให้ค่าความถูกต้องถึง 73.71% เมื่อทดสอบด้วยวิธี jackknife และเมื่อเพิ่มองค์ความรู้เกี่ยวกับการแทนที่ระหว่างกรดอะมิโน สามารถเพิ่มความถูกต้องได้ถึง 75.87% เมื่อทดสอบด้วยวิธี 10-fold cross validation

_____ _____ ___ / ___ / ___

ลายมือชื่อนิสิต ลายมือชื่ออาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

Siriwon Taewijit 2009: Enzyme Subfamily Classification and Structure Analysis Using Residue Contact and Profile HMMs. Master of Engineering (Computer Engineering), Major Field: Computer Engineering, Department of Computer Engineering. Thesis Advisor: Associate Professor Kitsana Waiyamai, Ph.D. 111 pages.

The number of enzyme sequence grows exponentially over time. Identifying of class for a new found enzyme is usually performed by experiments in laboratory. Unfortunately, it is both time-consuming and costly. Large number of researches on enzyme classification have been carried out for decade such as (1) homology-based, this approach got hight accuracy when similarity score between target and template is more than 60%. (2) subsequence-based, this approach aims to find useful portions that related to enzyme function. (3) feature-based, this approach tries to represent enzyme with the new form. Their results were able to generate good accuracy but they cannot exhibit relationships between enzyme and its structure.

In this research, we propose an enzyme subfamily classification method using residue contact and profile HMMs. Our hypothesis is based on the fact that enzyme function is directly related to its structure. First, each fragment from contact map matrix which is representative of 3D structure is extracted. Then, profile HMMs is used to build feature vector to represent each enzyme sequence. Notice that only enzymes with known PDB (11.21% of the trainning data) have been used to extract features. Finally, SVM was used to build a classifier. Our method yields high accuracy and provides structural description based on secondary structure contact on 3D domain. Empirical results show that our method yields 73.71% accuracy with jackknife test and the accuracy rises up to 75.87% when substitution group of amino acids is used for backround knowledge. These suggest that our feature extraction from contact map and Profile HMMs are sufficiently significant for enzyme subfamily classification.

_____ _____ ___ / ___ / ___

Student's signature Thesis Advisor's signature

กิตติกรรมประกาศ

ข้าพเจ้าขอกราบขอบพระคุณ รองศาสตราจารย์ ดร.กฤษณะ ไวยมัย ประธานกรรมการที่ปรึกษาวิทยานิพนธ์หลักที่ได้ประสิทธิ์ประสาทวิชาความรู้และทฤษฎีต่าง ๆ ตลอดจนเป็นที่ปรึกษาในการวางแนวทางวางแผนงาน แก้ปัญหา และตรวจสอบข้อบกพร่องต่าง ๆ จนงานวิจัยนี้สำเร็จลุล่วงด้วยดี

ขอกราบขอบพระคุณ รองศาสตราจารย์ ดร.พันธุ์ปิติ เปี่ยมสง่า ดร.สุภาวดี อิงศรีสว่าง กรรมการที่ปรึกษาวิทยานิพนธ์ร่วม รวมทั้งผู้ช่วยศาสตราจารย์ ดร.พีระวัฒน์ วัฒนพงศ์ ดร.สรรพฤทธิ์ มฤคทัต และคณาจารย์ภาควิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเกษตรศาสตร์ทุกท่าน ที่กรุณาให้ความรู้ ข้อเสนอแนะ คำปรึกษา คำแนะนำ เปิดโลกทัศน์และมุมมองต่าง ๆ เกี่ยวกับงานวิจัย

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

ขอขอบคุณเจ้าหน้าที่โครงการบัณฑิตศึกษา และเจ้าหน้าที่ธุรการภาควิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเกษตรศาสตร์ที่ช่วยเหลือในการประสานงาน และดำเนินงานด้านเอกสารต่างๆ ให้เป็นไปอย่างสะดวกลุล่วงไปด้วยดี

คุณงามความดี หรือประโยชน์อันใดเนื่องมาจากวิทยานิพนธ์ฉบับนี้ ขออุทิศแด่มารดา บิดา บุพการี และผู้มีพระคุณทุกท่าน

สิริวรรณ แต้วิจิตร

เมษายน 2552

สารบัญ

สารบัญตาราง

สารบัญภาพ

สารบัญภาพ (ต่อ)

สารบัญภาพ (ต่อ)

สารบัญภาพ (ต่อ)

คำอธิบายสัญลักษณ์และคำย่อ

Å	=	Angstrom
BLOSUM	=	Blocks of Amino Acid Substitution Matrix
CD	=	Covariant-discriminant predictor
CDA	=	Covariant-discriminant using Amphiphilic Pseudo Amino Acid Composition
HMMs	=	Hidden Markov Models
KEGG	=	Kyoto Encyclopedia of Genes and Genomes
NCBI	=	National Center for Biotechnology Information
NMR	=	Nuclear Magnetic Resonance
PDB	=	Protein Data Bank
PROSITE	=	Database of protein domains, families and functional site
SCOP	=	Structural Classification of Proteins
SOM	=	Self Organizing Map
SSCP	=	Sub-Structural Contacted Pattern
SVM	=	Support Vector Machine
UniProt	=	The Universal Protein Resource
WEKA	=	Waikato Environment for Knowledge Analysis

การจำแนกซับแฟมิลีและวิเคราะห์โครงสร้างของเอนไซม์ ด้วยวิธีการพิจารณาการติดกันของคู่ลำดับกรดอะมิโนและโพรไฟล์ฮิดเด้นมาร์คอฟโมเดล

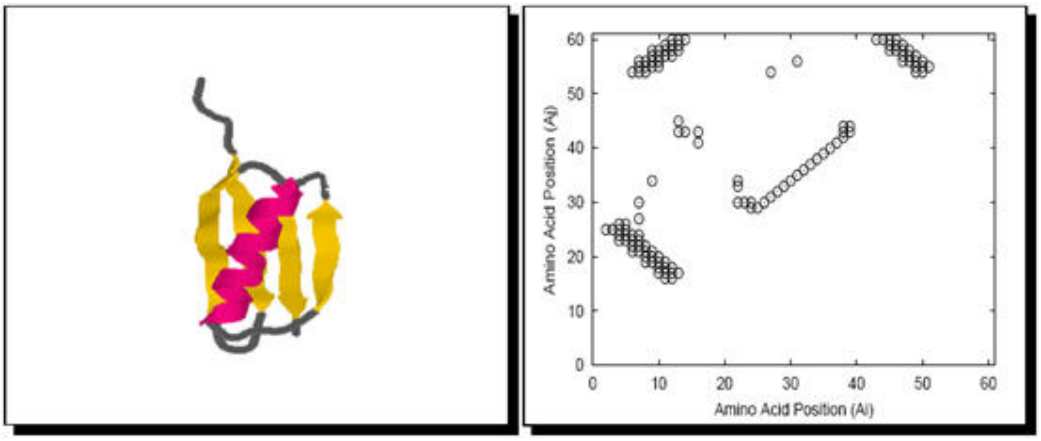
Enzyme Subfamily Classification and Structure Analysis Using Residue Contact and Profile HMMs

คำนำ

เอนไซม์คือโมเลกุลโปรตีนสำคัญที่ช่วยทำให้เซลล์ของอวัยวะต่างๆ ทำงานได้อย่างปกติ โดยเอนไซม์ทำหน้าที่ในการเร่งปฏิกิริยาเคมีภายในเซลล์สิ่งมีชีวิต หากขาดเอนไซม์อาจทำให้ผลิตภัณฑ์จากปฏิกิริยาเคมีกลายเป็นสารเคมีชนิดอื่น ส่งผลให้การทำงานของเซลล์ผิดปรกติได้ ปัจจุบันมีการค้นพบเอนไซม์ชนิดใหม่เพิ่มขึ้นอย่างรวดเร็วและเป็นปริมาณมาก การจำแนกประเภทเอนไซม์โดยวิธีการวิเคราะห์และทดสอบในห้องปฏิบัติการอาจใช้เวลานานและเสียค่าใช้จ่ายสูง ดังนั้นจึงมีการพัฒนาระบบอัตโนมัติเพื่อทำการจำแนกประเภทเอนไซม์ จากอดีตจนถึงปัจจุบันมีงานวิจัยมากมายเกี่ยวกับการจำแนกประเภทเอนไซม์โดยสามารถแบ่งกลุ่มตามวิธีการวิจัยได้ 3 กลุ่ม ดังนี้ (1) homology-based เป็นวิธีการเปรียบเทียบเอนไซม์ใหม่และเอนไซม์ต้นแบบที่ทราบฟังก์ชั่นแล้ว วิธีการดังกล่าวให้ความถูกต้องสูงเมื่อเอนไซม์ใหม่มีค่าความเหมือนกับเอนไซม์ต้นแบบมากกว่า 60% (2) subsequence-based เป็นวิธีการพิจารณาส่วนย่อยของลำดับกรดอะมิโน ซึ่งบริเวณส่วนย่อยนี้สามารถบ่งบอกฟังก์ชั่นของเอนไซม์ได้ หรือบริเวณที่เรียกว่า โมทีฟ หรือ โดเมน เป็นต้น (3) feature-based เป็นวิธีการสร้างฟีเจอร์เพื่อใช้เป็นตัวแทนเอนไซม์ เช่นพิจารณาความถี่ในการปรากฏกรดอะมิโนหนึ่งๆ (amino acid composition) หรือการปรากฏกลุ่มคุณสมบัติทางเคมีของอะมิโนบนสายโปรตีน วิธีทั้งหมดดังกล่าวข้างต้นให้ความถูกต้องได้ดีระดับหนึ่ง แต่กระนั้นผลลัพธ์ที่ได้ไม่สามารถแสดงความสัมพันธ์เกี่ยวกับโครงสร้างของเอนไซม์ใหม่ได้

ในงานวิจัยนี้มุ่งเน้นพัฒนาระบบจำแนกประเภทเอนไซม์ซับแฟมิลีที่ให้ความถูกต้องสูง โดยในงานวิจัยนี้สนใจเอนไซม์ในกลุ่มออกซิโดรีดักเทส (Oxidoreductases) ข้อมูลที่ใช้ทดสอบระบบนั้นมีค่าความเหมือนของสายโปรตีนไม่เกิน 30% นอกจากนี้ผลลัพทธ์ที่ได้จากระบบยังให้ข้อมูลที่สามารถนำไปวิเคราะห์โครงสร้างของโปรตีนได้ โดยอยู่ในรูปของการติดกันของโครงสร้างทุติยภูมิ จากงานวิจัยที่ผ่านมาพบว่าฟังก์ชั่นของเอนไซม์มีความสัมพันธ์เชิงวิวัฒนาการ

กับโครงสร้าง 3 มิติหรือโครงสร้างตติยภูมิอย่างใกล้ชิด (Gu *et al.*, 2008; Zhang *et al.*, 2008) กล่าวคือเอนไซม์ที่ถูกจัดอยู่ในประเภทเดียวกันนั้น พบว่ามีโครงสร้าง 3 มิติคล้ายคลึงกัน ซึ่งส่งผลให้เอนไซม์นั้นมีคุณสมบัติและหน้าที่การทำงานบางอย่างเหมือนกัน โดยงานวิจัยนี้มีสมมติฐานว่าส่วนย่อยของโครงสร้างเอนไซม์สามารถบ่งบอกหน้าที่การทำงานของเอนไซม์ได้ สำหรับส่วนย่อยของโครงสร้างนั้นสืบค้นจากตารางเมตริกซ์คอนแทคแมป (contact map matrix) ซึ่งเป็นการแปลงข้อมูลโครงสร้าง 3 มิติให้อยู่ในรูปตารางเมตริกซ์ที่เข้าใจง่ายและมีนัยสำคัญ แทนการพิจารณาข้อมูล 3 มิติของเอนไซม์โดยตรง เนื่องจากมีความซับซ้อนสูง โดยตารางเมตริกซ์คอนแทคแมปนั้นเป็นตารางแสดงการพิจารณาระยะห่างระหว่างคู่กรดอะมิโนใดๆ ในสายโปรตีนบนโดเมน 3 มิติ ดังนั้นรูปร่างที่ปรากฏบนตารางเมตริกซ์คอนแทคแมปนั้น แสดงถึงการวางตัวอยู่ติดกันของโครงสร้าง 2 มิติหรือโครงสร้างทุติยภูมิ (secondary structure ได้แก่ α-helix, β-sheet และ coil) ของโปรตีนบนโดเมน 3 มิตินั่นเอง (ภาพที่ 1) มีงานวิจัยมากมายใช้ประโยชน์จากตารางเมตริกซ์คอนแทคแมปในการเปรียบเทียบความคล้ายคลึงระหว่างโปรตีน (Gonzalez *et al.*, 2007; Pelta *et al.*, 2005 เป็นต้น) นอกจากนี้เรายังสามารถแปลงตำแหน่งต่างๆ บนตารางเมตริกซ์คอนแทคแมปกลับไปเป็นตำแหน่งโครงสร้างโปรตีนบนโดเมน 3 มิติได้อีกด้วย (Reconstruction of 3D structure) (Vassura *et al.*, 2008)



ภาพที่ 1 โครงสร้าง 3 มิติ (ภาพซ้าย) ของโปรตีน 2igd (PDB code) และตารางเมตริกซ์คอนแทคแมป (ภาพขวา) จากภาพเราสามารถพบรูปแบบโครงสร้าง 3 มิติได้ดังนี้ กลุ่มมุมบนขวาและมุมล่างซ้ายคือ anti parallel β-sheet, กลุ่มมุมบนซ้ายคือ parallel β-sheet และกลุ่มที่อยู่บริเวณแนวทแยงมุมคือ α-helix

ที่มา: Hu *et al.* (2002)

จากนั้นนำแต่ละส่วนย่อยมาประกอบกันเป็นตัวแทนสายโปรตีนด้วยโพรไฟล์ฮิดเด้นมาร์คอฟโมเดล ดังนั้นข้อมูลที่ใช้ในการสกัดฟีเจอร์จำเป็นต้องทราบโครงสร้าง 3 มิติแล้วเท่านั้น ซึ่งมีอยู่ปริมาณน้อยมากคิดเป็นสัดส่วนเพียง 11.21% ของข้อมูลทั้งหมดที่ใช้ในงานวิจัยนี้ โดยข้อมูลปฐมภูมิของเอนไซม์ได้มาจากฐานข้อมูลสวิสพรอท (SWISS-PROT) (Boeckmann *et al.*, 2003) ส่วนข้อมูลตติยภูมิของเอนไซม์ได้จากฐานข้อมูลพีดีบี (Protein Data Bank, PDB) (Berman *et al.*, 2000) สุดท้ายสร้างแบบจำลองด้วยอัลกอริทึมเรียนรู้ ซัพพอร์ตเวคเตอร์แมชชีน (SVM) จากผลการทดลองพบว่าฟีเจอร์จากตารางเมตริกซ์คอนแทคแมป มีความสัมพันธ์และมีนัยสำคัญเพียงพอในการจำแนกประเภทเอนไซม์ โดยแบบจำลองให้ค่าความถูกต้องถึง 73.71% เมื่อทดสอบด้วยวิธี jackknife และเมื่อเพิ่มองค์ความรู้เกี่ยวกับการแทนที่ระหว่างกรดอะมิโน สามารถเพิ่มความถูกต้องได้ถึง 75.87% เมื่อทดสอบด้วยวิธี 10-fold cross validation นอกจากนี้ผลลัพธ์ที่ได้จากระบบยังสามารถนำไปเป็นข้อมูลในการวิเคราะห์โครงสร้างของเอนไซม์ได้ โดยแสดงในรูปแบบของการติดกันของโครงสร้างทุติยภูมิบนโดเมน 3 มิติ ซึ่งเป็นข้อมูลที่มีประโยชน์ดังที่กล่าวแล้วข้างต้น

วัตถุประสงค์

1. ศึกษาความสัมพันธ์ระหว่างลำดับอะมิโนสายโปรตีนและโครงสร้างทุติยภูมิที่พบจากตารางเมตริกซ์คอนแทคแมป (contact map matrix)

2. พัฒนาวิธีการสร้างฟีเจอร์เพื่อใช้เป็นตัวแทนสายข้อมูลโปรตีน จากการสืบค้นรูปแบบที่ปรากฏบนตารางเมตริกซ์คอนแทคแมปที่ติดกันและต่อเนื่องมาที่สุด และโพรไฟล์ฮิดเด้นมาร์คอฟโมเดล

3. พัฒนาระบบในการจำแนกข้อมูลเอนไซม์ซับแฟมิลีได้อย่างมีประสิทธิภาพเหมาะสมและให้ความถูกต้องแม่นยำสูง โดยใช้ข้อมูลที่มีอยู่จำกัดเพียง 11.21% ของข้อมูลฝึกสอนระบบมาใช้ในการสร้างฟีเจอร์ตัวแทนสายโปรตีน

4. สืบค้นโครงสร้างทุติยภูมิบนตารางเมตริกซ์คอนแทคแมป สำหรับโปรตีนที่ไม่ทราบโครงสร้าง 3 มิติ

ขั้นตอนการวิจัย

1. ศึกษางานวิจัยและทฤษฎีต่างๆ ที่เกี่ยวข้องในการสร้างระบบจำแนกประเภทข้อมูลเอนไซม์ การสร้างตารางเมตริกซ์คอนแทคแมป โพรไฟล์ฮิดเด้นมาร์คอฟโมเดล อัลกอริทึมการเรียนรู้ รวมทั้งวิเคราะห์ข้อดีข้อด้วยของงานวิจัยที่ผ่านมา เพื่อที่จะนำความรู้ที่ได้มาใช้ในการวิจัยนี้

2. ออกแบบแนวทางการวิจัย วิธีการพัฒนาระบบ และวิธีการวัดผลการทดลอง

3. รวบรวมข้อมูลจากแหล่งข้อมูลมาตรฐาน ซึ่งในงานวิจัยนี้ใช้ข้อมูลจาก 2 แหล่งข้อมูล ได้แก่ ข้อมูลเอนไซม์นำมาจากฐานข้อมูลสวิสพรอท ส่วนข้อมูลโครงสร้าง 3 มิติที่จะนำมาใช้ในการสร้างตารางเมตริกซ์คอนแทคแมปนั้นนำมาจากฐานข้อมูลพีดีบี โดยข้อมูลทั้งหมดจะถูกนำมาใช้ในขั้นตอนการฝึกสอนและทดสอบระบบ

4. พัฒนาระบบจำแนกประเภทเอนไซม์

5. ทดสอบและวัดผลของระบบจำแนกประเภทเอนไซม์ที่พัฒนาขึ้น

6. วิเคราะห์ผลและสรุปผลการวิจัยและประโยชน์ที่ได้รับ

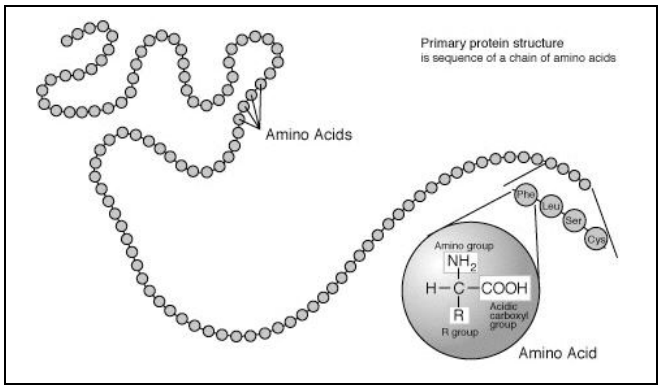
การตรวจเอกสาร

ในบทนี้กล่าวถึงความรู้พื้นฐานและงานวิจัยที่เกี่ยวข้องกับงานวิจัยนี้ โดยมีหัวข้อต่าง ๆ ได้แก่ ความรู้เบื้องต้นเกี่ยวกับโปรตีน ฐานข้อมูลโปรตีน ความรู้เกี่ยวกับการติดกันของคู่ลำดับอะมิโน (residue contact) วิธีการสร้างตารางเมตริกซ์คอนแทคแมป (contact map matrix) วิธีการสืบค้นรูปแบบย่อยที่ปรากฏบนตารางคอนแทคแมป การเทียบเรียงกลุ่มของสายโปรตีน (Multiple Sequence Alignment, MSA) และเทคนิคการแบ่งกลุ่มข้อมูล รวมทั้งอัลกอลิทึมพื้นฐานที่ใช้ในงานวิจัย ได้แก่ วิธีการเรียนรู้ ซัพพอร์ตเวคเตอร์แมชชีน (Support Vector Machine, SVM) และโพรไฟล์ฮิดเด้นมาร์คอฟโมเดล (Profile Hidden Markov Models, Profile HMMs)

1. ความรู้เบื้องต้นเกี่ยวกับโปรตีนและเอนไซม์

1.1 โปรตีน (Protein)

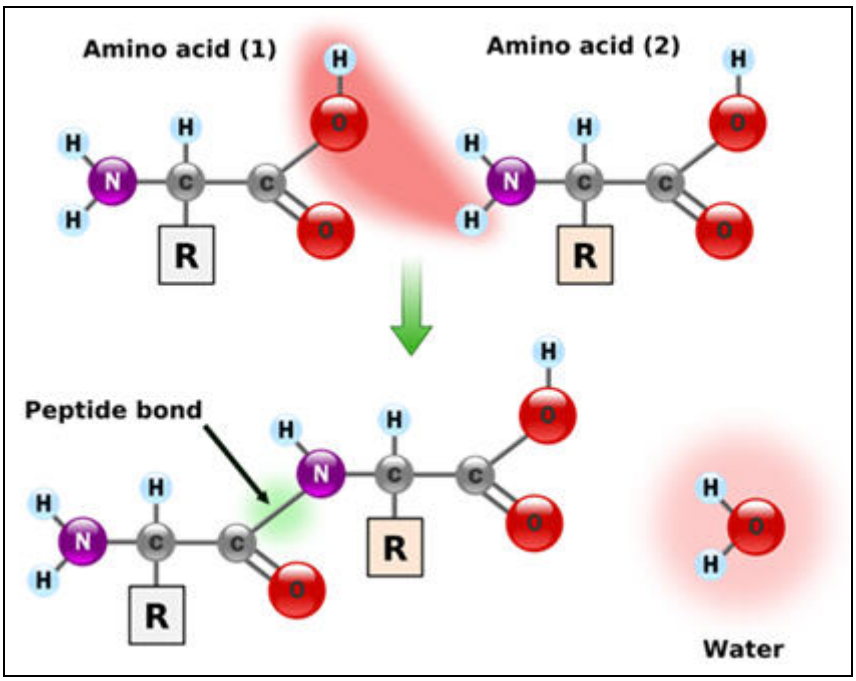
โปรตีนคือสารประกอบอินทรีย์ที่เกิดจากการเรียงตัวของกรดอะมิโนซึ่งเชื่อมต่อกันด้วยพันธะเปปไทด์ เรียกว่าสายโพลีเปปไทด์ (polypeptide) ดังภาพที่ 2 โปรตีนถูกพบครั้งแรกโดย Jöns Jakob Berzelius ในปี ค.ศ. 1838 โปรตีนจัดเป็นสารโมเลกุลขนาดใหญ่ เป็นส่วนประกอบสำคัญของโครงสร้างและกิจกรรมภายในเซลล์สิ่งมีชีวิตทุกชนิด



ภาพที่ 2 แสดงสายโปรตีนที่เกิดจากการเรียงตัวของกรดอะมิโนหลายชนิดต่อกัน

ที่มา: Wikipedia (2009)

ส่วนประกอบย่อยของสายโพลีเปปไทด์ คือกรดอะมิโน (amino acid) มีสูตรโครงสร้างเป็น NH2-CHR-COOH ดังภาพที่ 3 ในสูตรโครงสร้างนี้อะตอมคาร์บอนในตำแหน่งอัลฟา (α-Carbon) จะเป็นอะตอมคาร์บอนไม่สมมาตร (asymmetric carbon atom) ดังนั้นกรดอะมิโนทุกตัว (นอกจากไกลซีน ซึ่งมี R เป็นไฮโดรเจน) จะมีสเตริโอไอโซเมอร์ได้ สองชนิด คือ D- และ L- โดยที่กรดอะมิโนที่พบในธรรมชาติส่วนใหญ่เป็นชนิด L-



ภาพที่ 3 แสดงการเชื่อมต่อระหว่างกรดอะมิโน 2 ชนิด ด้วยพันธะเปปไทด์

ที่มา: Wikipedia (2009)

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

ตารางที่ 1 ตารางแสดงชื่อและสัญลักษณ์อักษรย่อ 3 ตัว และ 1 ตัว ของกรดอะมิโนมาตรฐาน 20 ชนิดที่พบในธรรมชาติ

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

ที่มา: Wikipedia (2009)

ข้อมูลโครงสร้างโปรตีนที่นักชีวเคมีให้ความสนใจมีอยู่ด้วยกัน 4 รูปแบบได้แก่ โครงสร้างปฐมภูมิ (primary structure) โครงสร้างทุติยภูมิ (secondary structure) โครงสร้างตติยภูมิ (tertiary structure) และโครงสร้างจตุรภูมิ (quaternary structure) แสดงดังภาพที่ 4, 5, 6 และ 7 ตามลำดับ สำหรับงานวิจัยนี้สนใจเฉพาะข้อมูลโครงสร้างปฐมภูมิและตติยภูมิเท่านั้น

โครงสร้างปฐมภูมิหรือสายลำดับกรดอะมิโน แสดงลำดับการเรียงตัวของกรดอะมิโนในสายโปรตีน (เป็นข้อมูลที่พบมากที่สุดในฐานข้อมูลโปรตีน)

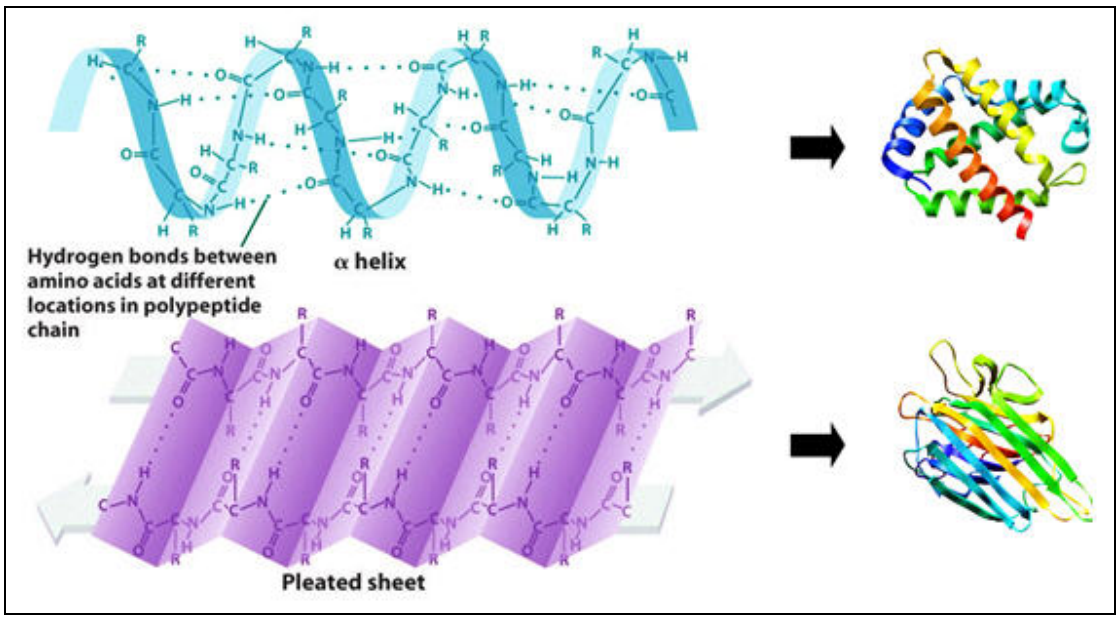
โครงสร้างทุติยภูมิหรือโครงสร้าง 2 มิติ แสดงการพันหรือม้วนตัว (fold) ของกรดอะมิโนที่อยู่ใกล้เคียงกันเพียงบางส่วนของสายโพลีเปปไทด์ เป็นรูปแบบเฉพาะอยู่ 2 ลักษณะคือ อัลฟา เฮลิกซ์ (alpha helix) สายเปปไทด์มีลักษณะขดเป็นเกลียว และเบต้า ชีท (beta sheet) หรือ เบต้า-พลีท ชีท (beta-pleated sheet) สายเปปไทด์มีลักษณะซิกแซกหรือพับเป็นแผ่น

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

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

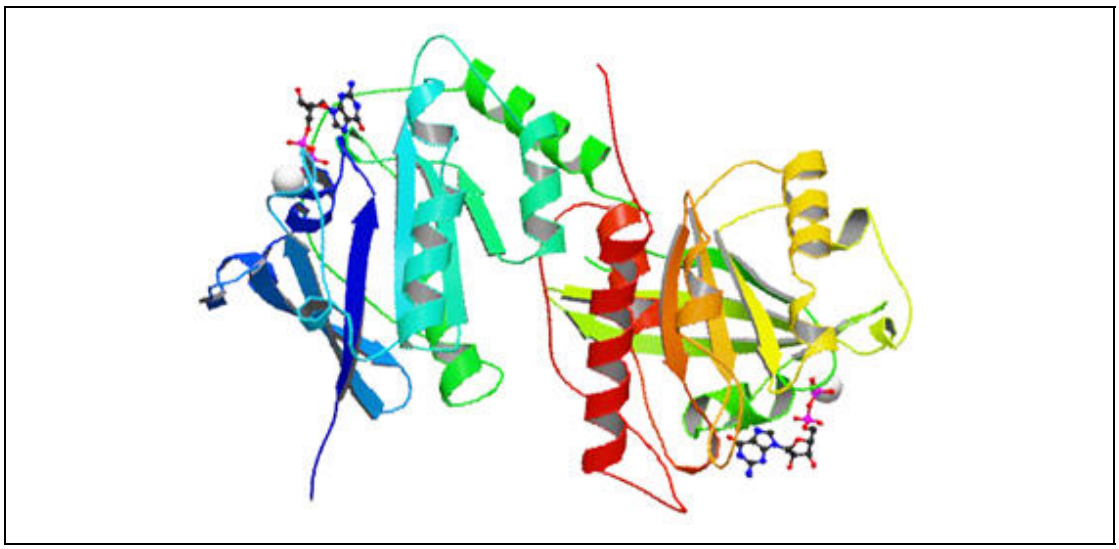
SFTLTNKNVIFVAGLGGIGLDTSKELLKRDLKNLVILDRIENPAAIAELKAINPKVTVTF
YPYDVTVPIAETTKLLKTIFAQLKTVDVLINGAGILDDHQIERTIAVNYTGLVNTTTAIL
DFWDKRKGGPGGIICNIGSVTGFNAIYQVPVYSGTKAAVVNFTSSLAKLAPITGVTAYTV
NPGITRTTLVHKFNSWLDVEPQVAEKLLAHPTQPSLACAENFVKAIELNQNGAIWKLDLG
TLEAIQWTKHWDSGI

ภาพที่ 4 โครงสร้างปฐมภูมิของโปรตีน Immunoglobulin-binding protein G



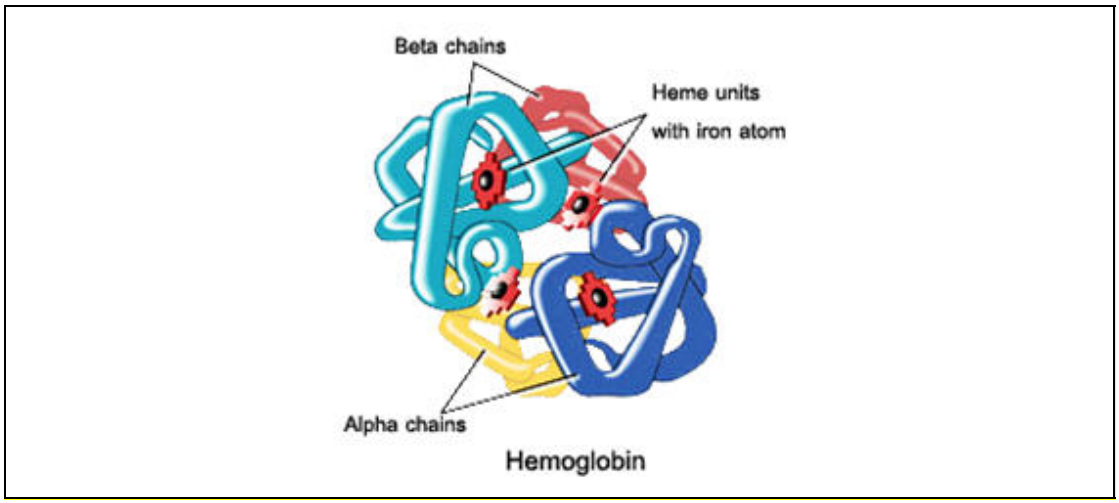
ภาพที่ 5 โครงสร้างทุติยภูมิ จากภาพคือ Alpha-helix (ซ้าย) และ Beta sheet หรือ Beta-pleated sheet (ขวา)

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



ภาพที่ 6 โครงสร้างตติยภูมิของโปรตีนแคไนน์ จีดีพี-แรน (CANINE GDP-RAN, PDB id: 1BYU)

ที่มา: Protein data bank (2009)

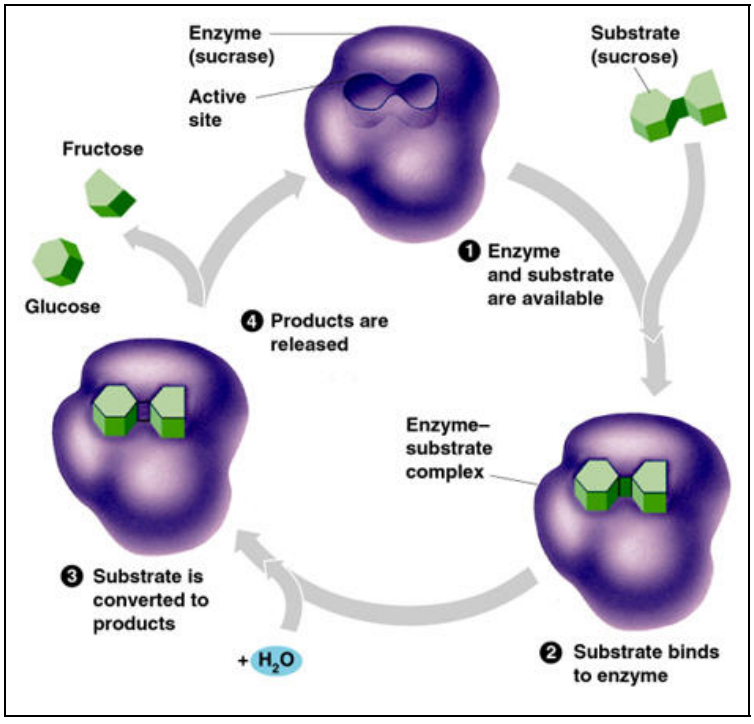


ภาพที่ 7 โครงสร้างจตุรภูมิของโปรตีนฮีโมโกลบิน (Hemoglobin)

ที่มา: biomed.brown.edu (2009)

1.2 ฟังก์ชั่นเอนไซม์ (Enzyme Function)

ฟังก์ชั่นเอนไซม์เป็นฟังก์ชั่นโปรตีนประเภทหนึ่งที่สำคัญอย่างยิ่งในสิ่งมีชีวิตทุกชนิด ลักษณะพิเศษของฟังก์ชั่นเอนไซม์คือความสามารถในการเร่งปฏิกิริยาที่สิ่งมีชีวิตต้องการได้อย่างรวดเร็ว เช่น การให้พลังงาน การย่อยอาหาร ฯลฯ การทำงานฟังก์ชั่นเอนไซม์ใช้บางส่วนของเอนไซม์ที่เรียกว่าบริเวณจับ (binding site) และบริเวณเร่ง (catalytic site) โดยบริเวณจับคือบริเวณที่เอนไซม์ใช้เข้าจับกับสารตั้งต้น (substrates) ให้อยู่ในสภาพที่พร้อมเกิดฟังก์ชั่น ส่วนบริเวณเร่งคือบริเวณที่เอนไซม์ใช้เหนี่ยวนำให้เกิดปฏิกิริยาเคมีกับสารตั้งต้นให้กลายเป็นผลิตภัณฑ์ (products) ดังนั้นโปรตีนที่มีโครงสร้างหรือลำดับกรดอะมิโนที่คล้ายคลึงกัน แต่บริเวณจับและบริเวณเร่งทำงานต่างกัน จึงไม่จำเป็นที่จะต้องมีฟังก์ชั่นเอนไซม์เหมือนกัน ในขณะที่โปรตีนที่มีโครงสร้างหรือลำดับกรดอะมิโนต่างกันมาก แต่มีลำดับกรดอะมิโนที่ทำหน้าที่บริเวณจับและบริเวณเร่งเหมือนกัน ย่อมสามารถมีฟังก์ชั่นเอนไซม์ที่เหมือนกันได้ (พีระ ลิ่วลม, 2551) รายละเอียดการทำงานของเอนไซม์แสดงดังภาพที่ 8



ภาพที่ 8 กลไกในบริเวณจับและบริเวณเร่งของเอนไซม์ซูเครส (sucrase)

ที่มา: porpax.bio.miami.edu (2009)

จากภาพที่ 8 แสดงการเข้าจับและเร่งปฏิกิริยาระหว่างเอนไซม์ซูเครส (sucrase) และสารตั้งต้นซูโครส (sucrose) จากภาพ หมายเลข 1 แสดงบริเวณเร่ง (active site) ของเอนไซม์ซูเครสและสารตั้งต้นซูโครส หมายเลข 2 แสดงการจับกัน (binding) ระหว่างเอนไซม์ซูเครสและสารตั้งต้นซูโครส ณ ตำแหน่งบริเวณเร่ง หมายเลข 3 แสดงการทำปฏิกิริยาของสารตั้งต้นซูโครสและน้ำ โดยมีตัวเร่งปฏิกิริยาคือเอนไซม์ซูเครส หมายเลข 4 แสดงผลลัพธ์จากปฏิกิริยาเคมีดังกล่าว ได้เป็นสารกลูโคส (glucose) และสารฟรัคโทส (fructose)

ดังนั้นงานด้านชีวสารสนเทศที่มุ่งเน้นศึกษาด้านเอนไซม์จึงไม่ได้มีเพียงความต้องการทำนายฟังก์ชั่นเอนไซม์ที่ถูกต้องแม่นยำเท่านั้น แต่หมายรวมไปถึงการใช้ข้อมูลจำนวนมากที่มีอยู่ในการศึกษาทำความเข้าใจการทำงานเอนไซม์ได้ดีขึ้น นั่นคือความสามารถในการระบุว่าส่วนใดของโปรตีนที่สามารถทำงานเป็นบริเวณจับหรือบริเวณเร่งที่ก่อให้เกิดกลไกการทำงานของฟังก์ชั่นเอนไซม์ได้อย่างสมเหตุผลน่าเชื่อถือในเชิงวิทยาศาสตร์ ซึ่งมีความสำคัญอย่างยิ่งต่อการนำไปใช้งานในระดับห้องปฏิบัติการ (หรือเรียกย่อว่า wet lab.) ในการทำความเข้าใจกลไกเอนไซม์และการออกแบบเอนไซม์ตัวใหม่

จากความสำคัญของเอนไซม์ ความซับซ้อนของฟังก์ชั่นเอนไซม์ และความต้องการในระดับห้องปฏิบัติการดังกล่าว ในงานวิจัยนี้จึงเน้นกลุ่มเป้าหมายคือเอนไซม์ และการพัฒนาตัวแทนสายโปรตีนที่ใช้ทำนายฟังก์ชั่นเอนไซม์ได้อย่างมีประสิทธิภาพ รวมทั้งยังสามารถหาความสัมพันธ์ระหว่างเอนไซม์และโครงสร้างทุติยภูมิได้โดยอาศัยองค์ความรู้จากตารางเมตริกซ์คอนแทคแมป

2. ฐานข้อมูลเอนไซม์และโปรตีน

ปัจจุบันฐานข้อมูลเอนไซม์และโปรตีนมีอยู่หลากหลายแหล่งข้อมูล มีทั้งฐานข้อมูลส่วนบุคคลและฐานข้อมูลสาธารณะ ได้แก่ Kyoto Encyclopedia of Genes and Genomes (KEGG), Comprehensive Enzyme Information system (BRENDA), The Universal Protein Resource (UniProt), National Center for Biotechnology Information (NCBI), Database of protein domains, families and functional site (PROSITE) เป็นต้น สำหรับในงานวิจัยนี้ใช้แหล่งข้อมูลเอนไซม์จากฐานข้อมูลสวิสพรอท (SWISS-PROT) และแหล่งข้อมูลโครงสร้าง 3 มิติที่จะนำมาใช้ในการสร้างตารางเมตริกซ์คอนแทคแมปนั้น นำมาจากฐานข้อมูลพีดีบี (Protein Data Bank, PDB)

2.1 ฐานข้อมูลสวิสพรอท (SWISS-PROT)

ฐานข้อมูลสวิสพรอท (Boeckmann, 2003) เกิดจากความร่วมมือระหว่างสถาบันชีวสารสนเทศแห่งประเทศสวิสเซอร์แลนด์ (The Swiss Institute of Bioinformatics, SIB) และสถาบันชีวสารสนเทศแห่งภูมิภาคยุโรป (European Bioinformatics Institute, EBI) เพื่อทำการจัดเก็บข้อมูลต่างๆ เกี่ยวกับโปรตีน อาทิเช่น ข้อมูลลำดับอะมิโน ข้อมูลหน้าที่การทำงานของโปรตีน ข้อมูลโครงสร้างของโปรตีน ข้อมูลเอนไซม์ชนิดต่างๆ เป็นต้น ฐานข้อมูลนี้เป็นฐานข้อมูลสาธารณะ ดังนั้นนักวิจัยสามารถเข้าใช้ฐานข้อมูลนี้ได้โดยตรงผ่านทางเว็บไซต์ www.expasy.org/sprot

2.2 ฐานข้อมูลพีดีบี (Protein Data Bank, PDB)

ฐานข้อมูลพีดีบี (Berman *et al*., 2002) เป็นแหล่งข้อมูลทางด้านโครงสร้าง 3 มิติ ของโปรตีนชนิดต่างๆ ทั้งที่ค้นพบจากกระบวนการ X-Ray Crystallographic และ Nuclear Magnetic Resonance (NMR) โดยมีการปรับปรุงฐานข้อมูลทุกสัปดาห์ ฐานข้อมูลนี้เป็นฐานข้อมูลสาธารณะภายใต้การดูแลของภาควิชาเคมีและชีวเคมี มหาวิทยาลัยรุทเกอร์ ในรัฐนิวเจอร์ซีย์ และศูนย์ซูเปอร์คอมพิวเตอร์ซานดิเอโก มหาวิทยาลัยแคลิฟอร์เนีย ประเทศสหรัฐอเมริกา ดังนั้นนักวิจัยสามารถเข้าใช้ฐานข้อมูลนี้ได้โดยตรงผ่านทางเว็บไซต์ www.rcsb.org

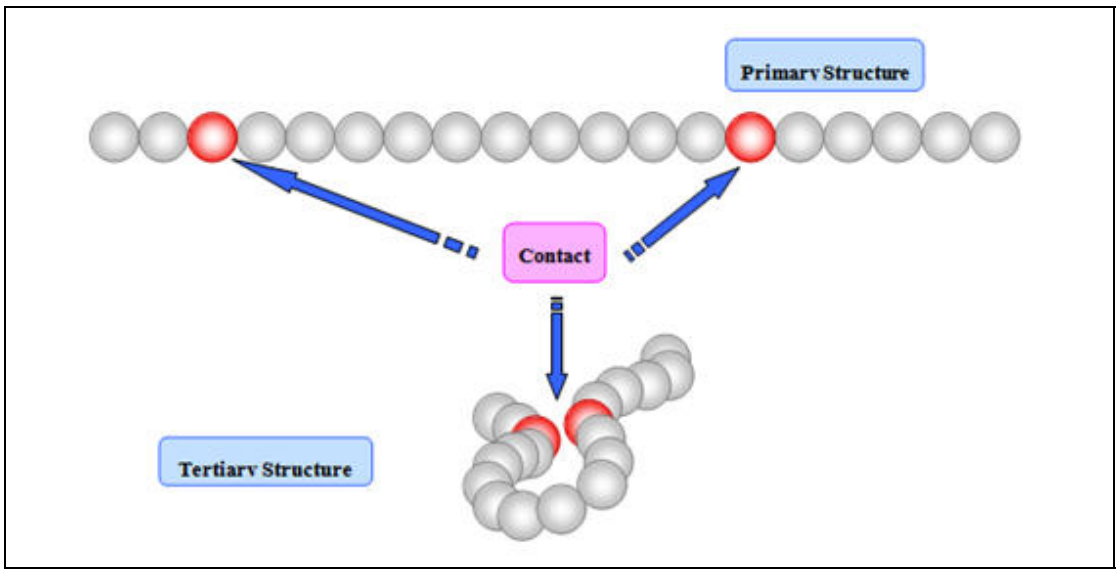
ปัจจุบันฐานข้อมูลพีดีบีมีข้อมูลโปรตีนที่ทราบโครงสร้าง 3 มิติทั้งสิ้น 55271 โครงสร้าง ภายในเว็บไซต์ประกอบด้วยข้อมูลที่เป็นประโยชน์มากมาย อาทิเช่น ข้อมูลปฐมภูมิซึ่งอยู่ในรูปแบบฟาสต้า (FASTA) ข้อมูลโครงสร้าง 3 มิติ นอกจากนี้ยังมีข้อมูลอ้างอิงกับฐานข้อมูลอื่น เช่นข้อมูลโครงสร้างทุติยภูมิ จากฐานข้อมูลสคอป (SCOP) และฐานข้อมูลแคช (CATH) ข้อมูลยีนออนโทโลจี (Gene Ontology) เป็นต้น ลักษณะข้อมูลของพีดีบีที่นำมาใช้ในงานวิจัยนี้ มีลักษณะดังภาพที่ 9

```
HEADER    TRANSFERASE                             22-MAR-04   1SR9
TITLE     CRYSTAL STRUCTURE OF LEUA FROM MYCOBACTERIUM TUBERCULOSIS
COMPND    MOL_ID: 1;
REMARK   1
REMARK   1 RESOLUTION. 2.00 ANGSTROMS.
SEQRES   1 A  644  MSE THR THR SER GLU SER PRO ASP ALA TYR THR GLU SER
SEQRES   2 A  644  PHE GLY ALA HIS THR ILE VAL LYS PRO ALA GLY PRO PRO
.....
SEQRES  50 B  644  ALA VAL ASN ARG ALA ALA ARG
ATOM      1  N   THR A  18      25.140  -9.574 -14.530  1.00 55.18           N
ATOM      2  CA  THR A  18      25.040  -8.843 -15.825  1.00 54.94           C
ATOM      3  C   THR A  18      23.656  -8.187 -16.042  1.00 53.26           C
.....
ATOM   8668  O   ALA B 643      36.391 -14.155  12.638  1.00 56.21           O
ATOM   8669  CB  ALA B 643      38.437 -16.997  11.989  1.00 53.23           C
TER    8670      ALA B 643
MASTER      462    0   27   50   34    0    0    6 9153    2  192  100
END
```

ภาพที่ 9 แสดงตัวอย่างรูปแบบข้อมูลของฐานข้อมูลพีดีบี (PDB format)

3. ความรู้เกี่ยวกับการติดกันของคู่ลำดับอะมิโน (Residue Contact) และตารางเมตริกซ์คอนแทคแมป (Contact Map Matrix)

Residue Contact เป็นการพิจารณาระยะห่างระหว่างคู่ลำดับกรดอะมิโนใดๆ บนโดเมน 3 มิติ ซึ่งบางกรณีพบว่ากรดอะมิโน 2 ตัวใดๆ เมื่อพิจารณาจากสายลำดับของโปรตีนหรือโครงสร้างปฐมภูมิ (primary Structure) พบว่าไม่อยู่ติดกัน แต่เมื่อนำมาพิจารณาบนโดเมน 3 มิติหรือโครงสร้างตติยภูมิ (tertiary Structure) กลับพบว่าอยู่ติดกัน แสดงดังภาพที่ 10



ภาพที่ 10 แสดงการติดกันระหว่างกรดอะมิโนบนโครงสร้างตติยภูมิ แต่ไม่อยู่ติดกันเมื่อพิจารณาบนโครงสร้างระดับปฐมภูมิ

ตารางเมตริกซ์คอนแทคแมปคือตารางเมตริกซ์ที่มีขนาด $N \times N$ (เมื่อ N คือความยาวของสายโปรตีน) โดยค่าในตารางเมตริกซ์นั้น ได้มาจากการพิจารณา Residue Contact ของทุกคู่ลำดับกรดอะมิโนในสายโปรตีน และเมื่อพิจารณารูปแบบที่ปรากฏบนตารางเมตริกซ์คอนแทคแมปจะพบว่ารูปแบบต่างๆ สามารถบ่งชี้ได้ว่ามีโครงสร้างทุติยภูมิใดบ้างที่อยู่ติดกันบนโดเมน 3 มิติ ซึ่งสามารถใช้แสดงความสัมพันธ์ของโครงสร้างโปรตีนและนำมาใช้ในการวิเคราะห์ได้ง่าย

สำหรับโปรตีนใดๆ ที่มีการค้นพบโครงสร้าง 3 มิติแล้ว โดยกระบวนการ X-Ray Crystallographic หรือ NMR สามารถทราบลักษณะการติดกันระหว่างกรดอะมิโนแต่ละคู่จากการคำนวณระยะห่างบนโดเมน 3 มิติ โดยพิจารณาจากตำแหน่งพิกัด (coordinate) ของอะตอมอัลฟา-คาร์บอน (α-Carbon) หรืออะตอมเบต้า-คาร์บอน (β-Carbon) ระหว่างกรดอะมิโนแต่ละคู่ โดยคำนวณได้ดังสมการที่ (1)

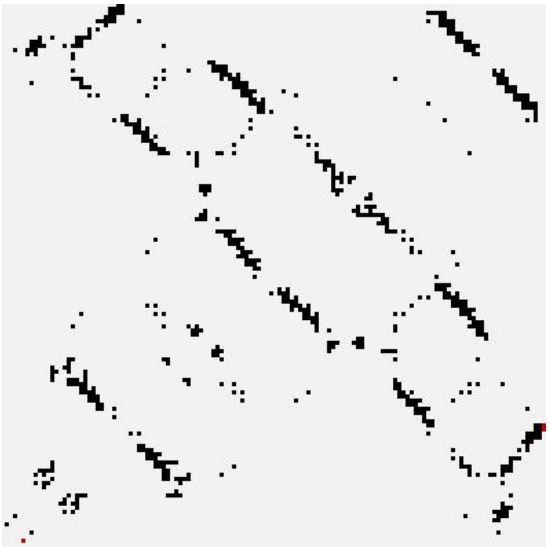
$$\delta(a_i, a_j) = \sqrt{(r_i - r_j)^2} = \left| r_i - r_j \right| \qquad \textbf{(1)}$$

โดยที่สัญลักษณ์ต่างๆ แสดงดังตารางที่ 2

ตารางที่ 2 แสดงสัญลักษณ์และคำอธิบายสัญลักษณ์ที่ใช้ในสมการการคำนวณระยะห่างระหว่างกรดอะมิโนบนโดเมน 3 มิติ

สัญลักษณ์	คำอธิบาย
$\delta(a_i, a_j)$	คือ ระยะห่างระหว่างคู่ลำดับกรดอะมิโนแต่ละคู่บนโดเมน 3 มิติ (หน่วย Angstroms, °A) ซึ่งคำนวณจากสมการ Euclidian Distance
r_i และ r_j	คือ ตำแหน่งพิกัดของอะตอม α-Carbon หรือ β-Carbon ของกรดอะมิโน a_i และ a_j ตามลำดับ

หมายเหตุ สำหรับกรดอะมิโนไกลซีน (glycine) จะพิจารณาตำแหน่งพิกัด (coordinate) ของอะตอม α-Carbon สำหรับกรดอะมิโนชนิดอื่น พิจารณาตำแหน่งพิกัดของอะตอม β-Carbon (CASP6, 2004)



ภาพที่ 11 แสดงตารางเมตริกซ์คอนแทคแมปของโปรตีน 2AMU (PDB code)

ในการสร้างตารางเมตริกซ์คอนแทคแมปนั้น ค่าที่อยู่ในตารางเมตริกซ์ $N \times N$ แสดงถึงการติดกันของกรดอะมิโนแต่ละคู่บนโดเมน 3 มิติ จากสมการที่ (1) พบว่า ถ้าค่าระยะห่างที่ได้มีค่าน้อยกว่าหรือเท่ากับค่าที่ผู้ใช้กำหนด (Threshold T) แสดงว่ากรดอะมิโนคู่นั้นถูกพิจารณาว่าติดกัน แต่ถ้าค่าระยะห่างที่ได้มีค่ามากกว่าค่าที่ผู้ใช้กำหนด นั่นคือกรดอะมิโนคู่นั้นไม่ถูกพิจารณาว่าติดกัน แสดงดังสมการที่ (2) เนื่องจากเมตริกซ์ $N \times N$ นั้นเป็นเมตริกซ์ที่สมมาตร (แสดงดังภาพที่ 11)

การพิจารณาตารางเมตริกซ์คอนแทคแมปจึงสามารถพิจารณาจากด้านแนวทแยงมุมเพียงด้านใดด้านหนึ่งก็เพียงพอ

$$\begin{aligned} &if\ \delta(a_i, a_j) \le T \quad then\ C_{ij} = 1 \\ &Otherwise \quad C_{ij} = 0 \end{aligned} \tag{2}$$

โดยที่สัญลักษณ์ต่างๆ แสดงดังตารางที่ 3

ตารางที่ 3 แสดงสัญลักษณ์และคำอธิบายสัญลักษณ์ที่ใช้ในสมการการคำนวณค่าในตารางเมตริกซ์คอนแทคแมป

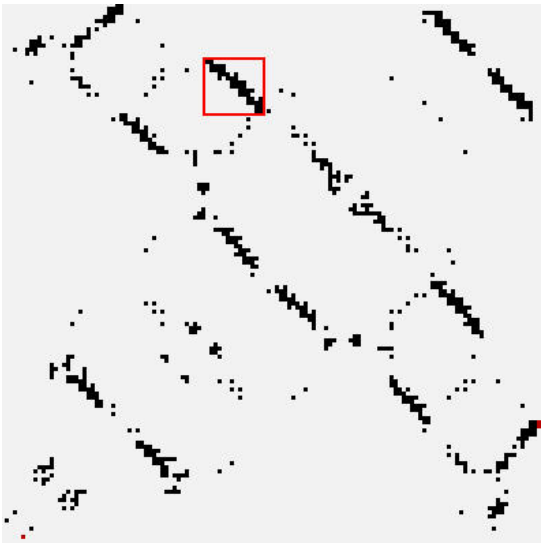
สัญลักษณ์	คำอธิบาย
C_{ij}	มีค่าเท่ากับ 1 เมื่อกรดอะมิโนที่ตำแหน่ง i และ j ติดกัน และ มีค่าเท่ากับ 0 เมื่อกรดอะมิโนที่ตำแหน่ง i และ j ไม่ติดกัน
T	คือ ค่า Threshold ของระยะห่างของคู่กรดอะมิโนใดๆ บนโดเมน 3 มิติ ถูกกำหนดโดยผู้ใช้งาน

4. วิธีการสืบค้นรูปแบบและการจัดกลุ่มโครงสร้างย่อยในตารางเมตริกซ์คอนแทคแมป

ในการสืบค้นรูปแบบโครงสร้างย่อยในตารางเมตริกซ์คอนแทคแมปและทำการจัดกลุ่มตามโครงสร้างย่อยนั้น ในงานวิจัยนี้ได้ประยุกต์ใช้บางส่วนของงานวิจัย (Yang *et al.*, 2004) ซึ่งเป็นงานวิจัยที่มีจุดมุ่งเน้นการสืบค้นความสัมพันธ์ระหว่างรูปแบบต่างๆ ที่เกิดขึ้นบนตารางเมตริกซ์คอนแทคแมป เพื่อนำไปใช้ในการเปรียบเทียบหาความเหมือนระหว่างโปรตีนได้ โดยประยุกต์ใช้ในขั้นตอนดังต่อไปนี้

4.1 การสืบค้นรูปแบบย่อยที่อยู่ติดและต่อเนื่องกันมากที่สุด (Mining Sub-Structural Contacted Patterns, SSCP)

เป็นการสืบค้นรูปแบบย่อยที่มีการติดกันของกรดอะมิโนที่มีขอบเขตมากที่สุด เมื่อพิจารณาจากตารางเมตริกซ์ ดังภาพที่ 12 โดยช่องที่ระบายสีดำในตารางเมตริกซ์ แทนการติดกันบนโดเมน 3 มิติ หรือ C_{ij} มีค่าเท่ากับ 1 นั่นเอง



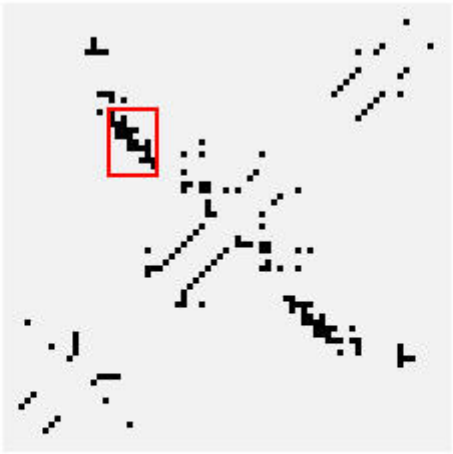
ภาพที่ 12 แสดงบริเวณ SSCP (กรอบสี่เหลี่ยมกลางภาพด้านบน) ของโปรตีน 2AMU (PDB code)

4.2 การสกัดฟีเจอร์เพื่อใช้เป็นตัวแทนของ SSCP ในการจัดกลุ่มตามรูปแบบที่เกิดขึ้นในตารางเมตริกซ์คอนแทคแมป

ทำการสกัดฟีเจอร์จาก SSCP ที่สืบค้นได้ในหัวข้อ 4.1 โดยฟีเจอร์ดังกล่าวมีอยู่ 6 ฟีเจอร์ด้วยกัน ดังแสดงในตารางที่ 4

ตารางที่ 4 แสดงฟีเจอร์ที่ต้องทำการสืบค้นจาก SSCP

ฟีเจอร์	คำอธิบาย
Height	ความสูงของ SSCP
Width	ความกว้างของ SSCP
NumOnes	จำนวนกรดอะมิโนที่อยู่ติดกันบนโดเมน 3 มิติ ที่อยู่ใน SSCP
Angle	มุมที่แสดงถึงการกระจายตัวของการติดกันของคู่กรดอะมิโนโดเมน 3 มิติ
xStdDev	ค่าการกระจายตัวของการติดกันของคู่กรดอะมิโนในแนวแกน X
yStdDev	ค่าการกระจายตัวของการติดกันของคู่กรดอะมิโนในแนวแกน Y



Height	Width	NumOnes	Angle	xStdDev	yStdDev
10	8	25	135.9657322	0.037859158	0.041880352

ภาพที่ 13 แสดงตัวอย่างของฟีเจอร์ที่สืบค้นได้จาก SSCP

5. การเทียบเรียงกลุ่มของสายโปรตีน (Multiple Sequence Alignment, MSA)

Multiple Sequence Alignment (MSA) ซึ่งเป็นเทคนิคที่พัฒนามาอย่างต่อเนื่องยาวนาน ส่วนใหญ่ประยุกต์ใช้ในการเปรียบเทียบความเหมือนของสายโปรตีน หรือใช้ในการสืบค้นโมทีฟ ยกตัวอย่างเช่น งานวิจัย Waterman *et al.* (1976) งานวิจัย Feng and Doolittle (1987) และงานวิจัย Barton (1990) เป็นต้น

ในการทำ MSA มีหลักการที่สำคัญคือการจัดเรียงลำดับสายโปรตีนของกลุ่มโปรตีนให้อยู่ในตำแหน่งที่มีค่าคะแนนความเหมือน (similarity) มากที่สุด โดยใช้ตารางคะแนนความเหมือน (similarity score table) ประเภทต่างๆ เช่น BLOSUM (Blocks of Amino Acid Substitution Matrix) (Henikoff and Henikoff, 1992) แสดงดังภาพที่ 14 และ PAM (Dayhoff *et al.*, 1978) แสดงดังภาพที่ 15 เข้ามาช่วยในการคำนวณ โดยองค์ประกอบของตารางประเภทนี้ประกอบด้วยคะแนนความเหมือนที่ได้จากการจับคู่กันของกรดอะมิโนแต่ละประเภทรวมทั้งแก๊ป (gap) (ใช้สัญลักษณ์ *) โดยจากภาพที่ 15 ได้ค่าคะแนนความเหมือนของกรดอะมิโน Y และ F เท่ากับ 7 คะแนน

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

ภาพที่ 14 แสดงเมตริกซ์การแทนที่ (substitution matrices) ของกรดอะมิโนในสายวิวัฒนาการของ BLOSUM62

ที่มา: National Center for Biotechnology Information (NCBI). (2009)

	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 15 แสดงเมตริกซ์การแทนที่ (substitution matrices) ของกรดอะมิโนในสายวิวัฒนาการของ PAM250

ที่มา: โปรแกรม BioEdit version 5.0.9 (Hall, 1999)

ในการจัดเรียงลำดับสายโปรตีนของกลุ่มโปรตีนให้อยู่ในตำแหน่งที่มีค่าคะแนนความเหมือนมากที่สุดอธิบายได้จากตัวอย่างในภาพที่ 16 แสดงการจัดเรียงสายโปรตีนสายสั้น 4 สายคือ NYLS, NKYLS, NFS, และ NFLS

```
N-YLS
NKYLS
N-F-S
N-FLS
★   ★
```

ภาพที่ 16 การทำ Multiple Sequence Alignment ของสายโปรตีนสายสั้น 4 สาย

จากภาพที่ 16 สัญลักษณ์ ★ หมายถึงบริเวณอนุรักษ์ โดยผลลัพธ์จากการทำ 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

6. วิธีการเรียนรู้ ซัพพอร์ตเวคเตอร์แมชชีน (Support Vector Machine, SVM)

ซัพพอร์ตเวคเตอร์แมชชีน (Vapnik, 1995) เป็นวิธีการเรียนรู้ (machine learning) ที่ใช้ในงานวิจัยอย่างแพร่หลาย โดยจะสร้างตัวจำแนกคลาสของข้อมูลจากการเรียนรู้คุณลักษณะของข้อมูลในแต่ละคลาสต้นแบบ โดยในหัวข้อนี้จะกล่าวถึงความรู้พื้นฐานของ SVM และวิธีการต่างๆ ที่ใช้ในการสร้างตัวจำแนกข้อมูลแบบหลายคลาส (multi-class) เพื่อรองรับการจำแนกข้อมูลสำหรับงานวิจัยนี้

6.1 พื้นฐานซัพพอร์ตเวคเตอร์แมชชีน (Support Vector Machine)

สมมติให้ชุดข้อมูล D มีสมาชิกทั้งหมด N ข้อมูล โดยข้อมูลแต่ละตัวอยู่ในรูปของเวคเตอร์ $\vec{x}$ ซึ่งประกอบด้วย p-dimension และสามารถจำแนกข้อมูลให้อยู่ในคลาส $c_i \in \{+1,-1\}$ โดยที่ $i \in \{1,\ldots,N\}$ ซึ่งแบ่งข้อมูลได้เป็น 2 คลาสคือคลาสที่หนึ่ง (+1) และคลาสที่สอง (-1) สามารถเขียนได้ดังสมการที่ (3)

$$D = \{(\vec{x}_i, c_i) \mid i \in \{1,\ldots,N\}, \vec{x}_i \in \Re^p, c_i \in \{+1,-1\}\} \tag{3}$$

หลักการของ SVM จะพยายามแปลงข้อมูลจากฟีเจอร์สเปซที่มีมิติ p ให้เป็นฟีเจอร์สเปซที่มีมิติที่สูงขึ้น (high-dimensional feature space) เพื่อให้ข้อมูลมีลักษณะเป็นเชิงเส้น (linear) โดยการโปรเจ็ค ϕ ลงบนมิติที่สูงขึ้น H แสดงเป็น $\phi: \Re^p \mapsto H$ เนื่องจากบางกรณีไม่สามารถหาค่าโปรเจ็ค ϕ ที่ชัดเจนได้ จึงใช้เคอร์เนลฟังก์ชั่น (kernel function) $K(\vec{x}_i, \vec{x}_j) = \phi(\vec{x}_i) \cdot \phi(\vec{x}_j)$ แทน จากนั้นทำการสร้างไฮเปอร์เพลน (hyperplane) ที่เหมาะสม เพื่อทำการแบ่งข้อมูลออกเป็น 2 คลาส โดยให้มีขอบเขต (margin) ของข้อมูลห่างกันมากที่สุด จากภาพที่ 17 จะได้ว่าเคอร์เนลฟังก์ชั่น $K(\vec{x}_i, \vec{x}_j)$ ทำการแปลงข้อมูลให้อยู่ในรูปแบบเชิงเส้นบนฟีเจอร์สเปซที่มีมิติสูงขึ้น และสมการในรูปทั่วไปของ SVM สามารถเขียนได้เป็น

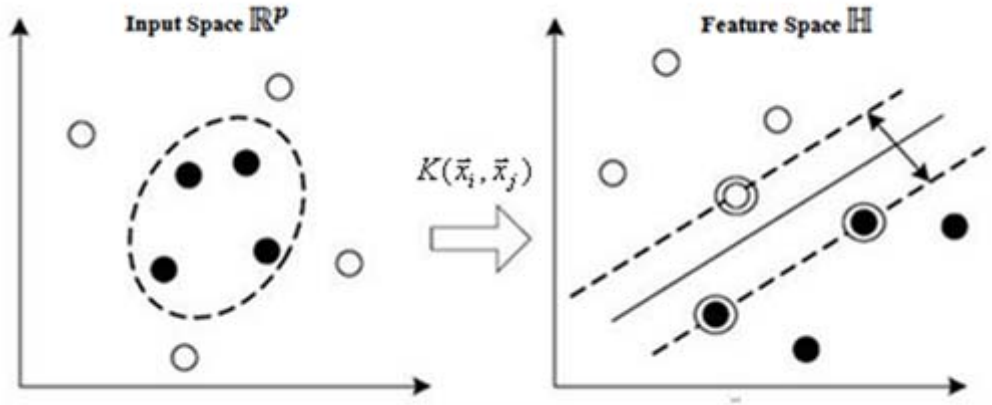
$$f(\vec{x}) = sign\left(\sum_{i=1}^{N} y_i \alpha_i \cdot K(\vec{x}_i, \vec{x}_j) + b\right) \quad \textbf{(4)}$$

โดยที่ค่าสัมประสิทธิ์ α_i คือการหาค่าที่มากที่สุดของสมการที่ (5) ซึ่งเป็นการแก้ปัญหา Quadratic Programming (QP)

$$\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \cdot y_i y_j \cdot K(\vec{x}_i, \vec{x}_j) \quad \textbf{(5)}$$

โดยที่ $0 \le \alpha_i \le C$ และ $\sum_{i=1}^{N} \alpha_i y_i = 0$ เมื่อ $(i = 1,2,\ldots,N)$

จากสมการที่ (5) จะได้ว่าค่า C คือพารามิเตอร์ที่ใช้เปรียบเทียบระหว่างการแบ่งขอบเขตให้กว้างที่สุดกับค่าความถูกต้องในการแบ่งข้อมูล และเรียก $\vec{x}_j$ ว่าเป็น Support Vector เมื่อ $\alpha_i > 0$



ภาพที่ 17 แสดงการแปลงข้อมูลให้อยู่ในรูปแบบเชิงเส้นบนฟีเจอร์สเปซที่มีมิติสูงขึ้น จากภาพวงกลมสีดำแสดงข้อมูลที่อยู่ในคลาสที่หนึ่ง และวงกลมสีขาวแสดงข้อมูลที่อยู่ในคลาสที่สอง วงกลมที่มีเส้นล้อมรอบจะแสดงซัพพอร์ตเวคเตอร์ (support vector)

สำหรับเคอร์เนลฟังก์ชั่นที่นิยมใช้ในการแปลงฟีเจอร์สเปซให้อยู่บนมิติที่สูงขึ้นมี 4 ฟังก์ชั่นได้แก่ (1) Linear kernel (2) Polynomial kernel (3) Radial Basis Function (RBF) kernel และ (4) Sigmoid kernel ดังสมการที่ 6, 7, 8 และ 9 ตามลำดับ

$$K(\vec{x}_i, \vec{x}_j) = \vec{x}_i \cdot \vec{x}_j \quad \textbf{(6)}$$

$$K(\vec{x}_i, \vec{x}_j) = (\vec{x}_i \cdot \vec{x}_j + \theta)^d \qquad (d = 1,2,\ldots) \quad \textbf{(7)}$$

$$K(\vec{x}_i, \vec{x}_j) = \exp(-\gamma \left\| \vec{x}_i - \vec{x}_j \right\|^2) \quad \textbf{(8)}$$

$$K(\vec{x}_i, \vec{x}_j) = \tanh[(\vec{x}_i \cdot \vec{x}_j) - c] \quad \textbf{(9)}$$

โดยที่สมการที่ (8) $\gamma = 1/\sigma^2$ และพารามิเตอร์ σ คือค่าความกว้างของเคอร์เนล ดังนั้นถ้า γ มีค่าต่ำ หรือ σ มีค่าสูง จะทำให้จำแนกคลาสของข้อมูลได้ดีขึ้น (สำหรับงานวิจัยนี้ใช้ RBF เป็นเคอร์เนลฟังก์ชั่น)

6.2 ซัพพอร์ตเวคเตอร์แมชชีนสำหรับการจำแนกข้อมูลแบบหลายคลาส (Multiclass Support Vector Machine)

จากความรู้พื้นฐานของซัพพอร์ตเวคเตอร์แมชชีนที่กล่าวข้างต้น เป็นวิธีการในการจำแนกข้อมูลแบบไบนารีคลาส สำหรับปัญหาการจำแนกคลาสข้อมูลออกเป็น k คลาสนั้น สามารถทำได้ 2 วิธีการได้แก่ (1) วิธี one versus-rest เป็นการสร้างตัวจำแนกคลาสของข้อมูลจำนวน k ตัว โดยแต่ละตัวจำแนกคลาสของข้อมูล i^{th} เกิดจากการฝึกสอนระบบ (train) ด้วยข้อมูลของคลาสนั้นให้เป็นข้อมูลของคลาสที่หนึ่งและข้อมูลของคลาสที่เหลือรวมกันเป็นข้อมูลของคลาสที่สอง และ (2) วิธี one-versus-one เป็นการฝึกสอนระบบด้วยข้อมูลทีละ 2 คลาสไปเรื่อยๆ จนครบทุกคู่ สุดท้ายแล้วจะได้ตัวจำแนกคลาสของข้อมูลทั้งหมดจำนวน $k(k-1)/2$ ตัว

7. แบบจำลองฮิดเด้นมาร์คอฟ (Hidden Markov Model, HMM)

7.1 มาร์คอฟโพรเซส (Markov Process)

เป็นกระบวนการพิจารณาความน่าจะเป็นของการเกิดลำดับเหตุการณ์ ณ เวลาหนึ่ง t ซึ่งขึ้นกับค่าความน่าจะเป็นของการเกิดเหตุการณ์ก่อนหน้า โดยมีสมมติฐานหลักในการสร้าง

แบบจำลองได้แก่ ต้องมีกลุ่มของสถานะ (state) ของสิ่งที่กำลังศึกษาที่มีจำนวนแน่นอน N สถานะ แทนด้วย $S_1,\ldots,S_n$ ความน่าจะเป็นของการเปลี่ยนสถานะ (state) จากสถานะต้นทาง i ไปสถานะปลายทาง j แทนด้วย a_{ij} และสถานะปัจจุบัน ณ เวลา t แทนด้วย q_t แต่เนื่องจากคุณสมบัติของมาร์คอฟ (markov property) นิยามไว้ว่าความน่าจะเป็นของการเปลี่ยนสถานะไปยังปลายทาง j ขึ้นกับความน่าจะเป็นของสถานะต้นทาง i เท่านั้น ซึ่งไม่เกี่ยวข้องกับกระบวนการก่อนหน้าสถานะ i ดังนั้นเราสามารถเขียนสมการคุณสมบัติของมาร์คอฟได้ดังนี้

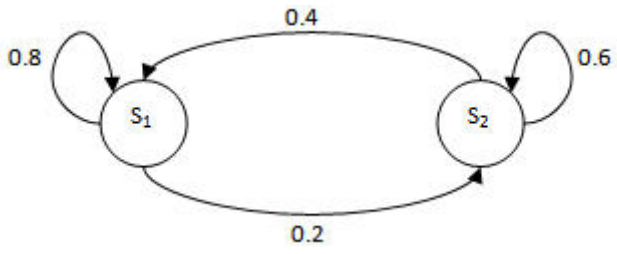
$$\Pr[q_t = S_t \mid q_{(t-1)} = S_{(t-1)}, q_{(t-2)} = S_{(t-2)},\ldots,q_0 = S_0] = \Pr[q_t = S_t \mid q_{(t-1)} = S_{(t-1)}] \qquad \textbf{(10)}$$

โดยที่ค่าความน่าจะเป็นของการเปลี่ยนสถานะจากสถานะต้นทาง i ไปยังสถานะปลายทาง j ได้ดังนี้

$$a_{ij} = \Pr[q_t = S_j \mid q_{(t-1)} = S_i], \quad 1 \le i, j \le N \qquad \textbf{(11)}$$

เมื่อ $a_{ij} \ge 0, \sum_{j=1}^{N} a_{ij} = 1$

พิจารณาตัวอย่างของการทอยเหรียญสองหน้าคือหัวและก้อยจำนวน 1 เหรียญ ดังนั้นสถานะจึงมี 2 สถานะคือ หัว แทนด้วย S_1 และก้อย แทนด้วย S_2 และมีค่าความน่าจะเป็นของการเปลี่ยนสถานะดังแสดงในภาพที่ 18



ภาพที่ 18 แสดงมาร์คอฟโพรเซส (markov process) ของการทอยเหรียญสองหน้าจำนวน 1 เหรียญ

ดังนั้นค่าความน่าจะเป็นของการเกิดเหตุการณ์ที่เหรียญจะออก “หัว-หัว-ก้อย-หัว-ก้อย-ก้อย-ก้อย” มีค่าเป็นเท่าไรนั้น เราสามารถคำนวณได้โดยกำหนดให้เหตุการณ์ที่เราสังเกต O (observe) ผ่านโมเดล M (model) มีค่าดังนี้

$$\begin{aligned}
\Pr(O \mid M) &= \Pr[S_1 S_1 S_2 S_1 S_2 S_2 S_2 \mid M] \\
&= \Pr[S_1] \cdot \Pr[S_1 \mid S_1] \cdot \Pr[S_2 \mid S_1] \cdot \Pr[S_1 \mid S_2] \cdot \Pr[S_2 \mid S_1] \\
&\quad \cdot \Pr[S_2 \mid S_2] \cdot \Pr[S_2 \mid S_2] \\
&= \pi_1 \cdot a_{11} \cdot a_{12} \cdot a_{21} \cdot a_{12} \cdot a_{22} \cdot a_{22} \\
&= 1 \cdot (0.8) \cdot (0.2) \cdot (0.4) \cdot (0.2) \cdot (0.6) \cdot (0.6) \\
&= 4.608 \times 10^{-3}
\end{aligned}$$

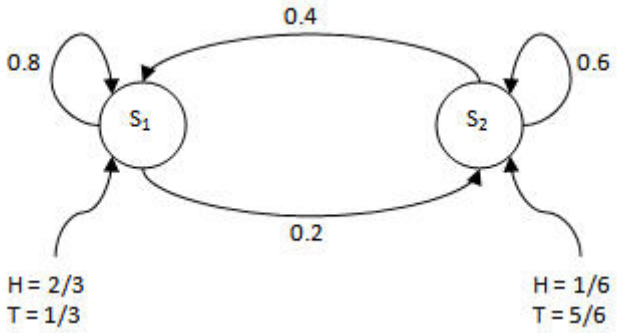
โดยที่ π_1 คือค่าความน่าจะเป็นตั้งต้นของสถานะ S_1 (initial state) มีค่าค่าเท่ากับ 1

7.2 แบบจำลองฮิดเด้นมาร์คอฟ (Hidden Markov Model, HMM)

จะเห็นได้ว่ามาร์คอฟโพรเซสนั้น เราสามารถสังเกตเหตุการณ์ที่เกิดขึ้นในการเปลี่ยนสถานะได้อย่างชัดเจน แต่ในบางกรณีเราไม่สามารถสังเกตเห็นเหตุการณ์ที่เกิดขึ้นในโพรเซสได้ ยกตัวอย่างเช่นการทอยเหรียญ 2 หน้าได้แก่ หน้าหัวและหน้าก้อย จำนวน 2 เหรียญโดยแต่ละเหรียญมีความน่าจะเป็นในการออกหัวและก้อยไม่เท่ากันเรียกว่ามีไบแอส ในกรณีนี้สถานะ (state) คือเหรียญที่ 1 และเหรียญที่ 2 แทนด้วย S_1 และ S_2 ตามลำดับ ผลลัพธ์จากการทอยเหรียญคือ หัวหรือก้อย แทนด้วย H และ T ตามลำดับ ส่วนความน่าจะเป็นในการเปลี่ยนสถานะและความน่าจะเป็นของการออกหัวหรือก้อยของแต่ละเหรียญ แสดงดังภาพที่ 19 ในที่นี้สิ่งที่เราสังเกตคือลำดับเหตุการณ์ของผลลัพธ์ในการออกหัวและก้อย ดังนั้นเราจะไม่ทราบลำดับในการเปลี่ยนสถานะ (hidden) ซึ่งลักษณะเช่นนี้เองที่เรียกว่าฮิดเด้นมาร์คอฟโมเดล (Hidden Markov Model) ซึ่งมีส่วนประกอบที่เกี่ยวข้อง 5 ตัว ดังตารางที่ 5

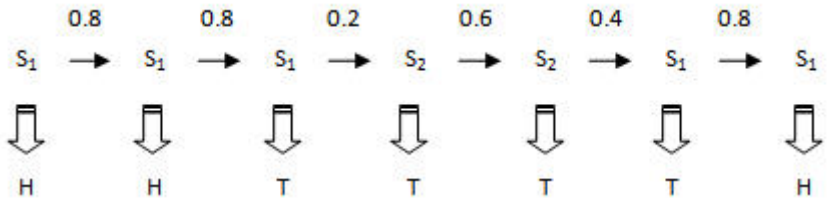
ตารางที่ 5 สัญลักษณ์ของส่วนประกอบของฮิดเด้นมาร์คอฟโมเดล

สัญลักษณ์	คำอธิบาย
Q	กลุ่มของสถานะ $(S_1, S_2, \ldots, S_n)$
V	เอาท์พุตของแต่ละสถานะ $(v_1, v_2, \ldots, v_m)$
$\pi(i)$	ค่าความน่าจะเป็นเริ่มต้นของสถานะ S_i ที่เวลา $t = 0$ (initial state)
A	ค่าความน่าจะเป็นของการเปลี่ยนสถานะ a_{ij} โดยที่ $a_{ij} = \Pr[q_t = S_j \mid q_{(t-1)} = S_i]$ และ a_{ij} ไม่ขึ้นกับสถานะของเวลาก่อนหน้า $t-1$ (เป็นคุณสมบัติของมาร์คอฟ)
B	ค่าความน่าจะเป็นของเอาท์พุต $b_j(k)$ โดยที่ $b_j(k) = \Pr[\,producing\, v_k\, at\, time\ \ t \mid in\, state\ \ S_j\, at\, time\ \ t]$



ภาพที่ 19 แสดงฮิดเด้นมาร์คอฟโมเดลของการทอยเหรียญสองหน้าที่มีไบแอส จำนวน 2 เหรียญ

จากภาพที่ 19 เราสามารถคำนวณหาค่าความน่าจะเป็นของลำดับเหตุการณ์ “H-H-T-T-T-T-H” ได้ดังนี้



ภาพที่ 20 แสดงรูปแบบลำดับเหตุการณ์เอาท์พุตที่ได้จากการเปลี่ยนสถานะ “S1 S1 S1 S2 S2 S1 S1”

ค่าความน่าจะเป็นของการเปลี่ยนสถานะหาได้จาก

$$\Pr[S_1S_1S_1S_2S_2S_1S_1] = \pi(S_1)a_{11}a_{11}a_{12}a_{22}a_{21}a_{11} \approx 0.025$$

ค่าความน่าจะเป็นของการเกิดลำดับเหตุการณ์ของเอาท์พุต “H-H-T-T-T-T-H” หาได้จาก

$$\Pr[(HHTTTTH)\,|\,(S_1S_1S_1S_2S_2S_1S_1)] = \frac{2}{3}\cdot\frac{2}{3}\cdot\frac{1}{3}\cdot\frac{5}{6}\cdot\frac{5}{6}\cdot\frac{1}{3}\cdot\frac{2}{3} \approx 0.023$$

ค่าความน่าจะเป็นของการเกิดลำดับเหตุการณ์เอาท์พุต “H-H-T-T-T-T-H” และการเปลี่ยนสถานะ “$S_1\ S_1\ S_1\ S_2\ S_2\ S_1\ S_1$” เป็นดังนี้

$$\Pr[(HHTTTTH) \wedge (S_1S_1S_1S_2S_2S_1S_1)] \approx (0.025)\cdot(0.023) \approx 5.7\times10^{-4}$$

ในการนำไปประยุกต์ใช้งานนั้น เราจะไม่ทราบลำดับของการเปลี่ยนสถานะที่แท้จริงว่าเป็นอย่างไร ดังนั้นให้ทำการพิจารณาหาค่าความน่าจะเป็นมากที่สุดของการเปลี่ยนสถานะเพื่อให้เกิดลำดับเหตุการณ์ที่เอาท์พุตที่กำลังสังเกตในโมเดลนั้น ด้วยไดนามิกอัลกอริทึม (dynamic algorithm) ดังสมการที่ (12)

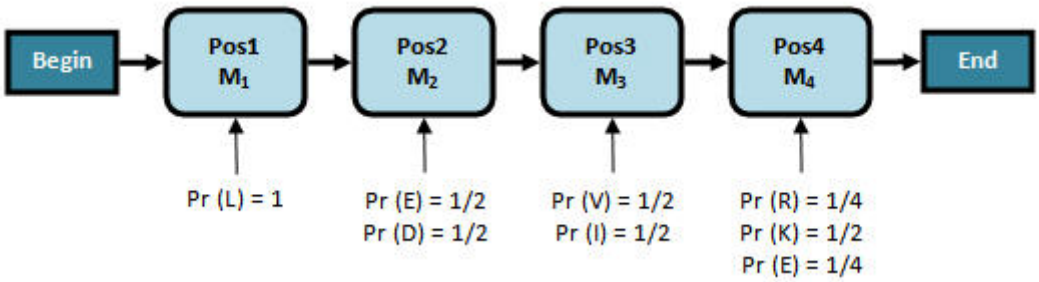
$$\max(\Pr[S_1, S_2, \ldots, S_T \,|\, O_1, O_2, \ldots, O_T, Model]) \qquad \textbf{(12)}$$

7.3 โพรไฟล์ฮิดเด้นมาร์คอฟโมเดล (Profile Hidden Markov Model, Profile HMM)

เป็นการนำโพรไฟล์ที่เกิดจากการเทียบเรียงสายโปรตีน (alignment) เข้าเป็นอินพุตของฮิดเด้นมาร์คอฟโมเดลแทนข้อมูลดิบ ซึ่งตำแหน่งของกรดอะมิโนที่ได้หลังจากการเทียบเรียงแล้วนั้นจะเกิดตำแหน่งที่เรียกว่าบริเวณอนุรักษ์ (conserve region) ซึ่งเป็นตำแหน่งที่บ่งชี้ความสัมพันธ์ทางชีววิทยาระหว่างชุดของสายโปรตีน โพรไฟล์ฮิดเด้นมาร์คอฟโมเดลนี้ถูกพัฒนาขึ้นโดย Krong *et al.* ในปี 1994 หลักการคือตำแหน่งคอลัมน์ที่ได้หลังจากการเทียบเรียงสายโปรตีน (consensus column) นั้น จะถูกใช้เป็นสถานะที่เรียกว่าสถานะที่ตรงกัน (match state) ค่าความน่าจะเป็นของการเปลี่ยนสถานะจากตำแหน่งไปยังอีกตำแหน่งหนึ่งมีค่าเท่ากับ 1 และทุก match state นั้นจะต้องมีค่าความน่าจะเป็นของการเกิดเอาท์พุต ซึ่งได้มาจากการคำนวณค่าความถี่ที่ปรากฏขึ้นของกรดอะมิโนที่ตำแหน่งนั้น ตัวอย่างโพรไฟล์ฮิดเด้นมาร์คอฟโมเดลอย่างง่ายแสดงดังภาพที่ 21 และ 22

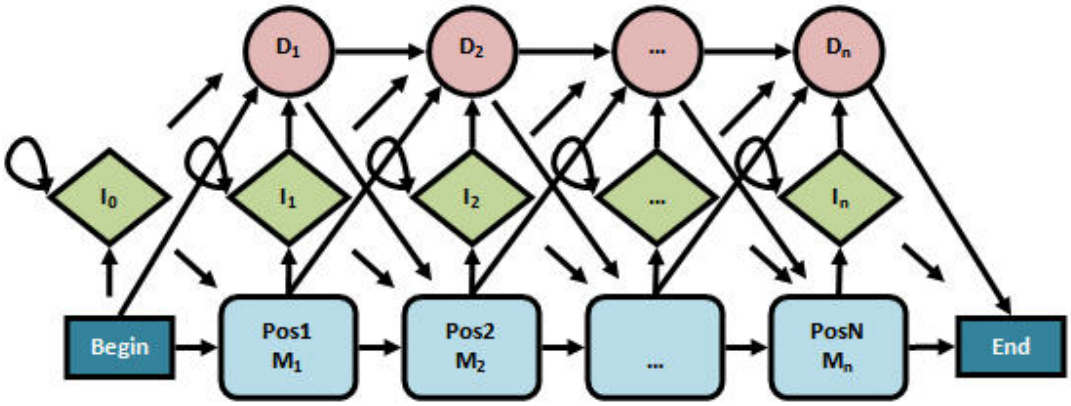
```
LEVK
LDIR
LEIK
LDVE
```

ภาพที่ 21 แสดงโพรไฟล์ที่ได้จากการเทียบเรียงสายโปรตีน 4 สาย



ภาพที่ 22 แสดงโพรไฟล์ฮิดเด้นมาร์คอฟโมเดลอย่างง่าย

คุณสมบัติที่สำคัญของฮิดเด้นมาร์คอฟโมเดลคือการเพิ่มหรือลบสถานะของการเกิดลำดับเหตุการณ์ ให้ I_j แทนการเพิ่มสถานะหลังคอมลัมน์ที่ j ซึ่งก็คือการยอมให้มีแก็ป (gap) ได้นั่นเอง ส่วนการลบสถานะนั้นแทนได้ด้วย D_j ดังภาพที่ 23



ภาพที่ 23 แสดงโพรไฟล์ฮิดเด้นมาร์คอฟโมเดลที่มีสถานะเพิ่ม (insert) และลบ (delete) โดยสี่เหลี่ยมโค้งมนคือสถานะของคอลัมน์ที่เป็นบริวณอนุรักษ์ สี่เหลี่ยมข้าวหลามตัดคือสถานะเพิ่ม และวงกลมคือสถานะลบ

พิจารณาตัวอย่างที่มีความซับซ้อนมากขึ้น โดยภาพที่ 24 คือโพรไฟล์ที่ได้จากการเทียบเรียงสายโปรตีน 7 เส้น จะเห็นได้โพรไฟล์ดังกล่าวจะมีทั้งสถานะเพิ่มและลบอยู่ด้วย

```
VGA--HAGEY
V----NVDEV
VEA--DVAGH
VKG------D
VYS--TYETS
FNA--NIPKH
IAGADNGAGY
```

ภาพที่ 24 แสดงโพรไฟล์ที่ได้จากการเทียบเรียงสายโปรตีน 7 สาย

หลักในการสร้างแบบจำลองนั้น สิ่งแรกที่ต้องพิจารณาคือสถานะแต่ละสถานะมีอะไรบ้าง วิธีการพื้นฐานคือให้พิจารณาตำแหน่งที่มีกรดอะมิโนอยู่อย่างน้อยเกินครึ่งหนึ่งของจำนวนสายโปรตีนให้เป็น match state ดังนั้นจะได้ว่า match state คือตำแหน่งที่ 1, 2, 3, 6, 7, 8, 9 และ 10 ขั้นตอนถัดมาคือคำนวณหาค่าความน่าจะเป็นของเอาท์พุตในแต่ละ match state กำหนดให้ b_{Mi} คือค่าความค่าความน่าจะเป็นของเอาท์พุตที่ตำแหน่ง i ดังนั้นจากตัวอย่างสามารถหาb_{M1} (ตำแหน่งที่ 1) ได้ดังนี้

$$b_{M1}V = \frac{5}{7}$$
$$b_{M1}F = \frac{1}{7}$$
$$b_{M1}I = \frac{1}{7}$$

สำหรับในกรณีที่ค่าความน่าจะเป็นของเอาท์พุตมีค่าเป็น 0 นั้น จะส่งผลต่อการคำนวณหาค่าความน่าจะเป็นของลำดับเหตุการณ์ที่เกิดขึ้นทั้งหมด ดังนั้นให้ทำการแทนค่าค่าความน่าจะเป็นของเอาท์พุตที่มีค่าเป็น 0 ด้วยค่าที่เล็กมากๆ 1 ค่า ซึ่งวิธีนี้เรียกว่ากฎ add-one (add-one rule) จากวิธีการดังกล่าวสามารถคำนวณค่าความน่าจะเป็นของเอาท์พุตได้ใหม่ดังนี้ (ค่าที่บวกเพิ่มคือ $\frac{1}{20}$)

$$b_{M1}V = \frac{5+1}{7+20} = \frac{6}{27}$$
$$b_{M1}F = \frac{1+1}{7+20} = \frac{2}{27}$$
$$b_{M1}I = \frac{1+1}{7+20} = \frac{2}{27}$$
$$all \quad else = \frac{0+1}{7+20} = \frac{1}{27}$$

การคำนวณค่าความน่าจะเป็นของการเปลี่ยนสถานะ (transition) a_{ij} จากสถานะ i ไปยังสถานะ j สามารถคำนวณได้ดังนี้

$$\frac{\text{ความถี่ในการเปลี่ยนสถานะจาก } i \text{ ไปยังสถานะ } j}{\text{ความถี่ในการเปลี่ยนสถานะจาก } i \text{ ไปยังสถานะอื่นทุกสถานะ}}$$

จากตัวอย่างสามารถคำนวณค่าความน่าจะเป็นของการเปลี่ยนสถานะจากตำแหน่งที่ 1 ไปตำแหน่งที่ 2 ได้เป็น (สมมติว่ามีตำแหน่งสถานะที่ 2 และ 3 เป็นการลบสถานะ และสถานะที่ 4 เป็นต้นไปเป็นการเพิ่มสถานะ)

$$a_{M1M2} = \frac{6}{7}$$
$$a_{M1D1} = \frac{1}{7}$$
$$a_{M1I1} = \frac{0}{7}$$

จะเห็นได้ว่าการเปลี่ยนสถานะนี้มีแก๊ปเท่ากับ 1 อย่างไรก็ตามค่าความน่าจะเป็นของการเปลี่ยนสถานะควรมีค่ามากกว่า 0 ดังนั้น เราจึงใช้กฎ add-one กับค่าความน่าจะเป็นของการเปลี่ยนสถานะเช่นเดียวกัน ซึ่งได้ผลลัพธ์ดังนี้ (ค่าที่บวกเพิ่มคือ $\frac{1}{3}$)

$$a_{M1M2} = \frac{6+1}{7+3} = \frac{7}{10}$$
$$a_{M1D1} = \frac{1+1}{7+3} = \frac{2}{10}$$
$$a_{M1I1} = \frac{0+1}{7+3} = \frac{1}{10}$$

ในขั้นตอนการเลือกสถานะที่ทำให้เกิดลำดับเหตุการณ์ที่เราสังเกตนั้น ใช้การพิจารณาค่าความน่าจะเป็นมากที่สุดของการเปลี่ยนสถานะเพื่อให้เกิดลำดับเหตุการณ์ที่เราสังเกตในโมเดลนั้น ซึ่งกล่าวไว้แล้วในสมการที่ (12) ข้างต้น

สำหรับงานวิจัยนี้ใช้ซอฟต์แวร์เพื่อสร้างโพรไฟล์ฮิดเด้นมาร์คอฟโมเดลที่ชื่อว่า HMMER (Eddy, 1998) version 2.3.2 ในซอฟต์แวร์ HMMER นี้จะมีฟังก์ชั่นที่งานวิจัยนี้เกี่ยวข้องอยู่ด้วยกัน 3 ส่วนด้วยกันได้แก่ ส่วนแรกคือ hmmbuild เป็นการสร้างโพรไฟล์ฮิดเด้นมาร์คอฟโมเดลจากเอาท์พุตของการเทียบเรียงสายโปรตีน (multiple sequence alignment, MSA) ส่วนที่สองคือ hmmcalibrate เป็นการปรับแต่งโมเดลให้มีประสิทธิภาพมากขึ้น และส่วนที่สามคือ hmmsearch เป็นการค้นหาความคล้ายคลึงกันของสายโปรตีนที่สนใจกับฐานข้อมูลที่สร้างขึ้นด้วย hmmbuild โดยค่าผลลัพธ์ที่ได้จะอยู่ในรูปของคะแนนความเหมือน (score) และค่าความคลาดเคลื่อน (e-value) ซึ่งงานวิจัยนี้ พิจารณาความคล้ายคลึงกันของสายโปรตีนกับฐานข้อมูลจากค่า e-value เป็นสำคัญ

8. การแบ่งกลุ่มข้อมูล (Data Clustering)

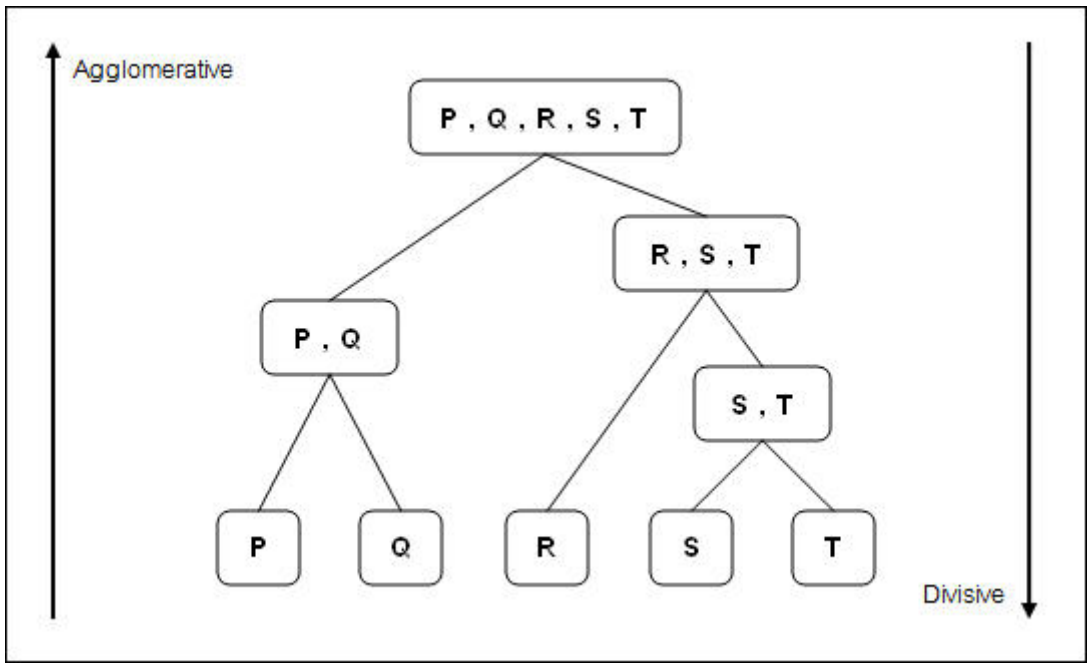
การแบ่งกลุ่มข้อมูล (Data Clustering) เป็นศาสตร์แขนงหนึ่งทางเดต้าไมน์นิ่ง (Data Mining) เป็นวิธีการแบ่งกลุ่มข้อมูลโดยไม่อาศัยตัวอย่างในการเรียนรู้ (unsupervised learning) โดยจะทำการแบ่งกลุ่มข้อมูลออกเป็นกลุ่มย่อยๆ เรียกว่าคลัสเตอร์ (cluster) โดยพิจาราณาจากความคล้ายคลึงกันของข้อมูลเป็นเกณฑ์หลักในการแบ่งกลุ่มข้อมูล ดังนั้นกลุ่มข้อมูลที่มีลักษณะคล้ายคลึงกันจะถูกจัดให้อยู่ในกลุ่มเดียวกัน และข้อมูลที่มีลักษณะแตกต่างกันจะถูกจัดไว้ต่างกลุ่มกัน วิธีการแบ่งกลุ่มนี้มีประโยชน์อย่างมากในการวิเคราะห์ความสัมพันธ์กันของชุดข้อมูล เทคนิคการแบ่งกลุ่มในปัจจุบันมีหลากหลายเทคนิค แต่ที่ใช้กันแพร่หลายได้แก่ การแบ่งกลุ่มแบบพาร์ทิชั่น (partition based) การแบ่งกลุ่มแบบใช้ความหนาแน่น (density based) การแบ่งกลุ่มโดยใช้กริด (grid based) และการแบ่งกลุ่มแบบมีลำดับชั้น (hierarchical based) ซึ่งในงานวิจัยนี้ใช้วิธีการแบ่งกลุ่มแบบมีลำดับชั้น ด้วยอัลกอริทึม Self Organization Map (SOM) ซึ่งจะกล่าวถึงรายละเอียดในส่วนถัดไป

8.1 พื้นฐานวิธีการแบ่งกลุ่มแบบมีลำดับชั้น (Hierarchical methods)

วิธีการแบ่งกลุ่มแบบมีลำดับชั้นนี้ มีลักษณะการทำงานคล้ายโครงสร้างต้นไม้ ซึ่งแบ่งออกเป็นสองประเภทตามลำดับการทำงานคือ การแบ่งกลุ่มแบบล่างขึ้นบน (agglomerative hierarchical clustering) และการแบ่งกลุ่มจากบนลงล่าง (divisive hierarchical clustering)

การแบ่งกลุ่มแบบล่างขึ้นบน (agglomerative hierarchical clustering) เริ่มด้วยการให้ข้อมูลทุกตัวอยู่ต่างกลุ่มกัน ในแต่ละลำดับชั้นจะทำการรวมกลุ่มที่มีความคล้ายคลึงกันมากที่สุดเข้าด้วยกันเกิดเป็นกลุ่มข้อมูลที่ใหญ่ขึ้นเรื่อยๆ

การแบ่งกลุ่มจากบนลงล่าง (divisive hierarchical clustering) มีแนวทางสลับกับแบบล่างขึ้นบน โดยเริ่มจากการให้ข้อมูลทุกตัวอยู่ในกลุ่มเดียวกัน จากนั้นจึงหาคู่ของข้อมูลที่มีความแตกต่างกันมากที่สุดภายในกลุ่ม แล้วจึงแยกข้อมูลทั้งคู่ให้อยู่ต่างกลุ่มกัน ข้อมูลที่เหลือจะถูกจัดให้อยู่ในกลุ่มที่ให้ความคล้ายคลึงมากที่สุด กลุ่มข้อมูลที่ได้ในแต่ละลำดับชั้นจะมีขนาดเล็กลงเรื่อยๆ



ภาพที่ 25 แสดงลักษณะการแบ่งกลุ่มข้อมูลแบบมีลำดับชั้น

ข้อดีของการแบ่งกลุ่มด้วยแบบลำดับชั้นคือ ไม่ต้องกำหนดจำนวนกลุ่มที่ตั้งแต่เริ่มทำงานเราสามารถหยุดการทำงานได้เมื่อเห็นว่าได้คุณภาพ หรือจำนวนกลุ่มตามที่ต้องการแล้ว และยังให้ความถูกต้องสูงเนื่องจากในแยกกลุ่มหรือรวมกลุ่มจะต้องมีการคำนวณค่าความคล้ายคลึงระหว่างกลุ่มย่อยทุกคู่ที่เป็นไปได้ แล้วจึงเลือกคู่ของกลุ่มที่ดีที่สุดในการรวมหรือแยกกลุ่ม แต่มีผลทำให้ใช้เวลาในการทำงานสูงขึ้นตามไปด้วย

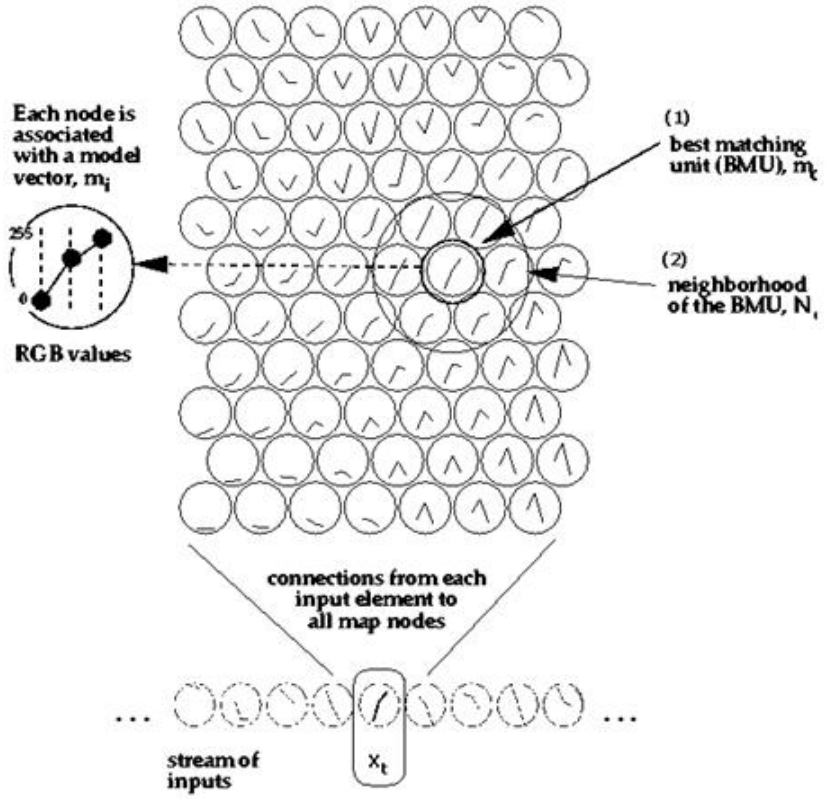
8.2 Self Organizing Map (SOM)

SOM (Kohonen, 1982) เป็นเทคนิคหนึ่งทางด้านโครงข่ายประสาทเทียม (artificial neuron network) เดิมทีถูกใช้ในขอบข่ายของงานการรู้จำเสียง (speech recognition) ต่อมาได้นำมาใช้ในงานจัดกลุ่มข้อมูล (clustering) ที่มีความซับซ้อนหรือหลายมิติ ผลลัพธ์ที่ได้จากการจัดจัดกลุ่มแสดงในลักษณะของกริด (grid) หลักการของ SOM นั้น จะทำการลดมิติหรือลดความซับซ้อนของข้อมูลหลายมิติให้อยู่ในรูปที่เข้าใจได้ง่าย เช่น 1 หรือ 2 มิติ โดยแสดงด้วยค่าความคล้ายคลึงกันระหว่างกลุ่มข้อมูล

สมมติให้กลุ่มตัวอย่างข้อมูลสีเป็นดังภาพที่ 26 สามารถเขียนแทนได้ด้วยเวคเตอร์จำนวนจริง (real vector) $x(t) \in R^n$ โดยที่ t คืออินเด็กซ์ของสมาชิกกลุ่มตัวอย่าง โหนด i แต่ละโหนดบนแมป (map) ประกอบด้วยเวคเตอร์โมเดล (vector model) $m_i(t) \in R^n$ ซึ่งมีจำนวนแอทริบิวต์ของค่าถ่วงน้ำหนัก (weight) เท่ากับจำนวนฟีเจอร์แอทริบิวต์ของอินพุต $x(t)$ แต่ละตัว สังเกตว่าจำนวนแมปโหนดก็คือจำนวนกลุ่มเอาท์พุตมากที่สุดนั่นเอง (จากภาพที่ 26 ข้อมูลอินพุตมี 3 ฟีเจอร์แอทริบิวต์ คือสีแดง-สีเขียว-สีน้ำเงิน) โครงสร้างพื้นฐานของ SOM แสดงดังภาพที่ 27

250	235	215	antique white
165	042	042	brown
222	184	135	burlywood
210	105	30	chocolate
255	127	80	coral
184	134	11	dark goldenrod
189	183	107	dark khaki
255	140		dark orange
233	150	122	dark salmon
...	...	...	...

ภาพที่ 26 แสดงกลุ่มข้อมูลตัวอย่างของแม่สี



ภาพที่ 27 แสดงโครงสร้างพื้นฐานของ SOM โดยแต่ละโหนด (วงกลมเรียงต่อกันบนกริด) เรียกว่า แมปโหนด (map node) ซึ่งมีจำนวนแอทริบิวต์ของค่าถ่วงน้ำหนัก (weight) เท่ากับฟีเจอร์แอทริบิวต์ของอินพุต ส่วนบริเวณหมายเลข (1) คือโหนดที่มีค่าใกล้เคียงกับโหนดอินพุต มากที่สุด (best matching unit, BMU) เรียกว่าวินนิ่งโหนด (wining node) และบริเวณที่ (2) คือโหนดเพื่อนบ้านของวินนิ่งโหนด

ที่มา: mlab.taik.fi/~timo/som/thesis-som.html (2009)

อัลกอริทึมพื้นฐานของ SOM มีอยู่ 5 ขั้นตอนได้แก่

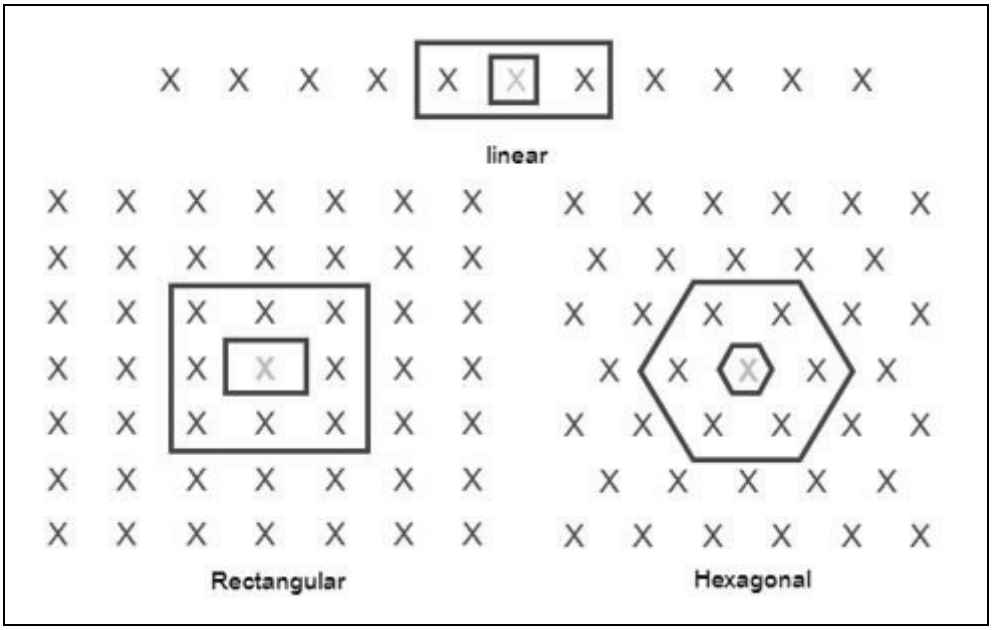
ขั้นตอนที่หนึ่ง กำหนดค่าเริ่มต้นให้กับแอทริบิวต์ของค่าถ่วงน้ำหนัก (weight) ของทุกแมปโหนดด้วยวิธีการสุ่ม (random)

ขั้นตอนที่สอง หาวินนิ่งโหนดสำหรับอินพุต $x(t)$ โดยทำการเปรียบเทียบค่าระยะห่างระหว่างอินพุต $x(t)$ กับแมปโหนด m_i ทุกโหนด ด้วยสมการยูคลีเดียนดิสแทน (Euclidean distance) และเลือกโหนดที่มีระยะห่างน้อยสุดเป็นวินนิ่งโหนด เมื่อกำหนดให้เมตริกซ์อินพุตแทน

ด้วย $X = [x_1, x_2, \ldots, x_k]$ เมื่อ k คือจำนวนอินพุตทั้งหมด ส่วนเมตริกซ์ค่าถ่วงน้ำหนักของแมปโหนดแทนด้วย $M = [m_1, m_2, \ldots, m_i]$ เมื่อ i คือจำนวนแมปโหนดทั้งหมด $(i = 1,2,\ldots,n)$ ดังนั้นวินนิ่งโหนดหาได้จากสมการดังนี้

$$\text{Wining node} = \text{MIN}\|x_t - m_i\|, \quad (t = 1,2,\ldots,k) \qquad \textbf{(13)}$$

ขั้นตอนที่สาม หาโหนดเพื่อนบ้าน (neighbor node) ของวินนิ่งโหนดว่ามีโหนดใดบ้าง โดยทั่วไปโครงสร้างแมปนั้นมีอยู่ด้วยกัน 3 แบบได้แก่ แบบเชิงเส้น (linear) แบบสีเหลี่ยม (rectangular) และแบบหกเหลี่ยม (hexagonal) ซึ่งจะมีผลต่อจำนวนของโหนดเพื่อนบ้านเช่น โครงสร้างแบบเชิงเส้น จะมีโหนดเพื่อนบ้าน 2 โหนด ในขณะที่โครงสร้างแบบสี่เหลี่ยม จะมีโหนดเพื่อนบ้าน 8 โหนด และโครงสร้างแบบหกเหลี่ยมจะมีโหนดเพื่อนบ้าน 6 โหนด เป็นต้น แสดงดังภาพที่ 28



ภาพที่ 28 แสดงโครงสร้างทั่วไปของ SOM โดยด้านบนคือโครงสร้างแบบเชิงเส้น (linear) ด้านล่างซ้ายคือโครงสร้างแบบสี่เหลี่ยม และด้านล่างขวาคือโครงสร้างแบบหกเหลี่ยม

ที่มา: www.cs.buu.ac.th/~prawit/ai (2009)

ขั้นตอนที่สี่ ทำการปรับค่าถ่วงน้ำหนักของทุกแอทริบิวต์ในแมปโหนดของวินนิ่งโหนด และโหนดเพื่อนบ้าน ด้วยสมการดังต่อไปนี้

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)] \quad \textbf{(14)}$$

โดยที่ $h_{ci}(t)$ คือ neighbourhood function ดังสมการต่อไปนี้

$$h_{ci} = \alpha(t) \cdot \exp(-\frac{\|r_c - r_i\|^2}{2\sigma^2(t)} \quad \textbf{(15)}$$

โดยที่ ค่า $\alpha(t)$ คือ learning rate function ค่า $\sigma(t)$ คือ เคอร์เนลฟังก์ชั่น (kernel width function) และ $\|r_c - r_i\|^2$ คือระยะห่างระหว่างวินนิ่งโหนด c และโหนดที่กำลังพิจารณา i

ขั้นตอนที่ห้า ทำซ้ำในขั้นตอนที่สอง สำหรับอินพุตใหม่ $x_{(t+1)}$

ข้อดีของการจัดกลุ่มข้อมูลด้วย SOM คือโครงสร้างและขั้นตอนการทำงานของ SOM นั้นไม่ซับซ้อน สามารถทำความเข้าใจได้โดยง่าย และสามารถจัดกลุ่มข้อมูลที่มีความซับซ้อนหรือมีหลายมิติได้ดี

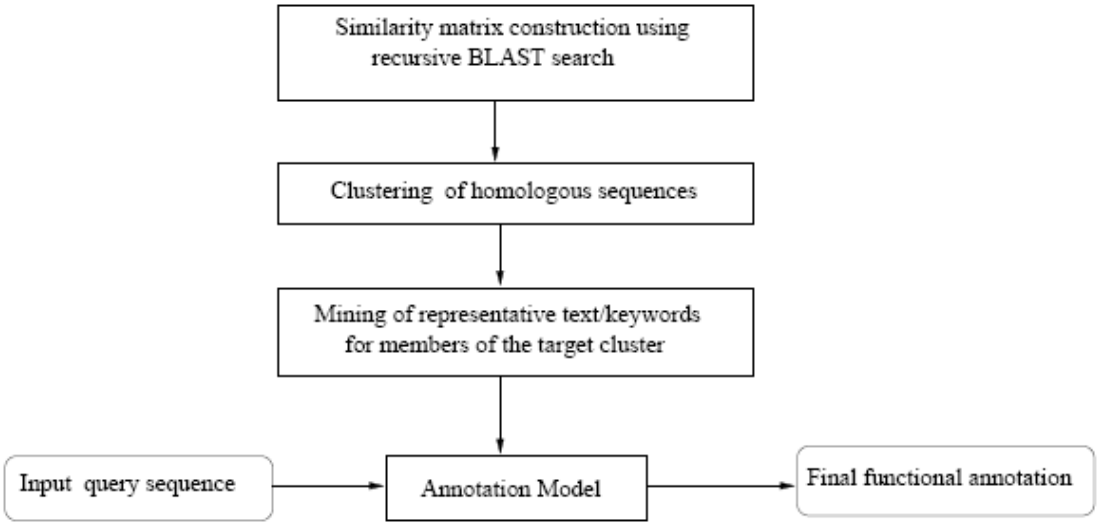
ข้อเสียของการจัดกลุ่มข้อมูลด้วย SOM คือการทำงานค่อนข้างช้า เนื่องจากในขั้นตอนการหาวินนิ่งโหนดนั้น จำเป็นต้องทำการคำนวณระยะห่างระหว่างอินพุตกับแมปโหนดทุกโหนด ดังนั้นยิ่งจำนวนแมปโหนดยิ่งมากจำทำให้ใช้เวลานานขึ้น

งานวิจัยที่เกี่ยวข้อง

ในวิทยานิพนธ์ฉบับนี้ ได้จำแนกงานวิจัยที่เกี่ยวข้องกับระบบจำแนกประเภทเอนไซม์เป็นกลุ่มหลัก 3 กลุ่มตามวิธีการที่ใช้ในงานวิจัยได้แก่ (1) การเปรียบเทียบความเหมือนของลำดับกรดอะมิโนในสายโปรตีน (homology-based approach) (2) การพิจารณาส่วนย่อยของสายโปรตีน (subsequence-based approach) และ (3) การสร้างตัวแทนสายโปรตีนหรือฟีเจอร์ (feature-based approach) ซึ่งมีรายละเอียดดังต่อไปนี้

1. ระบบจำแนกประเภทเอนไซม์โดยเปรียบเทียบความเหมือนของลำดับกรดอะมิโนในสายโปรตีน (Homology-based approach)

สมมติฐานของวิธีการนี้คือ เอนไซม์ที่มีลำดับของกรดอะมิโนหรือโครงสร้างปฐมภูมิที่คล้ายคลึงกัน (homology) จะมีฟังก์ชั่นที่เหมือนกัน วิธีการทำนายฟังก์ชั่นลักษณะนี้ เริ่มจากมีเอนไซม์ต้นแบบที่ทราบฟังก์ชั่นแล้ว จากนั้นทำการพิจารณาเอนไซม์ใหม่ที่ต้องการทราบฟังก์ชั่นว่ามีโครงสร้างปฐมภูมิคล้ายคลึงกับเอนไซม์ต้นแบบตัวใดมากที่สุด ก็จะทำนายได้ว่ามีฟังก์ชั่นการทำงานเหมือนกับเอนไซม์ต้นแบบตัวนั้นนั่นเอง สำหรับวิธีการพิจารณาความคล้ายคลึงของโครงสร้างปฐมภูมินั้นใช้วิธีการที่เป็นที่นิยมเช่น BLAST (Altschul *et al.*, 1990) PSI-BLAST (Altschul *et al.*, 1997) FASTA (Lipman and Pearson, 1985) และ HMM (Eddy, 1998) เป็นต้น ตัวอย่างงานวิจัยที่ใช้วิธีการนี้ได้แก่ GeneQuiz (Andrade, 1999) PEDANT (Riley *et al.*, 2005) และ Auto-FACT (Koski *et al.*, 2005) เป็นต้น นอกจากนี้ได้มีการพัฒนาวิธีการข้างต้น (Xie *et al.*, 2002; Abascal and Valencia, 2003; Sasson *et al.*, 2006) โดยพยายามจัดกลุ่มเอนไซม์ต้นแบบที่มีค่าความเหมือนใกล้เคียงกัน จากนั้นการสืบค้นตัวแทนกลุ่ม และนำไปสร้างเป็นโมเดลทำนายฟังก์ชั่น สำหรับในการทำนายฟังก์ชั่นเอนไซม์ทำได้โดย นำเอนไซม์ที่ไม่ทราบฟังก์ชั่นนั้นไปสืบค้นในโมเดลและถ้ามีบางส่วนตรงกับตัวแทนกลุ่มใดของเอนไซม์ต้นแบบ จะได้ว่ามีฟังก์ชั่นการทำงานตามเอนไซม์กลุ่มต้นแบบนั้น ดังภาพที่ 29



ภาพที่ 29 แสดงไดอะแกรมวิธีการสร้างโมเดลทำนายฟังก์ชั่นด้วยวิธีการ homology-based ที่มีการปรับเปลี่ยนวิธีการบางส่วน

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

ข้อดีของการทำนายฟังก์ชั่นของเอนไซม์ด้วยวิธีการนี้คือเป็นวิธีการที่ง่ายและให้ความถูกต้องสูงสำหรับสายโปรตีนที่มีค่าความเหมือนระหว่างสายโปรตีนสูงๆ (pairwise sequence similarity) ยกตัวอย่างเช่นในงานวิจัยของ Tian et *al.* (2004) กล่าวว่าสายโปรตีนที่ไม่ทราบฟังก์ชั่นที่มีค่าความเหมือนกันระหว่างสายโปรตีนต้นแบบ เฉลี่ยประมาณ 60% นั้น จะให้ค่าความถูกต้องของการทำนายฟังก์ชั่นได้สูงถึง 90% เป็นต้น

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

2. ระบบจำแนกประเภทเอนไซม์โดยพิจารณาส่วนย่อยของสายโปรตีน (Subsequence-based approach)

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

จากข้างต้นจะเห็นได้ว่า โมทีฟ หรือโดเมน นั้นเป็นส่วนที่สำคัญที่สามารถบ่งชี้ฟังก์ชั่นการทำงานของเอนไซม์ได้ งานวิจัยที่ใช้วิธีการนี้ได้แก่งานวิจัยของ Hannenhalli and Russell (2000) ได้พัฒนาวิธีการสืบค้นและระบุบริเวณส่วนย่อยของสายโปรตีนที่สามารถแบ่งแยกคลาสของเอนไซม์ได้ดีที่สุด โดยเริ่มจากทำการระบุตำแหน่งของบริเวณอนุรักษ์ ซึ่งได้จากการเทียบเรียงสายโปรตีนทุกสาย (multiple sequence alignment) ในกลุ่มแฟมิลีเดียวกัน จากนั้นทำการคำนวณค่าความสัมพันธ์ของเอ็นโทรปี (relative entropy) ในแต่ละตำแหน่งของกรดอะมิโนที่อยู่ในแฟมิลีเดียวกัน บริเวณที่มีค่าความสัมพันธ์ของเอ็นโทรปีรวมกันมากที่สุด จะเป็นบริเวณที่สามารถแบ่งแยกคลาสของเอนไซม์ได้ดีที่สุด จากนั้นนำมาเป็นฟีเจอร์ให้กับฮิดเด้นมาร์คอฟโมเดลเพื่อสร้างเป็นโมเดลจำแนกประเภทเอนไซม์ งานวิจัยของ Blekas *et al.* (2005) ใช้โมทีฟจากฐานข้อมูล PROSITE (Hulo, 2006) ซึ่งได้จากการทดสอบโดยผู้เชี่ยวชาญ และโมทีฟที่ได้จากการสืบค้นด้วยอัลกอริทึม MEME (Bailey *et al.*, 1999) เป็นฟีเจอร์ในการสร้างโมเดลจำแนกประเภทเอนไซม์ด้วยนิวรอน เน็ทเวิร์ค (neuron network) งานวิจัยของ Ben-Hur and Brutlag (2005) ใช้การพิจารณาความถี่ของการปรากฏโมทีฟแต่ละตัวในสายโปรตีนเป็นตัวแทนสายโปรตีนและสร้างเป็นฟีเจอร์เวคเตอร์นำเข้าอัลกอริทึมวิธีเรียนรู้ ซัพพอร์ตเวคเตอร์แมชชีน งานวิจัยของ Schug *et al.* (2002) เป็นงานวิจัยแรกที่เริ่มมีการนำโดเมนมาใช้ในการทำนายฟังก์ชั่น ในงานวิจัยนี้ทำการสกัดโดเมนจาก

ฐานข้อมูลโปรดอม (ProDom) (Servant *et al.*, 2002) และฐานข้อมูลซีดีดี (Conserved Domain Database, CDD) (Marchler-Bauer *et al.* 2005) จากนั้นสร้างโมเดลโดยใช้การสืบค้นด้วยอัลกอริทึม BLAST งานวิจัย WILMA (Prlic *et al.*, 2004) ใช้โดเมนจากฐานข้อมูล PROSITE และ Pfam (Sonnhammer *et al.*, 1997) และ PRINTS (Attwood *et al.*, 2003) ส่วนวิธีการทำนายฟังก์ชั่นใช้วิธีการสืบค้นร่วมกันระหว่างหลายอัลกอริทึม (ensemble) ได้แก่ RPS-BLAST (Altschul *et al.* 1997) PROSITE scans และ Finger-PRINTScan (Scordis *et al.*, 1999) เป็นต้น

ข้อดีของการทำนายฟังก์ชั่นของเอนไซม์ด้วยวิธีการนี้คือส่วนย่อยของลำดับกรดอะมิโนที่ได้จากโมทีฟหรือโดเมนนั้น มีความสัมพันธ์กับโครงสร้างโปรตีนที่ทำให้เกิดฟังก์ชั่นการทำงานเฉพาะอย่าง ซึ่งเป็นบริเวณที่สามารถบ่งบอกฟังก์ชั่นของเอนไซม์ได้เป็นอย่างดี

ข้อเสียของการทำนายฟังก์ชั่นของเอนไซม์ด้วยวิธีการนี้คือการนิยามส่วยย่อยของลำดับกรดอะมิโนว่าเป็นโมทีฟหรือโดเมนนั้นยังมีความคลาดเคลื่อนอยู่ และการใช้โมทีฟหรือโดเมนเป็นตัวแทนสายโปรตีนนั้น ในบางกรณีอาจยังไม่เพียงพอที่จะใช้เป็นตัวแทนสายโปรตีนทั้งเส้นได้

3. ระบบจำแนกประเภทเอนไซม์โดยการสร้างตัวแทนสายโปรตีนหรือฟีเจอร์ (Feature-based approach)

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

งานวิจัยที่ใช้วิธีการสร้างตัวแทนสายโปรตีนหรือฟีเจอร์ได้แก่ งานวิจัย SVM-Prot (Cai *et al.*, 2003) ทำการสร้างตัวแทนสายโปรตีนโดยพิจารณาคุณสมบัติของกรดอะมิโน เช่นแรงวัลเดอร์วาวล์ การมีขั้วไม่มีขั้ว การมีประจุไม่มีประจุ และแรงตึงผิวของกรดอะมิโน งานวิจัย CD (Chou and Elrod, 2003) ทำการพิจารณาความถี่ของการปรากฏของลำดับกรดอะมิโนในสายโปรตีน งานวิจัย CDA (Chou, 2005) เป็นการพิจารณาคุณสมัติทางเคมีร่วมกับพิจารณาความถี่ของการปรากฏของกรดอะมิโน และสร้างโมเดลจำแนกประเภทด้วยโคแวเรียนดิสคริมิแนนท์ งานวิจัยของ Zhou *et al.*

(2007) ใช้วิธีการสกัดฟีเจอร์เช่นเดียวกับงานวิจัย CDA แต่ต่างกันตรงที่สร้างโมเดลจำแนกประเภทด้วยอัลกอริทึมเรียนรู้ ซัพพอร์ตเวคเตอร์แมชชีน งานวิจัยของ King *et al.* (2000) สร้างฟีเจอร์เวคเตอร์จากชุดของกรดอะมิโนที่เกิดขึ้นบ่อย โดยสร้างโมเดลจำแนกประเภทด้วยอัลกอริทึม C4.5 เป็นต้น

สำหรับงานวิจัยที่ใช้ในการเปรียบเทียบเพื่อวัดผลการทดลองและประสิทธิภาพกับงานวิจัยในวิทยานิพนธ์ฉบับนี้ได้แก่ งานวิจัย CD และ CDA ซึ่งมีรายละเอียดดังต่อไปนี้

งานวิจัย Covariant Discriminant Algorithm (CD)

งานวิจัย CD นี้ มีวัตถุประสงค์เพื่อสร้างระบบจำแนกประเภทเอนไซม์ซับแฟมิลี่ โดยพิจารณาความถี่ในการปรากฏขึ้นของกรดอะมิโนหนึ่งๆ บนสายโปรตีน (amino acid composition) เปรียบเทียบกับความถี่ในการปรากฏขึ้นของกรดอะมิโนหนึ่งๆ ของทุกสายโปรตีนที่อยู่ในคลาสนั้น โดยวิธีการเปรียบเทียบความเหมือนนี้ ใช้ฟังก์ชั่นวัดระยะห่างที่เรียกว่า โคแวเรียนต์ ดิสคริมิแนนท์ ฟังก์ชั่น (covariant discriminant function) ซึ่งมีที่มาจากสูตรการคำนวณระยะห่างระหว่างข้อมูลคือ squared Mahalanobis distance (Mahalanobis, 1936) ดังรายละเอียดต่อไปนี้

สมมติให้โปรตีนตัวที่ k ในคลาส m สามารถเขียนให้อยู่ในรูปเวคเตอร์ได้ดังสมการต่อไปนี้

$$\mathrm{E}_k^m = \begin{bmatrix} e_{k,1}^m \\ e_{k,2}^m \\ \vdots \\ e_{k,20}^m \end{bmatrix}, \quad (k = 1,2,\ldots,n_m; m = 1,2,\ldots,16) \qquad \textbf{(16)}$$

โดยที่ $e_{k,1}^m, e_{k,2}^m, \ldots, e_{k,20}^m$ คือความถี่ในการปรากฏขึ้นของกรดอะมิโนหนึ่งๆ (amino acid composition) บนสายโปรตีน k ของคลาส m และ n_m คือจำนวนโปรตีนทั้งหมดในคลาส m สำหรับเวคเตอร์มาตรฐาน (standard vector) ของคลาส m สามารถนิยามได้ดังสมการต่อไปนี้

$$\overline{\mathrm{E}}^{m} = \begin{bmatrix} \overline{e}_1^{m} \\ \overline{e}_2^{m} \\ \vdots \\ \overline{e}_{20}^{m} \end{bmatrix}, \quad (m = 1,2,\ldots,16) \tag{17}$$

โดยที่

$$\overline{e}_i^{m} = \frac{1}{n}\sum_{k=1}^{n_m} e_{k,i}^{m}, \quad (i = 1,2,\ldots,20) \tag{18}$$

สมมติให้ E คือโปรตีนที่กำลังพิจารณาว่าอยู่คลาสใด เราสามารถสร้างตัวแทนสายโปรตีนซึ่งเกิดจากการคำนวณหาค่าความถี่ในการปรากฏขึ้นของกรดอะมิโนหนึ่งๆ ในสายโปรตีน E ให้อยู่ในรูปของเวคเตอร์ขนาด 20 มิติได้จากสมการที่ (16) สำหรับค่าความต่างระหว่างโปรตีนที่กำลังพิจารณา E กับค่ามาตรฐาน (norm) ของข้อมูลในคลาส m สามารถคำนวณได้โดยใช้ฟังก์ชั่นโคแวเรียนต์ ดิสครีมิแนนท์ (covariant discriminant function) ดังสมการต่อไปนี้

$$\Delta(\mathrm{E}, \overline{\mathrm{E}}^{m}) = D_M^2(\mathrm{E}, \overline{\mathrm{E}}^{m}) + \ln|S_m|, \quad (m = 1,2,\ldots,16) \tag{19}$$

โดยที่ $D_M^2(\mathrm{E}, \overline{\mathrm{E}}^{m}) = (\mathrm{E} - \overline{\mathrm{E}}^{m})^{\mathrm{T}} S_m^{-1} (\mathrm{E} - \overline{\mathrm{E}}^{m})$

จากสมการที่ (19) $D_M^2(\mathrm{E}, \overline{\mathrm{E}}^{m})$ คือสูตรการคำนวณระยะห่างระหว่างข้อมูลของ squared Mahalanobis distance T คือโอเปอร์เรเตอร์การทรานสโพสเมตริกซ์ (transposition operator) S_m และ S_m^{-1} คือเมตริกซ์ดีเทอร์มิแนนท์ และอินเวิร์สเมตริกซ์ดีเทอร์มิแนนท์ ตามลำดับ ซึ่ง S_m ก็คือโคแวเรียนต์เมตริกซ์ของคลาส m นั่นเอง สามารถเขียนได้ดังสมการต่อไปนี้

$$S_m = \begin{bmatrix} s_{1,1}^{m} & s_{1,2}^{m} & \cdots & s_{1,20}^{m} \\ s_{2,1}^{m} & s_{2,2}^{m} & \cdots & s_{2,20}^{m} \\ \vdots & \vdots & \ddots & \vdots \\ s_{20,1}^{m} & s_{20,2}^{m} & \cdots & s_{20,20}^{m} \end{bmatrix} \tag{20}$$

โดยที่สมาชิกของเมตริกซ์แต่ละค่าคือค่าการกระจายตัวของข้อมูลแต่ละคลาส (standard deviation) เพื่อทำการ normalize ข้อมูลให้อยู่ในสเกลเดียวกัน หาได้จาก

$$s_{i,j}^{m} = \frac{1}{n_m - 1}\sum_{k=1}^{n_m}[e_{k,i}^{m} - e_i^{-m}][e_{k,j}^{m} - e_j^{-m}], \quad (i, j = 1,2,\ldots,20) \qquad \textbf{(21)}$$

ดังนั้นแบบจำลองเพื่อจำแนกประเภทเอนไซม์ซับแฟมิลี่นี้ จะให้ผลลัพธ์เป็นคลาส m ก็ต่อเมื่อค่าความแตกต่างระหว่างโปรตีนที่กำลังพิจารณา E กับค่ามาตรฐานของข้อมูลในคลาส m มีค่าน้อยที่สุด ดังสมการต่อไปนี้

$$\Delta(\mathrm{E}, \overline{\mathrm{E}}^{u}) = \mathrm{MIN}\{\Delta(\mathrm{E}, \overline{\mathrm{E}}^{1}), \Delta(\mathrm{E}, \overline{\mathrm{E}}^{2}), \ldots, \Delta(\mathrm{E}, \overline{\mathrm{E}}^{16})\} \qquad \textbf{(22)}$$

โดยที่ μ มีค่าเป็น 1, 2, 3, …, 16 และโอเปอร์เรเตอร์ MIN คือค่าน้อยสุดของค่าที่อยู่ในวงเล็บ

การใช้ตัวแทนสายข้อมูลโปรตีนในงานวิจัย CD นี้ใช้การพิจารณาเพียงค่าความถี่ในการปรากฏของกรดอะมิโนหนึ่งๆ บนสายโปรตีนเท่านั้น ซึ่งยังขาดข้อมูลในเรื่องลำดับของกรดอะมิโน ซึ่งเป็นข้อมูลหนึ่งที่มีความสำคัญในการเกิดรูปร่างโครงสร้าง 3 มิติ ซึ่งเกี่ยวข้องกับการเกิดฟังก์ชั่นของเอนไซม์ นอกจากนี้ผลจากระบบจำแนกข้อมูลเอนไซม์ที่ได้นั้น ไม่สามารถบ่งบอกโครงสร้างในระดับทุติยภูมิได้

งานวิจัย Covariant Discriminant algorithm using Amphiphilic pseudo amino acid composition (Covariant-discirminant and Am-Pse-AA Composition, CDA)

งานวิจัย CDA นี้ พัฒนามาจากงานวิจัย CD โดยทำการปรับปรุงตัวแทนสายโปรตีน มีวัตถุประสงค์เพื่อสร้างตัวแทนสายข้อมูลโปรตีนที่อยู่ในรูปแบบดิสครีท (discrete) ซึ่งง่ายต่อการนำไปใช้เป็นข้อมูลอินพุตเพื่อสร้างระบบจำแนกประเภทข้อมูล และข้อมูลตัวแทนสายโปรตีนนั้นยังคงสามารถแสดงข้อมูลลำดับกรดอะมิโนได้อย่างเหมาะสม ดังนั้นตัวแทนสายข้อมูลโปรตีนที่สร้างขึ้นนี้ จึงใช้ข้อมูลความถี่ของการปรากฏขึ้นของกรดอะมิโนหนึ่งๆ บนสายโปรตีน ร่วมกับข้อมูลความถี่ของการเปลี่ยนคุณสมบัติทางเคมีภายในลำดับกรดอะมิโน โดยคุณสมบัติทางเคมีที่

พิจารณาได้แก่คุณสมบัติไฮโดรโฟบิกซิตี้ (hydrophobicity) และคุณสมบัติไฮโดรฟิลิกซิตี้ (hydrophilicity) เนื่องจากคุณสมบัติทางเคมีทั้งสองนี้มีบทบาทสำคัญในการเกิดรูปแบบโครงสร้าง 3 มิติของสายโปรตีน ซึ่งสารประกอบอินทรีย์ใดที่มีคุณสมบัติทั้งไฮโดรโฟบิกซิตี้และไฮโดรฟิลิกซิตี้ จะถูกเรียกว่ามีคุณสมบัติแอมฟีฟิลิกซ์นั่นเอง (amphiphilic) จากนั้นจึงนำตัวแทนสายโปรตีนที่ได้ เข้าเป็นอินพุตเพื่อสร้างระบบจำแนกประเภทเอนไซม์ซับแฟมิลี่ด้วยวิธีการคำนวณโคแวเรียนท์ดิสคริมิแนนท์ฟังก์ชั่น เช่นในงานวิจัย CD ตามลำดับ

โดยทั่วไปแล้วสามารถสร้างตัวแทนสายของโปรตีนได้ 2 แบบคือ แบบดิสครีท (Discrete) และแบบซีเควนเชียล (sequential) โดยแต่ละแบบนั้นมีข้อดีและข้อเสียแตกต่างกันคือ

แบบดิสครีท (Discrete) สายโปรตีนจะถูกแทนด้วยกลุ่มของตัวเลขที่ไม่ต่อเนื่อง เช่น 0, 1, 2 เป็นต้น หรืออยู่ในรูปของเวคเตอร์หลายมิติ (multiple dimension vector) เช่นอะมิโนแอซิดคอมโพสิชัน (amino acid composition) ซึ่งเป็นวิธีการที่ใช้กันอย่างแพร่หลาย

ข้อดีของตัวแทนสายโปรตีนแบบดิสครีทคือ สามารถสร้างข้อมูลเพื่อเป็นอินพุตเพื่อสร้างแบบจำลองการจำแนกประเภทข้อมูลเอนไซม์ซับแฟมิลี่ได้ง่าย

ข้อเสียของตัวแทนสายโปรตีนแบบดิสครีทคือ ข้อมูลที่ได้นั้นไม่สามารถบ่งบอกลำดับของกรดอะมิโนในสายโปรตีนได้

แบบซีเควนเชียล (sequential) สายโปรตีนจะถูกแทนด้วยการเรียงตัวของกรดอะมิโนตามตำแหน่งและลำดับบนสายโปรตีน เช่นข้อมูลสายลำดับโปรตีนในรูปแบบปฐมภูมิ ดังภาพที่ 30

```
MFAKTAAANLTKKGGLSLLSTTARRTKVTLPDLKWDFGALEPYISGQINELHYTKHHQTY
VNGFNTAVDQFQELSDLLAKEPSPANARKMIAIQQNIKFHGGGFTNHCLFWENLAPESQG
GGEPPTGALAKAIDEQFGSLDELIKLTNTKLAGVQGSGWAFIVKNLSNGGKLDVVQTYNQ
DTVTGPLVPLVAIDAWEHAYYLQYQNKKADYFKAIWNVVNWKEASRRFDAGKI
```

ภาพที่ 30 แสดงรูปแบบตัวแทนสายโปรตีนแบบซีเควนเชียล

ข้อดีของตัวแทนสายโปรตีนแบบซีเควนเชียลคือ ข้อมูลที่ได้สามารถแสดงลำดับและตำแหน่งของกรดอะมิโน และความยาวของสายโปรตีนได้

ข้อเสียของตัวแทนสายโปรตีนแบบซีเควนเชียลคือ ยากต่อการนำไปสร้างเป็นข้อมูลอินพุตเพื่อสร้างแบบจำลองการจำแนกประเภทข้อมูลเอนไซม์ซับแฟมิลี่

สำหรับขั้นตอนการสร้างตัวแทนสายข้อมูลโปรตีนนั้นมีรายละเอียดดังต่อไปนี้

สมมติให้โปรตีน P เกิดจากการเรียงตัวของกรดอะมิโน R ความยาว L แสดงเป็น $R_1R_2R_3R_4R_5R_6R_7...R_L$ โดยที่ R_1 คือกรดอะมิโนตำแหน่งที่ 1 R_2 คือกรดอะมิโนตำแหน่งที่ 2 และ R_L คือกรดอะมิโนตำแหน่งที่ L เป็นต้น

ไฮโดรโฟบิกซิตี้และไฮโดรฟิลิกซิตี้ เป็นคุณสมบัติทางเคมีที่เป็นส่วนประกอบหลักและสำคัญของกรดอะมิโน เนื่องจากมีบทบาทสำคัญในการหดพับและเกิดเป็นโครงสร้าง 3 มิติ ยกตัวอย่างเช่น โครงสร้างฮีลิกซ์ (helices) ที่มีลักษณะการม้วนตัวเป็นเกลียว เกิดจากกรดอะมิโนที่มีคุณสมบัติไฮโดรโฟบิกซิตี้และไฮโดรฟิลิกซิตี้นั่นเอง (Chou *et al.*, 1997) สำหรับลำดับของกรดอะมิโนในสายโปรตีนนั้นสามารถแทนได้ด้วยสมการ 23 ดังนี้

$$\begin{cases} \tau_1 = \frac{1}{L-1}\sum_{i=1}^{L-1} H_{i,i+1}^1 \\ \tau_2 = \frac{1}{L-1}\sum_{i=1}^{L-1} H_{i,i+1}^2 \\ \tau_3 = \frac{1}{L-2}\sum_{i=1}^{L-2} H_{i,i+2}^1 \\ \tau_4 = \frac{1}{L-2}\sum_{i=1}^{L-2} H_{i,i+2}^2 \quad , \quad \lambda < L \\ \quad \cdots\cdots\cdots\cdots \\ \tau_{2\lambda-1} = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda} H_{i,i+\lambda}^1 \\ \tau_{2\lambda} = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda} H_{i,i+\lambda}^2 \end{cases} \qquad \textbf{(23)}$$

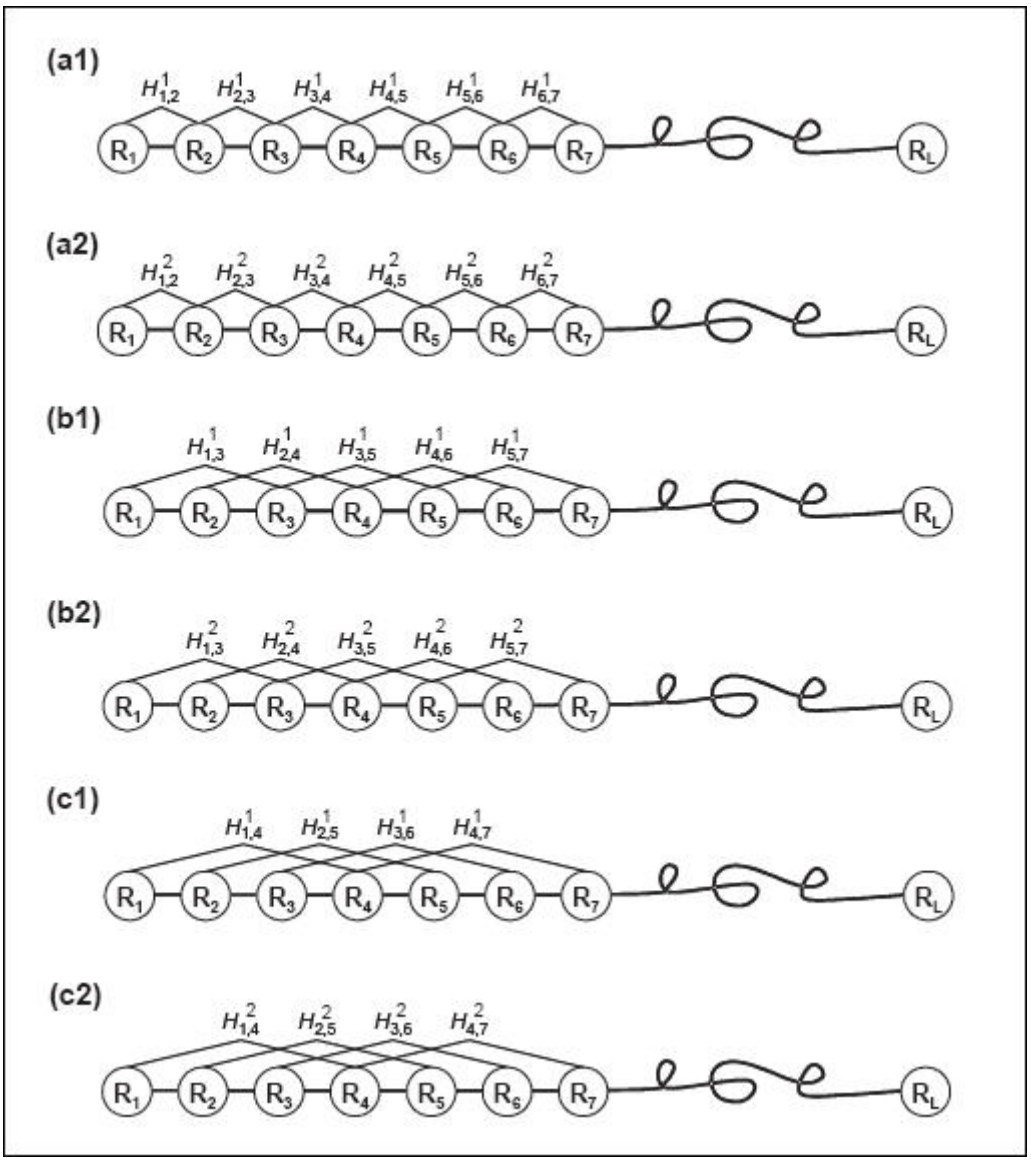
โดยที่ $H_{i,j}^1$ และ $H_{i,j}^2$ คือฟังก์ชั่นความสัมพันธ์ (correlation function) ของไฮโดรโฟบิกซิตี้และไฮโดรฟิลิกซิตี้ ดังสมการที่ (24)

$$H^1_{i,j} = h^1(R_i) \cdot h^1(R_j)$$
$$H^2_{i,j} = h^2(R_i) \cdot h^2(R_j) \quad \textbf{(24)}$$

โดยที่ $h^1(R_i)$ และ $h^2(R_i)$ คือค่าของไฮโดรโฟบิกซิตี้และไฮโดรฟิลิกซิตี้ของกรดอะมิโนตำแหน่งที่ i $(i = 1,2,...,L)$ สำหรับสัญลักษณ์ ดอท $(\cdot)$ คือการคูณ ส่วนสัญลักษณ์ τ_1 และ τ_2 ในสมการที่ (23) เรียกว่าแฟคเตอร์ความสัมพันธ์ลำดับชั้นที่ 1 (The first-tier correlation factors) ซึ่งเป็นสมการที่สะท้อนถึงความสัมพันธ์ของลำดับกรดอะมิโนที่อยู่ติดกันโดยใช้ข้อมูลไฮโดรโฟบิกซิตี้และไฮโดรฟิลิกซิตี้เป็นตัวแทน ดังภาพที่ 31 (ก) และ (ข) สัญลักษณ์ τ_3 และ τ_4 คือแฟคเตอร์ความสัมพันธ์ลำดับชั้นที่ 2 (The second-tier correlation factors) ซึ่งเป็นสมการที่สะท้อนถึงความสัมพันธ์ของลำดับกรดอะมิโนที่อยู่ติดกันโดยมีแก๊ป (gap) หรือระยะห่างเท่ากับ 1 อะมิโน ดังภาพที่ 31 (ค) และ (ง) สำหรับค่าของไฮโดรโฟบิกซิตี้และไฮโดรฟิลิกซิตี้ในสมการที่ (24) จะต้องทำการแปลงรูป ดังสมการที่ (25)

$$h^1(\mathrm{R}_i) = \frac{h_0^1(\mathrm{R}_i) - \sum_{k=1}^{20} h_0^1(\mathbb{R}_k)/20}{\sqrt{\sum_{u=1}^{20} \left[h_0^1(\mathbb{R}_u) - \sum_{k=1}^{20} h_0^1(\mathbb{R}_k)/20\right]^2 / 20}}$$
$$h^2(\mathrm{R}_i) = \frac{h_0^2(\mathrm{R}_i) - \sum_{k=1}^{20} h_0^2(\mathbb{R}_k)/20}{\sqrt{\sum_{u=1}^{20} \left[h_0^2(\mathbb{R}_u) - \sum_{k=1}^{20} h_0^2(\mathbb{R}_k)/20\right]^2 / 20}} \quad \textbf{(25)}$$

โดยที่ $\mathbb{R}_i$ $(i = 1,2,...,20)$ แทนกรดอะมิโนตัวที่ 1 ถึง 20 (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W และ Y) สัญลักษณ์ h_0^1 และ h_0^2 คือค่าเริ่มต้นของไฮโดรโฟบิกซิตี้และไฮโดรฟิลิกซิตี้ของกรดอะมิโน R_i โดยค่าดังกล่าวได้จาก Tanford (1962) และ Hopp and Woods (1981)



ภาพที่ 31 แสดงอัลกอริทึม Am-Pse-AA Composition ภาพ (ก) และ (ข) แสดงแฟคเตอร์ความสัมพันธ์ลำดับชั้นที่ 1 ภาพ (ค) และ (ง) แสดงแฟคเตอร์ความสัมพันธ์ลำดับชั้นที่ 2 (มีแก๊ปเท่ากับ 1 อะมิโน) ภาพ (จ) และ (ฉ) แสดงแฟคเตอร์ความสัมพันธ์ลำดับชั้นที่ 3 (มีแก๊ปเท่ากับ 2 อะมิโน)

จากสมการที่ 23-25 จะเห็นได้ว่าการพิจารณาลำดับของกรดอะมิโนโดยใช้คุณสมบัติของไฮโดรโฟบิกซิตี้และไฮโดรฟิลิกซิตี้ของกรดอะมิโนนั้น สามารถสร้างเวคเตอร์ได้เป็น 2λ มิติ ดังนั้นเมื่อรวมกับอะมิโนแอสิดคอมโพสิชัน (amino acid composition) พื้นฐานแล้ว จะได้เวคเตอร์ของตัวแทนสายข้อมูลโปรตีนที่มีขนาดเป็น $(20+2\lambda)$ มิติ ดังสมการที่ (26)

$$\mathrm{P} = \begin{bmatrix} p1 \\ \vdots \\ p20 \\ p20+1 \\ \vdots \\ p20+\lambda \\ p20+\lambda+1 \\ \vdots \\ p20+2\lambda \end{bmatrix} \tag{26}$$

โดยที่

$$p_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{2\lambda} \tau_j}, & 1 \le u \le 20 \\ \dfrac{w\tau_u}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{2\lambda} \tau_j}, & 20+1 \le u \le 20+2\lambda \end{cases} \tag{27}$$

เมื่อค่า $f_i\ (i = 1,2,\ldots,20)$ เป็นค่าความถี่ของการปรากฏขึ้นของอะมิโน i ต่อจำนวนกรดอะมิโนทั้งหมดในสายโปรตีน P (normalized occurrence frequencies) และค่า τ_j เป็นค่าแฟคเตอร์ความสัมพันธ์ลำดับชั้นที่ j โดยคำนวณจากสมการที่ (23) และ w คือค่าถ่วงน้ำหนัก (ในงานวิจัย CDA นี้ ใช้ค่า $w = 0.5$ และ $\lambda = 9$)

ดังนั้นในสมการที่ (26) ค่าของสมาชิก 20 ตัวแรกของเวคเตอร์จะเป็นค่าอะมิโนแอซิดคอมโพสิชัน และค่าของสมาชิก 2λ ตัวถัดมาเป็นค่าดิสครีทที่สะท้อนลำดับกรดอะมิโนโดยผ่านคุณสมบัติไฮโดรโฟบิกซิตี้และไฮโดรฟิลิกซิตี้ของกรดอะมิโน ดังนั้นสายตัวแทนโปรตีนที่เรียกว่า amphiphilic pseudo amino acid composition (Am-Pse-AA Composition) นั้น จะมีลักษณะข้อมูลและเวคเตอร์ตัวแทนคล้ายคลึงกับอะมิโนแอสิดคอมโฟสิชัน แต่จะมีข้อมูลที่มีความสัมพันธ์กับลำดับของกรดอะมิโนและการกระจายตัวของไฮโดรโฟบิกซิตี้และไฮโดรฟิลิกซิตี้ในสายโปรตีนเพิ่มเติม

อุปกรณ์และวิธีการ

อุปกรณ์

ฮาร์ดแวร์

1. เครื่องคอมพิวเตอร์โน๊ตบุค 1 เครื่อง ประกอบด้วยอุปกรณ์ดังต่อไปนี้

    1.1. ซีพียู (CPU) Intel Core 2 Duo ความเร็ว 1.66 GHz

    1.2. หน่วยความจำหลัก 1.5 GB

    1.3. ฮาร์ดดิกส์ขนาด 60 GB

ซอฟต์แวร์

1. ระบบปฏิบัติการ Windows XP Professional
2. ตัวแปรภาษา Perl – ActivePerl version 5.8.8.822 สำหรับระบบปฏิบัติการ Windows
3. โปรแกรม Weka version 3.5
4. โปรแกรม LibSVM version 2.88
5. โปรแกรม ClustalW version 1.83
6. โปรแกรม HMMER version 2.3.2

วิธีการ

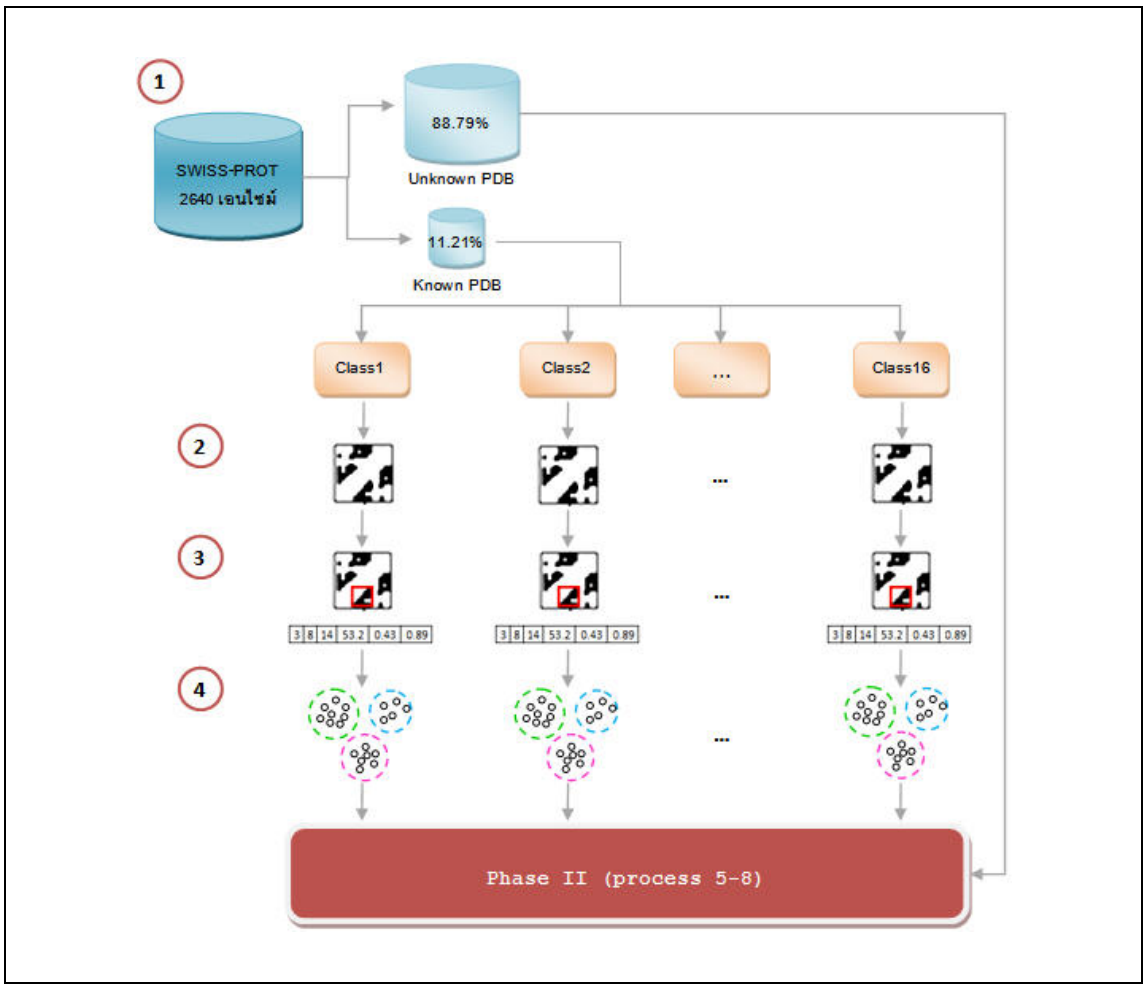
ภาพรวมของระบบ

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาระบบจำแนกประเภทเอนไซม์โดยอัตโนมัติ จากความรู้เกี่ยวกับการทำงานของเอนไซม์นั้น ถูกกำหนดขึ้นจากบริเวณส่วนย่อยของลำดับกรดอะมิโนในสายโปรตีน (Tian *et al.* 2004; Pandey *et al.*, 2006) และเนื่องจากฟังก์ชั่นเอนไซม์มีความสัมพันธ์อย่างใกล้ชิดกับโครงสร้าง 3 มิติ (Gu *et al.*, 2008; Zhang *et al.*, 2008) ดังนั้นแทนการพิจารณาสาย

ลำดับอะมิโนทั้งเส้น ในงานวิจัยนี้จึงตั้งสมมติฐานว่าบริเวณส่วนย่อยของลำดับกรดอะมิโนที่มีโครงสร้างระดับทุติยภูมิอยู่ติดกันเมื่อพิจารณาบนโดเมน 3 มิตินั้น เป็นบริเวณสำคัญที่ส่งผลต่อการเกิดฟังก์ชั่นของเอนไซม์นั่นเอง ดังนั้นเพื่อหาบริเวณส่วนย่อยของลำดับอะมิโนที่สำคัญนั้น ในงานวิจัยนี้ได้ทำการแปลงข้อมูล 3 มิติ ให้อยู่ในรูปตารางเมตริกซ์คอนแทคแมป ซึ่งเป็นตารางที่ให้ความรู้เรื่องการติดกันของโครงสร้างทุติยภูมิบนโดเมน 3 มิติและง่ายต่อความเข้าใจและนำไปพิจารณาต่อ สำหรับผลลัพธ์ที่ได้เพิ่มเติมจากระบบจำแนกประเภทเอนไซม์โดยอัตโนมัตินี้ สามารถให้ข้อมูลโครงสร้างของเอนไซม์ผ่านทางองค์ความรู้ตารางคอนแทคแมป ซึ่งเป็นข้อมูลที่มีนัยสำคัญทางชีววิทยา เป็นที่ทราบกันดีว่าเราสามารถสร้างโครงสร้าง 3 มิติของโปรตีนที่ไม่ทราบโครงสร้างได้จากตารางคอนแทคแมป (Vassura *et al.*, 2008) อย่างไรก็ดีในงานวิจัยนี้สามารถให้ข้อมูลคอนแทคแมปได้เพียงบางส่วนของกรดอะมิโนเท่านั้น

สำหรับข้อมูลเอนไซม์ที่ใช้ในการจำแนกประเภทนั้นยึดตามระบบอีซีนัมเบอร์ (EC Number, Enzyme Commission number) ในงานวิจัยนี้สนใจเอนไซม์ในกลุ่มของออกซิโดรีดักเทส (Oxidoreductases) ซึ่งมีหมายเลขอีซีนัมเบอร์คือ EC 1 ซึ่งมีคลาสย่อยที่ใช้ในการจำแนกประเภทอยู่ทั้งหมด 16 คลาสหรือซับแฟมิลี (คลาสทั้งหมดของออกซิโดรีดักเทสมีทั้งหมด 22 ซับแฟมิลี แต่ข้อมูลที่ใช้ในการทดสอบสามารถรองรับได้เพียง 16 ซับแฟมิลี) ดังภาพผนวกที่ 1 และรายชื่อเอนไซม์ที่ใช้ในงานวิจัยนี้แสดงดังตารางผนวกที่ 1 โดยเอนไซม์ทั้งหมดมีค่าความเหมือนระหว่างเอนไซม์เฉลี่ย 11%~27%

ขั้นตอนการวิจัยประกอบด้วย 2 เฟสหลัก (phase) โดยในเฟสแรกประกอบด้วยขั้นตอนย่อย 1-4 (ภาพที่ 32) เป็นขั้นตอนการเตรียมข้อมูลตลอดจนทำการค้นหารูปแบบโครงสร้างย่อยของโปรตีนที่มีนัยสำคัญหรือ SSCP บนตารางเมตริกซ์คอนแทคแมป ซึ่งเทียบได้กับการค้นหาบริเวณส่วนย่อยของลำดับกรดอะมิโนที่มีความสำคัญต่อการเกิดฟังก์ชั่นของเอนไซม์ สำหรับเฟสที่สองนั้นประกอบด้วยขั้นตอนย่อย 5-8 (ภาพที่ 33) เป็นขั้นตอนในการแปลงโครงสร้างย่อย SSCP ให้อยู่ในรูปลำดับกรดอะมิโนในโครงสร้างปฐมภูมิและนำไปสร้างเป็นตัวแทนของสายโปรตีน ตลอดจนสร้างโมเดลจำแนกประเภท ดังรายละเอียดต่อไปนี้



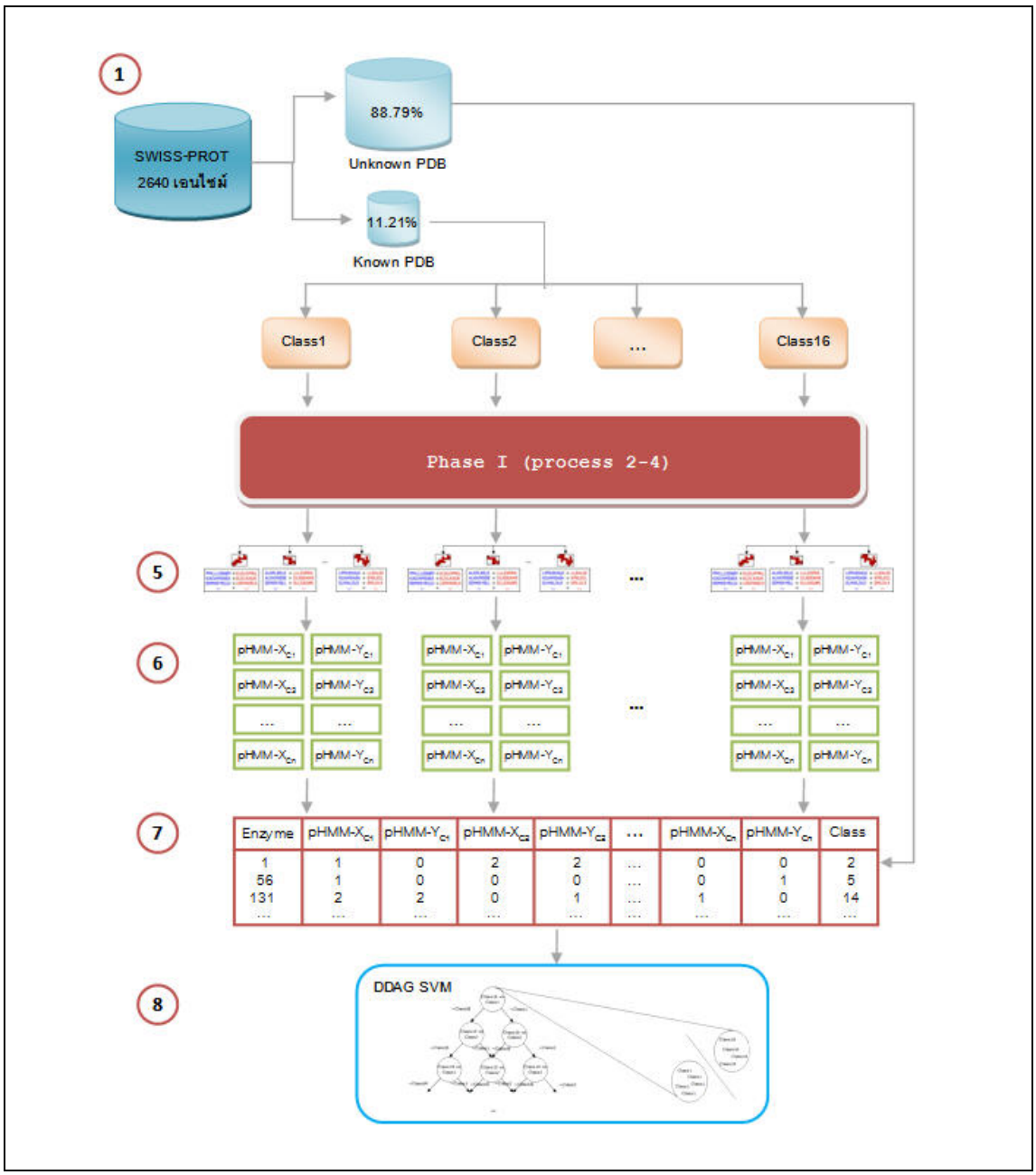
ภาพที่ 32 แสดงไดอะแกรมของขั้นตอนการวิจัยในเฟสที่ 1

ขั้นตอนการเตรียมข้อมูล โดยทำการรวบรวมข้อมูลจากฐานข้อมูลต่าง ๆ ที่เกี่ยวข้อง ได้แก่ ข้อมูลลำดับกรดอะมิโนและอีซีนัมเบอร์ของเอนไซม์ ได้มาจากฐานข้อมูลสวิสพรอท และข้อมูลโครงสร้าง 3 มิติของเอนไซม์ ได้มาจากฐานข้อมูลพีดีบี

ขั้นตอนการสร้างตารางเมตริกซ์คอนแทคแมป ในขั้นตอนนี้จำเป็นต้องใช้ข้อมูลโครงสร้าง 3 มิติของเอนไซม์ จากข้อมูลที่ใช้ในการทดสอบระบบทั้งหมด 2640 โปรตีน (ใช้ข้อมูลเดียวกับงานวิจัย CD ของ Chou and Elrod (2003)) พบว่ามีโปรตีนที่ทราบโครงสร้าง 3 มิติอยู่เพียง 11.21% เท่านั้น

ขั้นตอนการสืบค้นรูปแบบที่อยู่ติดกันและต่อเนื่องกันมากที่สุดบนตารางเมตริกซ์คอนแทคแมป (เรียกว่า sub-structural contactedpattern หรือ SSCP)

ขั้นตอนการจัดกลุ่มของ SSCP ตามลักษณะทางกายภาพที่คล้ายคลึงกันสำหรับแต่ละซับแฟมิลี่คลาส



ภาพที่ 33 แสดงไดอะแกรมของขั้นตอนการวิจัยในเฟสที่ 2

ขั้นตอนการสกัดส่วนลำดับย่อยของสายโปรตีนทั้งในแนวแกน X และในแนวแกน Y ตามโครงสร้าง SSCP

ขั้นตอนการสร้างโพรไฟล์ฮิดเด้นมาร์คอฟโมเดล (Profile Hidden Markov Models, Profile HMMs) เพื่อใช้เป็นตัวแทนของลำดับกรดอะมิโนทั้งหมด และใช้เป็นคุณสมบัติในการพิจารณาในการจำแนกซับแฟมิลี่คลาส

ขั้นตอนการนำโพรไฟล์ฮิดเด้นมาร์คอฟโมเดลและข้อมูลโปรตีนมาสร้างตารางเพื่อใช้เป็นตัวแทนสายโปรตีนทั้งเส้น เพื่อเป็นข้อมูลนำเข้าวิธีการเรียนรู้ซัพพอร์ตเวคเตอร์แมชชีน

ขั้นตอนการสร้างแบบจำลองจำแนกประเภทเอนไซม์ซับแฟมิลี่ เป็นขั้นตอนสุดท้ายที่ทำการรวบรวมข้อมูลทั้งหมดและทำการแปลงเป็นรูปแบบที่เหมาะสม เข้าสู่วิธีการเรียนรู้ซัพพอร์ตเวคเตอร์แมชชีน จนได้เป็นแบบจำลองที่ใช้ในการจำแนกประเภทเอนไซม์ซับแฟมิลี่ที่มีความถูกต้องสูง

ขั้นตอนวิธีวิจัย

1. การเตรียมข้อมูล

ขั้นตอนการเตรียมข้อมูลเป็นขั้นตอนหนึ่งที่มีความสำคัญมาก ในงานวิจัยนี้ใช้ข้อมูลทดสอบระบบจากงานวิจัย CD ของ Chou and Elrod (2005) ประกอบด้วยข้อมูล 2 ชุด (ตารางที่ 6) ได้แก่ (1) ข้อมูลชุดที่ 1 เป็นข้อมูลทดสอบระบบจำนวน 2640 โปรตีน โดยข้อมูลในชุดนี้จะถูกนำไปสกัดฟีเจอร์และสร้างแบบจำลองจำแนกประเภทเอนไซม์ (2) ข้อมูลชุดที่ 2 เป็นข้อมูลทดสอบระบบที่ไม่มีความสัมพันธ์กับข้อมูลในชุดแรก (unknown enzymes) จำนวน 2124 โปรตีน โดยข้อมูลชุดนี้จะถูกนำมาทดสอบระบบจำแนกประเภทเอนไซม์ที่ถูกสร้างจากชุดข้อมูลที่ 1 เท่านั้น สำหรับข้อมูลทั้ง 2 ชุดนี้ อยู่ในกลุ่มเอนไซม์อ็อกซิโดรีดักเทส ซึ่งมีคลาสหรือซับแฟมิลี (subfamily) จำนวน 22 คลาส แต่ในงานวิจัยนี้ทำการจำแนกประเภทเอนไซม์เพียง 16 คลาส เนื่องจากข้อมูลเอนไซม์ของชุดทดสอบระบบไม่สามารถรองรับคลาสได้ทั้งหมด โดยรายชื่อคลาสและเอนไซม์ทั้งหมดที่ใช้งานวิจัยนี้ แสดงไว้ที่ภาคผนวก

ตารางที่ 6 แสดงจำนวนข้อมูลฝึกสอนและทดสอบระบบของแต่ละซับแฟมิลี ที่ใช้ในงานวิจัยนี้

ซับแฟมิลีคลาส	EC. Number	จำนวนเอนไซม์ของข้อมูลชุดที่ 1 (2640)	จำนวนเอนไซม์ (unknown enzymes) ของข้อมูลชุดที่ 2 (2124)
Class1	EC 1.1	314	626
Class2	EC 1.2	216	216
Class3	EC 1.3	194	25
Class4	EC 1.4	130	17
Class5	EC 1.5	112	14
Class6	EC 1.6	305	608
Class7	EC 1.7	64	7
Class8	EC 1.8	59	6
Class9	EC 1.9	254	253
Class10	EC 1.10	94	12
Class11	EC 1.11	154	20
Class12	EC 1.13	94	12
Class13	EC 1.14	257	257
Class14	EC 1.15	155	20
Class15	EC 1.17	84	11
Class16	EC 1.18	154	20

เนื่องจากงานวิจัยนี้ มีสมมติฐานว่าหน้าที่การทำงานของเอนไซม์เกิดจากส่วนย่อยของโครงสร้างโปรตีน และมีความสัมพันธ์อย่างแนบแน่นกับโครงสร้าง (Zhang *et al.*, 2008) ดังนั้นจึงอาศัยองค์ความรู้โครงสร้างของโปรตีนมาใช้ในการสร้างตัวแทนกรดอะมิโน งานวิจัยนี้ทำการแบ่งข้อมูลชุดที่ 1 (จำนวน 2640 โปรตีน) ออกเป็น 2 กลุ่มคือ โดยกลุ่มแรกเป็นกลุ่มข้อมูลโปรตีนที่ทราบโครงสร้าง 3 มิติของโปรตีนแล้ว มีจำนวนทั้งสิ้น 296 โปรตีน (คิดเป็น 11.21% ของข้อมูลทั้งหมด) ใช้ในการสร้างพีเจอร์เพื่อเป็นนำมาเป็นตัวแทนสายโปรตีน กลุ่มที่สองเป็นกลุ่มข้อมูลโปรตีนที่ไม่ทราบโครงสร้าง 3 มิติ มีจำนวนทั้งสิ้น 2344 โปรตีน ในภายหลังข้อมูลกลุ่มที่ 1 และ 2

นี้จะถูกนำมารวมกันและใช้เป็นข้อมูลฝึกสอนและทดสอบระบบด้วยอัลกอริทึมการเรียนรู้ ซัพพอร์ตเวคเตอร์แมชชีน สำหรับข้อมูลกลุ่มที่สอง (unknown enzymes) จำนวน 2124 โปรตีน เป็นข้อมูลที่ไม่มีความสัมพันธ์กับข้อมูลชุดแรก เป็นเพียงข้อมูลที่นำมาทดสอบแบบจำลองที่สร้างได้จากข้อมูลชุดที่ 1 เท่านั้น สำหรับข้อมูลปฐมภูมิหรือข้อมูลลำดับกรดอะมิโนของสายโปรตีนทั้งหมดในงานวิจัยนี้ นำมาจากฐานข้อมูลสวิสพรอท ส่วนข้อมูลโครงสร้าง 3 มิติ ได้มาจากฐานข้อมูลพีดีบี

2. ขั้นตอนการสร้างตารางเมตริกซ์คอนแทคแมป (Contact Map Matrix)

ทำการสร้างตารางเมตริกซ์คอนแทคแมป จากข้อมูลโครงสร้าง 3 มิติของโปรตีน เพื่อทำการเปลี่ยนรูปแบบข้อมูลที่ซับซ้อนบนโดเมน 3 มิติ ให้สามารถเข้าใจได้ง่ายบนโดเมน 2 มิติ โดยใช้สมการที่ (1) และ (2) โดยในงานวิจัยนี้กำหนดให้

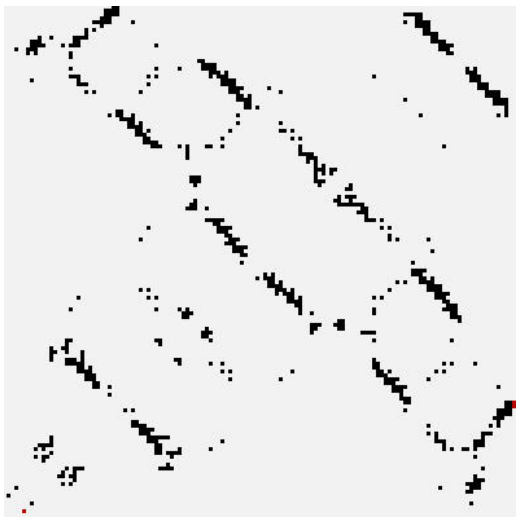
r_i และ r_j คือตำแหน่งของอะตอม β-carbon (α-carbon สำหรับกรดอะมิโน Glycine) ของกรดอะมิโน a_i และ a_j

Threshold T คือค่าที่ผู้ใช้งานกำหนด เพื่อเป็นจุดตัดสินว่าคู่ลำดับกรดอะมิโนนั้นอยู่ติดกันบนโดเมน 3 มิติหรือไม่ สำหรับงานวิจัยนี้ กำหนดค่า T เท่ากับ 8°A (Angstroms)

ข้อมูลอินพุตของระบบคือ ข้อมูลโปรตีนที่อยู่ในรูปแบบพีดีบี (PDB format) ดังภาพที่ 34 และผลลัพทธ์ที่ได้คือตารางเมตริกซ์คอนแทคแมป ขนาด $N \times N$ ดังภาพที่ 35 สำหรับอัลกอริทึมในการสร้างตารางเมตริกซ์คอนแทคแมป แสดงดังภาพที่ 36

```
HEADER    OXIDOREDUCTASE                          03-SEP-92   1ABN
TITLE     THE CRYSTAL STRUCTURE OF THE ALDOSE REDUCTASE NADPH BINARY
COMPND    MOL_ID: 1;
SOURCE    MOL_ID: 1;
SOURCE    2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
KEYWDS    OXIDOREDUCTASE
ATOM      1  CA  SER A   2      30.848  19.504  28.132  1.00 15.00           C
ATOM      2  CA  ARG A   3      28.654  21.084  30.618  1.00 15.00           C
ATOM      3  CA  LEU A   4      25.698  23.423  30.085  1.00 15.00           C
ATOM      4  CA  LEU A   5      25.102  26.733  31.988  1.00 15.00           C
.....
ATOM    299  CA  GLU A 313      -1.470  19.275  24.889  1.00 15.00           C
ATOM    300  CA  GLU A 314       0.320  18.990  21.747  1.00 15.00           C
ATOM    301  CA  PHE A 315       3.505  20.793  23.060  1.00 15.00           C
TER     302      PHE A 315
HETATM  303  PA  NDP  351      17.109   4.070  28.871  1.00 15.00           P
HETATM  304  O1A NDP  351      16.669   4.604  27.576  1.00 15.00           O
.....
CONECT  348  347
CONECT  349  347
CONECT  350  347
MASTER      649    0    1    9   11    1    0    6  349    1   48   25
END
```

ภาพที่ 34 แสดงข้อมูลอินพุตสำหรับอัลกอริทึมในการสร้างตารางเมตริกซ์คอนแทคแมป จากรูปคือข้อมูลโครงสร้าง 3 มิติของโปรตีนในรูปแบบพีดีบี (PDB format)



ภาพที่ 35 แสดงข้อมูลเอาท์พุตจากอัลกอริทึมในการสร้างตารางเมตริกซ์คอนแทคแมป จากภาพบริเวณที่เป็นสีดำคือบริเวณที่กรดอะมิโนอยู่ติดกันบนโดเมน 3 มิติ

```
METHOD :    GetCMAP (Get Contact Map Matrix)
INPUT :     PDB file pdb_file , Threshold t , Sequence separate seq_separate
OUTPUT :    Contact Map Matrix file cmap_file
 1| index = 0
 2| find_cmap = TRUE
 3| OPEN input file pdb_file
 4| WHILE (READ Each line of pdb_file TO line )
 5|     IF line CONTAINS multiple ranges of residue
 6|         find_cmap = FALSE
 7|         Exit
 8|     ELSE
 9|         record_name = substring ( line , 0, 6)
10|         IF record_name Equal TO "ATOM"
11|             atom_name = substring ( line , 12, 4)
12|             residue_name = substring ( line , 17, 4)
13|             IF ( residue_name Equal TO "GLY" AND atom_name Equal TO "CA") OR
                 ( residue_name NOT Equal TO "GLY" AND atom_name Equal TO "CB")
14|                 x_coordinate = substring ( line , 30, 8)
15|                 y_coordinate = substring ( line , 38, 8)
16|                 z_coordinate = substring ( line , 46, 8)
17|                 seq [ index ] = residue_name
18|                 coordinate [ index ][0] = x_coordinate
19|                 coordinate [ index ][1] = y_coordinate
20|                 coordinate [ index ][2] = z_coordinate
21|                 index ++
22|             END IF
23|         END IF
24|     END IF
25| END WHILE
26| IF find_cmap Equal TO TRUE
27|     OPEN output file cmap_file
28|     FOR row = Length( seq ) - 1; row >= 0; row - -
29|         PRINT seq [ row ] TO cmap_file
30|         FOR col = 0; col < Length( seq ); col ++
31|             cmap_val = plotCMAP
32|             PRINT tab and rms TO cmap_file
33|         END FOR
34|         PRINT new line TO cmap_file
35|     END FOR
36|     FOR col = 0; col < Length( seq ); col ++
37|         PRINT tab and seq [ col ] TO cmap_file
38|     END FOR
39| END IF
```

ภาพที่ 36 แสดงอัลกอริทึมในการสร้างตารางเมตริกซ์คอนแทคแมป

จากอัลกอริทึมพบว่าเวลาที่ใช้ในการสร้างตารางเมตริกซ์คอนแทคแมปเท่ากับ $O(N^2)$ เริ่มจากฟังก์ชั่น GetCMAP() เป็นฟังก์ชั่นหลักที่ทำหน้าที่เก็บข้อมูลตำแหน่งโคออร์ดิเนต (coordinate) ของอะตอมอัลฟา (α) หรือเบต้า (β) ของกรดอะมิโน เพื่อนำมาคำนวณระยะห่างระหว่างคู่กรดอะมิโนใดๆ ด้วยฟังก์ชั่น PlotCMAP() และแสดงผลเป็นตารางเมตริกซ์ขนาด $N \times N$ โดยมีรายละเอียดการทำงานดังนี้

ฟังก์ชั่น GetCMAP() ต้องการข้อมูลอินพุต 3 ชนิดด้วยกันได้แก่ ข้อมูลพีดีบีของโปรตีนที่ต้องการสร้างตาราเมตริกซ์คอนแทคแมป ค่าตัดสินใจ (threshold) เพื่อเป็นตัวบ่งชี้ว่าคู่กรดอะมิโนที่กำลังพิจารณานั้นอยู่ติดกันหรือไม่ มีหน่วยเป็น Angstrom (Å) (ในงานวิจัยนี้กำหนดค่า $T = 8$) และค่าแบ่งแยกสายลำดับ (sequence separation) เนื่องจากตารางเมตริกซ์คอนแทคแมปเป็นตารางที่สมมาตร เกิดจากการนำสายโปรตีนเส้นเดียวกันมาพิจารณาระยะห่างเป็นคู่ๆ ทุกคู่ ดังนั้นแนวแกนทแยงมุมก็คือกรดอะมิโนตัวเดียวกันซึ่งเราจะไม่พิจารณา แต่อย่างไรก็ตามกรดอะมิโนที่อยู่ใกล้กันมากๆ มีโอกาสสูงที่คู่กรดอะมิโนนั้นจะอยู่ติดกัน ซึ่งเป็นข้อมูลที่ไม่มีนัยสำคัญ ดังนั้นจึงทำการกำหนดค่าแบ่งแยกสายลำดับขึ้นมา โดยในงานวิจัยนี้ใช้ค่าสายลำดับเท่ากับ 4 กรดอะมิโน ส่วนเอาท์พุตที่ได้คือตารางเมตริกซ์คอนแทคแมปของโปรตีนอินพุตนั่นเอง

เริ่มจากบรรทัดที่ 1-2 เป็นการกำหนดค่าเริ่มต้นของฟังก์ชั่น บรรทัดที่ 3-4 เปิดไฟล์พีดีบีและอ่านข้อมูลพีดีบีทีละบรรทัด บรรทัดที่ 5-8 เป็นการตรวจสอบว่าข้อมูลพีดีบีนั้น เป็นข้อมูลของกรดอะมิโนที่ไม่สมบูรณ์หรือไม่ ถ้าพบว่าไม่สมบูรณ์เราจะไม่นำมาใช้เป็นอินพุตของระบบ เนื่องจากข้อมูลพีดีบี จะมีชื่อเรคคอร์ด (record name) เป็นตัวบอกว่าข้อมูลนั้นเกี่ยวกับอะไร บรรทัดที่ 9-23 สนใจเฉพาะบรรทัดที่มีชื่อเรคคอร์ดว่า “ATOM” ซึ่งเก็บข้อมูลพิกัดของอะตอมคาร์บอนไว้ โดยบรรทัดที่ 13 ทำการพิจารณาว่าถ้ากรดอะมิโนนั้นเป็นไกลซีน (glycine) ให้เก็บข้อมูลพิกัดของอัลฟาคาร์บอนอะตอม แต่ถ้าไม่ใช่ให้เก็บค่าพิกัดของเบต้าคาร์บอนอะตอมแทน โดยพิกัดที่เก็บไว้จะอยู่ในตัวแปรอาร์เรย์ 2 มิติชื่อ *coordinate*[][] เมื่อวนลูปอ่านไฟล์พีดีบีจนหมดทุกบรรทัด จากนั้นจึงทำกระบวนวนลูปแถวและคอลัมน์เพื่อเขียนข้อมูลเอาท์พุตซึ่งก็คือตารางเมตริกซ์คอนแทคแมปลงเอาท์พุตไฟล์ ดังบรรทัดที่ 27-39 สำหรับข้อมูลเอาท์พุตนั้นได้จากการเรียกใช้ฟังก์ชั่น PlotCMAP() ซึ่งมีรายละเอียดดังภาพที่ 37

METHOD : PlotCMAP (Calculate Root Mean Square, consider threshold and sequence separate values)

INPUT : row, $coordinate\ [row\ ][0]$, $coordinate\ [row\ ][1]$, $coordinate\ [row\ ][2]$,
col, $coordinate\ [col\ ][0]$, $coordinate\ [col\ ][1]$, $coordinate\ [col\ ][2]$,
t, $seq\_separate$

OUTPUT : $cmap\_val$ (0 - NOT Contact; 1 - Contact)

```
1| x1 = coordinate [row ][0]
2| y1 = coordinate [row ][1]
3| z1 = coordinate [row ][2]
4| x2 = coordinate [col ][0]
5| y2 = coordinate [col ][1]
6| z2 = coordinate [col ][2]
7| rms = sqrt((x1 - x2)^2 + (y1 - y2)^2 + (z1 - z2)^2)
8| residue_range = |row - col|
9| IF (residue_range ≥ seq_separate) AND (rms ≤ t)
10|     cmap_val = 1
11| ELSE
12|     cmap_val = 0
13| END IF
```

ภาพที่ 37 แสดงอัลกอริทึมในการคำนวณระยะห่างระหว่างคู่กรดอะมิโนใดๆ บนโดเมน 3 มิติ

ฟังก์ชั่นย่อย PlotCMAP() ทำหน้าที่คำนวณระยะห่างระหว่างคู่กรดอะมิโน และทำการตัดสินใจว่าคู่กรดอะมิโนนั้นอยู่ติดกันบนโดเมน 3 มิติหรือไม่ โดยใช้ค่าการตัดสินใจจากค่าที่ผู้ใช้กำหนด รวมทั้งยังพิจารณาค่าแบ่งแยกสายลำดับ เพื่อไม่ทำการพิจารณาคู่ลำดับอะมิโนที่มีตำแหน่งห่างกันน้อยกว่าค่าที่ผู้ใช้กำหนด ดังนั้นข้อมูลอินพุตจึงได้แก่ ตำแหน่งของกรดอะมิโนที่ต้องการพิจารณาตัวที่ 1, ข้อมูลพิกัดของกรดอะมิโนตัวที่ 1, ตำแหน่งของกรดอะมิโนที่ต้องการพิจารณาตัวที่ 2, ข้อมูลพิกัดของกรดอะมิโนตัวที่ 2, ค่าตัดสินใจ (Threshold) และค่าแบ่งแยกสายลำดับ (sequence separation) ส่วนเอาท์พุตที่ได้คือค่าคอนแทคแมป (มี 2 ค่าคือ “0” บ่งชี้ว่าคู่กรดอะมิโนไม่อยู่ติดกัน และ “1” บ่งชี้ว่าคู่กรดอะมิโนนั้นอยู่ติดกัน)

บรรทัดที่ 1-6 เก็บค่าพิกัดของกรดอะมิโนทั้งสองตัวในตัวแปรภายใน x_1, y_1, z_1 สำหรับกรดอะมิโนตัวที่ 1 และ x_2, y_2, z_2 สำหรับกรดอะมิโนตัวที่ 2 บรรทัดที่ 7 ทำการคำนวณระยะห่างระหว่างกรดอะมิโนทั้งสองตัวบนโดเมน 3 มิติ บรรทัดที่ 8 คำนวณระยะห่างระหว่างกรดอะมิโนทั้งสองตัวบนโดเมน 1 มิติ บรรทัดที่ 9-13 ตรวจสอบว่าถ้าระยะห่างบนโดเมน 1 มิติมีค่ามากกว่าหรือเท่ากับค่าที่ผู้ใช้กับหนด และค่าระยะห่างบนโดเมน 3 มิติ มีค่าน้อยกว่าหรือเท่ากับค่าที่ผู้ใช้กับหนด

แล้วนั้น เราจะกำหนดให้ค่าคอนแทคแมปเท่ากับ “1” นั่นหมายถึงกรดอะมิโนนั้นอยู่ติดกัน แต่ถ้าในทางกลับกันเราจะกำหนดค่าคอนแทคแมปเท่ากับ “0” นั่นหมายถึงกรดอะมิโนนั้นไม่อยู่ติดกันนั่นเอง

3. ขั้นตอนการสืบค้นรูปแบบที่อยู่ติดกันและต่อเนื่องกันมากที่สุด บนตารางเมตริกซ์คอนแทคแมป (Sub-Structural Contacted Pattern, SSCP)

ในขั้นตอนนี้จะประกอบด้วย 2 ขั้นตอนย่อยคือ การสืบค้น SSCP และการสกัดคุณลักษณะทางกายภาพของ SSCP

3.1 สืบค้น SSCP

เป็นการสืบค้นรูปแบบที่อยู่ติดกันและต่อเนื่องมากที่สุด บนตารางเมตริกซ์คอนแทคแมป เนื่องจากรูปแบบดังกล่าว สามารถบ่งบอกลักษณะทางกายภาพของโครงสร้างโปรตีนได้ว่ามีโครงสร้างทุติยภูมิลักษณะใดและวางตัวอย่างไรบนโดเมน 3 มิติ ภาพที่ 38 แสดงอัลกอริทึมที่ใช้ในการสืบค้น SSCP

```
METHOD :    FindSSCP (Mining Sub-Structural Contact Patterns)
INPUT :     Contact Map Matrix file cmap_file
OUTPUT :    SSCP file sscp_file
 1| row = 0
 2| OPEN input file cmap_file
 3| WHILE (READ Each line of cmap_file TO line AND split with tab)
 4|     IF line [0] NOT Equal TO empty
 5|         FOR col = 1; col <= Length( line ); col ++
 6|             couple_residue [row][col - 1] = line [col]
 7|         END FOR
 8|         row ++
 9|     END IF
10| END WHILE
11| CALL METHOD ChangeContactLocationToUniqueNo ()
12| CALL METHOD ReWriteToBoundaryNumber ()
13| PRINT new matrices which specify boundary TO SSCP file
```

ภาพที่ 38 แสดงอัลกอริทึมหลักที่ใช้ในการสืบค้น SSCP ประกอบด้วย 2 อัลกอริทึมย่อย

ฟังก์ชั่น FindSSCP() เป็นฟังก์ชั่นหลักในการสืบค้นรูปแบบที่อยู่ติดกันและต่อเนื่องกันมากที่สุดบนตารางเมตริกซ์คอนแทคแมป อินพุตของฟังก์ชั่นนี้ได้แก่ ไฟล์ตารางคอนเมตริกซ์แทคแมปที่ได้จากอัลกอริทึม GetCMAP() ส่วนเอาท์พุตคือไฟล์ SSCP นั่นเอง

บรรทัดที่ 1 เป็นการกำหนดค่าเริ่มต้นให้กับตัวแปรภายใน บรรทัดที่ 2-3 ทำการเปิดและอ่านไฟล์ตารางเมตริกซ์คอนแทคแมป บรรทัดที่ 5-6 ทำการเก็บค่าคอนแทค (0 หรือ 1) ของแต่ละคู่กรดอะมิโนลงในตัวแปรอาร์เรย์ 2 มิติชื่อ *couple_residue*[][] แต่เนื่องจากในไฟล์เมตริกซ์คอนแทคแมปนั้น คอลัมน์แรกจะเป็นชื่อของกรดอะมิโน ดังนั้นค่าของคอนแทคแมปจึงเริ่มจากคอลัมน์ถัดไป บรรทัดที่ 11-13 ทำการเรียกใช้ฟังก์ชั่น ChangeContactLocationToUniqueNo() และฟังก์ชั่น ReWriteToBoundaryNumber() เพื่อแปลงตารางเมตริกซ์คอนแทคแมปให้อยู่ในรูปแบบที่ง่ายต่อการสืบค้น SSCP จากนั้นทำการสืบค้น SSCP และเขียนเอาท์พุตลงไฟล์ ตามลำดับ

```
METHOD :      ChangeContactLocationToUniqueNo
INPUT :       seq_len , couple_residue
OUTPUT :      New contact map matrices with Unique number of contact value
1|  counting_no = 1
2|  FOR row = 0; row < seq_len ; row ++
3|      FOR col = 0; col < ( seq_len - row ); col ++
4|          IF couple_residue [ row ][ col ] Equal TO 1
5|              couple_residue [ row ][ col ] = counting_no
6|              counting_no ++
7|          END IF
8|      END FOR
9|  END FOR
```

ภาพที่ 39 แสดงอัลกอริทึมย่อยสำหรับแปลงรูปแบบข้อมูลบนตารางเมตริกซ์คอนแทคแมป ให้อยู่ในรูปแบบที่ง่ายต่อการสืบค้น SSCP

จากภาพที่ 39 ฟังก์ชั่นย่อย ChangeContactLocationToUniqueNo() เพื่อให้การสืบค้นรูปแบบทำได้ง่ายขึ้น จึงทำการเปลี่ยนค่าคอนแทคแมปที่มีค่าเท่ากับ “1” เป็นเลขลำดับจำนวนเต็มบวกที่ต่อเนื่องกัน โดยเริ่มจาก 1,2,3,… เป็นต้นไป ดังนั้นเอาท์พุตที่ได้คือตารางเมตริกซ์คอนแทคแมปที่มีค่าคอนแทคเป็นหมายเลขที่ไม่ซ้ำกันและเรียงต่อกัน

บรรทัดที่ 2-3 ทำการวนลูปแถวและคอลัมน์ของตารางเมตริกซ์ *couple _ residue*[][]
บรรทัดที่ 4-7 ทำการตรวจสอบค่าคอนแทคแมปในแต่ละเซลล์ (cell) ว่ามีค่าเท่ากับ “1” หรือไม่ ถ้ามีค่าเท่ากับ “1” ให้ทำการแทนที่ด้วยค่าที่อยู่ในตัวแปร *counting _ no* ซึ่งเป็นเลขลำดับ (running number) จำนวนเต็มบวก จากนั้นจึงบวกเพิ่มหนึ่งค่าให้กับตัวแปร *counting _ no*

```
METHOD :    ReWriteToBoundaryNumber
INPUT :     seq _len , couple _residue
OUTPUT :
1|  FOR row = 0; row < seq _len ; row ++
2|      FOR col = 0; col < ( seq _len - row ); col ++
3|          CALL METHOD GetMBR ()
4|      END FOR
5|  END FOR
```

ภาพที่ 40 แสดงอัลกอริทึมย่อยเพื่อทำการเรียกอัลกอริทึม GetMBR() เพื่อสืบค้น SSCP สำหรับทุกแถวและคอลัมน์ของตารางเมตริกซ์คอนแทคแมป

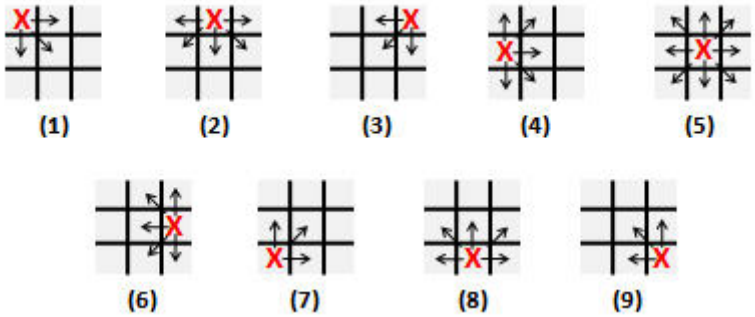
```
METHOD :    GetMBR (Recursive to find maximal boundary regtangle)
INPUT :     row, col, seq _len
OUTPUT :    SSCP
1|  IF (PREV_row >= 0) and (NEXT_col <= seq _len )
2|      IF ( couple _residue [PREV_row][NEXT_col] NOT Equal TO 0) and
        ( couple _residue [PREV_row][NEXT_col] NOT Equal TO couple _residue [CURR_row][CURR_col])
3|          couple _residue [PREV_row][NEXT_col] = couple _residue [CURR_row][CURR_col]
4|          CALL METHOD GetMBR(PREV_row, NEXT_col, seq _len )
5|      END IF
6|  END IF
7|  IF (PREV_row >= 0)
8|      IF ( couple _residue [PREV_row][CURR_col] NOT Equal TO 0) and
        ( couple _residue [PREV_row][CURR_col] NOT Equal TO couple _residue [CURR_row][CURR_col])
9|          couple _residue [PREV_row][CURR_col] = couple _residue [CURR_row][CURR_col]
10|         CALL METHOD GetMBR(PREV_row, CURR_col, seq _len )
11|     END IF
12| END IF
```

ภาพที่ 41 อัลกอริทึมย่อยเพื่อสืบค้น SSCP

```
13| IF (PREV_row >= 0) and (PREV_col >= 0)
14|     IF ( couple _ residue [PREV_row][PREV_col] NOT Equal TO 0) and
      ( couple _ residue [PREV_row][PREV_col] NOT Equal TO couple _ residue [CURR_row][CURR_col])
15|         couple _ residue [PREV_row][PREV_col] = couple _ residue [CURR_row][CURR_col]
16|         CALL METHOD GetMBR(PREV_row, PREV_col, seq _ len )
17|     END IF
18| END IF
19| IF (PREV_col >= 0)
20|     IF ( couple _ residue [CURR_row][PREV_col] NOT Equal TO 0) and
      ( couple _ residue [CURR_row][PREV_col] NOT Equal TO couple _ residue [CURR_row][CURR_col])
21|         couple _ residue [CURR_row][PREV_col] = couple _ residue [CURR_row][CURR_col]
22|         CALL METHOD GetMBR(CURR_row, PREV_col, seq _ len )
23|     END IF
24| END IF
25| IF (NEXT_row <= seq _ len ) and (PREV_col >= 0)
26|     IF ( couple _ residue [NEXT_row][PREV_col] NOT Equal TO 0) and
      ( couple _ residue [NEXT_row][PREV_col] NOT Equal TO couple _ residue [CURR_row][CURR_col])
27|         couple _ residue [NEXT_row][PREV_col] = couple _ residue [CURR_row][CURR_col]
28|         CALL METHOD GetMBR(NEXT_row, PREV_col, seq _ len )
29|     END IF
30| END IF
31| IF (NEXT_row <= seq _ len )
32|     IF ( couple _ residue [NEXT_row][CURR_col] NOT Equal TO 0) and
      ( couple _ residue [NEXT_row][CURR_col] NOT Equal TO couple _ residue [CURR_row][CURR_col])
33|         couple _ residue [NEXT_row][CURR_col] = couple _ residue [CURR_row][CURR_col]
34|         CALL METHOD GetMBR(NEXT_row, CURR_col, seq _ len )
35|     END IF
36| END IF
37| IF (NEXT_row <= seq _ len ) and (NEXT_col <= seq _ len )
38|     IF ( couple _ residue [NEXT_row][NEXT_col] NOT Equal TO 0) and
      ( couple _ residue [NEXT_row][NEXT_col] NOT Equal TO couple _ residue [CURR_row][CURR_col])
39|         couple _ residue [NEXT_row][NEXT_col] = couple _ residue [CURR_row][CURR_col]
40|         CALL METHOD GetMBR(NEXT_row, NEXT_col, seq _ len )
41|     END IF
42| END IF
43| IF (NEXT_col <= seq _ len )
44|     IF ( couple _ residue [CURR_row][NEXT_col] NOT Equal TO 0) and
      ( couple _ residue [CURR_row][NEXT_col] NOT Equal TO couple _ residue [CURR_row][CURR_col])
45|         couple _ residue [CURR_row][NEXT_col] = couple _ residue [CURR_row][CURR_col]
46|         CALL METHOD GetMBR(CURR_row, NEXT_col, seq _ len )
47|     END IF
48| END IF
```

ภาพที่ 41 อัลกอริทึมย่อยเพื่อสืบค้น SSCP (ต่อ)

ฟังก์ชั่นย่อย ReWriteToBoundaryNumber() ทำการวนลูปเพื่อแถวและคอลัมน์ของตารางเมตริกซ์ *couple_residue*[][] เพื่อเรียกใช้ฟังก์ชั่นย่อย GetMBR() เพื่อสืบค้นรูปแบบที่อยู่ติดกันและต่อเนื่องกันมากที่สุดและในแต่ละขั้นตอนการหา SSCP นั้น โปรแกรมจะมีการเรียกตัวเอง (Recursive) เพื่อสืบหาเซลล์ถัดไปที่มีค่าคอนแทคไม่เท่ากับ “0” ซึ่งอยู่ติดกับเซลล์ปัจจุบัน ดังภาพที่ 42 อินพุตของข้อมูลได้แก่ ตำแหน่งแถว ตำแหน่งคอลัมน์ และความยาวของสายโปรตีน ส่วนเอาท์พุตที่ได้คือบริเวณ SSCP นั่นเอง



ภาพที่ 42 แสดงตำแหน่งต่างๆ ที่เป็นไปได้บนตารางเมตริกซ์คอนแทคแมปและทิศทางของการหาเซลล์ที่อยู่ติดกัน

บรรทัดที่ 1-6 ใช้ในการพิจารณาว่าเซลล์ที่ติดกับเซลล์ปัจจุบันที่มีตำแหน่งอยู่ด้านบนขวานั้นเป็นเซลล์ที่มีค่าคอนแทคมากกว่า “0” หรือไม่ (ภาพที่ 42 หมายเลข 4, 5, 7, 8) ถ้าใช่ก็จะเรียกโปรแกรมเดิม (recursive) และส่งตำแหน่งใหม่เป็นอินพุตให้ ทำเช่นนี้ซ้ำไปเรื่อยๆ จนกว่าจะพบสถานะหยุด

บรรทัดที่ 7-12 ใช้ในการพิจารณาว่าเซลล์ที่ติดกับเซลล์ปัจจุบันที่มีตำแหน่งอยู่ด้านบนนั้นเป็นเซลล์ที่มีค่าคอนแทคมากกว่า “0” หรือไม่ (ภาพที่ 42 หมายเลข 4, 5, 6, 7, 8, 9)

บรรทัดที่ 13-18 ใช้ในการพิจารณาว่าเซลล์ที่ติดกับเซลล์ปัจจุบันที่มีตำแหน่งอยู่ด้านบนซ้ายนั้นเป็นเซลล์ที่มีค่าคอนแทคมากกว่า “0” หรือไม่ (ภาพที่ 42 หมายเลข 5, 6, 8, 9)

บรรทัดที่ 19-24 ใช้ในการพิจารณาว่าเซลล์ที่ติดกับเซลล์ปัจจุบันที่มีตำแหน่งอยู่ด้านซ้ายนั้นเป็นเซลล์ที่มีค่าคอนแทคมากกว่า “0” หรือไม่ (ภาพที่ 42 หมายเลข 2, 3, 5, 6, 8, 9)

บรรทัดที่ 25-30 ใช้ในการพิจารณาว่าเซลล์ที่ติดกับเซลล์ปัจจุบันที่มีตำแหน่งอยู่ด้านล่างซ้ายนั้นเป็นเซลล์ที่มีค่าคอนแทคมากกว่า “0” หรือไม่ (ภาพที่ 42 หมายเลข 2, 3, 5, 6)

บรรทัดที่ 31-36 ใช้ในการพิจารณาว่าเซลล์ที่ติดกับเซลล์ปัจจุบันที่มีตำแหน่งอยู่ด้านล่างนั้นเป็นเซลล์ที่มีค่าคอนแทคมากกว่า “0” หรือไม่ (ภาพที่ 42 หมายเลข 1, 2, 3, 4, 5, 6)

บรรทัดที่ 37-42 ใช้ในการพิจารณาว่าเซลล์ที่ติดกับเซลล์ปัจจุบันที่มีตำแหน่งอยู่ด้านล่างขวานั้นเป็นเซลล์ที่มีค่าคอนแทคมากกว่า “0” หรือไม่ (ภาพที่ 42 หมายเลข 1, 2, 4, 5)

บรรทัดที่ 43-48 ใช้ในการพิจารณาว่าเซลล์ที่ติดกับเซลล์ปัจจุบันที่มีตำแหน่งอยู่ด้านขวานั้นเป็นเซลล์ที่มีค่าคอนแทคมากกว่า “0” หรือไม่ (ภาพที่ 42 หมายเลข 1, 2, 4, 5, 7, 8)

3.2 สกัดคุณลักษณะ (Feature) ทางกายภาพของ SSCP

เมื่อได้รูปแบบ SSCP มาแล้ว ในแต่รูปแบบย่อยนั้นเราจะทำการสกัดคุณลักษณะทางกายภาพที่สำคัญของ SSCP เพื่อใช้เป็นตัวแทนในการจัดกลุ่มตามรูปแบบที่ปรากฏบนตารางเมตริกซ์คอนแทคแมป โดยฟีเจอร์ที่ใช้ในงานวิจัยนี้ได้ประยุกต์มาจากงานวิจัย Yang *et al*. (2004) ที่กล่าวไว้แล้วในส่วนงานวิจัยที่เกี่ยวข้อง ซึ่งเป็นฟีเจอร์ที่สามารถบ่งบอกคุณลักษณะทางกายภาพของ SSCP ได้อย่างเด่นชัด มีอยู่ด้วยกัน 5 ฟีเจอร์ คือ ความสูงของ SSCP, ความกว้างของ SSCP, จำนวนของการติดกันของคู่กรดอะมิโนบนโดเมน 3 มิติ, มุมที่แสดงถึงการกระจายตัวของการติดกันของคู่กรดอะมิโนโดเมน 3 มิติ, ค่าการกระจายตัวของการติดกันของคู่กรดอะมิโนในแนวแกน X และค่าการกระจายตัวของการติดกันของคู่กรดอะมิโนในแนวแกน Y

กำหนดให้ SSCP มีจำนวนของการติดกันของคู่กรดอะมิโนทุกคู่รวมกันเท่ากับ n และกำหนดให้ตำแหน่งของเซลล์แต่ละเซลล์ใน SSCP เขียนแทนด้วย $(x_1, y_1)(x_2, y_2)\ldots(x_n, y_n)$ ดังนั้นสามารถหามุมที่แสดงถึงการกระจายตัวได้ดังต่อไปนี้

$$\beta = \frac{\sum_{i=1}^{n}((x_i - \bar{x})\cdot(y_i - \bar{y}))}{\sum_{i=1}^{n}((x_i - \bar{x})^2)} \quad \textbf{(28)}$$

โดยที่ β คือความชันของการกระจายตัวของคู่อะมิโนที่ติดกันโดยค่า β มีค่าอยู่ระหว่าง $[-\infty,\infty]$ จากนั้นทำการแปลงให้อยู่ในรูปของมุมตามแนวแกน x ดังสมการที่ (29) ซึ่งจะได้ว่าค่า มุม α จะมีค่าอยู่ระหว่าง $[-90,90]$ องศา จะเห็นได้ว่าค่า α นั้นมีค่าอยู่ในควอดแรนท์ (Quadrant) ที่ 1 (0 ถึง 90 องศา) และ 4 (0 ถึง -90 องศา) ซึ่งอยากแก่การนำไปพิจารณา ดังนั้นจึงต้องทำการแปลงค่า α ที่อยู่ในควอดแรนท์ที่ 4 ให้มาอยู่ในควอดแรนท์ที่ 2 สุดท้ายจะได้มุม α' ที่มีช่วงอยู่ระหว่าง $[0,180)$ ดังสมการที่ (30)

$$\alpha = \frac{\arctan(\beta)\cdot 180}{\pi}, \quad (-90 \le \alpha \le 90; -\infty \le \beta \le \infty) \tag{29}$$

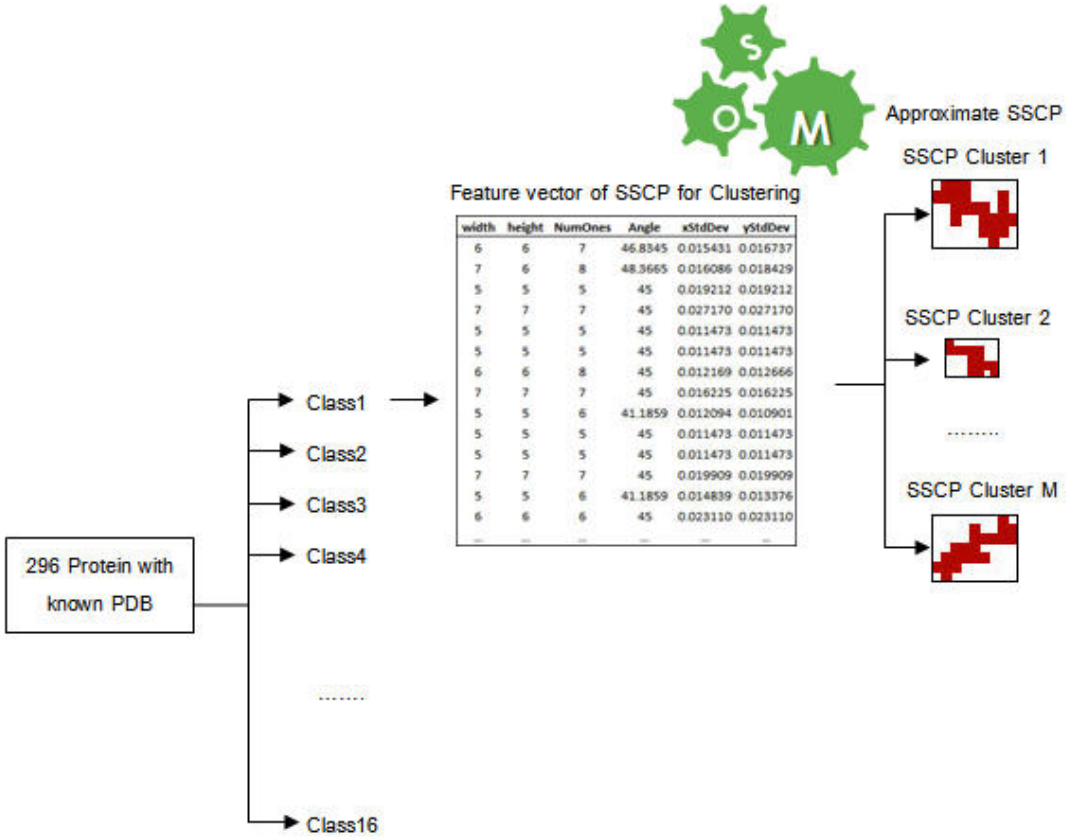
$$\begin{cases} \alpha' = \alpha + 180 & if \quad \alpha < 0 \\ \alpha' = \alpha & otherwise \end{cases} \tag{30}$$

3.3 กรอง SSCP ที่ไม่มีนัยสำคัญ

SSCP ที่สกัดได้นั้น มีขนาดกว้าง x ยาว ตั้งแต่ 1×1 เป็นต้นไป และแต่ละ SSCP จะมีจำนวนของการติดกันของคู่กรดอะมิโนบนโดเมน 3 มิตินั้นตั้งแต่ 1 เป็นต้นไปเช่นกัน จะสังเกตเห็นว่า SSCP บางรูปแบบนั้นไม่มีนัยสำคัญเพียงพอที่จะใช้เป็นตัวแทนของกรดอะมิโน ในงานวิจัยนี้จะทำการกรอง SSCP ที่ไม่มีนัยสำคัญออก โดยจะสนใจเฉพาะ SSCP ที่มีความหนาแน่นของการติดกันระหว่างคู่ลำดับอะมิโนตั้งแต่ 3 คู่ขึ้นไป (Yang *et al*., 2004)

4. ขั้นตอนการจัดกลุ่มของ SSCP ตามลักษณะทางกายภาพที่คล้ายคลึงกันสำหรับแต่ละซับแฟมิลี่คลาส

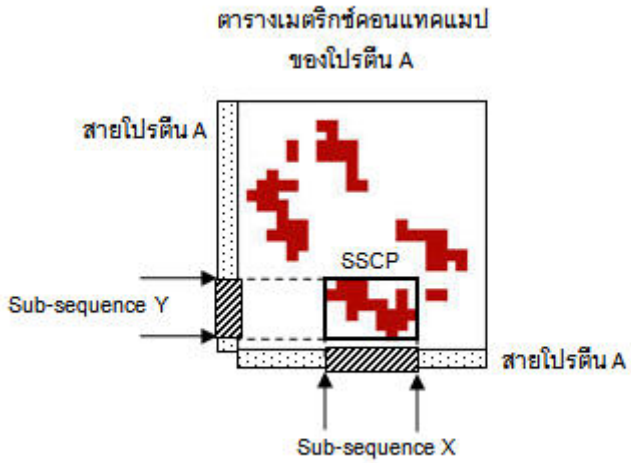
ทำการจัดกลุ่ม SSCP ของข้อมูลโปรตีนในแต่ละซับแฟมิลีคลาส (ภาพที่ 43) โดยงานวิจัยนี้ใช้อัลกอริทึมการจัดกลุ่ม (clustering algorithm) คือ แผนผังเรียนรู้การจัดตัวเอง (Self-Organizing Map, SOM) (Kohonen, 1997) ผ่านระบบ GEPAS (Gene Expression Pattern Analysis Suite) version 4.0 (Herrero *et al.*, 2003) ดังที่ได้กล่าวไว้แล้วในส่วนงานวิจัยที่เกี่ยวข้อง ดังนั้นกลุ่มที่ได้จะมีรูปแบบทางกายภาพบนตารางเมตริกซ์คอนแทคแมปที่คล้ายคลึงกัน ซึ่งก็คือเป็นการจัดกลุ่มตามโครงสร้างย่อย 3 มิติของโปรตีนนั่นเอง



ภาพที่ 43 แสดงวิธีการจัดกลุ่ม SSCP ตามลักษณะทางกายภาพที่คล้ายคลึงกันสำหรับแต่ละซับแฟมิลี่คลาส

5. ขั้นตอนการสกัดส่วนลำดับย่อยของสายโปรตีนทั้งในแนวแกน X และในแนวแกน Y ตามโครงสร้าง SSCP

ทำการแปลงรูปแบบส่วนย่อยของ SSCP ให้อยู่ในรูปของส่วนลำดับย่อยของกรดอะมิโน (sub-sequence) ในรูปโครงสร้างปฐมภูมิ โดยจาก SSCP ที่ได้จะทำการสกัดส่วนลำดับย่อยทั้งแนวแกน X และแนวแกน Y ดังภาพที่ 44 เพื่อใช้เป็นข้อมูลนำเข้าในขั้นตอนที่ 6 ต่อไป



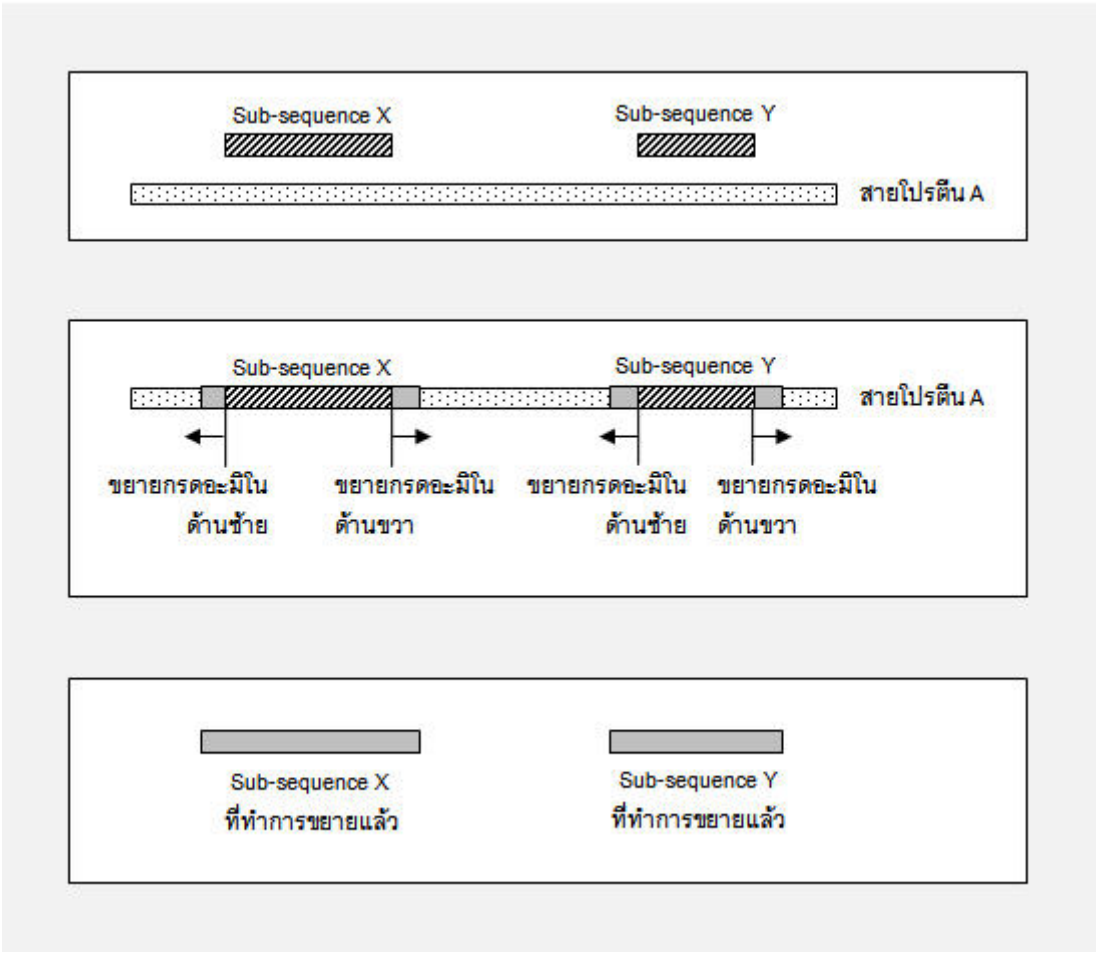
ภาพที่ 44 แสดงการสืบค้นสายลำดับย่อยในแนวแกน X และแนวแกน Y จากตารางเมตริกซ์คอนแทคแมป

6. ขั้นตอนการสร้างโพรไฟล์ฮิดเด้นมาร์คอฟโมเดล (Profile Hidden Markov Models, Profile HMMs)

ทำการสร้างโพรไฟลฮิดเด้นมาร์คอฟโมเดลเพื่อใช้เป็นตัวแทนของลำดับกรดอะมิโนทั้งหมด และใช้เป็นคุณสมบัติในการพิจารณาในการจำแนกซับแฟมิลี่คลาส ประกอบด้วย 2 ขั้นตอนย่อยดังนี้

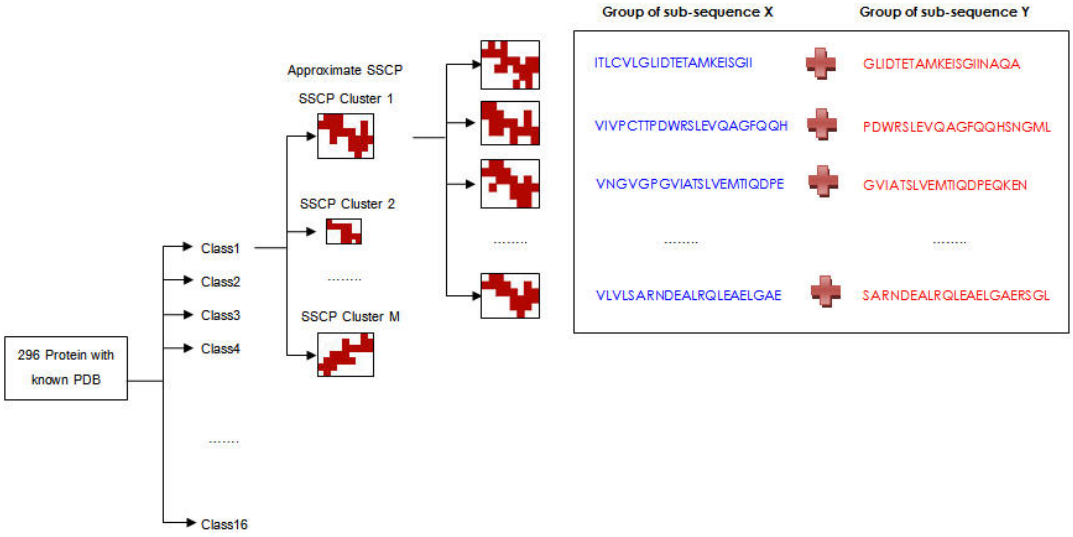
6.1 ขยายความยาวของสายลำดับย่อย (sub-sequence) ของแต่ละ SSCP

การที่กรดอะมิโนจะวางตัวติดกันหรือแยกห่างจากกันบนโดเมน 3 มิติได้นั้น เกิดจากแรงยึดเหนี่ยวและแรงผลักระหว่างอะตอมของกรดอะมิโนข้างเคียง ดังนั้นเพื่อให้ได้ตัวแทนกรดอะมิโนที่มีคุณภาพยิ่งขึ้น จึงทำการขยายความยาวของสายลำดับย่อยของแต่ละ SSCP ออกไปในทิศทางด้านซ้ายและขวาของกรดอะมิโนตัวสุดท้ายของสายลำดับย่อยของ SSCP ดังภาพที่ 45



ภาพที่ 45 แสดงวิธีการขยายความยาวของสายลำดับย่อยของ SSCP ทั้งแกน X และแกน Y ในงานวิจัยนี้ทำการขยายสายลำดับย่อยไปในทิศทางซ้ายและขวาด้านละ 7 อะมิโน

ดังนั้นภาพรวมสุดท้ายตั้งแต่การจัดกลุ่ม SSCP การสกัดส่วนลำดับย่อยของกรดอะมิโนตลอดจนการขยายความยาวของส่วนย่อยนั้น จะได้ภาพรวมดังภาพที่ 46



ภาพที่ 46 แสดงภาพรวมของการจัดกลุ่มตามโครงสร้าง SSCP และการสกัดส่วนย่อยของลำดับกรดอะมิโนตามโครงสร้าง SSCP นั้น

6.2 สร้างโพรไฟล์ฮิดเด้นมาร์คอฟโมเดลสำหรับแต่ละกลุ่มของ SSCP แยกตามซับแฟมิลี่

สายลำดับย่อยของ SSCP ที่ได้ทำการขยายแล้วทั้งแกน X และแกน Y จะถูกนำมาสร้างโพรไฟล์ฮิดเด้นมาร์คอฟโมเดล เพื่อทำการสร้างฐานข้อมูล SSCP และนำไปใช้ในการสร้างตารางเพื่อเป็นตัวแทนสายโปรตีนในลำดับถัดไป

เนื่องจากการสร้างโพรไฟล์ฮิดเด้นมาร์คอฟโมเดล ต้องมีข้อมูลนำเข้าเป็นผลลัพธ์จากการการจัดเรียงลำดับสายโปรตีนของกลุ่มโปรตีนให้อยู่ในตำแหน่งที่มีค่าคะแนนความเหมือน (similarity) มากที่สุด หรือเรียกว่าการทำ multiple sequence alignment (MSA) สำหรับงานวิจัยนี้ใช้โปรแกรม ClustalW (Thompson *et al.*, 1994) version 1.83 ในการทำ MSA จากนั้นทำการสร้างโพรไฟล์ฮิดเด้นมาร์คอฟโมเดล ด้วยโปรแกรม HMMER (Eddy, 1998) version 2.3.2 สำหรับรายละเอียดวิธีการทำงานของ MSA และ HMMER กล่าวไว้แล้วในส่วนการตรวจเอกสาร

6.3 ปรับแต่งโพรไฟล์ฮิดเด้นมาร์คอฟโมเดลด้วย HMM Calibrate

ทำการปรับแต่งโพรไฟล์ฮิดเด้นมาร์คอฟโมเดลที่ได้มาด้วย HMM Calibrate เพื่อเพิ่มประสิทธิภาพในการค้นหาข้อมูลผ่านฐานข้อมูลโพรไฟล์ที่สร้างขึ้น

ตารางที่ 7 แสดงจำนวนโพรไฟล์ฮิดเด้นมาร์คอฟโมเดลที่ได้จากข้อมูล

ซับแฟมิลี่คลาส	จำนวนข้อมูลชุดที่ 1 ที่ทราบโครงสร้าง 3 มิติ (296)	จำนวนโพรไฟล์ฮิดเด้นมาร์คอฟโมเดล (1030)
Class1	43	44
Class2	20	70
Class3	27	68
Class4	27	64
Class5	17	66
Class6	14	64
Class7	12	70
Class8	11	66
Class9	6	60
Class10	7	60
Class11	25	68
Class12	18	70
Class13	23	68
Class14	21	56
Class15	7	64
Class16	18	72

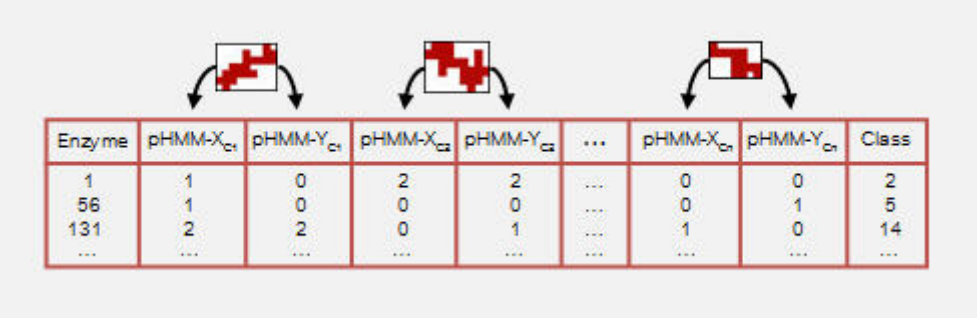
ดังนั้นจากขั้นตอนนี้จะพบว่าจำนวนโพรไฟล์ฮิดเด้นมาร์คอฟโมเดลที่ใช้เป็นตัวแทนสายลำดับย่อยมีจำนวนเท่ากับ 1056 ตัว โดยเป็นโพรไฟล์ฮิดเด้นมาร์คอฟโมเดลของสายลำดับย่อยของแกน X และแกน Y อย่างละเท่ากันคือ 528 ตัว แต่ฟีเจอร์ของบางคลาสอาจยังเป็นตัวแทนที่ไม่ดีพอ ดังนั้นจึงทำการกรองฟีเจอร์ของคลาสที่ 1 ซึ่งเป็นคลาสที่มีจำนวนข้อมูล PDB มากที่สุด โดยใช้เงื่อนไขว่าฟีเจอร์ที่ดีควรเป็นฟีเจอร์ที่พบอยู่ในโปรตีนของคลาสนั้น และไม่ควรปรากฏว่าเป็นฟีเจอร์ของคลาสอื่นๆ แต่ทั้งนี้เพื่อเพิ่มความยืดหยุ่นจึงกำหนดว่าให้ฟีเจอร์ที่ดีนั้น ควรพบในโปรตีนของคลาสอื่นๆ น้อยกว่า 50% สุดท้ายจะได้จำนวนฟีเจอร์ทั้งหมดเท่ากับ 1030 ฟีเจอร์ แสดงดังตารางที่ 7

7. ขั้นตอนการนำโพรไฟล์ฮิดเด้นมาร์คอฟโมเดลและข้อมูลโปรตีนมาสร้างตารางเพื่อใช้เป็นตัวแทนสายโปรตีนทั้งเส้น

ในขั้นตอนนี้จะทำการสร้างตารางฟีเจอร์ (feature table) เพื่อหาความสัมพันธ์ของโครงสร้างโปรตีนที่ได้แปลงให้อยู่ในรูปโพรไฟล์ฮิดเด้นมาร์คอฟโมเดลของสายลำดับย่อยแกน X และแกน Y (ให้สัญลักษณ์ย่อว่า $pHMM_x$ และ $pHMM_y$ ตามลำดับ) ดังนั้นแอทริบิวต์ของตารางคือโพรไฟล์ฮิดเด้นมาร์คอฟโมเดลจำนวน 1030 ตัว จากนั้นนำสายโปรตีนที่ต้องการตรวจสอบค่าแอทริบิวต์ มาค้นหาในฐานข้อมูลโพรไฟล์ฮิดเด้นมาร์คอฟโมเดล ถ้าผลลัพธ์ที่ได้จากการค้นหามีค่า E-value น้อยกว่าหรือเท่ากับค่าตัดสินใจ T เราจะให้ค่าคอลัมน์ของโพรไฟลฮิดเด้นมาร์คอฟมีค่าเท่ากับ “1” แต่ถ้าค่า E-value มีค่ามากกว่าค่าตัดสินใจ T เราจะให้ค่าคอลัมน์ของโพรไฟลฮิดเด้นมาร์คอฟมีค่าเท่ากับ “0”

แต่เนื่องจากการเกิดรูปแบบ SSCP บนตารางเมตริกซ์คอนแทคแมปนั้น ต้องอาศัยสายลำดับย่อยทั้งในแนวแกน X และแกน Y ดังนั้นจึงควรให้ค่าถ่วงน้ำหนักสำหรับแอทริบิวต์โพรไฟล์ฮิดเด้นมาร์คอฟโมเดลที่มีค่า $pHMM_x$ และ $pHMM_y$ เท่ากับ “1” ทั้งคู่ โดยให้ค่าแอทริบิวต์เป็น “2” ดังรายละเอียดวิธีการคำนวณในตารางที่ 8

จากภาพที่ 47 พบว่าถ้าเอนไซม์ใดมีฟีเจอร์ทั้งแกน X และแกน Y นั่นคือเอนไซม์นั้นมีองค์ประกอบของโครงสร้างย่อย SSCP ซึ่งส่วนย่อยของ SSCP หลายๆ ส่วนนี้ สามารถบ่งบอกฟังก์ชั่นการทำงานของเอนไซม์ได้



Enzyme	pHMM-X_{c_1}	pHMM-Y_{c_1}	pHMM-X_{c_2}	pHMM-Y_{c_2}	...	pHMM-X_{c_n}	pHMM-Y_{c_n}	Class
1	1	0	2	2	...	0	0	2
56	1	0	0	0	...	0	1	5
131	2	2	0	1	...	1	0	14
...	...	...	...	...	...	...	...	...

ภาพที่ 47 แสดงตาราฟีเจอร์ที่ใช้เป็นตัวแทนสายโปรตีน

ตารางที่ 8 แสดงเงื่อนไขการคำนวณค่าคอลัมน์ของแอทริบิวต์โพรไฟล์ฮิดเด้นมาร์คอฟโมเดล

ค่า $E(pHMM_x)$	ค่า $E(pHMM_y)$	ค่าแอทริบิวต์ของ $pHMM_x$	ค่าแอทริบิวต์ของ $pHMM_y$
$\leq T$	$\leq T$	2	2
$\leq T$	$> T$	1	0
$> T$	$\leq T$	0	1
$> T$	$> T$	0	0

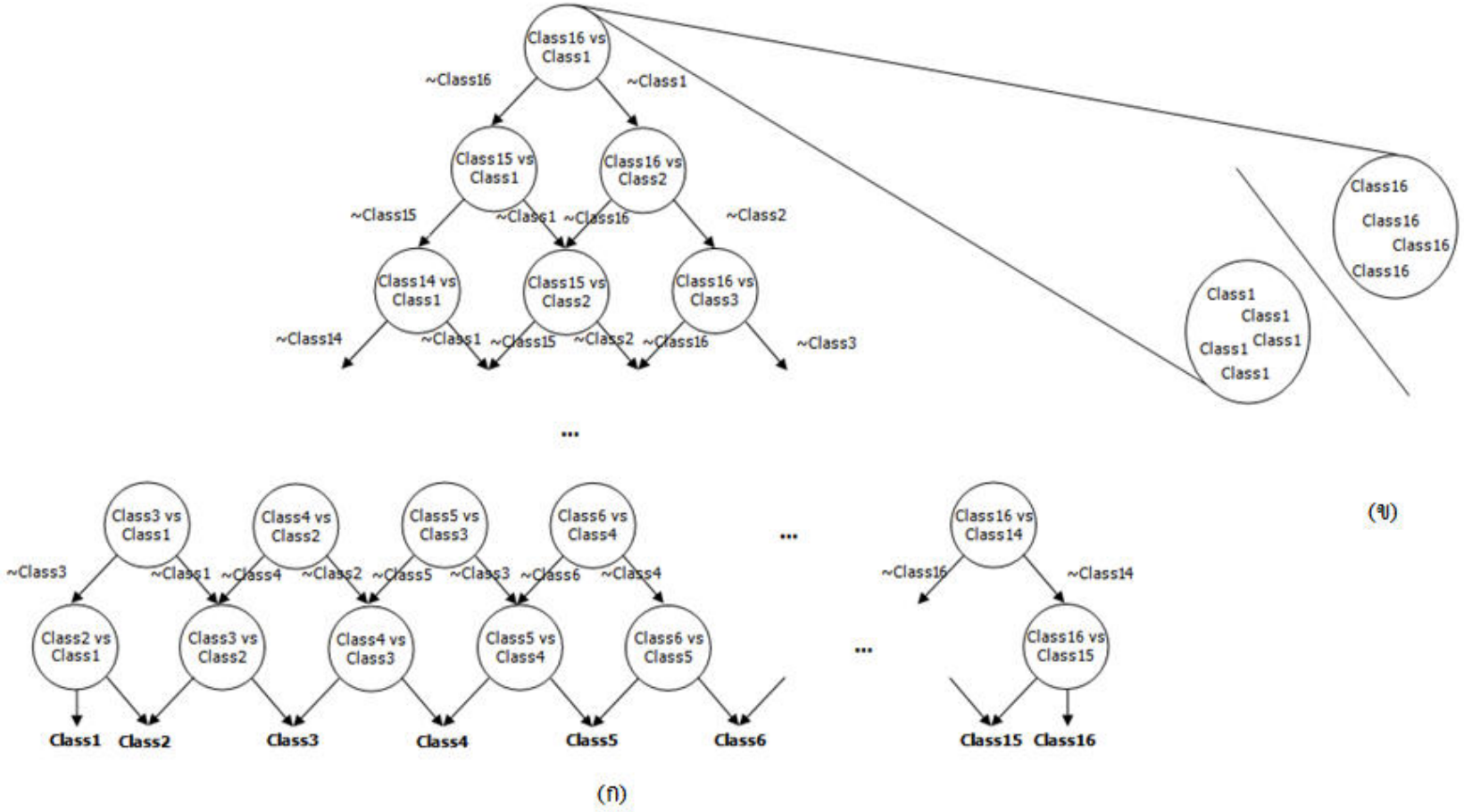
หมายเหตุ $E(pHMM_x)$ คือค่า E-value ของโพรไฟล์ฮิดเด้นมาร์คอฟโมเดลแกน X
$E(pHMM_y)$ คือค่า E-value ของโพรไฟล์ฮิดเด้นมาร์คอฟโมเดลแกน Y
T คือค่า Thershold ของ E-value (ในงานวิจัยนี้ใช้ค่า T = 10)

8. ขั้นตอนการสร้างแบบจำลองจำแนกประเภทเอนไซม์ซับแฟมิลี่

ขั้นตอนสุดท้ายจะเป็นการสร้างแบบจำลองเพื่อจำแนกประเภทเอนไซม์ซับแฟมิลี่ด้วยซัพพอร์ตเวคเตอร์แมชชีน (SVM) สำหรับหลักการทำงานของ SVM ได้กล่าวไว้ในส่วนของการตรวจเอกสาร ในงานวิจัยนี้ใช้โปรแกรม LibSVM (Chih-Chung and Chih-Jen, 2001) version 2.88 โดยใช้วิธีการจำแนกคลาสของข้อมูลออกเป็น N คลาสด้วยวิธี Decision Directed Acyclic Graph SVM (DDAG SVM) สุดท้ายแล้วได้โมเดลในการสร้างระบบจำแนกประเภทเอนไซม์ทั้งหมด 120 โมเดล โดยหลักการของ DDAG SVM คือใช้ตัวจำแนกคลาสของข้อมูลแบบไบนารีคลาส เพื่อจำแนกข้อมูลออกจากกันและส่งผลลัพธ์ที่ได้ไปยังตัวจำแนกคลาสของข้อมูลตัวถัดไป ซึ่งจะมีลักษณะกราฟเป็นแบบ Directed Acyclic Graph ดังภาพที่ 48 โดยแต่ละโหนดในภาพที่ 48 (ก) คือตัวจำแนกคลาสของข้อมูล SVM แบบ one-versus-one ซึ่งจำแนกข้อมูลออกจากกัน ดังภาพที่ 48 (ข)

จากภาพที่ 48 (ก) เริ่มจากการจำแนกข้อมูลคลาสของข้อมูลออกเป็น 2 คลาสคือ Class1 (CH-OH Group) และ Class16 (Reduced Ferredoxin) โดยข้อมูลที่ไม่ใช่ Class16 จะถูกส่งไปยังจัวจำแนกคลาสของข้อมูล SVM เพื่อจำแนกคลาสของข้อมูลออกเป็น 2 คลาสคือ Class1 (CH-OH Group) และ Class15 (-CH2 Group) และทำเช่นนี้ต่อไปเรื่อยๆ จนท้ายที่สุดจะได้ข้อมูลที่เป็น Class1 (CH-OH Group), Class2 (Aldehyd/Oxo Group), Class3 (CH-CH Group), Class4 (CH-NH2

Group), Class5 (CH-NH Group), Class6 (NADH/NADPH), Class7 (Other Nitrogenous Compounds), Class8 (Sulfur Group), Class9 (Heme Group), Class10 (Diphenols and Related Substances), Class11 (Peroxide), Class12 (Single Donors), Class13 (Paired Donors), Class14 (Superoxide Radicals), Class15 (-CH2 Groups) และ Class16 (Reduced Ferredoxin) ตามลำดับ



ภาพที่ 48 แสดงวิธีการสร้างแบบจำลองเพื่อจำแนกประเภทข้อมูลแบบหลายคลาสของ SVM (ก) การจำแนกคลาสของข้อมูลออกเป็น 16 คลาสด้วยวิธี DDAG SVM (ข) ไดอะแกรมฟีเจอร์สเปซ แสดงการจำแนกข้อมูล Class 1 และ Class

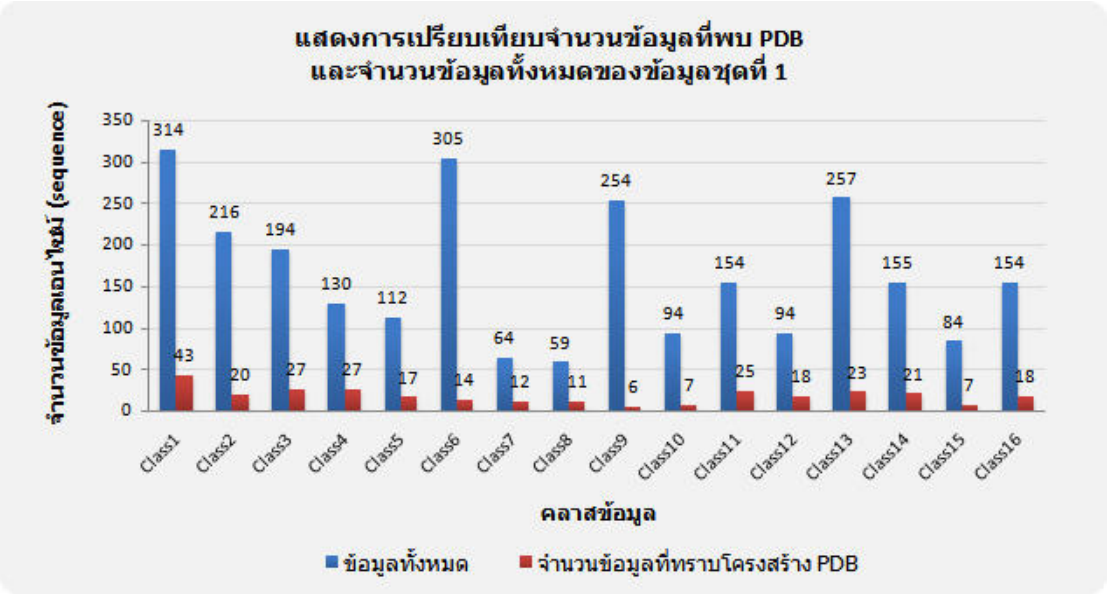
ผลและวิจารณ์

ผล

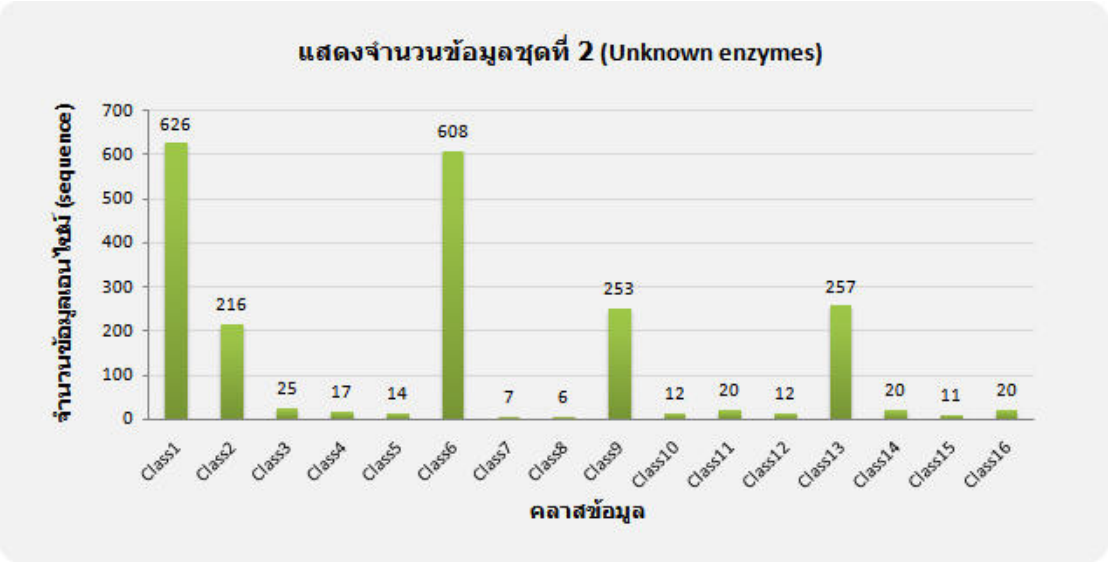
1. วิธีวัดผลการทดลอง

ในหัวข้อนี้จะกล่าวถึงผลการวิจัยของงานวิจัยนี้เปรียบเทียบกับงานวิจัย Chou and Elrod (2003) และงานวิจัย Chu (2005) โดยทั้ง 2 งานวิจัยนี้ ใช้ความน่าจะเป็นของการเกิดกรดอะมิโนที่อยู่ติดกัน หรือเรียกว่าอะมิโนแอซิดคอมโพสิชัน (amino acid composition) เป็นตัวแทนสายโปรตีนหรือฟีเจอร์ แต่ต่างกันสำหรับงานวิจัย Chu (2005) นั้น ไม่ได้ใช้ข้อมูลอะมิโนแอซิดคอมโพสิชันโดยตรง แต่จะใช้ซูโดอะมิโนแอซิด คอมโพสิชัน (pseudo amino acid composition) ซึ่งเป็นการพิจารณคู่กรดอะมิโนโดยที่มีระยะห่างหรือแก๊ปเท่ากับ 1, 2 หรือ 3 เป็นต้น และมีการคำนวณค่าไฮโดรโฟบิกซิตี้และไฮโดรฟิลิกซิติ้ระหว่างกรดอะมิโนร่วมด้วย จากนั้นนำฟีเจอร์ที่ได้ไปสร้างตารางตัวแทนสายข้อมูลโปรตีนทั้งเส้น เพื่อเป็นข้อมูลนำเข้าอัลกอริทึมวิธีการเรียนรู้ และทำการสร้างแบบจำลองเพื่อใช้ในการจำแนกเอนไซม์ซับแฟมิลี่คลาส

ข้อมูลที่ใช้ในงานวิจัยนี้ใช้ข้อมูลชุดเดียวกันกับงานวิจัย Chu and Elrod (2003) โดยแบ่งชุดข้อมูลออกเป็น 2 กลุ่มคือ (1) ข้อมูลชุดที่ 1 เป็นชุดข้อมูลทดสอบระบบ จำนวน 2640 โปรตีน โดยเป็นข้อมูลที่ใช้สกัดฟีเจอร์และสร้างแบบจำลอง (2) ข้อมูลชุดที่ 2 (unknown enzymes) เป็นชุดข้อมูลที่ไม่มีความสัมพันธ์กับข้อมูลชุดแรก ใช้เพื่อทดสอบแบบจำลองที่สร้างได้จากข้อมูลชุดที่ 1 เท่านั้น มีจำนวน 2124 โปรตีน ข้อมูลที่ใช้สร้างตัวแทนกรดอะมิโนใช้ข้อมูลจากข้อมูลชุดที่ 1 ที่ทราบโครงสร้าง 3 มิติแล้ว ซึ่งมีอยู่เพียง 11.21% เท่านั้น แสดงการเปรียบเทียบจำนวนข้อมูลด้วยกราฟในภาพที่ 49 และ 50 ดังนี้



ภาพที่ 49 แสดงจำนวนข้อมูลชุดที่ 1 เป็นข้อมูลที่ใช้ทดสอบระบบเปรียบเทียบกับจำนวนข้อมูลที่ทราบโครงสร้าง 3 มิติแล้ว



ภาพที่ 50 แสดงจำนวนข้อมูลชุดที่ 2 (unknown enzymes) เป็นข้อมูลทดสอบระบบที่ไม่เกี่ยวข้องกับข้อมูลชุดที่ 1 เป็นชุดข้อมูลที่ใช้ทดสอบแบบจำลองที่สร้างได้จากข้อมูลชุดที่ 1 เท่านั้น

2. ผลการทดลอง

สำหรับผลการทดลองของวิธีวิจัยในงานวิจัยนี้แสดงผลการทดลองแยกเป็น 2 วิธีคือ วิธีการที่ 1 (Approach 1) เป็นผลการทดลองตามวิธีวิจัยที่กล่าวแล้วในตอนต้น ส่วนวิธีการที่ 2 (Approach 2) เป็นผลการทดลองที่ได้มีการเพิ่มองค์ความรู้เกี่ยวกับการจัดกลุ่มการแทนที่ของกรดอะมิโน (resubstitution group) โดยพิจารณาจากตาราง BLOSUM62 เมื่อกำหนด Threshold เท่ากับ 0 ดังนั้นจะได้กลุ่มการแทนที่ทั้งหมด 16 กลุ่มคือ [P] [CA] [HY] [MQ] [WFY] [AGS] [NGS] [TAS] [TAV] [NTS] [ILMF] [ILMV] [DNQES] [RNQEH] [RNQEK] และ [SNQEK] โดยงานวิจัยนี้ได้เพิ่มองค์ความรู้ดังกล่าวในขั้นตอนการจัดกลุ่ม SSCP เพื่อเป็นการพิจารณาการแทนที่ระหว่างกรดอะมิโนสำหรับกลุ่มที่มีโครงสร้าง SSCP คล้ายคลึงกันร่วมด้วย ดังนั้นฟีเจอร์เวคเตอร์ของการจัดกลุ่ม SSCP จะเพิ่มขึ้นอีก 16 มิติ ตามจำนวนกลุ่มการแทนที่ของกรดอะมิโนนั่นเอง และค่าของฟีเจอร์แต่ละตัวคือการพิจารณาความถี่ในการปรากฏกรดอะมิโนที่อยู่ในกลุ่มการแทนที่นั้น

ผลการทดลองที่ได้ในงานวิจัยนี้แบ่งออกเป็น 2 หัวข้อหลักได้แก่ ผลทางด้านประสิทธิภาพของระบบจำแนกประเภทเอนไซม์ซับแฟมิลี่ โดยใช้ข้อมูลโครงสร้างโปรตีนจากเมตริกซ์คอนแทคแมปและโพรไฟล์ฮิดเด้นมาร์คอฟโมเดลเป็นตัวแทนสายข้อมูลโปรตีน และผลทางด้านองค์ความรู้ที่ได้จากระบบจำแนกประเภทข้อมูลเอนไซม์ เพื่อใช้ในการวิเคราะห์โครงสร้างของโปรตีนใหม่ที่ไม่ทราบโครงสร้าง 3 มิติมาก่อน

2.1 เปรียบเทียบด้านประสิทธิภาพของระบบจำแนกประเภทเอนไซม์ซับแฟมิลีที่ได้จากงานวิจัยนี้

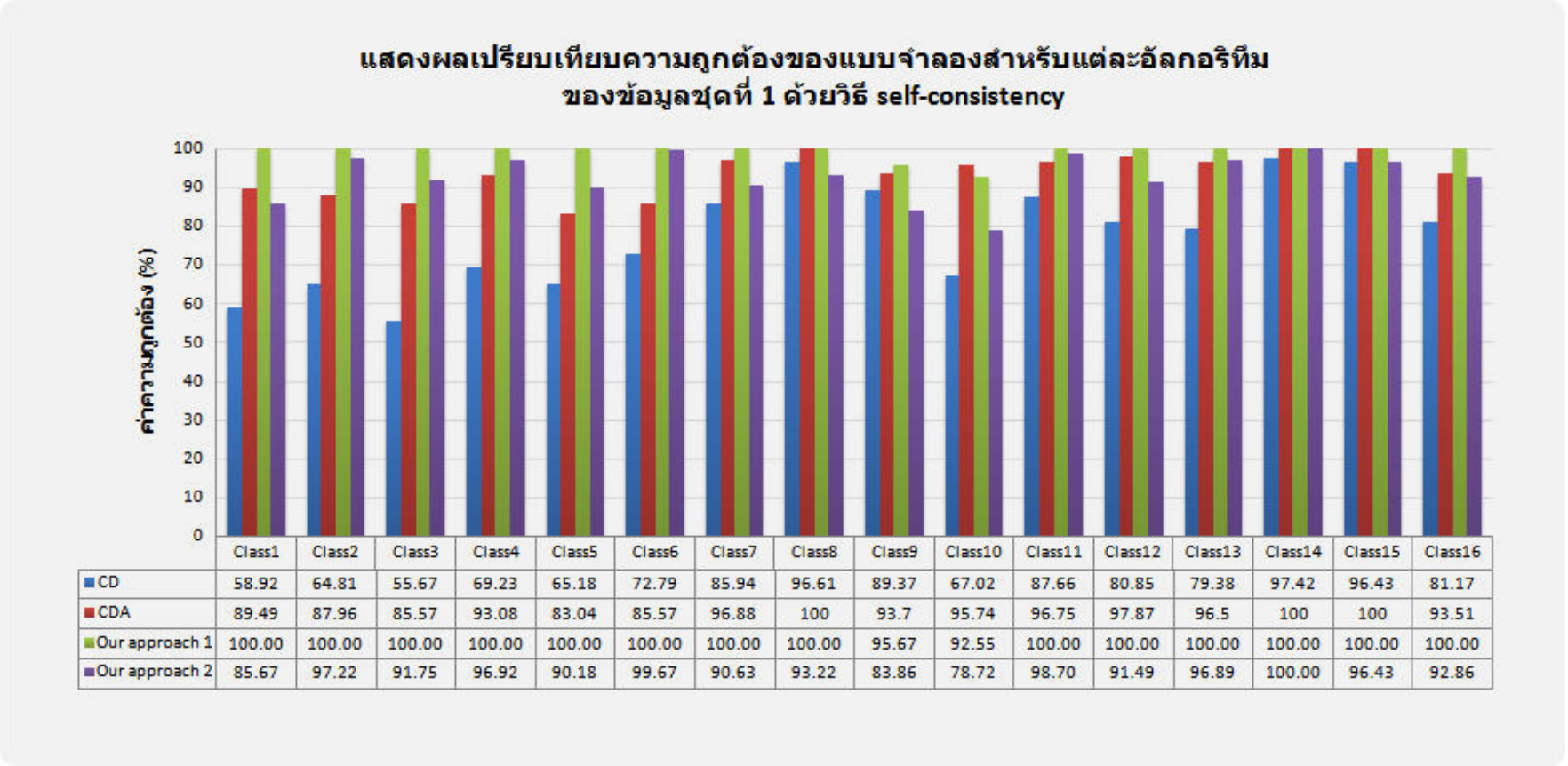
วิธีการวัดประสิทธิภาพของแบบจำลองนั้น มีอยู่ 3 วิธีที่มีการใช้อย่างกว้างขวาง ได้แก่ (1) วิธีการทดสอบด้วยข้อมูลที่ใช้สอนระบบ เรียกวิธีนี้ว่า resubstitution test หรือ self-consistency test (2) วิธีการทดสอบโดยแบ่งข้อมูลเป็นกลุ่มๆ จำนวน k กลุ่ม จากนั้นใช้ข้อมูลจำนวน k-1 กลุ่มในการฝึกสอนระบบและทดสอบระบบด้วยข้อมูล 1 กลุ่มที่เหลือ เรียกวิธีนี้ว่า k-fold cross validation สำหรับกรณีที่ k เท่าจำนวนข้อมูลทั้งหมดจะเรียกวิธีนี้ว่า jackknife test หรือ leave-one-out และ (3) คือวิธีการทดสอบด้วยข้อมูลที่ไม่เกี่ยวข้องกับชุดข้อมูลสอนระบบ เรียกวิธีนี้ว่า independent test ในงานวิจัยนี้ จะแสดงผลการทดสอบประสิทธิภาพแบบจำลองด้วยวิธีการทดสอบทั้ง 3 วิธีดังกล่าว

ข้างต้น และใช้วิธีการคำนวณค่าความถูกต้องของแบบจำลองสำหรับแต่ละคลาสและโดยรวม ดังสมการที่ (31-32)

$$accuracy(i) = \frac{p(i)}{obs(i)} \quad \textbf{(31)}$$

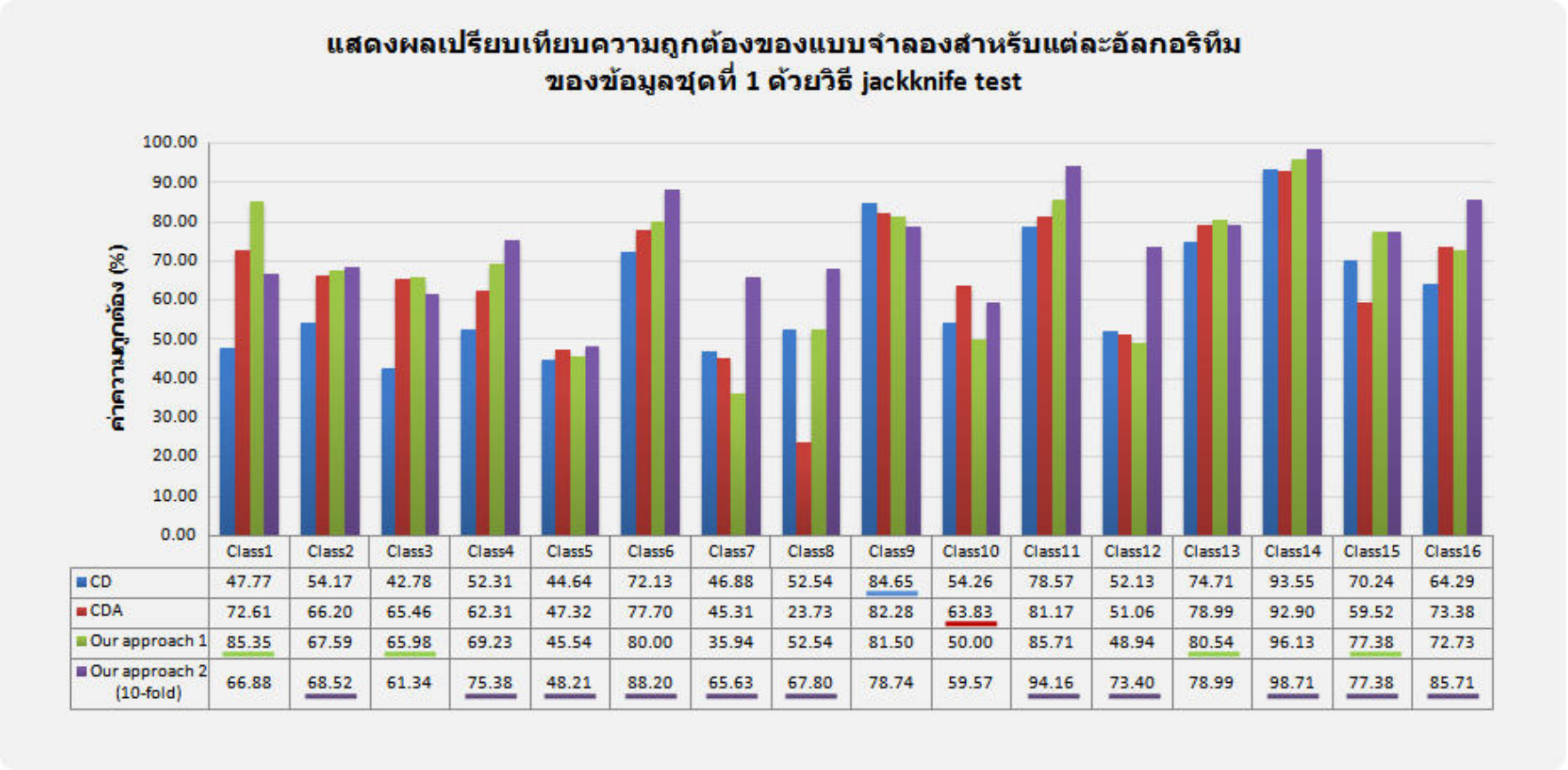
$$total \quad accuracy = \frac{\sum_{i=1}^{k} p(i)}{N} \quad \textbf{(32)}$$

โดยที่ ค่า $p(i)$ คือจำนวนโปรตีนที่ทำนายว่าอยู่คลาส i ค่า $obs(i)$ คือจำนวนโปรตีนทั้งหมดในคลาส i ค่า N คือจำนวนโปรตีนทั้งหมด และค่า k คือจำนวนคลาสของซับแฟมิลี



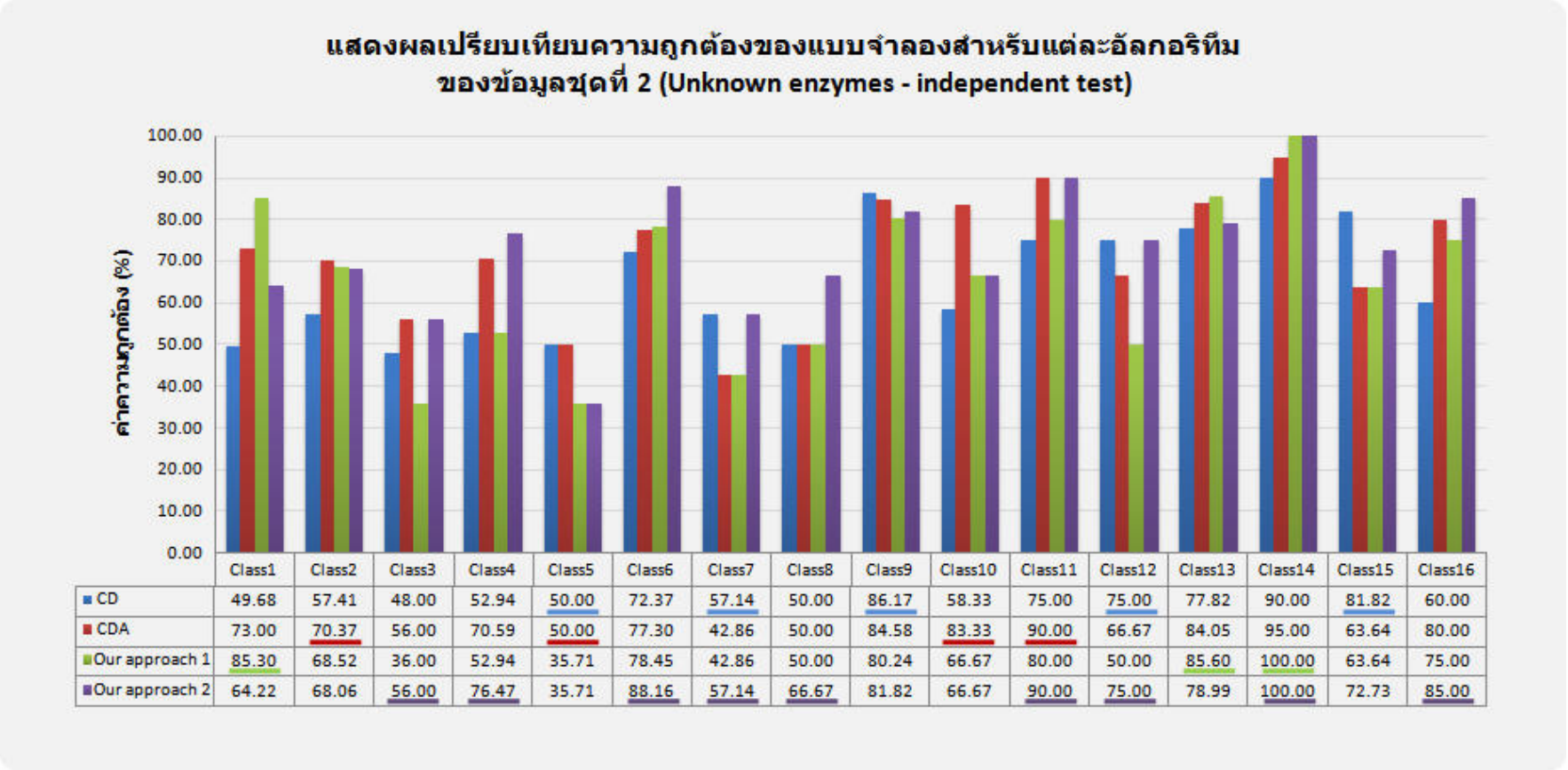
	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8	Class9	Class10	Class11	Class12	Class13	Class14	Class15	Class16
CD	58.92	64.81	55.67	69.23	65.18	72.79	85.94	96.61	89.37	67.02	87.66	80.85	79.38	97.42	96.43	81.17
CDA	89.49	87.96	85.57	93.08	83.04	85.57	96.88	100	93.7	95.74	96.75	97.87	96.5	100	100	93.51
Our approach 1	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	95.67	92.55	100.00	100.00	100.00	100.00	100.00	100.00
Our approach 2	85.67	97.22	91.75	96.92	90.18	99.67	90.63	93.22	83.86	78.72	98.70	91.49	96.89	100.00	96.43	92.86

ภาพที่ 51 แสดงความแม่นยำ (แยกตามคลาส) ของระบบจำแนกเอนไซม์ซับแฟมิลี ที่ได้จากข้อมูลทดสอบระบบชุดที่ 1 จำนวน 2640 เอนไซม์ ทดสอบด้วยวิธี self-consistency test



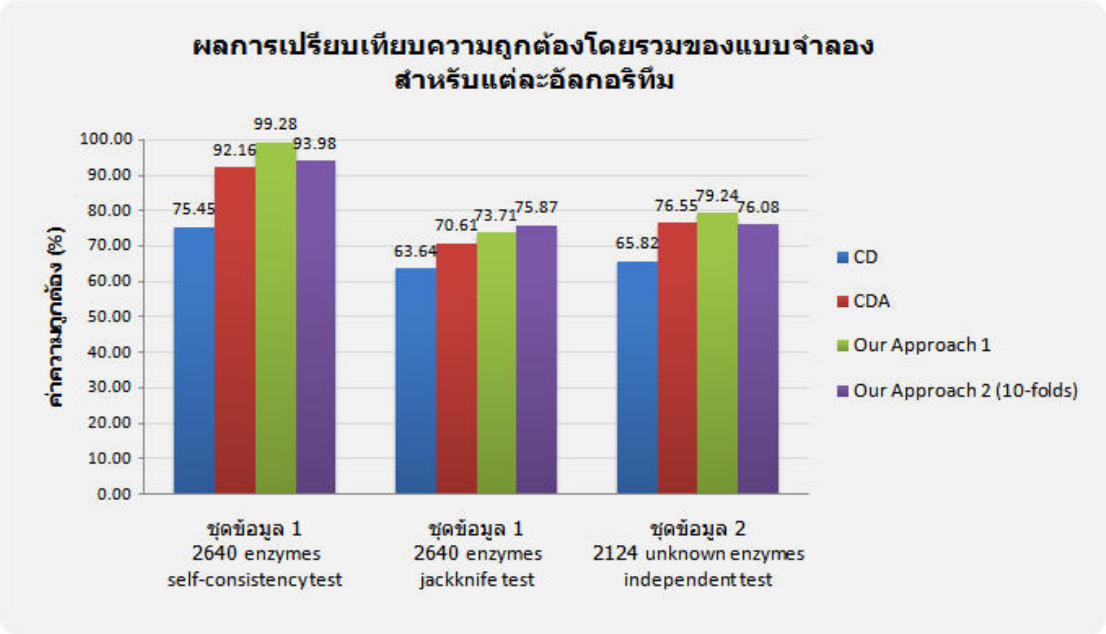
	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8	Class9	Class10	Class11	Class12	Class13	Class14	Class15	Class16
CD	47.77	54.17	42.78	52.31	44.64	72.13	46.88	52.54	84.65	54.26	78.57	52.13	74.71	93.55	70.24	64.29
CDA	72.61	66.20	65.46	62.31	47.32	77.70	45.31	23.73	82.28	63.83	81.17	51.06	78.99	92.90	59.52	73.38
Our approach 1	85.35	67.59	65.98	69.23	45.54	80.00	35.94	52.54	81.50	50.00	85.71	48.94	80.54	96.13	77.38	72.73
Our approach 2 (10-fold)	66.88	68.52	61.34	75.38	48.21	88.20	65.63	67.80	78.74	59.57	94.16	73.40	78.99	98.71	77.38	85.71

ภาพที่ 52 แสดงความแม่นยำ (แยกตามคลาส) ของระบบจำแนกเอนไซม์ซับแฟมิลี ที่ได้จากข้อมูลทดสอบระบบชุดที่ 1 จำนวน 2640 เอนไซม์ ทดสอบด้วยวิธี jackknife test



	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8	Class9	Class10	Class11	Class12	Class13	Class14	Class15	Class16
CD	49.68	57.41	48.00	52.94	50.00	72.37	57.14	50.00	86.17	58.33	75.00	75.00	77.82	90.00	81.82	60.00
CDA	73.00	70.37	56.00	70.59	50.00	77.30	42.86	50.00	84.58	83.33	90.00	66.67	84.05	95.00	63.64	80.00
Our approach 1	85.30	68.52	36.00	52.94	35.71	78.45	42.86	50.00	80.24	66.67	80.00	50.00	85.60	100.00	63.64	75.00
Our approach 2	64.22	68.06	56.00	76.47	35.71	88.16	57.14	66.67	81.82	66.67	90.00	75.00	78.99	100.00	72.73	85.00

ภาพที่ 53 แสดงความแม่นยำ (แยกตามคลาส) ของระบบจำแนกเอนไซม์ซับแฟมิลี สำหรับข้อมูลชุดที่ 2 (unknown enzymes) จำนวน 2124 เอนไซม์ (independent test)



ภาพที่ 54 แสดงความแม่นยำโดยเฉลี่ยของระบบจำแนกเอนไซม์ซับแฟมิลี ของข้อมูลชุดที่ 1 และ 2 ด้วยวิธีการทดสอบ 3 แบบ (self-consistency jackknife และ independent test)

ผลจากการทดสอบด้วยวิธี self-consistency (แสดงดังภาพที่ 51) พบว่าอัลกอริทึมของวิธีการ 1 และวิธีการที่ 2 ให้ค่าความถูกต้องแม่นยำถึง 99.28% และ 93.98% ตามลำดับ ซึ่งมีค่าสูงเนื่องจากข้อมูลที่ใช้ทดสอบระบบคือข้อมูลที่ใช้ในการฝึกสอนนั่นเอง ผลจากการทดลองด้วยวิธี self-consistency นั้น สามารถบ่งชี้ได้ว่าแบบจำลองนั้นมีประสิทธิภาพระดับหนึ่ง แต่อย่างไรก็ดีการทดสอบด้วยวิธีการนี้ไม่สามารถเป็นตัวชี้วัดถึงคุณภาพของแบบจำลองได้ทั้งหมด

ผลจากการทดสอบด้วยวิธี jackknife test (แสดงดังภาพที่ 52) เป็นวิธีการทดสอบที่มีน้ำหนักมากที่สุดที่ใช้ในการวัดประสิทธิภาพของแบบจำลอง (Chou and Elrod, 2003) จากการทดลองพบว่าอัลกอริทึมของงานวิจัยนี้ให้ค่าความถูกต้องแม่นยำโดยรวมของระบบ (แสดงดังภาพที่ 54) ถึง 73.71% สำหรับวิธีการที่ 1 และเพิ่มขึ้นเป็น 75.87 สำหรับวิธีการที่ 2 เมื่อเพิ่มองค์ความรู้ในการแทนที่กรดอะมิโนในขั้นตอนการจัดกลุ่ม SSCP ซึ่งทั้งสองอัลกอริทึมให้ความถูกต้องโดยรวมมากกว่าอัลกอริทึม CD, และ CDA เมื่อพิจารณาความถูกต้องในแต่ละคลาสพบว่าอัลกอริทึมของวิธีการที่ 1 ให้ผลค่าความถูกต้องมากกว่างานวิจัย CD และ CDA ถึง 10 คลาสได้แก่ คลาส 1, 2, 3, 4, 6, 8, 11, 13, 14 และ 15 สำหรับคลาสที่แม่นยำน้อยที่สุดคือ คลาส 7 (ให้ค่าความถูกต้องเพียง 35.94%) เนื่องจากข้อมูลฝึกสอนและข้อมูลที่ทราบโครงสร้าง 3 มิติ (known-PDB) ของคลาส 7 นั้น

มีจำนวนน้อยมาก ซึ่งผลการทดลองของอัลกอริทึม CD และ CDA ก็ให้ความถูกต้องน้อยเช่นกันคือ 46.88% และ 45.31% ตามลำดับ ในทางกลับกันคลาสที่ 1 ให้ความถูกต้องแม่นยำมากถึง 85.35% ซึ่งสูงกว่าอัลกอริทึม CDA ถึง 12.74% เนื่องจากข้อมูลฝึกสอนระบบและข้อมูลที่ทราบโครงสร้าง 3 มิตินั้นมีจำนวนมาก แต่อย่างไรก็ตามเมื่อพิจารณาที่คลาส 6, 9, 13 และ 15 แล้วนั้น พบว่าให้ค่าความถูกต้องสูงมาก (80.00%, 81.50%, 80.54% และ 77.38% ตามลำดับ) แม้ข้อมูลที่ทราบโครงสร้าง 3 มิติมีจำนวนน้อยมากเมื่อเทียบกับข้อมูลทั้งหมดของแต่คลาสนั้น คิดเป็นอัตราส่วนได้ดังนี้ คลาส 6 เท่ากับ 4.59% คลาส 9 เท่ากับ 2.36% คลาส 13 เท่ากับ 8.95% และคลาส 15 เท่ากับ 8.33% ตามลำดับ จากผลการทดลองพบว่าค่าความถูกต้องมีนัยสำคัญกับฟีเจอร์ที่สร้างจาก SSCP และฮิดเด้นมาร์คอฟโมเดล มากกว่าจำนวนข้อมูลฝึกสอนหรือข้อมูลโครงสร้าง 3 มิติ

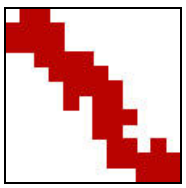
เมื่อพิจารณาค่าความถูกต้องแม่นยำในแต่ละคลาสของวิธีการที่ 2 เทียบกับงานวิจัย CD และ CDA พบว่าวิธีการที่ 2 ให้ผลค่าความถูกต้องมากกว่างานวิจัย CD และ CDA ถึง 11 คลาสได้แก่ คลาส 2, 4, 5, 6, 7, 8, 11, 12, 14, 15, และ 16 เสมอ 1 คลาสคือคลาส 13 เป็นที่น่าสังเกตว่าเมื่อเพิ่มองค์ความรู้การแทนที่ระหว่างกรดอะมิโนแล้วนั้น แบบจำลองให้ค่าความถูกต้องของคลาสที่ 7 และคลาสที่ 12 เพิ่มขึ้นถึง 20% และ 22% ตามลำดับ เนื่องจากในขั้นตอนการสร้างตัวแทนสายโปรตีนนั้น ยังคงอาศัยข้อมูลระดับปฐมภูมิซึ่งการเพิ่มองค์ความรู้การแทนที่ระหว่างกรดอะมิโนสามารถเพิ่มความชัดเจนในการแบ่งกลุ่มตามโครงสร้าง SSCP ได้ ในทางตรงกันข้ามกับคลาสที่ 1 นั้นสังเกตได้ว่าค่าความถูกต้องลดลงเหลือเพียง 66.88% ทั้งนี้การเพิ่มความรู้เกี่ยวกับการแทนที่กรดอะมิโนอาจส่งผลให้การแบ่งกลุ่ม SSCP มีจำนวนกลุ่มมากขึ้น ทำให้การแบ่งแยกคลาสที่ 1 ไม่ชัดเจน ส่วนคลาสที่ 14 เป็นคลาสที่ให้ค่าความถูกต้องสูงมาก ทั้งด้วยวิธีการที่ 1 และ 2 ทั้งนี้เนื่องจากข้อมูลในคลาสที่ 14 มีค่าความเหมือนระหว่างเอนไซม์มากที่สุด ประมาณ 26.92% นั่นเอง

ผลจากการทดสอบด้วยวิธี independent test พบว่าอัลกอริทึมของวิธีการที่ 1 ให้ค่าความถูกต้องถึง 79.24% และวิธีการที่ 2 ให้ค่าความถูกต้องถึง 76.08% ซึ่งเมื่อเทียบกับงานวิจัย CD และ CDA ถือว่าความถูกต้องดังกล่าวอยู่ในระดับดี อย่างไรก็ดีการทดสอบด้วยวิธีนี้เป็นการทดสอบเพื่อวัดประสิทธิภาพของระบบในเบื้องต้น โดยใช้ชุดข้อมูลที่ไม่เกี่ยวข้องกับข้อมูลฝึกสอนระบบ (unknown enzymes)

2.2 การวิเคราะห์โครงสร้างโปรตีนจากองค์ความรู้ที่ได้จากระบบจำแนกประเภทเอนไซม์ซับแฟมิลีจากงานวิจัยนี้

ผลลัพธ์อีกด้านหนึ่งที่ได้จากระบบจำแนกประเภทเอนไซม์ซับแฟมิลี่ของงานวิจัยนี้และไม่พบในงานวิจัยอื่นคือ องค์ความรู้เกี่ยวกับโครงสร้างโปรตีนบนตารางเมตริซ์คอนแทคแมป เนื่องจากฟีเจอร์ที่ใช้เป็นตัวแทนสายโปรตีนนั้น ได้มาจากโครงสร้างย่อยบนตารางเมตริกซ์คอนแทคแมป และนำมาสร้างตัวแทนสายโปรตีนด้วยฮิดเด้นมาร์คอฟโมเดล ดังนั้นเราสามารถใช้ประโยชน์จากข้อมูลที่ได้จากฮิดเด้นมาร์คอร์ฟโมเดล เพื่อสืบค้นกลับไปได้ว่าโปรตีนดังกล่าวมีโครงสร้างย่อยเป็นอย่างไร และอยู่บริเวณใดของตารางเมตริกซ์คอนแทคแมป

ยกตัวอย่างเช่น เอนไซม์ NADH Oxidase (มี accession number เป็น Q60049 และมี pdb id เป็น 1NOX) ความยาว 205 อะมิโน จากการแทนสายโปรตีนด้วยฟีเจอร์ที่สร้างขึ้นพบว่าฟีเจอร์ที่มีค่าเป็น 2 นั้นคือ $pHMM_{19}$-X และ $pHMM_{19}$-Y เมื่อทำการสืบค้นกลับไปยังกลุ่มของ SSCP ของฟีเจอร์นี้พบว่า มีฟีเจอร์คือความกว้างประมาณ 12 กรดอะมิโน ความสูงประมาณ 13 กรดอะมิโน จำนวนที่คู่กรดอะมิโนติดกันภายใน SSCP นี้มีจำนวนประมาณ 41 กรดอะมิโน มุมของ SSCP ประมาณ 136.25 องศาตามแนวแกน x ค่ากระจายตัวบนแนวแกน x ประมาณ 0.053 และค่ากระจายตัวบนแนวแกน y ประมาณ 0.056 ดังนั้นสามารถทราบโครงสร้างย่อยคร่าวๆ ได้เป็นดังภาพที่ 55 ซึ่งเป็นโครงสร้างแบบเบต้าชีทนั่นเอง



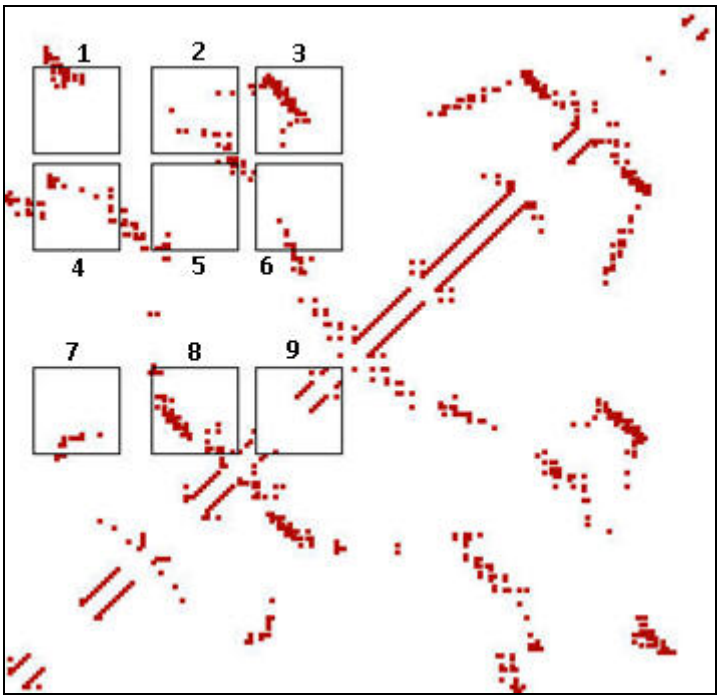
ภาพที่ 55 แสดง SSCP ที่สืบค้นกลับไปได้ จากแบบจำลองในงานวิจัยนี้

เพื่อให้ทราบตำแหน่งที่ SSCP นั้น จะปรากฏบนตารางเมตริกซ์คอนแทคแมป สามารถทำได้โดยสืบค้นกลับไปที่ฮิดเด้นมาร์คอฟโมเดลที่ 19-X และ 19-Y ซึ่งพบว่าฟีเจอร์ทั้งสองให้ตำแหน่งการสืบค้นพบรูปแบบลำดับอะมิโนที่ตรงกันกับสายโปรตีน 1NOX ได้ฟีเจอร์ละ 3 โดเมน แสดงดังตารางที่ 9

ตารางที่ 9 แสดงตำแหน่งที่โพรไฟล์ฮิดเด้นมาร์คอฟโมเดลสืบค้นพบรูปแบบอะมิโนบนโปรตีน 1NOX

โดเมนที่	ฟีเจอร์ $pHMM_{19}$-X		ฟีเจอร์ $pHMM_{19}$-Y	
	ตำแหน่งเริ่มต้น	ตำแหน่งสิ้นสุด	ตำแหน่งเริ่มต้น	ตำแหน่งสิ้นสุด
โดเมน 1	10	33	71	95
โดเมน 2	43	66	130	154
โดเมน 3	72	95	158	182

จากตารางที่ 9 สามารถสืบค้นกลับไปบนตารางเมตริกซ์คอนแทคแมปของโปรตีน 1NOX ได้ โดยทำการจับคู่ทุกคู่ของโดเมนที่ได้จากฟีเจอร์ $pHMM_{19}$-X และ $pHMM_{19}$-Y ดังนั้นจะได้ตำแหน่งที่เป็นไปได้ทั้งหมด 9 ตำแหน่งด้วยกัน แสดงดังภาพที่ 56



ภาพที่ 56 แสดงตำแหน่งของ SSCP ที่สามารถทำนายได้ บนตารางคอนแทคแมปเมตริกซ์ของโปรตีน 1NOX โดยกรอบสี่เหลี่ยม 6 กรอบคือบริเวณที่คาดว่าน่าจะพบ SSCP

จากภาพที่ 56 บริเวณกรอบสี่เหลี่ยม 9 บริเวณนั้น พบว่ามี 2 บริเวณ (หมายเลข 3 และหมายเลข 8) ที่ปรากฏโครงสร้าง SSCP ที่มีลักษณะคล้ายกับ SSCP ที่ประมาณได้ในภาพที่ 50 ส่วนบริเวณหมายเลข 1 นั้น ถึงแม้บริเวณที่ทำนายได้จะไม่ตรงกับตำแหน่งที่แท้จริง แต่ครอบคลุมรูปร่างบางส่วนของ SSCP ที่มีลักษณะคล้ายกับ SSCP ในภาพที่ 55 เช่นกัน ส่วนบริเวณอื่นนั้น แม้จะไม่สามารถทำนายตำแหน่งได้ถูกต้อง แต่ก็ยังคงพบว่าภายในบริเวณที่ทำนายนั้นพบตำแหน่งที่มีการติดกันของกรดอะมิโนได้ ทั้งนี้เนื่องมาจากการขยาย sub-sequence ด้านละ 7 อะมิโน ในขั้นตอนการสร้างฟีเจอร์นั่นเอง

วิจารณ์

จากผลการทดลองที่ได้นั้น ในด้านการวัดประสิทธิภาพของระบบจำแนกประเภทเอนไซม์ซับแฟมิลีนั้น สามารถวิเคราะห์และวิจารณ์ได้เป็นหัวข้อหลักดังนี้

1. ด้านประสิทธิภาพความถูกต้องแม่นยำของระบบจำแนกประเภทเอนไซม์ซับแฟมิลี

วิธีการวัดผลที่สำคัญและมีน้ำหนักมากที่สุดคือการวัดผลแบบ jackknife test (Chou and Elrod, 2003) จากผลการทดลองพบว่าอัลกอริทึมของงานวิจัยนี้ให้ค่าความถูกต้องแม่นยำมากที่สุด (73.71% สำหรับวิธีการที่ 1 และ 75.87% สำหรับวิธีการที่ 2) เมื่อเปรียบเทียบกับอัลกอริทึม CD และ CDA และเมื่อพิจารณาระดับคลาสพบว่าอัลกอริทึมของวิธีการที่ 1 ให้ค่าความถูกต้องมากกว่าอัลกอริทึม CD และ CDA ถึง 10 คลาส ส่วนวิธีการที่ 2 ให้ค่าความถูกต้องมากกว่าอัลกอริทึม CD และ CDA ถึง 11 คลาส เสมอ 1 คลาส และเมื่อทดสอบด้วยวิธี self-consistency test และ independent test ซึ่งเป็นวิธีการที่สำคัญในการบ่งชี้คุณภาพของแบบจำลองในเบื้องต้น แต่มีนัยสำคัญน้อยกว่าวิธี jackknife test พบว่าค่าความถูกต้องที่วัดได้จากวิธี self-consistency test ให้ค่าความถูกต้องถึง 99.28% และ 93.98% สำหรับวิธีการที่ 1 และ 2 ตามลำดับ ซึ่งให้ค่าความถูกต้องสูง ส่วนวิธี independent test นั้น ให้ความถูกต้องถึง 79.24% และ 76.08% ตามลำดับ ซึ่งให้ค่าความถูกต้องได้ดีระดับหนึ่งเมื่อเทียบกับอัลกอริทึม CD และ CDA

2. ความสัมพันธ์ของข้อมูลฟีเจอร์ที่ใช้เป็นตัวแทนสายเอนไซม์กับคลาสของเอนไซม์

งานวิจัยนี้ใช้ข้อมูลฟีเจอร์จากข้อมูลโครงสร้างย่อย (SSCP) ที่ปรากฏบนตารางคอนแทคแมป ซึ่งสามารถสืบค้นได้จากโครงสร้าง 3 มิติของโปรตีน แต่อย่างไรก็ตามเมื่อพิจารณาอัตราส่วนระหว่างจำนวนโปรตีนทั้งหมดที่ใช้ฝึกสอนระบบ (ข้อมูลชุดที่ 1) กับข้อมูลที่ทราบโครงสร้าง 3 มิติแล้ว พบว่าข้อมูลที่ทราบโครงสร้าง 3 มิตินั้น มีอยู่เพียง 11.21% (296/2640 โปรตีน) ซึ่งเป็นจำนวนที่น้อยมาก แต่ผลการทดสอบระบบพบว่าสามารถให้ค่าความถูกต้องได้สูงถึง 73.71% และ 75.87% สำหรับวิธีการที่ 1 และ 2 ตามลำดับ ซึ่งสูงกว่าอัลกอริทึมอื่นที่ใช้ในการเปรียบเทียบ นอกจากนี้เมื่อพิจารณาค่าความถูกต้องแม่นยำในระดับคลาสของวิธีการที่ 1 แล้ว พบว่าคลาส 6, 9, 13 และ 15 ให้ค่าความถูกต้องสูงมาก (80.00%, 81.50%, 80.54% และ 77.38% ตามลำดับ) แม้ข้อมูลที่ทราบโครงสร้าง 3 มิติมีจำนวนน้อยมากเมื่อเทียบกับข้อมูลทั้งหมดของแต่ละคลาสนั้น คิดเป็นอัตราส่วนได้ดังนี้ คลาส 6 เท่ากับ 4.59% คลาส 9 เท่ากับ 2.36% คลาส 13 เท่ากับ 8.95% และคลาส 15 เท่ากับ 8.33% ตามลำดับ จากผลการทดลองจึงสามารถสรุปได้ว่าฟีเจอร์ที่สร้างจาก SSCP และฮิดเด้นมาร์คอฟโมเดล มีนัยสำคัญต่อการทำนายฟังก์ชั่นของเอนไซม์

เมื่อพิจารณาค่าความถูกต้องแม่นยำในระดับคลาสของวิธีการที่ 2 ที่มีการเพิ่มองค์ความรู้การแทนที่ระหว่างกรดอะมิโนแล้ว พบว่าคลาส 6, 7, 8, 12 และ 16 สามารถเพิ่มค่าความถูกต้องขึ้นอีก 18.20%, 18.75%, 15.26%, 21.27% และ 12.33% ตามลำดับ

3. ด้านการวิเคราะห์โครงสร้างของโปรตีนโดยการสืบค้นกลับไปยังฟีเจอร์ที่สร้างจาก SSCP และโพรไฟล์ฮิดเด้นมาร์คอฟโมเดล

ข้อดีของการใช้โพรไฟล์ฮิดเด้นมาร์คอฟโมเดลเป็นตัวแทนสายโปรตีนคือสามารถสืบค้นกลับไปถึงตำแหน่งของ SSCP บนตารางเมตริกซ์คอนแทคแมปได้ แต่ตำแหน่งของ SSCP ที่ได้นั้นยังไม่ถูกต้องทั้งหมด แต่สามารถบ่งบอกบริเวณที่มีโอกาสปรากฏ SSCP ได้ ทั้งนี้เนื่องจากในขั้นตอนการสร้างตัวแทนสายโปรตีนด้วยโพรไฟล์ฮิดเด้นมาร์คอฟนั้น ใช้ข้อมูลลำดับกรดอะมิโนตรงๆ ทั้งนี้สามารถปรับปรุงฟีเจอร์ได้ โดยการเพิ่มหรือแปลงข้อมูลทางเคมีที่สำคัญและเป็นปัจจัยให้เกิดการติดกันของคู่ลำดับกรดอะมิโนเป็นต้น

สำหรับโครงสร้าง SSCP ที่ทำการสืบค้นกลับนั้น เกิดจากการหาค่าเฉลี่ยของฟีเจอร์แต่ละชนิดที่ใช้ในการจัดกลุ่ม ได้แก่ความกว้างของ SSCP ความสูงของ SSCP จำนวนคู่กรดอะมิโนที่ติดกันภายใน SSCP มุมของการกระจายตัวของการติดกันบนแนวแกน x ค่าการกระจายของการติดกันบนแนวแกน x และค่าการกระจายตัวของการติดกันบนแนวแกน y ทั้งนี้ SSCP ที่ได้ดังกล่าวจึงเป็นเพียงรูปร่างประมาณ แต่อย่างไรก็ตามรูปร่างดังกล่าว สามารถบ่งบอกชนิดของโครงสร้างทุติยภูมิของโปรตีนได้ว่าเป็นโครงสร้างแบบเบต้าหรืออัลฟา และเมื่อนำ SSCP ที่ทำนายได้ไปรวมกับการทำนายตำแหน่ง ทำให้ทราบโครงสร้างทุติยภูมิที่ชัดเจนขึ้น เช่นเป็นโครงสร้างทุติยภูมิแบบเบต้าชนิดทิศทางเดียวกัน หรือเบต้าชนิดทิศทางที่สวนทางกัน เป็นต้น

สรุปและข้อเสนอแนะ

สรุป

วิธีการพัฒนาระบบจำแนกประเภทเอนไซม์ซับแฟมิลีในงานวิจัยนี้นั้น อาศัยความรู้ที่ว่าการทำงานของเอนไซม์นั้น ถูกกำหนดขึ้นจากบริเวณส่วนย่อยของลำดับกรดอะมิโนในสายโปรตีน (Tian *et al.* 2004; Pandey *et al.*, 2006) และเนื่องจากฟังก์ชั่นเอนไซม์มีความสัมพันธ์อย่างใกล้ชิดกับโครงสร้าง 3 มิติ (Gu *et al.*, 2008; Zhang *et al.*, 2008) ดังนั้นแทนการพิจารณาสายลำดับอะมิโนทั้งเส้น ในงานวิจัยนี้จึงตั้งสมมติฐานว่าบริเวณส่วนย่อยของลำดับกรดอะมิโนที่มีโครงสร้างระดับทุติยภูมิอยู่ติดกันเมื่อพิจารณาบนโดเมน 3 มิตินั้น เป็นบริเวณสำคัญที่ส่งผลต่อการเกิดฟังก์ชั่นของเอนไซม์นั่นเอง ดังนั้นเพื่อหาบริเวณส่วนย่อยของลำดับอะมิโนที่สำคัญนั้น ในงานวิจัยนี้ได้พัฒนาวิธีการสืบค้นฟีเจอร์เพื่อใช้เป็นตัวแทนสายข้อมูลโปรตีนจากข้อมูลตารางเมตริกซ์คอนแทคแมป ซึ่งเป็นการพิจารณาการติดกันระหว่างคู่กรดอะมิโนทุกคู่บนโดเมน 3 มิติ และนำมาแสดงบนตารางเมตริกซ์ขนาด $N \times N$ โดย N คือความยาวของสายโปรตีน ซึ่งง่ายต่อการพิจารณาและนำไปใช้งานต่อ รูปร่างที่ปรากฏบนตารางเมตริกซ์คอนแทคแมปนั้น แสดงการติดกันของโครงสร้างทุติยภูมิของกรดอะมิโนนั่นเอง โดยตำแหน่งของการติดกัน สามารถบ่งบอกทิศทางของโครงสร้างทุติยภูมิชนิดเบต้าได้ด้วย ดังนั้นในงานวิจัยนี้จึงใช้ประโยชน์จากข้อมูลตารางเมตริกซ์คอนแทคแมปของโปรตีน ถึงแม้ข้อมูลที่ใช้ในการสืบค้นฟีเจอร์จะมีอยู่เป็นจำนวนน้อยมากเมื่อเทียบกับข้อมูลทดสอบระบบทั้งหมดที่ใช้ในงานวิจัยนี้ (11.21%) แต่กระนั้นด้วยวิธีการสืบค้น SSCP วิธีการจัดกลุ่มข้อมูลตามรูปแบบ SSCP เนื่องจากมีสมมติฐานว่า sub-sequence ที่มีโครงสร้าง SSCP ใกล้เคียงกัน จะมีกรดอะมิโนที่คล้ายคลึงกัน และตลอดจนถึงวิธีการสร้างตัวแทนสายโปรตีนสำหรับเอนไซม์ใหม่ที่ไม่ทราบโครงสร้างด้วยโพรไฟล์ฮิดเด้นมาร์คอฟโมเดลนั้น เป็นวิธีการที่มีประสิทธิภาพและส่งผลให้ระบบจำแนกประเภทข้อมูลเอนไซม์นี้มีความถูกต้องแม่นยำมากสุดถึง 75.87% ซึ่งมากกว่าอัลกอริทึม CD และ CDA ซึ่งเป็นวิธีการสืบค้นฟีเจอร์โดยพิจารณาความถี่ในการปรากฏกรดอะมิโน (amino acid composition) เป็นหลัก

นอกจากนี้ผลลัพธ์ที่ได้จากระบบจำแนกประเภทเอนไซม์ซับแฟมิลีของงานวิจัยนี้ ยังสามารถสืบค้นตำแหน่งและโครงสร้าง SSCP บนตารางเมตริกซ์คอนแทคแมปของโปรตีนใหม่ที่ไม่ทราบโครงสร้าง 3 มิติได้ด้วย ซึ่งทำให้ทราบชนิดและโครงสร้างทุติยภูมิโดยคร่าวของโปรตีนและนำไปวิเคราะห์ต่อได้ แต่เนื่องจากงานวิจัยนี้มุ่งเน้นพัฒนาระบบจำแนกประเภทเอนไซม์ซับแฟมิลี

ดังนั้นจึงยังไม่ได้ทำการพัฒนาความถูกต้องในส่วนการทำนายโครงสร้างและตำแหน่งหรือก็คือการทำนายคอนแทคแมปในส่วนงานวิจัยนี้

อย่างไรก็ดีจากผลการทดลองสามารถสรุปได้ว่าฟีเจอร์ที่สร้างได้จาก SSCP และโพรไฟล์ฮิดเด้นมาร์คอฟโมเดลนั้นมีนัยสำคัญเพียงพอสำหรับการนำไปใช้เป็นตัวแทนสายโปรตีนเพื่อใช้ในการสร้างระบบจำแนกประเภทเอนไซม์ซับแฟมมิลีที่มีค่าความเหมือน (similarity score) ระหว่างเอนไซม์น้อยกว่า 30%

ข้อเสนอแนะ

งานวิจัยนี้ยังสามารถทำการปรับปรุงเพื่อเพิ่มประสิทธิภาพได้ในหลายประเด็นดังนี้

ทำการปรับค่าตัดสินใจของการพิจารณาการติดกันระหว่างคู่กรดอะมิโนบนโดเมน 3 มิติให้น้อยลง (ในงานวิจัยนี้ใช้ค่าตัดสินใจเท่ากับ 8 Å โดยอ้างอิงจาก CASP6) เนื่องจากในขั้นตอนการสืบค้น SSCP นั้น จะทำให้ได้ SSCP ที่มีนัยสำคัญมากขึ้น กล่าวคือเป็นรูปร่างที่ชัดเจน สามารถระบุได้ว่าเป็นโครงสร้างทุติยภูมิชนิดใด เนื่องจากในงานวิจัยนี้ พบว่า SSCP ที่ได้นั้น บางส่วนมีขนาดเล็กและไม่สามารถระบุโครงสร้างทุติยภูมิได้ชัดเจน นอกจากจะพิจารณา SSCP ที่อยู่ใกล้กันร่วมด้วย จึงจะสามารถบ่งบอกโครงสร้างทุติยภูมิได้

ทำการพัฒนาวิธีการสืบค้นรูปแบบที่อยู่ติดกันและต่อเนื่องกันมากที่สุด โดยให้อัลกอริทึมพิจารณาบริเวณที่ไม่ติดกันร่วมด้วย โดยมีเงื่อนไขว่าบริเวณที่ไม่ติดกันนั้นต้องอยู่ห่างจากบริเวณที่ติดกันไม่เกินกี่อะมิโนเป็นต้น ทั้งนี้จะทำให้ได้ SSCP ที่มีนัยสำคัญมากขึ้นด้วยเช่นกัน

ทำการพัฒนาวิธีการสร้างตัวแทนสายข้อมูลด้วยฮิดเด้นมาร์คอฟโมเดล โดยทำการแปลงข้อมูลกรดอะมิโนเป็นข้อมูลคุณสมบัติทางเคมีที่สำคัญของกรดอะมิโนแทน เช่นคุณสมบัติไฮโดรโฟบิกซ์ ไฮโดรฟิลิกซ์ หรือแรงวัลเดอร์วาวล์ เป็นต้น เนื่องจากคุณสมบัติทางเคมีเหล่านี้มีผลต่อแรงผลักและแรงดูดระหว่างกรดอะมิโน

ทำการพัฒนาวิธีการกรองฟีเจอร์ฮิดเด้นมาร์คอฟโมเดล เนื่องจากฟีเจอร์บางตัวที่ได้นั้นมีความคลุมเครือไม่เหมาะสมเพียงพอเพื่อใช้ในการเป็นตัวแทนในการจำแนกคลาสของเอนไซม์ซับแฟมิลี

เอกสารและสิ่งอ้างอิง

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

สิริวรรณ แต้วิจิตร, ธนาวินท์ รักธรรมานนท์, ธนภัทร ฆังคะจิตร และ กฤษณะ ไวยมัย. 2551. การวิเคราะห์โครงสร้างและจำแนกประเภทเอนไซม์ซับแฟมิลี โดยพิจารณาการติดกันของคู่ลำดับอะมิโนและโพรไฟล์ฮิดเด้นมาร์คอฟโมเดล, น. 659-667. *ใน* N**ational Computer Science and Engineering Conference ครั้งที่ 11.**

เอกสิทธิ์ พัชรวงศักดา และ สุภาวดี อิงศรีสว่าง. 2548. การทำนายแหล่งที่อยู่ของโปรตีนภายในเซลล์แบคทีเรียชนิด แกรมลบ ด้วยวิธี Decision Directed Acyclic Graph Support Vector Machines, น. 161-170. *ใน* N**ational Computer Science and Engineering Conference ครั้งที่ 9**.

Abascal, F. and A. Valencia. 2003. Automatic annotation of protein function based on family identification. **Proteins: Structure, Function, and Bioinformatics** (53(3)): 683-692.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers and D.J. Lipman. 1990. Basic Local Alignment Search Tool. **J. Mol. Biol.** (215): 403-410.

Altschul, S.F., T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Research.** (25(17)): 3389–3402.

Andrade, M.A., N.P. Brown, C. Leroy, S. Hoersch, A. Daruvar, C. Reich, A. Franchini, J. Tamames, A. Valencia, C. Ouzounis and C. Sander. 1999. Automated genome sequence analysis and annotation. **Bioinformatics** (15(5)): 391-412.

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

Berman, H.M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov and P.E. Bourne. 2000. The Protein Data Bank. **Nucleic Acids Research**. 28: 235-242. Available Source: www.pdb.org, April 05, 2009.

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

Brigitte, B., A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout and M. Schneider**.** 2003. The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. **Nucleic Acids Research**. 31: 365-370. Available Source: www.expasy.ch/enzyme, April 05, 2009.

Chou, K.C. 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. **Bioinformatics**. 21(1): 10-19.

Chou, K.C. and D.W. Elrod. 2003. Prediction of Enzyme Family Classes. **J. Proteome Res**. 2(2): 183-190.

Chou, K.C., H.B. Shen. 2008. Cell-PLoc: a package of web-servers for predicting subcellular localization of proteins in various organisms. **Nature Protocols** (3): 153-162.

Dayhoff, M.O, R.M. Schwartz and B.C. Orcutt. 1978. **A model of evolutionary change in proteins.** National biomedical research foundation, Washington DC.

Eddy, S.R. 1998. Profile hidden Markov models. **Bioinformatics**. 14(9): 755-763.

Emanuelsson, O., S. Brunak, G.V. Heijne and H. Nielsen. 2007. Locating proteins in the cell using targetp, signalp and related tools. **Nature Protocols** (2): 953:971.

Feng, D.F. and R.F. Doolittle. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. **J Mol Evol** (25): 351-360.

Gu, F., C. Hang and J. Ni. 2008. Protein structural class prediction based on an improved statistical strategy. **BMC Bioinformatics** 9 (Suppl):S5.

Gonzalez, Juan Ramon, and David A Pelta. 2007. On Using Fuzzy Contact Maps for Protein Structure Comparison. *In* **Proceedings of FUZZ-IEEE 2007.**

Grana, O., D. Baker, R.M. MacCallum, J. Meiler, M. Punta, B. Rost, M.L. Tress and A. Valencia. 2004. CASP6 assessment of contact prediction, *In* **Critical assessment of methods of protein structure prediction (CASP) 6.**

Henikoff, S. and J.G. Henikoff. 1992. Amino acid substitution matrices from protein blocks, pp. 10915-10919. *In* **Proc. Natl. Acad. Sci.**

Herrero, J., F. Al-Shahrour, R. Díaz-Uriarte, A. Mateos, J.M. Vaquerizas, J. Santoyo and J. Dopazo. 2003. GEPAS: a web-based resource for microarray gene expression data analysis. **Nucleic Acids Research** (31(13)): 3461-3467.

Hu, J., X. Shen, Y. Shao, C. Bystroff and M.J. Zaki. 2002. Mining Protein Contact Maps. *In* **2nd BIOKDD Workshop on Data Mining in Bioinformatics.**

Kohonen, T. 1982. Self-organized formation of topologically correct feature maps. **Biological Cybernetics** (43): 59-69.

Koski, L.B., M.W. Gray, B.F. Lang and G. Burger. 2005. AutoFACT: An Automatic Functional Annotation and Classification Tool. **Bioinformatics** (6):151.

Krogh, A., M. Brown, I.S. Mian, K. Sjölander and D. Haussler. 1994. **Hidden Markov Models in Computational Biology: Applications to Protein Modeling UCSC-CRL-93-32.** 62.

Pelta, D., J.R. Gonzalez and N. Krasnogor. 2005. Protein Structure Comparison through Fuzzy Contact Maps, pp. 1124-1129. *In* **Proceedings of Fourth Conference of the European Society for Fuzzy Logic and Technology and 11 Rencontres Francophones sur la Logique Floue et ses Applications (EUSFLAT-LFA, 2005)**.

Riley, M.L.**,** T. Schmidt**,** C. Wagner**,** H.W. Mewes and D. Frishman**.** 2005. The PEDANT genome database in 2005. **Nucleic Acids Research (**33).

Sasson, O., N. Kaplan and M. Linial. 2006. Functional annotation prediction: All for one and one for all. **Protein Science** (15(6)): 1557-1562.

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

Vapnik, V.N. 1995. **The Nature of Statistical Learning Theory**. Springer. New York.

Vassura, M., L. Margara, P.D. Lena, F. Medri, P. Fariselli and R. Casadio. 2008. Reconstruction of 3D Structures from Protein Contact Maps. **IEEE/ACM Transactions on Computational Biology and Bioinformatics (**5(3)): 357-367.

Waterman, M.S., T.F. Smith and W.A. Beyer. 1976. Some biological sequence metrics. **Adv. Math** (20): 367–387.

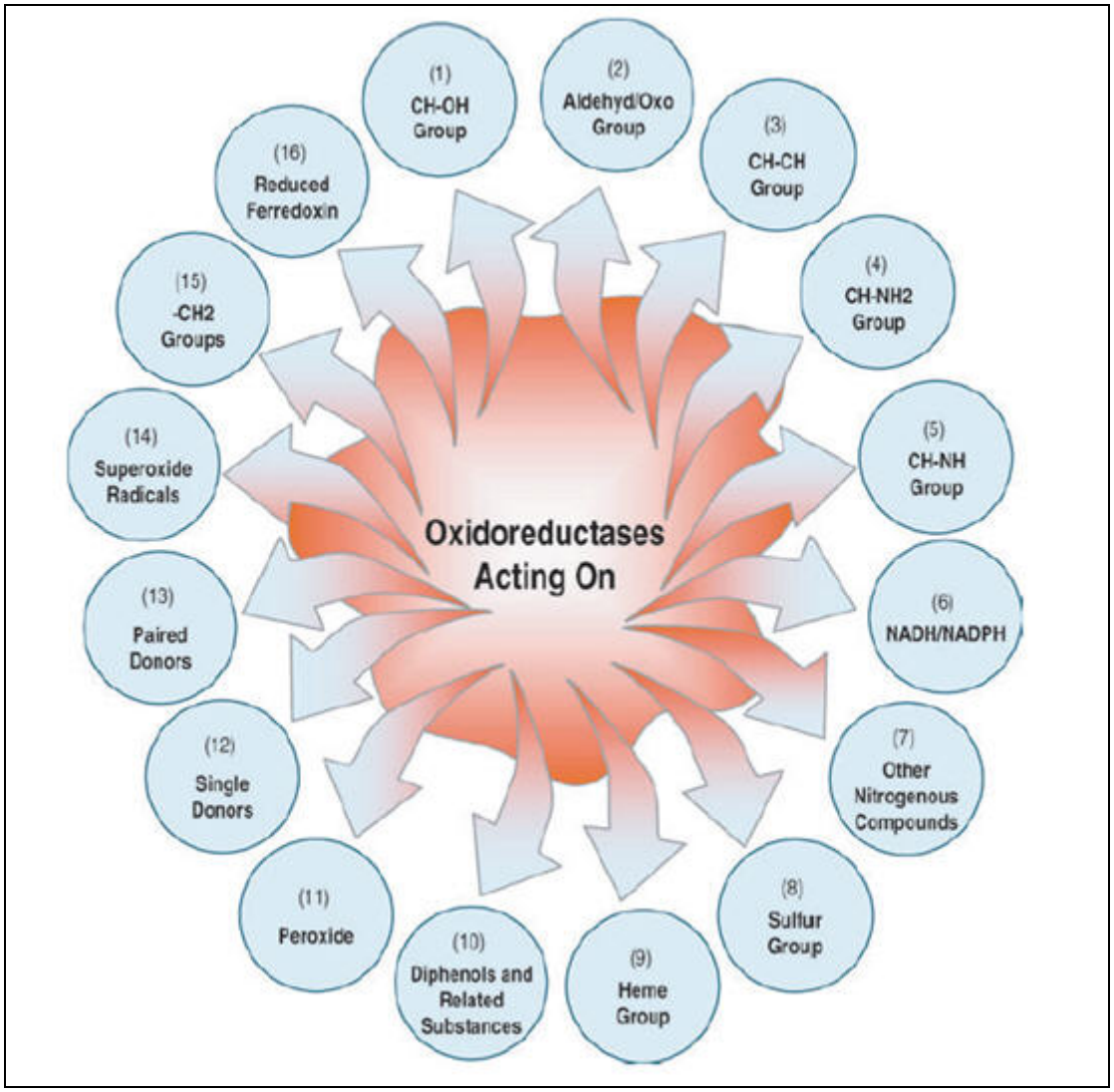
Xie, H., A. Wasserman, Z. Levine, A. Novik, V. Grebinskiy, A. Shoshan and L. Mintz. 2002. Large-scale protein annotation through Gene Ontology. **Genome Research** (12(5)): 785-794.

Yang, H., K. Marsolo, S. Parthasarathy and S. Mehta. 2004. Discovering Spatial Relationships Between Approximately Equivalent Patterns in Contact Maps, pp. 62-71. *In* **BIOKDD.**

Zhang, T.L., Y.S. Ding and K.C. Chou. 2008. Prediction protein structural classes with pseudo-amino acid composition: Approximate entropy and hydrophobicity pattern. **J Theor Biol** (250): 186–193.

ภาคผนวก

ข้อมูลเอนไซม์และข้อมูลซับแฟมิลีคลาสที่ใช้ในงานวิจัยนี้



ภาพผนวกที่ 1 ไดอะแกรมแสดงซับแฟมิลีคลาสของเอนไซม์อ็อกซิโดรีดักเทส (Oxidoreductases) ที่ใช้ในงานวิจัยนี้ จำนวนทั้งสิ้น 16 คลาส

ที่มา: Chou (2005)

ตารางผนวกที่ 1 แสดงข้อมูลเอนไซม์ (Assession number) ที่ใช้ในงานวิจัยนี้

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

ที่มา: Chou and Elrod (2003)

ตารางผนวกที่ 1 แสดงข้อมูลเอนไซม์ (Assession number) ที่ใช้ในงานวิจัยนี้ (ต่อ)

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

ที่มา: Chou and Elrod (2003)

ตารางผนวกที่ 1 แสดงข้อมูลเอนไซม์ (Assession number) ที่ใช้ในงานวิจัยนี้ (ต่อ)

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

ที่มา: Chou and Elrod (2003)

ตารางผนวกที่ 1 แสดงข้อมูลเอนไซม์ (Assession number) ที่ใช้ในงานวิจัยนี้ (ต่อ)

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

ที่มา: Chou and Elrod (2003)

ตารางผนวกที่ 1 แสดงข้อมูลเอนไซม์ (Assession number) ที่ใช้ในงานวิจัยนี้ (ต่อ)

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

ที่มา: Chou and Elrod (2003)

ตารางผนวกที่ 1 แสดงข้อมูลเอนไซม์ (Assession number) ที่ใช้ในงานวิจัยนี้ (ต่อ)

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

ที่มา: Chou and Elrod (2003)

ตารางผนวกที่ 1 แสดงข้อมูลเอนไซม์ (Assession number) ที่ใช้ในงานวิจัยนี้ (ต่อ)

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

ที่มา: Chou and Elrod (2003)

ตารางผนวกที่ 1 แสดงข้อมูลเอนไซม์ (Assession number) ที่ใช้ในงานวิจัยนี้ (ต่อ)

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

ที่มา: Chou and Elrod (2003)

ประวัติการศึกษา และการทำงาน

ชื่อ –นามสกุล	นางสาวสิริวรรณ แต้วิจิตร
วัน เดือน ปี ที่เกิด	วันที่ 15 ตุลาคม 2524
สถานที่เกิด	กรุงเทพมหานคร
ประวัติการศึกษา	วศ.บ. (วิศวกรรมคอมพิวเตอร์) คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตบางเขน (พ.ศ.2547)
ตำแหน่งหน้าที่การงานปัจจุบัน	นักวิเคราะห์อาวุโส
สถานที่ทำงานปัจจุบัน	บริษัท ไอทีวัน จำกัด
ผลงานดีเด่นและรางวัลทางวิชาการ	-
ทุนการศึกษาที่ได้รับ	-