เครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกัน
ด้วยวิธีการจัดเรียงกลุ่มอะมิโนที่ไม่ชอบน้ำแบบค้นหาเฉพาะส่วน

นายชนินท์ จันมา

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2552

PROTEIN HOMOLOGY SEARCH TOOL
THROUGH HYDROPHOBIC CLUSTER ALIGNMENT WITH LOCAL SEARCH

Mr. Chanin Chanma

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering Program in Computer Engineering
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2009

หัวข้อวิทยานิพนธ์	เครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกันด้วยวิธีการจัดเรียงกลุ่มอะมิโนที่ไม่ชอบน้ำแบบค้นหาเฉพาะส่วน
โดย	นายชนินท์ จันมา
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ศาสตราจารย์ ดร. ประภาส จงสถิตย์วัฒนา
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	ผู้ช่วยศาสตราจารย์ ดร. รัฐ พิชญางกูร

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร.บุญสม เลิศหิรัญวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.โชติรัตน์ รัตนามหัทธนะ)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ศาสตราจารย์ ดร. ประภาส จงสถิตย์วัฒนา)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม
(ผู้ช่วยศาสตราจารย์ ดร.รัฐ พิชญางกูร)

..... กรรมการภายนอกมหาวิทยาลัย
(อาจารย์ ดร. กมลทิพย์ ขัตติยะวงศ์)

ชนินท์ จันมา : เครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกันด้วยวิธีการจัดเรียงกลุ่มอะมิโนที่ไม่ชอบน้ำแบบค้นหาเฉพาะส่วน. (PROTEIN HOMOLOGY SEARCH TOOL THROUGH HYDROPHOBIC CLUSTER ALIGNMENT WITH LOCAL SEARCH) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ศ.ดร. ประภาส จงสถิตย์วัฒนา, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม: ผศ.ดร. รัฐ พิชญางกูร, 60 หน้า.

เทคนิคการเปรียบเทียบความคล้ายคลึงกันในปัจจุบัน ซึ่งใช้การจัดเรียงกรดอะมิโนในลักษณะ 1 มิตินั้นมีข้อจำกัดและมีความผิดพลาดเมื่อนำมาใช้เปรียบเทียบจัดเรียงลำดับกรดอะมิโนที่มีค่าความเหมือนต่ำแม้ว่าลำดับกรดอะมิโนนั้นจะมีหน้าที่การทำงานเหมือนกัน จึงมีการนำเทคนิคการวิเคราะห์กลุ่มกรดอะมิโนที่ไม่ชอบน้ำในลักษณะ 2 มิติ มาใช้ในการทำนายลักษณะโครงสร้างและหน้าที่การทำงานของโปรตีน ในงานวิจัยชิ้นนี้ได้นำเสนอเทคนิคที่ใช้ในการสร้างเครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกัน โดยมีพื้นฐานอยู่บนการจัดเรียงกลุ่มกรดอะมิโนที่ไม่ชอบน้ำในโครงสร้างระดับทุติยภูมิ เพื่อคิดคำนวณคะแนนความคล้ายคลึงกันของคู่โปรตีนโดยอัตโนมัติ ซึ่งช่วยลดภาระการค้นหาและจับคู่สายลำดับโปรตีนโดยผู้เชี่ยวชาญ เทคนิคที่ได้นำเสนอนี้ช่วยเพิ่มความเร็วในการคำนวณและแก้ไขข้อบกพร่องในการจับคู่กลุ่มกรดอะมิโนในงานวิจัยเก่าที่ใช้แนวความคิดเชิงละโมบ โดยได้ทำการทดสอบเทคนิคที่นำเสนอกับชุดข้อมูลจากฐานข้อมูล HOMSTRAD และฐานข้อมูล PIR ซึ่งผลการทดสอบแสดงให้เห็นถึงประสิทธิภาพที่พัฒนาขึ้นจากเทคนิคการค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกันแบบเก่า ทั้งทางด้านความถูกต้องและระยะเวลาที่ใช้คำนวณ

ภาควิชา.....วิศวกรรมคอมพิวเตอร์..	ลายมือชื่อนิสิต .....
สาขาวิชา.....วิศวกรรมคอมพิวเตอร์..	ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....
ปีการศึกษา.....2552.....	ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์ร่วม .....

5070254021 : MAJOR COMPUTER ENGINEERING

KEYWORDS: HYDROPHOBIC CLUSTER ANALYSIS / PROTEIN HOMOLOGY / PROTEIN SEQUENCE SEARCH TOOL

CHANIN CHANMA: PROTEIN HOMOLOGY SEARCH TOOL THROUGH HYDROPHOBIC CLUSTER ALIGNMENT WITH LOCAL SEARCH. THESIS ADVISOR: PROF. PRABHAS CHONGSTITVATANA, Ph.D., THESIS CO-ADVISOR: RATH PICHYANGKURA, 60 pp.

Current techniques in protein homology testing involve a 1-dimensional alignment of nucleotide or amino acid sequencing. Due to its various constraints and low sequence identity values, a 2-Dimensional Hydrophobic Cluster Alignment has increasingly been used to predict the structure and functionality of protein. This work proposed an algorithm based on a secondary-structure Hydrophobic Cluster Alignment to compute a similarity score of protein sequences automatically, which helps reduce interventions of a human expert for a manual alignment. Additional techniques are introduced to speed up the calculation, as well as to resolve some greedy-based alignment limitation in the previous work. The alignment results and the classification accuracies from the well-known HOMSTRAD and PIR database have demonstrated an improvement in both accuracy and the computation time.

Department : ....Computer Engineering... Student's signature : .....

Field of study:....Computer Engineering.. Advisor's signature : .....

Academic year : ..2009..... Co-advisor's signature : .....

กิตติกรรมประกาศ

ขอขอบคุณอาจารย์ที่ปรึกษา ศาสตราจารย์ ดร. ประภาส จงสถิตย์วัฒนา พร้อมด้วยอาจารย์ที่ปรึกษาร่วม ผู้ช่วยศาสตราจารย์ ดร. รัฐ พิชญางกูร ผู้เสียสละเวลาในการให้ความรู้และแนะนำแนวทางการทำวิทยานิพนธ์ ให้คำปรึกษาต่างๆที่เป็นประโยชน์ต่องานวิจัย รวมทั้งชี้แนะจุดบกพร่องที่ควรปรับปรุงแก้ไข จนกระทั่งทำให้วิทยานิพนธ์เล่มนี้สำเร็จสมบูรณ์ได้

ขอขอบคุณคณาจารย์ทุกท่านในภาควิชาคอมพิวเตอร์ ผู้อบรมสั่งสอนและให้ความรู้ แนวคิดอันเป็นรากฐานสำคัญ ที่เป็นประโยชน์สำหรับการทำวิทยานิพนธ์

ขอขอบคุณสมาชิกในห้องปฏิบัติการวิจัยระบบอัจฉริยะ ที่คอยให้คำปรึกษา ดูแลเอาใจใส่ และสร้างบรรยากาศในการทำงานที่ดีตลอดมา

สุดท้ายนี้ขอกราบขอบพระคุณคุณพ่อ คุณแม่และพี่ชาย เป็นอย่างสูงที่คอยเป็นกำลังใจ ให้ความรัก และสนับสนุนด้านการศึกษา ทำให้สามารถทำวิทยานิพนธ์ฉบับนี้จนประสบผลสำเร็จ

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

บทที่ 1

บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ในปัจจุบันความรู้ทางด้านชีววิทยาและความเข้าใจในกลไกการทำงานของสิ่งมีชีวิตได้มีการพัฒนาและมีจำนวนมากขึ้นอย่างรวดเร็ว รวมไปถึงข้อมูลจำนวนมากที่ได้มีการค้นพบและจัดเก็บลงฐานข้อมูลต่างๆ ทั่วโลก ฐานข้อมูลเหล่านี้เป็นฐานข้อมูลขนาดใหญ่ที่ได้เก็บรวบรวมข้อมูลลำดับนิวคลีโอไทด์ หรือข้อมูลลำดับกรดอะมิโนที่ได้มีการค้นพบใหม่ๆ ไว้ทั้งหมด ยกตัวอย่างเช่น ฐานข้อมูล GenBank [1] ได้รวบรวมลำดับนิวคลีโอไทด์ไว้มากกว่า 54 ล้านลำดับ หรือ ฐานข้อมูล PIR [2] ซึ่งได้เก็บลำดับของกรดอะมิโนไว้มากกว่า 3 ล้านลำดับ ฐานข้อมูลที่กล่าวมานั้นเป็นฐานข้อมูลดิบขนาดใหญ่ที่ยังไม่ได้ทำการวิเคราะห์ แต่เก็บข้อมูลไว้เพื่ออ้างอิงเพียงอย่างเดียว นอกจากนี้ยังมีกลุ่มนักวิจัยหลายสถาบันทั่วโลกได้ทำการวิเคราะห์ฐานข้อมูลเหล่านี้เพื่อจัดแบ่งข้อมูลตามลักษณะโครงสร้าง ความคล้ายคลึงทางพันธุกรรม หรือแม้กระทั่งหน้าที่การทำงาน สร้างเป็นฐานข้อมูลเฉพาะกลุ่มเพื่อให้นักวิจัยต่างๆเข้าถึงและนำไปใช้ในการศึกษาได้โดยง่าย ฐานข้อมูลที่จัดแบ่งลำดับกรดอะมิโน โดยใช้โครงสร้างและหน้าที่การทำงานเป็นหลัก เช่น SCOP [3] หรือ HOMSTRAD [4] ฐานข้อมูลเหล่านี้มีประโยชน์มากสำหรับนักวิจัยที่ทำการศึกษาและค้นหาโปรตีนชนิดใหม่ๆ นักวิจัยสามารถคาดเดาลักษณะโครงสร้างหรือหน้าที่การทำงานของโปรตีนชนิดใหม่นั้นได้จากความคล้ายคลึงของโปรตีนที่มีอยู่แล้วในฐานข้อมูล

แต่เดิมการวิเคราะห์หาหน้าที่การทำงานของโปรตีนต้องพึ่งพาการทดลองจากห้องปฏิบัติการทางเคมีและชีววิทยา ซึ่งการทดลองหาหน้าที่ของโปรตีนเหล่านี้จำเป็นต้องใช้ค่าใช้จ่ายในการดำเนินงานสูง และเวลาที่ใช้ในการทดลองแต่ละครั้งยังใช้เวลามากอีกด้วย นอกจากนี้เทคนิคและวิธีการทดลองยังมีหลากหลายไปตามความต้องการทดสอบว่า โปรตีนนั้นๆ มีหน้าที่อย่างที่ได้รับผลจากทดลองจริงหรือไม่ จะเห็นได้ว่าการวิเคราะห์ศึกษาหน้าที่การทำงานของโปรตีนโดยทำการทดลองในห้องปฏิบัติการนั้น แม้ว่าจะได้ความถูกต้องสูงแต่สิ้นเปลืองทรัพยากรเป็นอย่างมาก ดังนั้นเพื่อลดทั้งค่าใช้จ่าย เวลา และทรัพยากรมนุษย์ในการทำการทดลอง คอมพิวเตอร์จึงเข้ามามีบทบาทช่วยเหลือในการวิเคราะห์หน้าที่การทำงานของโปรตีน

เนื่องจากโปรตีนเกิดจากลำดับของกรดอะมิโนที่เรียงต่อกัน ซึ่งทางคอมพิวเตอร์สามารถมองให้อยู่ในรูปของ สายอักขระ และการวิเคราะห์หน้าที่หรือความคล้ายคลึงกันของโปรตีน ทำได้จากการวิเคราะห์ความคล้ายคลึงกับของสายอักขระนั้นๆ แต่เครื่องมือหาความคล้ายคลึงหรือจับคู่ลำดับในปัจจุบันนั้นยังคงมองลำดับของกรดอะมิโนในแบบโครงสร้าง 1 มิติ (Primary Structure) ซึ่งการเปรียบเทียบในลักษณะ 1 มิตินั้น เมื่อนำไปใช้กับโปรตีนซึ่งมีหน้าที่การทำงานเหมือนกัน แต่มาจากสิ่งมีชีวิตต่างชนิดกัน กลับให้ค่าความเหมือนของลำดับมีค่าต่ำ เนื่องจากสิ่งมีชีวิตเหล่านี้มีวิวัฒนาการที่แตกต่างกันตลอดเวลา ลำดับของกรดอะมิโนในโปรตีนที่ถูกสร้างตามข้อมูลของลำดับพันธุกรรมจึงแตกต่างกันไปตามเวลา นอกจากนี้ในแต่ละวิวัฒนาการอาจเกิดลำดับของกรดอะมิโนซึ่งไม่มีผลกับหน้าที่หรือโครงสร้างหลัก เพิ่มขึ้นมาในสายกรดอะมิโนทำให้การวิเคราะห์ในลักษณะ 1 มิติมีผลคลาดเคลื่อนได้ อีกเทคนิคในการศึกษาโครงสร้างของโปรตีนได้อย่างถูกต้องแม่นยำคือ การจำลองโครงสร้าง 3 มิติของโปรตีนขึ้นมาจากคุณสมบัติของกรดอะมิโนแต่ละตัวที่เรียงต่อกัน แต่เทคนิคนี้ต้องใช้เวลามากในการคำนวณเพื่อสร้างแบบจำลอง

เมื่อนักวิจัยได้ทำการศึกษาธรรมชาติการเกิดของโปรตีน จึงพบว่าคุณสมบัติการละลายน้ำของกรดอะมิโนเป็นปัจจัยหลักในการจัดตัวเป็นรูปร่างของลำดับกรดอะมิโน เนื่องจากการผลิตกรดอะมิโนนั้นเกิดขึ้นโดยมีน้ำเป็นสภาวะแวดล้อมหลัก ดังนั้นการจัดเรียงตัวของลำดับกรดอะมิโนจึงพยายามรักษาพลังงานให้มีระดับต่ำที่สุด โดยกรดอะมิโนที่ไม่ชอบน้ำจะรวมตัวกันเป็นโครงสร้างหลักอยู่ภายใน และพยายามให้กรดอะมิโนที่ชอบน้ำอยู่ด้านนอก ซึ่งนักวิจัยกลุ่มหนึ่งทำการวิเคราะห์โปรตีนโดยเน้นศึกษาเฉพาะลำดับกรดอะมิโนที่ไม่ชอบน้ำ แล้วสร้างแผนภูมิของลำดับกรดอะมิโนให้อยู่ในรูปแบบ 2 มิติ สามารถแสดงให้เห็นลักษณะโครงสร้างของกลุ่มกรดอะมิโนที่ชอบน้ำได้เด่นชัดขึ้น [5,6] การวิเคราะห์โปรตีนโดยอาศัยรูปร่างของกลุ่มกรดอะมิโนที่ไม่ชอบน้ำเป็นหลักนั้นสามารถนำไปใช้ในการทำนายโครงสร้างทุติยภูมิได้ เช่น งานวิจัย [7,8] ต่อมานักวิจัยจึงพัฒนานำเทคนิคการวิเคราะห์นี้ไปใช้การจับคู่ลำดับกรดอะมิโน [9-12] เพื่อหาความคล้ายคลึงของคู่โปรตีน และนำไปใช้ในการอ้างอิงหาโครงสร้างหรือหน้าที่การทำงานของโปรตีนต่อไป ซึ่งผลการหาความคล้ายคลึงด้วยการวิเคราะห์นี้ให้ผลลัพธ์ที่แม่นยำกว่าการจับคู่ในลักษณะ 1 มิติ แต่การจับคู่นั้นยังจำเป็นต้องอาศัยความรู้ของผู้เชี่ยวชาญและต้องจับคู่ด้วยมือ ดังนั้นการจับคู่กลุ่มกรดอะมิโนด้วยวิธีการนี้จึงไม่ได้นำไปพัฒนาและใช้กันอย่างกว้างขวาง

จากผลงานวิจัยและการทดลองในวิทยานิพนธ์ของ นายภิสิทธิ์ กรรณสูต [13] แสดงแนวทางการพัฒนาเครื่องมืออัตโนมัติเพื่อจับคู่และหาความคล้ายคลึงของโปรตีน โดยแนวทางการวิจัยนั้นอยู่บนพื้นฐานของการวิเคราะห์กลุ่มกรดอะมิโนที่ไม่ชอบน้ำและใช้วิธีกำหนดการพลวัต เป็นเทคนิคในการสร้างเครื่องมือ ซึ่งผลลัพธ์แสดงถึงความสามารถในการจับคู่และหาความคล้ายคลึงของโปรตีนโดยอัตโนมัติได้อย่างมีประสิทธิภาพ

ในงานวิจัยชิ้นนี้ ผู้วิจัยต้องการพัฒนาต่อยอดเพื่อเสนอแนวทางและเทคนิคการจับคู่และหาความคล้ายคลึงของโปรตีนโดยอัตโนมัติ ด้วยวิธีการจัดเรียงกลุ่มกรดอะมิโนที่ไม่ชอบน้ำแบบค้นหาเฉพาะส่วน เพื่อเพิ่มประสิทธิภาพความถูกต้อง และความเร็วในการจับคู่และหาความคล้ายคลึง และพัฒนาเป็นเครื่องมือค้นหาโปรตีนที่มีหน้าที่คล้ายคลึงมากที่สุดจากฐานข้อมูลต่อไป

1.2 วัตถุประสงค์ของการวิจัย

1. เพื่อนำเสนอแนวทางและพัฒนาเทคนิคการหาโปรตีนที่มีหน้าที่คล้ายคลึงกันได้อย่างมีประสิทธิภาพทั้งทางด้านความเร็วและความถูกต้อง ด้วยวิธีการจัดเรียงกลุ่มกรดอะมิโนที่ไม่ชอบน้ำแบบค้นหาเฉพาะส่วน
2. เพื่อสร้างเครื่องมือค้นหาโปรตีนที่มีหน้าที่คล้ายคลึงกันจากฐานข้อมูล ด้วยวิธีการจัดเรียงกลุ่มกรดอะมิโนที่ไม่ชอบน้ำแบบค้นหาเฉพาะส่วน

1.3 ขอบเขตของการวิจัย

1. เพื่อนำเสนอแนวทาง พัฒนาเทคนิค และสร้างเครื่องมือค้นหาโปรตีนที่มีหน้าที่คล้ายคลึงกันจากฐานข้อมูล ด้วยวิธีกำหนดการพลวัตแบบมีการค้นหาเฉพาะส่วน
2. ทดสอบประสิทธิภาพของเครื่องมือทั้งด้านความเร็วและความถูกต้อง

1.4 ขั้นตอนของการวิจัย

1. ศึกษาโครงสร้างและหน้าที่การทำงานของโปรตีน
2. ศึกษาวิธีการวิเคราะห์โปรตีนโดยเน้นเฉพาะกรดอะมิโนกลุ่มที่ไม่ชอบน้ำ
3. ศึกษาเทคนิคและวิธีการจับคู่แบบกำหนดการพลวัต
4. รวบรวมและเตรียมชุดข้อมูลที่ใช้ในการทดสอบ

5. สร้างเครื่องมือค้นหาโปรตีนที่มีหน้าที่คล้ายคลึงกันจากฐานข้อมูล
6. ทดสอบประสิทธิภาพของเครื่องมือทั้งด้านความเร็วและความถูกต้อง
7. วิเคราะห์และสรุปผลผลการทดลอง
8. เรียบเรียงวิทยานิพนธ์

1.5 ประโยชน์ที่ได้รับ

1. ได้เทคนิคการจับคู่ลำดับกรดอะมิโนโดยการวิเคราะห์กลุ่มที่ไม่ชอบน้ำที่มีประสิทธิภาพสูงขึ้น ด้วยวิธีจัดเรียงกลุ่มกรดอะมิโนที่ไม่ชอบน้ำแบบค้นหาเฉพาะส่วน
2. ได้เครื่องมือค้นหาโปรตีนที่มีหน้าที่คล้ายคลึงกันจากฐานข้อมูล ที่อาศัยเทคนิควิธีจัดเรียงกลุ่มกรดอะมิโนที่ไม่ชอบน้ำแบบค้นหาเฉพาะส่วนสร้างตามหลักการวิเคราะห์กลุ่มที่ไม่ชอบน้ำ

1.6 ผลงานที่ตีพิมพ์จากวิทยานิพนธ์

ส่วนหนึ่งของงานวิทยานิพนธ์นี้ได้รับการตอบรับให้ตีพิมพ์เป็นบทความวิชาการในหัวข้อเรื่อง “An Algorithm to Compute Protein Homology Based On Hydrophobic Cluster Analysis” โดย ชนินท์ จันมา, รัฐ พิชญางกูร, โชติรัตน์ รัตนามหัทธนะ และ ประภาส จงสถิตย์วัฒนา ในงานประชุมวิชาการ “The 6th International Joint Conferrence on Computer Science and Software Engineering (JCSSE2009)” ซึ่งจัดขึ้น ณ Laguna Resort จังหวัดภูเก็ต ประเทศไทย ในระหว่างวันที่ 13-15 พฤษภาคม 2552 ดังแสดงใน ภาคผนวก ข หน้า 54

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

งานวิจัยเกี่ยวกับเครื่องมือที่ใช้จับคู่กลุ่มกรดอะมิโนที่ไม่ชอบน้ำแบบ 2 มิติโดยอัตโนมัติ มีการใช้ทฤษฎีและงานวิจัยต่างๆ ซึ่งเป็นพื้นฐานต่อการทำการวิจัย โดยมีรายละเอียดดังต่อไปนี้

2.1 **โปรตีน** (Protein)

โปรตีนเป็นสารประกอบอินทรีย์ซึ่งเกิดจากกรดอะมิโน (Amino Acid) เรียงตัวต่อกันเป็นสายยาว ซึ่งกรดอะมิโนเหล่านี้เชื่อมต่อกันด้วยพันธะเปปไทด์ (Peptide Bond) โดยหน้าที่การทำงานของโปรตีนเกิดจากความแตกต่างทั้งชนิดของกรดอะมิโนและลำดับของกรดอะมิโน โปรตีนที่ทำหน้าที่คล้ายคลึงกันอาจจะมีลำดับของกรดอะมิโนแตกต่างกัน แต่ยังคงมีโครงสร้างที่คล้ายคลึงกัน สิ่งมีชีวิตเกือบทุกชนิดจะมีโปรตีนเป็นส่วนประกอบหลักมากเป็นอันดับสอง รองจากน้ำ เนื่องจากโปรตีนมีหน้าที่สำคัญหลายอย่างภายในสิ่งมีชีวิต ยกตัวอย่างเช่น เป็นเอนไซม์ (Enzyme) ซึ่งทำหน้าที่เป็นตัวเร่งปฏิกิริยาในการทางชีวเคมีต่างๆ ในสิ่งมีชีวิต, ทำหน้าที่เป็นโครงสร้างหลักของสิ่งมีชีวิต เช่น เส้นผม กล้ามเนื้อ อวัยวะ นอกจากนี้ยังทำหน้าที่สำคัญอื่นๆ อย่างการส่งสัญญาณระหว่างเซลล์ หรือทำหน้าเป็นภูมิคุ้มกันของระบบ

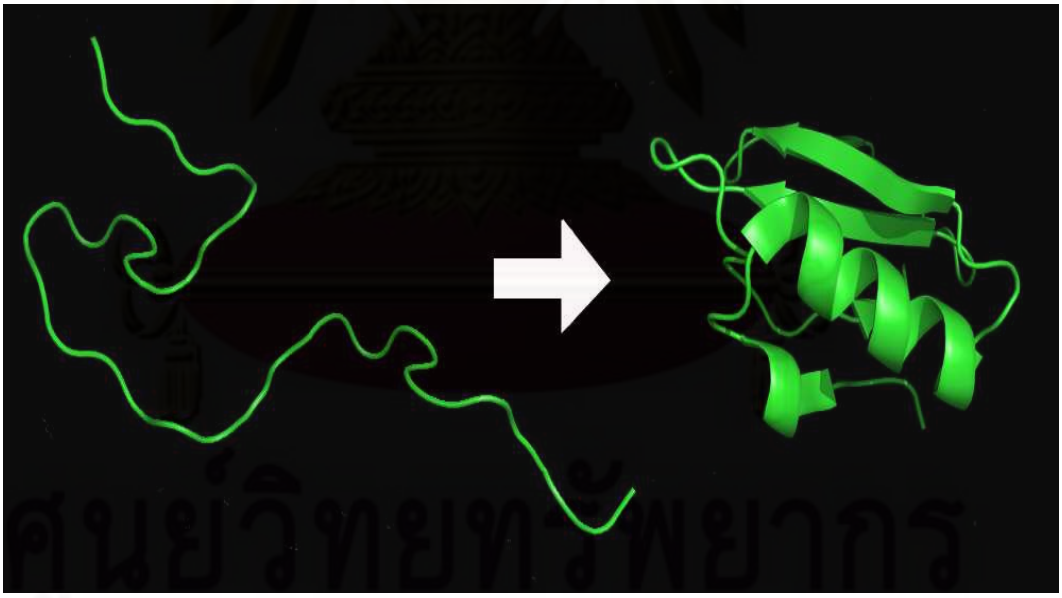
2.1.1 **การสังเคราะห์โปรตีน** (Protein Synthesis)

สายลำดับกรดอะมิโนที่เชื่อมต่อกันเป็นโปรตีนนั้นใช้ข้อมูลที่ถูกเข้ารหัสอยู่ในลำดับนิวคลีโอไทด์ (Nucleotide sequence) ภายในยีน (Gene) โดยการสังเคราะห์โปรตีนจะเริ่มจากการถอดรหัส (Transcription) ข้อมูลของลำดับนิวคลีโอไทด์ให้อยู่ในรูปของ mRNA (Messenger RNA) ซึ่งจะถูกส่งผ่านออกไปยัง ไซโทปลาสซึม (Cytoplasm) เพื่อทำการแปล (Translation) และสร้างเป็นสายลำดับของกรดอะมิโน โดยไรโบโซม (Ribosome) ต่อไป

จะเห็นได้ว่าลำดับของสายลำดับกรดอะมิโนนั้นขึ้นกับข้อมูลที่ถูกเข้ารหัสอยู่ในยีน และในธรรมชาติ ยีนของสิ่งมีชีวิตจะเกิดวิวัฒนาการเปลี่ยนแปลงไป ทำให้เกิดความแตกต่างทางพันธุกรรม ดังนั้นจึงมีโปรตีนที่มีสายลำดับโปรตีนแตกต่างกัน แต่มีหน้าที่การทำงานเหมือนกันอยู่มากมายในสิ่งมีชีวิตต่างๆ

2.1.2 **การม้วนพับของสายโปรตีน** (Protein Folding)

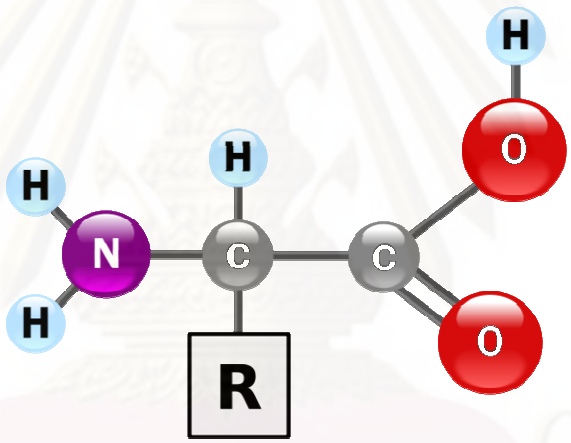
ทั้งในขณะการสร้างสายลำดับกรดอะมิโนของโปรตีนและภายหลังการสร้างสายลำดับนั้น เส้นสายลำดับของกรดอะมิโนจะทำการม้วนพับตัวเพื่อจัดรูปร่างสร้างโครงสร้าง 3 มิติ ดังแสดงในรูปที่1 โดยโครงสร้างของโปรตีนที่ก่อรูปขึ้นมานั้นขึ้นอยู่กับลำดับและคุณสมบัติของกรดอะมิโน เช่น คุณสมบัติความชอบและคุณสมบัติความไม่ชอบน้ำ (Hydrophilic and Hydrophobic) หรือ คุณสมบัติประจุทางไฟฟ้า (Electrically Charged) ซึ่งโครงสร้างของโปรตีนที่ทำการม้วนพับจนอยู่ตัวคงสภาพได้แล้ว โดยส่วนมากจะมีโครงสร้างหลักของกรดอะมิโนที่ไม่ชอบน้ำ (Hydrophobic Core) อยู่ภายใน และหันโครงสร้างด้านที่มีประจุหรือมีขั้ว ไว้ที่พื้นผิวภายนอกของโครงสร้าง เพื่อให้ทำปฏิกิริยาและละลายน้ำได้โดยง่าย เนื่องจากสภาวะแวดล้อมในขณะทำการม้วนพับนั้นจะมีน้ำเป็นองค์ประกอบหลัก การทำปฏิกิริยาหรือละลายน้ำได้โดยง่ายนั้นทำให้ใช้พลังงานในการม้วนพับน้อย



รูปที่ 2.1 ตัวอย่างการม้วนพับของสายโปรตีน
จากเส้นสายลำดับกรดอะมิโน เป็นโครงสร้าง 3 มิติ

2.2 **กรดอะมิโน** (Amino Acid)

กรดอะมิโนคือโมเลกุลที่ประกอบด้วยธาตุหลักทั้งสิ้น 4 ชนิด ได้แก่ คาร์บอน (C) ไฮโดรเจน (H) ออกซิเจน (O) และไนโตรเจน (N) นอกจากนั้นอาจมีธาตุอื่นๆ นอกเหนือจากธาตุหลักรวมอยู่ในโมเลกุลด้วย เช่น ฟอสฟอรัส (P) หรือกำมะถัน (S) โดยธาตุหลักจะประกอบกันอยู่ในรูปของหมู่กรดอะมิโน (Amino Group, $-NH_2$) หมู่คาร์บอกซิล (Carboxyl Group, -COOH) และหมู่อัลคิล (Alkyl Group, -R) ซึ่งมีสูตรในรูปทั่วคือ $H_2NCHRCOOH$ ดังแสดงในรูปที่ 2.2 โดยกรดอะมิโนในสิ่งมีชีวิตมีทั้งสิ้น 20 ชนิด ซึ่งคุณสมบัติและความแตกต่างของกรดอะมิโนในแต่ละชนิดนั้นขึ้นอยู่กับ หมู่อัลคิลที่แตกต่างกัน โดยคุณสมบัติที่งานวิจัยนี้สนใจเป็นพิเศษคือคุณสมบัติความชอบน้ำ และคุณสมบัติความไม่ชอบน้ำ



รูปที่ 2.2 โครงสร้างทั่วไปของกรดอะมิโน

โดยกรดอะมิโนทั้ง 20 ชนิดนั้นสามารถแบ่งตามคุณสมบัติความชอบ และไม่ชอบน้ำออกเป็นกลุ่มได้ตามตารางที่ 1 ดังนั้นกรดอะมิโนที่งานวิจัยนี้สนใจเป็นพิเศษคือกลุ่มกรดอะมิโนที่ไม่ชอบน้ำ ซึ่งมีทั้งสิ้น 7 ตัว ได้แก่ เวลีน (Valine) ไอโซลิวซีน (Isoleucine) ลิวซีน (Leucine) ฟีนิลอะลานีน (Phenylalanine) ทริปโตเฟน (Tryptophan) เมไธโอนีน (Methionine) และ ไทโรซีน (Tyrosine)

ตารางที่ 2.1 ตารางกรดอะมิโนแบ่งตามคุณสมบัติความไม่ชอบน้ำ [14]

กลุ่ม (Class)	กรดอะมิโน (Amino acid)	ค่าความไม่ชอบน้ำ (Hydrophobicity)
ไม่ชอบน้ำ (Hydrophobic)	เวลีน (Valine, V)	-0.31
	ไอโซลิวซีน (Isoleucine, I)	-0.60
	ลิวซีน (Leucine, L)	-0.55
	ฟีนิลอะลานีน (Phenylalanine, F)	-0.32
	ทริปโตเฟน (Tryptophan, W)	0.30
	เมไธโอนีน (Methionine, M)	-0.10
	ไทโรซีน (Tyrosine, Y)	0.68
ชอบน้ำ (Hydrophilic)	กรดแอสปาร์ติก (Aspartic acid, D)	3.49
	กรดกลูตามิก (Glutamic acid, E)	2.68
	แอสพาราจีน (Asparagine, N)	2.05
	กลูตามีน (Glutamine, Q)	2.36
	อาร์จินีน (Arginine, R)	2.58
	ไกลซีน (Glycine, G)	0.74
	ไลซีน (Lysine, K)	2.71
	โพรลีน (Proline, P)	2.23
	ซีรีน (Serine, S)	0.84
	ธรีโอนีน (Threonine, T)	0.52
เป็นกลาง (Neutral)	ฮีสติดีน (Histidine, H)	2.06
	อะลานีน (Alanine, A)	0.11
	ซีสเตอีน (Cysteine, C)	-0.13

2.3 **โครงสร้างของโปรตีน** (Protein Structure)

เนื่องจากหน้าที่และความสามารถในการทำงานของโปรตีนนั้น ถูกกำหนดด้วยโครงสร้าง 3 มิติของโปรตีนเป็นสำคัญ ดังนั้นนักวิจัยจึงได้ให้ความสนใจศึกษาโครงสร้างที่ซับซ้อนของโปรตีน โดยแบ่งระดับการศึกษาโครงสร้างของโปรตีนเป็น 4 ระดับขั้น ตามความซับซ้อนและความสัมพันธ์กันของลำดับกรดอะมิโน แสดงดังรูปที่ 3 ได้แก่

2.3.1 **โครงสร้างปฐมภูมิ** (Primary Structure)

เป็นโครงสร้างในระดับต่ำสุด โครงสร้างในระดับนี้แสดงรูปร่างของโปรตีนในลักษณะของลำดับกรดอะมิโนซึ่งเรียงต่อกันด้วยพันธะเปปไทด์เป็นเส้นตรง ข้อมูลที่ได้จากการศึกษาโครงสร้างในระดับปฐมภูมินี้ไม่สามารถแสดงถึงหน้าที่การทำงานของโปรตีนได้โดยตรง

2.3.2 **โครงสร้างทุติยภูมิ** (Secondary Structure)

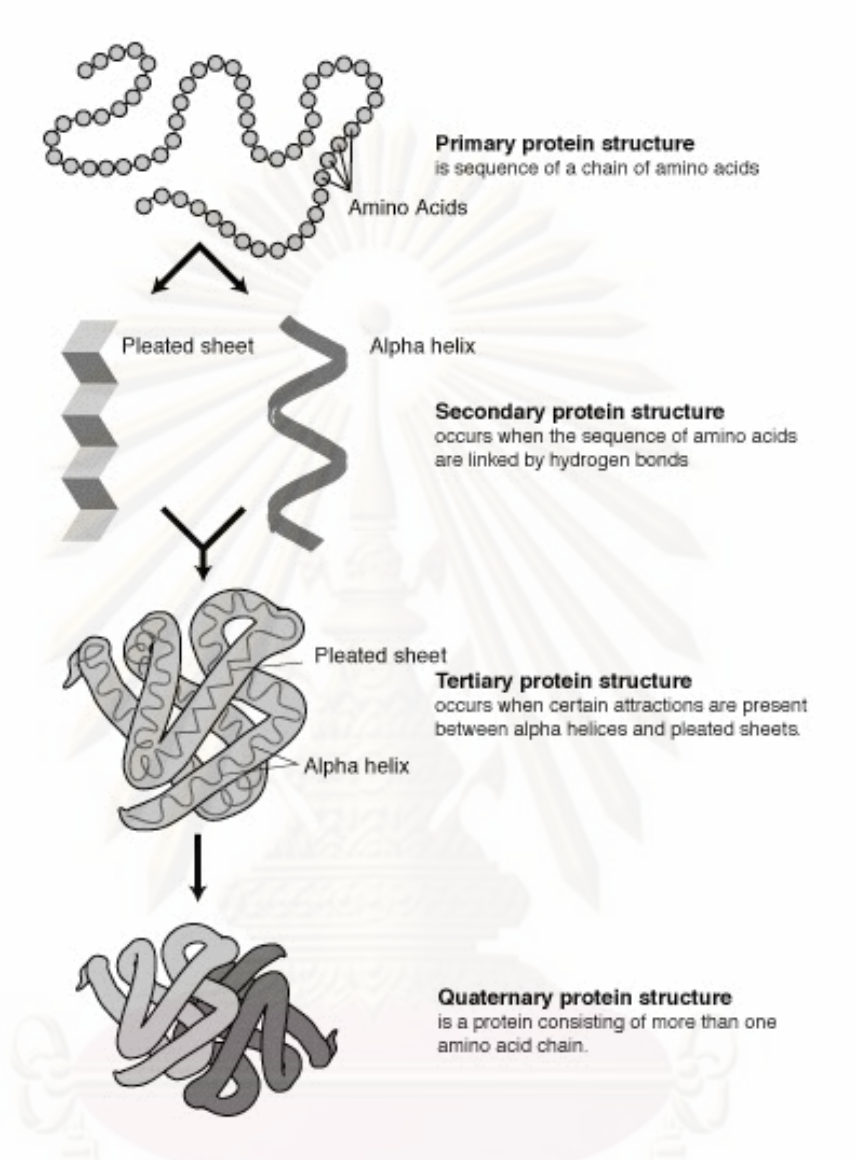
แสดงโครงสร้างของโปรตีน ในลักษณะการประกอบกันของโครงสร้างย่อยๆ ซึ่งเกิดจากแรงกระทำของพันธะไฮโดรเจนภายในแกนหลัก(Backbone) ของสายโปรตีน โดยเริ่มจากทำนายโครงสร้างย่อยของโปรตีนจากกลุ่มของลำดับกรดอะมิโนในช่วงสั้นๆ เป็นส่วนๆ ไปแล้วนำโครงสร้างย่อยๆ เหล่านั้นมาต่อกันแสดงเป็นโครงสร้างทั้งหมดของโปรตีนที่กำลังศึกษา ตัวอย่างโครงสร้างย่อยที่พบบ่อยได้แก่ เกลียวอัลฟา (Alpha-Helix) และแผ่นเบต้า (Beta-Sheet)

2.3.3 **โครงสร้างตติยภูมิ** (Tertiary Structure)

แสดงโครงสร้างของโปรตีนในรูปโครงสร้าง 3 มิติ ที่เกิดจากการม้วนพับของสายลำดับกรดอะมิโนโดยอาศัยแรงกระทำระหว่างหมู่อัลคิลของกรดอะมิโนแต่ละตัว ซึ่งโครงสร้างในระดับนี้สามารถแสดงถึงหน้าที่การทำงานของโปรตีนได้โดยตรง แต่ในขั้นตอนการศึกษาด้วยการจำลองโครงสร้างในระดับนี้จำเป็นต้องใช้การประมวลผลที่ซับซ้อนและใช้เวลาในการคำนวณสูง

2.3.4 **โครงสร้างจตุรภูมิ** (Quaternary Structure)

โครงสร้างระดับสูงสุดของโปรตีน แสดงโครงสร้างในระดับ 3 มิติ เช่นเดียวกับโครงสร้างตติยภูมิ แต่จะแสดงถึงตำแหน่งและลักษณะการรวมตัวกันของโปรตีนมากกว่าหนึ่งตัว โดยอาศัยแรงกระทำระหว่างพื้นผิวของโปรตีนแต่ละตัว เพื่อร่วมกันทำงานตามหน้าที่ต่างๆ



รูปที่ 2.3 โครงสร้างของโปรตีนในระดับต่างๆ [15]

จากระดับการศึกษาโครงสร้างของโปรตีนทั้ง 4 ระดับนั้น โครงสร้างตติยภูมิและโครงสร้างจตุรภูมิเป็นโครงสร้างที่สามารถแสดงถึงหน้าที่การทำงานของโปรตีนได้เป็นอย่างดี แต่การจำลองและสร้างโครงสร้างทั้งสองระดับในรูปแบบ 3 มิตินั้น จำเป็นต้องใช้การประมวลผลที่มีความซับซ้อนและใช้ระยะเวลาในการคำนวณเป็นจำนวนมาก การศึกษาหาหน้าที่การทำงานของโปรตีนจากโครงสร้างทุติยภูมิจึงเป็นแนวทางที่ผู้วิจัยเลือก แม้ว่าจะมีความแตกต่างของลำดับกรดอะมิโนภายในกลุ่มโปรตีนที่ทำหน้าเหมือนกัน แต่โครงสร้างระดับทุติยภูมิของโปรตีนยังคงมีโครงสร้างที่คล้ายคลึงกัน จึงเหมาะสมเป็นแนวทางในการนำมาศึกษาเทคนิคการค้นหาโปรตีนที่มีหน้าที่คล้ายคลึงกัน

2.4 **การจับคู่ลำดับกรดอะมิโน** (Amino Acid Sequence Alignment)

เทคนิคการจับคู่ลำดับกรดอะมิโนถูกนำมาใช้ในการหาความคล้ายคลึงของโปรตีน เพื่อศึกษาหาโปรตีนที่มีโครงสร้าง หรือหน้าที่การทำงานที่ใกล้เคียงกัน แทนการทดลองในห้องปฏิบัติการทางเคมี ซึ่งมีค่าใช้จ่ายและใช้ระยะเวลาในการดำเนินการทดลองสูง โดยเทคนิคการจับคู่ลำดับกรดอะมิโนได้ถูกพัฒนาและนำเสนอในหลายแนวทาง เช่น

2.4.1 **วิธีกำหนดการพลวัต** (Dynamic Programming Method)

เป็นเทคนิคแรกๆ ที่ถูกนำมาใช้ในการจับคู่กรดอะมิโน อัลกอริทึมที่เป็นที่นิยม เช่น Needleman-Wunsch Algorithm[16] และ Smith-Waterman Algorithm[17] โดยทั้งสองวิธีทำการจับคู่กรดอะมิโนโดยสร้างตารางคะแนนแทนทุกคู่ของลำดับกรดอะมิโนในแต่ละสาย แล้วคิดค่าคะแนนแต่ละคู่กรดอะมิโน โดยใช้ตารางคะแนนการแทนที่กรดอะมิโน (Substitution Matrix)[18] การคิดคะแนนจะเริ่มจากกรดอะมิโนคู่แรก แล้วคิดคะแนนทุกคู่กรดอะมิโนจนได้คะแนนสูงสุดที่คู่สุดท้ายของตาราง ซึ่งอาจกำหนดการหักคะแนนการเว้นช่องว่าง (Gap Penalty) เพื่อให้ได้ความถูกต้องมากขึ้น เมื่อได้คะแนนสูงสุดจากคู่ลำดับกรดอะมิโนคู่สุดท้ายแล้วจึงค้นหาเส้นทางย้อนกลับ เพื่อให้ได้การจับคู่กรดอะมิโนที่มีคะแนนมากที่สุด

เทคนิควิธีกำหนดการพลวัตนั้นรับประกันการจับคู่กรดอะมิโนที่ให้ความถูกต้องแม่นยำที่สุด แต่เทคนิคนี้ใช้ระยะการคำนวณในแต่ละคู่โปรตีนสูง เพราะต้องทำการให้คะแนนทุกคู่ลำดับกรดอะมิโน และเมื่อลำดับกรดอะมิโนมีความยาวมากขึ้น ระยะเวลาที่ต้องใช้ในการคำนวณจะยิ่งสูงขึ้นตามไปด้วย

2.4.2 **การเปรียบเทียบในลักษณะคำ** (Word Method, K-tuple Method)

เทคนิคนี้เป็นเทคนิคที่ไม่ได้รับประกันคำตอบที่มีค่าดีที่สุด แต่มีประสิทธิภาพด้านความเร็วมากกว่าวิธีกำหนดการพลวัตถึง 50 เท่า ดังนั้นจึงถูกนำไปพัฒนาใช้ในการค้นหาโปรตีนที่ใกล้เคียงมากที่สุดจากฐานข้อมูลขนาดใหญ่ เครื่องมือที่นิยมใช้เป็นอย่างมาก ได้แก่ BLAST [19] และ FASTA [20] เทคนิคที่ใช้คือ การใช้คะแนนการจับคู่กรดอะมิโนแบบเป็นชิ้นส่วน แล้วเลือกเก็บและให้คะแนนเฉพาะคู่ชิ้นส่วนที่มีคะแนนสูงเท่านั้นโดยใช้วิธีการวิเคราะห์ทางสถิติ แล้วจึงนำเฉพาะคู่ชิ้นส่วนที่เลือกมาต่อกันเป็นการจับคู่ที่สมบูรณ์ทั้งสายลำดับกรดอะมิโน ทำให้เทคนิคนี้มีความเร็วในการคำนวณสูง แต่ก็ไม่สามารถยืนยันได้ว่าเป็นคำตอบที่ถูกต้องมากที่สุด นอกจากนี้เทคนิคการให้คะแนนสำหรับโปรตีนที่มาจากสิ่งมีชีวิตคนละชนิด แต่มีหน้าที่การทำงานเหมือนกันยังไม่สามารถคิดคะแนนได้ถูกต้องเท่าที่ควร

2.4.3 **การค้นหาแบบข้อมูลฮิดเดนมาร์คอฟ** (Hidden Markov Profile Search)

เป็นอีกหนึ่งวิธีที่ได้รับความนิยมสำหรับการค้นหาคู่ลำดับกรดอะมิโนที่คล้ายคลึงกันมากที่สุด โดยการประยุกต์ใช้แบบจำลองฮิดเดนมาร์คอฟ (Hidden Markov Model) ซึ่งเป็นแบบจำลองทางสถิติเพื่อแทนลำดับของกรดอะมิโนของโปรตีนแต่ละชนิดด้วยแบบข้อมูล (Profile) โดยแบบข้อมูลนี้อาจสร้างได้จากการวิเคราะห์โปรตีนโดเมน (Protein Domain) แล้วนำมาสร้างเป็นแบบจำลองฮิดเดนมาร์คอฟแล้วเก็บไว้ในฐานข้อมูลเพื่อทำการค้นหาต่อไป ตัวอย่างแนวทางที่ใช้เทคนิคนี้ได้แก่ SAM[21] และ HMMER [22] โดยเทคนิคนี้จำเป็นต้องค้นหาโปรตีนโดเมนในแต่ละสายลำดับกรดอะมิโนขึ้นมาก่อนแล้วจึงนำไปสร้างแบบจำลองฮิดเดนมาร์คอฟ จึงสามารถทำการเปรียบเทียบและทำการค้นหาได้ จึงไม่สามารถใช้กับโปรตีนชนิดใหม่ๆ ที่ยังไม่ได้ค้นหาโปรตีนโดเมน หรือ ไม่มีแบบจำลองฮิดเดนมาร์คอฟอยู่ในฐานข้อมูล

2.5 **การวิเคราะห์กลุ่มที่ไม่ชอบน้ำ** (Hydrophobic Cluster Analysis)

เนื่องจากการจับคู่ลำดับกรดอะมิโนโดยใช้ข้อมูลจากโครงสร้างระดับปฐมภูมิไม่สามารถให้คำตอบที่ดีในการค้นหาโปรตีนที่มีหน้าที่เหมือนกันแต่มาจากสิ่งมีชีวิตคนละชนิดกัน นักวิจัยจึงศึกษาการพับม้วนตัวของโปรตีนซึ่งเกิดขึ้นในน้ำ แล้วเสนอแนวคิดการวิเคราะห์โครงสร้างของโปรตีนจากกลุ่มกรดอะมิโนที่ไม่ชอบน้ำเป็นหลัก แนวความคิดนี้ได้พัฒนาไปถึงการแสดงผลโครงสร้างระดับทุติยภูมิในรูปแบบของแผนภูมิ 2 มิติ [5, 6] และนำไปใช้ในการวิเคราะห์โปรตีนที่มีหน้าคล้ายกันจากการศึกษาโครงสร้างของกลุ่มกรดอะมิโนที่ไม่ชอบน้ำ [9-12] แสดงให้เห็นว่าโปรตีนที่มีหน้าที่เหมือนกันจะมีโครงสร้างของกลุ่มกรดอะมิโนที่ไม่ชอบน้ำคล้ายคลึงกัน แต่แนวทางการวิเคราะห์และจับคู่ลำดับโปรตีนเพื่อหาหน้าที่การทำงาน โดยอาศัยการศึกษากลุ่มกรดอะมิโนที่ไม่ชอบน้ำนี้ไม่ได้รับความนิยมและพัฒนาต่อเนื่องมากนัก เนื่องจากการวิเคราะห์และจับคู่กลุ่มกรดอะมิโนที่ไม่ชอบน้ำจำเป็นต้องอาศัยความรู้ความเข้าใจของผู้เชี่ยวชาญและต้องทำการเลือกและจับคู่ลำดับกรดอะมิโนด้วยมือ

2.6 **การจัดเรียงกลุ่มกรดอะมิโนที่ไม่ชอบน้ำแบบ 2 มิติโดยอัตโนมัติ** (Automatic 2-D Hydrophobic Cluster Analysis Alignment)

วิทยานิพนธ์ของ นายภิสิทธิ์ กรรณสูต [13] ได้ทำการศึกษาและทดลองสร้างเครื่องมือจับคู่ลำดับกรดอะมิโนแบบอัตโนมัติ โดยใช้พื้นฐานการจับคู่ของกลุ่มกรดอะมิโนที่ไม่ชอบน้ำในลักษณะ 2 มิติ และเทคนิควิธีกำหนดการพลวัตเป็นหลัก ซึ่งผลวิจัยแสดงให้เห็นแนวทางการค้นหาคู่โปรตีนที่ทำหน้าที่เหมือนกันจากการวิเคราะห์กลุ่มกรดอะมิโนที่ไม่ชอบน้ำและให้ผลลัพธ์ที่ดีกว่าการจับคู่ในลักษณะ 1 มิติ

ขั้นตอนวิธีการหาคะแนนความเหมือนของคู่ลำดับอะมิโนมีทั้งหมด 3 ขั้นตอนดังนี้

2.6.1 **การแบ่งลำดับกรดอะมิโนให้อยู่ในรูปของกลุ่มกรดอะมิโนที่ไม่ชอบน้ำ**

ในขั้นตอนนี้จะทำการแปลงสัญลักษณ์สายลำดับของกรดอะมิโน ให้อยู่ในรูปแบบการคำนวณที่ต้องการ และตัดแบ่งสายลำดับออกเป็น กลุ่มของกรดอะมิโนที่ไม่ชอบน้ำหลายกลุ่ม (Hydrophobic cluster) โดยมีสองขั้นตอนย่อยดังนี้

1. แทนสัญลักษณ์กรดอะมิโนที่ชอบน้ำทั้งหมดด้วยเลข 0 และคงสัญลักษณ์กรดอะมิโนที่ไม่ชอบน้ำทั้ง 7 ชนิดไว้

2. ตัดสายลำดับกรดอะมิโนแบ่งออกให้เป็นกลุ่มย่อยของกรดอะมิโนที่ไม่ชอบน้ำ โดยมีหลักเกณฑ์ในการตัดแบ่งกลุ่มดังนี้

   2.1. เมื่อเจอกรดอะมิโนโพรลีน (Proline, P) จะทำการแบ่งสายกรดอะมิโนออกเป็น 2 กลุ่ม

   2.2. จากการแทนที่สัญลักษณ์กรดอะมิโนที่ชอบน้ำด้วย เลข 0 ถ้ามีเลข 0 เรียงติดกันมากกว่า 4 ตัวในสายลำดับ จะทำการแบ่งสายลำดับออกเป็น 2 กลุ่ม เช่นเดียวกัน

ก)	MASFKIALLLGVIAFVNACSQAPGTTTTTVTTTVTTVSADDGSEAGLLS
ข)	M00F0I0LLL0VI0FV0000000000000V000V00V000000000LL0
ค)	M00F0I0LLL0VI0FV V000V00V LL

รูปที่ 2.4 ตัวอย่างการแบ่งสายลำดับโปรตีนเป็น กลุ่มย่อยๆของกรดอะมิโนที่ไม่ชอบน้ำ

จากรูปที่ 2.4 ก) แสดงสายลำดับกรดอะมิโน ข) ทำการแทนที่สัญลักษณ์กรดอะมิโนที่ชอบน้ำด้วยเลข 0 ค) ทำการตัดแบ่งกลุ่มกรดอะมิโนที่ไม่ชอบน้ำ

2.6.2 การคิดคะแนนระหว่างกลุ่มกรดอะมิโนที่ไม่ชอบน้ำ

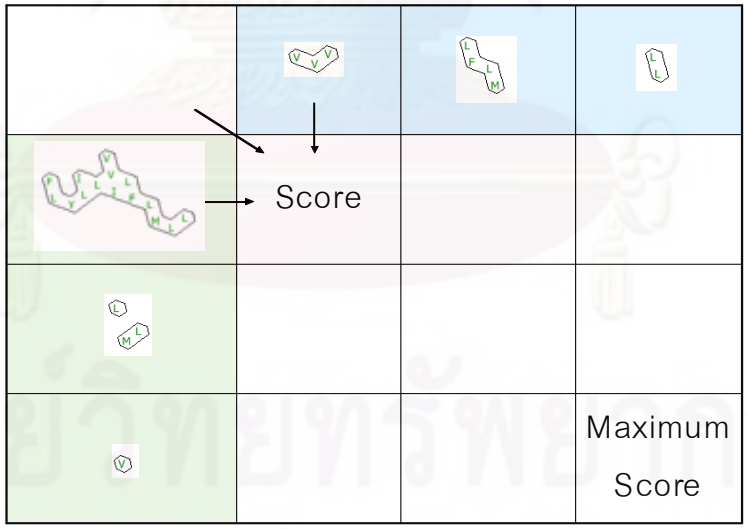
ขั้นตอนการคิดคะแนนระหว่างกลุ่มกรดอะมิโนที่ไม่ชอบน้ำเป็นการให้คะแนนระหว่างคู่กลุ่มกรดอะมิโนที่ไม่ชอบน้ำ ว่ามีความเหมือนคล้ายคลึงกันมากน้อยแค่ไหน โดยวิธีการให้คะแนนจะเลือกคะแนนที่มากที่สุด จากการเลื่อนกลุ่มกรดอะมิโนเปรียบเทียบกันจากซ้ายไปขวา การคิดคะแนนของการเลื่อนในตำแหน่งปัจจุบันทำโดยใช้ผลรวมทั้งหมดจากการเปลี่ยนแปลงกรดอะมิโน ซึ่งคะแนนการเปลี่ยนแปลงระหว่างคู่กรดอะมิโนอ้างอิงจาก ตารางคะแนนในรูปที่ 2.5 โดยตารางคะแนนนี้ปรับปรุงมาจาก ตารางคะแนน BLOSUM62 โดยคะแนนการแทนที่ระหว่างกรดอะมิโนที่ไม่ชอบน้ำกับกรดอะมิโนที่ชอบน้ำจะใช้คะแนนเฉลี่ยทุกคู่การแทนที่ของลำดับกรดอะมิโนที่ชอบน้ำที่เหลืออยู่ทั้งหมด

	M	I	L	V	F	Y	W	Other
M	**5**	1	2	1	0	-1	-1	-2
I		**4**	2	3	0	-1	-3	-3
L			**4**	1	0	-1	-2	-2
V				**4**	-1	-1	-3	-2
F					**6**	3	1	-2
Y						**7**	2	-2
W							**11**	-3

รูปที่ 2.5 ตารางคะแนน BLOSUM62 ที่ปรับปรุงจากเดิม

2.6.3 การคิดคะแนนความคล้ายคลึงกันของสายลำดับโปรตีน

ขั้นตอนการคิดคะแนนใน 2.6.2 นั้นเป็นเพียงการให้คะแนนระหว่างกลุ่มกรดอะมิโนที่ไม่ชอบน้ำ ซึ่งเป็นเพียงส่วนย่อยของสายลำดับโปรตีนเท่านั้น การให้คะแนนความคล้ายคลึงกันของโปรตีนทั้งสายลำดับนั้นจะเกิดขึ้นในขั้นตอนนี้ โดยเริ่มจากสร้างตารางคะแนนขนาดเท่ากับจำนวนของกลุ่มกรดอะมิโนในแต่ละสายลำดับ แล้วเริ่มหาคะแนนความเหมือนของแต่ละคู่กลุ่มกรดอะมิโนที่ไม่ชอบน้ำด้วยวิธีในขั้นตอน 2.6.2 โดยเริ่มหาคะแนนจากคู่กลุ่มกรดอะมิโนแรก ไปจนถึงคู่กลุ่มกรดอะมิโนสุดท้าย ทุกคู่กลุ่มกรดอะมิโนจะถูกหาคะแนนระหว่างคู่ทั้งหมด ดังแสดงในรูปที่ 2.6 โดยการหาคะแนนระหว่างคู่ปัจจุบันนั้นจะรวมคะแนนที่มากที่สุดจากคู่กลุ่มกรดอะมิโนที่อยู่ติดกันก่อนหน้า 3 กลุ่ม ได้แก่ กลุ่มซ้าย บน และแนวทะแยงของตารางและคะแนนปัจจุบันของคู่กลุ่มกรดอะมิโนเข้าไว้ด้วยกัน เมื่อคะแนนของคู่กลุ่มปัจจุบันถูกคำนวณ ลำดับของกรดอะมิโนในแต่ละกลุ่มกรดอะมิโนจะแก้ไขให้เหลือเพียงลำดับกรดอะมิโนที่เหลือจากการจับคู่ที่ให้คะแนนมากที่สุดเท่านั้น เมื่อทำการคิดคะแนนทุกคู่กลุ่มอะมิโมจนได้คะแนนสูงสุดที่คู่กลุ่มสุดท้ายของตาราง จึงค้นหาเส้นทางย้อนกลับเพื่อให้ได้การจับคู่ลำดับกรดอะมิโนที่สมบูรณ์



รูปที่ 2.6 ตัวอย่างการจับคู่กลุ่มกรดอะมิโนที่ไม่ชอบน้ำด้วยวิธีกำหนดการพลวัต [13]

บทที่ 3

วิธีดำเนินงานวิจัย

เนื่องจากผลของงานวิจัยที่กล่าวมาในบทที่ 2 ได้แสดงให้เห็นถึงประสิทธิภาพของการจับคู่ลำดับกรดอะมิโนด้วยการจับคู่กลุ่มกรดอะมิโนที่ไม่ชอบน้ำเป็นหลัก สามารถค้นหาโปรตีนที่มีหน้าที่คล้ายคลึงกันได้ถูกต้องมากกว่า การจับคู่ลำดับกรดอะมิโนในลักษณะ 1 มิติ แต่ยังเป็นการใช้เทคนิคที่สิ้นเปลืองเวลาในการประมวลผล และเทคนิคที่ใช้ยังมีลักษณะการคิดแบบละโมบ (Greedy) ซึ่งอาจทำให้ผลลัพธ์มีความผิดพลาดได้จากการเลือกจับคู่กลุ่มกรดอะมิโนที่ผิดพลาดในลำดับแรกๆ จะส่งผลต่อการจับคู่กลุ่มกรดอะมิโนในลำดับถัดไปเป็นลูกโซ่ ทำให้คะแนนความคล้ายคลึงผิดพลาดจากความเป็นจริงได้ง่าย จึงมีแนวความคิดเพิ่มเติมเทคนิคเสนอแนวทางการให้คะแนนเพื่อพัฒนาความเร็วในการประมวลผลและความถูกต้องของการเปรียบเทียบความคล้ายคลึงกับของโปรตีนได้อย่างมีประสิทธิภาพ และนำไปสร้างเป็นเครื่องมือค้นหาโปรตีนที่มีหน้าที่คล้ายคลึงกันจากฐานข้อมูลได้จริง

ขั้นตอนการพัฒนาเครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกันที่นำเสนอนี้สามารถแบ่งออกเป็น 3 ส่วน ได้แก่ การพัฒนาการให้คะแนนระหว่างกลุ่มกรดอะมิโนที่ไม่ชอบน้ำ การพัฒนาเทคนิคการค้นหาและสร้างเครื่องมือค้นหาโปรตีนที่มีหน้าที่คล้ายคลึงกันจากฐานข้อมูล และการทดสอบประสิทธิภาพของเครื่องมือค้นหาโปรตีนที่มีหน้าที่คล้ายคลึงกันจากฐานข้อมูล

3.1 การพัฒนาการให้คะแนนระหว่างกลุ่มกรดอะมิโนที่ไม่ชอบน้ำ

ในการพัฒนาสร้างเครื่องมือค้นหาโปรตีนที่มีหน้าที่เหมือนกันจากฐานข้อมูลได้อย่างมีประสิทธิภาพ ความเร็วในการเปรียบเทียบและให้คะแนนระหว่างกลุ่มกรดอะมิโนที่ไม่ชอบน้ำนั้นมีความสำคัญมาก เนื่องจากสายลำดับกรดอะมิโนหนึ่งเส้นที่ต้องการค้นหานั้นจะถูกเปรียบเทียบกับลำดับกรดอะมิโนทุกตัวที่อยู่ในฐานข้อมูล เพื่อค้นหาลำดับกรดอะมิโนที่คล้ายคลึงมากที่สุด เมื่อฐานข้อมูลมีขนาดใหญ่จึงทำให้มีจำนวนครั้งที่ต้องทำการเปรียบเทียบให้คะแนนระหว่างกลุ่มกรดอะมิโนที่ไม่ชอบน้ำเป็นจำนวนมากตามไปด้วย ซึ่งการเปรียบเทียบให้คะแนนนี้เป็นการประมวลผลที่ใช้เวลาโดยรวมมากที่สุดในระบบ การลดการคำนวณเหล่านี้ลงจะทำให้ระบบมีความเร็วมากขึ้น เทคนิคที่นำมาใช้คือ การคิดคะแนนแบบค้นหาในพจนานุกรม

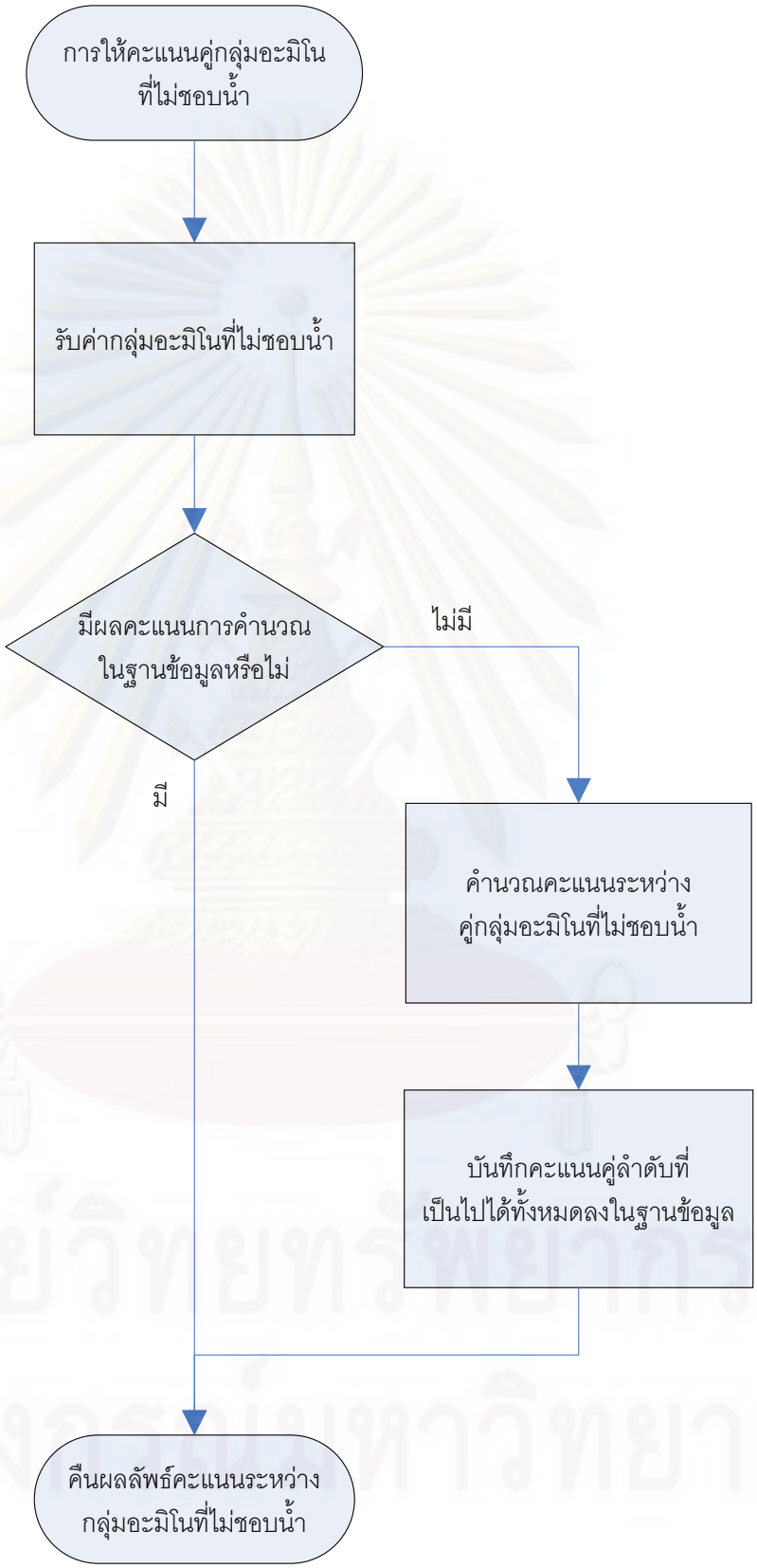
3.1.1 การคิดคะแนนแบบค้นหาในพจนานุกรม

การคิดคะแนนแบบค้นหาในพจนานุกรม จะทำการบันทึกคะแนนคู่ลำดับกรดอะมิโนที่ได้คิดคำนวณไว้แล้วลงในพจนานุกรม แล้วนำคะแนนที่บันทึกไว้มาใช้ทันที ไม่ต้องคิดคำนวณขึ้นใหม่ เมื่อพบการเปรียบเทียบกลุ่มกรดอะมิโนที่เคยคิดคะแนนไว้แล้ว โดยในแต่ละครั้งที่ต้องการเปรียบเทียบคู่กลุ่มกรดอะมิโนที่ไม่ชอบน้ำนั้น จะทำการวิเคราะห์คู่กลุ่มกรดอะมิโนนั้นว่า มีส่วนหนึ่งส่วนใด หรือกลุ่มคู่ลำดับใดบ้างที่เคยทำการคำนวณคะแนนและบันทึกคะแนนเก็บไว้ เมื่อพบก็จะดึงคะแนนส่วนนั้นมาใช้ทันที ไม่ทำการคำนวณใหม่ ทำให้เหลือการคำนวณเฉพาะเพียงบางส่วนที่ยังไม่เคยทำการคำนวณเท่านั้น เมื่อทำการเปรียบเทียบให้คะแนนคู่กลุ่มกรดอะมิโนที่ไม่ชอบน้ำนั้นเสร็จ จะทำการวิเคราะห์คู่ลำดับกรดอะมิโนที่เป็นไปได้ทั้งหมด ตัวอย่างคู่ลำดับกรดอะมิโนที่เป็นไปได้แสดงในรูปที่ 3.1 เพื่อบันทึกคู่ลำดับหรือกลุ่มกรดอะมิโนที่เคยคำนวณคะแนนแล้วไว้ในพจนานุกรมต่อไป

ก)	M00F0 LL0VI
ข)	“M,L” , “M0,LL” , “M00,LL0” , “M00F,LL0V” , “M00F0,LL0VI” , “0,L” , “00,L0” , “00F,L0V” , “00F0,L0VI” , “0,0” , “0F,0V” , “0F0,0VI”, “F,V” , “F0,VI” , “0,I”

รูปที่ 3.1 ตัวอย่างการสร้างคู่ลำดับกรดอะมิโนที่เป็นไปได้ทั้งหมด

จากรูปที่ 3.1 ตัวอย่างการสร้างคู่ลำดับกรดอะมิโนที่เป็นไปได้ทั้งหมด ก) แสดงถึงคู่กลุ่มกรดอะมิโนที่นำมาคิดคะแนน ข) แสดงผลคู่ลำดับกรดอะมิโนที่เป็นไปได้ทั้งหมดที่สร้างและนำไปบันทึก



รูปที่ 3.2 แผนภูมิการขั้นตอนการทำงาน การคิดคะแนนแบบค้นหาในพจนานุกรม

3.2 การพัฒนาเทคนิคการค้นหาและสร้างเครื่องมือค้นหาโปรตีนที่มีหน้าที่คล้ายคลึงกันจากฐานข้อมูล

จากงานวิจัยของนายภิสิทธิ์ กรรณสูต[13] การคิดคะแนนด้วยเทคนิควิธีกำหนดการพลวัตนั้น ทำการคิดคำนวณครั้งเดียวและเกิดขึ้นทั้งตารางคะแนน อีกทั้งการให้คะแนนระหว่างคู่กลุ่มกรดอะมิโนที่ไม่ชอบน้ำเป็นแบบเชิงละโมบ อาจทำให้เกิดสายลำดับกรดอะมิโนที่เหลือในการคำนวณคะแนนคู่ลำดับกลุ่มกรดอะมิโนที่ถูกนำไปใช้ในขั้นตอนต่อๆไปผิดพลาดและคะแนนผลลัพธ์สุดท้ายจึงมีโอกาสผิดไปจากความเป็นจริง

เทคนิคและแนวความคิดที่พัฒนาเพื่อเพิ่มประสิทธิภาพความถูกต้องที่วิทยานิพนธ์นี้นำเสนอ มี 2 เทคนิคคือ

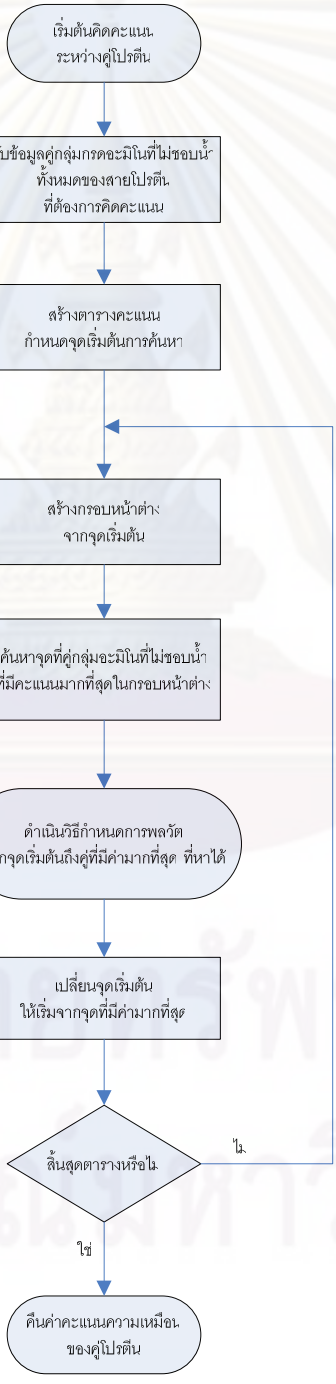
1. เทคนิคการค้นหาค่ามากที่สุดภายในกรอบหน้าต่าง (Local Window Maximum Search)
2. เทคนิคการค้นหาค่ามากที่สุดจากบนลงล่าง (Top-down Maximum Search)

โดยเทคนิคทั้ง 2 นี้ จะถูกนำไปประยุกต์ใช้กับเทคนิคการคิดคะแนนโปรตีนที่ทำหน้าที่คล้ายคลึงกัน ด้วยกำหนดการพลวัตของ นายภิสิทธิ์ กรรณสูต เพื่อลดพื้นที่ของข้อมูลทำการค้นหา และกรองข้อมูลทำให้ผลการค้นหามีประสิทธิภาพสูงขึ้น โดยนำเทคในไปใช้ในขั้นตอนภายหลังจากการแบ่งสายลำดับกรดอะมิโนให้อยู่ในรูปของกลุ่มกรดอะมิโนที่ไม่ชอบน้ำด้วยวิธีในข้อ 2.6.1 ก่อนการทำงานของวิธีกำหนดการพลวัต และใช้วิธีการคิดคะแนนระหว่างคู่กลุ่มกรดอะมิโนที่ไม่ชอบน้ำด้วยวิธีในข้อ 2.6.2

3.2.1 เทคนิคการค้นหาค่ามากที่สุดภายในกรอบหน้าต่าง (Local Window Maximum Search)

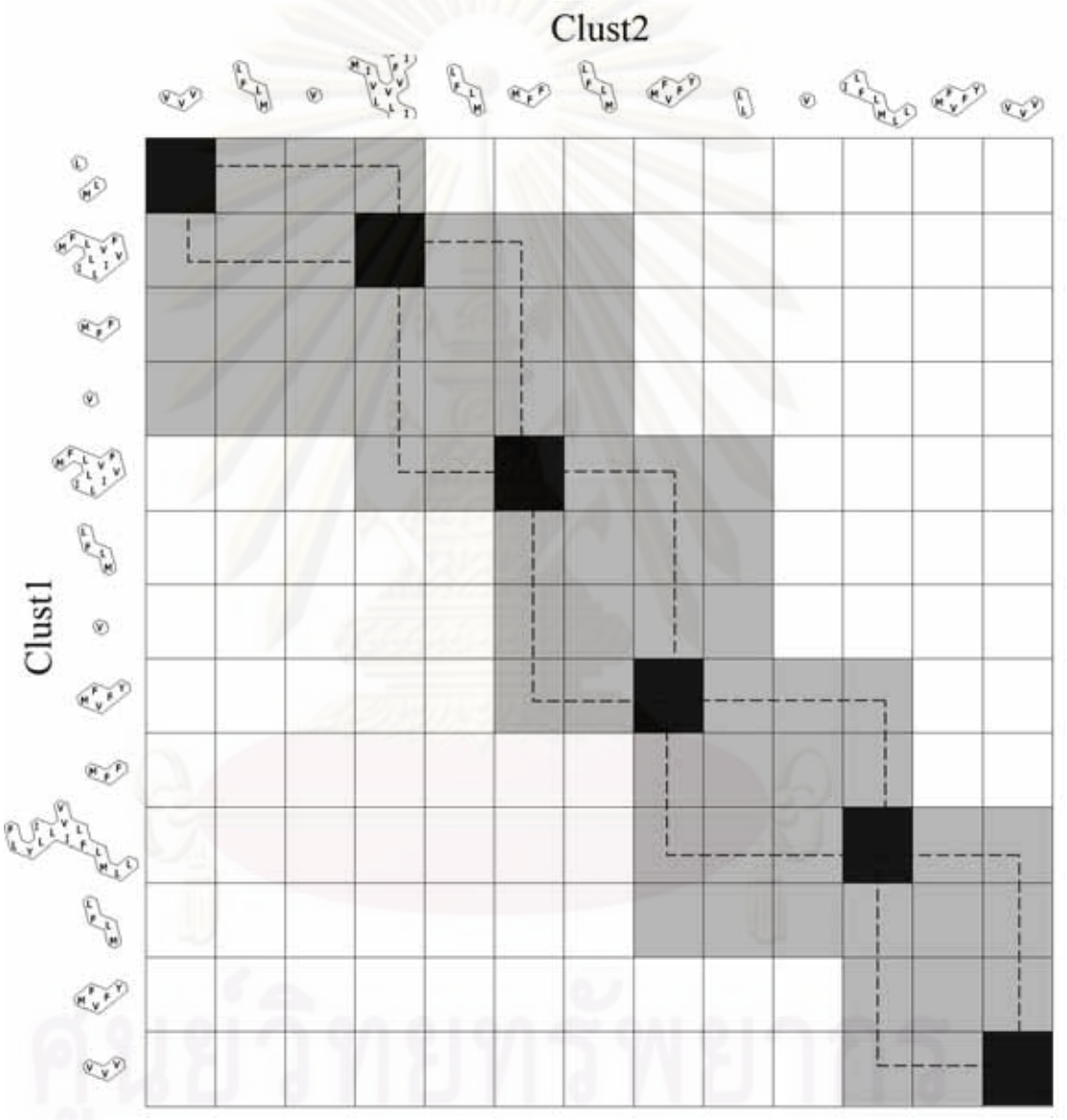
เทคนิคการค้นหาค่ามากที่สุดภายในกรอบหน้าต่างเป็นการเลือกทำวิธีกำหนดการพลวัตเพียงเฉพาะส่วน คือไม่ทำวิธีกำหนดการพลวัตครั้งเดียวทั้งตารางคะแนน แต่จะทำเป็นส่วนๆ ภายในกรอบหน้าต่างที่กำหนด โดยค้นหาคู่กลุ่มกรดอะมิโนที่ให้คะแนนความเหมือนมากที่สุดภายในกรอบหน้าต่างก่อน โดยขนาดของกรอบหน้าต่างนั้น มีขนาดเป็น 10% ของจำนวนกลุ่มกรดอะมิโนที่ไม่ชอบน้ำในโปรตีนแต่ละเส้น และการให้คะแนนความเหมือนครั้งแรกนี้ คิดทั้งกลุ่มกรดอะมิโน ไม่ได้คิดจากกรดอะมิโนที่เหลืออยู่ เมื่อได้คู่กลุ่มที่ให้คะแนนสูงที่สุดภายในกรอบหน้าต่างแล้วจึงดำเนินวิธีกำหนดการพลวัตจากจุดเริ่มต้นไปยังคู่กลุ่มที่ให้คะแนนสูงที่สุดนั้น

จากนั้นจึงเลื่อนจุดเริ่มต้นของกรอบหน้าต่างให้ไปอยู่ที่คู่กลุ่มกรดอะมิโนที่ให้คะแนนสูงสุดในครั้งสุดท้ายแล้วดำเนินการหาคู่กลุ่มคะแนนที่ให้คะแนนสูงสุดคู่ใหม่ต่อไป การให้คะแนนจะทำการเลื่อนกรอบหน้าต่าง หาคู่คะแนนสูงสุด ดำเนินวิธีกำหนดการพลวัต เช่นนี้ไปเรื่อยๆ จนถึงคู่กลุ่มสุดท้ายของลำดับกรดอะมิโน จึงได้คำตอบการจับคู่กลุ่มกรดอะมิโนที่สูงที่สุดในสายลำดับ แผนภูมิขั้นตอนการทำงานแสดงในรูปที่ 3.3



รูปที่ 3.3 แผนภูมิขั้นตอนการทำงานของเทคนิคการค้นหาค่ามากที่สุดภายในกรอบหน้าต่าง

รูปที่ 3.4 เทคนิคนี้นอกจากลดความคิดเชิงละโมบลงแล้วยังแสดงให้เห็นพื้นที่ตารางที่ทำการให้คะแนนและทำการคำนวณลดลงอีกด้วย จึงสามารถเพิ่มทั้งประสิทธิภาพทั้งด้านความเร็วและความถูกต้องของการจับคู่สายลำดับกรดอะมิโน

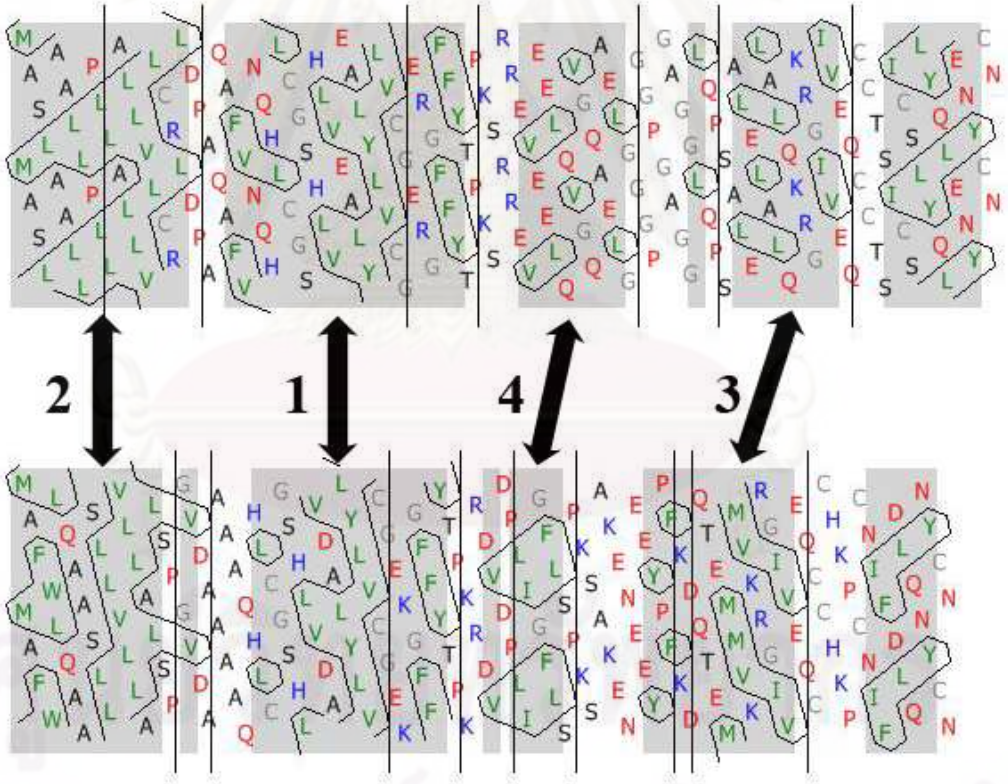


รูปที่ 3.4 พื้นที่การคำนวณของเทคนิคการค้นหาค่ามากที่สุดภายในกรอบหน้าต่าง

จากรูปที่ 3.4 กรอบสีเทาแสดงถึงพื้นที่กรอบหน้าต่างในการค้นหาคู่กลุ่มที่ให้คะแนนสูงที่สุด ตำแหน่งสีดำแสดงถึง คู่กลุ่มที่ให้คะแนนสูงที่สุดซึ่งเป็นจุดเริ่มต้นของกรอบหน้าต่างที่เลื่อนไป กรอบเส้นประแสดงถึง พื้นที่ดำเนินวิธีกำหนดการพลวัต

3.2.2 **เทคนิคการค้นหาค่ามากที่สุดจากบนลงล่าง** (Top-down Maximum Search)

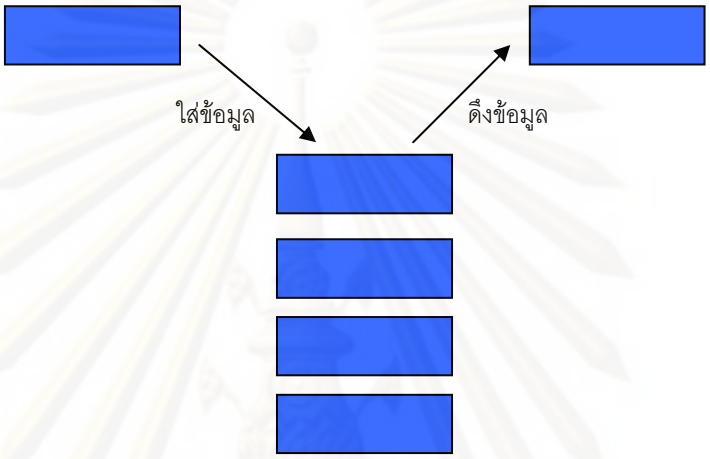
เทคนิคการค้นหาค่ามากที่สุดจากบนลงล่างใช้แนวความคิดคล้ายคลึงกับวิธีการคิดในการจับคู่กลุ่มกรดอะมิโนที่ไม่ชอบน้ำโดยมนุษย์ คือแก้ปัญหาจากบนลงล่าง หรือ มองจากปัญหาใหญ่ไปยังปัญหาย่อย ซึ่งแนวทางการจับคู่กรดอะมิโนที่ไม่ชอบน้ำโดยมนุษย์นั้นจะพยายามจับคู่กลุ่มกรดอะมิโนที่มีความคล้ายคลึงกันมากที่สุดจากภาพรวมทั้งสายลำดับของโปรตีนก่อน จากนั้นจึงจับคู่กลุ่มกรดอะมิโนที่มีความคล้ายคลึงมากรองลงมา ภายในช่วงการค้นหาที่เล็กลงไปเรื่อยๆ จนครบทั้งสายลำดับกรดอะมิโน ดังตัวอย่างที่แสดงในรูปที่ 3.5 การจับคู่จะเริ่มจากการหาคู่กลุ่มกรดอะมิโนที่คล้ายคลึงกันมากที่สุดจากทั้งสายโปรตีน คือคู่หมายเลข 1 แล้วจึงหาคู่รองลงมาคือหมายเลข 2 และหมายเลข 3 ตามลำดับ จากนั้น จึงหาคู่ที่คล้ายคลึงมากที่สุด ระหว่างคู่หมายเลข 1 กับหมายเลข 3 จึงได้การจับคู่ หมายเลข 4



รูปที่ 3.5 ตัวอย่างขั้นตอนการจับคู่กลุ่มกรดอะมิโนจากบนลงล่าง

การพัฒนาแนวความคิดนี้ ต้องใช้โครงสร้างข้อมูลแบบกองซ้อน (Stack) ตัวอย่างโครงสร้างแสดงในรูปที่ 3.6 ซึ่งมีลักษณะการเก็บและดึงข้อมูลกลับดังนี้

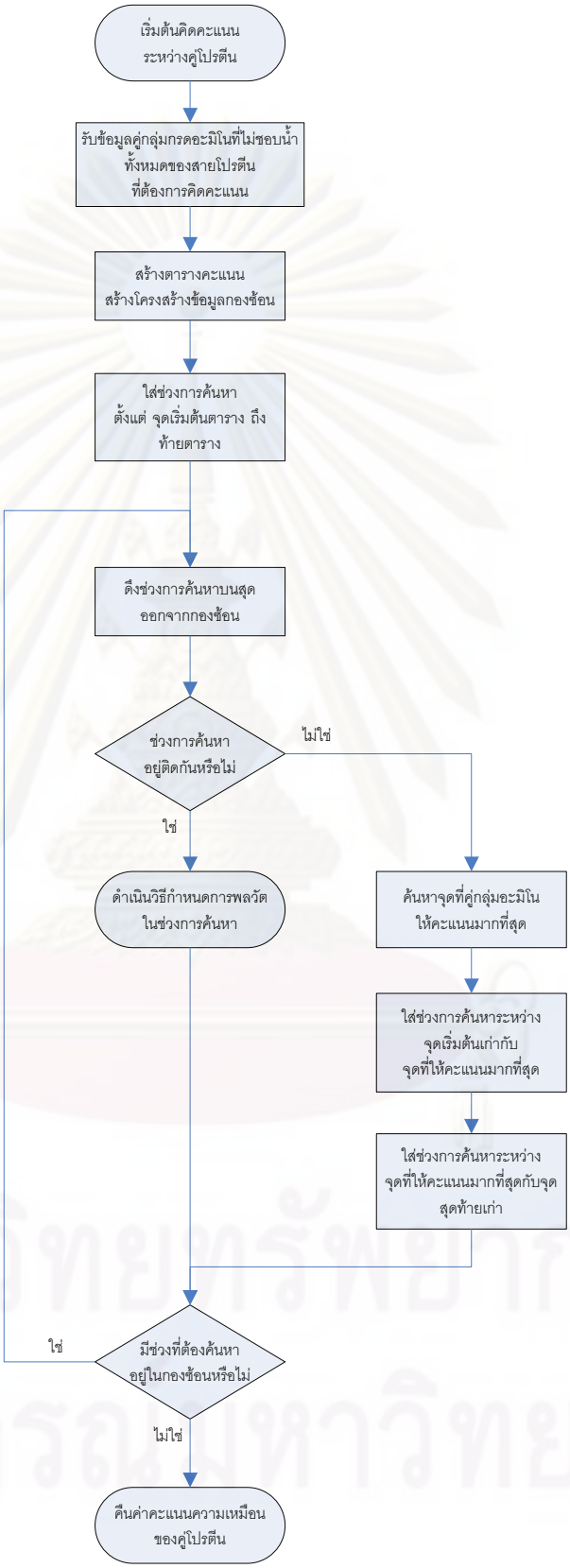
1. เมื่อใส่ข้อมูลลงกองซ้อน ข้อมูลจะอยู่ด้านบนกองซ้อนเสมอ
2. สามารถดึงข้อมูลได้จากตัวบนสุดของกองซ้อนเท่านั้น



รูปที่ 3.6 โครงสร้างข้อมูลแบบกองซ้อน

ซึ่งคุณสมบัติของโครงสร้างข้อมูลแบบกองซ้อนนี้ทำให้สามารถออกแบบการค้นหาค่ามากที่สุดจากบนลงล่าง โดยขั้นตอนการทำงานดังนี้ และแผนภูมิขั้นตอนการทำงานแสดงในรูปที่ 3.7

1. เริ่มจากสร้างโครงสร้างข้อมูลแบบกองซ้อน แล้วใส่ช่วงการค้นหาแรก คือ ช่วงตั้งแต่จุดเริ่มต้นของตาราง ไปถึง จุดสุดท้ายของตาราง
2. ดึงข้อมูลที่อยู่บนสุดของกองซ้อนและตั้งช่วงการค้นหาให้เท่ากับข้อมูลที่ดึงขึ้นมา
3. ถ้าช่วงการค้นหานั้น เป็นช่วงที่อยู่ติดกัน ให้ทำการคิดคะแนนด้วยวิธีให้คะแนนด้วยกำหนดการพลวัต
4. แต่ถ้าไม่อยู่ติดกัน ให้ทำการค้นหาจุดคู่ลำดับที่ให้คะแนนมากที่สุดภายในช่วงการค้นหานั้น และทำการสร้างช่วงการค้นหาเพิ่มขึ้น 2 ช่วงคือ
   – จุดเริ่มต้นเก่าถึงจุดที่คะแนนมากที่สุด
   - จุดที่คะแนนมากที่สุดถึงจุดสุดท้ายเก่า
   และเก็บช่วงการค้นหาทั้งสอง ไว้ส่วนบนสุดของกองซ้อน
5. ดำเนินการตั้งแต่ ข้อ2 ซ้ำ จนกระทั่ง หมดทุกช่วงการค้นหาที่เก็บอยู่ในกองซ้อน



รูปที่ 3.7 แผนภูมิขั้นตอนการทำงานของเทคนิคการค้นหาค่ามากที่สุดจากบนลงล่าง

บทที่ 4

การทดลองและผลการทดลอง

ในขั้นตอนวิธีการทดลองและผลการทดลองของวิทยานิพนธ์เรื่อง เครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกันด้วยวิธีการจัดเรียงกลุ่มกรดอะมิโนที่ไม่ชอบน้ำแบบค้นหาเฉพาะส่วน ทดสอบถึงประสิทธิภาพในด้านความเร็วที่พัฒนาขึ้นเมื่อนำเทคนิคการคิดคะแนนแบบค้นหาในพจนานุกรม ในหัวข้อ3.1.1 มาใช้ และทดสอบความถูกต้องในการค้นหาหน้าที่การทำงานของโปรตีนจากฐานข้อมูลด้วยเทคนิคที่นำเสนอในหัวข้อ 3.2 โดยฐานข้อมูลโปรตีนที่ใช้ในการทดลองนี้มาจาก 2 ฐานข้อมูลหลัก ได้แก่ PIR และ HOMSTRAD ซึ่งเป็นฐานข้อมูลที่ถูกแบ่งกลุ่มของโปรตีนตามหน้าที่การทำงาน

4.1 ขั้นตอนวิธีการเตรียมชุดข้อมูลที่ใช้ในการทดลอง

สำหรับชุดข้อมูลที่ใช้ในการทดลองนี้แบ่งเป็น 3 ชุดข้อมูล คือ ชุดข้อมูลสำหรับการทดสอบประสิทธิภาพด้านความเร็วเมื่อใช้เทคนิคการคิดคะแนนแบบค้นหาในพจนานุกรม ชุดข้อมูลจากฐานข้อมูล HOMSTRAD เพื่อใช้ในการทดสอบความถูกต้องของเครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกัน และ ชุดข้อมูลจากฐานข้อมูล PIR เพื่อใช้ในการทดสอบความถูกต้องของเครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกัน รายละเอียดขั้นตอนการเตรียมชุดข้อมูล และลักษณะของชุดข้อมูล เป็นดังต่อไปนี้

4.1.1 ชุดข้อมูลสำหรับการทดสอบประสิทธิภาพด้านความเร็วเมื่อใช้เทคนิคการคิดคะแนนแบบค้นหาในพจนานุกรม

ชุดข้อมูลสำหรับการทดสอบประสิทธิภาพด้านความเร็วเมื่อใช้เทคนิคการคิดคะแนนแบบค้นหาในพจนานุกรม เป็นชุดข้อมูลที่นำไปใช้ในการทดสอบด้านความเร็วเท่านั้น ไม่ได้ใช้ในการทดสอบความถูกต้องของเครื่องมือ ดังนั้นชุดข้อมูลนี้ จึงเป็นชุดข้อมูลที่ไม่ต้องมีการแบ่งกลุ่มหน้าที่การทำงานของโปรตีน ข้อมูลในชุดข้อมูลจึงถูกทำการเลือกแบบสุ่มโดยเลือกเป็นชุดข้อมูลในการทดสอบ 100 สายลำดับกรดอะมิโน และสำหรับชุดข้อมูลในการเปรียบเทียบมีขนาดแตกต่างกันดังแสดงในตารางที่ 4.1

ตารางที่ 4.1 ชุดข้อมูลสำหรับการทดสอบประสิทธิภาพด้านความเร็ว
เมื่อใช้เทคนิคการคิดคะแนนแบบค้นหาในพจนานุกรม

ลำดับที่	จำนวนข้อมูลเปรียบเทียบ	จำนวนข้อมูลทดสอบ
1	50	100
2	100	100
3	500	100
4	1000	100
5	5000	100
6	10000	100
7	15000	100
8	30000	100

4.1.2 ชุดข้อมูลจากฐานข้อมูล HOMSTRAD เพื่อใช้ในการทดสอบความถูกต้องของเครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกัน

ชุดข้อมูลนี้ เป็นชุดข้อมูลที่เตรียมให้คล้ายกับชุดทดลองที่ใช้วิทยานิพนธ์ของนายภิสิทธิ์ กรรณสูต โดยชุดข้อมูลถูกเตรียมจากฐานข้อมูล HOMSTRAD ซึ่งมีการแบ่งกลุ่มโปรตีนตามหน้าที่การทำงานเป็น 1032 กลุ่ม โดยรายละเอียดของกลุ่มโปรตีนแสดงในภาคผนวก ก.1 โดยชุดข้อมูลจะแบ่งเป็นชุดข้อมูลทดสอบ และชุดข้อมูลเปรียบเทียบ สำหรับชุดข้อมูลทดสอบนั้นจะทำการสุ่มเลือกมาจากทุกกลุ่มโปรตีน กลุ่มละ 1 สายลำดับ ทำให้ชุดข้อมูลทดสอบมีขนาด 1032 สายลำดับคงที่ และสำหรับชุดข้อมูลที่ใช้ในการเปรียบเทียบ เกิดจากการสุ่มเลือกสายลำดับจากฐานข้อมูลที่ไม่ซ้ำกับชุดข้อมูลที่ใช้ในการทดสอบ และแต่ละกลุ่มโปรตีนต้องถูกเลือกอย่างน้อย 1 สายลำดับ รายละเอียดจำนวนสายลำดับข้อมูลและจำนวนกลุ่มโปรตีนแสดงไว้ในตารางที่ 4.2

ตารางที่ 4.2 ชุดข้อมูลจากฐานข้อมูล HOMSTRAD เพื่อใช้ในการทดสอบความถูกต้องของเครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกัน

ลำดับที่	จำนวนข้อมูลเปรียบเทียบ	จำนวนข้อมูลทดสอบ	จำนวนกลุ่ม
1	3448	1032	1032
2	44813	1032	1032
3	86678	1032	1032
4	187991	1032	1032
5	336827	1032	1032

4.1.3 ชุดข้อมูลจากฐานข้อมูล PIRเพื่อใช้ในการทดสอบความถูกต้องของเครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกัน

เพื่อให้ชุดข้อมูลที่ใช้ในการทดสอบมีความเป็นมาตรฐานมากขึ้น ชุดข้อมูลนี้จึงถูกเตรียมจากฐานข้อมูล PIR โดยเลือกการแบ่งกลุ่มของโปรตีนเป็น 67 กลุ่ม ตามชุดข้อมูลในการทดลองของ William R. Person [23] ซึ่งใช้เป็นมาตรฐานในการทดสอบเทคนิคการจับคู่สายลำดับกรดอะมิโนอื่นๆ โดยรายละเอียดของกลุ่มโปรตีนได้แสดงในภาคผนวก ก.2 โดยชุดข้อมูลจะแบ่งเป็นชุดข้อมูลทดสอบ และชุดข้อมูลเปรียบเทียบ ชุดข้อมูลทั้ง 2 จะมีข้อมูลแบบเดียวกัน โดยทำการเลือกแบบสุ่มมาจากฐานข้อมูล โดยให้แต่ละกลุ่มโปรตีน ต้องถูกเลือกอย่างน้อย 1 สายลำดับ รายละเอียดจำนวนสายลำดับข้อมูลและจำนวนกลุ่มโปรตีนแสดงไว้ในตารางที่ 4.3

ตารางที่ 4.3 ชุดข้อมูลจากฐานข้อมูล PIR เพื่อใช้ในการทดสอบความถูกต้องของเครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกัน

ลำดับที่	จำนวนข้อมูลเปรียบเทียบ	จำนวนข้อมูลทดสอบ	จำนวนกลุ่ม
1	500	500	67
2	1000	1000	67
3	2500	2500	67
4	5000	5000	67
5	10000	10000	67

4.2 ขั้นตอนวิธีทำการทดลอง

ในวิทยานิพนธ์นี้แบ่งการทดสอบประสิทธิภาพของเครื่องมือเป็น 3 การทดลอง ได้แก่ การทดสอบประสิทธิภาพด้านความเร็วเมื่อใช้เทคนิคการคิดคะแนนแบบค้นหาในพจนานุกรม การทดสอบความถูกต้องของเครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกันบนชุดข้อมูลจากฐานข้อมูล HOMSTRAD และ การทดสอบความถูกต้องของเครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงบนชุดข้อมูลจากฐานข้อมูล PIR

4.2.1 การทดสอบประสิทธิภาพด้านความเร็วของเทคนิคการคิดคะแนนแบบค้นหาในพจนานุกรม

การทดสอบนี้เป็นการทดสอบประสิทธิภาพด้านความเร็วของเทคนิคการคิดคะแนนแบบค้นหาในพจนานุกรม การทดลองนี้ทำการเปรียบเทียบระยะเวลาที่ใช้ในการคิดคะแนนความคล้ายคลึงกันของคู่โปรตีน ระหว่างการใช้เทคนิคการคิดคะแนนแบบค้นหาในพจนานุกรมร่วมด้วยกับไม่ใช้เทคนิคการคิดคะแนนแบบค้นหาในพจนานุกรม โดยใช้ชุดข้อมูลทดสอบประสิทธิภาพด้านความเร็วในข้อ 4.1.1 ซึ่งมีจำนวนข้อมูลที่ใช้ในการทดสอบคงที่ และเปลี่ยนแปลงจำนวนข้อมูลที่ใช้ในการเปรียบเทียบให้มากขึ้น โดยวิธีการทดลองมีขั้นตอนดังนี้

1. นำสายลำดับโปรตีนทั้งหมดของชุดข้อมูลทดสอบ เปรียบเทียบคิดหาคะแนนความคล้ายคลึงกับสายลำดับโปรตีนทั้งหมดในชุดข้อมูลเปรียบเทียบ
2. จับระยะเวลาทั้งหมดที่ใช้ในการคำนวณคิดคะแนนความคล้ายคลึงกัน
3. คำนวณระยะเวลาเฉลี่ยที่ใช้ในการคิดคะแนนความคล้ายคลึงกันของคู่โปรตีน โดยนำระยะเวลาที่ใช้ทั้งหมดหารด้วยผลคูณของจำนวนชุดข้อมูลทดสอบและชุดข้อมูลเปรียบเทียบ
4. สร้างกราฟเปรียบเทียบระยะเวลาเฉลี่ยที่ใช้ในการคิดคะแนนระหว่าง ระยะเวลาเฉลี่ยเมื่อใช้การคิดคะแนนแบบค้นหาในพจนานุกรม กับ ระยะเวลาเฉลี่ยเมื่อไม่ใช้

4.2.2 การทดสอบความถูกต้องของเครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกันบนชุดข้อมูลจากฐานข้อมูล HOMSTRAD

การทดลองนี้ทำเพื่อทดสอบความถูกต้องของเครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกัน โดยใช้ชุดข้อมูลจากฐานข้อมูล HOMSTRAD ในข้อ 4.1.2 การทดลองนี้เป็นการเปรียบเทียบความถูกต้องในการค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกัน ระหว่างเครื่องมือที่ใช้กำหนดการพลวัตของนายภิสิทธิ์ กรรณสูต กับกำหนดการพลวัตที่เพิ่มวิธีการค้นหาค่ามากที่สุดภายในกรอบหน้าต่าง และวิธีการค้นหาค่ามากที่สุดจากบนลงล่าง ขั้นตอนวิธีการทดลองมีรายละเอียดดังนี้

1. นำสายลำดับโปรตีนแต่ละสายลำดับในชุดข้อมูลทดสอบไปค้นหาในชุดข้อมูลเปรียบเทียบเพื่อหาโปรตีนที่ทำหน้าที่คล้ายคลึงกันมากที่สุด ด้วยเทคนิควิธีการคิดคะแนนและค้นหาทั้ง 3 วิธี
2. ตรวจความถูกต้องของโปรตีนที่พบจากการค้นหา ว่าจัดอยู่ในกลุ่มเดียวกันหรือไม่ ถ้าจัดอยู่ในกลุ่มเดียวกัน ให้ถือว่าการค้นหาถูกต้อง
3. คำนวณหาประสิทธิภาพร้อยละความถูกต้องของแต่ละวิธี
4. รายงานผลการทดลองและสร้างกราฟเปรียบเทียบประสิทธิภาพความถูกต้องของแต่ละวิธี

4.2.3 การทดสอบความถูกต้องของเครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกันบนชุดข้อมูลจากฐานข้อมูล PIR

การทดลองนี้ทำเพื่อทดสอบความถูกต้องของเครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกัน โดยใช้ชุดข้อมูลจากฐานข้อมูล PIR ในข้อ 4.1.3 การทดลองนี้เป็นการเปรียบเทียบความถูกต้องในการค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกัน ระหว่างเครื่องมือที่ใช้กำหนดการพลวัตของนายภิสิทธิ์ กรรณสูต กับกำหนดการพลวัตที่เพิ่มวิธีการค้นหาค่ามากที่สุดภายในกรอบหน้าต่าง และวิธีการค้นหาค่ามากที่สุดจากบนลงล่าง ขั้นตอนวิธีการทดลองมีรายละเอียดดังนี้

1. นำสายลำดับโปรตีนแต่ละสายลำดับในชุดข้อมูลทดสอบไปค้นหาในชุดข้อมูลเปรียบเทียบยกเว้นสายลำดับเดียวกับสายลำดับที่นำมาค้นหา เพื่อหาโปรตีนที่ทำหน้าที่คล้ายคลึงกันมากที่สุด ทั้ง 3 วิธี

2. ตรวจความถูกต้องของโปรตีนที่พบจากการค้นหา ว่าจัดอยู่ในกลุ่มเดียวกันหรือไม่ ถ้าจัดอยู่ในกลุ่มเดียวกัน ให้ถือว่าการค้นหาถูกต้อง
3. คำนวณหาประสิทธิภาพร้อยละความถูกต้องของแต่ละวิธี
4. รายงานผลการทดลองและสร้างกราฟเปรียบเทียบประสิทธิภาพความถูกต้องของแต่ละวิธี

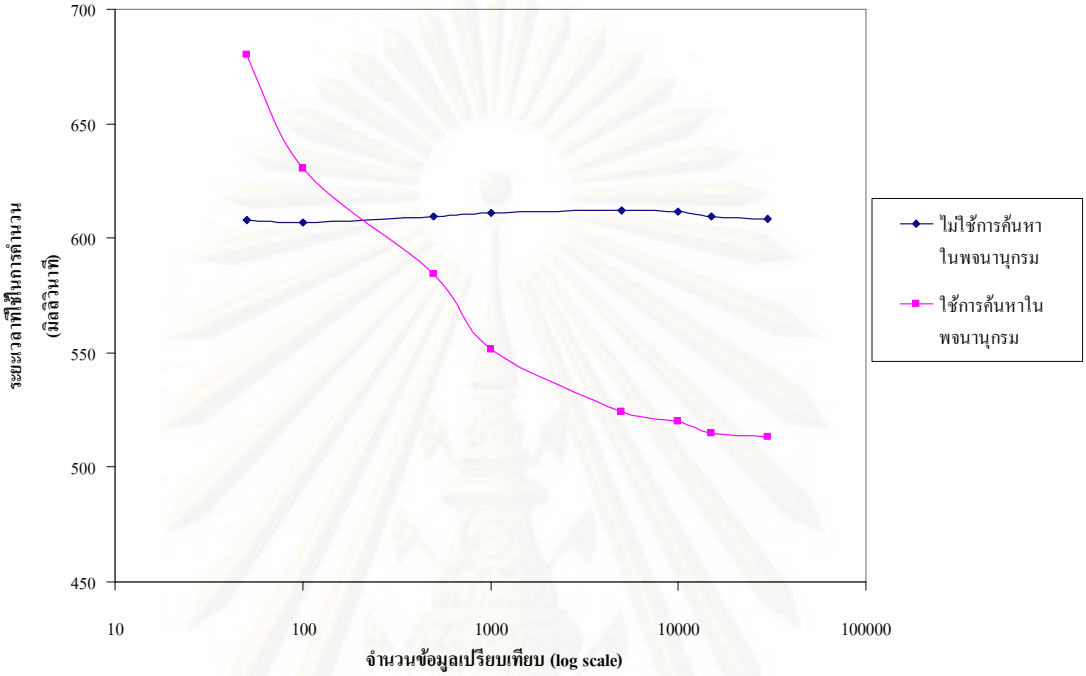
4.3 ผลการทดลอง

จากขั้นตอนวิธีทำการทดลองข้างต้น ผลการทดลองสามารถแบ่งแยกตามหัวข้อที่ได้ทำการทดลองดังต่อไปนี้

4.3.1 ผลการทดสอบประสิทธิภาพด้านความเร็วของการคิดคะแนนแบบค้นหาในพจนานุกรม

ตารางที่ 4.4 ผลการทดสอบประสิทธิภาพด้านความเร็ว
ของการคิดคะแนนแบบค้นหาในพจนานุกรม

จำนวนข้อมูลเปรียบเทียบ	จำนวนข้อมูลทดสอบ	ระยะเวลาเฉลี่ยที่ใช้คำนวณคะแนนความคล้ายคลึงกันระหว่างคู่โปรตีน (มิลลิวินาที)	
		ไม่ใช้เทคนิคการค้นหาในพจนานุกรม	ใช้เทคนิคการค้นหาในพจนานุกรม
50	100	607.86	680.34
100	100	607.06	630.31
500	100	609.51	584.44
1000	100	611.19	551.46
5000	100	612.11	524.10
10000	100	611.49	520.07
15000	100	609.34	514.64
30000	100	608.65	513.24



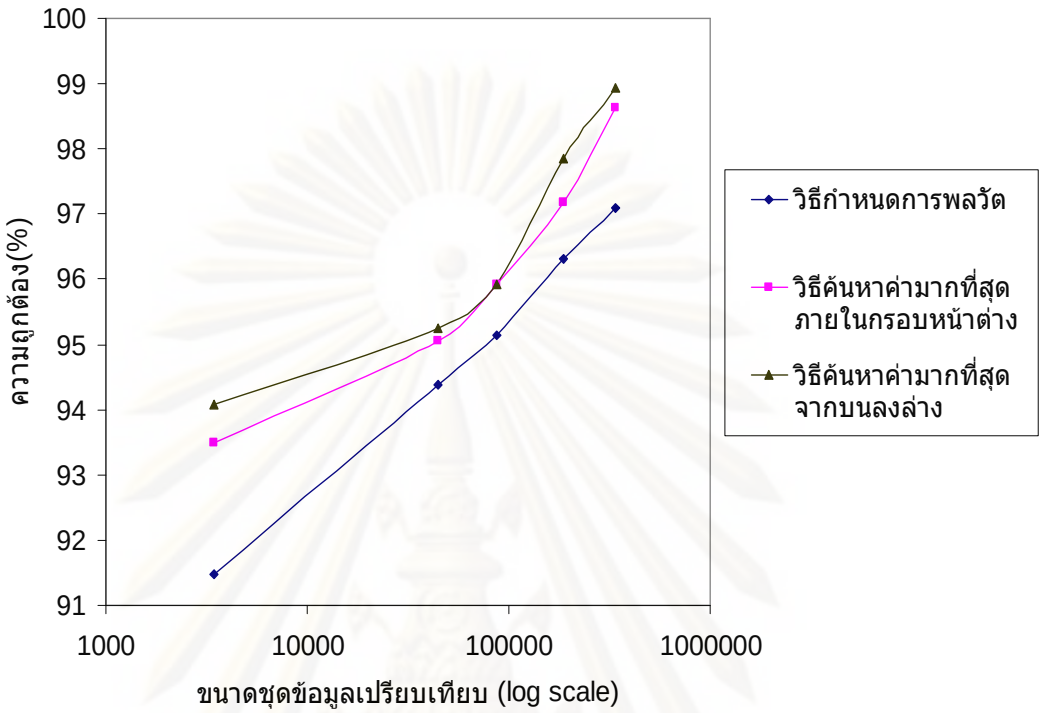
รูปที่ 4.1 กราฟเปรียบเทียบระยะเวลาเฉลี่ยที่ใช้ในการคำนวณคะแนนความคล้ายคลึงกันของคู่โปรตีน

จากผลการทดลองข้างต้นแสดงให้เห็นว่า เมื่อนำการค้นหาในพจนานุกรมมาใช้ร่วมในการคิดคำนวณคะแนนความคล้ายคลึงกันของคู่โปรตีน ทำให้สามารถคิดคำนวณคะแนนได้เร็วขึ้น เมื่อมีจำนวนข้อมูลที่เปรียบเทียบมากขึ้น แต่เมื่อใช้ข้อมูลเปรียบเทียบที่มีจำนวนน้อยจะใช้ระยะเวลาเฉลี่ยในการคำนวณมากกว่า เนื่องจากเสียเวลาในการคำนวณให้กับการสร้างคู่ลำดับที่เป็นไปได้ทั้งหมดและเก็บบันทึกลงในพจนานุกรม โดยเมื่อใช้จำนวนของข้อมูลเปรียบเทียบมากที่สุด การค้นหาในพจนานุกรม ทำให้ประสิทธิภาพการคิดคำนวณคะแนนเร็วขึ้น 15.67%

4.3.2 ผลการทดสอบความถูกต้องของเครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกันบนชุดข้อมูลจากฐานข้อมูล HOMSTRAD

ตารางที่ 4.5 ผลการทดสอบความถูกต้องของเครื่องมือค้นหาโปรตีน
ที่ทำหน้าที่คล้ายคลึงกันบนชุดข้อมูลจากฐานข้อมูล HOMSTRAD

ขนาดชุดข้อมูลเปรียบเทียบ	ขนาดชุดข้อมูลทดสอบ	เทคนิคที่ใช้ในการทดลอง	ความถูกต้อง
3448	1032	กำหนดการพลวัต	91.47%
		การค้นหาค่ามากที่สุดภายในกรอบหน้าต่าง	93.50%
		การค้นหาค่ามากที่สุดจากบนลงล่าง	**94.09%**
44813	1032	กำหนดการพลวัต	94.38%
		การค้นหาค่ามากที่สุดภายในกรอบหน้าต่าง	95.05%
		การค้นหาค่ามากที่สุดจากบนลงล่าง	**95.25%**
86678	1032	กำหนดการพลวัต	95.15%
		การค้นหาค่ามากที่สุดภายในกรอบหน้าต่าง	95.93%
		การค้นหาค่ามากที่สุดจากบนลงล่าง	**95.93%**
187991	1032	กำหนดการพลวัต	96.32%
		การค้นหาค่ามากที่สุดภายในกรอบหน้าต่าง	97.18%
		การค้นหาค่ามากที่สุดจากบนลงล่าง	**97.86%**
336827	1032	กำหนดการพลวัต	97.09%
		การค้นหาค่ามากที่สุดภายในกรอบหน้าต่าง	98.64%
		การค้นหาค่ามากที่สุดจากบนลงล่าง	**98.93%**



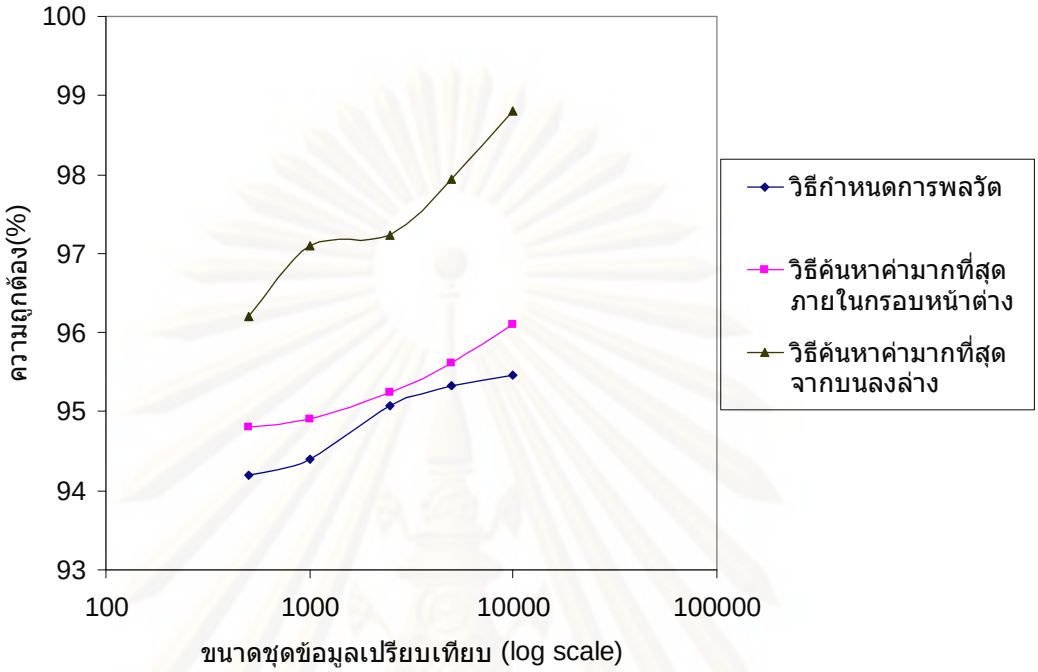
รูปที่ 4.2 กราฟเปรียบเทียบความถูกต้องของแต่ละวิธีที่ใช้ในการพัฒนาเครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกัน บนชุดข้อมูลจากฐานข้อมูล HOMSTRAD

จากผลการทดลองแสดงให้เห็นว่า การนำวิธีการค้นหาค่ามากที่สุดภายในกรอบหน้าต่าง และ วิธีการค้นหาค่ามากที่สุดจากบนลงล่างมาใช้ร่วมกับวิธีกำหนดการพลวัต ทำให้สามารถพัฒนาเครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกันได้มีประสิทธิภาพความถูกต้องที่สูงขึ้น โดยวิธีการค้นหาค่ามากที่สุดจากบนลงล่าง ให้ร้อยละความถูกต้องในการค้นหาสูงที่สุด โดยมีร้อยละความถูกต้องอยู่ที่ 98.93%

4.3.3 ผลการทดสอบความถูกต้องของเครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกันบนชุดข้อมูลจากฐานข้อมูล PIR

ตารางที่ 4.6 ผลการทดสอบความถูกต้องของเครื่องมือค้นหาโปรตีน
ที่ทำหน้าที่คล้ายคลึงกันบนชุดข้อมูลจากฐานข้อมูล PIR

ขนาดชุดข้อมูลเปรียบเทียบ	ขนาดชุดข้อมูลทดสอบ	เทคนิคที่ใช้ในการทดลอง	ความถูกต้อง
500	500	กำหนดการพลวัต	94.20%
		การค้นหาค่ามากที่สุดภายในกรอบหน้าต่าง	94.80%
		การค้นหาค่ามากที่สุดจากบนลงล่าง	**96.20%**
1000	1000	กำหนดการพลวัต	94.40%
		การค้นหาค่ามากที่สุดภายในกรอบหน้าต่าง	94.90%
		การค้นหาค่ามากที่สุดจากบนลงล่าง	**97.10%**
2500	2500	กำหนดการพลวัต	95.08%
		การค้นหาค่ามากที่สุดภายในกรอบหน้าต่าง	95.24%
		การค้นหาค่ามากที่สุดจากบนลงล่าง	**97.24%**
5000	5000	กำหนดการพลวัต	95.32%
		การค้นหาค่ามากที่สุดภายในกรอบหน้าต่าง	95.62%
		การค้นหาค่ามากที่สุดจากบนลงล่าง	**97. 94%**
10000	10000	กำหนดการพลวัต	95.47%
		การค้นหาค่ามากที่สุดภายในกรอบหน้าต่าง	96.11%
		การค้นหาค่ามากที่สุดจากบนลงล่าง	**98.81%**



รูปที่ 4.3 กราฟเปรียบเทียบความถูกต้องของแต่ละวิธีที่ใช้ในการพัฒนาเครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกัน บนชุดข้อมูลจากฐานข้อมูล PIR

จากผลการทดลองแสดงให้เห็นว่า การนำวิธีการค้นหาค่ามากที่สุดจากบนลงล่างมาใช้ร่วมกับวิธีกำหนดการพลวัต ทำให้ได้เครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกันได้มีประสิทธิภาพความถูกต้องสูงที่สุด โดยมีความถูกต้องอยู่ที่ 98.81% และการใช้วิธีการค้นหาค่ามากที่สุดภายในกรอบหน้าต่างให้ความถูกต้องรองลงมาคือ 96.11% และการใช้กำหนดการพลวัตเพียงอย่างเดียวทำให้ได้ประสิทธิภาพต่ำที่สุดมีความถูกต้องอยู่ที่ 95.47%

แม้ว่าผลการทดลองแสดงถึงความถูกต้องที่สูงมาก แต่เมื่อวิเคราะห์ถึงรายละเอียดความผิดพลาดในการค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกันด้วยเทคนิคที่นำเสนอจะพบว่าความผิดพลาดเกือบทั้งหมดเกิดขึ้นกับกลุ่มโปรตีนที่มีขนาดเล็ก และเมื่อแบ่งสายลำดับกรดอะมิโนให้อยู่ในรูปของกลุ่มย่อยของกรดอะมิโนที่ไม่ชอบน้ำแล้ว จะมีจำนวนกลุ่มกรดอะมิโนที่ไม่ชอบน้ำเป็นจำนวนน้อย ทำให้เครื่องมือที่นำเสนอมีความผิดพลาด ไม่สามารถทำการค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกันได้อย่างถูกต้อง

บทที่ 5

สรุปผลการวิจัยและข้อเสนอแนะ

ในวิทยานิพนธ์ฉบับนี้ ผู้วิจัยเสนอแนวทางเพื่อการพัฒนาเทคนิคการจับคู่และหาความคล้ายคลึงของโปรตีนโดยอัตโนมัติ ด้วยวิธีการจัดเรียงกลุ่มกรดอะมิโนที่ไม่ชอบน้ำแบบค้นหาเฉพาะส่วน โดยจากผลการทดลองในบทที่ 4 ผลการทดลองสามารถสรุปได้ดังนี้

5.1 สรุปผลการวิจัย

เครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกัน โดยการจับคู่กลุ่มที่ไม่ชอบน้ำแบบ 2 มิติโดยอัตโนมัติ ซึ่งอาศัยกำหนดการพลวัตร่วมกับการวิเคราะห์กลุ่มที่ไม่ชอบน้ำนั้นสามารถค้นหาโปรตีนที่มีหน้าที่คล้ายคลึงกันได้ถูกต้องมากกว่าการจับคู่ลำดับกรดอะมิโนในลักษณะ 1 มิติ แต่เทคนิคกำหนดการพลวัตและวิธีการคิดคำนวณคะแนนความคล้ายคลึงกันของคู่โปรตีนที่ใช้ยังคงตรงไปตรงมา ทำให้สิ้นเปลืองเวลาในการประมวลผล และวิธีที่ใช้เป็นการคิดจับคู่จากต้นไปท้ายเพียงอย่างเดียว ซึ่งอาจทำให้ผลลัพธ์ผิดพลาด ดังนั้นวิทยานิพนธ์นี้จึงนำเสนอเทคนิคในการพัฒนาประสิทธิภาพทั้งด้านความเร็วและความถูกต้องในการค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกัน

วิทยานิพนธ์นี้นำเสนอวิธีพัฒนาประสิทธิภาพด้านความเร็ว คือการคิดคะแนนแบบค้นหาในพจนานุกรมร่วมกับการคิดคะแนนความคล้ายคลึงกันของคู่โปรตีน เพื่อลดระยะเวลาที่สูญเสียไปกับการคิดคำนวณที่ซ้ำซ้อน เมื่อฐานข้อมูลมีขนาดใหญ่ขึ้น จากผลการทดลองประสิทธิภาพเมื่อใช้ฐานข้อมูลที่มีขนาดใหญ่มากที่สุด การคิดคะแนนแบบค้นหาในพจนานุกรมนี้ทำให้ระยะเวลาในการคำนวณเร็วขึ้น 15.67% แต่ถ้าฐานข้อมูลเปรียบเทียบมีขนาดเล็กการใช้วิธีนี้จะเสียเวลาในการคำนวณให้กับการสร้างคู่ลำดับที่เป็นไปได้ทั้งหมดและเก็บบันทึกลงในพจนานุกรม ทำให้ระยะเวลาเฉลี่ยในการคิดคะแนนไม่ต่างกัน

ในด้านความถูกต้องของการค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกัน วิทยานิพนธ์นี้ได้นำเสนอ 2 วิธี คือ การค้นหาค่ามากที่สุดภายในกรอบหน้าต่าง และการค้นหาค่ามากที่สุดจากบนลงล่าง โดยทั้ง 2 วิธีที่นำเสนอนี้ เมื่อนำมาใช้ร่วมกับกำหนดการพลวัต ทำให้ประสิทธิภาพในการค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกัน มีความถูกต้องเพิ่มสูงขึ้นมากกว่าการใช้กำหนดการพลวัตเพียงอย่างเดียว โดยจากผลการทดสอบค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกัน ด้วยชุดข้อมูลจากฐานข้อมูล HOMSTRAD ซึ่งเป็นชุดข้อมูลที่ใช้ในการทดลองเดิมของภิสิทธิ์ กรรณสูต พบว่าการค้นหาค่ามากที่สุดจากบนลงล่างให้ประสิทธิภาพการค้นหาถูกต้องมากที่สุดที่ 98.93% และเมื่อใช้ชุดข้อมูลจากฐานข้อมูล PIR เพื่อทำการทดสอบกับชุดข้อมูลที่มีข้อมูลที่ต้องการค้นหาสูงขึ้น การค้นหาค่ามากที่สุดจากบนลงล่างยังคงให้ประสิทธิภาพการค้นหาถูกต้องมากที่สุดที่ 98.81% แสดงให้เห็นถึงประสิทธิภาพในการค้นหาที่สูงขึ้นเมื่อเปรียบเทียบกับวิธีของนายภิสิทธิ์ กรรณสูตซึ่งใช้วิธีกำหนดการพลวัตเพียงอย่างเดียว

เนื่องจากการหาค่ามากที่สุดจากบนลงล่าง มีวิธีการแก้ปัญหาการจับกลุ่มกรดอะมิโนที่ไม่ชอบน้ำคล้ายคลึงกับวิธีการคิดของมนุษย์ คือพยายามแก้ปัญหาจากปัญหาใหญ่ไปปัญหาย่อย หรือจับคู่กลุ่มกรดอะมิโนจากภาพรวมก่อน จึงจับคู่กลุ่มกรดอะมิโนในส่วนย่อยๆ จึงเป็นสาเหตุที่ทำให้มีความถูกต้องสูงที่สุด แต่ข้อจำกัดของวิธีนี้เกิดขึ้นเมื่อต้องการค้นหาโปรตีนโดยใช้โปรตีนที่มีสายลำดับของกรดอะมิโนสั้น และ เมื่อแบ่งสายลำดับให้อยู่ในรูปกลุ่มย่อยๆ ของกรดอะมิโนที่ไม่ชอบน้ำแล้วมีจำนวนกลุ่มของกรดอะมิโนที่ไม่ชอบน้ำเป็นจำนวนน้อย จะทำให้การค้นหาผิดพลาด ไม่สามารถค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกันได้อย่างถูกต้อง

วิทยานิพนธ์นี้ได้นำเสนอเครื่องมือค้นหาโปรตีนที่หน้าที่คล้ายคลึงกันด้วยวิธีการจัดเรียงกลุ่มกรดอะมิโนที่ไม่ชอบน้ำแบบค้นหาเฉพาะส่วน ซึ่งมีประสิทธิภาพสูงขึ้นกว่าการใช้กำหนดการพลวัตเพียงอย่างเดียว ทั้งด้านความเร็วและความถูกต้องในการค้นหา ซึ่งความถูกต้องในการค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกันนั้นมีความสำคัญมากและเป็นประโยชน์ในการศึกษาสายลำดับกรดอะมิโนที่ค้นพบขึ้นใหม่ในอนาคต

5.2 ข้อเสนอแนะ

วิทยานิพนธ์นี้ได้นำเสนอเครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกันด้วยวิธีการจัดเรียงกลุ่มกรดอะมิโนที่ไม่ชอบน้ำแบบค้นหาเฉพาะส่วน ซึ่งมีประสิทธิภาพในการค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกัน แต่อย่างไรก็ตาม ยังมีข้อเสนอแนะบางประการที่สามารถพัฒนาเพื่อเพิ่มประสิทธิภาพของวิธีการค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกันให้ดีขึ้น

1. เนื่องจากเวลาที่ใช้ในการค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกัน ขึ้นกับขนาดของฐานข้อมูลที่ทำการค้นหาเป็นหลัก ดังนั้นการพัฒนาตัวกรอง หรือ การทำดัชนีเพื่อใช้ในฐานข้อมูลจะทำให้การค้นหาทำได้เร็วขึ้น
2. การวิเคราะห์กลุ่มกรดอะมิโนที่ไม่ชอบน้ำนั้นสามารถพัฒนานำไปเพื่อสร้างแบบจำลองที่เป็นตัวแทนของกลุ่มโปรตีน ทำให้การค้นหาหน้าที่ของโปรตีนมีประสิทธิภาพมากขึ้นทั้งด้านความถูกต้องและความเร็ว
3. ข้อจำกัดของเครื่องมือค้นหาโปรตีนที่ทำหน้าที่คล้ายคลึงกันด้วยวิธีการจัดเรียงกลุ่มกรดอะมิโนที่ไม่ชอบน้ำแบบค้นหาเฉพาะส่วน เมื่อทำการค้นหาโปรตีนที่มีขนาดเล็กหรือกลุ่มกรดอะมิโนที่ไม่ชอบน้ำเป็นจำนวนน้อย สามารถแก้ไขด้วยตัวกรองที่เหมาะสม เช่น ขนาดของโปรตีนและจำนวนกลุ่มกรดอะมิโนที่ไม่ชอบ หรือนำวิธีอื่นมาประยุกต์ใช้เพิ่มเติมกับสถานการณ์เฉพาะเช่นนี้

รายการอ้างอิง

[1] Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler, GenBank. Nucl. Acids Res 36 (2008): D25-30.

[2] Barker, W. C., J. S. Garavelli, H. Huang, P. B. McGarvey, B. C. Orcutt, G. Y. Srinivasarao, C. Xiao, L.-S. L. Yeh, R. S. Ledley, J. F. Janda, F. Pfeiffer, H.-W. Mewes, A. Tsugita, and C. Wu, The Protein Information Resource (PIR). Nucl. Acids Res 28 (2000): 41-44.

[3] Murzin, A. G., S. E. Brenner, T. Hubbard, and C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247 (1995): 536-540.

[4] Mizuguchi, K., C. M. Deane, T. L. Blundell, and J. P. Overington, Homstrad: a database of protein structure alignments for homologous families. Protein Sci 7 11(November 1998): 2469-2471.

[5] Lemesle-Varloot, L., et al., Hydrophobic cluster analysis: procedures to derive structural and functional information from 2-D-representation of protein sequence. Biochimie 72 (1990): 555-74.

[6] Lemesle-Varloot, L., et al., MANSEK and SUNHCA. Two interactive programs for the hydrophobic cluster analysis of protein sequences. Computer Applications in the Biosciences 9 (1993): 37-44.

[7] Woodcock, S., J.P. Mornon and B. Henrissat, Detection of secondary structure elements in proteins by hydrophobic cluster analysis. Protein engineering 5 (1992): 629-635.

[8] Mechin, M.C., Y. Bertin and J.P. Girardeau, Hydrophobic cluster analysis and secondary structure predictions revealed that major and minor structural subunits of K88-related adhesins of Escherichia coli share a common overall fold and differ structurally from other fimbrial subunits. FEBS Lett 364 (1995): 319-324.

[9] Henrissat, B., Y. Popineau and J.-C. Kader, Hydrophobic-cluster analysis of plant protein sequences. A domain homology between storage and lipid-transfer proteins. Biochem J 255 (1988): 901-905.

[10] Callebaut, I., K. Prat, E. Meurice, J.-P. Mornon and S. Tomavo, Prediction of the general transcription factors associated with RNA polymerase II in Plasmodium falciparum: conserved features and differences relative to other eukaryotes. BMC Genomics 6 (2005): 100.

[11] Girardeau, J.P., Y. Bertin and I. Callebaut, Conserved structural features in class I major fimbrial subunits (Pilin) in gram-negative bacteria. Molecular basis of classification in seven subfamilies and identification of intrasubfamily sequence signature motifs which might be implicated in quaternary structure. J Mol Evol 50 (2000): 424-442.

[12] Callebaut, I., J.C. Courvalin, H.J. Worman and J.-P. Mornon, Hydrophobic cluster analysis reveals a third chromodomain in the Tetrahymena Pdd1p protein of the chromo superfamily. Biochem Biophys Res Commun 235 (1997):103-107.

[13] Phisit Kannasut, Detecting Protein Homology using Automatic 2-D Hydrophobic Cluster Alignment, Master's Thesis, Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, 2008.

[14] Hessa, T., et al., Recognition of transmembrane helices by the endoplasmic reticulam translocon. Nature 433 (2005): 377-381.

[15] National Human Genome Research Institute, Protein structure, from primary to quaternary structure. [Online]. Available from: http://www.genome.gov/Pages/Hyperion/DIR/VIP/Glossary/Illustration/protein.cfm [11 February 2009].

[16] Needlemanand, S.B., and C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48 (1970): 443-453.

[17] Smith, T.F., and M.S. Waterman, Identification of common molecular subsequences. Journal of molecular biology 147 (1981): 195-197.

[18] Henikoff, S., and J. G. Henikoff, Amino acid substation matrices from protein blocks. Proc. Natl. Acad. Sci. USA 89 (1992): 10915-10919.

[19] Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, Basic local alignment search tool. J Mol Biol 215 (1990): 403-410.

[20] Pearson, W. R., and D. J. Lipman, Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85 (1988): 2444-2448.

[21] Karchin, R., and A. Hughey, Weighting Hidden Markov Models for Maximum Discrimination. Bioinformatics 14 (1998): 772-782.

[22] Eddy, S. R., Profile Hidden Markov Models. Bioinformatics 14 (1998): 755-763.

[23] Pearson, W. R., and D. J. Lipman, Comparison of methods for searching protein sequence databases. Protein Sci. 4 (1995): 1145-1160.

ภาคผนวก

ภาคผนวก ก

ฐานข้อมูลโปรตีนที่ใช้ในการทดลอง

ชุดข้อมูลสายลำดับกรดอะมิโนที่นำมาใช้ในการทดลองในวิทยานิพนธ์นี้ถูกเตรียมมาจากฐานข้อมูล 2ฐานข้อมูล คือ ฐานข้อมูล PIR และ ฐานข้อมูล HOMSTRAD โดยรายละเอียดชุดข้อมูลที่เตรียมขึ้นเป็นดังนี้

ก.1 ชุดข้อมูลจากฐานข้อมูล HOMSTRAD

ฐานข้อมูล HOMSTRAD เป็นฐานข้อมูลที่แบ่งโปรตีนตามหน้าที่การทำงานและโครงสร้าง โดยแบ่งออกเป็นกลุ่มย่อยๆ 1032 กลุ่ม โดยโปรตีนแต่ละตัวที่ถูกคัดเลือกมาเป็นโปรตีนที่ผ่านการวิเคราะห์โครงสร้างอย่างถูกต้องแล้วจากฐานข้อมูล PIR โดยรายละเอียดตัวอย่างการแบ่งกลุ่มโปรตีนแสดงไว้ในตารางที่ ก.

ตารางที่ ก.1 ตัวอย่างการแบ่งกลุ่มโปรตีนในฐานข้อมูล HOMSTRAD

Family Name	Length
3,4-dihydroxy-2-butanone 4-phosphate synthase	210
3C cysteine protease (picornain 3C)	191
4\'-phosphopantetheinyl transferase superfamily	109
5\'-3\' exonuclease	265
6,7-dimethyl-8-ribityllumazine synthase	158
6-O-methylguanine DNA methyltransferase	166
6-phosphofructo-2-kinase	213
6-phosphogluconate dehydrogenases	475
7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase (HPPK)	159
7S seed storage protein	353
7kD DNA-binding domain	65
ABC transporter	255
ACT domain	90

ตารางที่ ก.1 ตัวอย่างการแบ่งกลุ่มโปรตีนในฐานข้อมูล HOMSTRAD (ต่อ)

Family Name	Length
ADP-specific Phosphofructokinase/Glucokinase conserved region	450
AMP-binding enzyme	514
AP endonuclease family 1	271
AP endonuclease family 2	271
ARID/BRIGHT DNA binding domain	117
ATP dependent DNA ligase C-terminal domain	90
ATP dependent DNA ligase N-terminal domain	207
ATP synthase	481
ATP synthase, Delta/Epsilon chain, beta-sandwich domain	86
ATP synthase, gamma subunit	241
ATP-guanido phosphotransferase	367
ATP-sulfurylase	541
ATPase family associated with various cellular activities (AAA)	314
Acetyltransferase (GNAT) family	161
Aconitase family (aconitate hydratase)	445
actin-depolymerizing proteins	147
actin/heat-shock cognate	377
Activin types I and II receptor domain	87
Acyl CoA binding protein	87
acyl-CoA dehydrogenase	385
acylphosphatase	98
Adenosylmethionine decarboxylase	309
Adenovirus fiber protein head domain (knob domain)	188
Adenylosuccinate synthetase	430
Adenylyl- / guanylyl cyclase, catalytic domain	189
Agglutinin	149

ตารางที่ ก.1 ตัวอย่างการแบ่งกลุ่มโปรตีนในฐานข้อมูล HOMSTRAD (ต่อ)

Family Name	Length
Alanine dehydrogenase/pyridine nucleotide transhydrogenase	364
Alanine dehydrogenase/pyridine nucleotide transhydrogenase domain 1	193
Alanine dehydrogenase/pyridine nucleotide transhydrogenase domain 2	171
Alanine racemase, N-terminal domain	229
albumin	194
alcohol dehydrogenase	373
Aldehyde dehydrogenase	484
Aldehyde ferredoxin oxidoreductase	608
Aldehyde oxidase and xanthine dehydrogenase, domains 1-2	172
Aldehyde oxidase and xanthine dehydrogenase, domains 3-4	764
aldo/keto reductase	310
Alkaline phosphatase	464
Alpha adaptin AP2, C-terminal domain of C-terminal region	113
Alpha adaptin AP2, C-terminal region (consists of 2 domains)	240
Alpha adaptin AP2, N-terminal domain of C-terminal region	126
Alpha amylase, C-terminal domain	85
Alpha amylase, N-terminal ig-like domain	121
Alpha amylase, catalytic and C-terminal domains	486
Alpha amylase, catalytic domain	401
alpha beta-hydrolase	533
Alpha-2-macroglobulin family A	132
Alpha-2-macroglobulin family B	285
Alphavirus core protein	150
Amidase	450
Amidinotransferase	354
Amino acid kinase family	310

ตารางที่ ก.1 ตัวอย่างการแบ่งกลุ่มโปรตีนในฐานข้อมูล HOMSTRAD (ต่อ)

Family Name	Length
Aminotransferase class IV	286
aminotransferase class-III	431
Aminotransferases class-V	360
Anaphase-promoting complex, subunit 10 (APC10)	170
Anaphylatoxin homologous domain	69
animal haem peroxidase	540
ankyrin repeats	127
annexin	317
antibacterial protein	111
Anticodon binding domain	104
Antifreeze protein	67
Apocytochrome F	250
Apoptosis regulator proteins, Bcl-2 family	166
Arginase family	303
Arginine repressor (ArgR), N-terminal DNA-binding domain	76
arginine repressor, C-terminal domain	71
Arginosuccinate synthase	412
Armadillo/beta-catenin-like repeats	431
Arrestin (or S-antigen)	364
Arrestin (or S-antigen), C-terminal domain	191
Arrestin (or S-antigen), N-terminal domain	173
Arthropod defensin	39
Asp/Glu/Hydontoin racemase	240
Asp/Glu/Hydontoin racemase domain	109
Asparagine synthase	493
Asparagine synthase, C-terminal domain	298
Aspartate carbamoyltransferase regulatory chain	152

ตารางที่ ก.1 ตัวอย่างการแบ่งกลุ่มโปรตีนในฐานข้อมูล HOMSTRAD (ต่อ)

Family Name	Length
Aspartate carbamoyltransferase regulatory chain N-terminal domain	85
Aspartate/ornithine carbamoyltransferase	37
aspartic proteinase	37
Assemblin (Peptidase family S21)	37
Astacin (Peptidase family M12A)	32
avidin	32
azurin/plastocyanin	33
B domain	78
BAG domain	81
BRCA1 C Terminus (BRCT) domain	30
Bacterial DNA recombination protein, RuvA	33
Bacterial DNA-binding protein	33
Bacterial RNA polymerase, alpha chain	47
Bacterial RNA polymerase, alpha chain C-terminal domain	31
bacterial exopeptidases	20
Bacterial extracellular solute-binding protein, family 1	34
bacterial extracellular solute-binding proteins, family 3	43
Bacterial extracellular solute-binding proteins, family 5	25
bacterial lipase	84
Bacterial luciferase	30
Bacterial regulatory helix-turn-helix protein, lysR family	20
Bacterial regulatory helix-turn-helix proteins, araC family, single structural repeat	27
Bacterial regulatory helix-turn-helix proteins, araC family, the two structural repeats	46
Bacterial regulatory proteins, luxR family	32

ตารางที่ ก.1 ตัวอย่างการแบ่งกลุ่มโปรตีนในฐานข้อมูล HOMSTRAD (ต่อ)

Family Name	Length
Bacterial transferase hexapeptide repeats	160
Bacteriochlorophyll A protein	353
Bacterioferritin	158
Bacteriorhodopsin	226
Beta-eliminating lyase	445
beta-lactamase	261
beta/gamma crystallins	174
Biopterin-dependent aromatic amino acid hydroxylase	316
Biotin carboxylase/Carbamoyl phosphate synthetase	240
Biotin-requiring enzymes	83
Bromodomain	120
Bulb-type mannose-specific lectin	108
C-5 cytosine-specific DNA methylase	325
C-terminal domain of Threonine dehydratase	81
C-terminal tandem repeated domains in type 4 procollagen	226
C-type lectin	126
C1q domain	127
CAP-Gly domain	100
CAT RNA binding domain	55
CBS domain	53
CIDE-N domain	104
cadherin	104
calcium-binding protein -- calmodulin-like	159
calcium-binding protein -- parvalbumin-like	107
Calpain family cysteine protease, catalytic domain (domain II)	337

ตารางที่ ก.1 ตัวอย่างการแบ่งกลุ่มโปรตีนในฐานข้อมูล HOMSTRAD (ต่อ)

Family Name	Length
Calponin homology domain	115
Capsid protein (F protein)	421
Carbamoyl-phosphate synthase L chain and Phosphoribosylglycinamide synthetase, N-terminal domain	115
carbohydrate binding module family 12	52
carbohydrate binding module family 3	153
Carbon-nitrogen hydrolase	286
Carboxypeptidase activation peptide	92
Caspase recruitment domain	96
Caspase, interleukin-1 beta converting enzyme (ICE) homologues	243
catalase	566
CcmE	108
Cdc48-like, domains 1 and 2	184
Cellulose binding domain	152
Cellulose binding domain family 2	98
Cereal trypsin/alpha-amylase inhibitor family	114
Chalcone and stilbene synthases	345
Chaperonin 10 kD subunit	98
Chitin binding domain	42
Chitinase class I	242
chorismate binding enzyme	481
Chorismate mutase	218
chromo (CHRomatin Organization MOdifier domain)	71
ciliate pheromone	39
Citrate synthase	394
Class II Aldolase	214

ตารางที่ ก.1 ตัวอย่างการแบ่งกลุ่มโปรตีนในฐานข้อมูล HOMSTRAD (ต่อ)

Family Name	Length
Class II histocompatibility antigen, alpha chain	181
Class II histocompatibility antigen, beta chain	192
Clostridial binary toxin A	204
Clostridial neurotoxin zinc protease	529
Clostridium neurotoxins	438
Clp amino terminal domain	137
CoA binding domain	125
CoA-dependent acetyltransferase	232
CoA-ligase	389
CoA-ligase C-terminal domain	152
Cobalamin (vitamin B12)-binding domain	149
Coenzyme A transferase	267
Cohesin domain	140
cold-shock DNA-binding domain	67
Colicin immunity protein / pyocin immunity protein	85
Colicin pore forming domain	194
Common central domain of tyrosinase	279
complement control protein module (SUSHI) (SCR)	59
Copper amine oxidase	637
Copper resistance protein CopC	102
copper-containing nitrite reductase	334
Cu/Zn superoxide dismutase	152
Cucumovirus coat protein	166
Cutinase	202
Cyclic nucleotide-monophosphate binding domain	123
Cyclin	252

ก.2 ชุดข้อมูลจากฐานข้อมูล PIR

ชุดข้อมูลจากฐานข้อมูล PIR เป็นชุดที่ทำการเลือกสายลำดับกรดอะมิโนมาจากฐานข้อมูล PIR โดยเลือกเฉพาะกลุ่มของโปรตีน จำนวน 67 กลุ่ม ตาม ชุดข้อมูลที่ใช้ในการทดลองของ William R. Person [23] ซึ่งใช้เป็นมาตรฐานในการทดสอบเทคนิคการจับคู่สายลำดับกรดอะมิโนโดยทั่วไป โดยรายละเอียดของกลุ่มโปรตีนได้แสดงในตาราง ก.2

ตารางที่ ก.1 ตัวอย่างการแบ่งกลุ่มโปรตีนในฐานข้อมูล HOMSTRAD

Superfamily	Length
Hemoglobin alpha/beta	141
Ig Kappa chain V-I region	108
G-prot. Coupled receptors	348
Cytochrome C	105
Snake neurotoxin	74
Calcium binding EF-hand	159
Glutathione transferase	222
Protein kinase, cAMP-dependent	351
Ferredoxin	54
Ribulose-bisphosphate carboxylase	139
Ig Kappa chain C region	106
Hemagglutinin	567
Histocompatibility antigen	338
Insulin	110
Alpha-Crystallin chain A	173
Phospholipase A2	148
Glyceraldehyde-3-P-DH	335
Transforming prot. (N-ras)	189
Serine protease	246
Glucagon precursor	180

ตารางที่ ก.1 ตัวอย่างการแบ่งกลุ่มโปรตีนในฐานข้อมูล HOMSTRAD (ต่อ)

Superfamily	Length
H^+-transporting ATP synthase alpha chain precursor	553
Hemagglutinin-neuraminidase	576
Ribonuclease	124
Interferon alpha-I-6	189
Glutamate-ammonia ligase	373
Azurin	129
Fusion protein—Sendai virus	565
Cytochrome P450	497
Outer capsid protein VP8	280
Gag polyprotein	512
Keratin	471
Nucleoprotein-influenza A	498
Acidic ribosomal protein P2	115
E6 protein papillomavirus	158
Lysozyme	130
N-Cadherin	906
Exo-alpha-sialidase	454
L2 protein papillomavirus	507
Scorpion neurotoxin	64
E7 protein papillomavirus	98
H^+-transporting ATP synthase lipid-binding	75
L-Lactate dehydrogenase	333
E2 protein papillomavirus	322
Core antigen—hepatitis B	183
Antithrombin-III	464
Thymidine kinase	376
Phycocyanin	162

ตารางที่ ก.1 ตัวอย่างการแบ่งกลุ่มโปรตีนในฐานข้อมูล HOMSTRAD (ต่อ)

Superfamily	Length
Protamine Y2	34
Transforming prot. (myc)	439
Matrix protein	348
H^+-transporting ATP synthase P6	226
Alcohol dehydrogenase A	375
Glycoprotein B	857
Ionotropic acetylcholine receptor	457
Non-structural protein NS2	121
Annexin I	346
Histone H1b	218
Metallothionein	61
Beta-Crystallin chain Bp	204
Proteinase inhibitor	71
Hepatic lectin H1	291
E2 glycoprotein precursor	1447
Alpha-2u-Globulin precursor	181
Pepsin	388
DNA-directed DNA polymerase	1462
Prolactin	227
Vitamin B12 trans. btuD	249

ภาคผนวก ข

ผลงานตีพิมพ์

งานประชุมวิชาการ “The 6th International Joint Conferrence on Computer Science and Software Engineering (JCSSE2009)” ซึ่งจัดขึ้น ณ Laguna Resort จังหวัดภูเก็ต ประเทศไทย ในระหว่างวันที่ 13-15 พฤษภาคม 2552 ในหัวข้อเรื่อง “An Algorithm to Compute Protein Homology Based On Hydrophobic Cluster Analysis” โดย ชนินท์ จันมา , ผู้ช่วยศาตราจารย์ ดร. รัฐ พิชญางกูร, ผู้ช่วยศาสตราจารย์ ดร. โชติรัตน์ รัตนามหัทธนะ และศาตราจารย์ ดร. ประภาส จงสถิตย์วัฒนา

An Algorithm to Compute Protein Homology Based On Hydrophobic Cluster Analysis

Chanin Chanma[1], Rath Pichyangkura[2],
Chotirat Ann Ratanamahatana[1], Prabhas Chongstitvatana[1]
[1]Department of Computer Engineering
[2]Department of Biochemistry
Chulalongkorn University
Phayathai Rd., Pathumwan, Bangkok, Thailand, 10330
chanin.c@student.chula.ac.th, {prath, chotirat.r, prabhas}@chula.ac.th

Abstract

Current techniques in protein homology testing involve a 1-dimensional alignment of Nucleotide or Amino acid sequencing. Due to its various constraints and low sequence identity values, a 2-Dimensional Hydrophobic Cluster Alignment has increasingly been used to predict the structure and functionality of protein. This work proposed an algorithm based on a secondary-structure Hydrophobic Cluster Alignment to compute a similarity score of protein sequences automatically, which helps reduce interventions of a human expert for a manual alignment. Additional techniques are introduced to speed up the calculation, as well as to resolve some greedy-based alignment limitation in the previous work. The alignment results and the classification accuracies from the well-known HOMSTRAD database have demonstrated an improvement in both accuracy and the computation time.

Key Words: bioinformatic; hydrophobic cluster analysis; protien homology; automatic alignment

1. Introduction

As biological data have grown tremendously in the past decade, they provide an avenue of researches in analyzing and extracting useful knowledge that give us better understanding of the rules of nature. Genome projects are parts of the largest resources of life science data, which mainly include nucleotide and amino acid sequences. However, effective retrieval of these data is still a great challenge. More specifically, we need a high-quality tool to determine protein homology via sequence alignment. Detection of protein homology has become a large research field in bioinformatics. Several crucial analyzed protein databases, such as UniProt [1], PDB [2], SCOP [3], and PFam [4] have been created. These databases contain useful knowledge, e.g., protein homology, structure, or functionalities. By detecting protein similarity, the newly discovered protein sequences can be used to predict their functionalities from the known information in the database.

Early methods, such as Maximum Matching, Basic Alignment Search Tool (BLAST) [5], and FASTA [6], measure protein homology from protein's primary structure information. These methods still have major limitations and drawbacks; they are unable to provide a proof of sequence homology if the sequence identity appears to be too low, a situation that typically occurs in proteins of the same functionality, but belong to organism from different species.

The similarity can generally be measured from an alignment of either nucleotide sequences or amino acid sequences. However, nucleotide sequence similarity is not suitable for protein function discovery; amino acid sequences are typically exploited instead since they contain much more information, such as hydrophobic and hydrophilic properties. Unfortunately, the current one-dimensional alignment tools mentioned earlier all have some limitation that yields poor alignment results. Therefore, higher level structures (Secondary (2-dimensional), Tertiary (3-dimensional), and Quaternary (4-dimensional)) have been increasingly put into consideration. Functionality of a protein is generally based on its 3-Dimensional structure. Some researchers have attempted to predict this protein structure by amino sequence folding based on each amino acid property, but this approach turns out to be unfeasible in practice that extremely high computational power is needed.

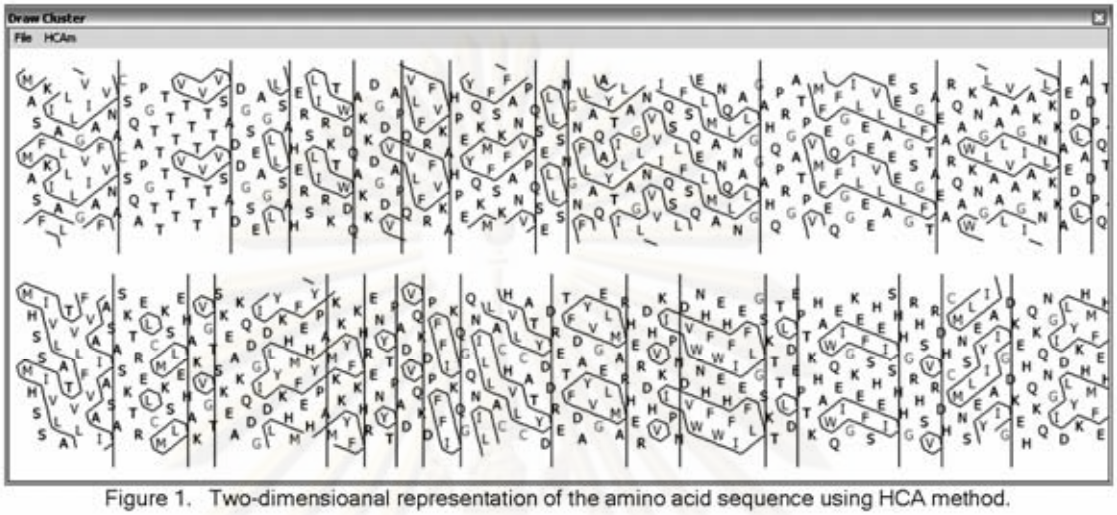


Figure 1. Two-dimensioanal representation of the amino acid sequence using HCA method.

Instead, our proposed method is based on the idea of a 2-Dimensional structure, Hydrophobic Cluster Analysis (HCA). Hydrophobicity property is the key in protein folding. The reason is that as protein creating and folding occur in water, they try to preserve the structure by compacting and turning their hydrophobic part inside and resting the hydrophilic part outside for an easy access to water. Hydrophobic Cluster Analysis (HCA) approach has been developed from this belief. HCA approach visualizes amino acid sequence as a 2-D helical pattern. In a Hydrophobic Cluster Analysis, amino acid sequences are laid into a 2-D helical pattern by twisting the protein into a smoothed helix, where each twist will contain 3.6 amino acids [7]. Then, this cylinder is cut lengthwise and spread into a 2-Dimensional plane, and the hydrophobic amino acid will be highlighted and grouped together [8], as shown in Fig. 1. This representation is then used in protein alignment. However the actual hydrophobic cluster alignment requires a human expert to perform the alignment manually.

An earlier work of Automatic 2-D Hydrophobic Cluster Alignment [9] introduced the new representation of amino acid sequences and used dynamic programming approach to measure their similarity. The homologous proteins with the same functionalities will have a high sequence identity score and others with different functionalities will have a low sequence score.

In more detail, the algorithm in [9] is based on a dynamic programming approach. Every cluster block from the test sequence will be compared with all cluster blocks from another sequence to discover the best alignment and score. First, a matrix as large as the number of cluster blocks in each sequence is created. Then each cell in the matrix is filled accordingly, starting from the first pair to the last pair. To determine a cumulative value in each current cell, the maximum score of the three neighboring cells (Top, Left, and Diagonal) is added to the current cell's alignment score. This cell's alignment score reflects the best alignment score by shifting residues one-by-one from left to right and updating the remaining residues in the matrix that will be used later. After the score of the last pair is calculated, the best score of the alignment is obtained. The actual alignment can be constructed by tracing the path back to the first alignment pair. Even though this approach achieves good accuracy improvement over the one-dimensional alignment approaches, its alignment is still not optimal and its computational complexity can be much improved.

Therefore, this work proposes an extension of the previous technique in [9]. We revise the representation of the amino acid sequence and improve the dynamic programming algorithm. The goal is to offer a better accuracy alignment score and improve the efficiency of the computation.

2. Proposed Method

2.1 New representation

We improve the representation proposed in [9], whose amino acids are simply transformed into binary symbols; "1" represents hydrophobic amino acid and "0" represent hydrophilic amino acid. As a result, the information regarding different hydrophobic amino acids is lost. Instead, our representation replaces only hydrophilic amino acid into "0" since it has no role during the alignment, and keeps the original representation of all 7 types of the hydrophobic amino acids (Valine (V), Isoleucine (I), Leucine (L), Phenylalanine (F), Tryptophan (W),

Methionine (M), and Tyrosine (Y)). Preserving this hydrophobic information enables our approach to achieve a more accurate score since we can assign different scoring to different types of hydrophobic amino acids. Our substitution matrix is extended from BLOSUM62 [10]. Specifically, we use only the scores of hydrophobic amino acid pairs and the hydrophilic amino acid is substituted with the average score of the hydrophobic amino acids, as shown in Fig. 2.

	M	I	L	V	F	Y	W	Other
M	**5**	1	2	1	0	-1	-1	-2
I		**4**	2	3	0	-1	-3	-3
L			**4**	1	0	-1	-2	-2
V				**4**	-1	-1	-3	-2
F					**6**	3	1	-2
Y						**7**	2	-2
W							**11**	-3

Figure 2. The substitution matrix for hydrophobic cluster alignment.

To extract an amino acid sequence into an individual hydrophobic cluster, we follow the previously proposed method whose streak of five "0"s or more in the sequence, as well as the Proline amino acid symbol ("P"), indicate the end of the cluster. Now our amino acid sequences are transformed into cluster blocks and are ready to be used in our alignment algorithm. Fig. 3 shows our new representation.

(a) MASFKIALLLGVIAFVNACSQAPGTTTTTVTTTVTTVSADDGSEAGLLS

(b) M00F0I0LLL0VI0FV0000000000000V000V00V000000000LL0

(c) M00F0I0LLL0VI0FV V000V00V LL

Figure 3. Our proposed representation. (a) Original sequence. (b) Hydrophilic replaced. (c) Break to cluster block.

In order to recognize, identify, and understand features of cluster blocks easily, we use the visualization tool in [9]. An example visualization of the 2-D clusters in our representation is shown in Fig. 4.

Example	2-D Cluster	Representation
1		M00F0I0LLL0VI0FV
2		V000V00V

Figure 4. A Visualization corresponded to our representation

To calculate the alignment score of two cluster blocks, we search the alignment position that gives the highest score by shifting residues one-by-one from left to right. Current position alignment score is the summation of all aligned amino acid pairs score by using our substitution matrix in Fig. 1. An example is shown in Fig. 5. Cluster1 and Cluster2 are used in alignment and the result maximum score is 4.

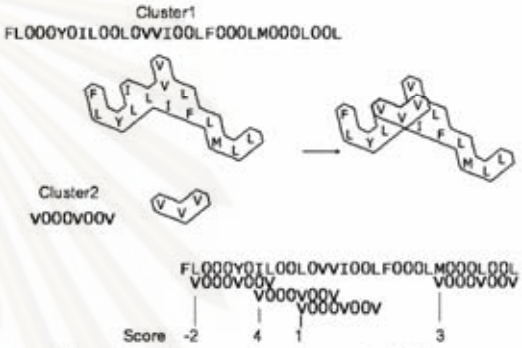


Figure 5. An example of our calculation alingment score.

2.2 Cluster Alignment Algorithm

In this work, our cluster alignment algorithm also uses a dynamic programming approach and is based on the 2-dimensional string matching technique similar to that in [9]. However, we include additional techniques as follows to increase accuracy and to reduce computation cost.

2.2.1 Score Dictionary Lookup

The computation time for an alignment of amino sequences or cluster blocks is spent mostly in score calculations of the sequence pairs. Generally, the scores will be calculated between the new query sequence and each of the sequences in a very large database. Calculating the score of the same sequence repeatedly clearly wastes a lot of computational resources. So, we introduce the score dictionary lookup technique to retrieve the already existed intermediate scores.

After we calculate the score of each sequence pair, we add all possible subsequences and their corresponding scores into a dictionary. Next time a pair of the sequences with the same substring in the dictionary, we simply lookup without the need of the score's recalculation. Fig. 6 shows an example of some possible subsequence added to the dictionary. After we calculate alignment score between "M00F0" and "LL0VI" cluster blocks. All other possible subsequences score we get will add to dictionary.

a)	(M00F0,LL0VI = -4)
b)	(M,L = 2) , (0,L = -2) , (0,0 = 0) , (F,V = -1) , (0,I = -3) , (M0,LL = 0) , (00,L0 = -2) , (0F,0V = -1) , (F0,VI = -4), (M00,LL0 = 0) , (00F,L0V = -3) , (0F0,0VI = -4) , (M00F,LL0V = -1) , (00F0,L0VI = -6) , *(M00F0,LL0VI = -4)*

Figure 6. Examples of some possible subsequences created and added to a dictionary a) Query Sequence. b) Possible subsequences add to the dictionary.

2.2.2 Local Window Search

The approach in [9] scores all pairs of cluster blocks in each sequence and finds the best score and its path from the entire table, which evidently requires extremely high computation complexity. Moreover their approach uses a greedy choice in the search, where each score depends on the remaining residue from the previous cluster used. This leads to confusion when each protein sequence has similar cluster blocks but their positions in the sequences are very different. The subsequent cluster will get a mismatch score caused by wrong remaining residue.

Therefore, our work introduces a Local window search to reduce bias from a greedy choice and to create more efficient remaining residue. Before we perform a dynamic programming step, we find a pair of clusters that obtains the maximum score in the window. Then we do a dynamic programming from the starting point to this maximum point in the window. After the dynamic programming step is finished, we move the starting point to the last maximum point. Fig. 7 illustrates an example of our local window search space. A table of m by n is constructed, where m is the number of cluster blocks in Clust1 and n is the number of cluster blocks in Clust2 that we break from Representation Step.

In a dynamic programming procedure, we start filling in the table from starting point to the maximum point in the window. Each cell in the table is a cumulative value of the maximum score from the three neighboring cells (Top, Left, and Diagonal) and the current cell score (calculate similarly to the example in Fig. 5). An example is shown in Fig. 8.

As a result, this local window search greatly improves the quality of the scores and reduces the computation time, as will be demonstrated in the experiment section.

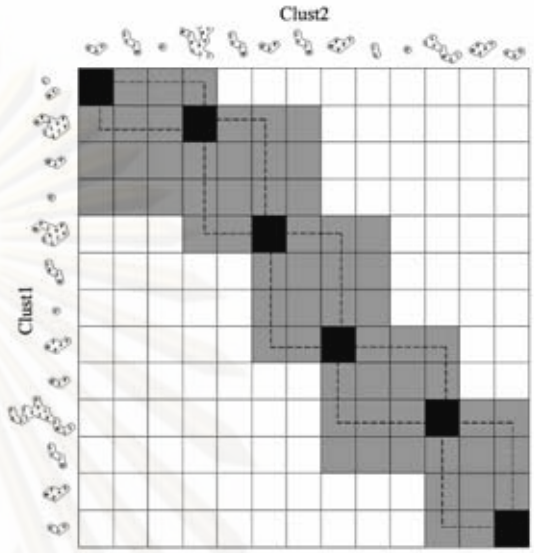


Figure 7. Local search space in the table. The gray areas denote local windows. The black areas denote the maximum score node in each local window. The starting point of a new local window is moved to the previous maximum score node. The dashed rectangles are the areas where we perform dynamic programming.

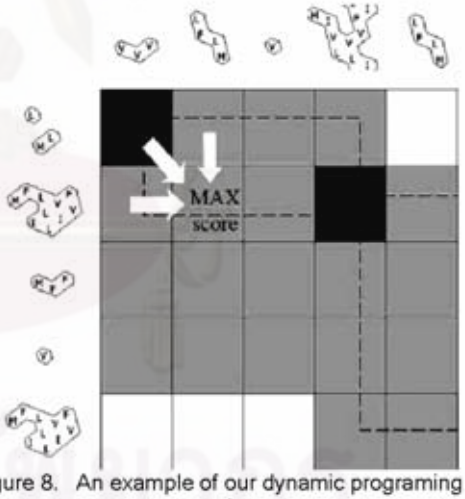


Figure 8. An example of our dynamic programing procedure.

3. Experiments and Results

We test our approach on HOMSTRAD [11] database, which contains more than 300,000 annotated protein sequences. This database classifies their protein sequences into 1032 families based on their structure alignments. We separate this database into training and test set.

In our experiments, test sets include one random sequence from each of the protein family, resulting in a total of 1032 sequences. The training set also randomly select and each protein family must be selected at least one sequence. Sizes of the training set are varied from 3448 sequences to the entire HOMSTRAD database of 336827 sequences. Note that both training and test data are distinct, where no sequences in the test set overlaps with those in the training set.

To validate our proposed method for each of the sequence in the test set, we search the training set to find the most similar sequence based on our alignment scores. If the best sequence we discover is in the same family as the test sequence, we denote it as a correct answer. We compare our classification accuracy with the previous method [9]; results are shown in Table 1.

Table 1. Experimnetal result

Training size	Testing size	Method	Accuracy
3448	1032	**Our aprroach** Previous approach[9]	93.50% 91.47%
44813	1032	**Our aprroach** Previous approach[9]	95.05% 94.38%
86678	1032	**Our aprroach** Previous approach[9]	95.93% 95.15%
187991	1032	**Our aprroach** Previous approach[9]	97.18% 96.32%
336827	1032	**Our aprroach** Previous approach[9]	98.64% 97.09%

From the results, our proposed method reports higher classification accuracy based on protein's secondary structure and its homology. Our running time is also significantly improved. However, we decide not to report the raw running time here since it would be unfair to the previous approach as they are implemented under different platforms. However, a simple analysis of our algorithm theoretically confirms that our newly proposed approach will reduce the time complexity by a large margin, especially in massive databases.

4. Conclusion

In this work, we propose an improved automatic 2-Dimensional Hydrophobic Cluster Alignment algorithm that yields higher classification accuracy with reduced time complexity. As a result, our proposed technique could facilitate and unveil a new opportunity of research in protein's homology, its functionality, and numerous Bioinformatics applications.

5. References

[1] R. Leinonen, F. G. Diez, D. Binns, W. Fleischmann, R. Lopez, and R. Apweiler, "UniProt archive," Bioinformatics, vol. 20, pp. 3236-3237, 2004.

[2] H. M. Berman, et al., "The Protein Data Bank, " Nucl. Acids Res., vol. 28, pp. 235-242, 2000.

[3] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," J Mol Biol, vol. 247, pp. 536-540, 1995.

[4] A. Bateman, et al., "The Pfam protein families database," Nucl. Acids Res., vol. 32, pp. D138-141, 2004.

[5] S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, "Basic local alignment search tool," J Mol Biol, vol. 215, pp. 403-410, 1990.

[6] W. R. Pearson and D. J. Lipman, "Improved tools for biological sequence comparison," Proc Natl Acad Sci U S A, vol. 85, pp. 2444-2448, 1988.

[7] Christine Gaboriaud, et al., "Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences," FEBS Lett, vol. 224, no. 1, pp. 149-55, 1987.

[8] L. Lemesle-Varloot, et al., "Hydrophobic cluster analysis: procedures to derive structural and functional information from 2-D-representation of protein sequence," Biochimie, vol. 72, pp. 555-74, 1990.

[9] P. Kannasut, R. Pichyangkura and C. A. Ratanamahatana, "Automatic 2-D Hydrophobic Cluster Alignment," Int. J. Biomedical Engineering and Technology., inpress.

[10] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks." Proc Natl Acad Sci U S A, vol. 89, no. 22, pp. 10 915-10 919, November 1992.

[11] K. Mizuguchi, C. M. Deane, T. L. Blundell, and J. P. Overington, "Homstrad: a database of protein structure alignments for homologous families." Protein Sci, vol. 7, no. 11, pp. 2469-2471, November 1998.

ประวัติผู้เขียนวิทยานิพนธ์

นายชนินท์ จันมา เกิดเมื่อวันที่ 14 เมษายน พ.ศ. 2528 ที่จังหวัดกรุงเทพมหานคร สำเร็จการศึกษาระดับปริญญาวิศวกรรมศาสตรบัณฑิต (วศ.บ.) จากคณะวิศวกรรมศาสตร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2549 และได้เข้าศึกษาต่อในหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต (วศ.ม.) สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2550