ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

วิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)

ปริญญา

วิศวกรรมคอมพิวเตอร์	วิศวกรรมคอมพิวเตอร์
สาขา	**ภาควิชา**

เรื่อง ขั้นตอนวิธีการเทียบเรียงกลุ่มลำดับข้อมูลชีวภาพโดยพิจารณาความถี่ส่วนย่อยของลำดับ

A Sub-Segment Frequency-Based Multiple Alignment Algorithm for Biological Sequences

นามผู้วิจัย นายคมสัน จันมา

ได้พิจารณาเห็นชอบโดย

ประธานกรรมการ .....

(ผู้ช่วยศาสตราจารย์พันธุ์ปิติ เปี่ยมสง่า, D.Sc.)

กรรมการ .....

(ผู้ช่วยศาสตราจารย์พีรวัฒน์ วัฒนพงศ์, Ph.D.)

กรรมการ .....

(ผู้ช่วยศาสตราจารย์เขมะฑัต วิภาตะวนิช, Ph.D.)

หัวหน้าภาควิชา .....

(ผู้ช่วยศาสตราจารย์เขมะฑัต วิภาตะวนิช, Ph.D.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

.....

(รองศาสตราจารย์วินัย อาจคงหาญ, M.A.)

คณบดีบัณฑิตวิทยาลัย

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

วิทยานิพนธ์

เรื่อง

ขั้นตอนวิธีการเทียบเรียงกลุ่มลำดับข้อมูลชีวภาพโดยพิจารณาความถี่ส่วนย่อยของลำดับ

A Sub-Segment Frequency-Based Multiple Alignment Algorithm for Biological Sequences

โดย

นายคมสัน จันมา

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

เพื่อความสมบูรณ์แห่งปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)

พ.ศ. 2550

คมสัน จันมา 2550: ขั้นตอนวิธีการเทียบเรียงกลุ่มลำดับข้อมูลชีวภาพโดยพิจารณาความถี่ส่วนย่อยของลำดับ ปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์) สาขาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ ประธานกรรมการที่ปรึกษา: ผู้ช่วยศาสตราจารย์พันธุ์ปิติ เปี่ยมสง่า, D.Sc. 100 หน้า

การเทียบเรียงกลุ่มลำดับข้อมูลทางชีวภาพเป็นงานที่จำเป็นสำหรับนักชีววิทยา ซึ่งในปัจจุบันจำนวนของข้อมูลทางด้านชีวภาพได้เพิ่มขึ้นอย่างรวดเร็ว ดังนั้นเครื่องมือสำหรับการเทียบเรียงกลุ่มลำดับข้อมูลทางชีวภาพจึงจำเป็นต้องมีความถูกต้องในระดับที่นักชีววิทยาพอใจ และใช้เวลาในการประมวลผลน้อย งานวิจัยนี้จึงนำเสนอวิธีการเทียบเรียงกลุ่มลำดับข้อมูลทางชีวภาพที่ใช้เวลาในการทำงานน้อยมาก โดยที่ความถูกต้องเทียบเคียงได้กับเครื่องมือที่ใช้ในปัจจุบัน โดยใช้วิธี n-gram ตัดข้อมูลเป็นส่วนย่อย และพิจารณาจากความถี่ของส่วนย่อยของแต่ละคู่ลำดับข้อมูลเพื่อใช้เป็นแนวทางในการเทียบเรียง วัดความถูกต้อง และเวลาการประมวลผลโดยใช้ชุดข้อมูลทดสอบ BAliBASE และ PREFAB ผลลัพธ์ที่ได้จากการเทียบเรียงชุดทดสอบ PREFAB สามารถเทียบเรียงได้เร็วกว่า ClustalW 6.24 เท่า โดยที่ความถูกต้องอยู่ในระดับเดียวกัน

_____ _____ ___ / ___ / ___

ลายมือชื่อนิสิต ลายมือชื่อประธานกรรมการ

Comsan Chanma 2007: A Sub-Segment Frequency-Based Multiple Alignment Algorithm for Biological Sequences. Master of Engineering (Computer Engineering), Major Field: Computer Engineering, Department of Computer Engineering. Thesis Advisor: Assistant Professor Panpiti Piamsa-nga, D.Sc. 100 pages.

Multiple sequence alignment has become a necessary task of biologists for sequence analysis. Because of fast increasing of biology data, biologists not only need accurate tools but fast tools are also required for multiple sequence alignment of biology data. In this paper, we propose a sub-segment's frequency-based processing technique to improve speed of alignment process. This technique is pre-process a sub-segment's frequency matrix between all pair of sequences with n-gram technique. This provides us with a distance matrix that can be used to guide the progressive alignment. Accuracy and computing time are measured using BAliBASE and PREFAB benchmark. The CPU time is drastically reduced as compared with ClustalW. On the PREFAB benchmark alignment databases, our result is 6.24 times faster than ClustalW with comparable accuracy.

_____ _____ ___ / ___ / ___

Student's signature Thesis Advisor's signature

กิตติกรรมประกาศ

ข้าพเจ้าขอขอบพระคุณ ผู้ช่วยศาสตราจารย์พันธุ์ปิติ เปี่ยมสง่า ประธานกรรมการที่ปรึกษาที่ให้คำแนะนำและให้ความช่วยเหลือในการทำวิทยานิพนธ์ฉบับนี้เป็นอย่างสูง ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์พีรวัฒน์ วัฒนพงศ์ กรรมการสาขาวิชาเอก ผู้ช่วยศาสตราจารย์เขมะฑัต วิภาตะวนิช กรรมการสาขาวิชารอง และผู้ที่เกี่ยวข้องกับวิทยานิพนธ์ฉบับนี้ที่ให้คำแนะนำเกี่ยวกับการนำเสนองานวิจัย และช่วยตรวจสอบแก้ไขข้อบกพร่องต่าง ๆ ของวิทยานิพนธ์ฉบับนี้

ขอขอบคุณคุณจิติมนต์ เขียนดวงจันทร์ และคุณวีรวุฒิ คงบุญเกียรติ ที่ให้แนวคิดและแนวทางในการทำงานวิจัย และขอบคุณสมาชิกในห้องปฏิบัติการวิจัยการวิเคราะห์และค้นพบเนื้อหาประสมที่เอื้อเฟื้อสถาที่ในการทำวิทยานิพนธ์ และให้กำลังใจในการทำวิทยานิพนธ์เสมอมา

สุดท้ายนี้ขอขอบคุณคุณพ่อ คุณแม่ และทุกๆคนในครอบครัวที่คอยเป็นกำลังใจที่ดีที่สุด ขอบคุณรุ่นพี่ รุ่นน้อง และเพื่อนๆทั้งที่วิศวกรรมศาตร์ มหาวิทยาลัยเกษตรศาสตร์ และที่บริษัท ที. เอ็น. อินฟอร์เมชั่น ซิสเท็มส์ จำกัด ที่ให้ความช่วยเหลือ และเป็นกำลังใจพลักดันให้วิทยานิพนธ์ฉบับนี้เสร็จสมบูรณ์ได้

คมสัน จันมา

มีนาคม 2550

สารบัญ

หน้า

สารบัญตาราง

สารบัญภาพ

สารบัญภาพ (ต่อ)

ภาพที่ **หน้า**

สารบัญภาพ (ต่อ)

ภาพที่ **หน้า**

ภาพผนวกที่

สารบัญภาพ (ต่อ)

สารบัญภาพ (ต่อ)

สารบัญภาพ (ต่อ)

สารบัญภาพ (ต่อ)

สารบัญภาพ (ต่อ)

ขั้นตอนวิธีการเทียบเรียงกลุ่มลำดับข้อมูลชีวภาพโดยพิจารณาความถี่ส่วนย่อยของลำดับ

A Sub-Segment Frequency-Based Multiple Alignment Algorithm for Biological Sequences

คำนำ

ในปัจจุบันการนำศาสตร์ทางด้านคอมพิวเตอร์เข้ามาช่วยงานทางด้านชีวภาพนั้นแพร่หลายมากขึ้นเนื่องจากมีการทำงานที่รวดเร็ว และถูกต้อง ซึ่งประโยชน์ที่ได้ก่อให้เกิดความก้าวหน้าและการค้นพบสิ่งใหม่ในทางด้านชีววิทยาเป็นอันมาก การเทียบเรียงกลุ่มลำดับข้อมูลเป็นงานวิจัยด้านหนึ่งที่ได้รับความสนใจ และมีความสำคัญ เนื่องจากเป็นขั้นตอนที่จำเป็นสำหรับการวิเคราะห์ลำดับข้อมูลทางด้านชีววิทยา ซึ่งมีจำนวนข้อมูลมาก และเพิ่มขึ้นอย่างรวดเร็ว ผลที่ได้จากการเทียบเรียงกลุ่มลำดับข้อมูลทางชีวภาพนั้นถูกใช้ประโยชน์ในหลายๆด้าน ได้แก่ การวิเคราะห์วิวัฒนาการของสิ่งมีชีวิต, การระบุส่วนจำเพาะของลำดับข้อมูล หรือแม้กระทั่งใช้ในการทำนายโครงสร้างของลำดับโปรตีนนั้นๆ วิธีการและขั้นตอนวิธีทางด้านการโปรแกรมคอมพิวเตอร์ต่างๆ ถูกนำมาประยุกต์ใช้ในงานวิจัยเกี่ยวกับการเทียบเรียงกลุ่มลำดับข้อมูลทางชีวภาพ เพื่อพัฒนา 2 ปัจจัยหลัก คือความถูกต้องของการเทียบเรียง และความเร็วในการประมวลผล

การเทียบเรียงกลุ่มลำดับข้อมูลทางชีวภาพนั้น มีความยากระดับ NP complete การเทียบเรียงนั้นเป็นการเทียบเรียงลำดับที่มีความยาวสูง จำนวนมาก เป็นผลให้การประมวลผลใช้เวลานาน โดยที่ปกติค่าความซับซ้อนของการประมวลผลคือ O(NL) เมื่อ L คือ ความยาวของแต่ละสายอักขระ และ N คือ จำนวนสายอักขระ ดังนั้นแนวทางการพัฒนาในปัจจุบันจึงมุ่งเน้นให้การเทียบเรียงใช้มีความเร็วสูงขึ้น ในขณะที่ความถูกต้องอยู่ในระดับที่ยอมรับได้

วิทยานิพนธ์นี้นำเสนอการเทียบเรียงกลุ่มลำดับข้อมูลชีวภาพในแบบวิธีฮิวริสติก โดยพิจารณาความถี่ของลำดับย่อยร่วมของลำดับข้อมูล เพื่อพัฒนาขั้นตอนวิธีการเทียบเรียงกลุ่มลำดับข้อมูลทางชีวภาพให้มีประสิทธิภาพดีขึ้น รองรับข้อมูลจำนวนมากได้ โดยเน้นพัฒนาด้านความเร็ว เปรียบเทียบกับโปรแกรม ClustalW

วัตถุประสงค์

1. ศึกษา และเปรียบเทียบขั้นตอนวิธีการเทียบเรียงกลุ่มลำดับข้อมูลชีวภาพเพื่อเพิ่มประสิทธิภาพของการเทียบเรียง

2. พัฒนาขั้นตอนวิธีการเทียบเรียงกลุ่มลำดับข้อมูลชีวภาพ โดยใช้เทคนิคการพิจารณาความถี่ส่วนย่อยของลำดับเข้าช่วย เพื่อให้การเทียบเรียงกลุ่มลำดับข้อมูลมีความเร็วเพิ่มขึ้น และสามารถรองรับการเทียบเรียงกลุ่มลำดับข้อมูลขนาดใหญ่ได้ โดยคำนึงถึงความถูกต้องของการเทียบเรียงด้วย

การตรวจเอกสาร

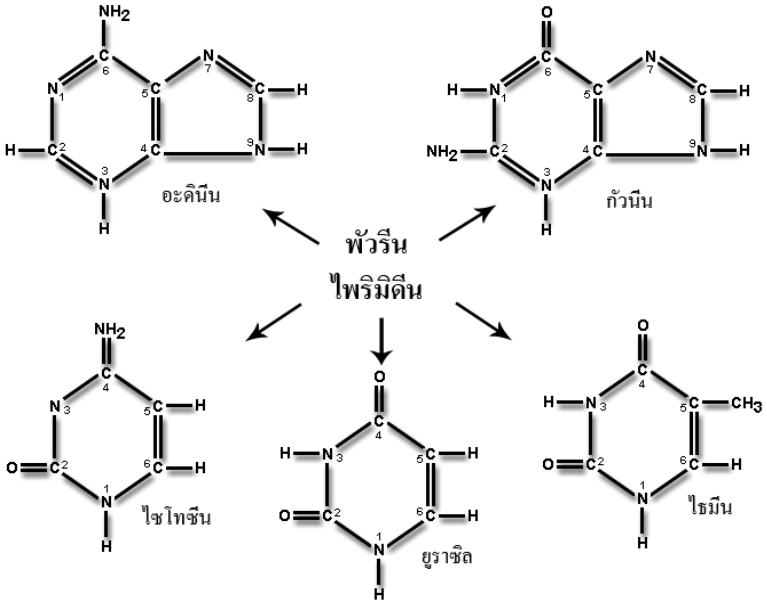
ความรู้เบื้องต้นเกี่ยวกับข้อมูลชีวภาพ

ในส่วนนี้กล่าวถึงความรู้พื้นฐานของข้อมูลชีวภาพที่ใช้ในการเทียบเรียงกลุ่มลำดับข้อมูล ข้อมูลชีวภาพที่ใช้ในการเทียบเรียงแบ่งได้เป็น 2 กลุ่ม คือ ข้อมูลกรดนิวคลีอิก ได้แก่ ดีเอ็นเอ และอาร์เอ็นเอ และข้อมูลโปรตีน

1. กรดนิวคลีอิก

กรดนิวคลีอิก(Nucleic acid) (ไพศาล, 2539; อมรา, 2540; พจน์ และคณะ, 2543) พบอยู่ในนิวเคลียส(Nucleus), ไซโทพลาซึม(Cytoplasm), และในออร์แกเนลล์(organelle) บางชนิดของเซลล์ สามารถแบ่งได้เป็น 2 ชนิดด้วยกันคือ ดีเอ็นเอ(Deoxyribonucleic acid: DNA) และอาร์เอ็นเอ (Ribonucleic acid: RNA) กรดนิวคลีอิกแต่ละชนิดประกอบด้วยเส้นสายที่เกิดจากการเชื่อมต่อของนิวคลีโอไทด์(nucleotide) หลายชนิด ในแต่ละนิวคลีโอไทด์ยังประกอบด้วยเบสที่มีไนโตรเจน (nitrogenous base) น้ำตาลที่มีคาร์บอน 5 อะตอม(pentose) และกรดฟอสฟอริก(phosphoric acid)

สารประกอบเบสที่มีไนโตรเจนที่พบในนิวคลีโอไทด์ เป็นสารที่มีวงแหวนคาร์บอน-ไนโตรเจนจำนวน 1 หรือ 2 วง เบสในนิวคลีโอไทด์แบ่งออกได้เป็น 2 กลุ่มตามจำนวนของวง-แหวน คือ เบสที่มีวงแหวนเดียว หรือไพริมิดีน(Pyrimidine) มีอยู่ 3 ชนิด ได้แก่ ไธมีน(Thymine: T), ไซโทซีน(Cytosine: C), และยูราซิล(Uracil: U) ส่วนเบสที่มี 2 วงแหวน หรือพัวรีน(Purine) มีอยู่ 2 ชนิด ได้แก่ อะดินีน(Adenine: A) และกัวนีน(Guanine: G) เบสในนิวคลีโอไทด์มีโครงสร้างโมเลกุลดังภาพที่ 1

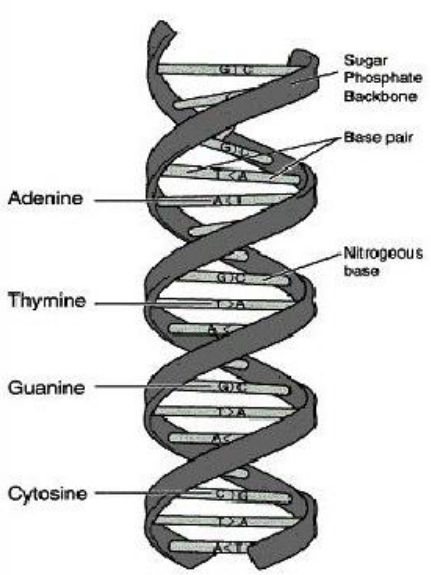


ภาพที่ 1 โครงสร้างโมเลกุลของเบสอันเป็นองค์ประกอบของนิวคลีโอไทด์

1.1 ดีเอ็นเอ เป็นสารชีวโมเลกุลที่มีขนาดใหญ่มาก เกิดจากการเรียงต่อกันของนิวคลีโอไทด์ บางโมเลกุลของดีเอ็นเออาจมีความยาวถึง 200,000 นิวคลีโอไทด์ ในเซลล์ของมนุษย์สามารถพบดีเอ็นเอได้ทั้งในนิวเคลียส และในไมโทคอนเดรีย สำหรับในพืชนอกจากจะพบดีเอ็นเอในนิวเคลียสแล้ว ยังพบดีเอ็นเอได้อีกในคลอโรพลาสต์อีกด้วย ดีเอ็นเอทำหน้าที่เป็นสารพันธุกรรม(Genetic Material) โดยจัดเรียงตามข้อกำหนดจำเพาะ เปรียบเสมือนเป็นแบบพิมพ์ชีวิต

โมเลกุลของดีเอ็นเอประกอบด้วยนิวคลีโอไทด์ 4 ชนิด ได้แก่ อะดินีน, กัวนีน, ไธมีน, และไซโทซีน โมเลกุลของดีเอ็นเอนั้นอยู่ในสภาพที่เป็นเกลียวคู่ขนาน(double helix) คล้ายกับบันไดเวียนดังภาพที่ 2 โดยมีน้ำตาลและกรดฟอสฟอริกเป็นราวบันไดทั้งสองข้าง พันธะไฮโดรเจนซึ่งยึดระหว่างพัวรีน และไพริมิดีนเป็นขั้นบันได โดยที่บันไดดังกล่าวหมุนเป็นเกลียวอย่างมีระเบียบและลำดับการเรียงตัวของเบสมีความเฉพาะเจาะจง ซึ่งเป็นหัวใจสำคัญในการควบคุมลักษณะทางพันธุกรรมและจะถูกถ่ายทอดต่อไป

ข้อมูลของดีเอ็นเอถูกเก็บในรูปแบบของตัวอักษร 4 ตัวคือ A G C และ T แทนนิวคลีโอไทด์ทั้ง 4 ชนิด และตัวอักษร N แทนเบสที่ไม่รู้จัก โดยที่ตัวอักษรจะเรียงเป็นสายอักษรตามลำดับของโครงสร้างโมเลกุลของดีเอ็นเอ



ภาพที่ 2 โครงสร้างของดีเอ็นเอ เป็นแบบเกลียวคู่ขนาน (double helix)

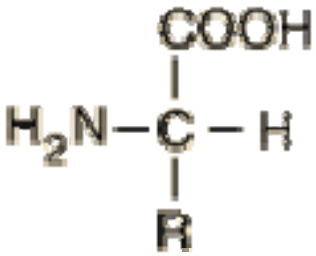
1.2 อาร์เอ็นเอ สามารถพบได้ในนิวเคลียส, ไซโทพลาซึม และไมโทคอนเดรีย มีหน้าที่เกี่ยวข้องกับการสังเคราะห์โปรตีน(protein synthesis) โครงสร้างโมเลกุลของอาร์เอ็นเอมีลักษณะคล้ายกับดีเอ็นเอ คือเป็นสายของนิวคลีโอไทด์ แต่แตกต่างจากดีเอ็นเอคือเป็นสายเดี่ยว และโมเลกุลของอาร์เอ็นเอจะประกอบด้วยนิคลีโอไทด์ 4 ชนิด คือ อะดินีน, กัวนีน, ไซโทซีน, และ ยูราซิล การเก็บข้อมูลของอาร์เอ็นเอมีรูปแบบเหมือนกับดีเอ็นเอ คือเป็นสายของตัวอักษร สายอักษรของข้อมูลอาร์เอ็นเอจะประกอบด้วยตัวอักษร 4 ตัว คือ A G C และ U

2. โปรตีน

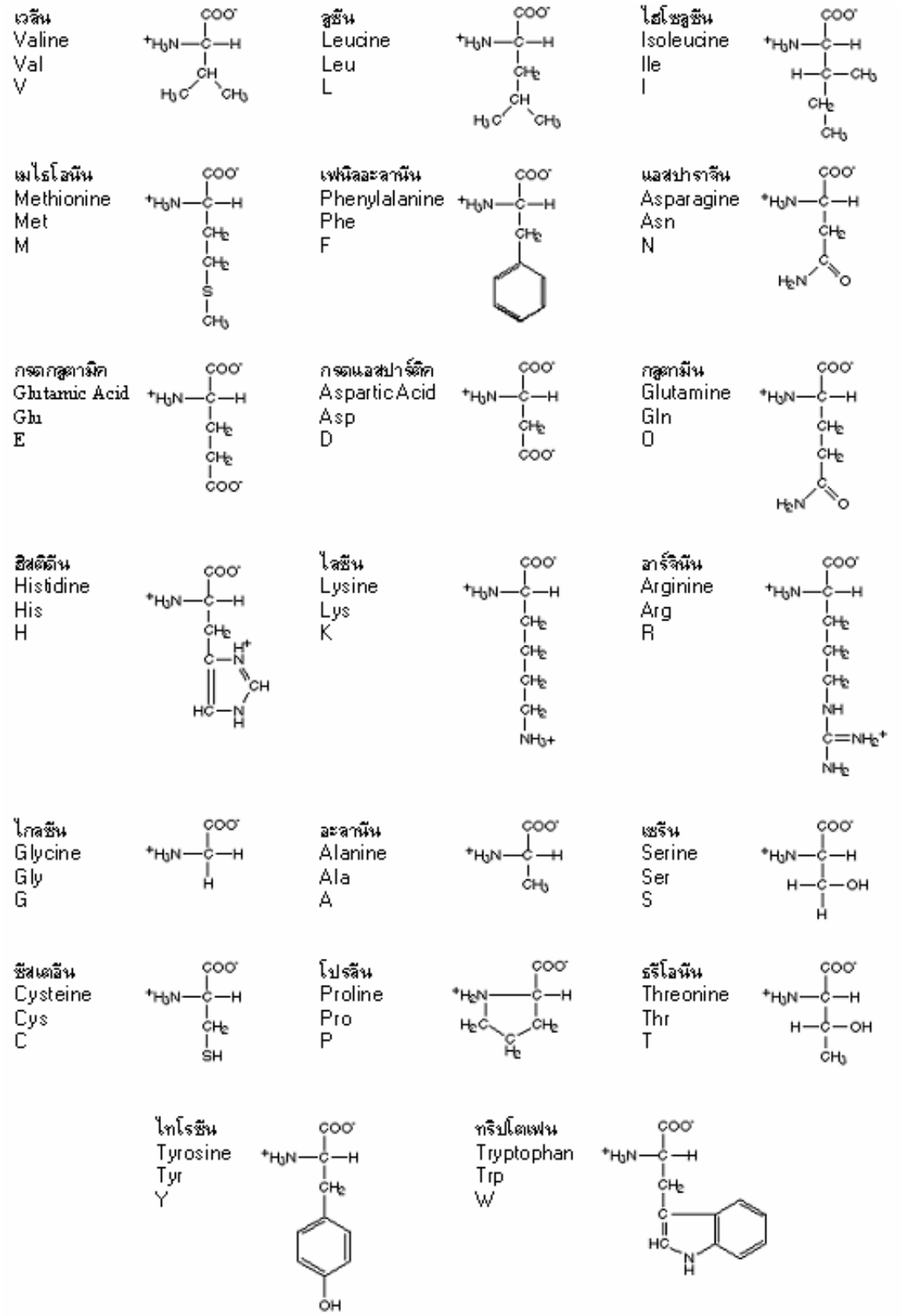
โปรตีน(ไพศาล, 2539; พจน์ และคณะ, 2543) จัดเป็นส่วนประกอบที่สำคัญของเซลล์ และเป็นส่วนประกอบของเอนไซม์ทุกชนิด โปรตีนและเอนไซม์จัดว่ามีความสำคัญในการก่อให้เกิดลักษณะต่าง ๆ ในสิ่งมีชีวิต โปรตีนจัดเป็นโมเลกุลขนาดใหญ่ โปรตีนแต่ละชนิดประกอบด้วยเส้นโพลีเปปไทด์(Polypeptide) จำนวน 1 เส้น หรือมากกว่า 1 เส้น โพลีเปปไทด์ประกอบด้วยกรดอะมิโน(Amino acid) หลายชนิดมาเรียงต่อกัน กรดอะมิโนเป็นกรดอินทรีย์ที่ประกอบด้วยหมู่คาร์บอกซิลิก(Carboxylic group, -COOH), หมู่อะมิโน(Amino group, -NH−2), อะตอมไฮโดรเจน, และหมู่ R (side chain) ต่ออยู่กับอะตอมคาร์บอนในตำแหน่งแอลฟา(α-carbon) ดังภาพที่ 3 เนื่องจากหมู่ทั้งสี่ที่ต่อกับอะตอมคาร์บอนนี้ไม่เหมือนกัน คาร์บอนนี้จึงเป็นคาร์บอนชนิดอสมมาตร(Asymmetric carbon) ดังนั้นกรดอะมิโนทุกตัว(ยกเว้นกรดอะมิโนไกลซีนซึ่งมีหมู่ R เป็นอะตอมไฮโดรเจน) จะ

มีสเตริโอไอโซเมอร์ได้ 2 ชนิด คือ D- และ L- กรดอะมิโนที่พบในร่างกายมนุษย์เป็นชนิด L- กรดอะมิโนมีอยู่มากกว่า 80 ชนิด แต่ชนิดที่พบทั่วไปในโปรตีนชนิดต่าง ๆ และถือว่ามีความสำคัญมีอยู่เพียง 20 ชนิดเท่านั้น ซึ่งมีโครงสร้างโมเลกุล ชื่อ และชื่อย่อดังภาพที่ 4

ข้อมูลของโปรตีนถูกเก็บอยู่ในรูปแบบตัวอักษร 20 ตัวคือ A R N D C Q E G H I L K M F P S T W Y และ V แทนกรดอะมิโนทั้ง 20 ชนิด ตามชื่อย่อแบบหนึ่งตัวอักษร โดยตัวอักษรจะเรียงตัวเป็นสายอักษรตามการเรียงตัวของกรดอะมิโนภายในเส้นของโพลีเปปไทด์



ภาพที่ 3 โครงสร้างทั่วไปทางเคมีของกรดอะมิโน



ภาพที่ 4 โครงสร้างโมเลกุล และชื่อของกรดอะมิโนทั้ง 20 ชนิด

รูปแบบการจัดเก็บข้อมูลชีวภาพ

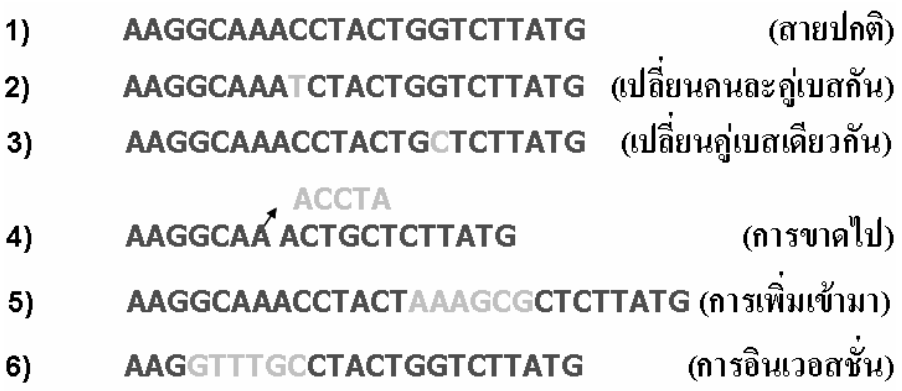
รูปแบบการจัดเก็บข้อมูลชีวภาพแสดงในงานวิจัยนี้จัดเก็บอยู่ในรูปแบบ FASTA (David, 2001; วีรวุฒิ, 2545) ดังภาพที่ 5 โดยมี 3 ส่วน คือส่วนที่ 1 มีเครื่องหมาย > ตามด้วยชื่อและแหล่งที่มาของข้อมูลแต่ละสายอักษร ส่วนที่ 2 เป็นข้อมูลลำดับนิวคลีโอไทด์หรือกรออะมิโน ซึ่งใช้อักษรตัวเดียวเป็นสัญลักษณ์ ส่วนที่ 3 เป็นเครื่องหมาย * แสดงถึงจุดสิ้นสุดของข้อมูล โดยส่วนนี้จะมีหรือไม่ก็ได้ เครื่องหมาย * อาจจำเป็นต้องใช้ในบางโปรแกรม

```
>YCZ2_YEAST protein in HMR 3′ region
MKAVVIEDGKAVVKEGVPIPELEEGFV
GNPTDWAHIDYKVGPQGSILGCDAAGQ
IVKLGPAVDPKDFSIGDYIYGFIHGSS
VRFPSNGAFAEYSAISTVVAYKSPNEL
KFLGEDVLPAGPVRSLEGAATIPVSLT*
```

ภาพที่ 5 ตัวอย่างการจัดเก็บข้อมูลชีวภาพแบบ FASTA

ลักษณะการเปลี่ยนแปลงของข้อมูลชีวภาพ

การเปลี่ยนแปลงในดีเอ็นเอ หรือโปรตีน จะมีลักษณะเหมือนกันคือจะเกิดจากการแทรก การลบ หรือการเปลี่ยนของตัวอักษร ดังตัวอย่างภาพที่ 6 แสดงการเปลี่ยนของลำดับเบสในดีเอ็นเอ การเปลี่ยนแปลงของข้อมูลชีวภาพในลักษณะต่าง ๆ เหล่านี้ส่งผลให้ลักษณะของสิ่งมีชีวิตแตกต่างไปจากเดิมด้วย(ไพศาล, 2539; อมรา, 2540; วีรวุฒิ, 2545)

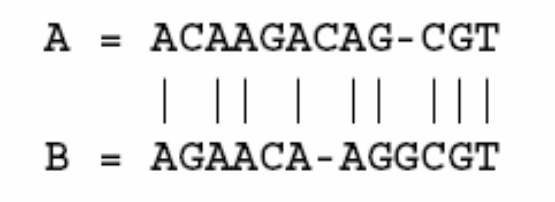


ภาพที่ 6 รูปแบบการเปลี่ยนแปลงของเบสในดีเอ็นเอ โดย 1) คือลำดับเบสของดีเอ็นเอสายปกติ 2) คือการเปลี่ยนจาก C ไป T เป็นการเปลี่ยนของคนละคู่เบสกัน(Transition) 3) คือการเปลี่ยนจาก G ไป C เป็นการเปลี่ยนกับคู่คอมพลีเมนกัน (Transversion) 4) คือการขาดไปของช่วงลำดับเบส ACCTA 5) คือการเพิ่มเข้ามาของช่วงลำดับเบส AAAGC 6) คือการอินเวอสชั่น จาก 5-GCAAAC-3 เปลี่ยนเป็น 5-GTTTGC-3

การเทียบเรียงข้อมูล

การเทียบเรียงข้อมูล(Sequence Alignment) (Crochemore, 2003) เป็นวิธีสำหรับการเปรียบเทียบข้อมูล เพื่อระบุส่วนที่คล้าย และส่วนที่แตกต่าง ของข้อมูลที่ต้องการเปรียบเทียบ การเทียบเรียงจะพิจารณาจากความคล้ายกันของอักขระในแต่ละตำแหน่งของลำดับข้อมูล ซึ่งค่าความคล้ายกันจะคำนวณจากรูปแบบการให้คะแนนที่เหมาะสม

รูปแบบเบื้องต้นในการเทียบเรียงข้อมูลจำนวน 2 ลำดับเริ่มจากเขียนลำดับแรกไว้บรรทัดบนของที่ 2 แล้วพิจารณาส่วนย่อยที่มีความเหมือนกันระหว่าง 2 ลำดับ ในแต่ละลำดับสามารถแยกส่วนย่อยออกจากกันได้ด้วยช่องว่าง ซึ่งโดยปกติแล้วการแทรกช่องว่างจะไม่แทรกในตำแหน่งที่ตรงกันของทั้งสองลำดับ ผลลัพธ์สุดท้ายที่ได้หลังจากเทียบเรียงแล้วลำดับที่เทียบเรียงจะมีความยาวเท่ากัน ตัวอย่างการเทียบเรียงลำดับ A = “ACAAGACAGCGT” และลำดับ B = “AGAACAAGGCGT” เป็นดังภาพที่ 7



ภาพที่ 7 ตัวอย่างการเทียบเรียงลำดับ 2 ลำดับ

จากตัวอย่างภาพที่ 7 จะเห็นได้ว่าผลการเทียบเรียงมีอักขระที่เหมือนกันในตำแหน่งเดียวกันทั้งหมด 9 ตัว อย่างไรก็ตามการเทียบเรียงให้อักขระที่ต่างกันอยู่ในตำแหน่งเดียวกันสามารถเกิดขึ้นได้ ดังเช่นในตัวอย่างตำแหน่งที่ 2 อักขระ C ของลำดับ A ถูกจัดวางในตำแหน่งที่ตรงกับอักขระ G ของลำดับ B และตำแหน่งที่ 4 อักขระ G ของลำดับ A ก็ถูกจัดวางในตำแหน่งที่ตรงกับอักขระ C ของลำดับ B เช่นเดียวกัน นอกจากการเทียบเรียงอักขระที่เหมือนและไม่เหมือนแล้ว การแทรกช่องว่างภายในลำดับก็สามารถทำได้เช่นเดียวกัน เช่นในตำแหน่งที่ 10 ของลำดับ A การแทรกช่องว่างทำให้ได้การเทียบเรียงที่ดีสำหรับอักขระ 3 ตัวสุดท้าย

หากมองในมุมมองของการเปลี่ยนแปลงของข้อมูลแล้ว ผลของการเทียบเรียงข้อมูลสามารถแสดงให้เห็นถึงรูปแบบการเปลี่ยนแปลงจากลำดับหนึ่งไปเป็นอีกลำดับหนึ่งได้ โดยในตำแหน่งที่อักขระไม่ตรงกันถือว่าเป็นการแทนที่(substitution) ของอักขระ ในตำแหน่งที่เป็นช่องว่างในลำดับแรกถือว่าเป็นการแทรก(insertion) อักขระที่ลำดับที่ 2 และในตำแหน่งที่เป็นช่องว่างในลำดับที่ 2 ถือว่าเป็นการลบ(deletion) อักขระที่ลำดับแรกออก จากตัวอย่างการเทียบเรียงในภาพที่ 7 สามารถสรุปได้ว่าการเปลี่ยนแปลงจากลำดับ A ไปเป็นลำดับ B มี 4 ขั้นตอน คือ 1) แทนที่อักขระ C ด้วย G ในตำแหน่งที่ 2; 3) แทนที่อักขระ G ด้วย C ในตำแหน่งที่ 5; 4) ลบอักขระ C ในตำแหน่งที่ 7; และ 4) แทรกอักขระ G ในตำแหน่งที่ 10

การเทียบเรียงข้อมูลนั้นสามารถมีผลการเทียบเรียงได้มากมายหลายรูปแบบ แต่ผลการเทียบเรียงที่สนใจนั้น คือผลที่เทียบเรียงข้อมูลให้มีความคล้ายกันมากที่สุด วิธีการวัดผลความคล้ายกันของการเทียบเรียงข้อมูลสามารถคำนวณได้จากค่าคะแนนความคล้ายของแต่ละคู่อักขระซึ่งมีหลายรูปแบบ โดยปกติแล้วค่าคะแนนความคล้ายจะให้คะแนนที่ดีสำหรับการเทียบเรียงอักขระเดียวกัน และให้คะแนนที่ไม่ดี หรือหักคะแนนสำหรับการเทียบเรียงอักขระที่ต่างกันหรือช่องว่าง คะแนนรวมของการเทียบเรียงจะคำนวณจากผลรวมคะแนนในแต่ละตำแหน่งของผลการเทียบเรียง ตัวอย่างเช่นหากวิธีการให้คะแนนอยู่ในรูปแบบให้คะแนน +1 สำหรับการเทียบเรียงอักขระเดียวกัน

และให้คะแนน -1 สำหรับการเทียบเรียงอักขระต่างกันหรือช่องว่าง จากตัวอย่างการเทียบเรียงในภาพที่ 7 มีการเทียบเรียงอักขระเดียวกัน 9 ตัว และมีการเทียบเรียงอักขระที่ต่างกัน หรือช่องว่าง 4 ตัว ดังนั้นคะแนนความคล้ายกันจะมีค่าเท่ากับ 5 เป็นต้น

เมื่อพิจารณาถึงจำนวนข้อมูลที่ต้องการเทียบเรียง การเทียบเรียงข้อมูลสามารถแบ่งได้เป็น 2 แบบ คือ การเทียบเรียงคู่ลำดับข้อมูล(Pairwise Alignment) และการเทียบเรียงกลุ่มลำดับข้อมูล (Multiple Alignment)

การเทียบเรียงคู่ลำดับข้อมูล คือ การเทียบเรียงลำดับข้อมูลจำนวน 2 ลำดับ เพื่อหาความคล้าย, ความแตกต่าง หรือการเปลี่ยนแปลงระหว่างข้อมูล และนอกจากนั้นยังสามารถประยุกต์ไปใช้กับการค้นหากลุ่มลำดับที่ใกล้เคียง(Homologous sequence) จากฐานข้อมูล สำหรับลำดับที่ต้องการได้อีกด้วย

การเทียบเรียงกลุ่มลำดับข้อมูล(Gusfield, 1999) คือการเทียบเรียงลำดับข้อมูลจำนวนตั้งแต่ 3 ลำดับขึ้นไป เพื่อหาความสัมพันธ์โดยรวมของข้อมูลทั้งหมดที่เทียบเรียง โดยจุดประสงค์หลักคือการหาส่วนที่คล้ายกันจากทุกลำดับที่เทียบเรียง ซึ่งเป็นวิธีการที่ขยายมาจากการเทียบเรียงคู่ลำดับข้อมูล

เมื่อพิจารณาถึงช่วงของลำดับที่ต้องการเทียบเรียงข้อมูล การเทียบเรียงข้อมูลสามารถแบ่งได้เป็น 2 แบบ คือ การเทียบเรียงแบบโกลบอล(Global Alignment) และการเทียบเรียงแบบโลคอล (Local Alignment)

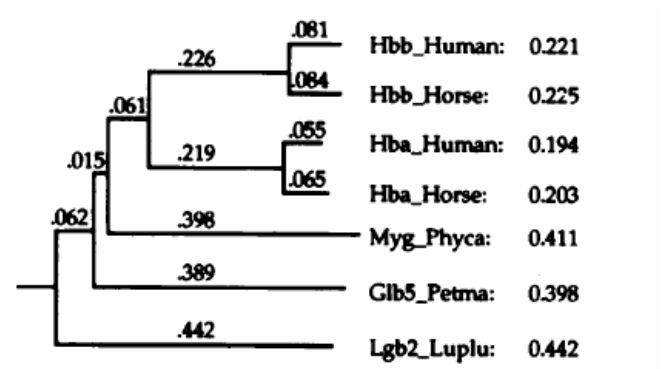
การเทียบเรียงแบบโกลบอล(Gusfield, 1999:216-217) เป็นการหาการเทียบเรียงที่ดีที่สุดของทั้งสายลำดับข้อมูล โดยการเทียบเรียงแบบโกลบอลมีนิยาม คือ การเทียบเรียงแบบโกลบอลของลำดับข้อมูล S1 และ S2 จำนวน 2 ลำดับ ทำได้โดยแทรกช่องว่างภายในลำดับ หรือที่จุดสิ้นสุดของลำดับ ของทั้งลำดับ S1 และ S2 หลังจากนั้นจัดวางผลการเทียบเรียงของทั้ง 2 ลำดับ โดยให้ลำดับหนึ่งอยู่เหนืออีกลำดับหนึ่ง ผลที่ได้จากการเทียบเรียงคือทุกๆอักขระหรือช่องว่างในแต่ละลำดับต้องตรงข้ามกับอักขระ หรือช่องว่างของอีกลำดับเพียงตัวเดียวเท่านั้น

จากนิยามของการเทียบเรียงแบบโกลบอล คือผลที่ได้จากการเทียบเรียง ทุกลำดับจะมีความยาวเท่ากัน และแสดงให้เห็นถึงการจัดเรียงอักขระของทั้งลำดับข้อมูลให้มีความคล้ายกันมากที่สุด ซึ่งผลที่ได้จะแตกต่างกันไปขึ้นอยู่กับวิธีการ และค่าการให้คะแนนที่เลือกใช้

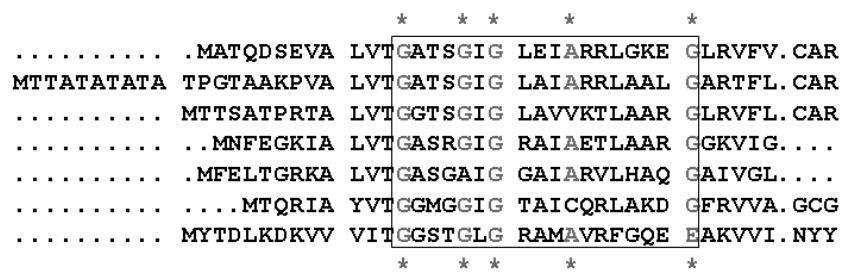
การเทียบเรียงแบบโลคอล(Gusfield, 1999:230-232) เป็นการหาส่วนย่อยของลำดับที่ให้ผลการเทียบเรียงที่ดีที่สุด โดยการเทียบเรียงแบบโลคอลมีนิยาม คือ การเทียบเรียงแบบโลคอลของลำดับข้อมูล S1 และ S2 คือการหาส่วนย่อยของลำดับข้อมูล S1 และ S2 ที่ให้ผลการเทียบเรียงแบบโกลบอลระหว่างส่วนย่อยสูงที่สุด และดีกว่าทุกคู่ของส่วนย่อยอื่นที่ได้จาก S1 และ S2

ประโยชน์ของการเทียบเรียงกลุ่มข้อมูลชีวภาพ

การเทียบเรียงกลุ่มลำดับข้อมูลทางชีวภาพเป็นวิธีการที่ทำให้ทราบถึงความสัมพันธ์ภายในโครงสร้างของสายโปรตีน ซึ่งสามารถนำไปประยุกต์ใช้กับงานทางด้านชีววิทยาได้หลากหลาย งานทางด้านชีววิทยาที่นิยมใช้วิธีการเทียบเรียงกลุ่มลำดับข้อมูลทางชีวภาพมีหลักๆ 3 แนว คือ วิเคราะห์วิวัฒนาการของสิ่งมีชีวิต(Phylogenetic analyses) ดังตัวอย่างภาพที่ 8 ระบุส่วนจำเพาะ (Conserved motifs and domains) ของกลุ่มข้อมูลดังตัวอย่างภาพที่ 9 และทำนายโครงสร้างทุติยภูมิและตติยภูมิ(Secondary and tertiary structure prediction) ของข้อมูลโปรตีน(Notredame, 2002)



ภาพที่ 8 ตัวอย่างแผนภูมิต้นไม้แสดงความสัมพันธ์ทางวิวัฒนาการที่วิเคราะห์ได้จากการเทียบเรียงกลุ่มลำดับข้อมูลชีวภาพ



ภาพที่ 9 ตัวอย่างผลของการเทียบเรียงกลุ่มลำดับข้อมูลเพื่อระบุส่วนจำเพาะของกลุ่มลำดับข้อมูล โดยส่วนที่มีเครื่องหมาย * คือส่วนจำเพาะของกลุ่มลำดับข้อมูล

ตารางค่าคะแนน

ตารางค่าคะแนนในการเทียบเรียงกลุ่มลำดับข้อมูล(วีรวุฒิ, 2545) เป็นค่าคะแนนที่บอกความเหมือนของแต่ละตัวอักษรของข้อมูลชีวภาพ สร้างค่าคะแนนนโดยการหาทางเคมีโดยแต่ละตารางมีค่าคะแนนที่ต่างกัน การเลือกใช้ตารางค่าคะแนนก็มีผลต่อความถูกต้องของการเทียบเรียง โดยงานวิจัยนี้ใช้ตารางค่าคะแนน Gonnet250 ซึ่งเป็นตารางค่าคะแนนที่ปกติโปรแกรม ClustalW (Thomson et al., 1994) ใช้ นอกจากนี้ยังสามารถเลือกใช้ตารางค่าคะแนนได้หลายแบบซึ่งเปลี่ยนได้โดยผู้ใช้ ได้แก่ BLOSUM PAM และ Gonnet ซึ่งเป็นตารางค่าคะแนนที่นิยมใช้

1. BLOSUM

BLOSUM เป็นตารางค่าคะแนนที่คิดค้นโดย Henikoff ในปี ค.ศ. 1992 ซึ่งเหมาะกับการค้นหาความเหมือน (similarity searches) เหมาะกับโปรแกรมในการค้นหา ได้แก่ FASTA และ BLAST โดยที่ตารางค่าคะแนน BLOSUM มีหลายระดับ แต่ละระดับมีค่าน้ำหนักที่ต่างกัน เช่น BLOSUM62 คำนวณจากชุดโปรตีนที่มีความเหมือนกันมากกว่า 62% เป็นต้น

2. PAM

PAM เป็นตารางค่าคะแนนโดย Dayhoff ซึ่งเป็นที่นิยมตั้งแต่ช่วงปี ค.ศ. 1978 โดยมีพื้นฐานจากร้อยละของการยอมรับการเปลี่ยนรูปของสายอักษร (mutation) และการอ้างถึงการเปลี่ยนในโปรตีน โดยการทดสอบความถูกต้อง Dayhoff พบว่าตารางค่าคะแนน PAM เหมาะสำหรับการหาค่าความห่างของความสัมพันธ์

3. Gonnet

Gonnet เป็นตารางค่าคะแนนโดย Cohen และ Benner ในปี ค.ศ. 1992 ซึ่งตารางค่าคะแนนนี้พัฒนามาจากตารางค่าคะแนน PAM โดยตารางค่าคะแนน Gonnet จะอิงชุดข้อมูลที่มีขนาดใหญ่

เทคนิคที่ใช้ในการเทียบเรียงกลุ่มข้อมูล

ในส่วนนี้จะกล่าวถึงเทคนิคที่ใช้ในการเทียบเรียงกลุ่มลำดับข้อมูล ซึ่งจะกล่าวถึง 2 เทคนิคหลัก คือ การโปรแกรมแบบพลวัต และฮิวริสติก

1. การโปรแกรมแบบพลวัต(Dynamic Programming)

การโปรแกรมแบบพลวัต (Dynamic Programming) (Gusfield, 1999; Notredame, 2002; วรวุฒิ, 2545) เป็นวิธีการที่ให้ค่าการเทียบเรียงที่ดีที่สุด โดยการคำนวณอ้างอิงจากตารางค่าคะแนนที่เลือกใช้ ซึ่งวิธีการหาเริ่มจากการเลือกตารางค่าคะแนน และค่าของช่องว่างที่ต้องการใช้ และสร้างตารางการคำนวณ โดยให้ลำดับข้อมูลแรกกำกับแนวสดมภ์ และลำดับข้อมูลที่สองกำกับแนวแถวของตารางการคำนวณ และมีเงื่อนไขพื้นฐานตามสมการที่ 1 และ สมการที่ 2

$$F(i,0) = d \tag{1}$$

$$F(0,j) = d \tag{2}$$

โดยที่ i และ j คือตำแหน่งของสดมภ์ และแถวในตารางการคำนวณ $F(i,j)$ คือค่าในตารางที่ได้จากการคำนวณแบบพลวัต และ d คือค่าของช่องว่างที่เลือกใช้

จากเงื่อนไขพื้นฐานตามสมการที่ 1 และ สมการที่ 2 การคำนวณต่อมาเป็นการเติมค่าในตารางการคำนวณให้ครบ โดยการเทียบเรียงแบบโกลบอลใช้สมการที่ 3 ส่วนการเทียบเรียงแบบโลคอลจะเริ่มต้นค่าพื้นฐานเป็นศูนย์ และใช้สมการที่ 4 หาค่าเพื่อใส่ในตารางการคำนวณ วิธีการคำนวณค่าในตารางเป็นดังภาพที่ 10

กำหนดให้

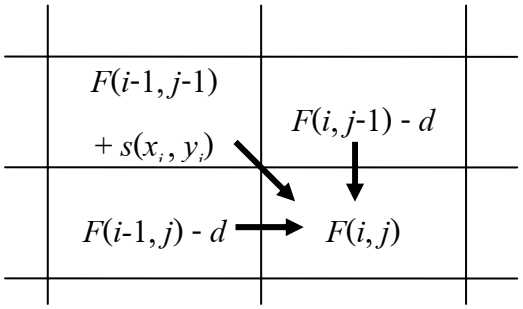
$F(i,j)$ เป็นค่าที่คำนวณได้ในตารางการคำนวณ

$s(x_i, y_j)$ เป็นค่าเปรียบเทียบตัวอักษรในตำแหน่งนั้น อ้างอิงจากตารางค่าคะแนน

d เป็นค่าช่องว่างที่เลือกใช้

$$F(i,j) = \max \begin{cases} F(i-1,j\text{-}1) + s(x_i, y_j), \\ F(i-1,j) - d, \\ F(i,j-1) - d. \end{cases} \quad (3)$$

$$F(i,j) = \max \begin{cases} 0, \\ F(i-1,j\text{-}1) + s(x_i, y_j), \\ F(i-1,j) - d, \\ F(i,j-1) - d. \end{cases} \quad (4)$$



ภาพที่ 10 การคำนวณค่าในตารางการคำนวณ

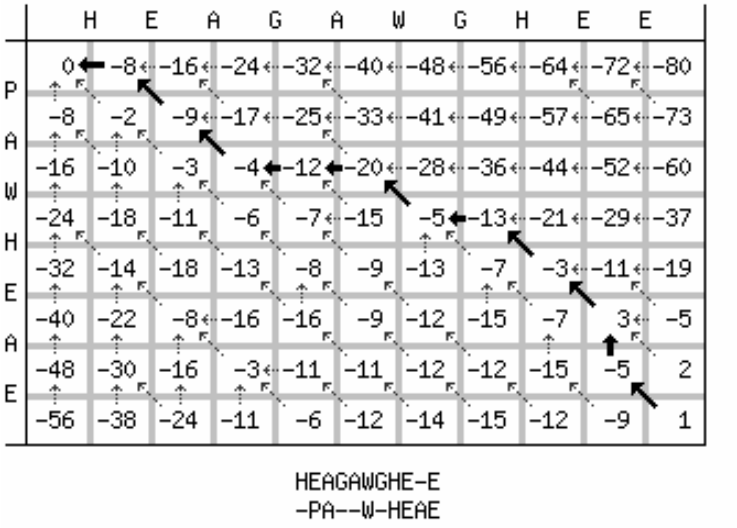
เมื่อคำนวณจนได้ค่าในตารางการคำนวณครบแล้ว ผลลัพธ์ของการเทียบเรียงข้อมูลได้จากการหาเส้นทางย้อนกลับจากมุมล่างขวาของตารางการคำนวณย้อนทางค่าที่คำนวณได้จนถึงตำแหน่งบนซ้ายของตารางก็จะได้การเทียบเรียงข้อมูลที่ให้ค่าคะแนนที่ดีที่สุด โดยหากเส้นทางการย้อนกลับเป็น i – 1 และ j – 1 หมายถึงอักขระตำแหน่ง i และ j ถูกเทียบเรียงอยู่ในตำแหน่งที่ตรงกัน หากการย้อนกลับเป็น i – 1 และ j หมายถึงอักขระในสดมภ์ตำแหน่ง i ถูกเทียบเรียงคู่กับช่องว่าง และหากการย้อนกลับเป็น i และ j – 1 หมายถึงอักขระในแนวแถวตำแหน่งที่ j ถูกเทียบเรียงคู่กับช่องว่าง ซึ่งตัวอย่างการคำนวณแบบพลวัตเป็นดังตัวอย่างที่ 1

ตัวอย่างที่ 1 การเทียบเรียงแบบโกลบอล และการเทียบเรียงแบบโลคอล ของลำดับข้อมูล HEAGAWGHEE และ PAWHEAE โดยวิธีการโปรแกรมแบบพลวัต และใช้ตารางค่าคะแนน BLOSUM50 ดังตารางที่ 1 และใช้ค่าของช่องว่างคือ -8

ตารางที่ 1 ตารางค่าคะแนน BLOSUM50

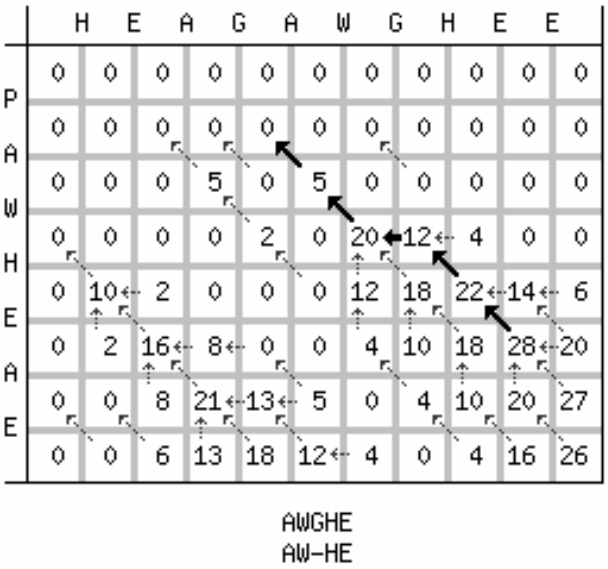
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5	-2	-1	-2	-1	-1	-1	0	-2	-1	-2	-1	-1	-3	-1	1	0	-3	-2	0
R	-2	7	-1	-2	-4	1	0	-3	0	-4	-3	3	-2	-3	-3	-1	-1	-3	-1	-3
N	-1	-1	7	2	-2	0	0	0	1	-3	-4	0	-2	-4	-2	1	0	-4	-2	-3
D	-2	-2	2	8	-4	0	2	-1	-1	-4	-4	-1	-4	-5	-1	0	-1	-5	-3	-4
C	-1	-4	-2	-4	13	-3	-3	-3	-3	-2	-2	-3	-2	-2	-4	-1	-1	-5	-3	-1
Q	-1	1	0	0	-3	7	2	-2	1	-3	-2	2	0	-4	-1	0	-1	-1	-1	-3
E	-1	0	0	2	-3	2	6	-3	0	-4	-3	1	-2	-3	-1	-1	-1	-3	-2	-3
G	0	-3	0	-1	-3	-2	-3	8	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4
H	-2	0	1	-1	-3	1	0	-2	10	-4	-3	0	-1	-1	-2	-1	-2	-3	2	-4
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	5	2	-3	2	0	-3	-3	-1	-3	-1	4
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	5	-3	3	1	-4	-3	-1	-2	-1	1
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6	-2	-4	-1	0	-1	-3	-2	-3
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7	0	-3	-2	-1	-1	0	1
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8	-4	-3	-2	1	4	-1
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10	-1	-1	-4	-3	-3
S	1	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	5	2	-4	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	5	-3	-2	0
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-4	15	2	-3
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	8	-1
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	5

การเทียบเรียงแบบโกลบอลเริ่มจากการสร้างตารางการคำนวณจากเงื่อนไขพื้นฐานตามสมการ 1 และสมการที่ 2 และคำนวณค่าในแต่ละตำแหน่งของตารางการคำนวณโดยใช้สมการ 3 เมื่อได้คำนวณค่าในตารางจนครบจึงหาเส้นทางย้อนกลับจากมุมล่างขวาของตาราง ก็จะได้ผลการเทียบเรียงดังภาพที่ 11



ภาพที่ 11 การเทียบเรียงแบบโกลบอลโดยใช้การโปรแกรมพลวัต และผลลัพธ์ที่ได้

การเทียบเรียงแบบโลคอลเริ่มจากการสร้างตารางการคำนวณจากเงื่อนไขพื้นฐานเท่ากับศูนย์ และคำนวณค่าในแต่ละตำแหน่งของตารางการคำนวณโดยใช้สมการที่ 4 เมื่อคำนวณค่าในตารางจนครบแล้ว ผลการเทียบเรียงแบบโลคอลหาได้จากการหาเส้นทางย้อนกลับจากจุดที่มีค่าคะแนนสูงที่สุดในตารางย้อนกลับไปจนกระทั่งมีค่าเป็นศูนย์ ดังภาพที่ 12 ค่าคะแนนสูงสุดในตารางคือ 28 และหาเส้นทางย้อนกลับไปจนถึงค่าศูนย์ก็จะได้ผลการเทียบเรียงแบบโลคอลของทั้ง 2 ลำดับ



ภาพที่ 12 การเทียบเรียงแบบโลคอลโดยใช้การโปรแกรมพลวัต และผลลัพธ์ที่ได้

วิธีการเทียบเรียงลำดับข้อมูลโดยวิธีการโปรแกรมพลวัตเป็นวิธีที่รับประกันผลลัพธ์ที่ดีที่สุด โดยมีค่าความซับซ้อนของเวลา (Time Complexity) คือ $O(L^N)$ เมื่อ L คือความยาวเฉลี่ยของแต่ละลำดับ และ N คือจำนวนลำดับที่เข้าเทียบเรียง ดังนั้นหากใช้วิธีการโปรแกรมพลวัตเทียบเรียงลำดับข้อมูลที่มีความยาวของลำดับมาก หรือเทียบเรียงลำดับจำนวนมากจะใช้เวลาในการประมวลผลมาก วิธีการโปรแกรมพลวัตจึงไม่เหมาะสมกับการเทียบเรียงกลุ่มลำดับข้อมูลในระดับปริมาณมากกว่า 5 ลำดับที่มีความยาวของลำดับ 100 อักขระขึ้นไป (Gusfield, 1999)

2. ฮิวริสติค(Heuristic)

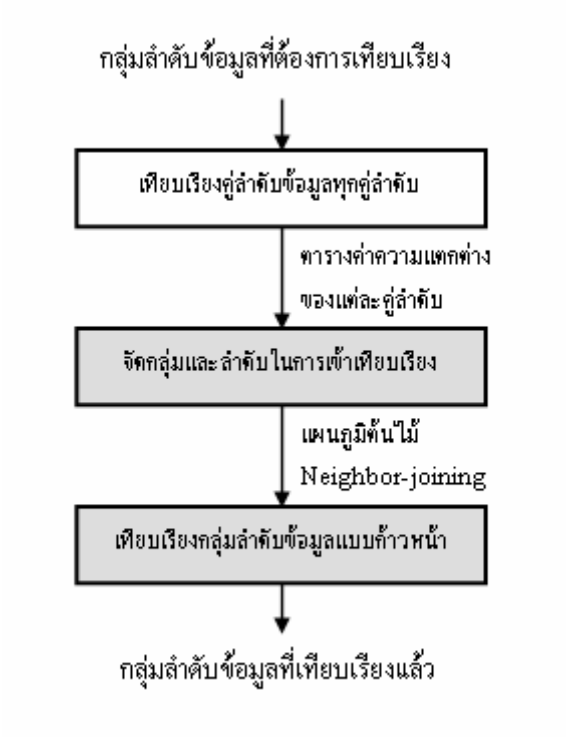
การเทียบเรียงลำดับข้อมูลโดยวิธีการฮิวริสติค เป็นการเทียบเรียงลำดับข้อมูลแบบโดยอ้างอิงหลักการประมาณ พยายามให้ผลการเทียบเรียงที่ได้มีคะแนนการเทียบเรียงใกล้เคียงกับค่าคะแนนที่ดีที่สุด วิธีในกลุ่มฮิวริสติคที่นิยมใช้ในการเทียบเรียงลำดับข้อมูลคือ การเทียบเรียงแบบก้าวหน้า(Progressive Alignment) (Notredame, 2002) ซึ่งเป็นวิธีที่เรียบง่าย และมีประสิทธิภาพสำหรับการเทียบเรียงกลุ่มลำดับข้อมูลโดยใช้เวลา และหน่วยความจำในการประมวลผลไม่มาก ทำให้วิธีนี้เป็นที่นิยมในการใช้งานในการเทียบเรียงปัจจุบัน เนื่องจากใช้เวลาในประมวลผลน้อย และรองรับการเทียบเรียงข้อมูลขนาดใหญ่ หรือปริมาณมากได้(Thomson et al., 1999) โปรแกรมที่ใช้วิธีการในกลุ่มของฮิวริสติคและเป็นที่นิยมได้แก่ Pileup, MultiAlign, ClustalW และ ClustalX ซึ่งโปรแกรม ClustalX เป็นโปรแกรม ClustalW ที่แสดงผลเป็นแบบกราฟฟิก

โครงสร้างการเทียบเรียงกลุ่มลำดับข้อมูลของโปรแกรม ClustalW

โปรแกรม ClustalW เป็นโปรแกรมเทียบเรียงกลุ่มลำดับข้อมูลโดยใช้วิธีการเทียบเรียงข้อมูลแบบก้าวหน้าที่ได้รับความนิยมเพราะเป็นโปรแกรมไม่คิดมูลค่า มีความถูกต้องสูง ใช้กับลำดับที่มากได้ และใช้ได้กับหลายระบบปฏิบัติการ ได้แก่ SUN Solaris, IRIX5.3 บนเครื่อง Silicon Graphics, Digital UNIX บน DEC Stations, Microsoft Windows(32 bit) สำหรับเครื่องคอมพิวเตอร์ส่วนบุคคล, Linux สำหรับเครื่องคอมพิวเตอร์ส่วนบุคคล และ Macintosh PowerMac (วีรวุฒิ, 2545)

โปรแกรม ClustalW สามารถเทียบเรียงได้ทั้งดีเอ็นเอ และโปรตีน โดยทุกลำดับที่ต้องการเทียบเรียงต้องอยู่ในไฟล์เดียวกัน โปรแกรม ClustalW รับไฟล์ได้ 7 รูปแบบ คือ NBRF/PIR, EMBL/SWISSPROT, Pearson (Fasta), Clustal (*.aln), GCG/MSF(Pileup), GCG9/RSF และ

GDE (Thomson *et al*., 1994) การเทียบเรียงกลุ่มลำดับข้อมูลของโปรแกรม ClustalW แบ่งเป็น 3 ขั้นตอนดังภาพที่ 13



ภาพที่ 13 ขั้นตอนในการเทียบเรียงกลุ่มลำดับข้อมูลของโปรแกรม ClustalW

ขั้นตอนแรกเป็นการสร้างตารางค่าความแตกต่างของแต่ละคู่ลำดับข้อมูล (Distance matrix) เพื่อใช้เป็นค่าคะแนนสำหรับจัดลำดับการเทียบเรียงกลุ่มลำดับข้อมูลแบบก้าวหน้า การหาค่าคะแนนความแตกต่างนั้นใช้วิธีเทียบเรียงคู่ลำดับข้อมูลแต่ละคู่ โดยใช้เทคนิคการโปรแกรมพลวัตในการเทียบเรียงแต่ละคู่ลำดับข้อมูล เพื่อคำนวณหาค่าความคล้ายกันของแต่ละคู่ข้อมูล และการคำนวณค่าคะแนนความแตกต่างของแต่ละคู่ข้อมูล ($D_{ClustalW}(x , y)$) จะเป็นดังสมการที่ 5

$$D_{ClustalW}(x , y) = 1\text{-}S_{ClustalW}(x , y) \quad (5)$$

โดยที่ $D_{ClustalW}(x , y)$ คือค่าคะแนนความแตกต่างของคู่ลำดับข้อมูล x และ y และ $S_{ClustalW}(x , y)$ คือค่าคะแนนความคล้ายกันของคู่ลำดับข้อมูล x และ y

การทำงานในขั้นตอนแรกใช้วิธีการโปรแกรมพลวัตทีละคู่ลำดับ ซึ่งมีความซับซ้อนของเวลา $O(L^2)$ เมื่อ L คือความยาวของลำดับ และต้องเทียบเรียงคู่ลำดับให้ครบทุกคู่ลำดับ ซึ่งมีจำนวนการคำนวณทั้งหมด $N(N\text{-}1)/2$ ครั้ง หรือประมาณ $O(N^2)$ เมื่อ N คือจำนวนของลำดับข้อมูลที่ต้องการเทียบเรียง ดังนั้นในขึ้นตอนนี้มีความซับซ้อนของเวลาทั้งหมด $O(N^2L^2)$

ในขั้นตอนที่ 2 เป็นการนำตารางค่าความแตกต่างของแต่ละคู่ลำดับข้อมูลมาสร้างเป็นแผนภูมิต้นไม้โดยใช้วิธีเนเบอร์- จอยนิ่ง (neighbor-joining) (Saitou, 1987) เพื่อจัดกลุ่ม และจัดลำดับการเข้าเทียบเรียงกลุ่มลำดับข้อมูลแบบก้าวหน้าสำหรับแต่ละลำดับข้อมูล

ขั้นตอนสุดท้ายเป็นการเทียบเรียงกลุ่มลำดับข้อมูลแบบก้าวหน้าโดยลำดับการเข้าเทียบเรียงอ้างอิงจากแผนภูมิต้นไม้ที่สร้างในขั้นตอนที่ 2 การเทียบเรียงนั้นเริ่มจากคู่ลำดับข้อมูลที่มีค่าคะแนนความแตกต่างกันน้อยที่สุด หลังจากนั้นจึงนำลำดับข้อมูลที่มีค่าคะแนนความแตกต่างมากกว่าเข้ามาร่วมเทียบเรียงกับกลุ่มลำดับข้อมูลที่เทียบเรียงไปก่อนหน้านี้ ทำจนครบทุกลำดับก็จะได้ผลลัพธ์การเทียบเรียงกลุ่มลำดับข้อมูล

ปัญหาของการเทียบเรียงกลุ่มลำดับข้อมูลชีวภาพ

การเทียบเรียงกลุ่มลำดับข้อมูลทั้งวิธีการโปรแกรมแบบพลวัต และวิธีในกลุ่มฮิวริสติกนั้นได้ผลที่มีความถูกต้องในระดับที่ยอมรับได้ แต่เมื่อนำมาเทียบเรียงกลุ่มลำดับที่มีความยาวของข้อมูลมาก หรือมีจำนวนมากก็จะเกิดปัญหาทางด้านเวลาในการประมวลผลได้

วิธีการโปรแกรมพลวัตเมื่อนำมาใช้กับการเทียบเรียงคู่ลำดับจะมีความซับซ้อนของเวลาคือ $O(L^2)$ และเมื่อใช้กับการเทียบเรียงกลุ่มลำดับข้อมูลจะมีค่าความซับซ้อนของเวลา คือ $O(L^N)$ เมื่อ L คือความยาวเฉลี่ยของแต่ละลำดับ และ N คือจำนวนลำดับที่เข้าเทียบเรียง ซึ่งวิธีนี้ใช้ได้กับการเทียบเรียงกลุ่มลำดับที่มีจำนวนน้อย และมีความยาวไม่มากนัก (วีรวุฒิ, 2545)วิธีการโปรแกรมพลวัตจึงไม่เหมาะสมกับการเทียบเรียงกลุ่มลำดับข้อมูลในระดับปริมาณมากกว่า 5 ลำดับที่มีความยาวของลำดับ 100 อักขระขึ้นไป (Gusfield, 1999)

วิธีการในกลุ่มฮิวริสติกได้รับความนิยมมาก โดยเฉพาะโปรแกรม ClustalW เนื่องจากสามารถเทียบเรียงกลุ่มลำดับข้อมูลที่มีความยาว และจำนวนมากกว่าการโปรแกรมแบบพลวัต แต่หากเทียบเรียงกลุ่มลำดับข้อมูลที่มีขนาดใหญ่มากก็จะใช้เวลานานในการประมวลผลเช่นเดียวกัน

ดังนั้นการพัฒนาเทคนิคที่มีความเร็วในการประมวลผลเพิ่มขึ้นจึงเป็นสิ่งจำเป็นสำหรับปัจจุบันที่ข้อมูลทางชีวภาพเพิ่มขึ้นอย่างรวดเร็ว

อุปกรณ์และวิธีการ

อุปกรณ์

1. ซอฟท์แวร์

1.1 ระบบปฏิบัติการลินุกส์

1.2 โปรแกรม ClustalW เวอร์ชั่น 1.83 (สามารถหาได้จาก ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/)

1.3 โปรแกรม Bali_score สำหรับใช้วัดค่าคะแนนชุดทดสอบ BAliBASE (สามารถหาได้จาก http://www-igbmc.u-strasbg.fr/BioInfo/BAliBASE2/)

1.4 ฐานข้อมูลชีวภาพ BAliBASE เวอร์ชั่น 2 (สามารถหาได้จาก http://www-igbmc.u-strasbg.fr/BioInfo/BAliBASE2/)

1.5 โปรแกรม qscore สำหรับใช้วัดค่าคะแนนชุดทดสอบ PREFAB (สามารถหาได้จาก http://www.drive5.com/muscle/prefab.htm)

1.6 ฐานข้อมูลชีวภาพ PREFAB เวอร์ชั่น 4.0 (สามารถหาได้จาก http://www.drive5.com/muscle/prefab.htm)

1.7 ตัวแปลภาษา C

2. ฮาร์ดแวร์

2.1 เครื่องคอมพิวเตอร์ Athlron XP 2000+ ที่มี CPU Clock rate 1.67 GHz

2.2 หน่วยความจำหลัก 512 MB

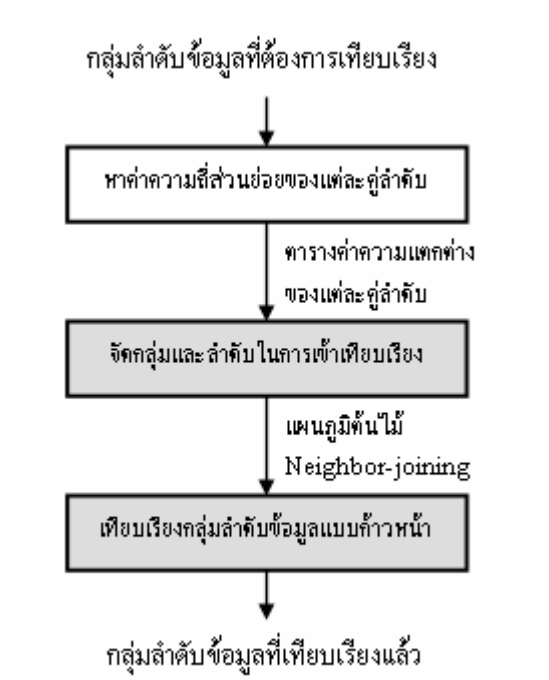
2.3 ฮาร์ดดิสก์ชนิด IDE ขนาด 80 GB จำนวน 1 ตัว

วิธีการ

ในวิทยานิพนธ์นำเสนอขั้นตอนเทคนิคการเทียบเรียงกลุ่มลำดับข้อมูลโดยพิจารณาความถี่ส่วนย่อยของลำดับ (Sub-segment Frequency Alignment :SSFA) (คมสัน, 2548) ซึ่งได้แนวคิดมาจากความรู้เทคนิคการเทียบเรียงกลุ่มลำดับข้อมูลโดยใช้โปรแกรม ClustalW (Thompson *et al*., 1999)

1. ขั้นตอนวิธีการเทียบเรียงกลุ่มลำดับข้อมูล

ขั้นตอนการเทียบเรียงกลุ่มลำดับข้อมูลโดยใช้ความถี่แบบพิจารณาค่าน้ำหนักแบ่งเป็น 3 ขั้นตอนดังภาพที่ 14



ภาพที่ 14 การเทียบเรียงกลุ่มลำดับข้อมูลโดยพิจารณาความถี่ส่วนย่อยของลำดับ

ขั้นตอนแรกเป็นการสร้างตารางค่าความแตกต่างของแต่ละคู่ลำดับข้อมูลโดยพิจารณาความถี่ส่วนย่อยซึ่งแตกต่างจากวิธีของโปรแกรม ClustalW มีรายละเอียดขั้นตอนการคำนวณดังนี้

ขั้นแรกเริ่มจากนำลำดับข้อมูลแต่ละลำดับมาแบ่งเป็นส่วนย่อยโดยใช้วิธี n-gram โดยส่วนย่อยที่ใช้ในงานวิจัยนี้มีความยาว 3 ตัวอักษร ดังตัวอย่างที่ 2

ตัวอย่างที่ 2 การแบ่งส่วนย่อยของแต่ละลำดับ
สมมติให้ข้อมูลที่ต้องการเทียบเรียงมีทั้งหมดจำนวน 3 ลำดับ ได้แก่

ลำดับ a = ACATG

ลำดับ b = TACAG

ลำดับ c = AGCATG

แบ่งเป็นส่วนย่อยได้ดังนี้

ส่วนย่อยลำดับ a คือ a1={ACA}, a2={CAT}, a3={ATG}

ส่วนย่อยลำดับ b คือ b1={TAC}, b2={ACA}, b3={CAG}

ส่วนย่อยลำดับ c คือ c1={AGC}, c2={GCA}, c3={CAT}, c4={ATG}

เมื่อได้ส่วนย่อยของลำดับแล้วจึงนำแต่ละส่วนย่อยของลำดับมาแปลงให้เป็นค่าตัวเลขโดยแทนค่าตัวอักษรในแต่ละหลักด้วยเลข 0-20 ตามตารางที่ 2 และนำมาผ่านสมการแฮช ดังสมการที่ 6

ตารางที่ 2 ตารางค่าตัวเลขสำหรับแทนค่าอักขระในส่วนย่อยของลำดับ

-	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

$$H(x) = (x_2 \text{ x } 21^2) + (x_1 \text{ x } 21) + x_0 \qquad (6)$$

โดยที่ $H(x)$ คือค่าตัวเลขที่ได้จากการแปลงส่วนย่อยของลำดับ x ส่วน x_2 คือค่าตัวเลขที่แทนค่าอักขระตัวแรกของส่วนย่อย x x_1 คือค่าตัวเลขที่แทนค่าอักขระตัวที่ 2 ของส่วนย่อย x และ x_0 คือค่าตัวเลขที่แทนค่าอักขระตัวที่ 3 ของส่วนย่อย x ซึ่งค่า $H(x)$ ที่คำนวณได้สำหรับแต่ละส่วนย่อยของลำดับนั้นจะได้ค่าเท่ากันก็ต่อเมื่อส่วนย่อยทั้ง 2 มีอักขระเดียวกัน และเรียงตัวเหมือนกันเท่านั้น ตัวอย่างการแปลงส่วนย่อยของลำดับเป็นดังตัวอย่างที่ 3

ตัวอย่างที่ 3 การแทนค่าส่วนย่อยของลำดับด้วยตัวเลข จากตัวอย่างที่ 2 เมื่อแบ่งลำดับออกเป็นส่วนย่อยแล้ว นำแต่ละส่วนย่อยมาผ่านสมการแฮชได้ค่าดังนี้

ส่วนย่อยของลำดับ a ได้แก่

H(a1) = H({ACA}) = (1 x 21^2) + (5 x 21) + 1 = 547

H(a2) = H({CAT}) = (5 x 21^2) + (1 x 21) + 17 = 2243

H(a3) = H({ATG}) = (1 x 21^2) + (17 x 21) + 8 = 806

ส่วนย่อยของลำดับ b ได้แก่

H(b1) = H({TAC}) = (17 x 21^2) + (1 x 21) + 5 = 7523

H(b2) = H({ACA}) = (1 x 21^2) + (5 x 21) + 1 = 547

H(b3) = H({CAG}) = (5 x 21^2) + (1 x 21) + 8 = 2234

ส่วนย่อยของลำดับ c ได้แก่

H(c1) = H({AGC}) = (1 x 21^2) + (8 x 21) + 5 = 614

H(c2) = H({GCA}) = (8 x 21^2) + (5 x 21) + 1 = 3634

H(c3) = H({CAT}) = (5 x 21^2) + (1 x 21) + 17 = 2243

H(c4) = H({ATG}) = (1 x 21^2) + (17 x 21) + 8 = 806

เมื่อแปลงส่วนย่อยของลำดับเป็นตัวเลขโดยผ่านฟังก์ชั่นแฮชแล้วจึงเก็บข้อมูลลงในตารางแบบแฮช ส่วนย่อยที่เหมือนกันจะถูกเก็บอยู่ในตำแหน่งของตารางแฮชเดียวกัน โดยข้อมูลที่เก็บในตารางแฮชจะมีการเก็บจำนวนของส่วนย่อยของแต่ละลำดับข้อมูลไว้ด้วย

ในระหว่างการเก็บข้อมูลส่วนย่อยลงในตารางแฮชนั้น จะมีการเก็บข้อมูลความถี่ของส่วนย่อยที่ซ้ำกันของแต่ละคู่ลำดับข้อมูลเพื่อสร้างตารางค่าความคล้ายกันของแต่ละคู่ลำดับข้อมูล ซึ่งค่าความคล้ายกันคำนวณจากอัตราส่วนของความถี่ของส่วนย่อยที่ซ้ำกันของคู่ลำดับข้อมูลต่อจำนวนส่วนย่อยทั้งหมดของคู่ลำดับข้อมูล ดังสมการที่ 7

$$S_{SSFA}(x , y) = \frac{\sum_{£} (k(£))}{(n_x + n_y)} \qquad (7)$$

โดยที่ $S_{SSFA}(x,y)$ คือค่าคะแนนความคล้ายกันพิจารณาจากความถี่ของส่วนย่อยที่ซ้ำกันของลำดับ x และ y, £ คือส่วนย่อย n-gram ที่ซ้ำกัน, n_x และ n_y คือจำนวนของส่วนย่อยแบบ 3-gram ของลำดับ x และ y และ k(£) คือความถี่ของส่วนย่อย £ ที่ปรากฏในลำดับ x และ y

ค่าความคล้ายกันของคู่ลำดับข้อมูลมีค่าอยู่ระหว่าง 0 ถึง 1 โดยค่า 0 หมายถึงคู่ลำดับข้อมูลที่ไม่มีความคล้ายกันเลย และ 1 คือคู่ลำดับข้อมูลที่เหมือนกันทุกประการ ตัวอย่างการคำนวณค่าความคล้ายกันของคู่ลำดับข้อมูลเป็นดังตัวอย่างที่ 4

ตัวอย่างที่ 4 การคำนวณค่าความคล้ายกันสำหรับแต่ละคู่ลำดับข้อมูลจากตัวอย่างที่ 3 สามารถคำนวณได้โดยใช้สมการที่ 7 ดังนี้

ค่าความคล้ายกันของลำดับ a และลำดับ b คือ

$$S_{SSFA}(a,b) = \frac{k(\{ACA\})}{n_a + n_b} = \frac{2}{3+3} = 0.333$$

ค่าความคล้ายกันของลำดับ a และลำดับ c คือ

$$S_{SSFA}(a,c) = \frac{k(\{CAT\}) + k(\{ATG\})}{n_a + n_c} = \frac{2+2}{3+4} = 0.571$$

ค่าความคล้ายกันของลำดับ b และลำดับ c คือ

$$S_{SSFA}(b,c) = \frac{0}{n_b + n_c} = \frac{0}{3+4} = 0$$

เมื่อได้ค่าความคล้ายกันของแต่ละคู่ลำดับข้อมูลแล้ว จึงสร้างตารางค่าความแตกต่างโดยคำนวณจากค่าความคล้ายกันดังสมการที่ 8 และจากตัวอย่างที่ 4 เมื่อคำนวณค่าความแตกต่างแล้วจะได้ตารางค่าความแตกต่างของแต่ละคู่ลำดับข้อมูลดังตารางที่ 3

$$D_{SSFA}(x,y) = 1 - S_{SSFA}(x,y) \quad (8)$$

เมื่อ $D_{SSFA}(x, y)$ คือค่าความแตกต่างของลำดับ x และลำดับ y

ตารางที่ 3 ตารางค่าความแตกต่างของลำดับ a, b และ c ที่เข้าเทียบเรียง

	ลำดับ a	ลำดับ b	ลำดับ c
ลำดับ a	0	0.667	0.429
ลำดับ b	0.667	0	1
ลำดับ c	0.429	1	0

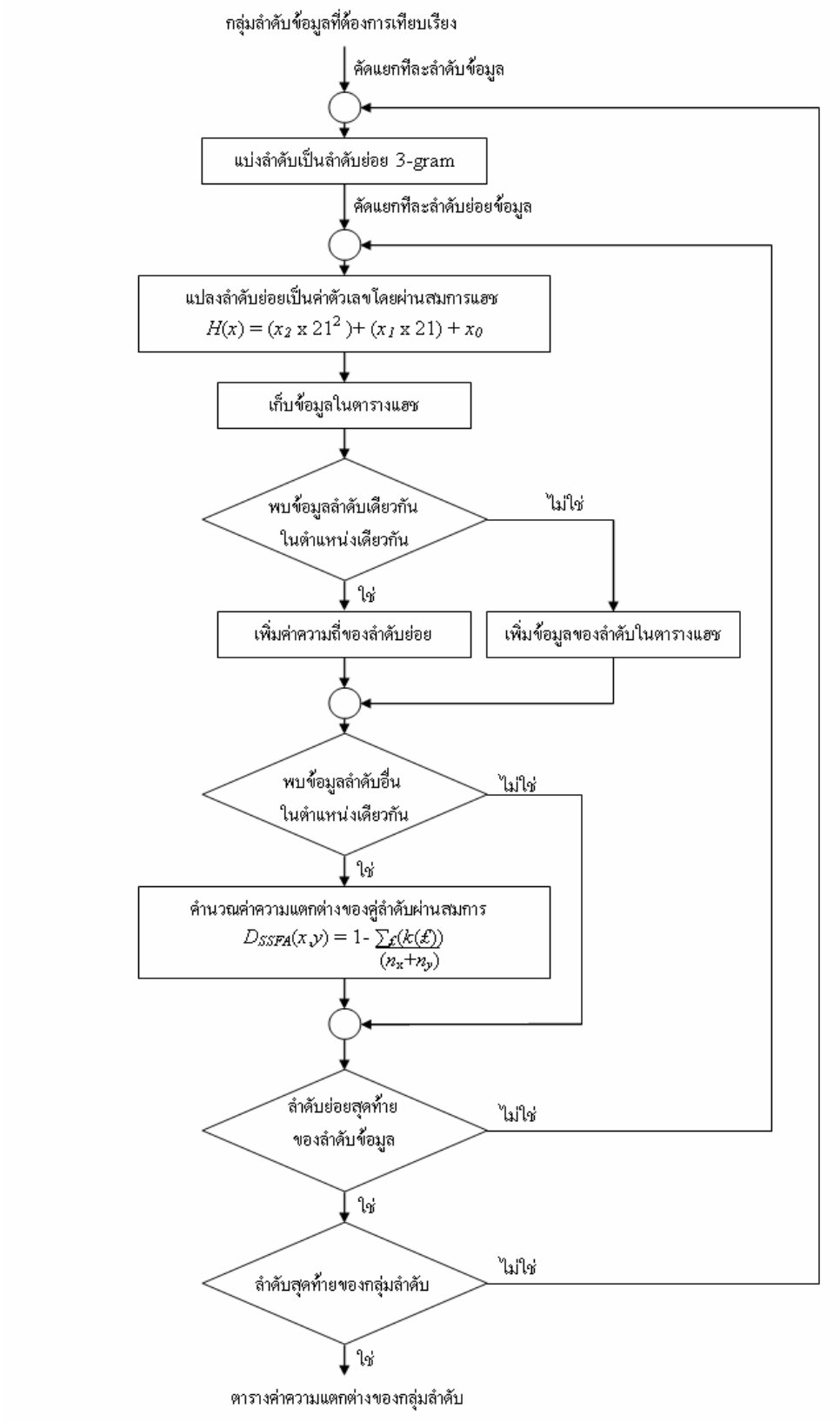
การคำนวณค่าความแตกต่างสามารถคำนวณสะสมได้พร้อมกับการเก็บข้อมูลแต่ละส่วนย่อยในตารางแฮช โดยใช้สมการที่ 9 ซึ่งเกิดจากการแทนค่า $S_{SSFA}(x,y)$ ในสมการที่ 8 ด้วยสมการที่ 7 ดังนั้นขั้นโปรแกรม SSFA จึงมีขั้นตอนการทำงานดังภาพที่ 15 และมีตัวอย่างการคำนวณดังภาพที่ 16

$$D_{SSFA}(x , y) = 1 - \frac{\sum_{£} (k(£))}{(n_x + n_y)} \tag{9}$$

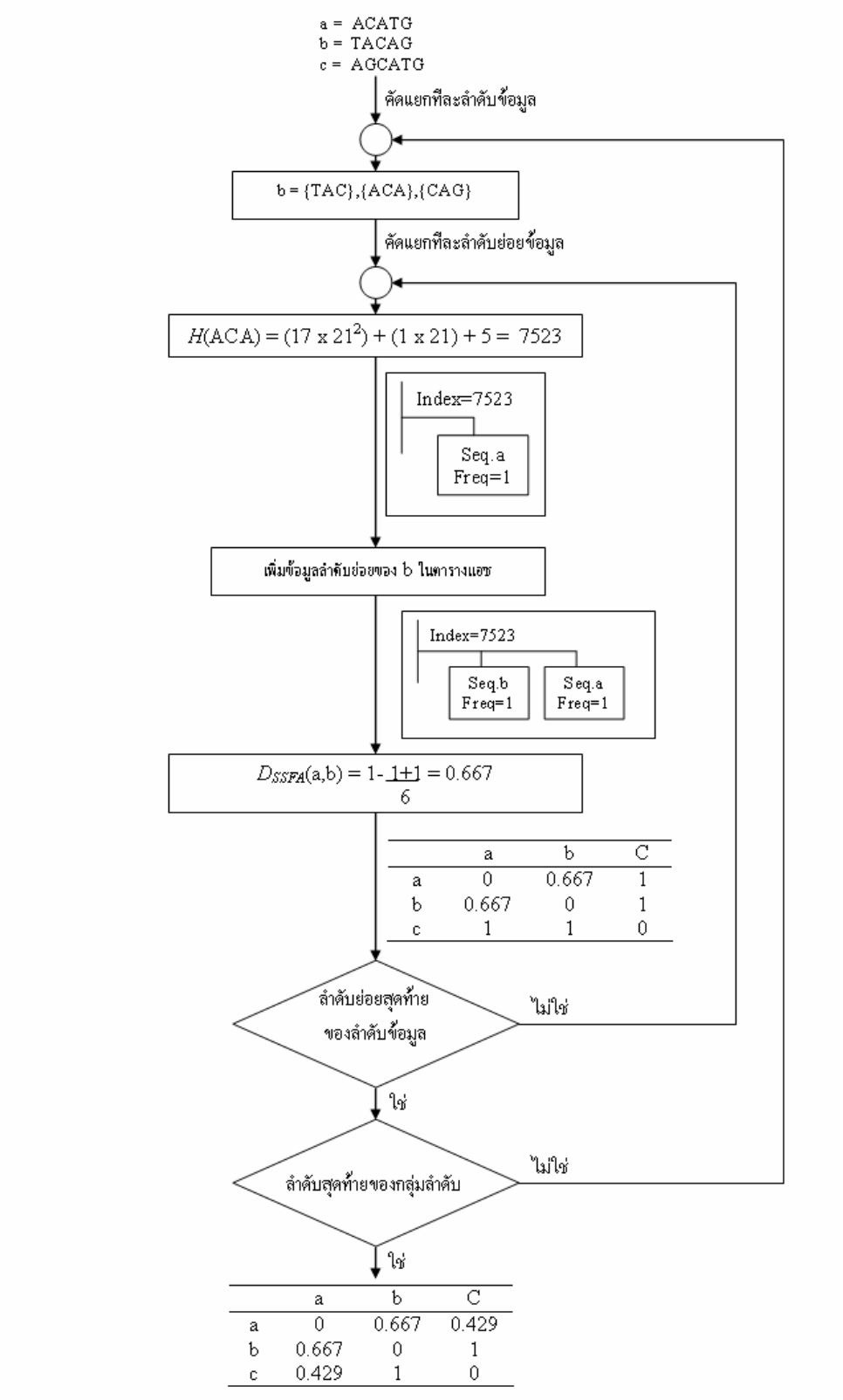
เมื่อได้ตารางค่าความแตกต่างของแต่ละคู่ลำดับข้อมูลแล้ว จึงทำขั้นตอนจัดกลุ่มและลำดับที่เข้าเทียบเรียง และขั้นตอนเทียบเรียงกลุ่มลำดับข้อมูลแบบก้าวหน้าดังภาพที่ 14 โดยทั้ง 2 ขั้นตอนนี้ใช้วิธีการเดียวกับโปรแกรม ClustalW แต่อ้างอิงจากตารางค่าความแตกต่างที่คำนวณจากความถี่ส่วนย่อยของข้อมูลแทนการเทียบเรียงคู่ลำดับของโปรแกรม ClustalW

การแบ่งส่วนย่อยลำดับข้อมูลจำนวน N ลำดับโดยวิธี n-gram นั้นมีค่าความซับซ้อนของเวลา O(N(L-n+1)) หรือประมาณ O(NL) เมื่อ L คือความยาวของลำดับข้อมูล และ n คือความยาวส่วนย่อย n-gram ในส่วนของการคำนวณค่าความคล้ายกันนั้นใช้วิธีคำนวณแบบสะสมจากการเก็บข้อมูลใน hash เดียว ทำให้การสร้างตารางค่าความแตกต่างของคู่ลำดับข้อมูลนั้นมีการประมวลผลแต่ละลำดับข้อมูลเพียงรอบเดียว

การทำงานของโปรแกรม SSFA และการทำงานของโปรแกรม ClustalW นั้นใช้เวลาในการประมวลผลแตกต่างกันในขั้นตอนการสร้างตารางค่าความแตกต่างของคู่ลำดับข้อมูล โดยโปรแกรม SSFA ใช้วิธีการพิจารณาความถี่ของส่วนย่อย และ มีค่าซับซ้อนของเวลา O(NL) ในขณะที่โปรแกรม ClustalW ใช้วิธีเทียบเรียงแต่ละคู่ลำดับข้อมูล และมีค่าซับซ้อนของเวลา O(N^2L^2) ซึ่งทำให้โปรแกรม SSFA ใช้เวลาเทียบเรียงกลุ่มลำดับข้อมูลน้อยกว่าโปรแกรม ClustalW มาก



ภาพที่ 15 ขั้นตอนการสร้างตารางค่าความแตกต่างโดยพิจารณาความถี่ส่วนย่อย



ภาพที่ 16 ขั้นตอนการสร้างตารางค่าความแตกต่างโดยพิจารณาความถี่ส่วนย่อย

2. การเตรียมข้อมูล

ฐานข้อมูลที่ใช้ในงานวิจัยนี้ คือฐานข้อมูล BAliBASE เวอร์ชั่น 2.01 และฐานข้อมูล PREFAB เวอร์ชั่น 4.0

ฐานข้อมูล BAliBASE เวอร์ชั่น 2.01 (Thompson *et al*., 1999; วีรวุฒิ, 2545) เป็นฐานข้อมูลสำหรับวัดผลการเทียบเรียงกลุ่มลำดับข้อมูลโปรตีน ประกอบด้วยกลุ่มลำดับข้อมูลที่เทียบเรียงแล้วสำหรับอ้างอิง 141 กลุ่ม แบ่งเป็นกลุ่มอ้างอิง 8 ชุด โดยในงานวิจัยนี้ใช้กลุ่มอ้างอิงทั้งหมด 5 กลุ่มที่สามารถวัดผลการเทียบเรียงได้ คือ ชุดอ้างอิงที่ 1 เป็นชุดที่มีค่าความห่างเท่ากันและหลายระดับส่วน, ชุดอ้างอิงที่ 2 เป็นการเทียบเรียงกับสายอักขระที่มีการเปลี่ยนแปลงสูง, ชุดอ้างอิงที่ 3 เป็นกลุ่มย่อยที่น้อยกว่า 25% ของส่วนที่เหลือระหว่างกลุ่ม, ชุดอ้างอิงที่ 4 แต่ละกลุ่มข้อมูลประกอบด้วยลำดับข้อมูลที่เป็นส่วนขยายขนาดใหญ่ ที่จุดสิ้นสุด N/C(N/C-terminal extensions) และชุดอ้างอิงที่ 5 แต่ละกลุ่มข้อมูลประกอบด้วยลำดับข้อมูลที่เกิดการแทรกภายในช่วงกลางลำดับ โดยมีรายละเอียดแต่ละชุดอ้างอิงดังตารางที่ 4 – ตารางที่ 9

ตารางที่ 4 ชุดอ้างอิง 1 เป็นชุดที่มีค่าความห่างเท่ากันและหลายระดับส่วนสงวน (ขนาดสั้น)

คุณลักษณะ	ชื่อ(จำนวนสายอักขระ)
ความเหมือนกัน(identity) น้อยกว่า 25%	1aboA (5), 1idy (5), 1r69 (4), 1tvxA (4), 1ubi (4), 1wit (5), 2trx (4) รวม 31 สายอักขระ
ความเหมือนกัน 20-40%	1aab (4), 1fjlA (6), 1hfh (5), 1hpi (4), 1csy (5), 1pfc (5), 1tgxA (4), 1ycc (4), 3cyr (4), 451c (5) รวม 46 สายอักขระ
ความเหมือนกันมากกว่า 35%	1aho (5), 1csp (5), 1dox (4), 1fkj (5), 1fmb (4), 1krn (5), 1plc (5), 2fxb (5), 2mhr (5), 9rnt (5) รวม 48 สายอักขระ

ตารางที่ 5 ชุดอ้างอิง 1 เป็นชุดที่ค่าความห่างเท่ากันและหลายระดับส่วนสงวน(ขนาดกลาง)

คุณลักษณะ	ชื่อ(จำนวนสายอักขระ)
ความเหมือนกัน(identity) น้อยกว่า 25%	1bbt3 (5), 1sbp (5), 1havA (5), 1uky (4), 2hsdA (4), 2pia (4), 3grs (4), kinase (5) รวม 36 สายอักขระ
ความเหมือนกัน 20-40%	1ad2(4),1aym3(4),1gdoA(4), 1ldg(4),1mrj(4),1pgtA(4), 1pii(4) รวม 38 สายอักขระ
ความเหมือนกันมากกว่า 35%	1amk (5), 1ar5A (4), 1ezm (5), 1led (4), 1ppn (5), 1pysA (4), 1thm (4), 1tis (5), 1zin (4), 5ptp (5) รวม 45 สายอักขระ

ตารางที่ 6 ชุดอ้างอิง 1 เป็นชุดที่ค่าความห่างเท่ากันและหลายระดับส่วนสงวน (ขนาดยาว)

คุณลักษณะ	ชื่อ(จำนวนสายอักขระ)
ความเหมือนกัน (identity) น้อยกว่า 25%	1ajsA (4), 1cpt (4), 1lvl (4), 1pamA (5), 1ped (3), 2myr (4), 4enl (3), gal4 (5) รวม 32 สายอักขระ
ความเหมือนกัน 20-40%	1ac5 (4), 1adj (4), 1bgl (4), 1dlc (4), 1eft (4), 1fieA (4), 1gowA (4), 1pkm (4), 1sesA (5), 2ack (5), arp (5), glg (5) รวม 52 สายอักขระ
ความเหมือนกันมากกว่า 35%	1ad3 (4), 1gpb (5), 1gtr (5), 1lcf (6), 1rthA (5), 1taq (5), 3pmg (4), actin (5) รวม 39 สายอักขระ

ตารางที่ 7 ชุดอ้างอิง 2 เป็นการเทียบเรียงกับสายอักขระที่มีการเปลี่ยนแปลงสูง

คุณลักษณะ	ชื่อ(จำนวนสายอักขระ)
ขนาดสั้น	1aboA (15), 1csy (19), 1idy (21), 1r69 (22), 1tgxA (20), 1tvxA (17), 1ubi (17), 1wit (22), 2trx (20) รวม 173 สายอักขระ
ขนาดกลาง	1sbp (18), 1havA (18), 1uky (24), 2hsdA (22), 2pia (18), 3grs (16), kinase (18) รวม 134 สายอักขระ
ขนาดยาว	1ajsA (20), 1cpt (15), 1lvl (24), 1pamA (19), 1ped (19), 2myr (19), 4enl (18) รวม 134 สายอักขระ

ตารางที่ 8 ชุดอ้างอิง 3 เป็นกลุ่มย่อยที่น้อยกว่า 25% ของส่วนที่เหลือระหว่างกลุ่ม

คุณลักษณะ	ชื่อ(จำนวนสายอักขระ)
ขนาดสั้น	1idy (27), 1r69 (23), 1ubi (22), 1wit (19) รวม 91 สายอักขระ
ขนาดกลาง	1uky (24), 2pia (20), kinase (23) รวม 67 สายอักขระ
ขนาดยาว	1ajsA (28), 1pamA (19), 1ped (21), 2myr (21), 4enl (19) รวม 108 สายอักขระ

ตารางที่ 9 ชุดอ้างอิง 4 และชุดอ้างอิง 5

ชุดอ้างอิง	คุณลักษณะ	ชื่อ(จำนวนสายอักขระ)
4	สายอักขระ N/C (N/C-terminal extensions)	1ckaA (10), 1csp (6), 1dynA (16), 1lkl (8), 1mfa (8), 1pfc (10), 1pysA (4), 1vln (14), 1ycc (9), 2abk (7), kinase1 (7), kinase2 (18) รวม 107 สายอักขระ
5	สายอักขระที่มีการแทรกภายใน	1eft (8), 1ivy (7), 1pysA (10), 1qpg (5), 1thm1 (11), 1thm2 (7), 2cba (8), s51 (15), s52 (5), kinase1 (5), kinase2 (12), kinase3 (19) รวม 112 สายอักขระ

ข้อมูลทดสอบมาจาก http://www-igbmc.u-strasbg.fr/BioInfo/BAliBASE2/BAliBASE2.01.tar.gz มีขนาด 7.81 MB และเมื่อทำการขยายไฟล์จะเป็นชุดทดสอบที่มีขนาด 82.4 MB ประกอบด้วย 8 กลุ่มย่อยตามชุดอ้างอิง โดยมีชื่อ ref1, ref2, ref3, ref4, ref5, ref6, ref7 และ ref8 ตามลำดับ ในงานวิจัยนี้ใช้กลุ่มย่อย ref1 – ref5 โดยใช้ชุดอ้างอิงเท่ากับ ref เช่น ชุดอ้างอิง: 1 ตามโครงสร้างคือ ref1 และชุดทดสอบเท่ากับ test เช่น ชุดทดสอบ 1 ตามโครงสร้างคือ test1 ชื่อชุดข้อมูลแสดงได้ดังตารางที่ 10 และตารางที่ 11

ตารางที่ 10 ชุดอ้างอิง 1

ชุดทดสอบ	ชื่อชุดข้อมูล
1	1aab_ref1, 1aboA_ref1, 1aho_ref1, 1csp_ref1, 1csy_ref1, 1dox_ref1, 1fjlA_ref1, 1fkj_ref1, 1fmb_ref1, 1hfh_ref1, 1hpi_ref1, 1idy_ref1, 1krn_ref1, 1pfc_ref1, 1plc_ref1, 1r69_ref1, 1tgxA_ref1, 1tvxA_ref1, 1ubi_ref1, 1wit_ref1, 1ycc_ref1, 2fxb_ref1, 2mhr_ref1, 2trx_ref1, 3cyr_ref1, 451c_ref1, 9rnt_ref1
2	1ad2_ref1, 1amk_ref1, 1ar5A_ref1, 1aym3_ref1, 1bbt3_ref1, 1ezm_ref1, 1gdoA_ref1, 1havA_ref1, 1ldg_ref1, 1led_ref1, 1mrj_ref1, 1pgt_ref1, 1pii_ref1, 1ppn_ref1, 1pysA_ref1, 1sbp_ref1, 1thm_ref1, 1tis_ref1, 1ton_ref1, 1uky_ref1, 1zin_ref1, 2cba_ref1, 2hsdA_ref1, 2pia_ref1, 3grs_ref1
3	1ac5_ref1, 1ad3_ref1, 1adj_ref1, 1ajsA_ref1, 1bgl_ref1, 1cpt_ref1, 1dlc_ref1, 1eft_ref1, 1fieA_ref1, 1gowA_ref1, 1gpb_ref1, 1gtr_ref1, 1lcf_ref1, 1lvl_ref1, 1pamA_ref1, 1ped_ref1, 1pkm_ref1, 1rthA_ref1, 1sesA_ref1, 1taq_ref1, 2ack_ref1, 2myr_ref1, 3pmg_ref1, 4enl_ref1, actin_ref1, arp_ref1, gal4_ref1,glg_ref1

ตารางที่ 11 ชุดอ้างอิง 2 - 5

ชุดอ้างอิง	ชื่อชุดข้อมูล
2	1aboA_ref2, 1ajsA_ref2, 1cpt_ref2, 1csy_ref2, 1havA_ref2, 1idy, 1idy_ref2, 1lvl_ref2, 1pamA_ref2, 1ped_ref2, 1r69_ref2, 1sbp_ref2, 1tgxA_ref2, 1tvxA_ref2, 1ubi_ref2, 1uky_ref2, 1wit_ref2, 2hsdA_ref2, 2myr_ref2, 2pia_ref2, 2trx_ref2, 3grs_ref2, 4enl_ref2, kinase_ref2
3	1ajsA_ref3, 1idy_ref3, 1pamA_ref3, 1ped_ref3, 1r69_ref3, 1ubi_ref3, 1uky_ref3, 1wit_ref3, 2myr_ref3, 2pia_ref3, 4enl_ref3, kinase_ref3
4	1ckaA_ref4, 1csp_ref4, 1dynA_ref4, 1lkl_ref4, 1mfa_ref4, 1pfc_ref4, 1pysA_ref4, 1vln_ref4, 1ycc_ref4, 2abk_ref4, kinase1_ref4, kinase2_ref4
5	1eft_ref5, 1ivy_ref5, 1pysA_ref5, 1qpg_ref5, 1thm1_ref5, 1thm2_ref5, 2cba_ref5, kinase1_ref5, kinase2_ref5, kinase3_ref5, s51_ref5, s52_ref5

ฐานข้อมูล PREFAB (Edgar, 2004) เวอร์ชั่น 4.0 ประกอบด้วยกลุ่มข้อมูลสำหรับเทียบเรียง 1682 ชุดข้อมูล โดยไม่ได้แบ่งเป็นกลุ่มอ้างอิงเช่นเดียวกับ BAliBASE แต่ละชุดข้อมูลประกอบด้วยคู่ลำดับข้อมูลโปรตีนที่มาจากการค้นหาลำดับข้อมูลที่เทียบเคียงกันได้ผ่านโปรแกรม PSI-BLAST จากฐานข้อมูลโปรตีน NCBI จำนวนลำดับข้อมูลในแต่ละชุดข้อมูลโดยเฉลี่ยมีจำนวน 45 ลำดับ

ฐานข้อมูล PREFAB เวอร์ชั่น 4.0 มาจาก http://www.drive5.com/muscle/ prefab.htm โดยมีขนาด 7.6 MB เมื่อขยายไฟล์ออกมาจะเป็นชุดทดสอบที่มีขนาด 31 MB ประกอบด้วยชุดข้อมูล 1682 ชุด ในงานวิจัยนี้จะแบ่งออกเป็นกลุ่มข้อมูลจำนวน 68 กลุ่มดังตารางที่ 12 – ตารางที่ 21

ตารางที่ 12 กลุ่มข้อมูลที่ 1-7

กลุ่มข้อมูล	ชื่อชุดข้อมูล
1	12asA_1atiA, 153l_1cnsA, 16pk_1qpg, 19hcA_1duwA, 1a02N_1a3qA, 1a04A_1dz3A, 1a04A_1ibjA, 1a04A_3chy, 1a0cA_1a0dA, 1a0cA_1bxbD, 1a0cA_1ubpC, 1a0fA_1hqoA, 1a0fA_1ljrA, 1a0fA_1pd21, 1a0hA_5hpgA, 1a0p_1a36A, 1a0p_1ae9A, 1a0tP_2mprA, 1a12A_1jtdB, 1a1z_1d2zA, 1a28A_2prgA, 1a28A_3erdA, 1a2vA_1ksiB, 1a36A_1a41, 1a3aA_1a6jA
2	1a3c_1tc1A, 1a3k_1c1lA, 1a3k_1lcl, 1a3qA_1gof, 1a44_1qouB, 1a49A_1dxeA, 1a49A_1pkyC, 1a49A_2tpsA, 1a4iA_1b0aA, 1a4iA_1ee9A, 1a53_1dvjA, 1a53_1hg3A, 1a53_1nsj, 1a53_1qo2A, 1a62_1mjc, 1a65A_1aozA, 1a65A_1nif, 1a6cA_1bmv1, 1a6dA_1ass, 1a6dA_1derA, 1a6jA_1hynP, 1a6l_2fdn, 1a6m_1ash, 1a6m_1cg5B, 1a6m_1ewaA
3	1a6m_1flp, 1a6m_1h97A, 1a6m_1hlb, 1a6m_1ithA, 1a6m_2fal, 1a6m_2gdm, 1a6m_2hbg, 1a6m_2vhbA, 1a6m_3sdhA, 1a6o_1buhA, 1a75A_1b8cB, 1a7s_1agjA, 1a7s_2cgaB, 1a7tA_1e5dA, 1a7tA_1smlA, 1a7w_1aoiA, 1a7w_1aoiB, 1a7w_1b67B, 1a7w_1bh9B, 1a7w_1tafA, 1a7w_1tafB, 1a8d_1wba, 1a8h_1qqtA, 1a8l_1hyuA, 1a8o_1qrjB
4	1a8rA_1fb1D, 1a9nA_1dceA, 1aba_1h75A, 1aba_1kte, 1aba_1qfnA, 1aba_3grx, 1abrB_1dqgA, 1abwA_1cqxA, 1ac5_1auoA, 1ac5_1cpy, 1ad3A_1bpwA, 1ad3A_1euhA, 1adeA_1dj3A, 1adjA_1atiA, 1adjA_1httD, 1ae9A_1a36A, 1ae9A_1aihA, 1aerA_1aerB, 1ag4_1amm, 1agdA_1bqsA, 1agdA_1iakA, 1agdB_1iakA, 1agjA_1cqqA, 1agqA_1vpfA, 1ah1_1cid
5	1ah1_1f97A, 1ah1_1kb5B, 1ah1_1qfoA, 1ah1_1tlk, 1ah1_1vcaA, 1ah1_2ncm, 1ah7_1ca1, 1ah9_1bkb, 1aihA_1a0p, 1air_1pcl, 1aisB_1guxB, 1aj6_1b63A, 1aj8A_1csh, 1ajqA_1cp9A, 1ajqB_1nedA, 1ajsA_1aam, 1ajsA_1ibjA, 1ak1_1hrkA, 1ako_1bix, 1alu_1bgc, 1alu_1cnt3, 1alu_1i1rB, 1aly_1d4vB, 1aly_1dyoA, 1aly_2tnfA
6	1am2_1at0, 1am7A_3lzt, 1amj_1zymA, 1amk_1hg3A, 1amoA_1ddgB, 1amoA_1ep2B, 1amoA_1fnc, 1amoA_1qfjA, 1amoA_1rcf, 1amp_1b8oA, 1amp_1cg2A, 1amp_1cx8B, 1amp_1xjo, 1amuA_1lci, 1an8_1aw7A, 1an8_1eu4A, 1an8_1f77B, 1an9A_1c0pA, 1aoa_1bkrA, 1aocA_1bndA, 1aoeA_1d1gA, 1aoeA_1vdrA, 1aoeA_3dfr, 1aohA_1anu, 1aohA_1g1kB
7	1aoiA_1bh9B, 1aoiA_1tafA, 1aoiB_1a7w, 1aoiB_1aoiD, 1aoiB_1jfiB, 1aoiB_1tafA, 1aoiC_1f66G, 1aoiD_1aoiB, 1aoiD_1jfiB, 1aoxA_1atzA, 1aoxA_1auq, 1aoxA_1ido, 1aoy_1b4aA, 1ap0_1dz1A, 1ap8_1ejhA, 1apyA_1ayyC, 1apyB_2gawD, 1aq0A_1eokA, 1aq0A_1ghsB,1aqb_1avgI,1aqb_1bebA,1aqb_1epaA,1aqb_1h91A,1aqt_1e79H,1aquA_1fmlB

ตารางที่ 13 กลุ่มข้อมูลที่ 8-14

กลุ่มข้อมูล	ชื่อชุดข้อมูล
8	1aqzA_1de3A, 1aqzA_9rnt, 1arb_2hrvA, 1ash_1a6m, 1ash_1flp, 1ash_1ithA, 1ash_2fal, 1at3A_1fl1A, 1at3A_1lay, 1atg_1ixh, 1atg_1mrp, 1atg_1wod, 1atiA_1g5hA, 1atiA_1qf6A, 1atlA_1bkcA, 1atzA_1auq, 1atzA_1ido, 1atzA_2scuB, 1au7A_2hddA, 1auiB_1dguA, 1auk_1fsu, 1auoA_1brt, 1auoA_1fj2B, 1auoA_1i6wA, 1auoA_1jfrA
9	1auq_1ido, 1auvA_1iow, 1auyA_1e57B, 1auyA_1qqp3, 1auz_1dciA, 1avaC_1wba, 1avgI_1epaA, 1avmA_3mdsA, 1avpA_1euvA, 1aw0_1cc8A, 1aw8B_1cr5A, 1awcA_1bc8C, 1awcA_2hts, 1awcA_2irfG, 1awd_1ayfA, 1awe_1btkA, 1awe_1dbhA, 1awe_1faoA, 1awe_1pls, 1ax4A_2dkb, 1ax4A_2tplA, 1ax8_1f6fA, 1ax8_1rcb, 1axiB_1bp3B, 1axiB_1j7vR
10	1axkA_2nlrA, 1aye_1qmuA, 1ayfA_1put, 1aym1_2hwf1, 1aym2_1hri2, 1aym2_1qqp2, 1aym3_1b35A, 1aym3_1ihmA, 1aym3_1rud3, 1az9_1bn5, 1az9_1c24A, 1azpA_1bf4A, 1azsA_1culB, 1azsA_1fx2A, 1b0nA_1r69, 1b0uA_1jj7A, 1b0uA_1skyE, 1b16A_1bdb, 1b16A_1bsvA, 1b16A_1e6wA, 1b16A_1fds, 1b16A_1he2A, 1b16A_1xel, 1b16A_1ybvA, 1b16A_3chy
11	1b20A_1rgeA, 1b25A_1aorA, 1b2pA_1msaA, 1b34B_1d3bA, 1b34B_1d3bB, 1b34B_1i8fA, 1b35A_1b35B, 1b35A_1ihmA, 1b35A_2mev1, 1b37B_1gpeA, 1b3aA_1hfgA, 1b3aA_1tvxA, 1b3jA_1agdA, 1b3mA_1b8sA, 1b4kA_1qmlA, 1b5fB_1smrA, 1b5l_1rmi, 1b64_1fjfJ, 1b6cB_1ia8A, 1b6e_1fm5A, 1b6e_1hyrB, 1b6g_1bn6A, 1b6g_1cqwA, 1b6rA_1gsoA, 1b6tA_1f9aA
12	1b7eA_1f3iA, 1b7fA_1b64, 1b7gO_1drw, 1b7gO_1gcuA, 1b7gO_1ofgA, 1b87A_1i21A, 1b8aA_1aszA, 1b8aA_1lylA, 1b8gA_1bjwA, 1b8gA_1bw0A, 1b8gA_1iaxA, 1b8oA_1cb0A, 1b8xA_1pd21, 1b8xA_1pgtA, 1b93A_1g8mA, 1b9hA_1ibjA, 1b9lA_1dhn, 1b9yC_1a0rP, 1b9zA_1a49A, 1ba1_1dkgD, 1ba1_1glcG, 1bag_1smd, 1bak_1btkA, 1bak_1btn, 1bak_1dynA
13	1bak_1faoA, 1bak_1mai, 1bb9_2semA, 1bbhA_1jafB, 1bbhA_2ccyA, 1bbzA_1ckaA, 1bbzA_1gcpA, 1bcfA_1dpsA, 1bcfA_1eumA, 1bcfA_1qghA, 1bcfA_2fha, 1bcpA_1lt3A, 1bcpD_1prtF, 1bcpD_1tiiD, 1bcpD_3chbD, 1bd3A_1dkuA, 1bd3A_1nulA, 1bdb_1cydA, 1bdb_1e6wA, 1bdb_1e7wA, 1bdb_1hdcA, 1bdb_1oaa, 1bdb_1ybvA, 1bdo_1ghj, 1bdo_1iyu
14	1bdyA_1rlw, 1be3A_1be3B, 1be3A_1ezvA, 1be3A_1ezvB, 1be3A_1hr6A, 1be3B_1bccB, 1be3B_1ezvB, 1be3B_1hr6A, 1be3C_3bccC, 1bebA_1bj7, 1bebA_1epaA, 1bebA_1mup, 1bec_1iakB, 1bec_1tvdA, 1bec_3mcg1, 1befA_1jxpA, 1befA_2hrvA, 1bf2_1ehaA, 1bfd_1d4oA, 1bfg_1afcH, 1bfg_1jlyA, 1bg2_1f9tA, 1bglA_1bhgA, 1bh9B_1jfiB,1bh9B_1tafA

ตารางที่ 14 กลุ่มข้อมูลที่ 15 – 21

กลุ่มข้อมูล	ชื่อชุดข้อมูล
15	1bhe_1czfA, 1bhe_1rmg, 1bhgA_1qnoA, 1bhtA_5hpgA, 1bi0_1fx7B, 1bi0_1smtA, 1bi5A_1dd8A, 1bi5A_1hn9A, 1bif_1qhfA, 1bj4A_1bjwA, 1bj7_1epaA, 1bj7_1obpA, 1bjnA_1bt4A, 1bjwA_1bw0A, 1bjwA_1cl2A, 1bjwA_1d2fA, 1bjx_1qgvA, 1bk5A_1ialA, 1bk7A_1bolA, 1bkjA_1f5vA, 1bkjA_1vfrA, 1bkpA_1tlcA, 1bkrA_1aoa, 1bkrA_1bhdA, 1bli_1jdc
16	1bm9A_1cf7B, 1bmdA_1b8pA, 1bmdA_1ceqA, 1bmdA_1hyhA, 1bmdA_2cmd, 1bmfG_1mabG, 1bmlC_1qqrC, 1bn5_1c24A, 1bn6A_1cv2A, 1bndA_1aocA, 1bndA_1nt3A, 1bo4A_1cjwA, 1bo4A_1i21A, 1bob_1qsmA, 1bob_1yghA, 1bolA_1bk7A, 1booA_1eg2A, 1boy_1a21A, 1bp7A_1af5, 1bpi_1brcI, 1bpv_1bquA, 1bpv_2fnbA, 1bpwA_1euhA, 1bpwA_1eyyA, 1bqcA_7a3hA
17	1bqg_1ec7B, 1bqg_1mucA, 1bqg_1qumA, 1bqg_2mnr, 1bqk_2mtaA, 1bqsA_1cvsC, 1bquA_1bj8, 1br9_1d2bA, 1br9_1jc7A, 1brmA_1drw, 1brmA_1gcuA, 1brmA_1qrrA, 1brt_1i6wA, 1brt_1jfrA, 1brt_1qlwA, 1bs0A_1b8gA, 1bs0A_1elqA, 1bs9_1cex, 1bs9_1ei9A, 1bsvA_1bwsA, 1bsvA_1bxkA, 1bsvA_1he2A, 1bsvA_1xel, 1bt3A_1lla, 1btkA_1btn
18	1btkA_1dynA, 1btkA_1faoA, 1btkA_1pls, 1btkA_1rrpB, 1btl_1e25A, 1btl_1ei5A, 1btl_1omeB, 1btl_1skf, 1btn_1dynA, 1btn_1mai, 1btn_1pls, 1btn_1qqgA, 1btn_1rrpB, 1bu2A_1g3nC, 1bu2A_1vin, 1bu7A_1ea1A, 1burA_1rusB, 1bv1_1em2A, 1bv1_3pmgA, 1bvwA_1dysA, 1bvwA_1tml, 1bvzA_1cxlA, 1bvzA_1ehaA, 1bvzA_1smaA, 1bvzA_7taa
19	1bw0A_1d2fA, 1bw9A_1qorA, 1bw9A_1qp8A, 1bx4A_1dgyA, 1bx4A_1rkd, 1bxkA_1db3A, 1bxkA_1eq2A, 1bxkA_1he2A, 1bxkA_1xel, 1byb_1fa2A, 1byfA_1esl, 1byfA_1fm5A, 1byfA_1htn, 1byfA_1rtm1, 1byi_1eg7A, 1byi_1nksA, 1bykA_1gca, 1bykA_1qpzA, 1bykA_1tlfA, 1bykA_2dri, 1bylA_1ecsA, 1bylA_1qipA, 1bylA_1qtoA, 1byuA_1ctqA, 1bywA_1drmA
20	1bywA_1g28D, 1bywA_3pyp, 1c05A_1dm9A, 1c0aA_1adjA, 1c0aA_1lylA, 1c0nA_1ecxA, 1c0nA_1elqA, 1c20A_1ig6A, 1c24A_1chmA, 1c24A_1xgmA, 1c3cA_1f1oA, 1c3d_1qsjD, 1c3oB_1qdlB, 1c3wA_1brd, 1c3wA_1jgjA, 1c4rA_1qu0A, 1c4rA_1sacA, 1c4xA_1jfrA, 1c53_2dvh,1c8bA_1cfzA,1c8oA_1hleA,1c8oA_1ovaA,1c8oA_1sek,1c8pA_1cto,1c9kB_1cbuC
21	1c9kB_1d2nA, 1c9kB_1esc, 1c9kB_1g5rA, 1c9kB_2dhqA, 1ca1_1ah7, 1caxB_1dgwA, 1cb8A_1eguA, 1cczA_1fltX, 1cczA_1hngB, 1cczA_1qfoA, 1cczA_1tlk, 1cd31_1cd32, 1cd8_1hxmD, 1cd8_1tvdA, 1cd8_1wit, 1cdkA_1apmE, 1cdkA_1csn, 1cdkA_1hcl, 1cdkA_1ia8A,1ceo_1cz1A,1ceo_1edg,1ceqA_1hyhA,1ceqA_2cmd, 1ceqA_3ldh,1cewI_1eqkA

ตารางที่ 15 กลุ่มข้อมูลที่ 22 - 28

กลุ่มข้อมูล	ชื่อชุดข้อมูล
22	1cf7A_1cf7B, 1cf9A_1g2iA, 1cg2A_1cx8B, 1cg2A_1xjo, 1cg5B_1flp, 1cg5B_1gcwC, 1cg5B_1hlb, 1cg5B_1ithA, 1cg5B_2fal, 1cg5B_2hbg, 1cg5B_3sdhA, 1chkA_1l92, 1chmA_1ihoA,1cid_1kb5B,1cid_1qa9A,1cipA_1ctqA,1cipA_1hurA,1cjwA_1i21A, 1cjxA_1qipA,1ckeA_1dekA,1ckuA_1eytA,1ckv_1g10A,1cl2A_1qgnH,1clc_1tf4A,1clqA_1nozB
23	1cmbA_1mjoB, 1cmoA_1hjbC, 1cnt3_1bgc, 1cnt3_1cnt2, 1cnt3_1lki, 1cnv_1llo, 1cnzA_1iso, 1cof_1cnuA, 1cozA_1ihoA, 1cozA_2ts1, 1cp2A_1eg7A, 1cp2A_2nipA, 1cpcA_1allA, 1cpcA_2hbg, 1cpq_256bA, 1cpt_1f4tA, 1cqkA_1fc1B, 1cqkA_1hxmD, 1cqkA_1iakA, 1cqqA_1havA, 1cqxA_1dlyA, 1cr1A_2reb, 1cr5A_1e32A, 1cr5A_1qcsA, 1cr6B_1ehyA
24	1cs6A_1fltX, 1cs6A_1tit, 1cs8A_1cv8, 1cs8A_1qmyA, 1csh_4ctsA, 1csn_1ckiA, 1ct9A_1gdoA, 1ctj_1e29A, 1ctj_2mtaC, 1ctn_1d2kA, 1cto_1bpv, 1ctqA_1am4D, 1ctqA_1d4aA, 1ctqA_1d5cA, 1cv8_1dkiA, 1cviA_1rpt, 1cvjB_2u1a, 1cvl_1tca, 1cvsC_1ev2E, 1cvsC_1fltX, 1cvsC_1qfoA, 1cvsC_1wit, 1cvsC_2ncm, 1cwpA_1f15B, 1cwvA_1qfhA
25	1cwyA_1bag, 1cx8B_1xjo, 1cxlA_1cgu, 1cxlA_7taa, 1cxqA_1b9dA, 1cxqA_1c0mC, 1cy5A_3ygsP, 1cydA_1e6wA, 1cydA_1e7wA, 1cydA_1enp, 1cydA_1eny, 1cydA_1fds, 1cydA_1fjhB, 1cydA_1gegG, 1cydA_1h5qL, 1cydA_1he2A, 1cydA_1oaa, 1cydA_1ybvA, 1cydA_2ae2B, 1cyo_1cxyA, 1cyx_1plc, 1cyx_2occB, 1czfA_1rmg, 1cztA_1eut, 1cztA_1ulo
26	1d0nA_2vil, 1d1dA_1qrjB, 1d1gA_1vdrA, 1d1gA_3dfr, 1d1rA_2if1, 1d2fA_1c7nF, 1d2fA_1c7oH, 1d2iA_1es8A, 1d2kA_1ctn, 1d2kA_1e9lA, 1d2mA_1pjr, 1d2nA_1g6oA, 1d2sA_1dykA, 1d2sA_1sacA, 1d2tA_1eoiA, 1d2zA_1d2zB, 1d2zA_1fadA, 1d2zA_1ngr, 1d2zB_1fadA, 1d2zB_1ngr, 1d3bA_1d3bB, 1d3bA_1i8fA, 1d3bB_1i8fA, 1d3gA_2dorA, 1d4oA_1j8fA
27	1d4tA_1csyA, 1d4tA_1jwoA, 1d4vA_1extA, 1d4vB_2tnfA, 1d5rA_1fpzA, 1d5rA_1vhrA, 1d5yA_1bl0A, 1d6aB_1dm0A, 1d6jA_1f48A, 1d6jA_1nksA, 1d6jA_1nstA, 1d6jA_3tmkA, 1d7yA_1nhp, 1d8jA_1d8kA, 1d9cA_1ekuB, 1d9eA_1dvjA, 1d9eA_1fx6B, 1d9eA_1qr7A, 1daaA_1ekfA, 1daaA_1et0A, 1daaA_1i1mC, 1dapA_1dssG, 1dapA_1hyhA, 1dar_1efcA, 1db1A_2lbd
28	1db1A_3erdA, 1db3A_1bg6, 1db3A_1fds, 1db3A_1he2A, 1db3A_1qrrA, 1dbtA_1dvjA, 1dbwA_1tmy, 1dbwA_3chy, 1dceB_1ft1B, 1dcfA_1dbwA, 1dcfA_3chy, 1dciA_1ef8A, 1dciA_2dubF,1dcs_1bk0,1ddbA_2bidA,1ddzA_1i6pA,1debA_1fe6A,1dekA_1gky, 1dekA_3tmkA,1deoA_1wab,1derA_1a6dA,1dfuP_1feuA,1dg9A_1d2aB,1dg9A_1jfvA, 1dgnA_3crd

ตารางที่ 16 กลุ่มข้อมูลที่ 29 - 35

กลุ่มข้อมูล	ชื่อชุดข้อมูล
29	1dhn_1b9lA, 1dhpA_1nal3, 1dhpA_2tpsA, 1dhr_1ybvA, 1dhx_1ruxA, 1di0A_1rvv1, 1di1A_1ps1A, 1di6A_1ihcA, 1din_1ei9A, 1djxB_1rlw, 1dk4A_1imbA, 1dk4A_1inp, 1dk5A_1aow, 1dkiA_1cv8, 1dklA_1qfxA, 1dkuA_1qb7A, 1dl2A_1hcuA, 1dl5A_1dusA, 1dl5A_1fbnA, 1dl5A_1g6q2, 1dl5A_1vid, 1dl5A_1xvaA, 1dleA_1qq4A, 1dleA_1svpA, 1dlyA_1dlwA
30	1dlyA_2gdm, 1dlyA_2vhbA, 1dmgA_1ffkC, 1dmhA_3pchA, 1dmhA_3pchM, 1dn1A_1dcfA, 1do0A_1e32A, 1do0A_1g8pA, 1dokA_1tvxA, 1dosA_1zen, 1dpe_1jevA, 1dpsA_1qghA, 1dptA_1ca7A, 1dptA_1otgA, 1dpuA_1hstA, 1dpuA_1qbjA, 1dpuA_1smtA, 1dqaA_1qaxA, 1dqnA_1tc1A, 1dqrA_1iatA, 1dqrA_2pgi, 1dquA_1pymA, 1dqwA_1dvjA, 1dqyA_1ei9A, 1dqyA_1ivyA
31	1dr9A_1i85A, 1drmA_1ew0A, 1drmA_3pyp, 1drw_1gcuA, 1drw_1id1A, 1drw_1ofgA, 1dssG_1ceqA, 1dssG_1he2A, 1dt4A_1vih, 1dt6A_1bu7A, 1dtyA_2dkb, 1dtyA_2gsaA, 1dulA_1hq1A, 1dun_1dupA, 1dusA_1i4wA, 1dusA_2dpmA, 1dvjA_1fwrA, 1dvjA_1ho1A, 1dvjA_2tpsA,1dvpA_1elkA,1dwnA_1msc,1dwnA_1qbeA,1dxy_1gdhA,1dxy_1psdA,1dxy_2dldA
32	1dykA_1a3k, 1dynA_1mai, 1dynA_1pls, 1dyoA_1ulo, 1dz1A_1ap0, 1dz3A_1iow, 1dz3A_3chy, 1e0cA_1rhs, 1e15A_1ctn, 1e15A_1e9lA, 1e20A_1g5qA, 1e2tA_1f13A, 1e2xA_1qbjA, 1e32A_1e69B, 1e3jA_1pedA, 1e3jA_1qr6A, 1e4fT_1g99A, 1e54A_2omf, 1e54A_3prn, 1e5dA_5nul, 1e69B_1f2tA, 1e69B_1f2tB, 1e69B_1gajA, 1e69B_1qhlA, 1e6bA_1aw9
33	1e6bA_1eemA, 1e6bA_1gnwA, 1e6wA_1e7wA, 1e6wA_1enp, 1e6wA_1fds, 1e6wA_1fjhB, 1e6wA_1oaa, 1e6wA_1ybvA, 1e70M_1qvbA, 1e70M_1tr1B, 1e7wA_1ybvA, 1e8xA_1e8zA, 1ea1A_1bu7A, 1eaf_1c4tA, 1eaf_3cla, 1eagA_1smrA, 1eagA_2er7E, 1eagA_2rmpA, 1eagA_3pep,1eaiC_1ate,1ebmA_1mun,1ebuA_1gcuA,1ecfB_1gdoA,1ecpA_1a2zA,1ecpA_1k3fA
34	1ecxA_1elqA, 1ecxA_2gsaA, 1eduA_1dvpA, 1eduA_1hg5A, 1eemA_1gnwA, 1eemA_1hqoA, 1eerA_1f6fA, 1eerB_1iarB, 1ef1A_1a5r, 1ef1A_1gg3C, 1efdN_1qguA, 1efuB_1tfe, 1efvA_1efpA, 1efvA_1mjhA, 1eg7A_1hyqA, 1ehaA_1uok, 1ehyA_1qo7A, 1ehyA_2bce, 1ei5A_3pte, 1ei9A_1cr6B, 1ei9A_1cvl, 1eia_1d1dA, 1eiwA_1egaA, 1ej0A_1dhr, 1ej0A_1fbnA
35	1ejeA_1axj, 1ejeA_1i0rA, 1ejhA_1ap8, 1ekeA_2rn2, 1ekfA_1et0A, 1ekrA_1mla, 1elqA_2dkb, 1ema_1ggxA, 1emsA_1erzA, 1enp_1ej0A, 1enp_1eny, 1enp_1qsgA, 1enp_1ybvA, 1eny_1ybvA, 1eokA_1d2kA, 1ep2B_1fnc, 1epaA_1bj7, 1epaA_1mup, 1epaA_1np1A, 1eq2A_1xel, 1eq3A_1pinA, 1eqkA_1cewI, 1eqkA_1molA, 1erv_1thx, 1erzA_1f89A

ตารางที่ 17 กลุ่มข้อมูลที่ 36 - 42

กลุ่มข้อมูล	ชื่อชุดข้อมูล
36	1esc_1eny, 1esl_1h8uA, 1esl_1qddA, 1esl_1qo3C, 1esl_1rtm1, 1etb1_1gkeC, 1eteA_1hmcB, 1etpA_1fcdC, 1eu8A_4mbp, 1euhA_1eyyA, 1eumA_1qghA, 1eumA_2fha, 1eut_2sli, 1eut_3sil, 1evhA_1i7aC, 1evhA_1qc6A, 1evsA_1lki, 1ew4A_1dlxA, 1ewaA_1flp, 1ewaA_1ithA, 1ewaA_2fal, 1ewxA_1gp1A, 1ewxA_1jfuA, 1ex1A_3chy, 1exg_1xbd
37	1extA_1d4vA, 1exzA_1hmcB, 1eyrA_1fwyA, 1eyrA_1ga8A, 1eyrA_1h7eA, 1eyyA_1ad3A, 1ezvB_1be3B, 1f13A_1g0dA, 1f2dA_1oasA, 1f2nA_1smvA, 1f2nA_2tbvA, 1f37A_1b9yC, 1f3yA_1g0sA, 1f4tA_1egyA, 1f4tA_1phd, 1f4tA_1rom, 1f53A_1g6eA, 1f5qB_1guxB, 1f5wA_1neu, 1f5xA_1foeA, 1f7cA_1pbwA, 1f7cA_1tx4A, 1f94A_1cds, 1f94A_3ebx, 1f97A_1qfoA
38	1f9aA_1hybA, 1faoA_1dynA, 1faoA_1ef1A, 1faoA_1fhoA, 1faoA_1fhxA, 1faoA_1mai, 1faoA_1rrpB, 1fc6A_1pdr, 1fcdC_1e29A, 1fchA_1hxiA, 1fdo_1dmr, 1fdr_1a8p, 1fdr_1fnc, 1fdr_1qfjA, 1fds_1he2A, 1fepA_1by5A, 1ff9A_1gcuA, 1ff9A_1id1A, 1ff9A_2scuA, 1ffkE_1e7kB, 1fftC_1ocrC, 1fggA_1fgxA, 1fggA_1fr9A, 1fggA_1ga8A, 1fgkA_1csn
39	1fgkA_1fgiB, 1fgkA_1ia8A, 1fgxA_1foaA, 1fgxA_1g8oA, 1fgxA_1ga8A, 1fh6A_1fh6G, 1fhoA_1pls, 1fhoA_1qqgA, 1fhuA_1mucA, 1fi2A_1caxB, 1fi2A_1dzrA, 1fit_1kpf, 1fj7A_1cvjB, 1fjcA_1cvjB, 1fjfL_1fjfQ, 1fjhB_1eq2A, 1fjjA_1qouA, 1fknA_1hvc, 1fknA_1lywA, 1fknA_1smrA, 1fknA_2rmpA, 1floA_1ae9A, 1floA_1aihA, 1flp_1h97A, 1flp_1hlb
40	1flp_1ithA, 1flp_2fal, 1flp_3sdhA, 1fltX_1qa9A, 1fltX_1tit, 1fltX_2ncm, 1fm5A_1htn, 1fm5A_1qddA, 1fm5A_1qo3C, 1fmb_1b11A, 1fmb_2hpeA, 1fmb_2rmpA, 1fmk_1hcl, 1fmtA_2gar, 1fnc_1fb3B, 1fnc_1fdr, 1fnc_1qfjA, 1fnc_2pia, 1fo5A_1hyuA, 1fofA_1e25A, 1fpqA_1dusA, 1fpqA_1fp1D, 1fpzA_1pty, 1fpzA_1vhrA, 1fqtA_1rfs
41	1fqtA_1rie, 1fqvD_1vcbB, 1fr9A_1fgxA, 1fr9A_1h7eA, 1fr9A_1i52A, 1frb_1qrqA, 1frpA_1dk4A, 1fshA_1hstA, 1ft9A_1qbjA, 1ft9A_2cgpA, 1fukA_1qvaA, 1fvaA_1qd1A, 1fvkA_1bed, 1fvzA_1bv1, 1fwkA_1fi4A, 1fxd_1vjw, 1fxkA_1fxkC, 1fxkC_1fxkA, 1fzgD_2eboA,1g0rA_1qg8A,1g10A_1ckv,1g1eB_1e91A,1g24A_1qs1A,1g4uS_1ytw, 1g4uS_2shpA
42	1g55A_6mhtA, 1g5rA_1g6oA, 1g5rA_1g8yA, 1g5rA_2reb, 1g61A_1g62A, 1g6gA_1qu5A, 1g6q2_1f3lA, 1g6q2_1vid, 1g6q2_2admA, 1g7eA_1prxA, 1g8fA_1cozA, 1g8fA_1f9aA, 1g8jA_2napA, 1g8jB_1fqtA, 1g8jB_1rfs, 1g8jB_1rie, 1g8oA_1qg8A, 1g8qA_1g8qB, 1g8yA_1g5rA,1ga3A_1rcb,1ga8A_1fgxA,1ga8A_1g8oA,1ga8A_1qg8A,1gakA_1lis,1gaxA_1ile

ตารางที่ 18 กลุ่มข้อมูลที่ 43 - 49

กลุ่มข้อมูล	ชื่อชุดข้อมูล
43	1gc1G_1g9nG, 1gca_1qpzA, 1gca_1tlfA, 1gca_2dri, 1gceA_3pte, 1gci_1sud, 1gdhA_1psdA, 1gefA_1hh1A, 1gen_1hxn, 1gen_1pex, 1ggqA_1f1mC, 1ggxA_1h4uA, 1ghj_1iyu, 1ghqB_1e5gA, 1gkxA_1id0A, 1gky_1nksA, 1gky_1qhsA, 1gky_3tmkA, 1gnwA_1ljrA, 1gnwA_1pgtA, 1gp1A_1erv, 1gp1A_1jfuA, 1gpl_1bu8A, 1gpmA_1qdlB, 1gr2A_1ii5A
44	1gsa_2hgsA, 1gtxA_2dkb, 1gtxA_2gsaA, 1h70A_1jdw, 1h75A_1kte, 1h7wA_2dorA, 1h8cA_1c1yB, 1h8cA_1ubi, 1h8uA_1htn, 1h8uA_1qddA, 1h8uA_1rtm1, 1h97A_1ithA, 1h97A_2fal, 1h97A_2hbg, 1h97A_3sdhA, 1h9jA_1g291, 1ha1_1f9fA, 1ha1_1hd0A, 1ha1_1jmtA, 1ha1_2u2fA, 1han_1cjxA, 1han_1eirA, 1han_1jc4A, 1havA_1cqqA, 1hcl_1csn
45	1hd2A_1qq2A, 1hdmB_1b3jA, 1he2A_1hyhA, 1he2A_1qrrA, 1hg3A_1nsj, 1hg3A_2tpsA, 1hgeB_1flcB, 1hjp_1bvsA, 1hlb_1dlyA, 1hlb_2fal, 1hmcB_1jli, 1hmt_1lfo, 1hn9A_1bi5A, 1hnoA_1nzyA, 1ho1A_1nsj, 1hp4A_1qba, 1hqoA_1g6wA, 1hqoA_1gnwA, 1hqoA_1ljrA, 1hqoA_1pgtA, 1hs6A_1hyt, 1hssA_1bea, 1hssA_1hyp, 1htn_1ixxB, 1htn_1rdo1
46	1htn_1rtm1, 1huuA_1hueA, 1huw_1a22A, 1hv8A_1c9kB, 1hwyA_1b26F, 1hwyA_1k89, 1hx1B_2a3dA, 1hxmA_1tvdA, 1hxmD_1tvdA, 1hyhA_1hygB, 1hyhA_1ldnF, 1hyhA_2cmd, 1hyhA_3ldh, 1hyt_1ezm, 1hyt_1hs6A, 1hyuA_1tde, 1i1jA_1ckaA, 1i21A_1i12D, 1i21A_1nmtA, 1i21A_1yghA, 1i4wA_1vid, 1i4wA_2dpmA, 1i5nA_1qspA, 1i69A_1pda, 1i6wA_1jfrA
47	1i6wA_1tca, 1iae_1kuh, 1iakA_1hdmA, 1iakA_1tlk, 1iarB_1cto, 1ibjA_1bs0A, 1ibjA_1cs1A, 1ibzA_1kcw, 1ibzA_1plc, 1ibzA_1rcy, 1iciA_1nbaA, 1id1A_1eq2A, 1id1A_1ofgA, 1ifa_1huw, 1ifa_1rmi, 1ig0A_1ig3A, 1igtB_1fe8J, 1ihmA_1a6cA, 1ihp_1bif, 1ihp_1qfxA, 1ii5A_1lst, 1ijqA_1qlgA, 1im3D_1tvdA, 1imbA_1jp4A, 1imbA_1qgxA
48	1iow_1ehiB, 1iow_1gpmA, 1iq3A_2cblA, 1irp_2irtA, 1iso_1cnzA, 1itbB_1kb5B, 1itbB_1tit, 1itbB_1tlk, 1itbB_1wit, 1itbB_2ncm, 1ithA_1cg5B, 1ithA_2fal, 1ithA_3sdhA, 1ixh_1pot, 1ixxB_1eggA, 1iyu_1dczA, 1iyu_1ghj, 1j9yA_1oyc, 1jdc_1amy, 1jfiB_1tafA, 1jfiB_1tafB, 1jfuA_1prxA, 1jhjA_1xnaA, 1jhnA_1led, 1jk0A_1xikA
49	1jkmA_1evqA, 1jmtA_2u2fA, 1jn5A_1ounA, 1jn5B_1opy, 1jotA_1c3nA, 1jotA_1vmoA, 1jp4A_1qgxA, 1jxpA_1svpA, 1k89_1bw9A, 1kb5B_1fo0B, 1kb5B_1nkr, 1kb5B_1qa9A, 1kb5B_1qfoA, 1kb5B_1tlk, 1kb5B_1vcaA, 1kb5B_2ncm, 1kit_2sli, 1kjs_1c5a, 1krs_1b8aA, 1krs_1lylA, 1ksiB_1spuB, 1kuh_1hfc, 1kuh_1iae, 1kum_1cxlA, 1kwaA_1kwaB

ตารางที่ 19 กลุ่มข้อมูลที่ 50 - 56

กลุ่มข้อมูล	ชื่อชุดข้อมูล
50	1kwaA_1pdr, 1kwaA_1qauA, 1lam_1ecpA, 1larA_1d5rA, 1larA_2shpA, 1lay_1at3A, 1lbd_2lbd, 1lckA_1griA, 1lcl_1sacA, 1lea_1qbjA, 1lfb_2hddA, 1lfdA_2rgf, 1lis_1gakA, 1ljrA_1pd21, 1lki_1exzA, 1lla_1hc2, 1lpbA_1pcn, 1lst_1nnt, 1lt3A_1bcpA, 1lt3A_1xtcA, 1lvk_1br2F, 1lvl_1geuB, 1lxa_1qreA, 1lxa_2xat, 1mai_1pls
51	1maz_1f16A, 1maz_1g5jA, 1maz_2bidA, 1mfa_1neu, 1mfmA_1xsoB, 1mfmA_2apsB, 1mgtA_1sfe, 1mh1_2cmd, 1mnmA_1c7uB, 1moq_1bvyF, 1mpgA_1mun, 1mpgA_2abk, 1mrj_1qcjB, 1mroA_1mroB, 1mroB_1e6vE, 1mrp_1d9yA, 1msc_1qbeA, 1mspA_1qpxA, 1mtyB_1mhyB, 1mtyB_1qq8A, 1mucA_2mnr, 1mugA_1bjt, 1mugA_3eugA, 1mun_1ebmA, 1mup_1bj7
52	1mup_1qftA, 1nar_1d2kA, 1nbaA_1yacA, 1nbcA_1g43A, 1nbcA_1tf4A, 1ndoA_1rie, 1nedA_1g3iJ,1nedA_1pmaP, 1neu_1kacB, 1neu_1tvdA, 1nf1A_1wer, 1ng1_1j8yF, 1nksA_1nstA,1nksA_1qhiA,1nksA_1shkA,1nmtA_1qsmA,1nnt_1dsn,1nox_1bkjA,1npk_1nueD, 1nseA_1qomB, 1nsgB_256bA,1nukA_1ulo,1nwpA_1qhqA,1nzyA_1dubE,1oaa_1ybvA
53	1oasA_1tdj, 1obpA_1mup, 1ocrC_1qleC, 1opc_1qqiA, 1opy_3stdA, 1orc_4croE, 1otcB_1quqA, 1ovaA_1sek, 1ovaA_2achA, 1oyc_2tmdA, 1p35A_1i4eA, 1pauB_1qduL, 1pcl_1czfA, 1pcl_1qcxA, 1pd21_1pgtA, 1pdgA_1vpfA, 1pdr_1kwaA, 1pdr_2pdzA, 1pedA_1bxzD, 1pedA_2ohxA, 1pgtA_1gsdB, 1pgtA_1gumH, 1pii_1a53, 1pjr_1hv8A, 1pjr_1qvaA
54	1pjr_1uaaA, 1plc_1rcy, 1plc_2occB, 1plq_1axcE, 1plq_1ge8A, 1plq_2polA, 1pls_1foeA, 1pls_1rrpB, 1pmaA_1nedA, 1pmaA_1pmaP, 1pne_3nul, 1pot_1lst, 1poxA_1bfd, 1preA_1bcpB, 1prs_1g6eA, 1prtF_2bosA, 1prtF_3chbD, 1pscA_1b5tA, 1psrA_1cll, 1psrA_1qlsA, 1psrA_4icb, 1ptf_2hid, 1pty_1ytw, 1pty_2shpA, 1pvc4_1ar94
55	1pvl_7ahlA, 1qa9A_1qfoA, 1qa9A_1tcrA, 1qa9A_1tit, 1qa9A_1tlk, 1qa9A_1wit, 1qa9A_2ncm, 1qazA_1cem, 1qbhA_1g73D, 1qbjA_2cgpA, 1qbzC_2ezoA, 1qdlB_1i7qB, 1qdlB_1tmy, 1qfjA_1fnc, 1qfoA_1tit, 1qfoA_1tvdA, 1qfoA_1vcaA, 1qfoA_1wit, 1qfoA_2fcbA, 1qfoA_2ncm, 1qghA_1bcfA, 1qghA_2fha, 1qguB_3csuA, 1qgvA_1b9yC, 1qgvA_1bjx
56	1qgwA_1qgwB, 1qgwC_1allA, 1qgwC_1cpcL, 1qhaA_1g99A, 1qhiA_1qhsA, 1qhqA_1bqk, 1qhsA_3tmkA, 1qhuA_1ck7A, 1qhvA_1nobE, 1qipA_1cjxA, 1qipA_1fa5B, 1qj2A_1ffvA, 1qj2B_1ffvE, 1qj2C_1ffuF, 1qj4A_1ei9A, 1qjdA_1qo8D, 1qk3A_1gca, 1qk3A_1nulA, 1qksA_1n50A,1qlaA_1fumA,1qlaB_1c1yB,1qndA_1c44A,1qniA_1ibzA,1qnoA_1eceA, 1qnxA_1cfe

ตารางที่ 20 กลุ่มข้อมูลที่ 57 - 63

กลุ่มข้อมูล	ชื่อชุดข้อมูล
57	1qo0D_1ybvA, 1qo2A_1thfD, 1qo3C_1rtm1, 1qorA_2ohxA, 1qoxN_1e70M, 1qpxA_1bf8, 1qpzA_1tlfA, 1qpzA_2dri, 1qq4A_1arb, 1qq4A_1svpA, 1qq4A_2hrvA, 1qq8A_1j77A, 1qqp1_2mev1, 1qqp2_1pov1, 1qqp3_1mec3, 1qqp3_1pov1, 1qqp3_2mev1, 1qr4B_1f97A, 1qrjB_1ak4C, 1qs1A_1g24A, 1qsmA_1yghA, 1qtrA_1hlgA, 1qtrA_1qfmA, 1qu5A_1ygs, 1qu9A_1qd9C
58	1qupA_1yaiA, 1quqA_1quqB, 1r2fA_1xikA, 1r2fA_1xsm, 1r69_1b0nA, 1rcb_1lki, 1rcb_3inkC, 1rcf_1d04A, 1rcf_5nul, 1rcy_1qhqA, 1rcy_2cuaA, 1rec_1bjfB, 1rfs_1rie, 1rhoA_1ft3A, 1rhs_1e0cA, 1rie_1ezvE, 1rom_1jipA, 1rpxA_2dorA, 1rsy_1a25B, 1rthA_2rn2, 1rtm1_1qo3C, 1rzl_1fk3A, 1sacA_1b09C, 1sat_1hfc, 1sayA_1hzzB
59	1sbp_1wod, 1scjB_1b64, 1seiA_1qd7G, 1sek_1dvmD, 1sek_1psi, 1sfe_1eh7A, 1shcA_1x11A, 1shkA_1ng1, 1shkA_3tmkA, 1shsA_1ejfA, 1skyB_1a5t, 1skyE_1e32A, 1skyE_1skyB, 1sluA_1fi8E, 1smrA_2rmpA, 1smvA_1c8nC, 1smvA_2tbvA, 1smvA_4sbvC, 1sro_1mjc, 1stfI_1cewI, 1sur_2nsyA, 1svpA_5ptp, 1svy_1svq, 1svy_2vil, 1swuA_2aviA
60	1t1dA_1buoA, 1taq_5ktqA, 1taxA_1xas, 1tbgE_1tbgF, 1tc1A_1qk3A, 1tcrA_1ah1, 1tcrA_1qsfD, 1tfb_1c9bA, 1tfe_1efuB, 1tgxA_1cdtA, 1theB_1ef7A, 1theB_1qmyA, 1tig_2ifeA, 1tit_1wit, 1tit_2ncm, 1tlfA_2dri, 1tlk_1wit, 1tlk_1wwcA, 1trb_1f6mF, 1tul_1dun, 1tvdA_1bzqN, 1tvdA_1ivlB, 1tvdA_1qfwL, 1tvxA_1a15A, 1tx4A_1f7cA
61	1tyfA_1nzyA, 1u2fA_1fjcA, 1u9aA_1c4zD, 1u9aA_1i7kB, 1ubpC_1a4mA, 1uch_1cmxA, 1ulo_1cx1A, 1uok_1bvzA, 1urnA_1u2fA, 1urnA_2u1a, 1uteA_1qhwA, 1uteA_4kbpA, 1vcaA_2fcbA, 1vdrA_3dfr, 1vfrA_1icuB, 1vfrA_1nox, 1vfyA_1hyiA, 1vid_2admA, 1vid_2dpmA, 1vid_3mag, 1vpfA_1pdgA, 1vpsA_1dzlA, 1vsgA_2vsgA, 1wab_1deoA, 1wdcC_1br4B
62	1wer_1nf1A, 1wgjA_2prd, 1wit_2ncm, 1wwcA_1he7A, 1x11A_2nmbA, 1xbd_1ayoA, 1xel_1i3nA, 1xikA_1qghA, 1xikA_1xsm, 1xvaA_2dpmA, 1xvaA_2ercA, 1xwl_1t7pA, 1xwl_2kfzA, 1xxaA_1b4aA, 1ybvA_1g6q2, 1ybvA_2ae2B, 1yer_1b63A, 1yge_1byt, 1yghA_1bob,1yghA_1i21A,1yghA_1qsrA,1ygs_1devA,1ytbA_1qnaB,1zin_5ukdA, 1zpdA_1qpbB
63	1zxq_1ic1A, 2a3dA_1fpoA, 2abk_1ebmA, 2admA_3mag, 2afpA_1ixxB, 2bbkH_1qfmA, 2bbvA_1f2nA, 2bce_1acl, 2bce_1qonA, 2bopA_1ris, 2btvA_2btvB, 2cba_1koqB, 2cblA_1psrA, 2cbp_1f56C, 2ccyA_256bA, 2cmd_1mldD, 2cmd_3ldh, 2cpl_1c5fC, 2cpl_2nul, 2dhqA_1c9kB, 2dkb_2gsaA, 2dkb_2oatC, 2dorA_1ep3A, 2dri_8abp, 2eboA_1mof

ตารางที่ 21 กลุ่มข้อมูลที่ 64 - 68

กลุ่มข้อมูล	ชื่อชุดข้อมูล
64	2er7E_2rmpA, 2ercA_3mag, 2fal_1ebt, 2fal_3sdhA, 2fcbA_1iisC, 2fcbA_2ncm, 2fnbA_1bpv, 2gar_1bxkA, 2gdm_1a6m, 2gdm_2vhbA, 2gmfA_1f6fA, 2hbg_2vhbA, 2hbg_3sdhA, 2hddA_1b8iA, 2hdhA_3hdhB, 2hrvA_1agjA, 2hrvA_5ptp, 2hts_2irfG, 2i1b_1iltA, 2i1b_1iraX, 2i1b_2ila, 2if1_1d1rA, 2ila_1abrB, 2ila_1hce, 2ilk_1vlk
65	2lbd_2prgA, 2masA_1ezrA, 2mbr_1qltA, 2mev1_1tmf1, 2nlrA_1h8vF, 2nmbA_1shcA, 2nsyA_1sur, 2occB_1qleB, 2ohxA_3hudA, 2omf_1osmC, 2omf_2por, 2pgd_1pgjA, 2pii_1cc8A, 2pkaA_1a7s, 2pkaA_1aksA, 2por_3prn, 2prgA_3erdA, 2pth_1c8bA, 2qwc_1b9vA, 2qwc_1nsdB, 2reb_1cr1A, 2reb_1g19A, 2rmpA_1avfA, 2rn2_1ekeA, 2scuA_1drw
66	2scuA_1eudA, 2scuB_1eucB, 2shpA_1i9sA, 2tbvA_1pov1, 2tct_1a6i, 2tgi_1es7C, 2thiA_1eu8A, 2tnfA_1d4vB, 2tpsA_1qpoA, 2tysA_1pii, 2tysB_1tdj, 2u2fA_1u2fA, 2xat_3tdt, 3bct_1g3jC, 3chbD_1prtF, 3chy_1b00B, 3chy_1tmy, 3chy_4tmyB, 3cla_1nocB, 3crd_3ygsP, 3erdA_1dkfA, 3erdA_1ereB, 3eugA_1lauE, 3inkC_1irl, 3ldh_1hlpA
67	3lzt_1gbzA, 3nul_1acf, 3nul_1fil, 3nul_1ifqA, 3pchM_3pchA, 3pte_1pmd, 3pte_1skf, 3pyp_1drmA, 3sdhA_1sctF, 3seb_1enfA, 3stdA_2std, 3tgl_1dt5A, 3ullA_1kawB, 3ullA_1prtF, 3vub_2vubH, 4aahA_1flgB, 4bcl_1ksaA, 4crxA_1floA, 4icb_1mho, 4mbp_1eljA, 4mbp_1eu8A, 4mbp_1ezpA, 4mbp_1gggB, 4pgaA_1ho3A, 4uagA_1eehA
68	5hpgA_2pk4, 5nul_1akt, 5nul_1j9gA, 5tmpA_3tmkG, 6prcL_6prcM, 7taa_2taaA, 8fabA_1tetH

รวมทั้งหมด 68 กลุ่มข้อมูลมี 1682 ชุดข้อมูลที่ใช้ในการทดลอง

เนื่องจากการชุดฐานข้อมูล BAliBASE และ PREFAB ในแต่ละชุดที่เทียบเรียงมีจำนวนและความยาวของลำดับข้อมูลไม่มากซึ่งทำให้ผลความแตกต่างทางด้านเวลาที่ใช้ในการประมวลผลในแต่ละชุดข้อมูลไม่เด่นชัด ดังนั้นจึงได้มีการวัดผลทดลองการเทียบเรียงกลุ่มลำดับข้อมูลโปรตีนจากฐานข้อมูล NCBI ที่มีจำนวนข้อมูลตั้งแต่ 10 – 500 ลำดับ และมีความยาวของลำดับเฉลี่ยประมาณ 280 ตัวอักษร เพื่อศึกษาแนวโน้ม และเปรียบเทียบเวลาที่ใช้ในการเทียบเรียงของทั้ง 2 โปรแกรมเมื่อจำนวนของลำดับที่เทียบเรียงเพิ่มขึ้นอีกด้วย

3. การวัดความถูกต้องของการเทียบเรียง

ความถูกต้องของผลการเทียบเรียงในงานวิจัยนี้วัดผลโดยพิจารณาจากค่าคะแนนความถูกต้อง ซึ่งสำหรับฐานข้อมูล BAliBASE วัดผลจากค่าคะแนน SP และค่าคะแนน TC และสำหรับฐานข้อมูล PREFAB วัดผลจากค่าคะแนน Q โดยค่าคะแนนแต่ละแบบมีความหมายดังนี้

3.1 ค่าคะแนน SP (Sum of Pair score) คำนวณโดยโปรแกรมของคณะวิจัยผู้สร้างฐานข้อมูล BAliBASE เป็นค่าคะแนนที่ได้จากการหาอัตราส่วนระหว่างจำนวนอักขระของส่วนจำเพาะที่มีตำแหน่งถูกต้องของผลการเทียบเรียงของโปรแกรมเปรียบเทียบกับจำนวนอักขระส่วนจำเพาะทั้งหมดของผลการเทียบเรียงอ้างอิง ค่าคะแนน SP ใช้วัดความถูกต้องของผลการเทียบเรียงฐานข้อมูล BAliBASE บ่งบอกถึงความถูกต้องของผลการเทียบเรียงกลุ่มลำดับข้อมูลเมื่อเปรียบเทียบกับผลเทียบเรียงอ้างอิงซึ่งเทียบเรียงโดยผู้เชี่ยวชาญ ซึ่งค่าคะแนน SP จะมีค่าระหว่าง 0.0-1.0 โดยที่ค่าคะแนนจะเพิ่มขึ้นตามความถูกต้องของผลการเทียบเรียง

3.2 ค่าคะแนน TC (Total Column score) คำนวณโดยโปรแกรมของคณะวิจัยผู้สร้างฐานข้อมูล BAliBASE เป็นค่าคะแนนที่ได้จากการหาอัตราส่วนระหว่างจำนวนของสดมภ์ในส่วนจำเพาะที่ถูกต้องของผลการเทียบเรียงของโปรแกรมเปรียบเทียบกับจำนวนสดมภ์ในส่วนจำเพาะทั้งหมดของผลการเทียบเรียงอ้างอิง ค่าคะแนน TC ใช้วัดความถูกต้องของผลการเทียบเรียงฐานข้อมูล BAliBASE บ่งบอกถึงความถูกต้องของผลการเทียบเรียงในแนวสดมภ์เมื่อเปรียบเทียบกับผลเทียบเรียงอ้างอิงซึ่งเทียบเรียงโดยผู้เชี่ยวชาญเช่นกัน ค่าคะแนน TC จะมีค่าระหว่าง 0.0-1.0 โดยที่ค่าคะแนนจะเพิ่มขึ้นตามความถูกต้องของผลการเทียบเรียงเช่นเดียวกับค่าคะแนน SP

3.3 ค่าคะแนน Q คำนวณโดยโปรแกรมของคณะวิจัยสร้างฐานข้อมูล PREFAB เวอร์ชั่น 4.0 ไว้สำหรับวัดผลการเทียบเรียงฐานข้อมูล PREFAB ค่าคะแนน Q คำนวณโดยเริ่มจากนำผลการเทียบเรียงของโปรแกรมไปหาตำแหน่งที่ตรงกันทีละคู่ลำดับ และทำเช่นเดียวกันนี้กับผลการเทียบเรียงอ้างอิง หลังจากนั้นจึงเอาข้อมูลตำแหน่งตรงกันที่หาได้ของแต่ละคู่จากผลการเทียบเรียงของโปรแกรม และผลการเทียบเรียงอ้างอิงมาเทียบกัน โดยคำนวณเป็นค่าเฉลี่ยของอัตราส่วนระหว่างจำนวนของตำแหน่งที่ถูกต้องของผลการเทียบเรียงของโปรแกรม กับจำนวนตำแหน่งที่ตรงกันของแต่ละคู่ของการเทียบเรียงอ้างอิง โดยค่าคะแนน Q จะมีค่าระหว่าง 0 ถึง 1 เช่นเดียวกับค่าคะแนน SP และ TC ของฐานข้อมูล BAliBASE

โปรแกรมสำหรับการคำนวณค่าคะแนน SP และค่าคะแนน TC ชื่อโปรแกรม Bali_score.c สามารถหาได้จาก http://bips.u-strasbg.fr/fr/Products/Databases/BAliBASE/bali_score.c และโปรแกรมสำหรับการคำนวณค่าคะแนน Q สามารถหาได้จาก http://www.drive5.com/qscore/qscore_src.tar.gz

ผลและวิจารณ์

ระบบที่ใช้ในงานวิจัยนี้ เราใช้เครื่องคอมพิวเตอร์ Athlron XP 2000+ ที่มี CPU Clock rate 1.67 GHz หน่วยความจำหลัก 512 MB ในส่วนของการทดลอง เราใช้โปรแกรม ClustalW และโปรแกรม SSFA เทียบเรียงชุดลำดับข้อมูลของฐานข้อมูล BAliBASE และ PREFAB เพื่อเปรียบเทียบประสิทธิภาพทางด้านเวลาที่ใช้ประมวลผล และค่าความถูกต้องของผลลัพธ์ที่เทียบเรียงได้ โดยที่ ClustalW คือผลที่ได้จากโปรแกรม ClustalW และ SSFA คือผลที่ได้จากโปรแกรม SSFA

1. ผลการเทียบเรียงชุดข้อมูลของฐานข้อมูล BAliBASE

การเทียบเรียงชุดข้อมูลของฐานข้อมูล BAliBASE วัดผลประสิทธิภาพทางด้านเวลาด้วยเวลารวมในการเทียบเรียงทุกชุดของฐานข้อมูล เนื่องจากแต่ละชุดข้อมูลของฐานข้อมูล BAliBASE มีจำนวนข้อมูลไม่มากส่งผลให้เวลาในการประมวลผลของแต่ละชุดใช้เวลาน้อย และวัดประสิทธิภาพได้ยาก ส่วนในด้านความถูกต้องพิจารณาจากค่าคะแนน SP และค่าคะแนน TC ทั้งของแต่ละชุดข้อมูล และค่าคะแนนเฉลี่ยของแต่ละชุดอ้างอิง

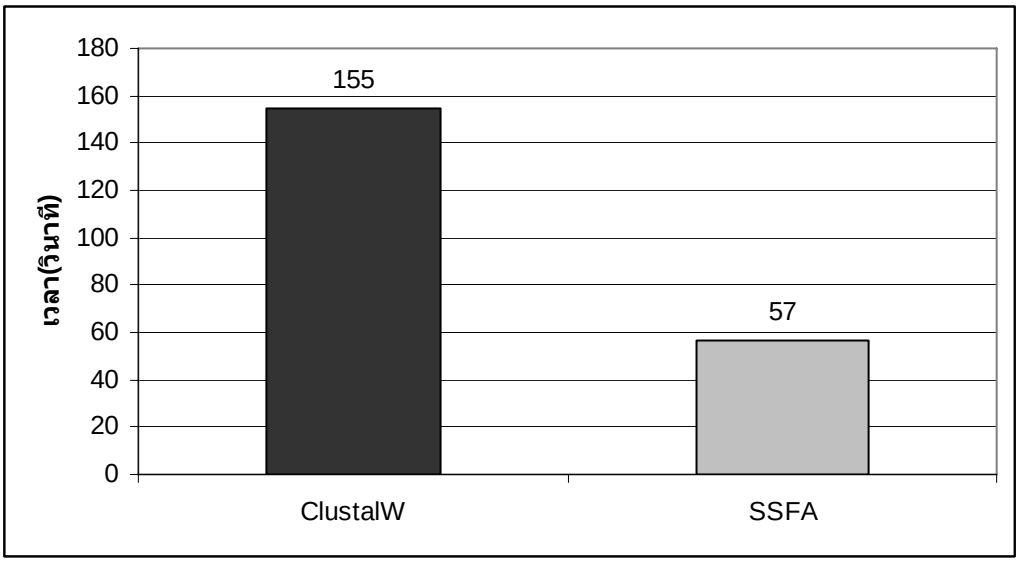
ภาพที่ 17 แสดงเวลาที่ใช้ในการเทียบเรียงลำดับข้อมูลชุดอ้างอิงทั้ง 5 ชุด แสดงให้เห็นว่าที่โปรแกรม ClustalW ใช้เวลารวมในการประมวลผลมากกว่าโปรแกรม SSFA อย่างเห็นได้ชัด เนื่องจากเวลาที่ใช้ในขั้นตอนที่ 1 ของการเทียบเรียงลดลง ซึ่งเวลาที่โปรแกรม ClustalW ใช้นั้นมากกว่าโปรแกรม SSFA ถึง 2.72 เท่า

ภาพที่ 18 - ภาพที่ 31 แสดงการเปรียบเทียบค่าคะแนน SP และค่าคะแนน TC ของผลลัพธ์ที่ได้จากการเทียบเรียงด้วยโปรแกรม ClustalW และโปรแกรม SSFA สำหรับแต่ละชุดข้อมูล และภาพที่ 32 และภาพที่ 33 แสดงการเปรียบเทียบค่าคะแนน SP และค่าคะแนน TC เฉลี่ยของแต่ละชุดอ้างอิงเมื่อเทียบเรียงด้วยโปรแกรม ClustalW และโปรแกรม SSFA

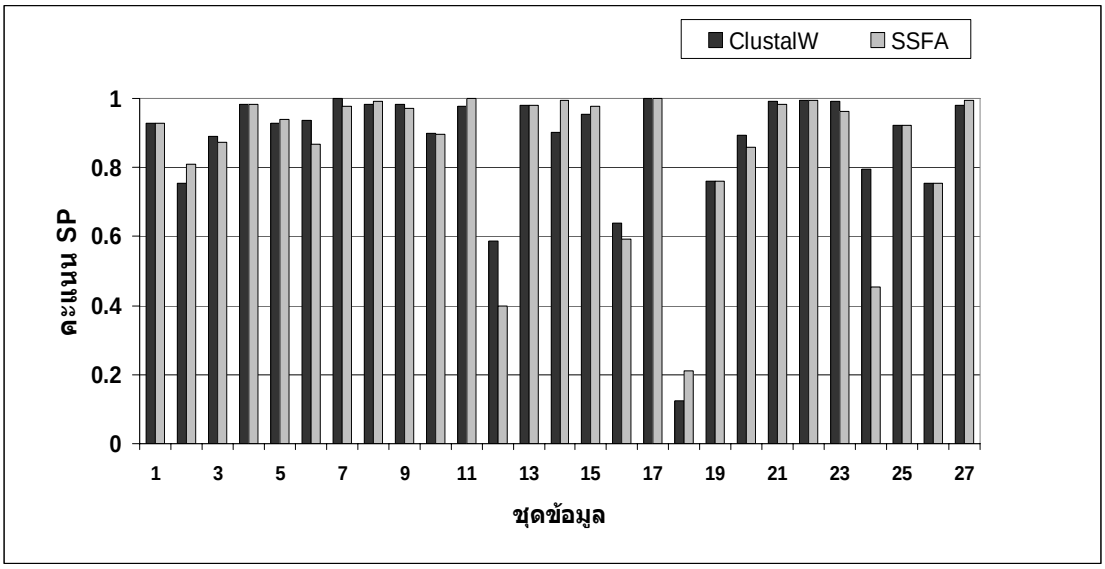
เมื่อพิจารณาค่าคะแนน SP ในภาพคู่ (ภาพที่ 18 – ภาพที่ 32) แสดงให้เห็นว่าโปรแกรม SSFA และโปรแกรม ClustalW ให้ค่าคะแนน SP มีแนวโน้มไปในทิศทางเดียวกัน และมีค่าคะแนนใกล้เคียงกันอยู่ในระดับที่ยอมรับได้ โดยเฉพาะชุดอ้างอิง: 5 ซึ่งเป็นชุดข้อมูลที่มีการแทรก

ภายในโปรแกรม SSFA ได้ค่าคะแนน SP ที่ดีกว่าโปรแกรม ClustalW ในหลายๆชุดข้อมูล และโปรแกรม SSFA มีค่าคะแนน SP เฉลี่ยของชุดอ้างอิง: 5 มากกว่าโปรแกรม ClustalW เล็กน้อย

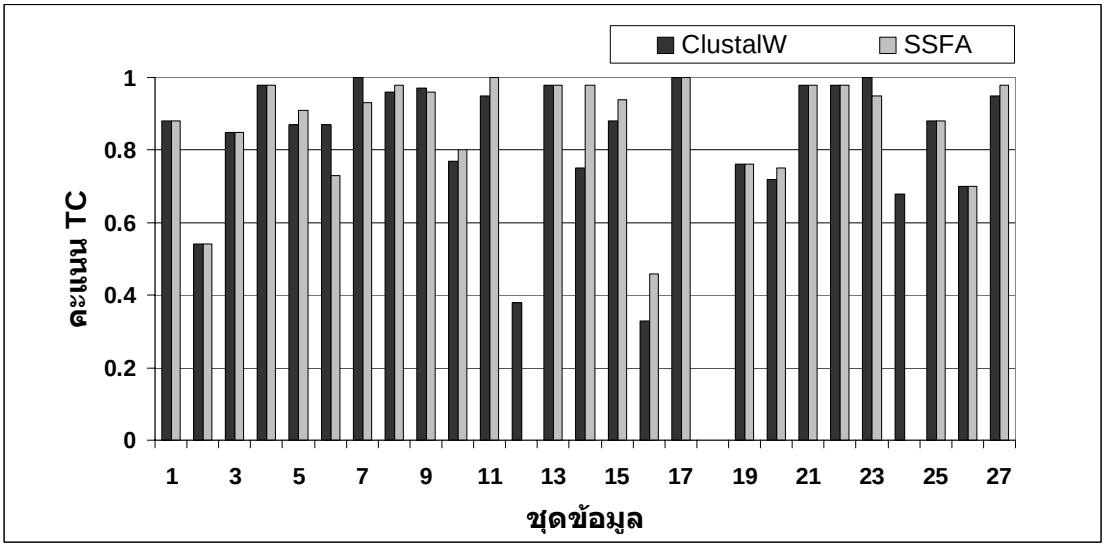
เมื่อพิจารณาค่าคะแนน TC ซึ่งเป็นค่าคะแนนความถูกต้องของสดมภ์เมื่อเปรียบเทียบกับผลลัพธ์อ้างอิง ในภาพคู่ (ภาพที่ 19 – ภาพที่ 33) แสดงให้เห็นว่าโปรแกรม SSFA และโปรแกรม ClustalW ให้ค่าคะแนน TC ที่ใกล้เคียงกันอยู่ในระดับที่ยอมรับได้เช่นเดียวกัน ยกเว้นชุดอ้างอิง: 4 ซึ่งโปรแกรม SSFA ให้ค่าคะแนน TC น้อยกว่าโปรแกรม ClustalW มากพอสมควร เกิดจากการคำนวณค่าคะแนน TC สนใจเฉพาะความถูกต้องในส่วนจำเพาะ และชุดข้อมูลที่ 4 มีลักษณะการมีข้อมูลขนาดใหญ่เพิ่มเติมที่ตอนต้นหรือตอนท้ายของข้อมูล ซึ่งลักษณะข้อมูลแบบชุดข้อมูลที่ 4 นี้ทำให้ผลการเทียบเรียงของโปรแกรม SSFA ได้ส่วนจำเพาะที่กระจายไปในข้อมูลขนาดใหญ่ที่เพิ่มเติมเข้ามา ทำให้ผลการเทียบเรียงแตกต่างจากผลการเทียบเรียงอ้างอิงมากกว่าโปรแกรม ClustalW แต่หากพิจารณาค่าคะแนน SP ของชุดอ้างอิง: 4 ที่มีค่าใกล้เคียงกัน สามารถสรุปได้ว่าโปรแกรม SSFA สามารถเทียบเรียงชุดอ้างอิง: 4 ให้ผลลัพธ์ที่มีความถูกต้องในระดับที่ยอมรับได้ แต่ผลลัพธ์มีความแตกต่างจากผลลัพธ์อ้างอิงมากกว่าโปรแกรม ClustalW



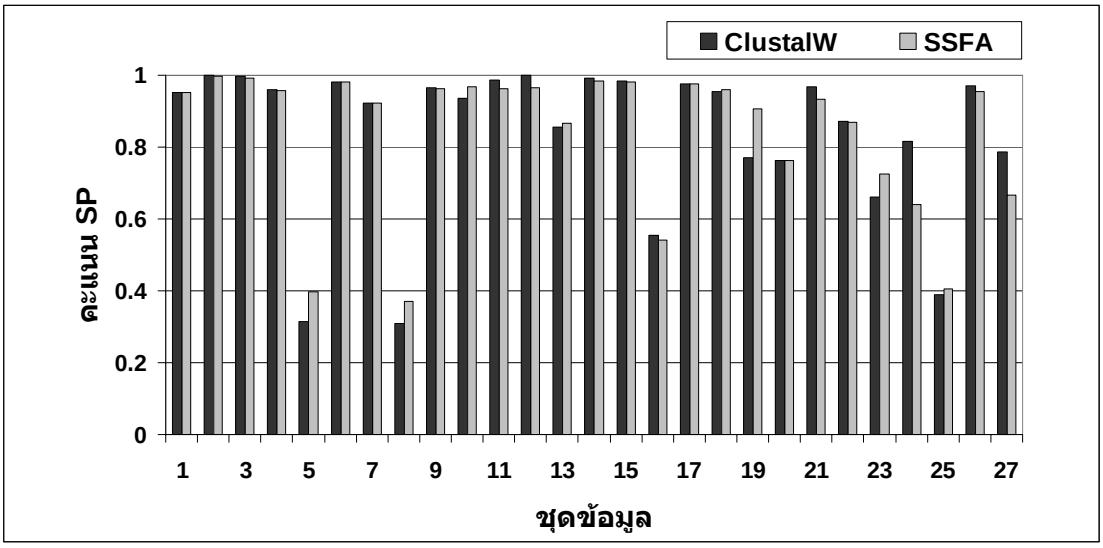
ภาพที่ 17 เวลาที่โปรแกรม ClustalW และโปรแกรม SSFA ใช้ในการเทียบเรียงกลุ่มลำดับข้อมูลของชุดอ้างอิงทั้ง 5 ชุด



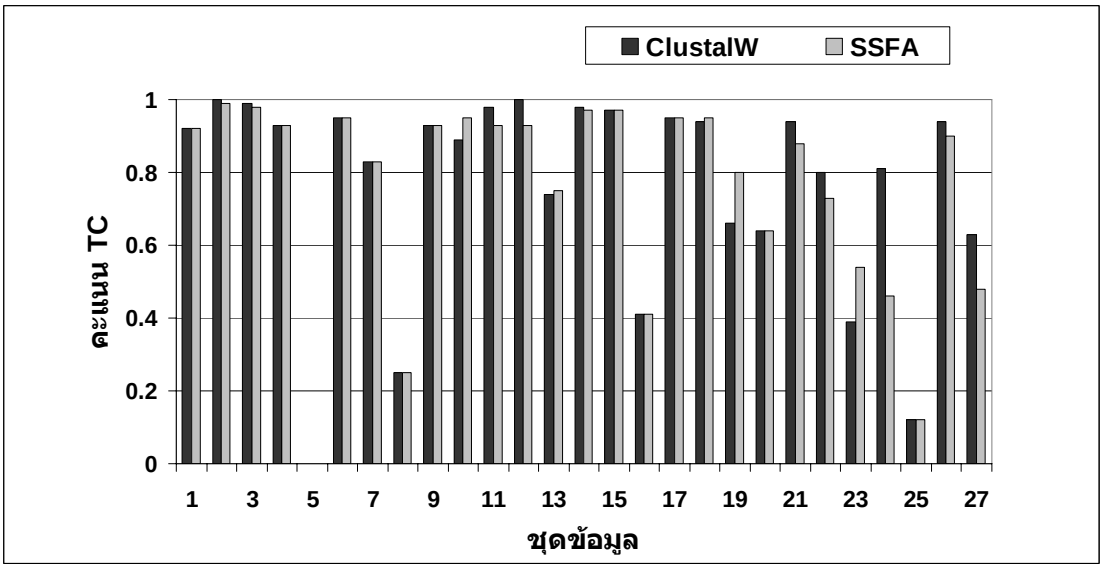
ภาพที่ 18 ค่าคะแนน SP ในการเทียบเรียงกลุ่มลำดับข้อมูลของโปรแกรม ClustalW และโปรแกรม SSFA กับชุดอ้างอิง: 1 ชุดทดสอบ 1



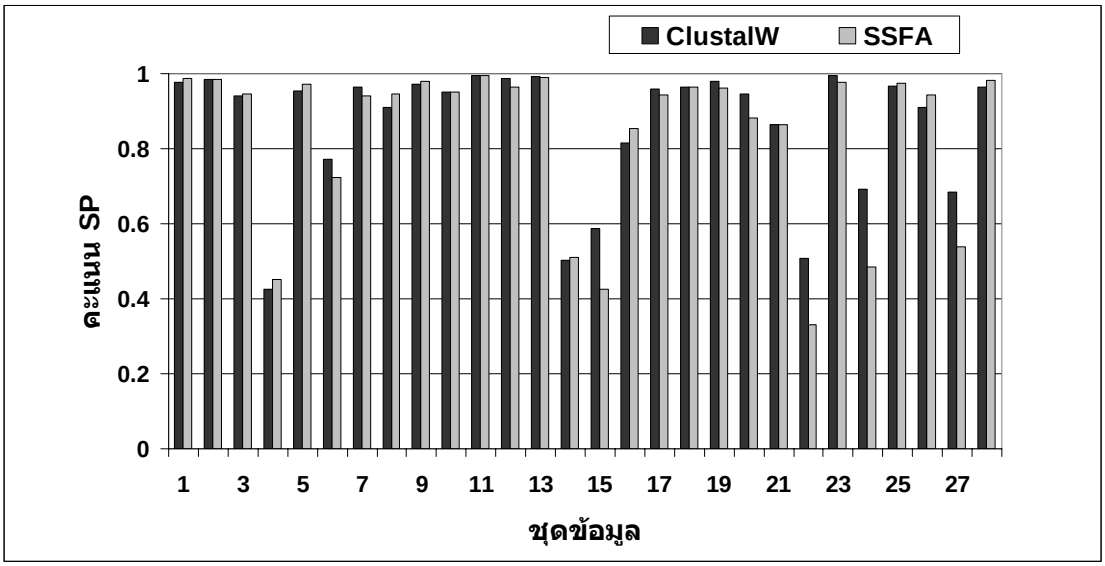
ภาพที่ 19 ค่าคะแนน TC ในการเทียบเรียงกลุ่มลำดับข้อมูลของโปรแกรม ClustalW และโปรแกรม SSFA กับชุดอ้างอิง: 1 ชุดทดสอบ 1



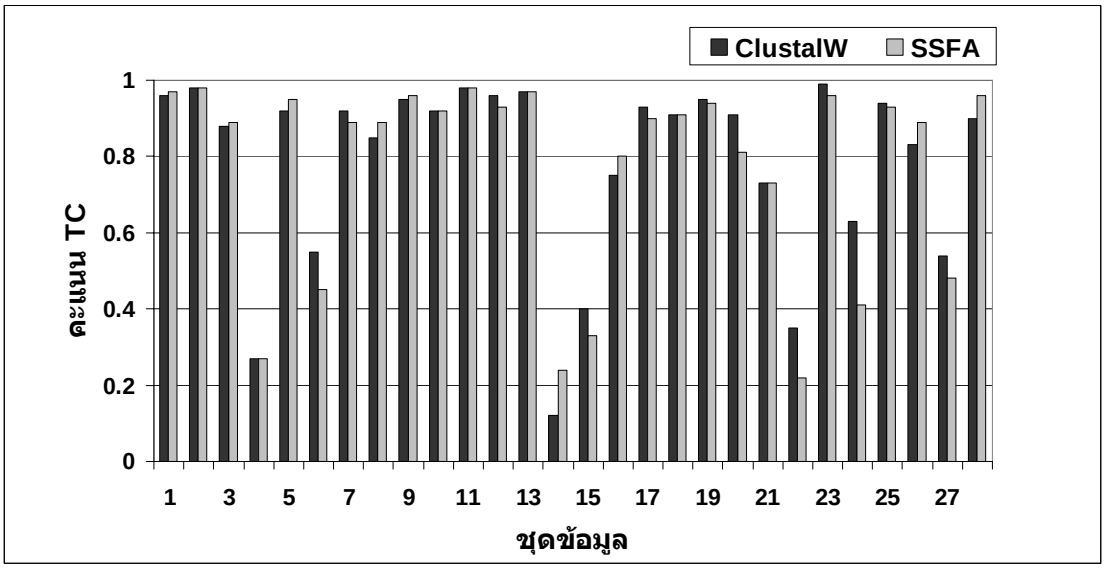
ภาพที่ 20 ค่าคะแนน SP ในการเทียบเรียงกลุ่มลำดับข้อมูลของโปรแกรม ClustalW และโปรแกรม SSFA กับชุดอ้างอิง: 1 ชุดทดสอบ 2



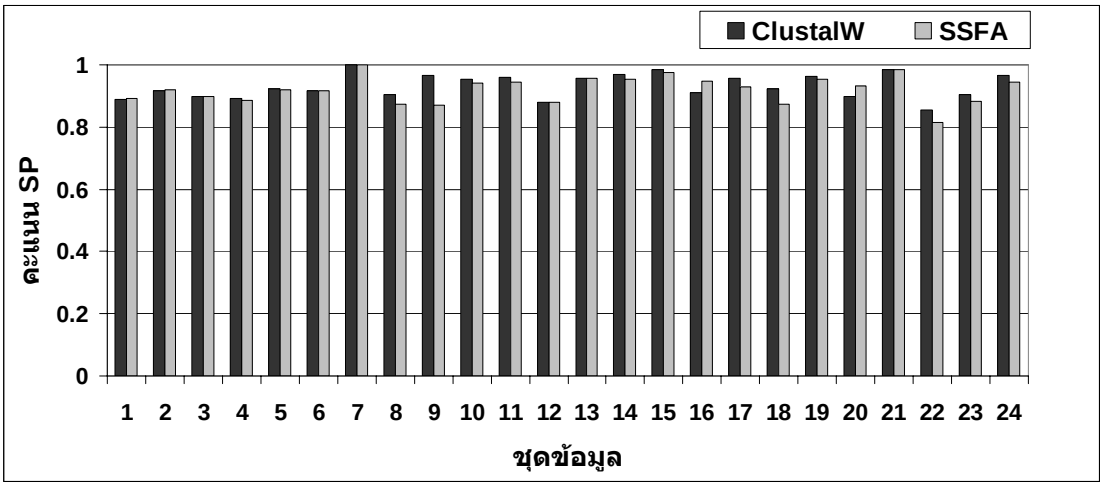
ภาพที่ 21 ค่าคะแนน TC ในการเทียบเรียงกลุ่มลำดับข้อมูลของโปรแกรม ClustalW และโปรแกรม SSFA กับชุดอ้างอิง: 1 ชุดทดสอบ 2



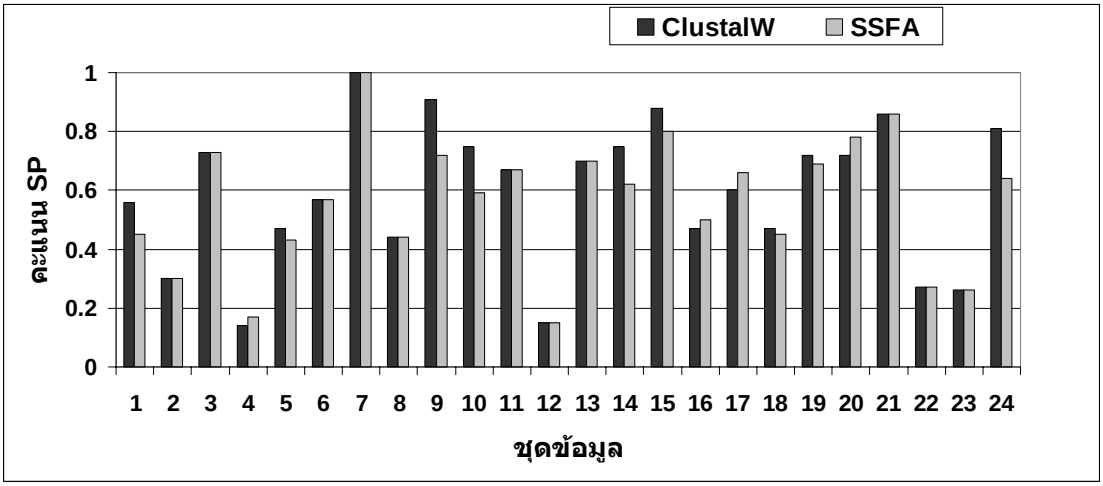
ภาพที่ 22 ค่าคะแนน SP ในการเทียบเรียงกลุ่มลำดับข้อมูลของโปรแกรม ClustalW และโปรแกรม SSFA กับชุดอ้างอิง: 1 ชุดทดสอบ 3



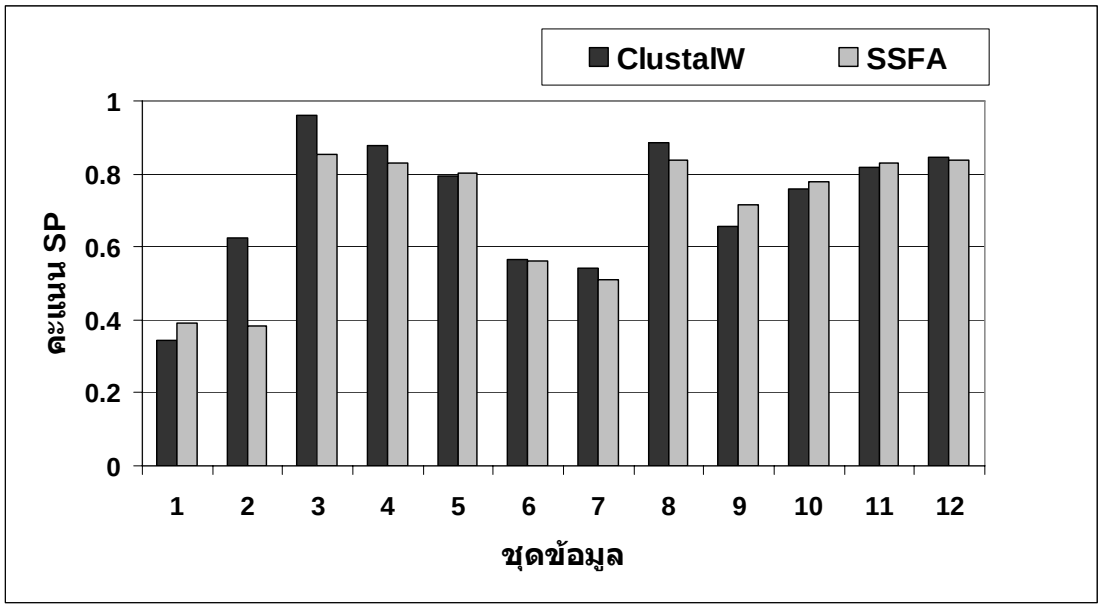
ภาพที่ 23 ค่าคะแนน TC ในการเทียบเรียงกลุ่มลำดับข้อมูลของโปรแกรม ClustalW และโปรแกรม SSFA กับชุดอ้างอิง: 1 ชุดทดสอบ 3



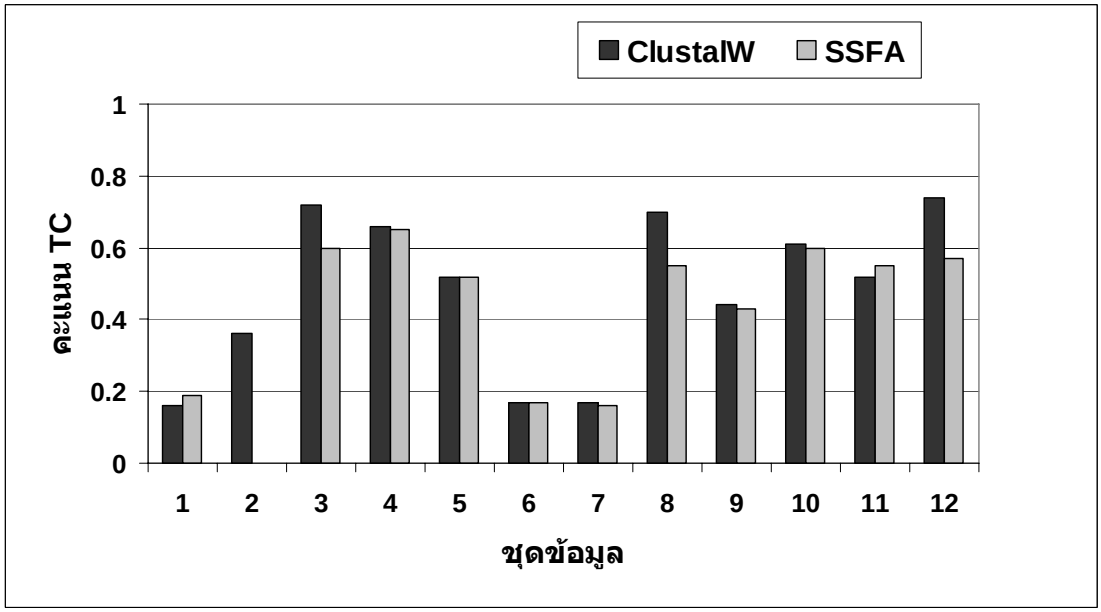
ภาพที่ 24 ค่าคะแนน SP ในการเทียบเรียงกลุ่มลำดับข้อมูลของโปรแกรม ClustalW และโปรแกรม SSFA กับชุดอ้างอิง: 2



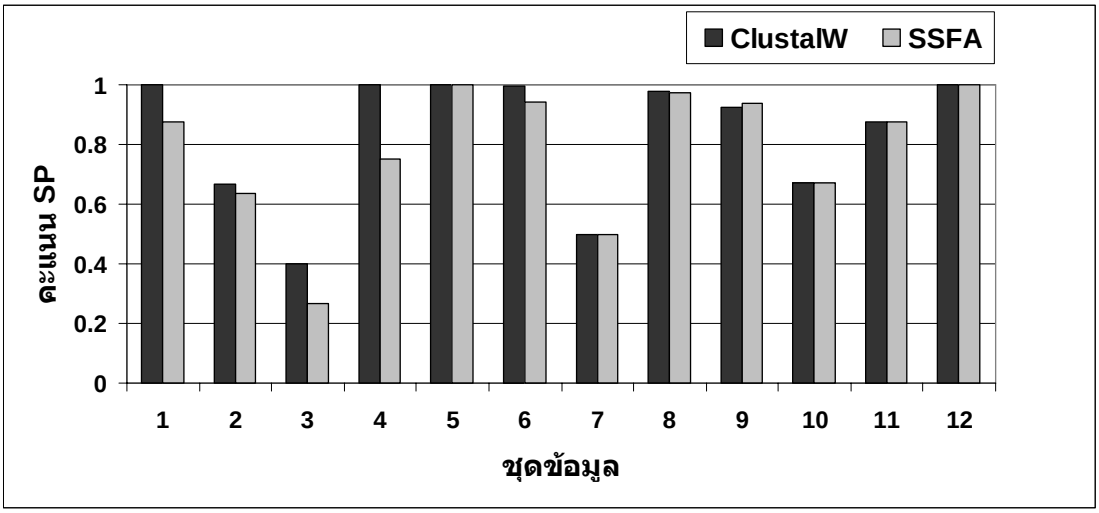
ภาพที่ 25 ค่าคะแนน TC ในการเทียบเรียงกลุ่มลำดับข้อมูลของโปรแกรม ClustalW และโปรแกรม SSFA กับชุดอ้างอิง: 2



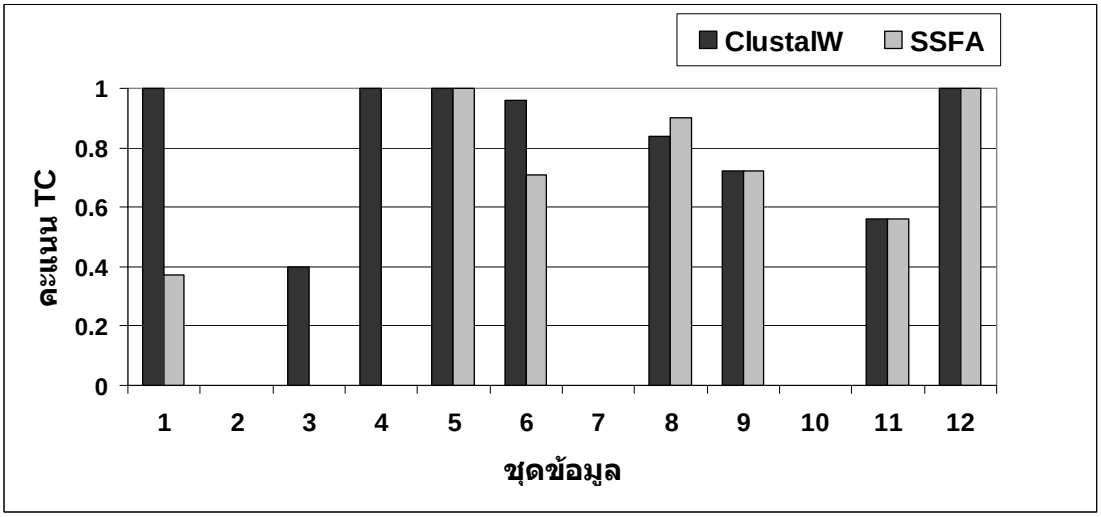
ภาพที่ 26 ค่าคะแนน SP ในการเทียบเรียงกลุ่มลำดับข้อมูลของโปรแกรม ClustalW และโปรแกรม SSFA กับชุดอ้างอิง: 3



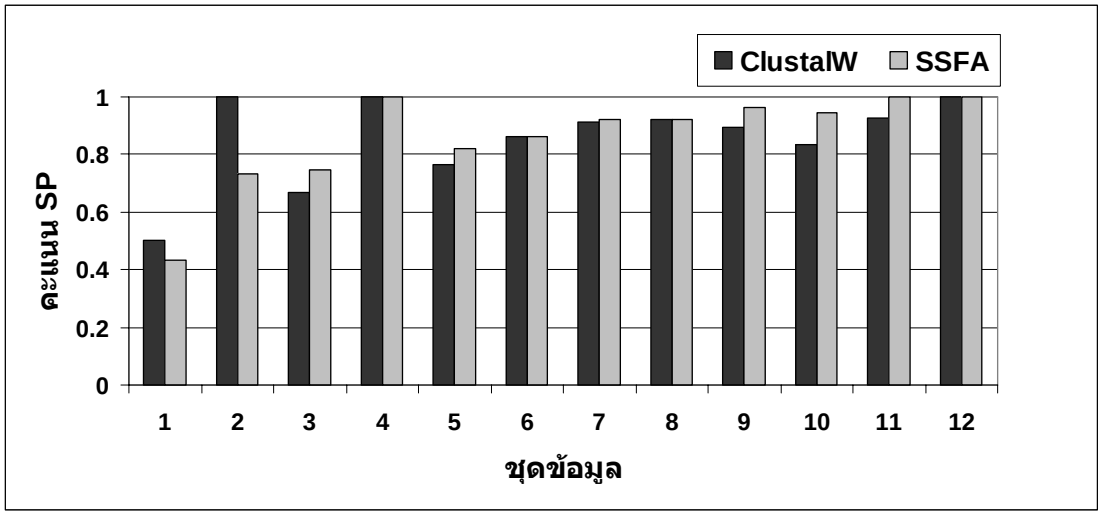
ภาพที่ 27 ค่าคะแนน TC ในการเทียบเรียงกลุ่มลำดับข้อมูลของโปรแกรม ClustalW และโปรแกรม SSFA กับชุดอ้างอิง: 3



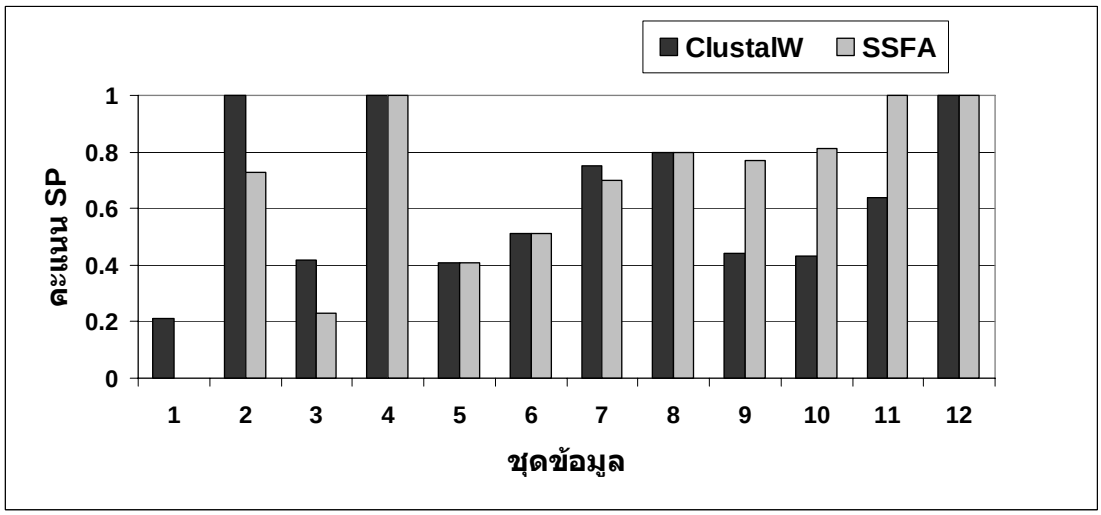
ภาพที่ 28 ค่าคะแนน SP ในการเทียบเรียงกลุ่มลำดับข้อมูลของโปรแกรม ClustalW และโปรแกรม SSFA กับชุดอ้างอิง: 4



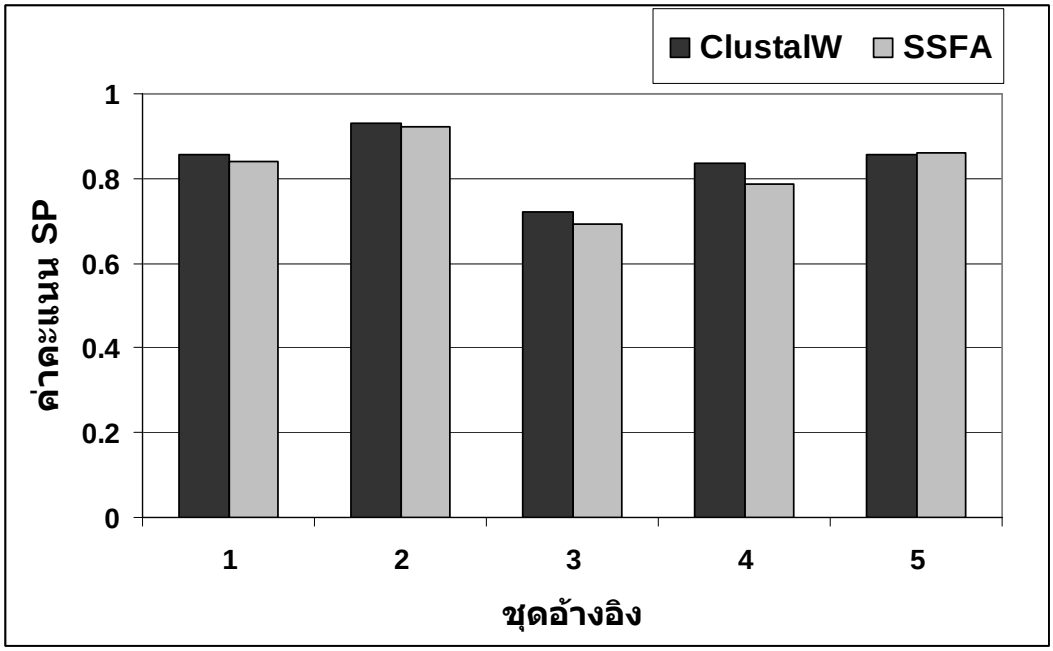
ภาพที่ 29 ค่าคะแนน TC ในการเทียบเรียงกลุ่มลำดับข้อมูลของโปรแกรม ClustalW และโปรแกรม SSFA กับชุดอ้างอิง: 4



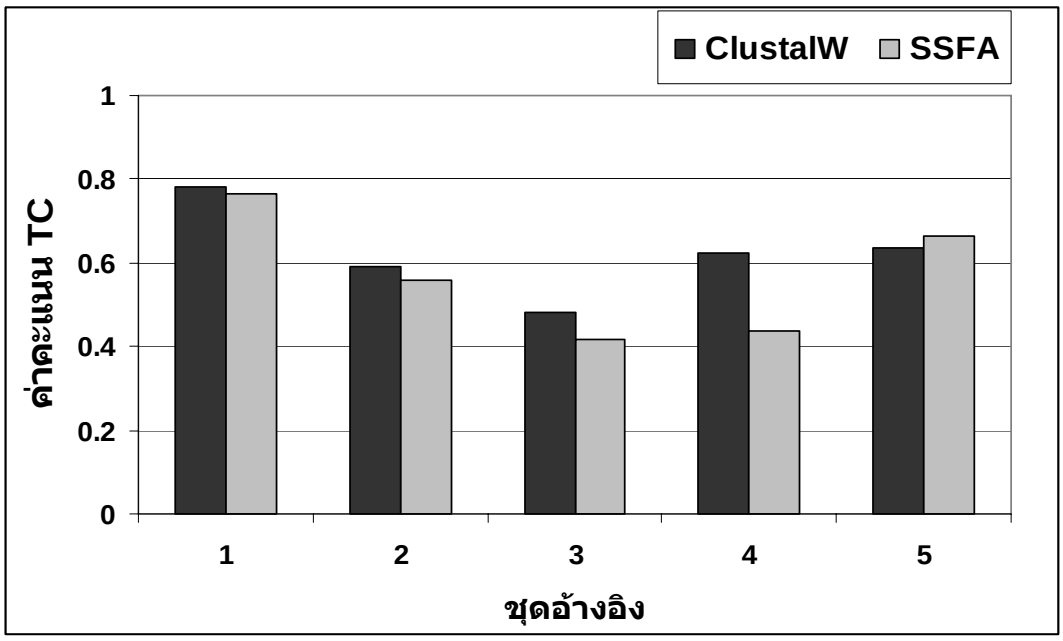
ภาพที่ 30 ค่าคะแนน SP ในการเทียบเรียงกลุ่มลำดับข้อมูลของโปรแกรม ClustalW และโปรแกรม SSFA กับชุดอ้างอิง: 5



ภาพที่ 31 ค่าคะแนน TC ในการเทียบเรียงกลุ่มลำดับข้อมูลของโปรแกรม ClustalW และโปรแกรม SSFA กับชุดอ้างอิง: 5



ภาพที่ 32 ค่าคะแนน SP เฉลี่ยในการเทียบเรียงกลุ่มลำดับข้อมูลของโปรแกรม ClustalW และโปรแกรม SSFA ของแต่ละชุดอ้างอิง

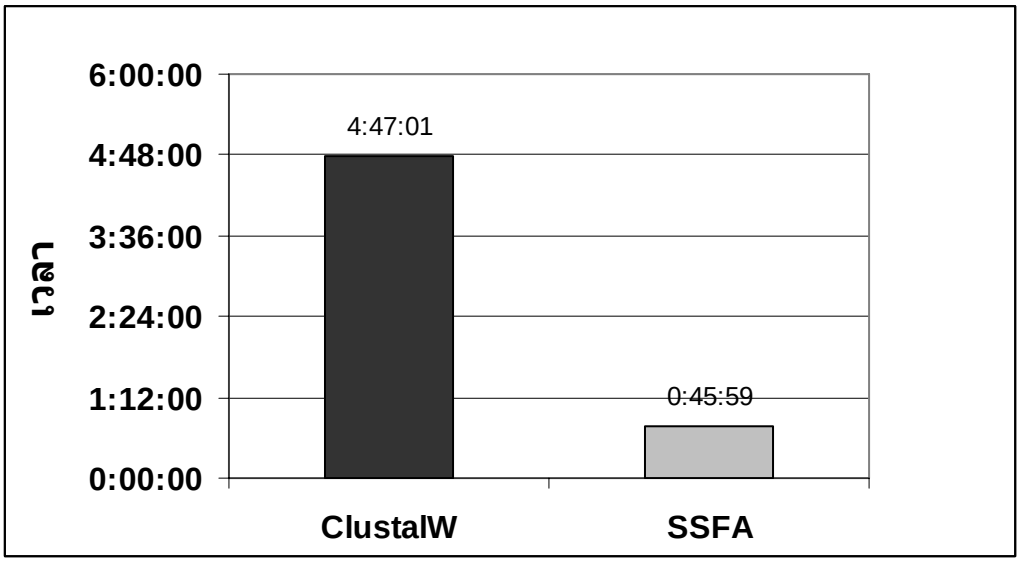


ภาพที่ 33 ค่าคะแนน TC เฉลี่ยในการเทียบเรียงกลุ่มลำดับข้อมูลของโปรแกรม ClustalW และโปรแกรม SSFA ของแต่ละชุดอ้างอิง

2. ผลการเทียบเรียงชุดข้อมูลของฐานข้อมูล PREFAB

การเทียบเรียงชุดข้อมูลของฐานข้อมูล PREFAB เปรียบเทียบประสิทธิภาพทางด้านเวลาด้วยเวลารวมในการเทียบเรียงทุกชุดของฐานข้อมูลเช่นเดียวกับฐานข้อมูล BAliBASE ส่วนการเปรียบเทียบด้านความถูกต้องพิจารณาจากค่าคะแนน Q ทั้งจากการเทียบเรียงแต่ละชุดข้อมูล และค่าคะแนนเฉลี่ยของทุกชุดข้อมูล

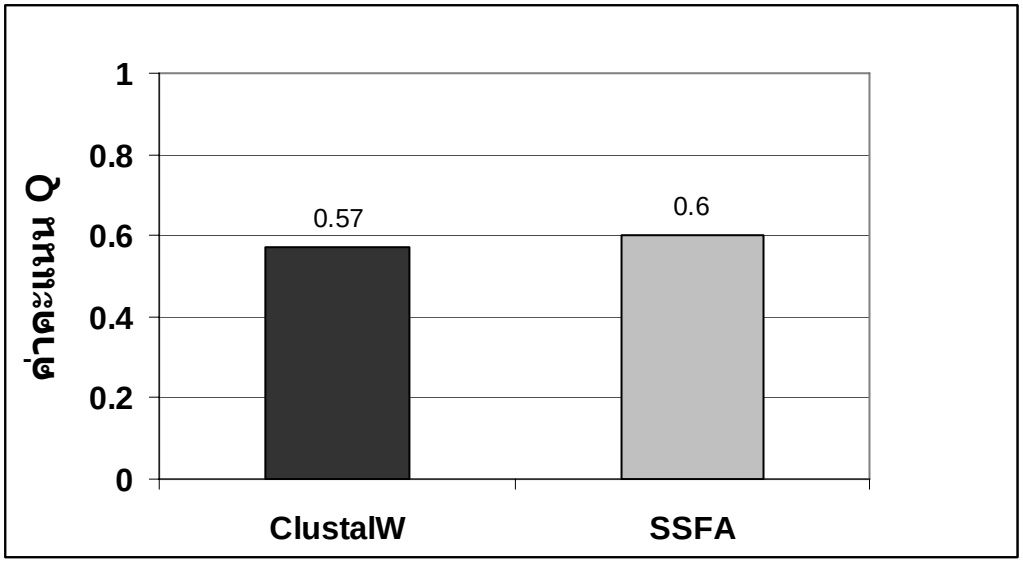
ภาพที่ 34 แสดงให้เห็นถึงความแตกต่างของเวลาที่โปรแกรม ClustalW และ โปรแกรม SSFA ใช้ในการเทียบเรียงชุดข้อมูลทั้ง 1682 ชุดของฐานข้อมูล PREFAB จะเห็นได้ว่าโปรแกรม ClustalW ใช้เวลาในการเทียบเรียงมากกว่าโปรแกรม SSFA มากถึง 6.24 เท่า ซึ่งความแตกต่างของเวลาที่ใช้ในการเทียบเรียงฐานข้อมูล PREFAB เห็นได้ชัดกว่าฐานข้อมูล BAliBASE เนื่องจากฐานข้อมูล PREFAB มีจำนวนชุดข้อมูลมากกว่า และในแต่ละชุดข้อมูลมีจำนวนลำดับข้อมูลมากกว่าฐานข้อมูล BAliBASE



ภาพที่ 34 เวลาที่โปรแกรม ClustalW และโปรแกรม SSFA ใช้ในการเทียบเรียงฐานข้อมูล PREFAB

ในด้านความถูกต้องเมื่อพิจารณาจากค่าคะแนน Q เฉลี่ยของผลลัพธ์ที่ได้จากโปรแกรม ClustalW และโปรแกรม SSFA ดังภาพที่ 35 ซึ่งมีความใกล้เคียงกัน โดยที่โปรแกรม SSFA ได้ค่าคะแนน Q เฉลี่ย 0.60 โดยมีค่าความเบี่ยงเบนมาตรฐานอยู่ที่ 0.34 ส่วนโปรแกรม ClustalW ได้ค่า

คะแนน Q เฉลี่ย 0.57 โดยมีค่าความเบี่ยงเบนมาตรฐานอยู่ที่ 0.33 ซึ่งถือว่ามีความใกล้เคียง และมีแนวโน้มของค่าคะแนนที่คล้ายกัน แสดงให้เห็นว่าโปรแกรม SSFA ให้ความถูกต้องในการเทียบเรียงกลุ่มลำดับข้อมูลในระดับเดียวกับโปรแกรม ClustalW

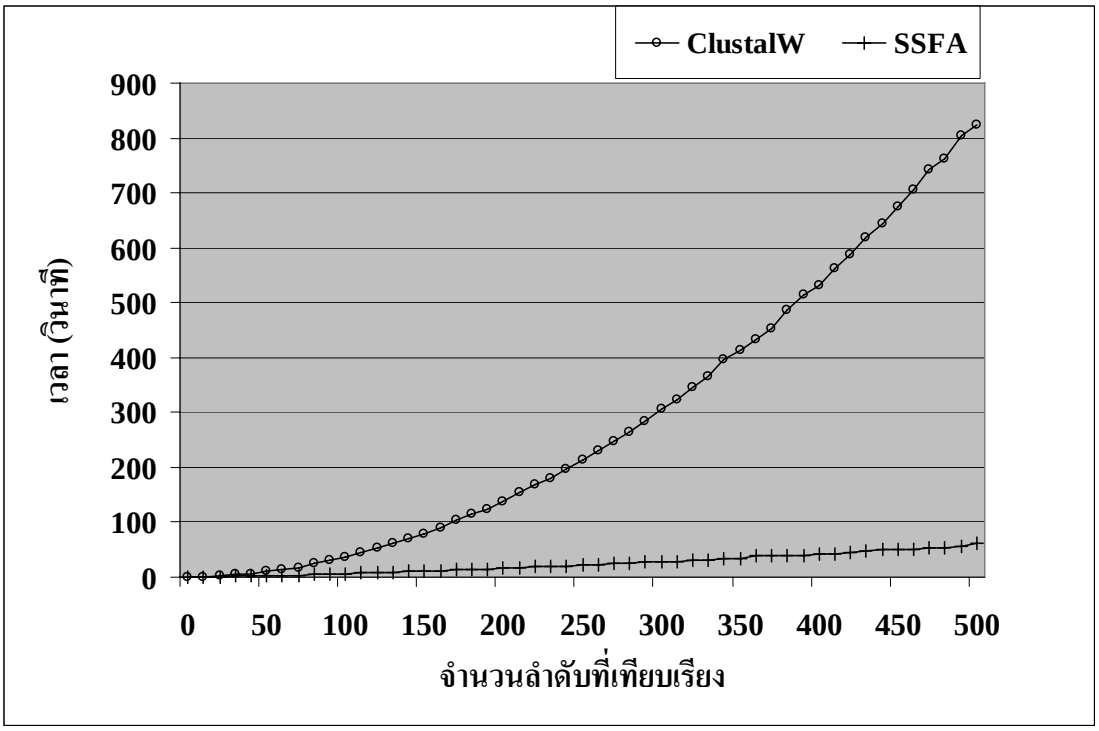


ภาพที่ 35 ค่าคะแนน Q เฉลี่ยในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA

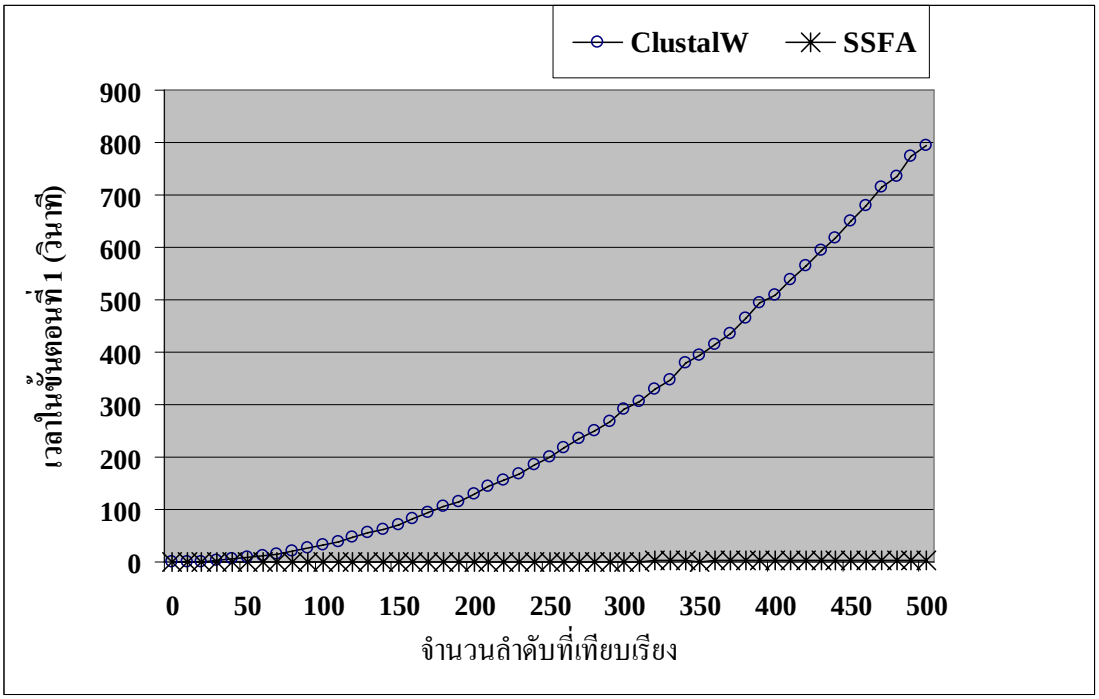
3. แนวโน้มทางด้านเวลาในการเทียบเรียงกลุ่มลำดับข้อมูลขนาดต่าง ๆ

ปัจจัยหลักที่ทำให้เวลาในการเทียบเรียงกลุ่มลำดับข้อมูลแตกต่างกัน คือจำนวนข้อมูลที่เข้าเทียบเรียง เมื่อจำนวนข้อมูลในกลุ่มลำดับข้อมูลมากขึ้นเวลาที่ใช้ในการเทียบเรียงจะมากขึ้นอย่างเห็นได้ชัด ในงานวิจัยนี้ได้ใช้ตัวอย่างข้อมูลโปรตีนจากฐานข้อมูล NCBI จัดกลุ่มลำดับที่เข้าเทียบเรียงให้มีจำนวนลำดับตั้งแต่ 10 – 500 ลำดับ โดยมีความยาวของลำดับเฉลี่ย 280 ตัวอักษร เมื่อนำมาเทียบเรียงด้วยโปรแกรม ClustalW และโปรแกรม SSFA เวลาที่ใช้ในการเทียบเรียงเป็นดังภาพที่ 36 และเวลาที่ใช้ในการทำงานขั้นตอนที่ 1 ของการเทียบเรียงข้อมูลเป็นดังภาพที่ 37 ซึ่งแสดงให้เห็นว่าเมื่อจำนวนข้อมูลที่เข้าเทียบเรียงมากขึ้น โปรแกรม ClustalW ใช้เวลาเพิ่มมากขึ้นมากเมื่อจำนวนลำดับที่เทียบเรียงเพิ่มมากขึ้น เนื่องจากในขั้นตอนการสร้างตารางค่าความแตกต่างใช้วิธีเทียบเรียงแต่ละคู่ลำดับข้อมูลซึ่งมีความซับซ้อนของเวลา $O(N^2L^2)$ เมื่อ N คือจำนวนลำดับที่เทียบเรียง และ L คือความยาวของลำดับที่เทียบเรียง ซึ่งแตกต่างกับโปรแกรม SSFA ที่ใช้การพิจารณาความถี่ส่วนย่อยของข้อมูลในการสร้างตารางค่าความแตกต่างซึ่งมีความซับซ้อนของเวลา $O(NL)$ ทำ

ให้อัตราการเพิ่มขึ้นของเวลาที่ใช้ต่อจำนวนของลำดับนั้นน้อย และเวลาที่ใช้ในการประมวลผลน้อยกว่าโปรแกรม ClustalW มาก โดยในการเทียบเรียงกลุ่มลำดับข้อมูลจำนวน 500 ลำดับที่มีความยาวเฉลี่ย 280 ตัวอักษร โปรแกรม SSFA สามารถสร้างตารางค่าความแตกต่างได้เร็วกว่าโปรแกรม CluatalW 198.75 เท่า และเทียบเรียงได้เร็วกว่าโปรแกรม ClustalW 13.14 เท่าซึ่งแสดงให้เห็นว่าโปรแกรม SSFA สามารถรองรับการเทียบเรียงกลุ่มลำดับข้อมูลที่มีจำนวนมากได้แม้ทำงานบนเครื่องคอมพิวเตอร์ทั่วไป



ภาพที่ 36 เวลาที่ใช้ในการเทียบเรียงของโปรแกรม ClustalW และโปรแกรม SSFA ในจำนวนลำดับที่ต่างกัน



ภาพที่ 37 เวลาที่ใช้ในขั้นตอนที่ 1 ของการเทียบเรียงของโปรแกรม ClustalW และโปรแกรม SSFA ในจำนวนลำดับที่ต่างกัน

สรุปและข้อเสนอแนะ

สรุป

การเทียบเรียงกลุ่มลำดับข้อมูลทางชีวภาพในปัจจุบันนอกจากคำนึงถึงความถูกต้องของการเทียบเรียงแล้ว ยังต้องคำนึงถึงเวลาในการเทียบเรียงกลุ่มลำดับข้อมูลซึ่งแปรผันตรงกับจำนวนข้อมูลที่เพิ่มขึ้นอย่างรวดเร็วในปัจจุบัน งานวิจัยนี้จึงนำเสนอวิธีการเทียบเรียงกลุ่มลำดับข้อมูลแบบก้าวหน้าโดยพิจารณาความถี่ส่วนย่อยของลำดับข้อมูล ซึ่งงานวิจัยนี้ใช้เวลาในการประมวลผลน้อยมากโดยที่ความถูกต้องสามารถเทียบเคียงได้กับวิธีการเทียบเรียงกลุ่มลำดับข้อมูลที่นิยมในปัจจุบัน งานวิจัยนี้ใช้เทคนิค n-gram แบ่งส่วนย่อยของข้อมูลเพื่อหาความถี่ของแต่ละส่วนย่อยของคู่ลำดับข้อมูล สำหรับนำไปสร้างตารางค่าความแตกต่าง เพื่อใช้อ้างอิงในการเทียบเรียงกลุ่มลำดับข้อมูลแบบก้าวหน้า ซึ่งเทคนิคนี้ใช้เวลาในการทำงานน้อย ทำให้สามารถรองรับการเทียบเรียงกลุ่มลำดับข้อมูลจำนวนมาก และลำดับข้อมูลที่มีขนาดยาวได้

ผลลัพธ์การเทียบเรียงของงานวิจัยนี้เปรียบเทียบกับโปรแกรม ClustalW ซึ่งเป็นโปรแกรมในกลุ่มวิธีเทียบเรียงกลุ่มลำดับข้อมูลแบบก้าวหน้าที่เป็นที่นิยมที่สุดในปัจจุบัน ผลของการเทียบเรียงฐานข้อมูล BAliBASE โปรแกรม SSFA สามารถเทียบเรียงได้เร็วกว่าโปรแกรม ClustalW 2.72 เท่า โดยที่ความถูกต้องของการเทียบเรียงใกล้เคียงกันอยู่ในระดับที่ยอมรับได้ โปรแกรม SSFA ได้เปรียบโปรแกรม ClustalW ในข้อมูลชุดที่ 5 ที่เป็นกลุ่มข้อมูลที่มีลักษณะการแทรกภายในและโปรแกรม SSFA จะเสียเปรียบโปรแกรม ClustalW ในข้อมูลชุดที่ 4 ซึ่งมีข้อมูลขนาดใหญ่เพิ่มเติมที่ตอนต้นหรือตอนท้ายของข้อมูล ซึ่งข้อมูลลักษณะนี้มีผลทำให้ส่วนจำเพาะที่โปรแกรม SSFA หาได้กระจายไปในส่วนข้อมูลที่เพิ่มเติมขึ้นมาทำให้ผลที่ได้แตกต่างจากผลเทียบเรียงอ้างอิง

สำหรับผลการเทียบเรียงเมื่อเทียบเรียงฐานข้อมูล PREFAB โปรแกรม SSFA สามารถเทียบเรียงได้เร็วกว่าโปรแกรม ClustalW 6.24 เท่า ซึ่งมากกว่าการเทียบเรียงฐานข้อมูล BAliBASE เนื่องจากฐานข้อมูล PREFAB มีชุดข้อมูลที่มากกว่า และในแต่ละชุดข้อมูลมีจำนวนลำดับและความยาวของสายลำดับมากกว่าฐานข้อมูล BAliBASE ซึ่งแสดงให้เห็นว่าเมื่อกลุ่มข้อมูลที่เข้าเทียบเรียงมีขนาดใหญ่ขึ้นประสิทธิภาพทางด้านเวลาของโปรแกรม SSFA จะดีกว่าโปรแกรม ClustalW เพิ่มขึ้นด้วย ในส่วนความถูกต้องของผลการเทียบเรียงนั้นโปรแกรม SSFA ให้ค่าความถูกต้องเฉลี่ยและค่าเบี่ยงเบนมาตรฐานใกล้เคียงกับ โปรแกรม ClustalW แสดงให้เห็นว่าความถูกต้องของโปรแกรม SSFA นั้นอยู่ในระดับเดียวกับโปรแกรม ClustalW

เมื่อพิจารณาประสิทธิภาพทางด้านเวลาเมื่อเทียบเรียงกลุ่มข้อมูลขนาดต่างๆ แสดงให้เห็นว่าเมื่อกลุ่มข้อมูลที่เข้าเทียบเรียงมีขนาดใหญ่ขึ้นจะยิ่งเห็นความแตกต่างของเวลาที่ใช้เทียบเรียงมากขึ้น โดยเมื่อเทียบเรียงกลุ่มลำดับข้อมูลจำนวน 500 ลำดับ โปรแกรม SSFA สามารถเทียบเรียงได้เร็วกว่าโปรแกรม ClustalW ถึง 13.14 เท่า ซึ่งแสดงให้เห็นว่าเทคนิคการเทียบเรียงกลุ่มลำดับข้อมูลโดยพิจารณาความถี่ส่วนย่อยของลำดับสามารถรองรับการเทียบเรียงกลุ่มข้อมูลขนาดใหญ่ได้ดีกว่าโปรแกรม ClustalW

ข้อเสนอแนะ

วิทยานิพนธ์ฉบับนี้นำเสนอวิธีการเทียบเรียงกลุ่มลำดับข้อมูลโดยพิจารณาความถี่ส่วนย่อยของลำดับ ซึ่งเป็นวิธีที่ใช้เวลาในการประมวลผลน้อย และรองรับการเทียบเรียงกลุ่มลำดับข้อมูลขนาดใหญ่ โดยที่มีความถูกต้องในระดับที่ยอมรับได้ อย่างไรก็ตามการเทียบเรียงกลุ่มลำดับข้อมูลยังคงต้องการความถูกต้องแม่นยำที่สูงกว่าปัจจุบัน ซึ่งแนวทางการพัฒนาในอนาคตควรเป็นการพัฒนาในขั้นตอนที่เหลือเพื่อเพิ่มความถูกต้องแม่นยำในการเทียบเรียงให้สูงขึ้น โดยที่ยังใช้เวลาประมวลผลน้อย และรองรับการเทียบเรียงข้อมูลขนาดใหญ่เช่นกัน

เอกสารและสิ่งอ้างอิง

คมสัน จันมา และ พันธุ์ปิติ เปี่ยมสง่า. 2548. การเทียบเรียงกลุ่มลำดับข้อมูลทางชีวภาพโดยพิจารณาความถี่ส่วนย่อยของลำดับ, น. 231-240. *ใน* **National Computer Science and Engineering Conference ครั้งที่ 9.** มหาวิทยาลัยหอการค้าไทย, กรุงเทพฯ.

จิติมนต์ เขียนดวงจันทร์. 2544. **ขั้นตอนวิธีในการจับคู่ลำดับเบสดีเอ็นเอ.** วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยเกษตรศาสตร์.

พจน์ ศรีบุญลือ, โสพิศ วงศ์ดำ และ พัชรี บุญศิริ. 2543. **ตำราชีวเคมี.** ภาควิชาชีวเคมี คณะแพทย์ศาสตร์ มหาวิทยาลัยขอนแก่น, ขอนแก่น.

ไพศาล เหล่าสุวรรณ. 2539. **พันธุศาสตร์.** ไทยวัฒนาพานิช, กรุงเทพฯ.

มุกดา ณัฏฐสมบูรณ์. 2526. **พันธุศาสตร์พื้นฐาน.** ภาควิชาพฤกษศาสตร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ.

วีรวุฒิ คงบุญเกียรติ. 2545. **การเทียบเรียงกลุ่มลำดับข้อมูลชีวภาพโดยใช้ความถี่แบบพิจารณาค่าน้ำหนัก.** วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยเกษตรศาสตร์.

อมรา คัมภิรานนท์. 2540. **พันธุศาสตร์ของเซลล์.** มหาวิทยาลัยเกษตรศาสตร์, กรุงเทพฯ.

Baldi, P. and S. Brunak. 1999. **Bioinformatics : the machine learning approch.** 3 ed. The MIT press, London, England.

Crochemore, M., G.M. Landau and M. Ziv-Ukelson. 2003. A Subquadratic Sequence Alignment Algorithm for Unrestricted Scoring Matrices. **SIAM Journal on Computing** 32 (6): 1654-1673.

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high troughput. **Nucleic Acids Research** 32 (5): 1792-1797.

Gusfield, D. 1999. **Algorithms on strings, trees, and sequences : computer science and computation biology..** 2 ed. The press syndicate of The University of Cambridge, New York, USA.

Notredame, C. 2002. Recent progress in multiple sequence alignment: a survey. **Pharmacogenomics** 3 (1): 131-144.

Saitou, N. and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees.. **Molecular Biology and Evolution** 4 (4): 406-425.

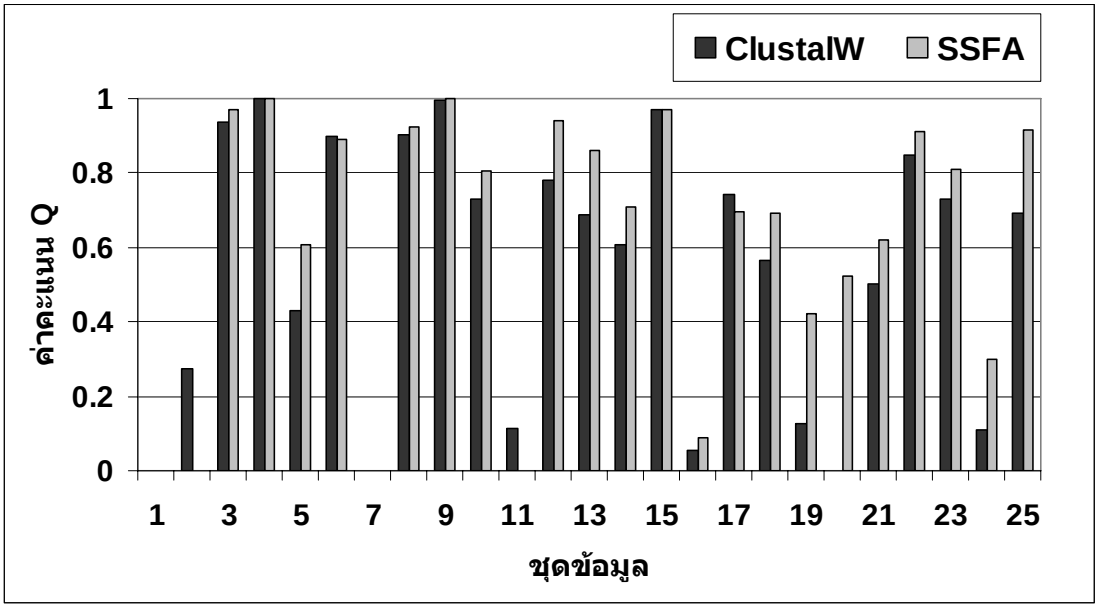
Thompson, J.D., D.G. Higgins and T.J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.. **Nucleic Acids Research** 22 (22): 4673-4680.

_____, _____, _____, F. Plewniak and F. Jeanmougin. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. **Nucleic Acids Research** 25 (24): 4876-4882

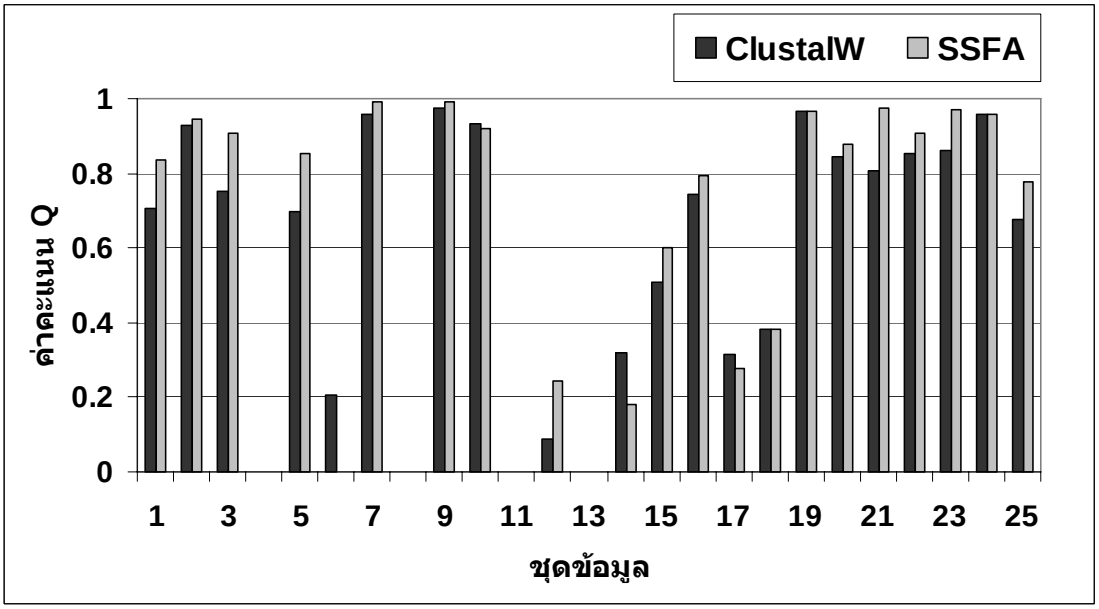
_____, _____ and O. Poch. 1999. A comprehensive comparison of multiple sequence alignment programs. **Nucleic Acids Research** 27 (13): 2682-2690.

ภาคผนวก

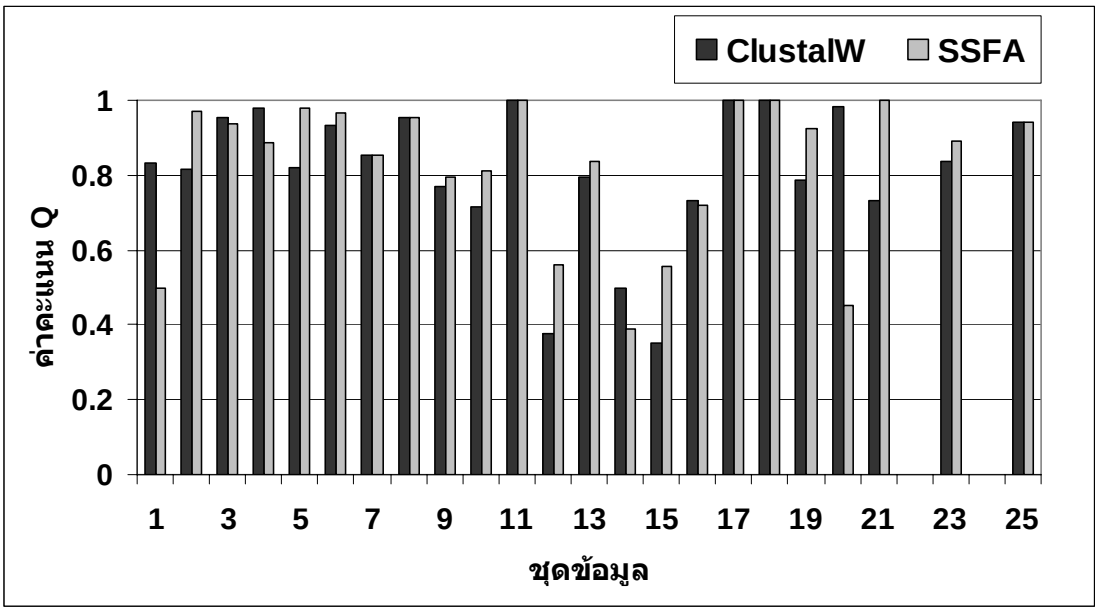
ผลการเทียบเรียงชุดข้อมูลของฐานข้อมูล PREFAB แยกตามกลุ่มข้อมูล



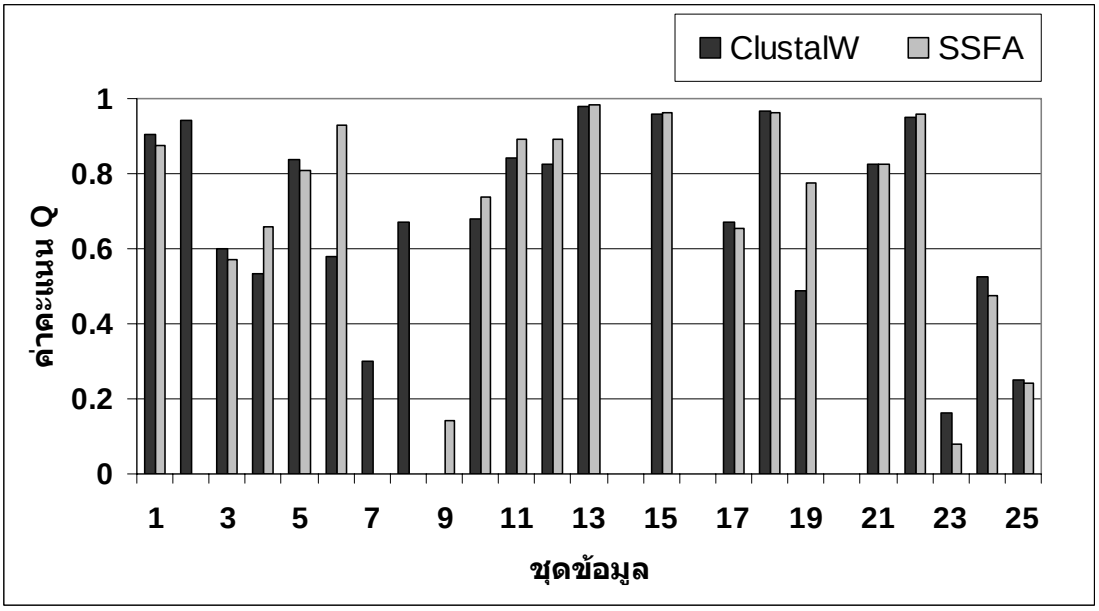
ภาพผนวกที่ 1 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 1



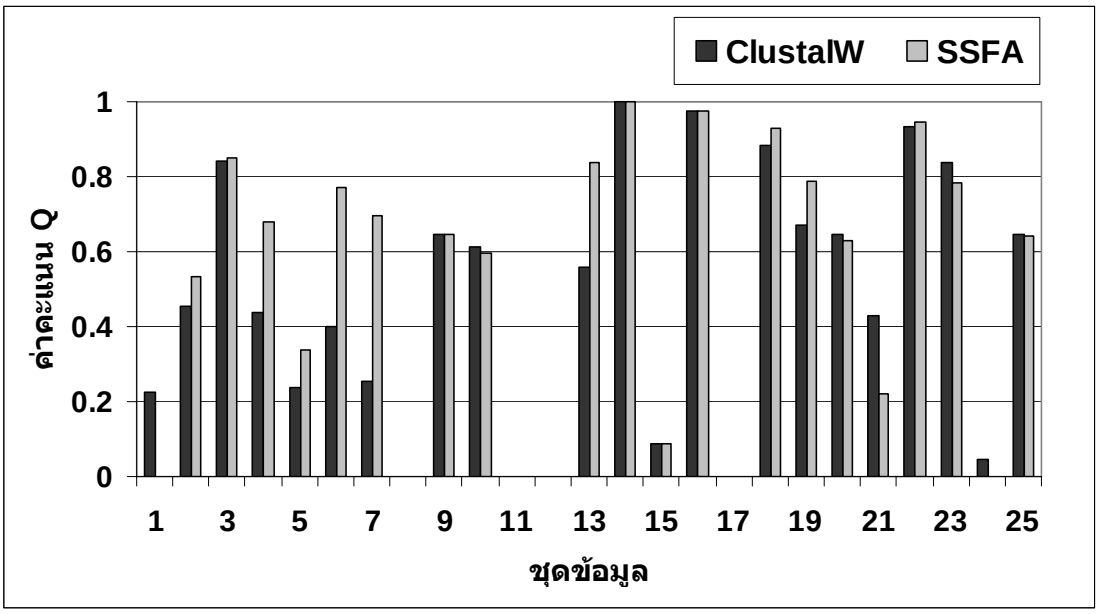
ภาพผนวกที่ 2 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 2



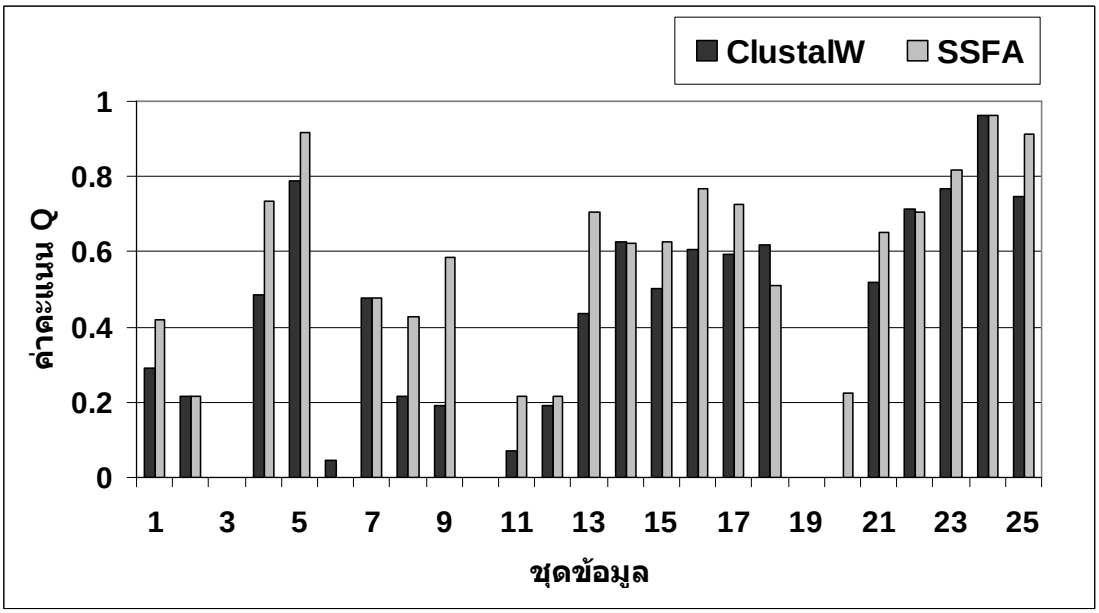
ภาพผนวกที่ 3 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 3



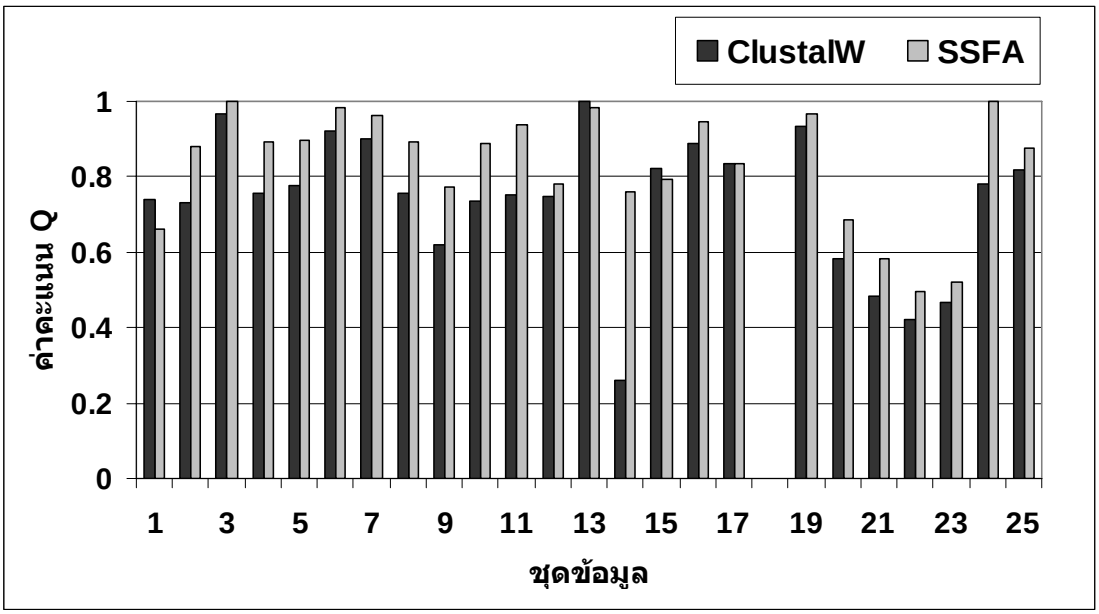
ภาพผนวกที่ 4 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 4



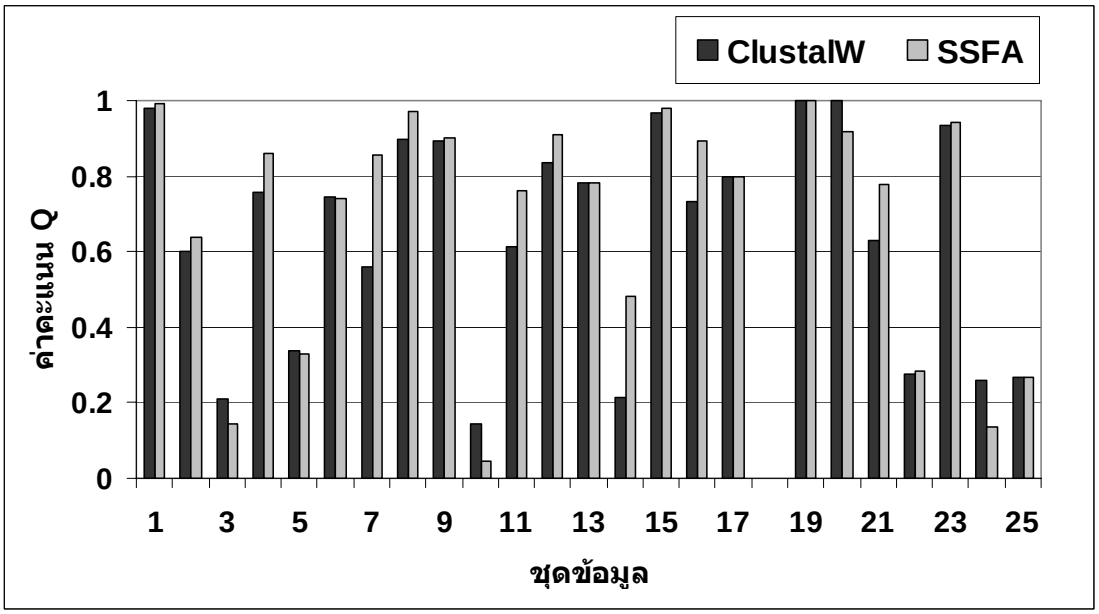
ภาพผนวกที่ 5 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 5



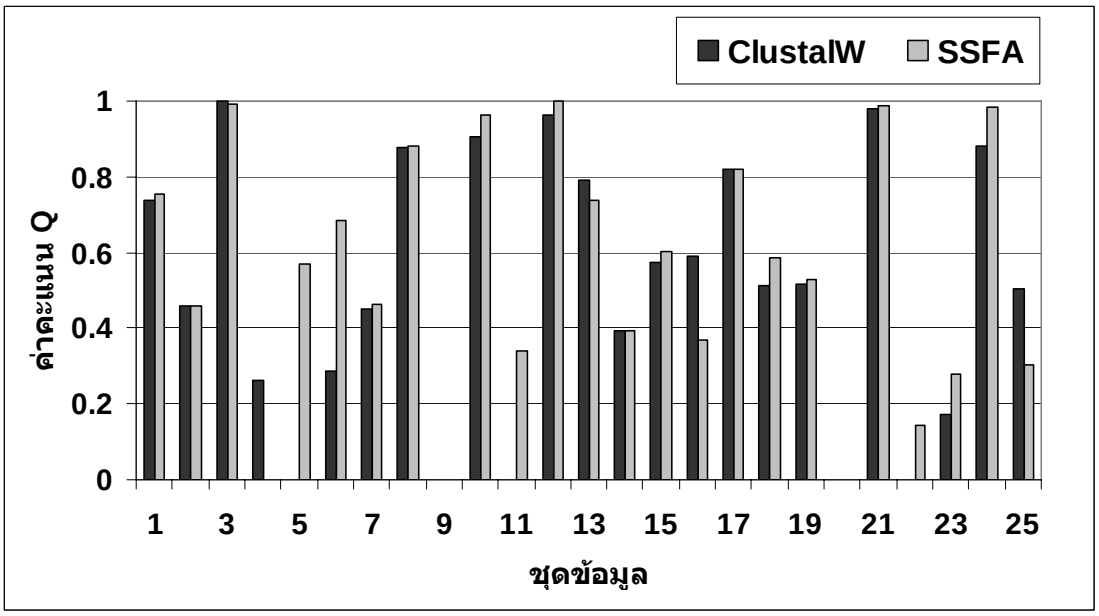
ภาพผนวกที่ 6 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 6



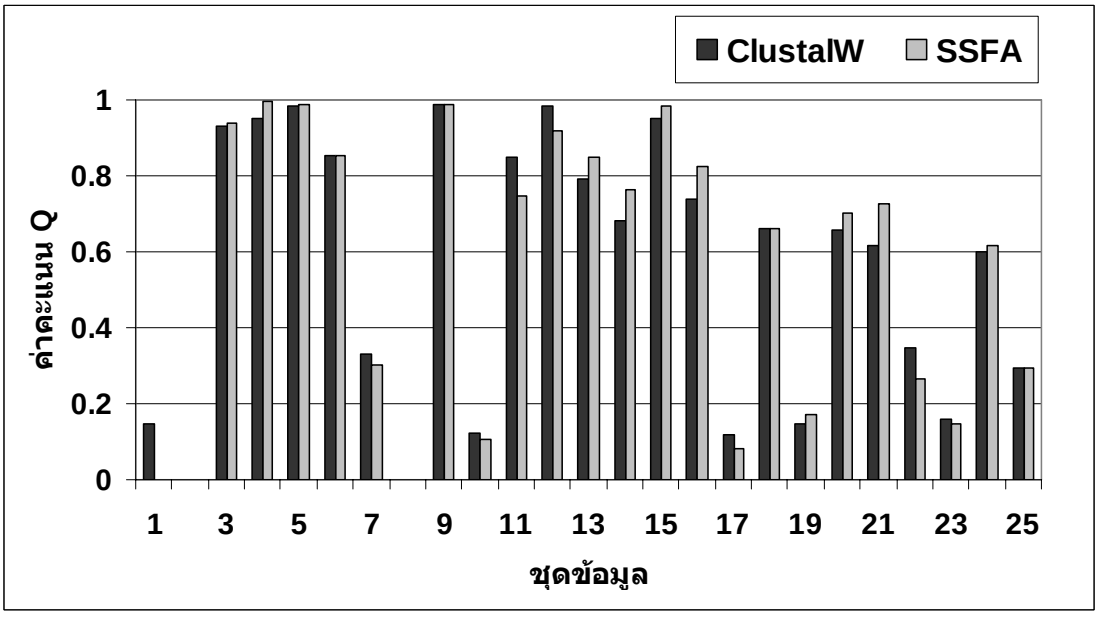
ภาพผนวกที่ 7 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 7



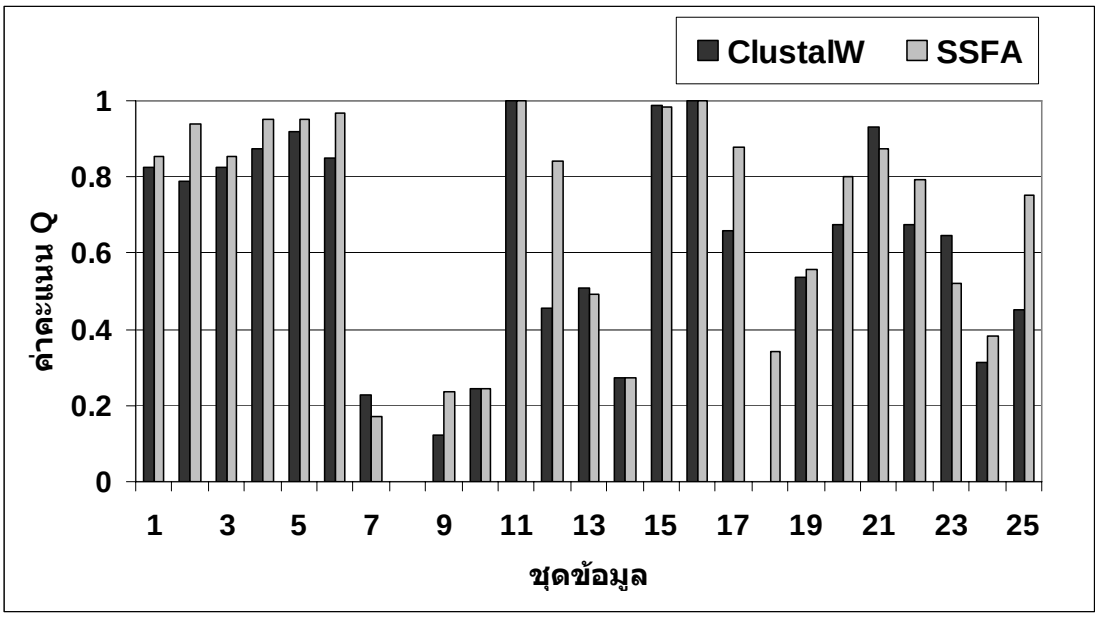
ภาพผนวกที่ 8 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 8



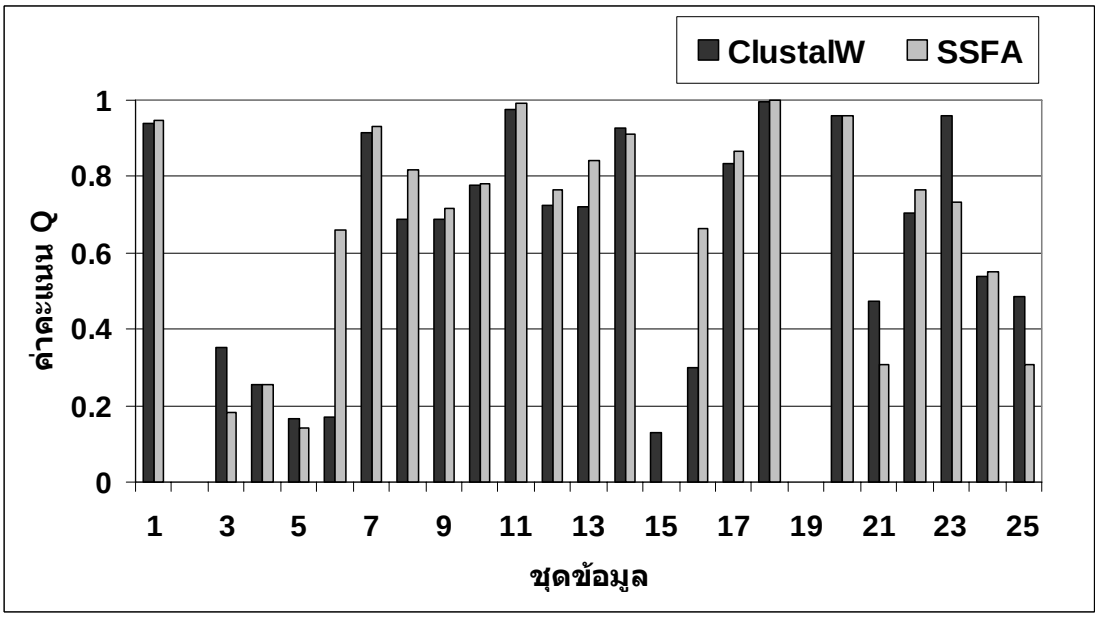
ภาพผนวกที่ 9 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 9



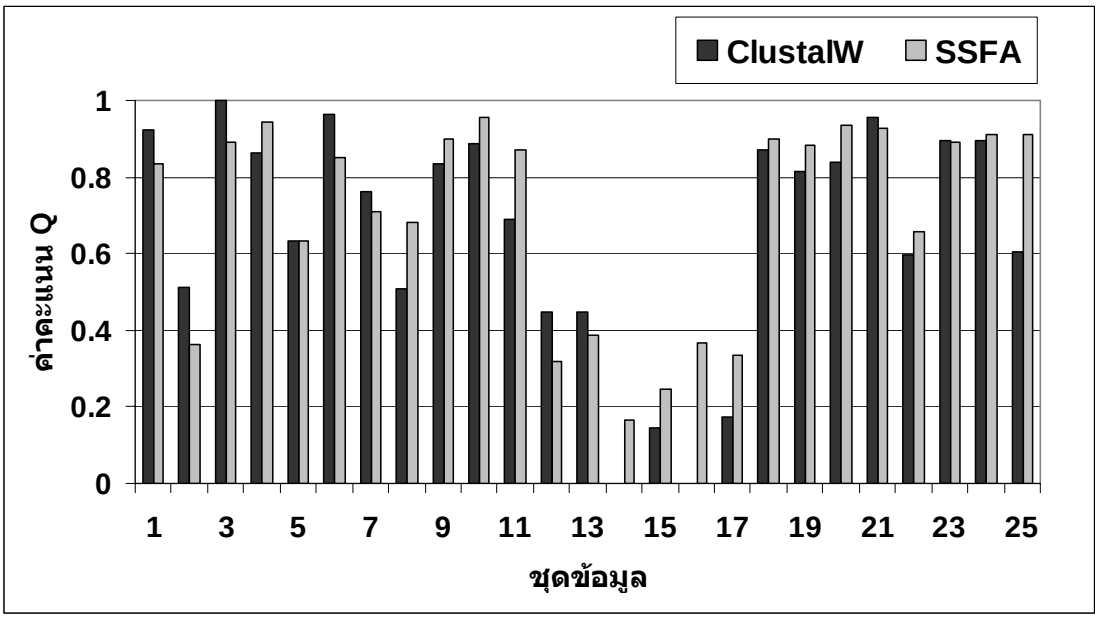
ภาพผนวกที่ 10 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 10



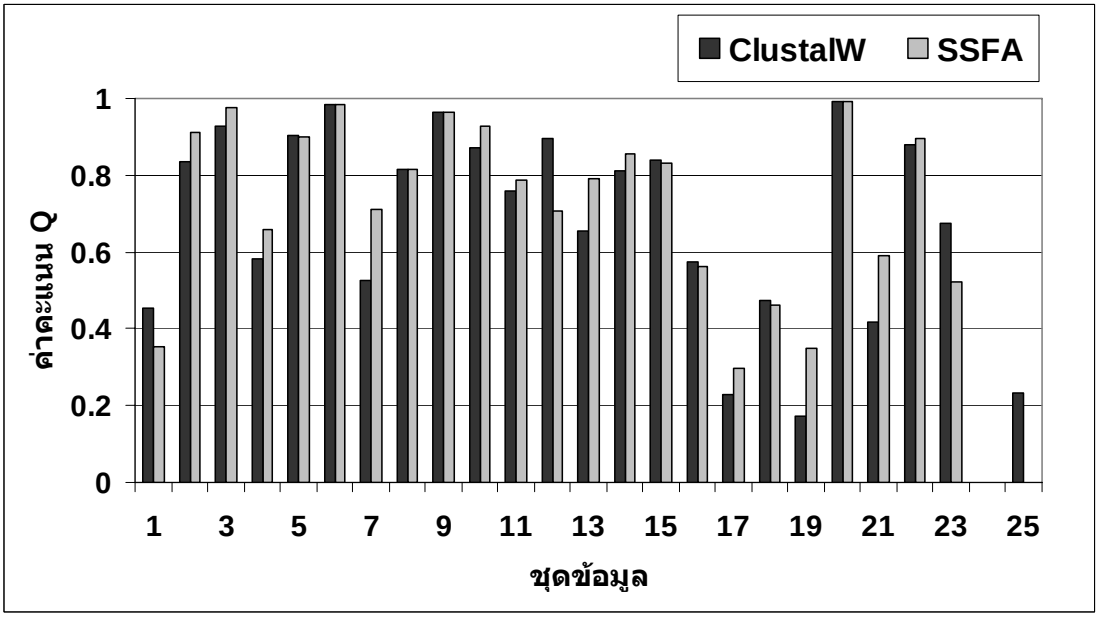
ภาพผนวกที่ 11 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 11



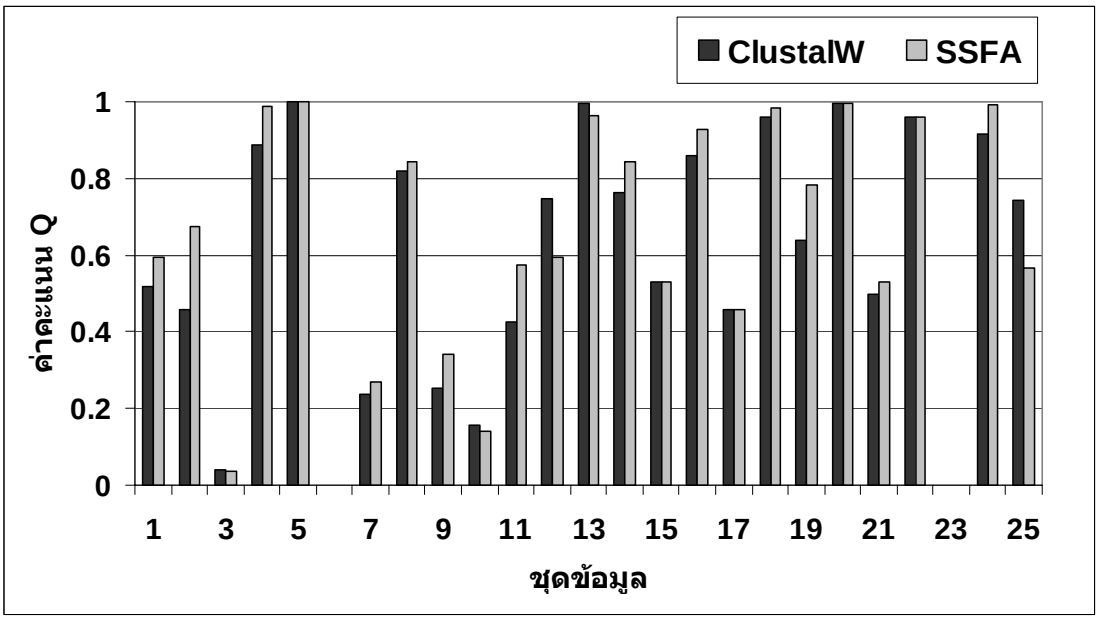
ภาพผนวกที่ 12 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 12



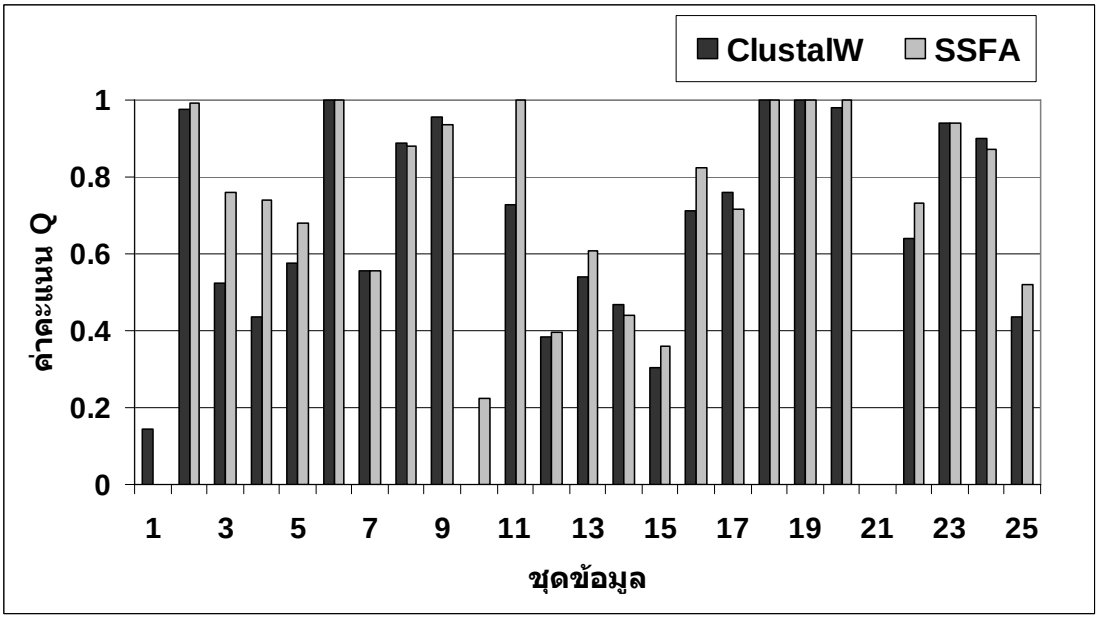
ภาพผนวกที่ 13 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 13



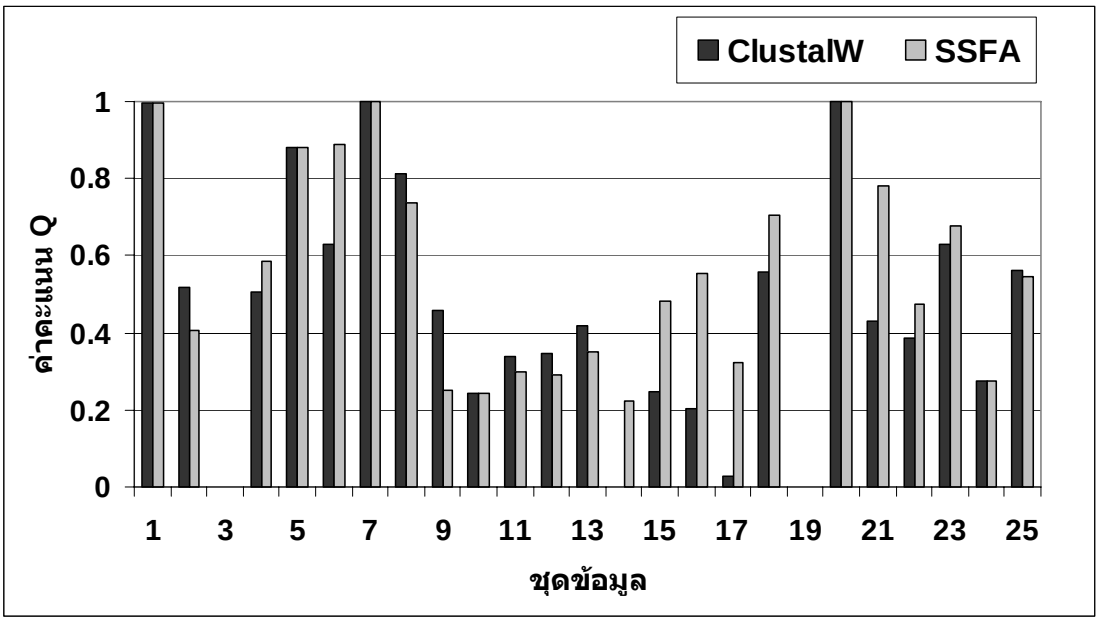
ภาพผนวกที่ 14 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 14



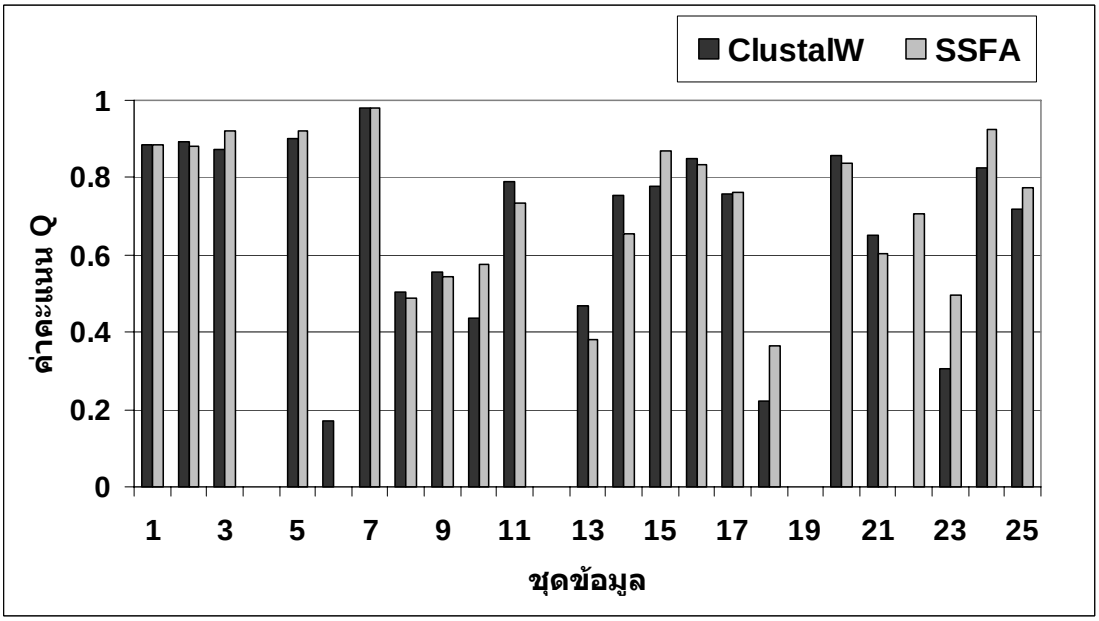
ภาพผนวกที่ 15 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 15



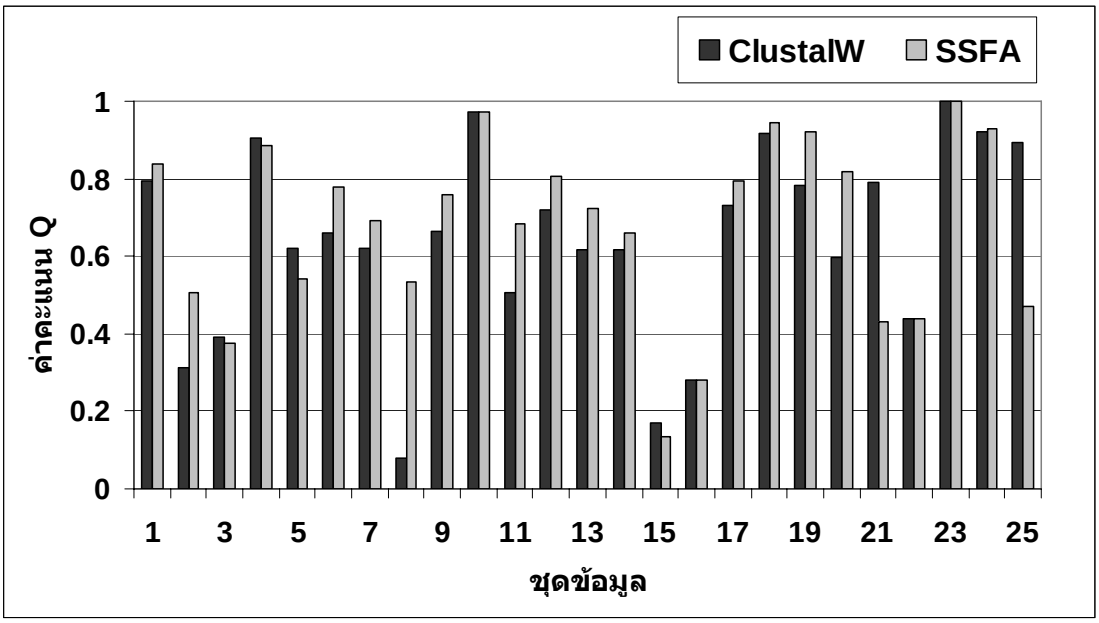
ภาพผนวกที่ 16 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 16



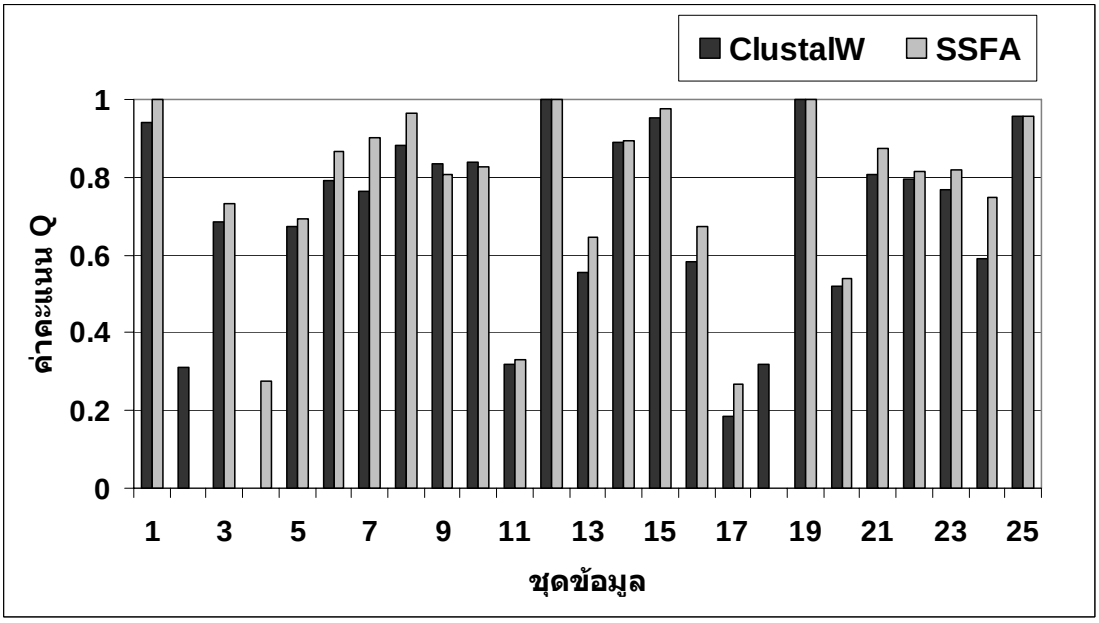
ภาพผนวกที่ 17 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 17



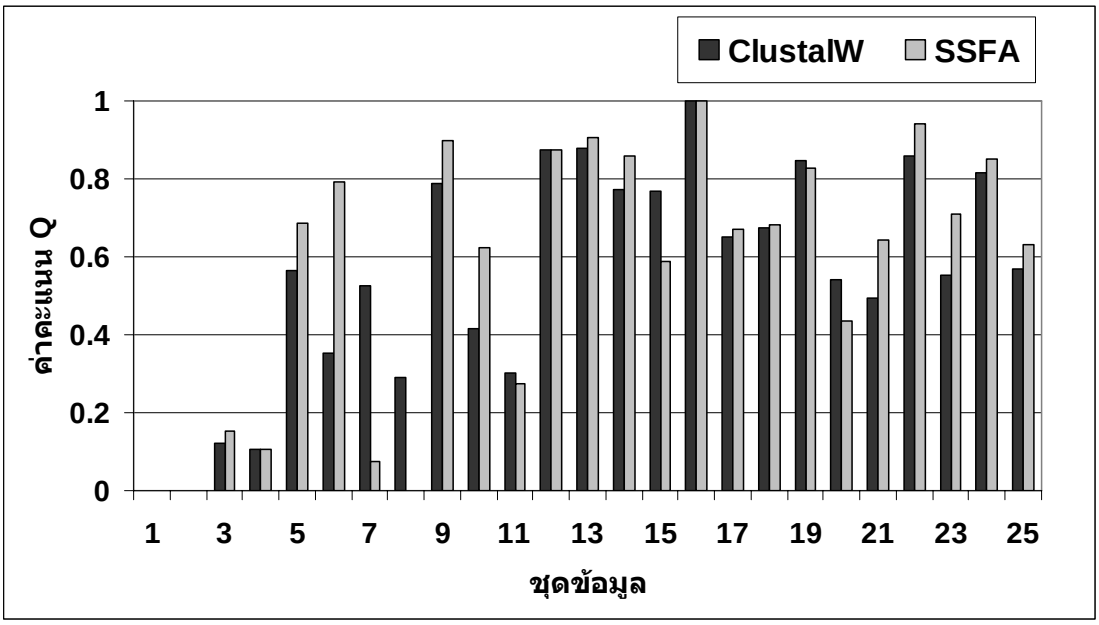
ภาพผนวกที่ 18 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 18



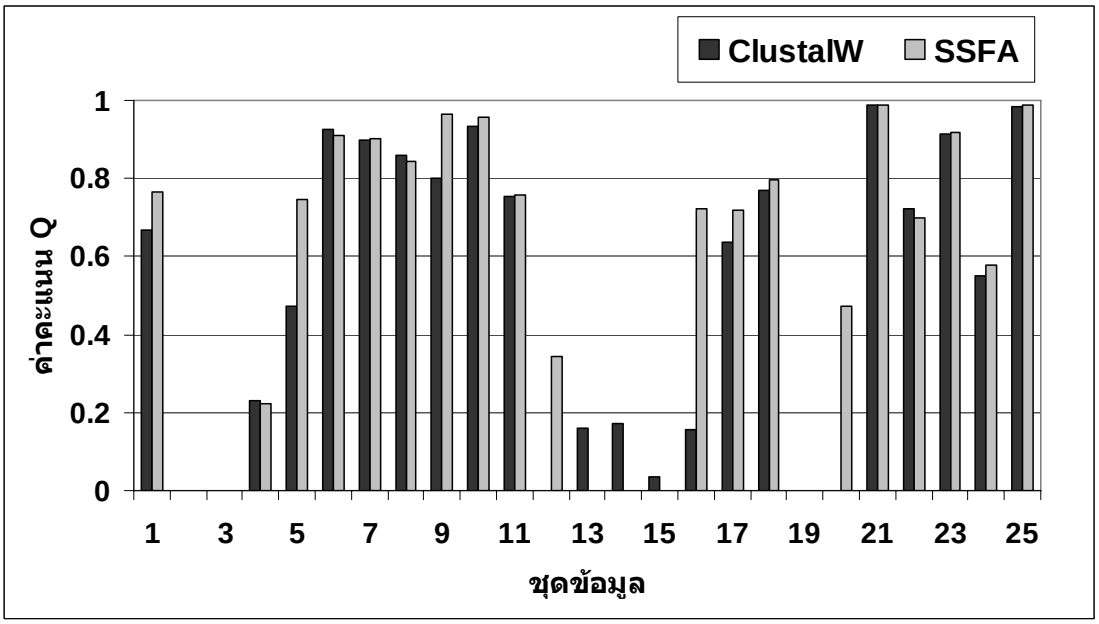
ภาพผนวกที่ 19 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 19



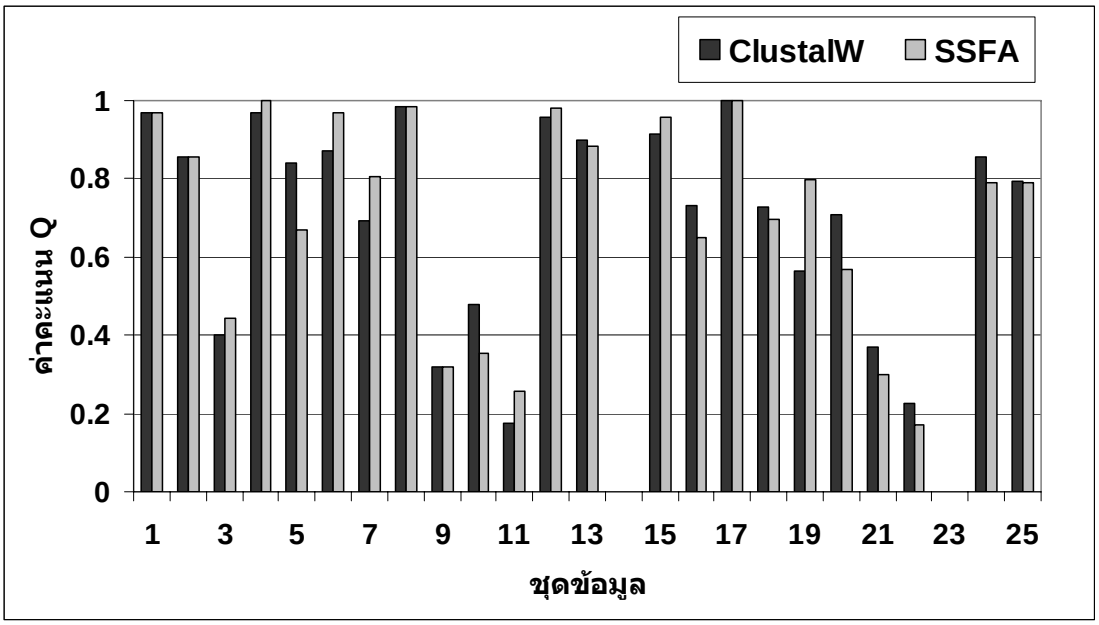
ภาพผนวกที่ 20 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 20



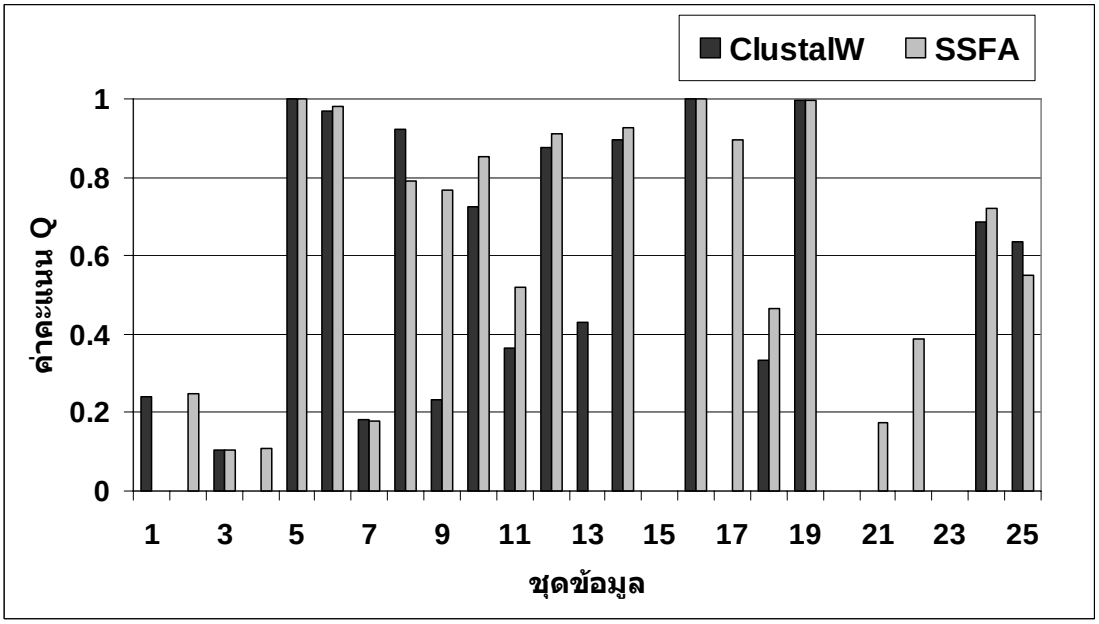
ภาพผนวกที่ 21 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 21



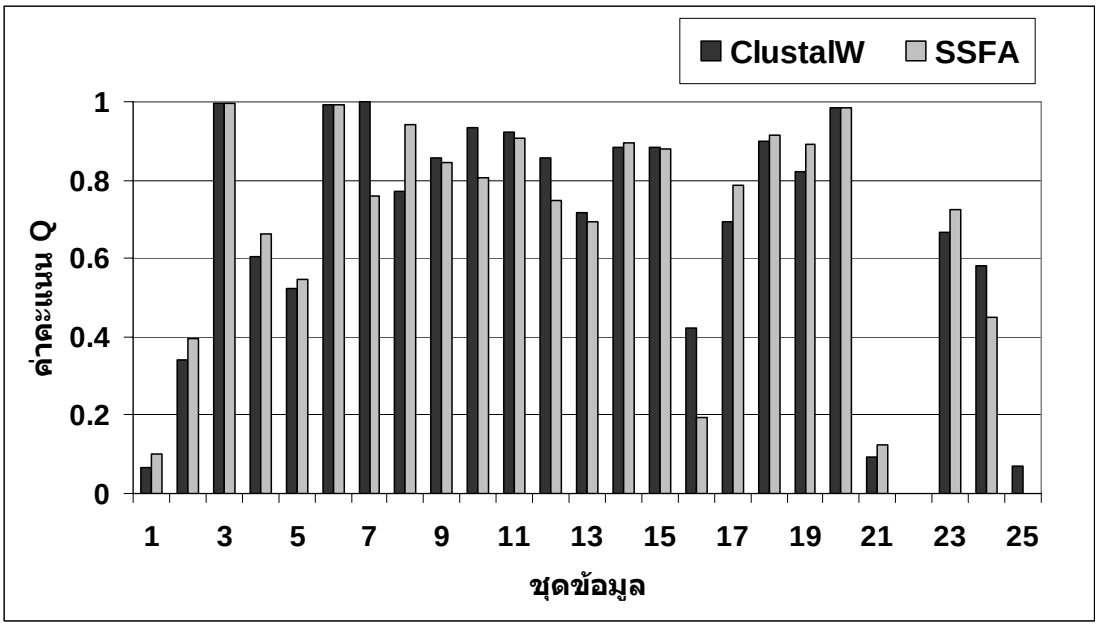
ภาพผนวกที่ 22 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 22



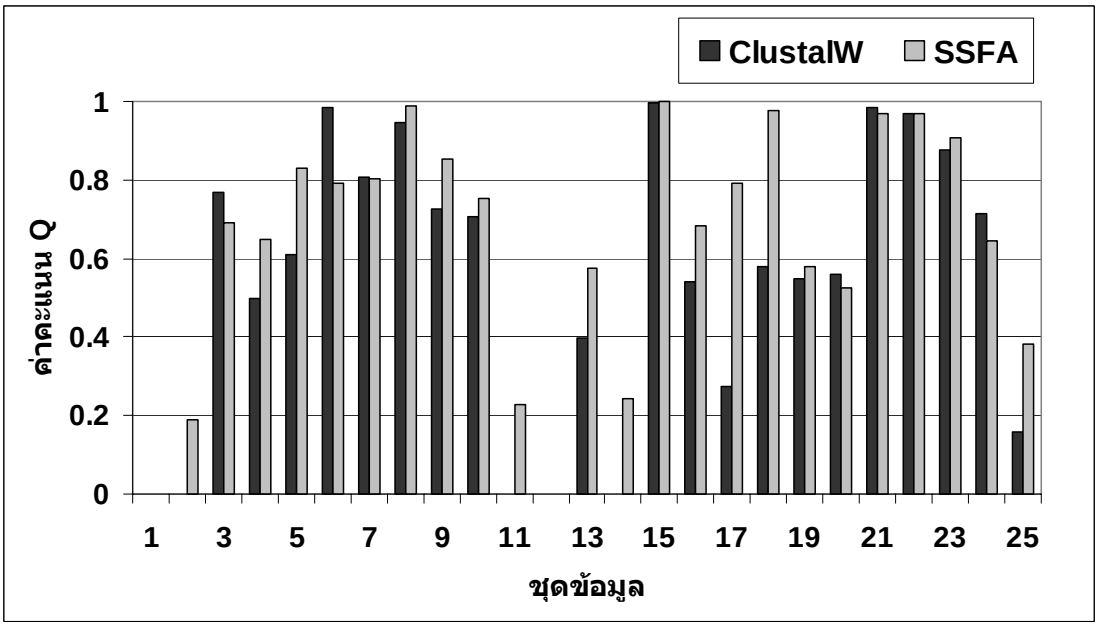
ภาพผนวกที่ 23 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 23



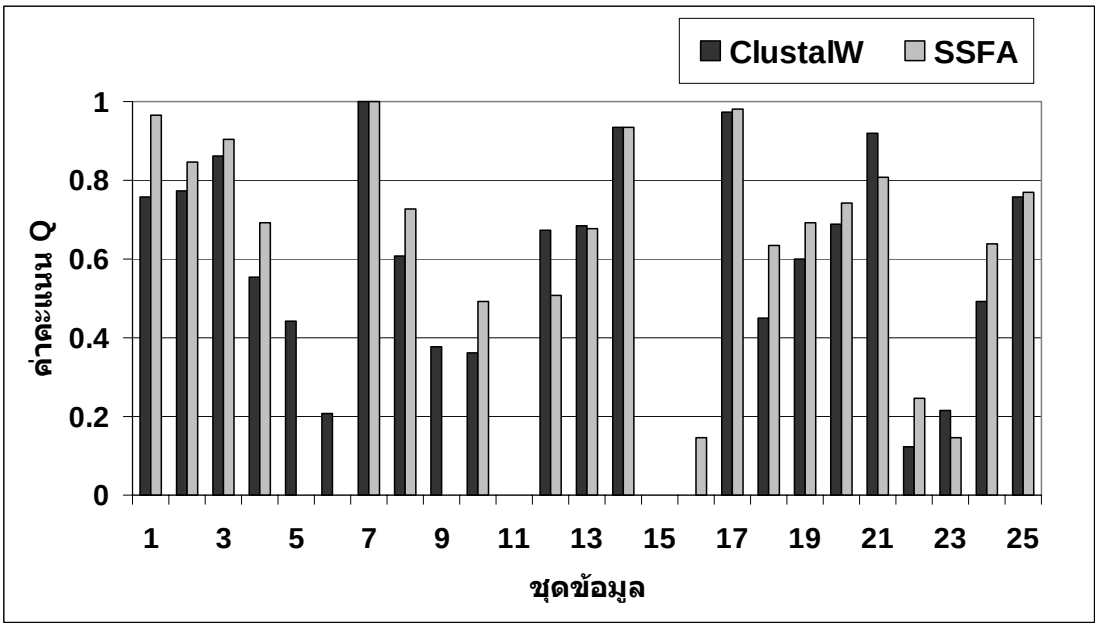
ภาพผนวกที่ 24 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 24



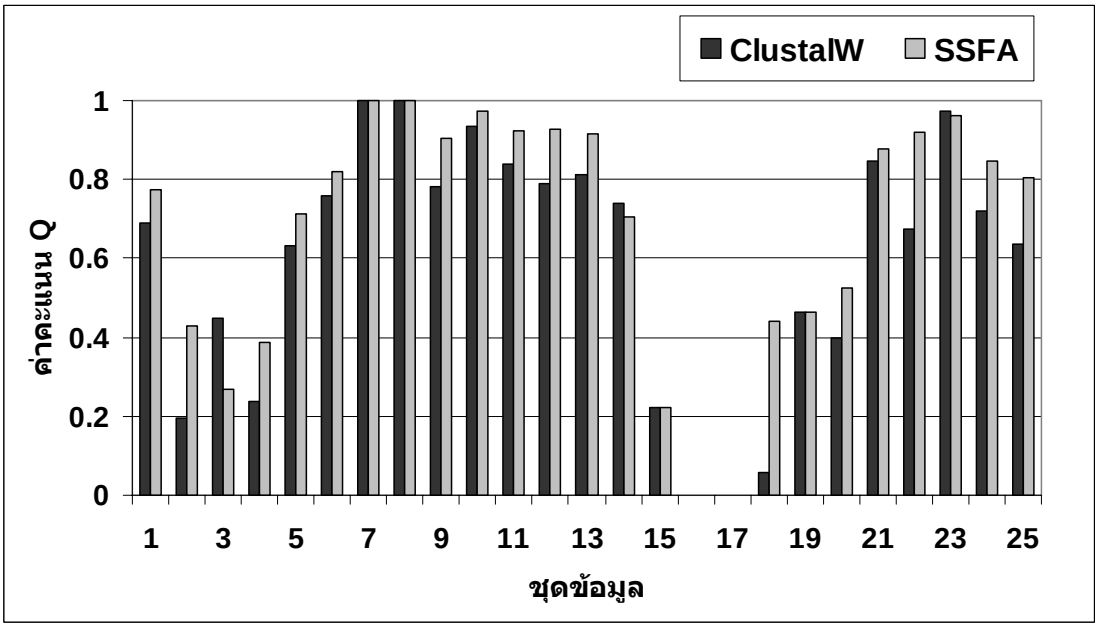
ภาพผนวกที่ 25 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 25



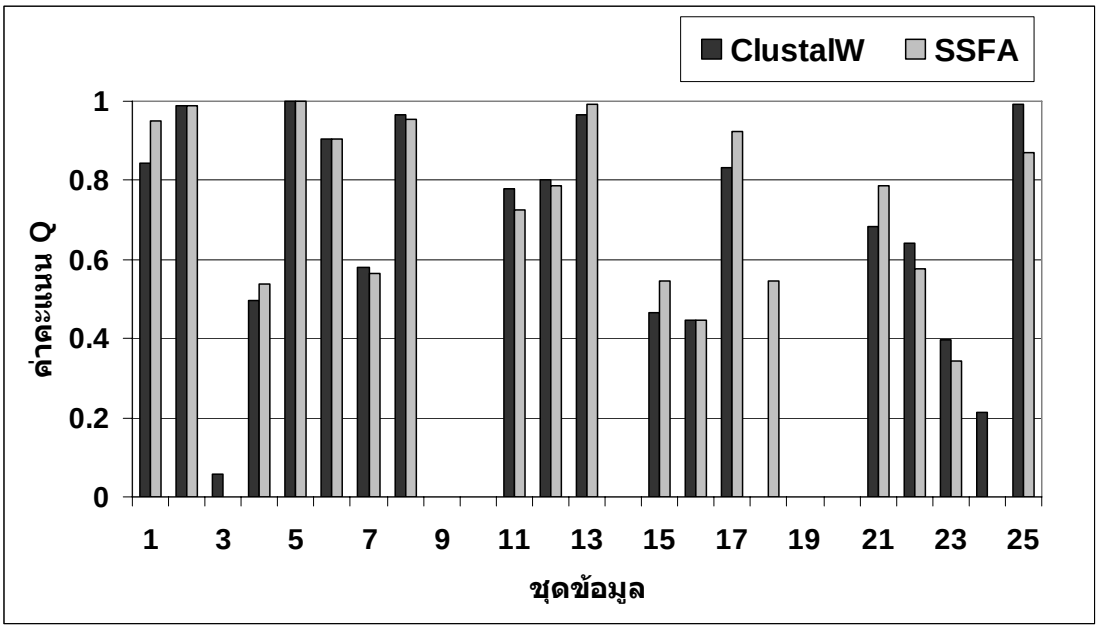
ภาพผนวกที่ 26 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 26



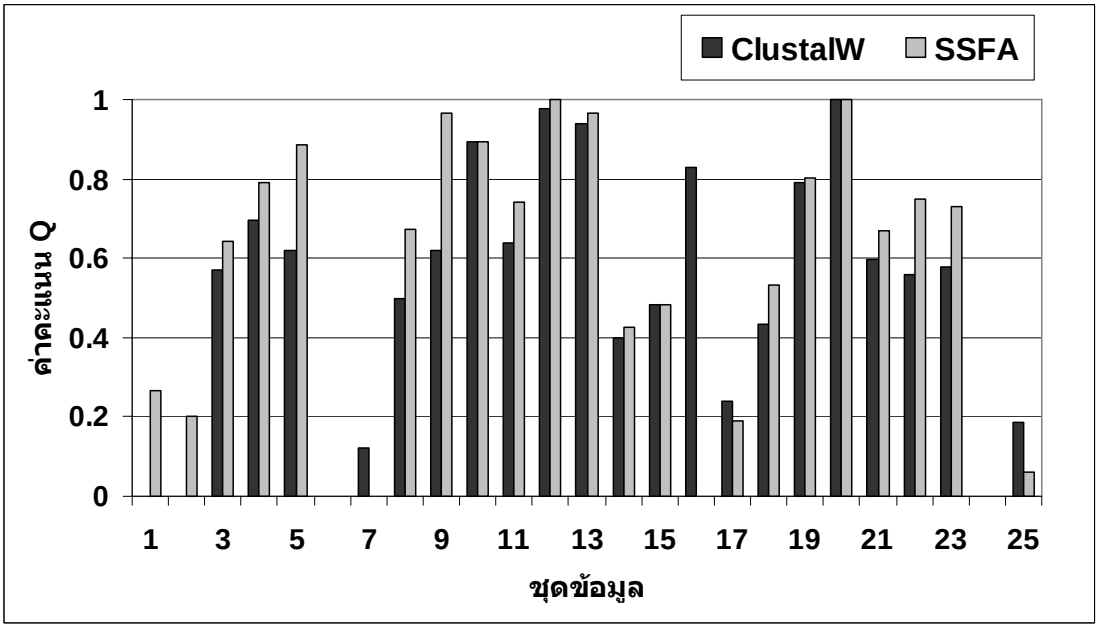
ภาพผนวกที่ 27 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 27



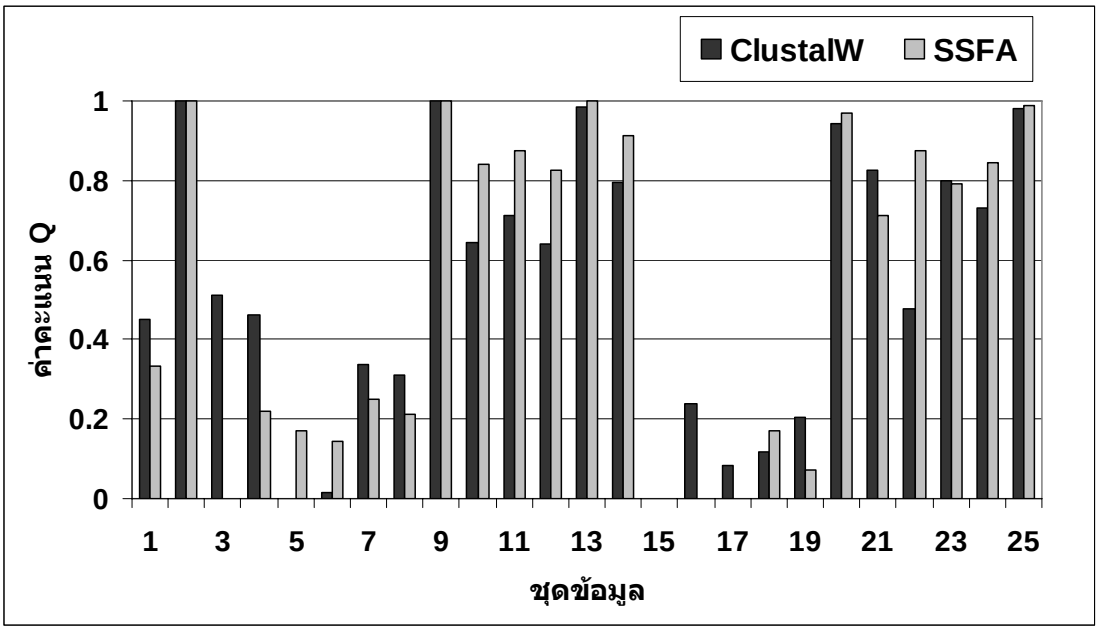
ภาพผนวกที่ 28 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 28



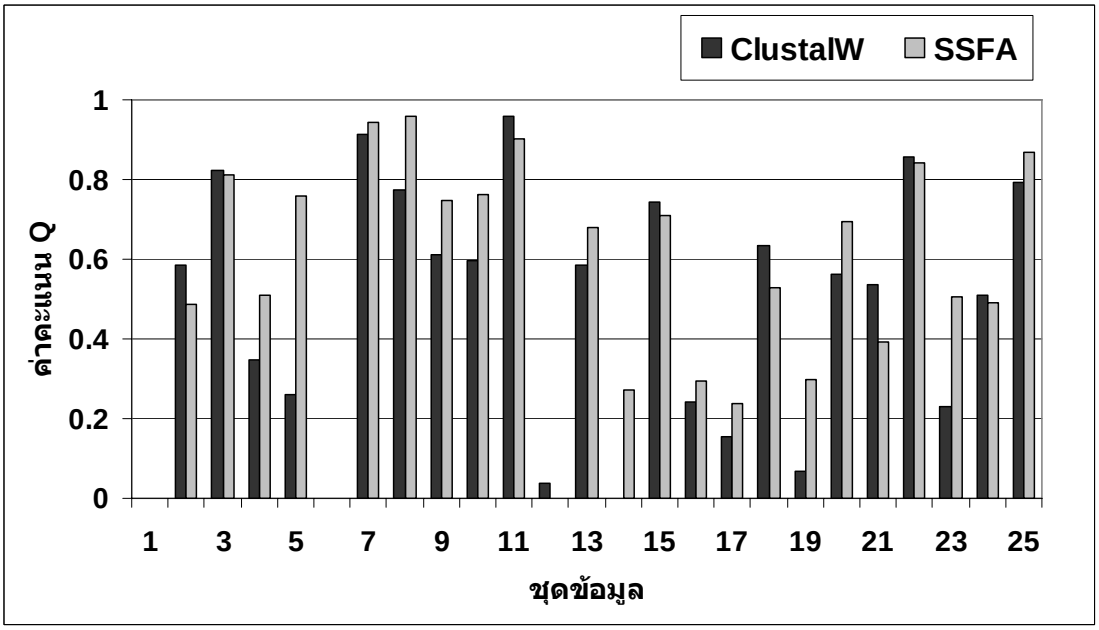
ภาพผนวกที่ 29 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 29



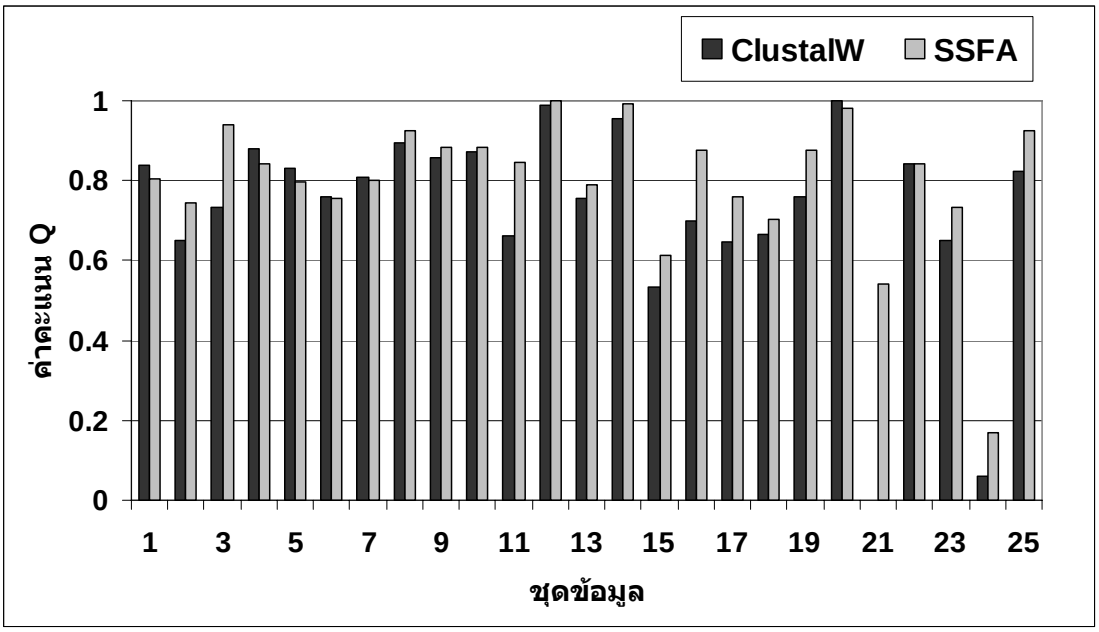
ภาพผนวกที่ 30 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 30



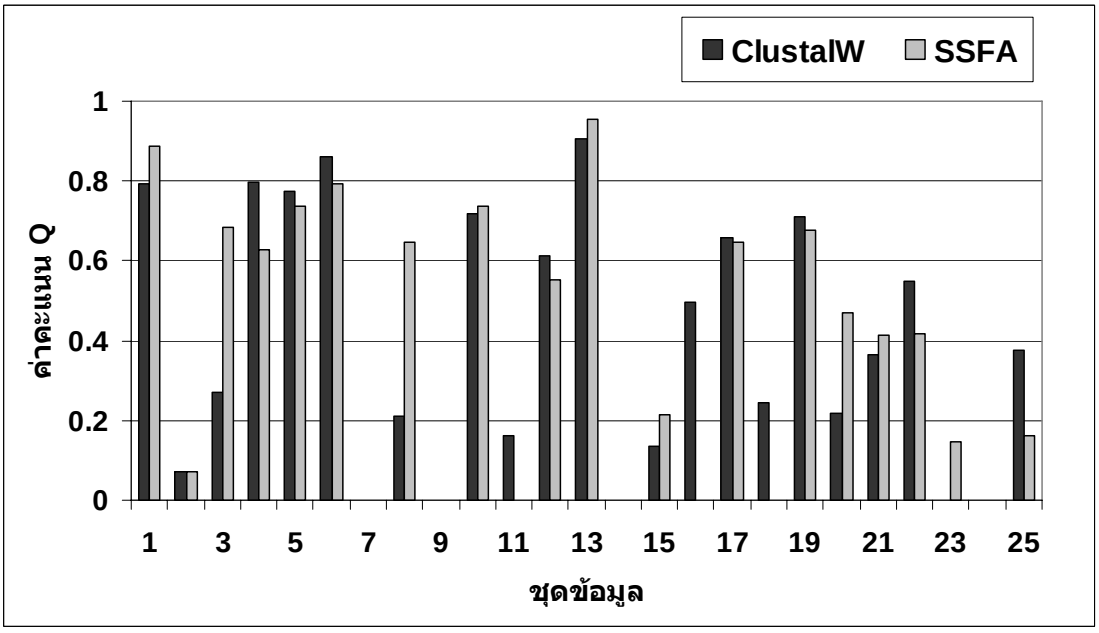
ภาพผนวกที่ 31 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 31



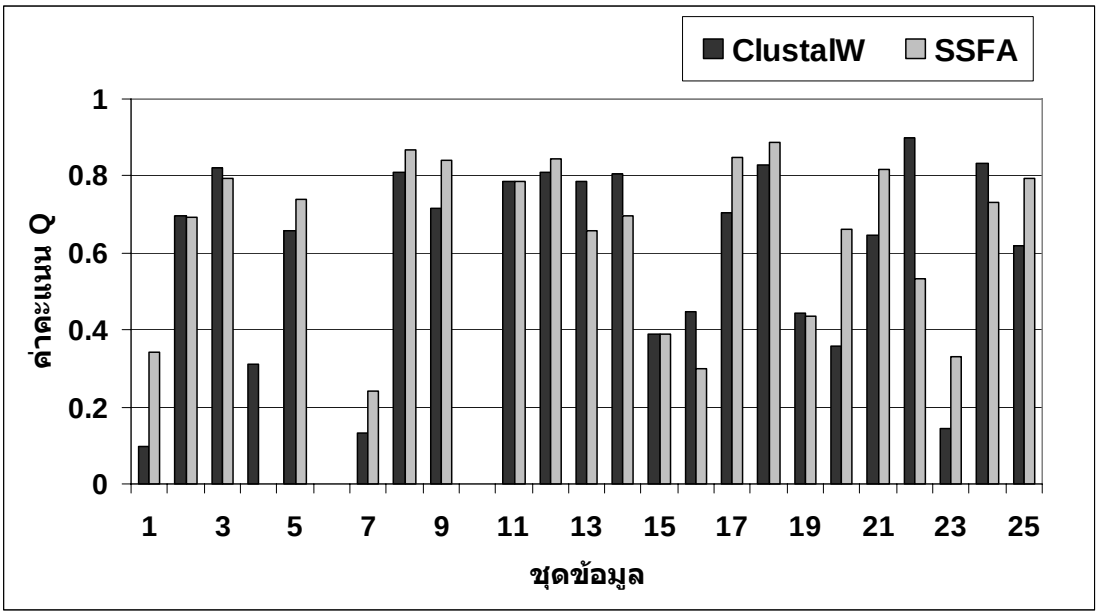
ภาพผนวกที่ 32 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 32



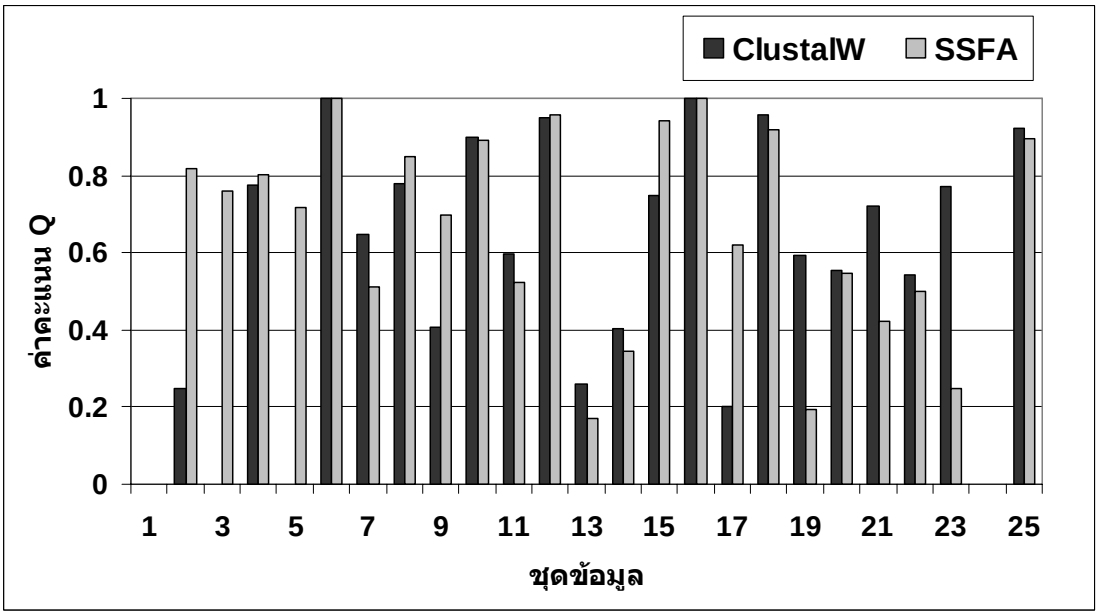
ภาพผนวกที่ 33 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 33



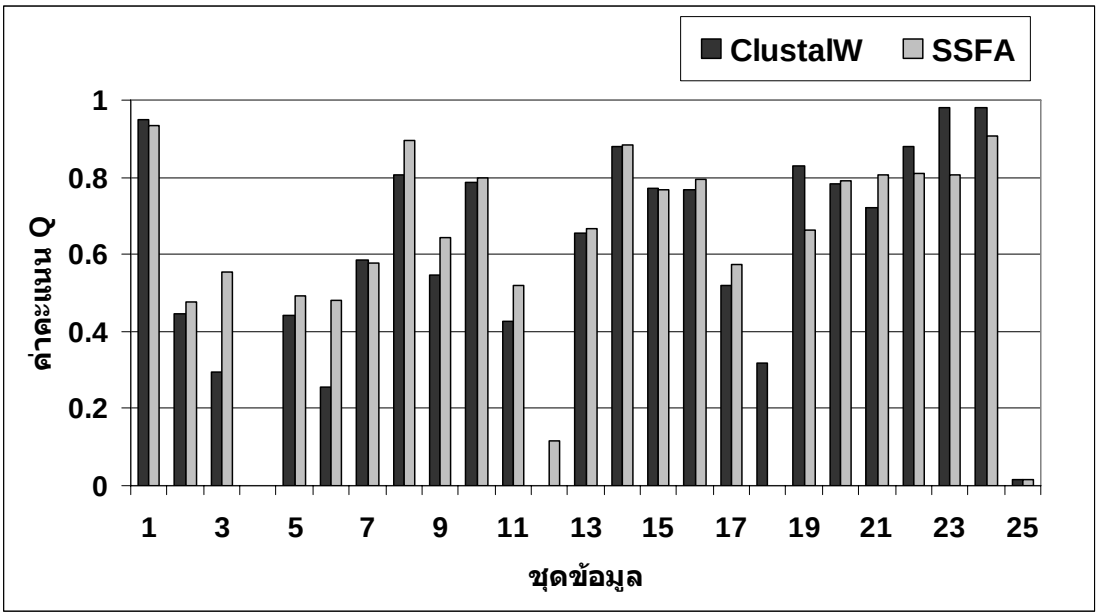
ภาพผนวกที่ 34 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 34



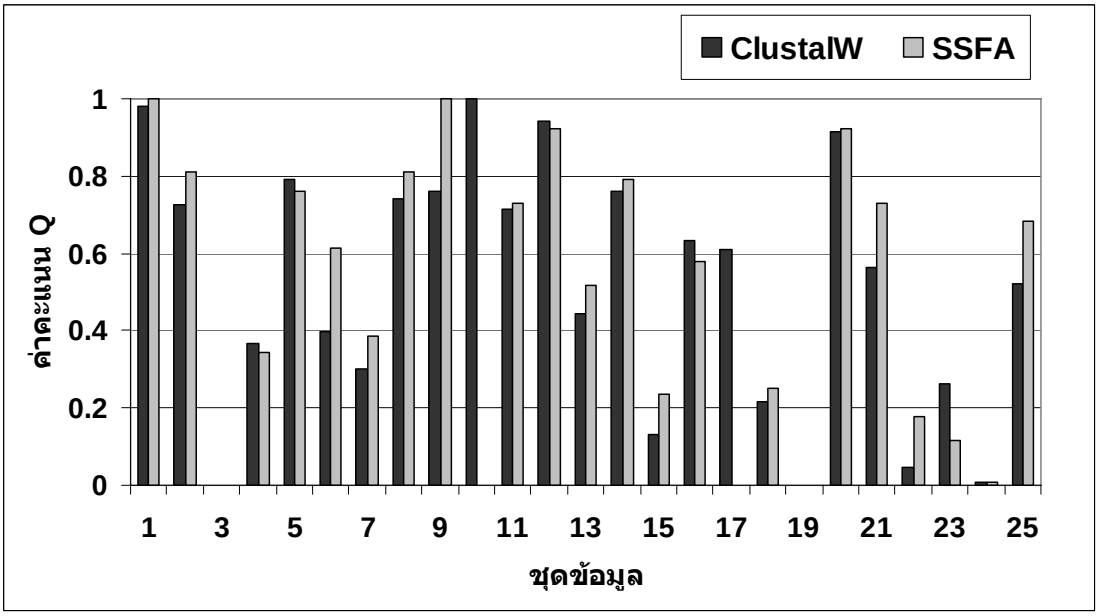
ภาพผนวกที่ 35 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 35



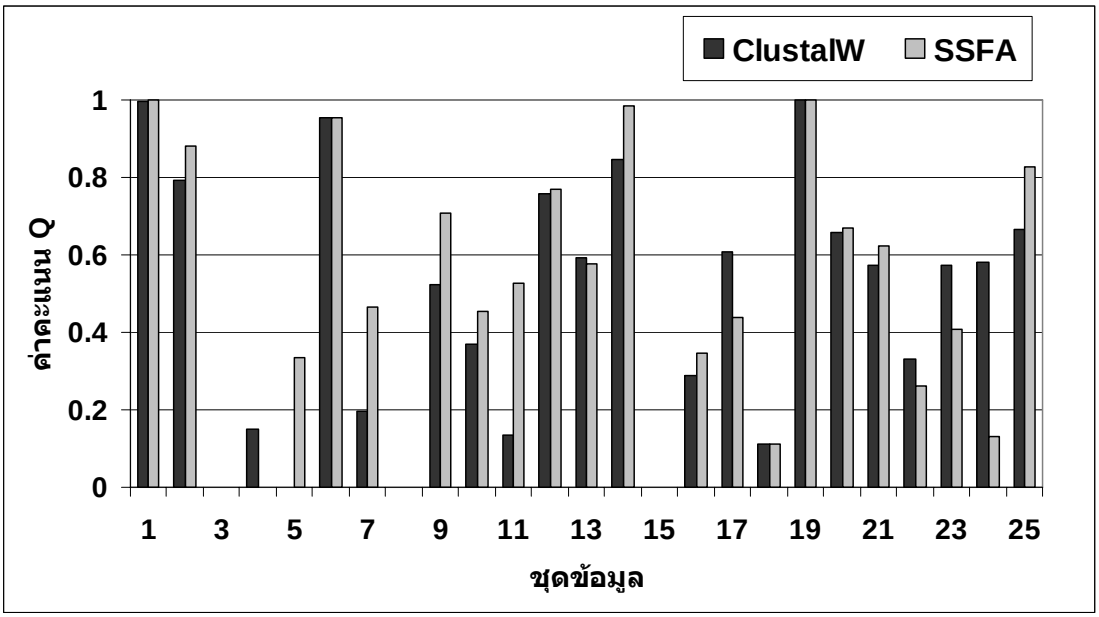
ภาพผนวกที่ 36 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 36



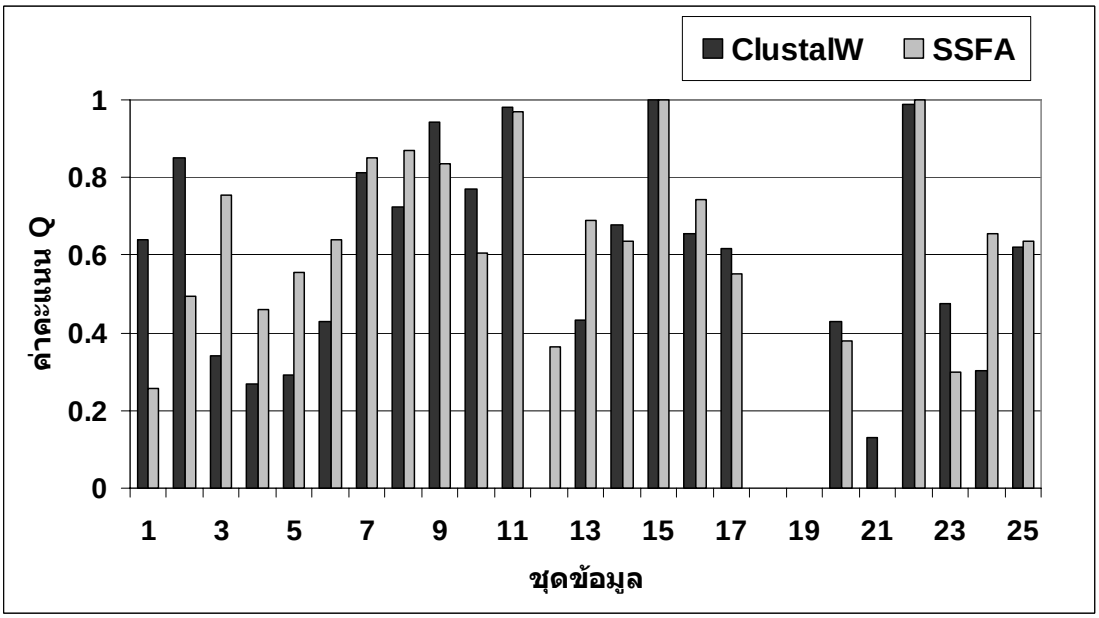
ภาพผนวกที่ 37 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 37



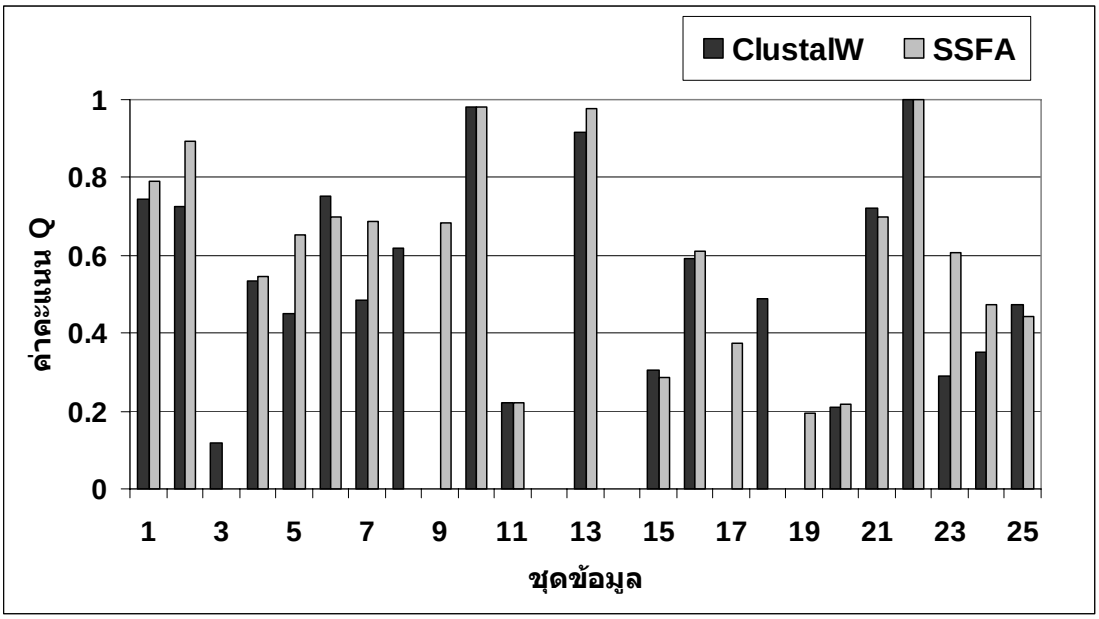
ภาพผนวกที่ 38 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 38



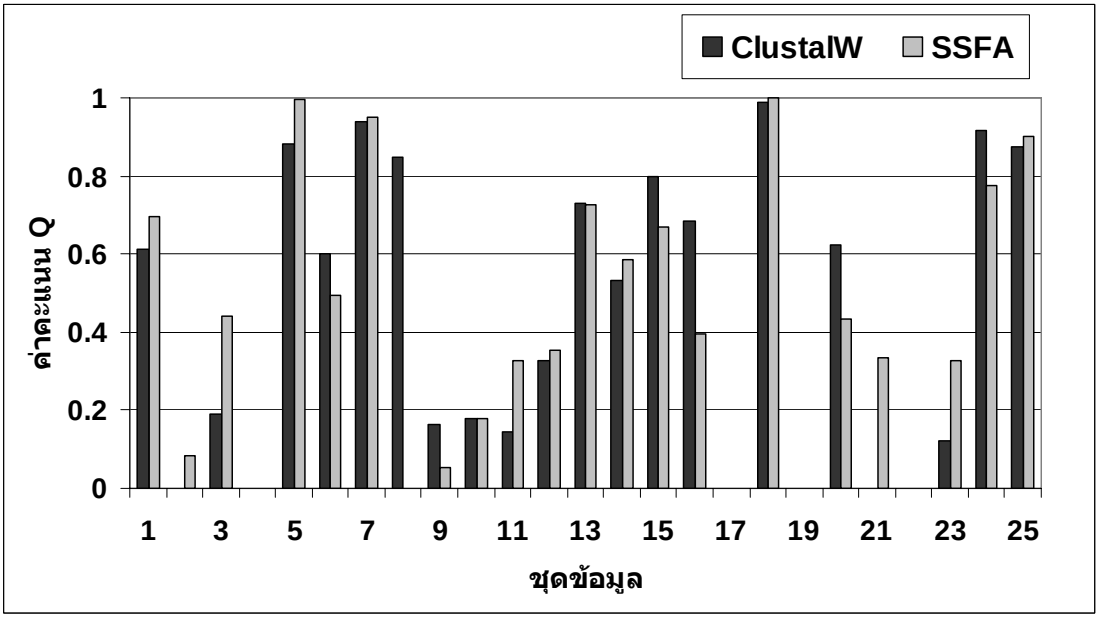
ภาพผนวกที่ 39 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 39



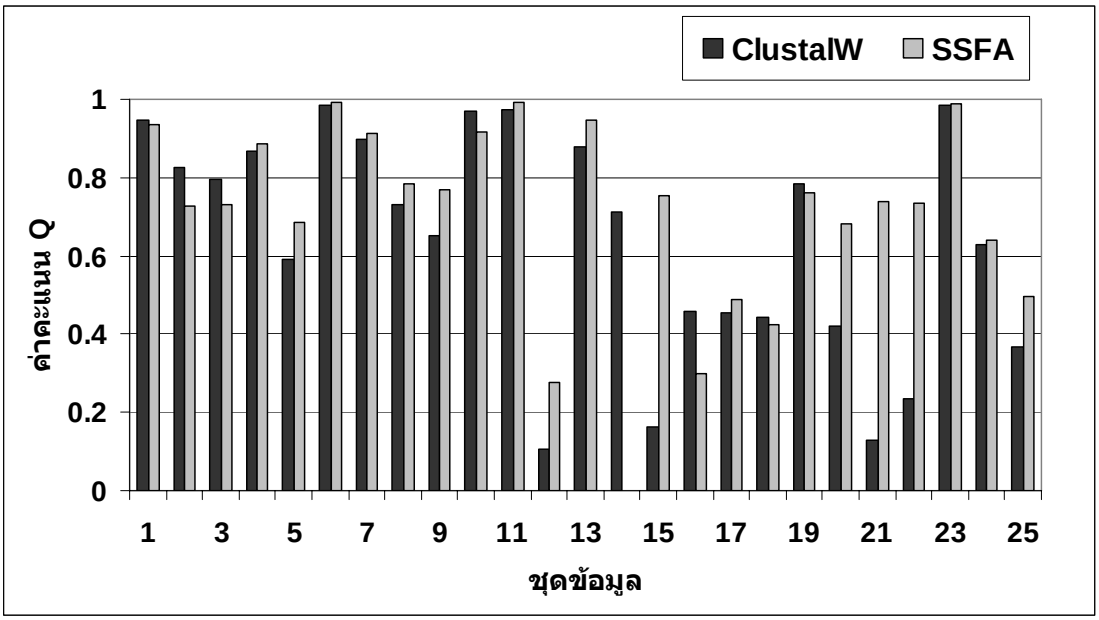
ภาพผนวกที่ 40 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 40



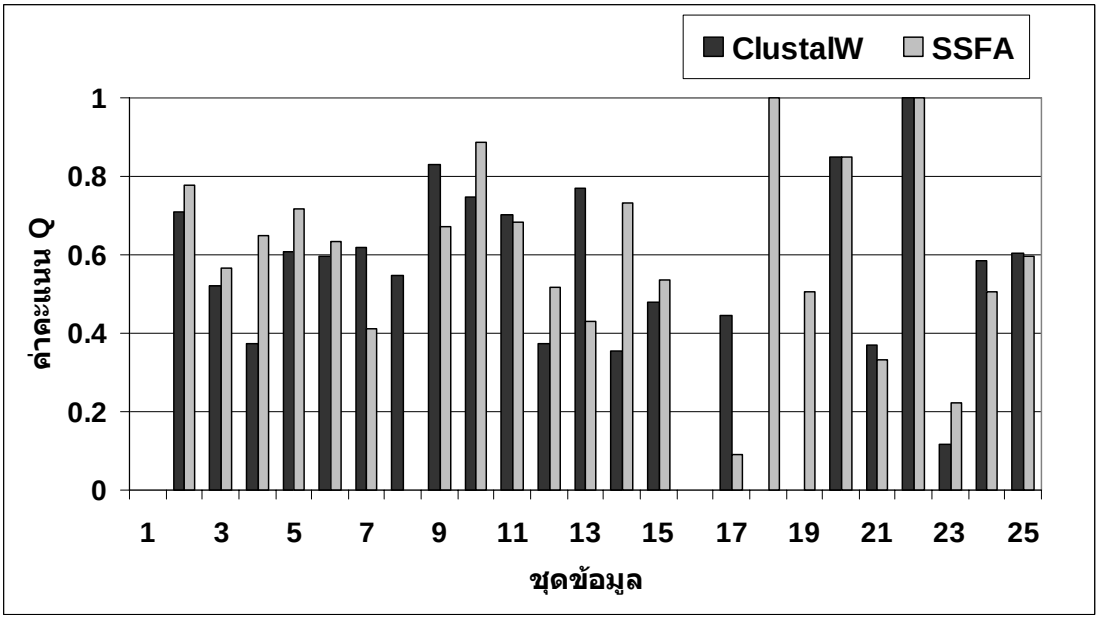
ภาพผนวกที่ 41 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 41



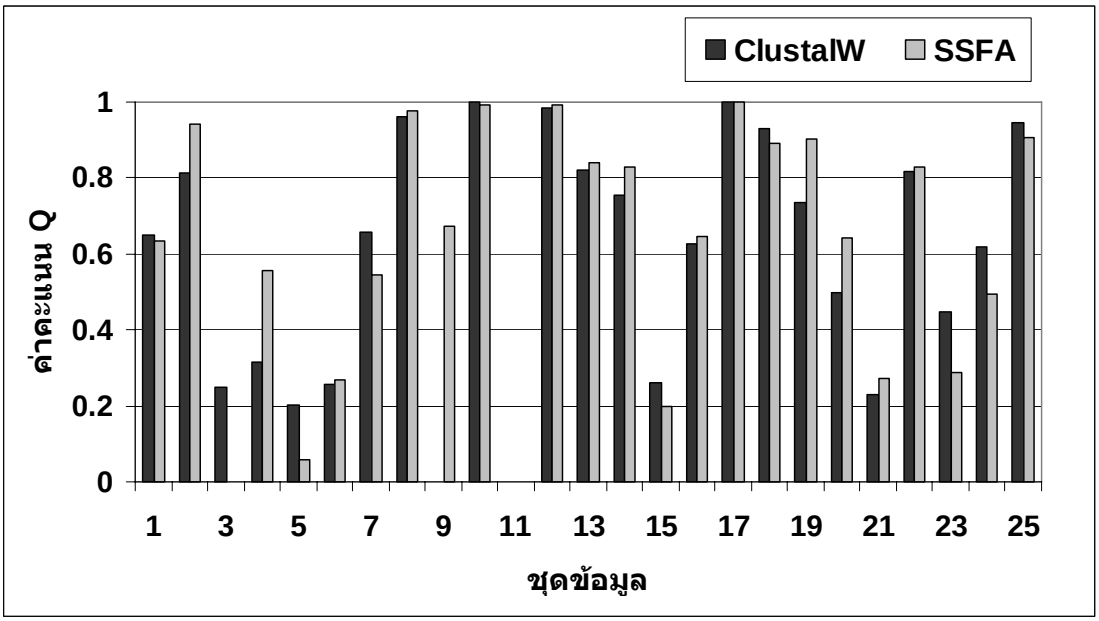
ภาพผนวกที่ 42 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 42



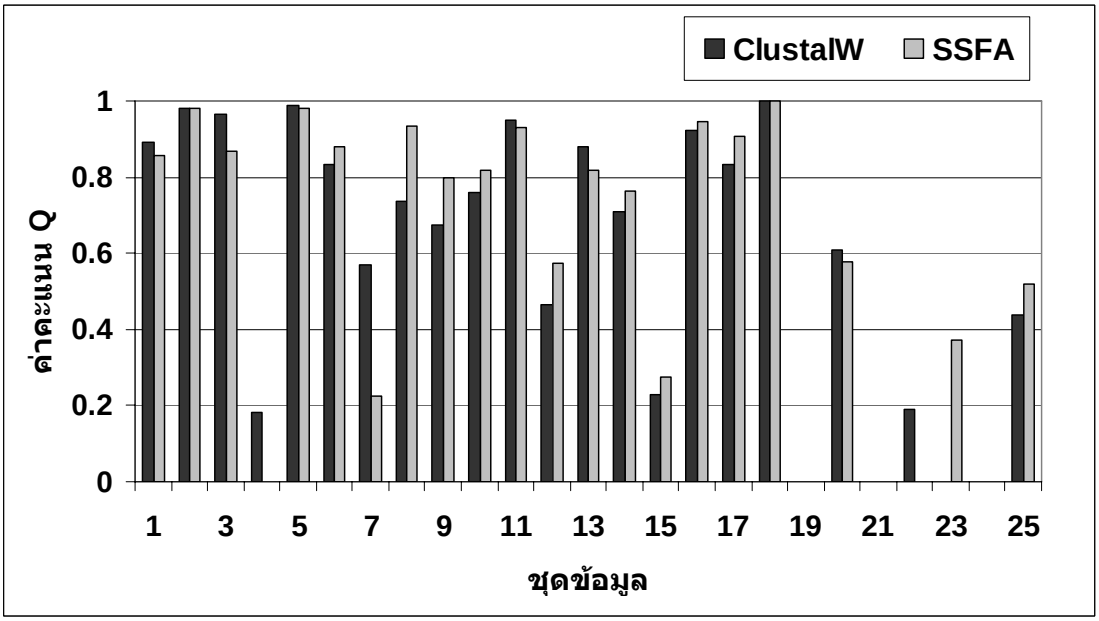
ภาพผนวกที่ 43 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 43



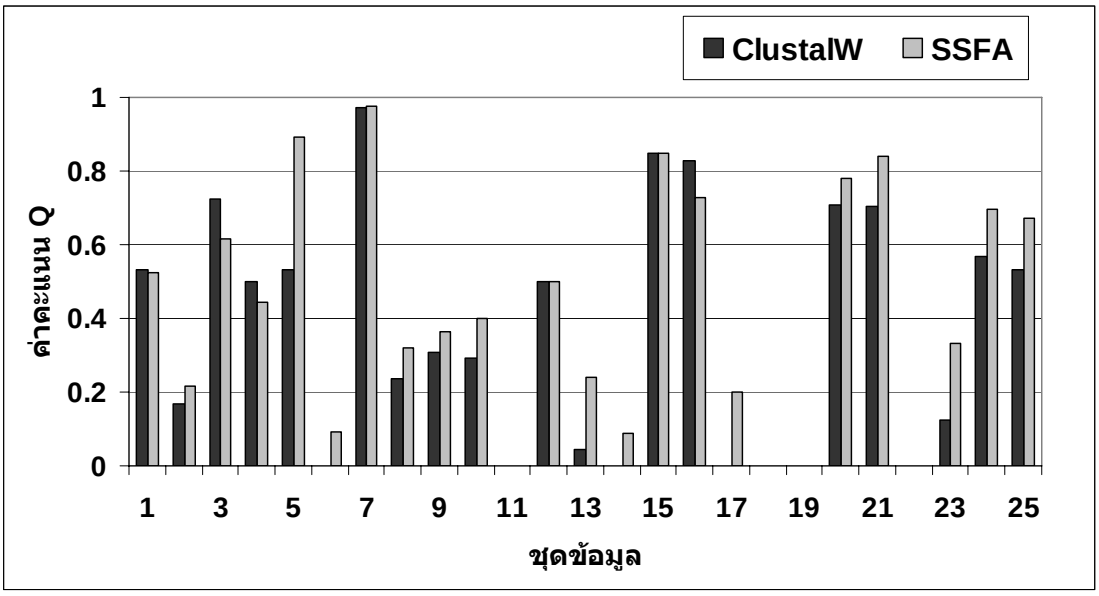
ภาพผนวกที่ 44 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 44



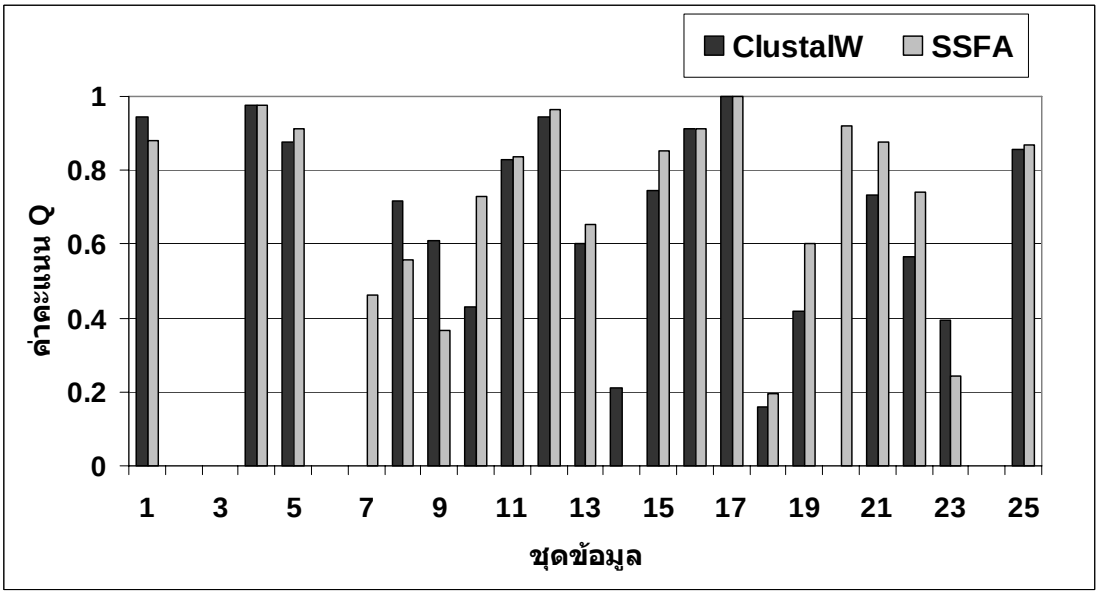
ภาพผนวกที่ 45 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 45



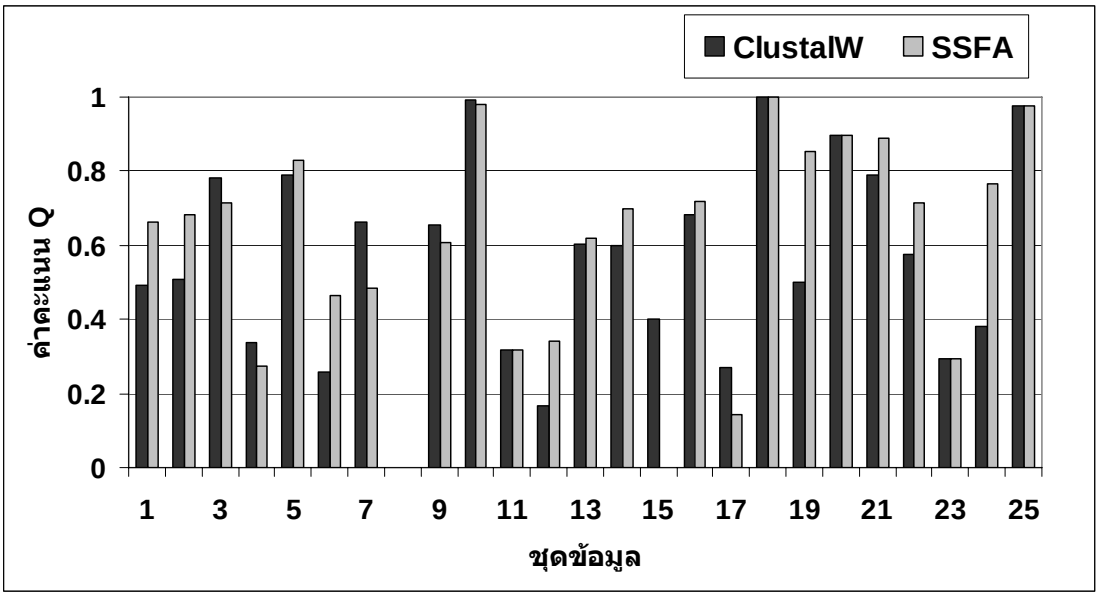
ภาพผนวกที่ 46 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 46



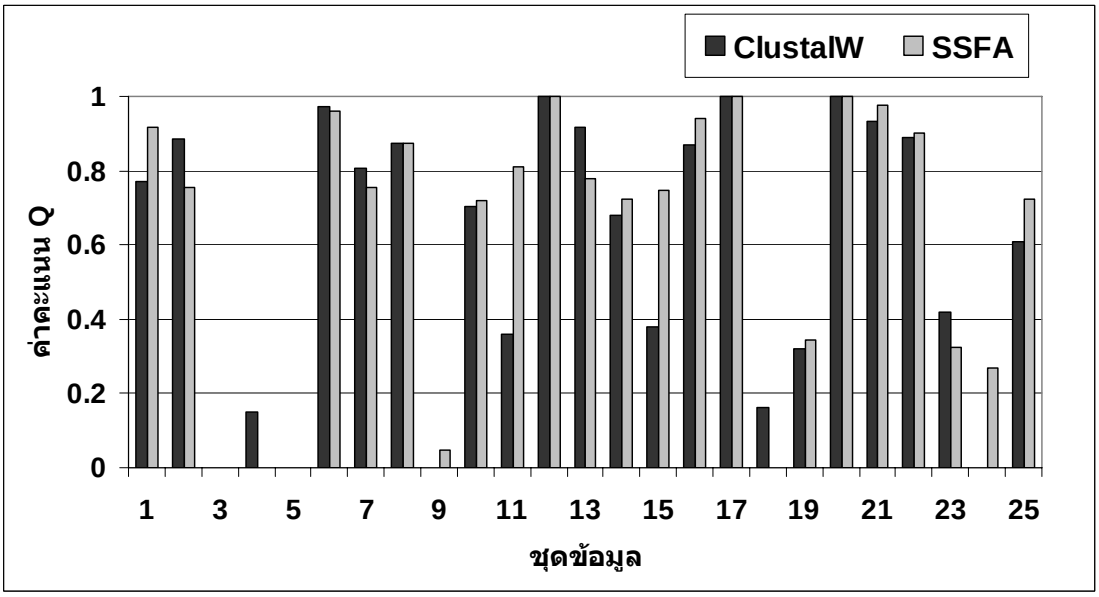
ภาพผนวกที่ 47 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 47



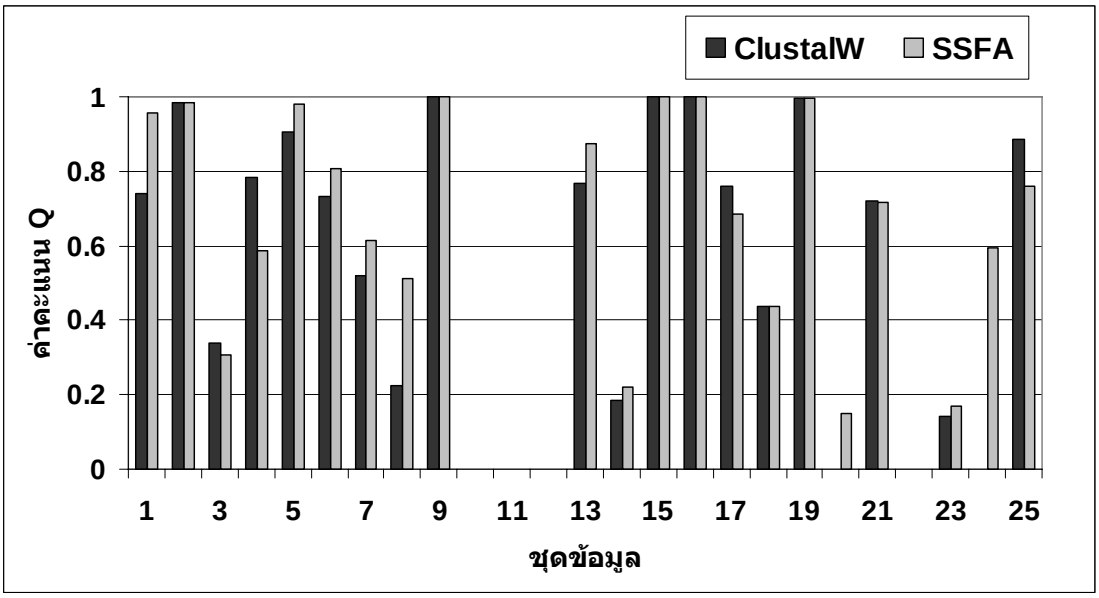
ภาพผนวกที่ 48 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 48



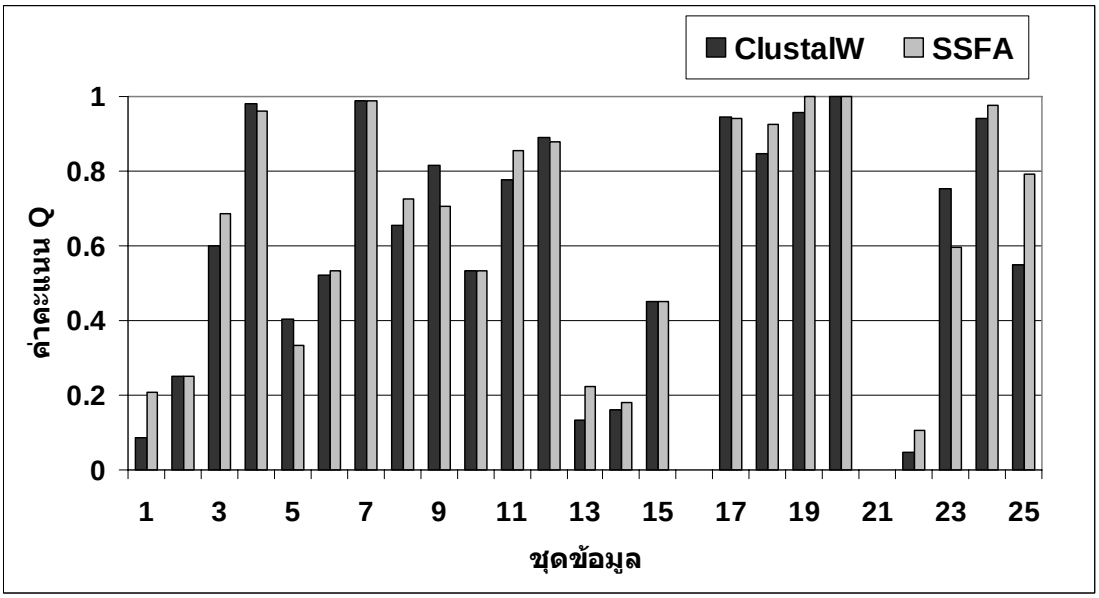
ภาพผนวกที่ 49 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 49



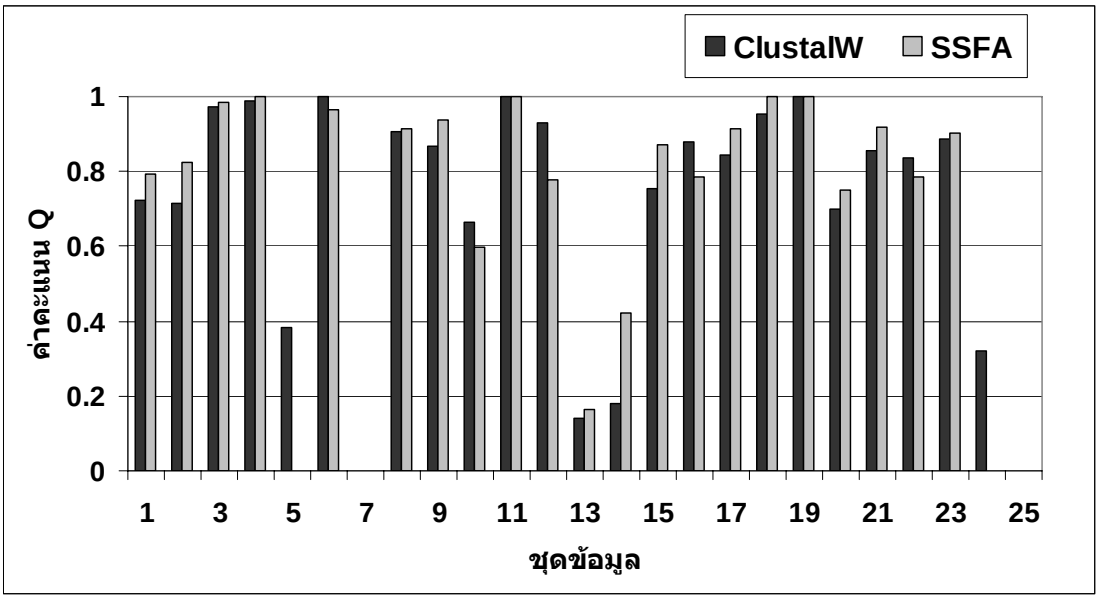
ภาพผนวกที่ 50 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 50



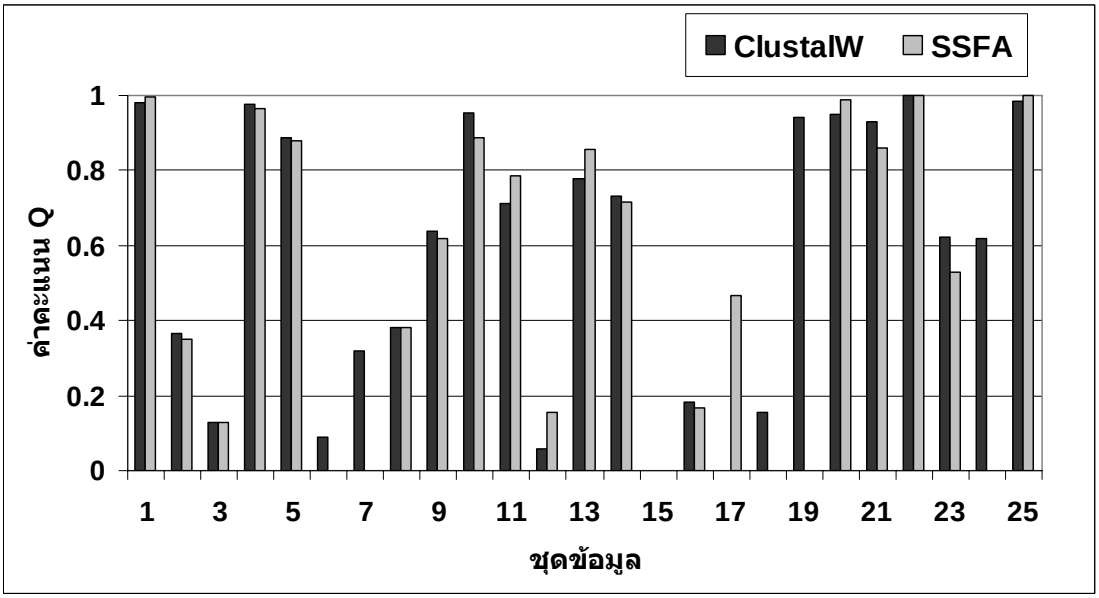
ภาพผนวกที่ 51 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 51



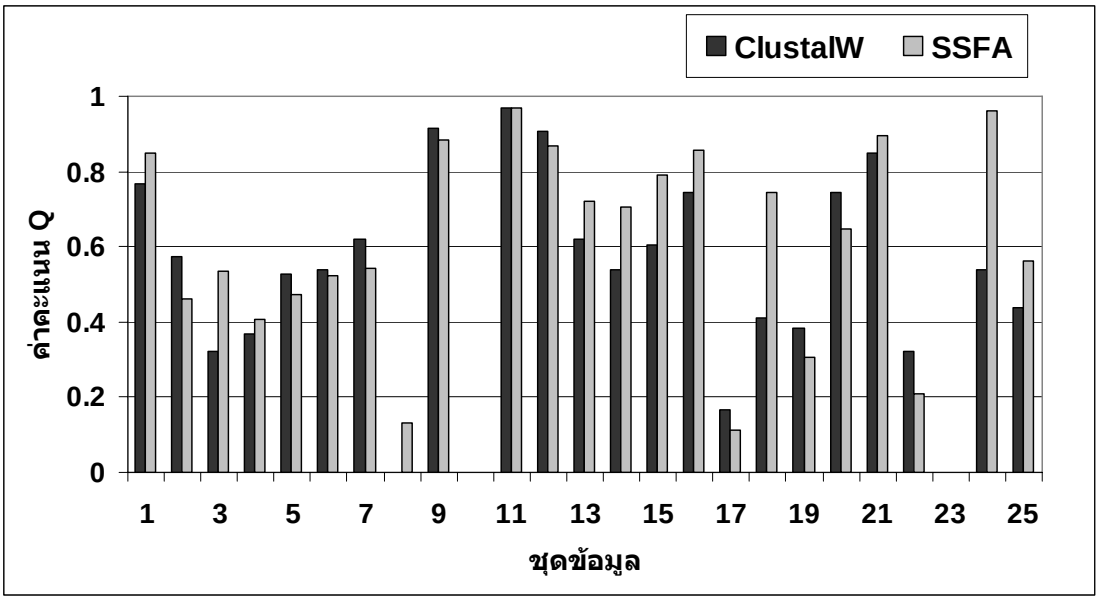
ภาพผนวกที่ 52 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 52



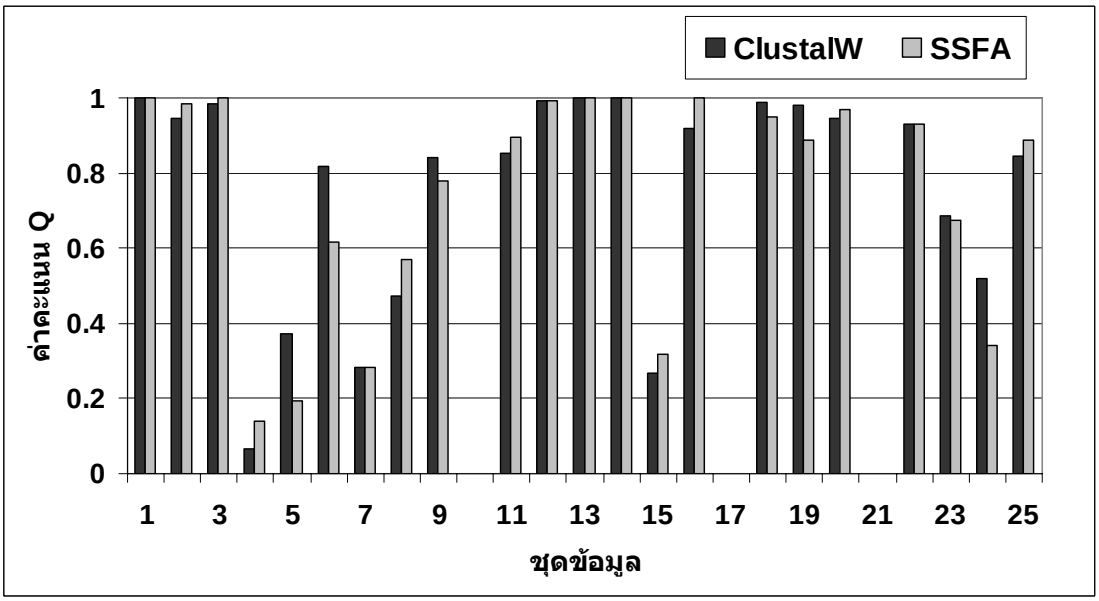
ภาพผนวกที่ 53 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 53



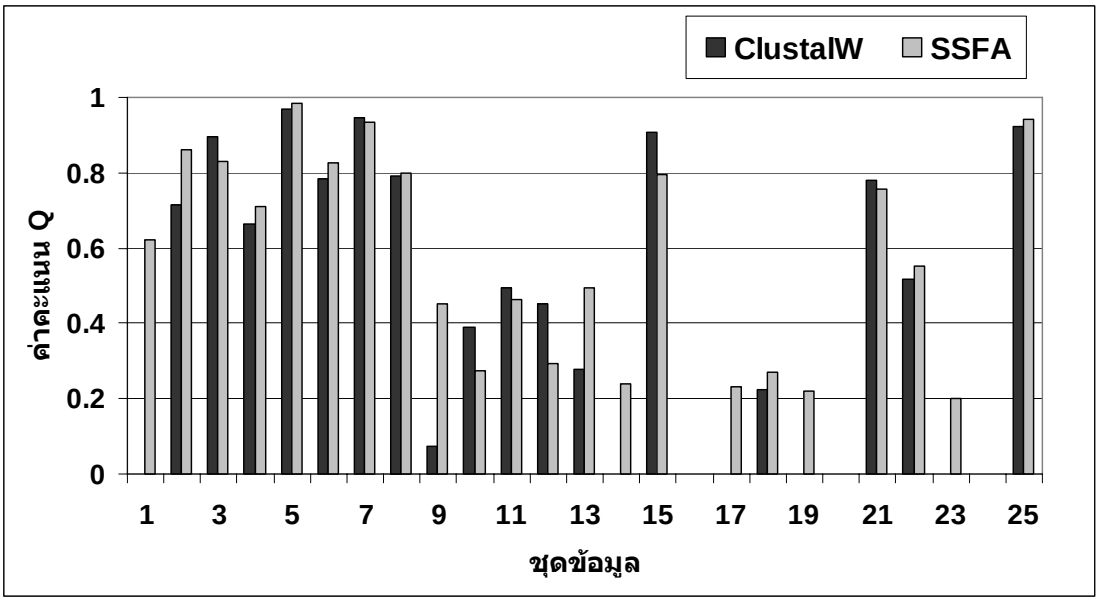
ภาพผนวกที่ 54 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 54



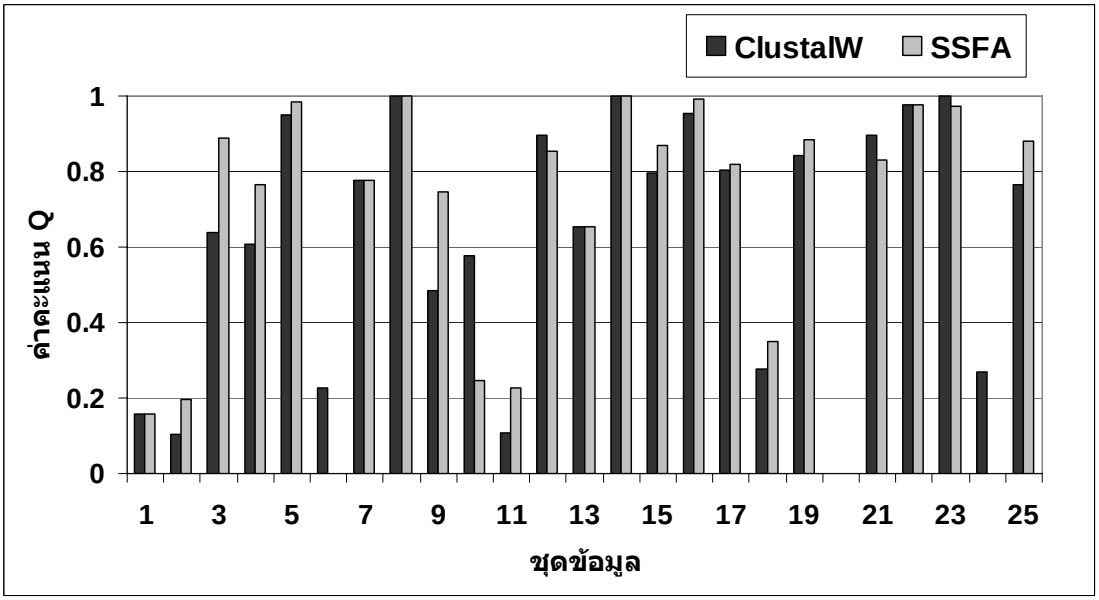
ภาพผนวกที่ 55 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 55



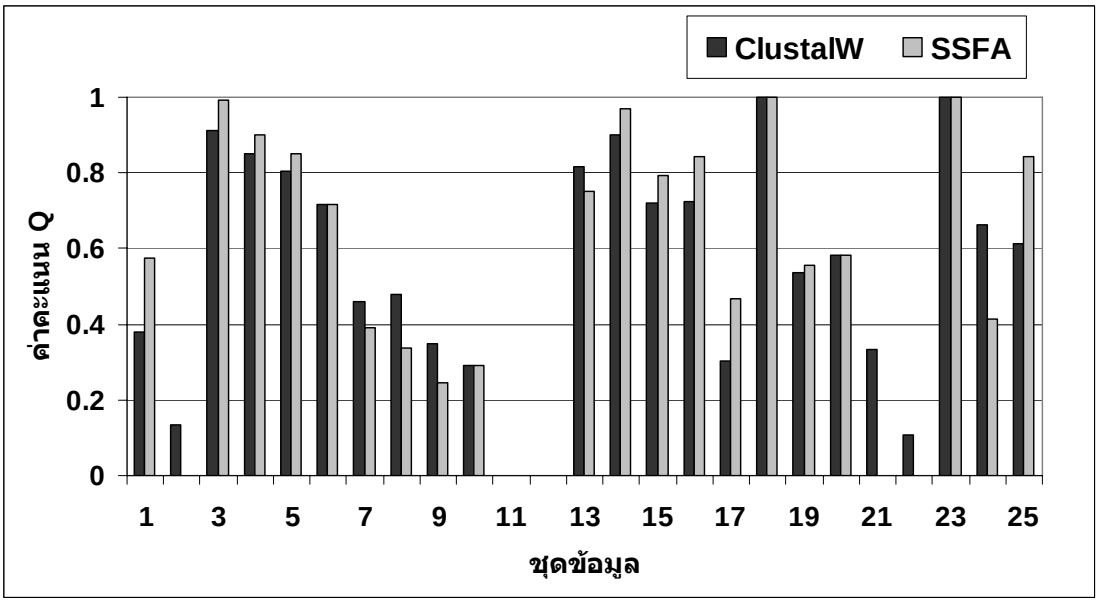
ภาพผนวกที่ 56 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 56



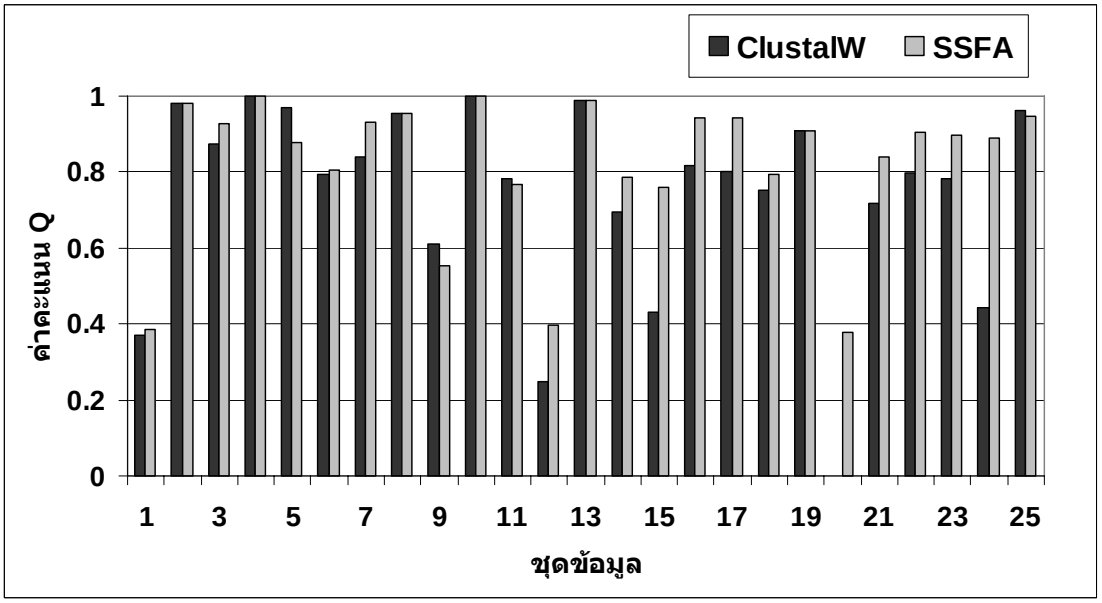
ภาพผนวกที่ 57 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 57



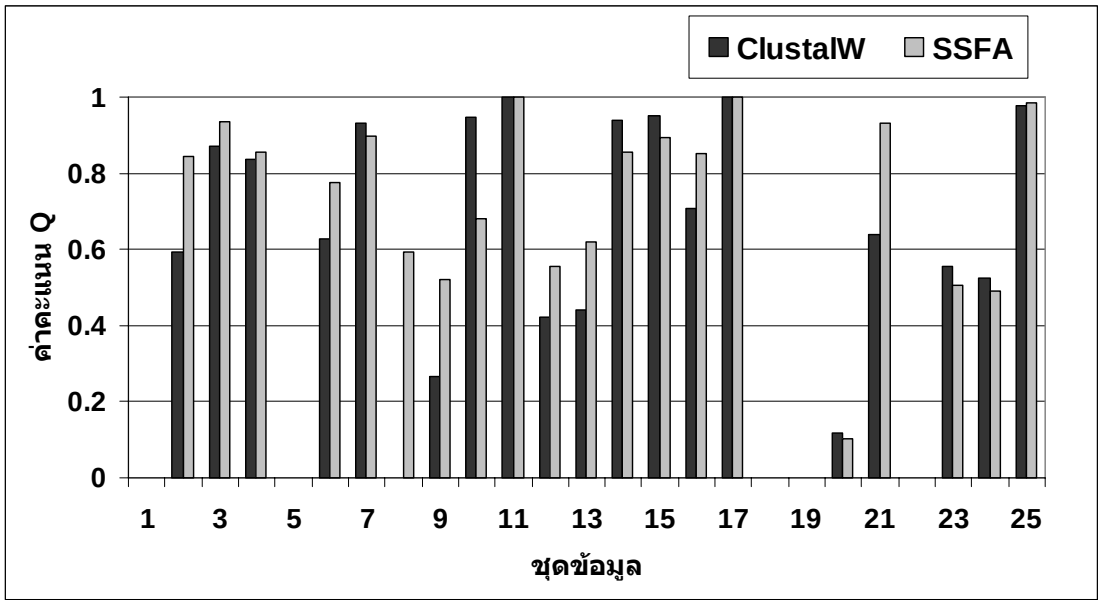
ภาพผนวกที่ 58 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 58



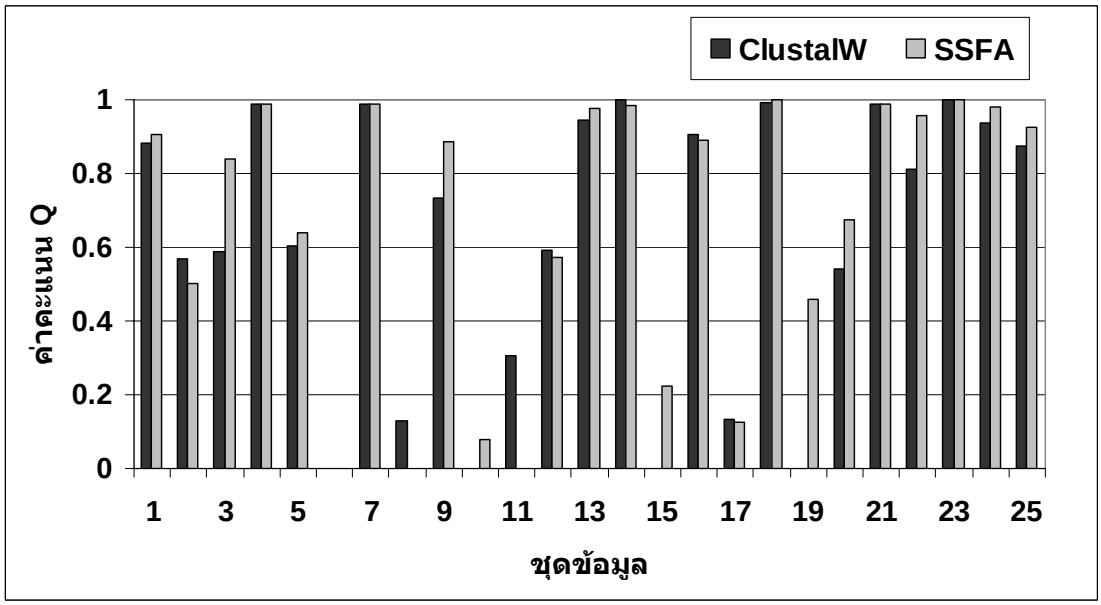
ภาพผนวกที่ 59 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 59



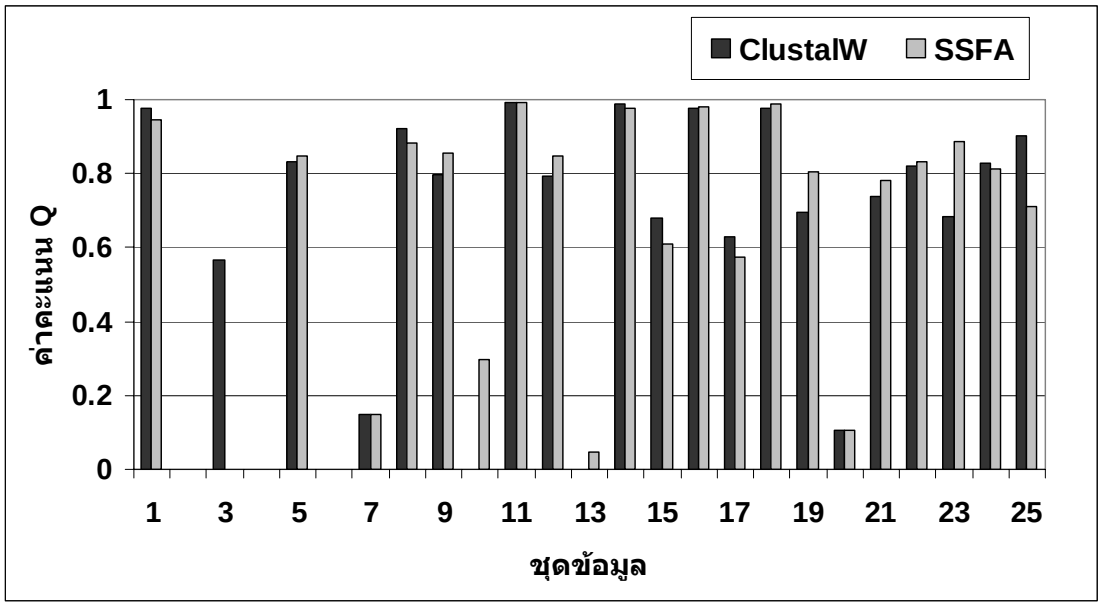
ภาพผนวกที่ 60 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 60



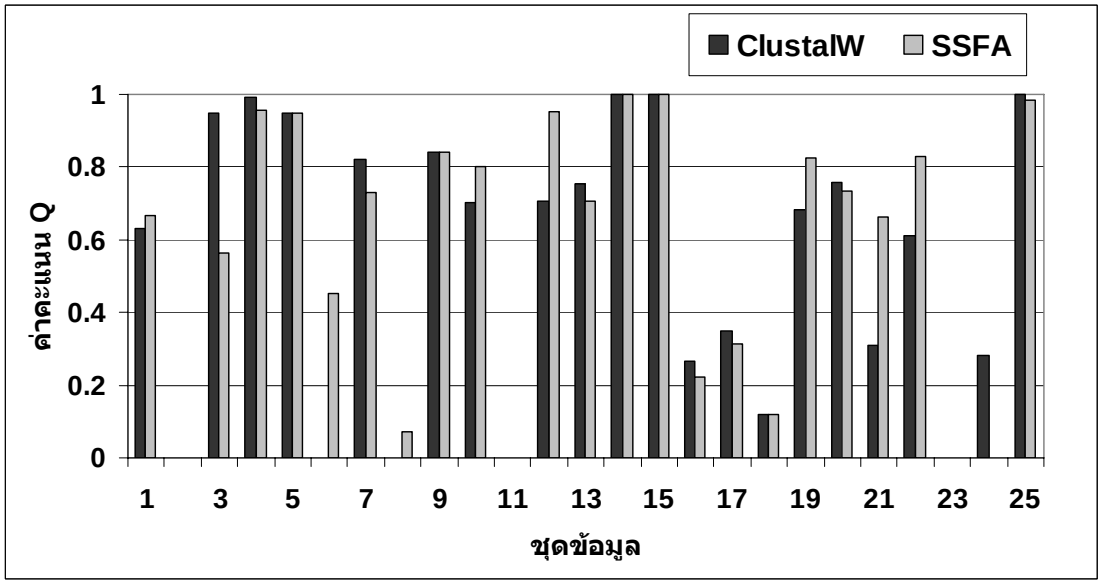
ภาพผนวกที่ 61 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 61



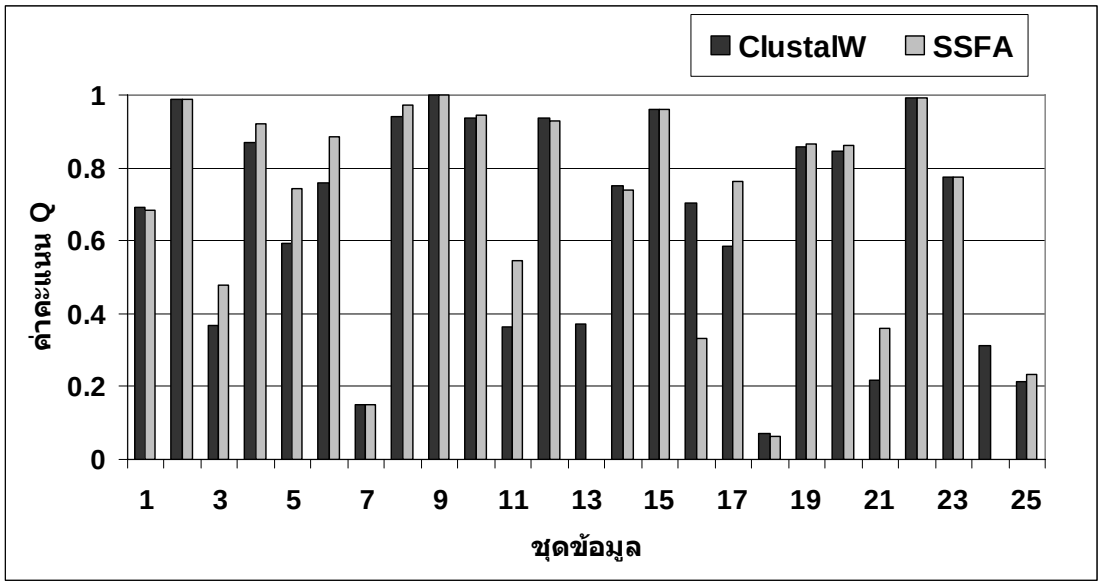
ภาพผนวกที่ 62 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 62



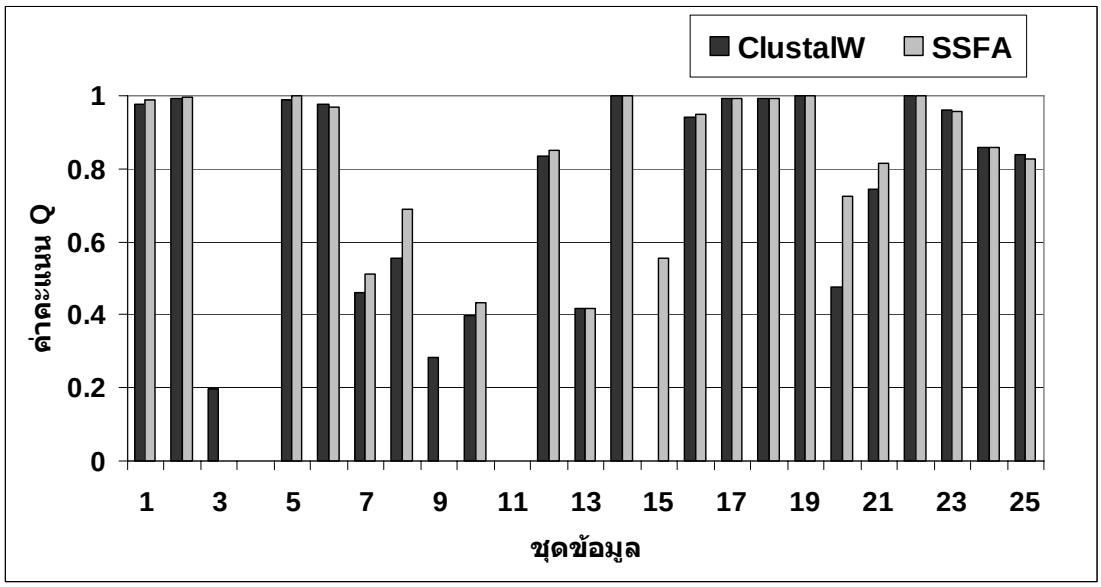
ภาพผนวกที่ 63 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 63



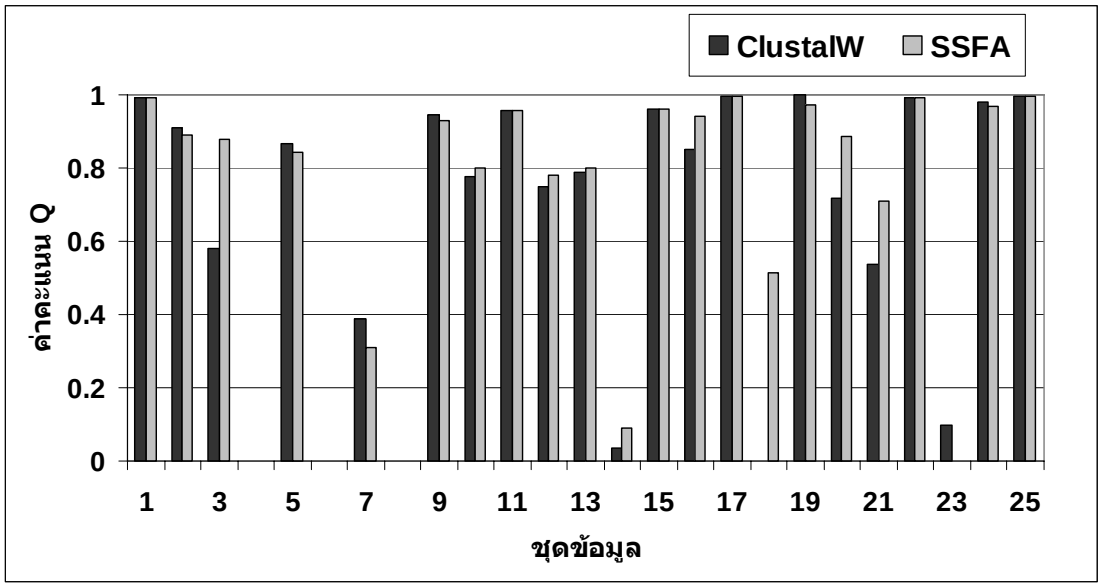
ภาพผนวกที่ 64 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 64



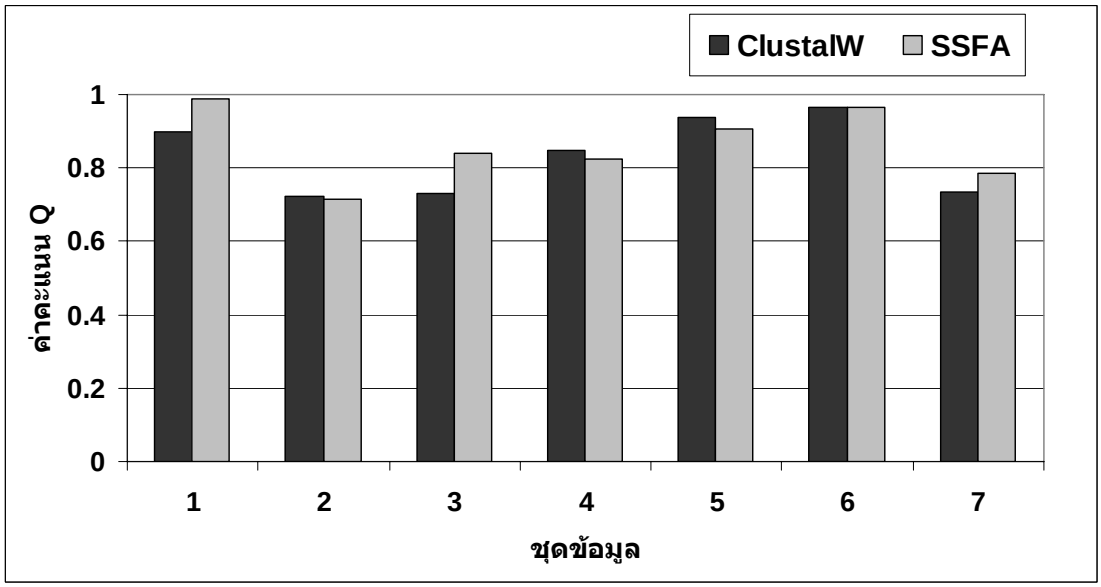
ภาพผนวกที่ 65 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 65



ภาพผนวกที่ 66 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 66



ภาพผนวกที่ 67 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 67



ภาพผนวกที่ 68 ค่าคะแนน Q ในการเทียบเรียงฐานข้อมูล PREFAB ของโปรแกรม ClustalW และโปรแกรม SSFA กับกลุ่มข้อมูลที่ 68

ประวัติการศึกษา และการทำงาน

ชื่อ –นามสกุล	นายคมสัน จันมา
วัน เดือน ปี ที่เกิด	14 พฤศจิกายน 2523
สถานที่เกิด	จังหวัดกรุงเทพมหานคร
ประวัติการศึกษา	ปริญญาวิศวกรรมศาสตรบัณฑิต (วิศวกรรมคอมพิวเตอร์)
ตำแหน่งหน้าที่การงานปัจจุบัน	วิศวกรระบบ
สถานที่ทำงานปัจจุบัน	บริษัท ที.เอ็น. อินฟอร์เมชั่น ซิสเท็มส์ จำกัด
ผลงานดีเด่นและรางวัลทางวิชาการ	-
ทุนการศึกษาที่ได้รับ	-