การศึกษาประสิทธิภาพกระบวนการสร้างโพรไฟล์ด้วยการวิเคราะห์ค่าความคล้ายคลึง
The Efficiency of profiling process of sequence by similarity rate analysis

ณลักขณา คิดเหมาะ
คณะเศรษฐศาสตร์และบริหารธุรกิจ มหาวิทยาลัยทักษิณ วิทยาเขตสงขลา

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อวิเคราะห์ผลกระทบของการกำหนดค่าความคล้ายคลึง (Similarity rate) ของสายลำดับ เพื่อนำไปใช้ในกระบวนการสร้างโพรไฟล์ให้แก่สายลำดับอินพุตและสายลำดับในฐานข้อมูล โดยการศึกษาครั้งนี้ได้นำข้อมูลจากฐานข้อมูล Greengenes ประกอบด้วย รหัสสายลำดับ ชื่อสายลำดับและลำดับดีเอ็นเอ โดยที่ความยาวของสายลำดับอินพุตจะต้องไม่เกินความยาวสายลำดับในฐานข้อมูล การประเมินผลจะวัดประสิทธิภาพจากจำนวนที่ลดลงของสายลำดับในฐานข้อมูล ความไว (Sensitivity) ของแต่ละเงื่อนไขและค่าเฉลี่ยเวลาที่ใช้ในการระบุชื่อสายลำดับ ซึ่งจะทำการเปรียบเทียบการระบุชื่อสายลำดับด้วยวิธีการแบบดั้งเดิม (Original Search) และการระบุชื่อสายลำดับด้วยโปรแกรม BLAST ร่วมกับเทคนิคที่นำเสนอ ผลการศึกษาพบว่า การกำหนดอัตราค่าความคล้ายคลึงที่ >80% ให้ผลลัพธ์ดีที่สุด สามารถลดเวลาค้นหาโดยเฉลี่ย 51.96% (1.59 วินาทีต่อสายลำดับจาก 3.31 วินาทีต่อสายลำดับ) และค่าความไวอยู่ที่ 1.00 จากผลการเปรียบเทียบประสิทธิภาพนี้สามารถนำค่าความคล้ายคลึงที่ >80% มาใช้ในกระบวนการสร้างโพรไฟล์ให้แก่สายลำดับอินพุตและสายลำดับในฐานข้อมูล

คำสำคัญ : โปรแกรม BLAST, การระบุชื่อสายลำดับ, ค่าความคล้ายคลึง

Abstract

The purpose of this study is to analyze the effects of the similarity rates of the sequences for generating the profiles of the input sequences and the sequences in the database. In this study, the data from Greengenes database included the sequence ids, sequence names and DNA sequences. The lengths of the input sequences must not exceed the lengths of the sequences in the database. For the evaluation, the efficiency was measured from the reduction of the sequences in the database, sensitivity of each condition and average time for identifying the sequence names. The sequence names were compared with the original search method, and the sequence names were identified with BLAST program with the presented technique(s). According to the findings, it was found that the similarity rate of >80% resulted in the best result since the average search time could be reduced for 51.96% (1.59 seconds per sequence from 3.31 seconds per sequence). The sensitivity was 1.00. By

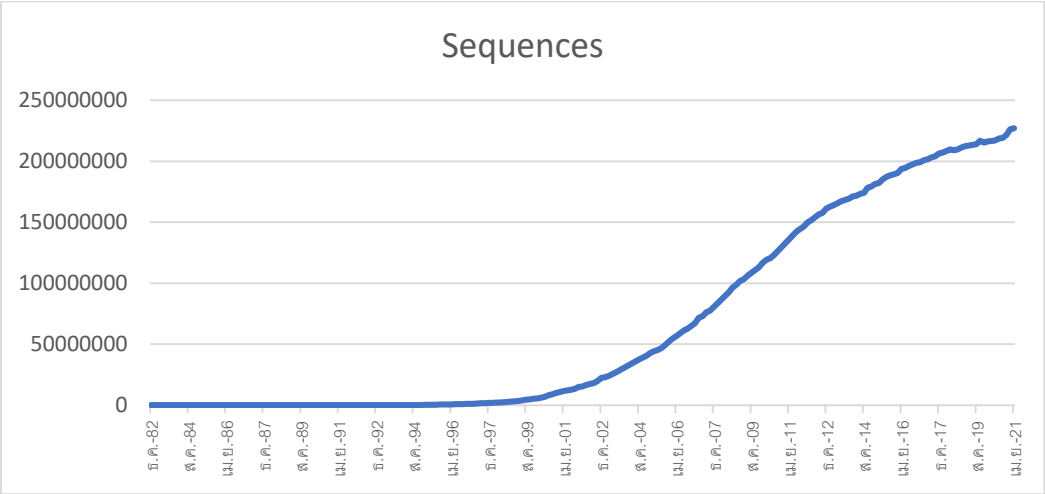
comparing the efficiency results, the similarity rate of >80% can be used for generating the profiles of the input sequences and the sequences in the database.

Keywords The BLAST, Sequence search, Similarity rate

บทนำ

การเชื่อมโยงข้อมูลทางชีววิทยากับเทคนิคการจัดเก็บข้อมูลหรือการค้นคืนข้อมูล เพื่อช่วยในการตอบคำถามทางชีววิทยา เป็นแนวคิดทางวิทยาศาสตร์ที่เรียกว่า ชีวสารสนเทศศาสตร์ (Bioinformatics) เป็นการนำข้อมูลที่สร้างขึ้นจากการทดลองในห้องปฏิบัติการมาประยุกต์ใช้กับเทคโนโลยีคอมพิวเตอร์ (Pareek, Smoczynski, & Tretyn, 2011) ซึ่งข้อมูลดังกล่าวสามารถนำไปสู่การค้นพบทางวิทยาศาสตร์ โดยเฉพาะด้านการแพทย์ อาทิ การระบุความสัมพันธ์ระหว่างลำดับของยีนกับโรคที่สามารถใช้ทำนายโครงสร้างโปรตีนจากลำดับกรดอะมิโน การออกแบบยารักษาโรค รวมถึงการปรับแต่งการรักษาให้กับผู้ป่วยโดยใช้ลำดับดีเอ็นเอของตัวเอง (Ng & Kirkness, 2010)

ปัญหาด้านชีววิทยาที่สำคัญสำหรับโรคที่เกิดขึ้นในมนุษย์มีมากมาย ซึ่งสามารถใช้ชีวสารเทศในการตอบปัญหาเหล่านี้ได้ อาทิ การทำความเข้าใจจีโนไทป์ – ฟีโนไทป์ และการทำความเข้าใจโครงสร้างโปรตีน เป็นต้น จากการศึกษาจำนวนของลำดับข้อมูลชีวภาพโดย GenBank ได้ทำการบันทึกไว้ตั้งแต่ปี ค.ศ. 1982 – เมษายน 2021 พบอัตราการเจริญเติบโตของลำดับข้อมูลทางชีวภาพจำนวน 227,123,201 สายลำดับ (NCBI, 1988) ดังภาพที่ 1 สะท้อนให้เห็นว่าการค้นพบสายลำดับนั้นเพิ่มขึ้นอย่างรวดเร็ว เนื่องจากความก้าวหน้าของเทคโนโลยี ซึ่งการเพิ่มขึ้นของขนาดฐานข้อมูลสายลำดับชีวภาพ เกิดเป็นความท้าทายในการคำนวณเพื่อค้นคืนสายลำดับ ทำให้มีการพัฒนาอัลกอริทึมในการค้นคืนสายลำดับเพื่อลดระยะเวลาในการระบุชื่อสายลำดับ มีหลากหลายโปรแกรมที่ใช้เพื่อค้นคืนสายลำดับ อาทิ FASTA (William R. Pearson, 2014) SSEARCH (William R. Pearson, 1991) CAFÉ (Williams & Zobel, 2002) Genoogle (Albrecht, 2015) เป็นต้น แต่โปรแกรมที่ได้รับความนิยมมากที่สุดในการค้นคืนสายลำดับ คือ โปรแกรม BLAST โดย BLAST เป็นเครื่องมือที่ใช้ค้นหาส่วนของดีเอ็นเอหรือโปรตีนที่มีความคล้ายคลึงกันของสายลำดับ ซึ่งมีการเปรียบเทียบสายลำดับที่ต้องการระบุชื่อกับฐานข้อมูล และมีการคำนวณค่านัยสำคัญทางสถิติของความเหมือนของสายลำดับ โดยสามารถระบุชนิดของยีนหรือจำแนกสายพันธุ์ของสิ่งมีชีวิตได้ กลไกการทำงานของ BLAST อนุญาตให้มีการกลายพันธุ์เกิดขึ้น ดังนั้น BLAST จึงไม่สามารถค้นคืนสายลำดับได้ 100% แต่สามารถทำการเปรียบเทียบความคล้ายคลึงกับสายลำดับที่ต้องการระบุชื่อได้ และ BLAST ยังมีข้อจำกัดในเรื่องของเวลาในการระบุชื่อค่อนข้างนาน เนื่องจากข้อมูลมีขนาดใหญ่ขึ้น (Altschul, Gish, Miller, Myers, & Lipman, 1990) ด้วยเหตุผลดังกล่าวข้างต้น จึงนำไปสู่การพัฒนาเทคนิคการค้นคืนข้อมูลลำดับเบสในฐานข้อมูลจีโนมด้วยโปรแกรม BLAST เพื่อเพิ่มความเร็วในการค้นคืนสายลำดับ



ภาพที่ 1 สถิติการเติบโตของข้อมูลลำดับชีวภาพ

ดังนั้นงานวิจัยนี้มีวัตถุประสงค์เพื่อวิเคราะห์ผลกระทบการกำหนดค่าความคล้ายคลึง เพื่อนำไปใช้ในกระบวนการสร้างโพรไฟล์ ซึ่งทำการทดลองกับข้อมูลสายลำดับอินพุตที่มีความยาวประมาณ 400 ตัวอักษร จำนวน 100 สายลำดับ และฐานข้อมูล Greengenes จำนวน 1,075,170 สายลำดับ โดยกำหนดการทดลอง 3 เงื่อนไข คือ 1) ค่าความคล้ายคลึงที่ >60% 2) ค่าความคล้ายคลึงที่ >70% และ 3) ค่าความคล้ายคลึงที่ >80% ทั้งนี้ในการทดลองแต่ละเงื่อนไขจะใช้ข้อมูลชุดเดียวกันทั้งหมด จากนั้นเปรียบเทียบประสิทธิภาพในการระบุชื่อสายลำดับด้วยวิธีการแบบดั้งเดิม (Original Search) และการระบุชื่อสายลำดับด้วยโปรแกรม BLAST ร่วมกับเทคนิคที่นำเสนอ

วัตถุประสงค์ของการวิจัย

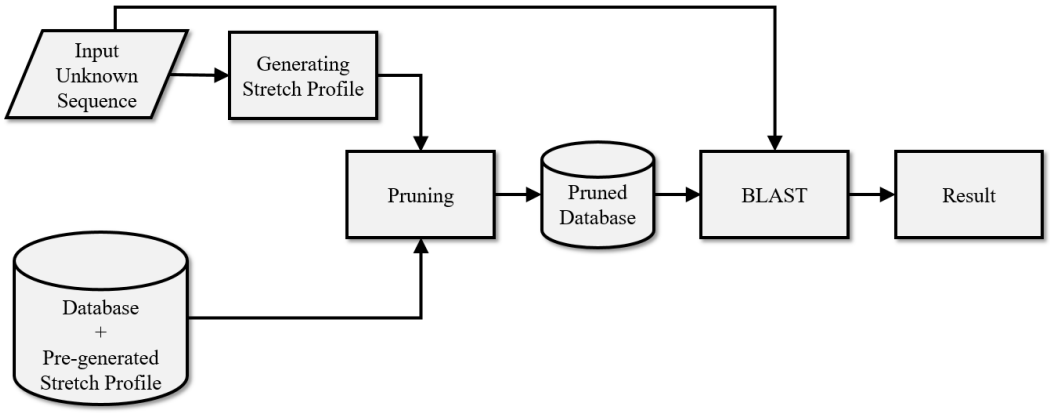
เพื่อวิเคราะห์ผลกระทบการกำหนดค่าความคล้ายคลึง ที่จะนำไปใช้ในกระบวนการสร้างโพรไฟล์ของสายลำดับอินพุตและสายลำดับในฐานข้อมูล

ขอบเขตการวิจัย

การศึกษาอัตราค่าความคล้ายคลึงเพื่อใช้ในกระบวนการสร้างโพรไฟล์ให้กับสายลำดับในครั้งนี้ ผู้วิจัยใช้ข้อมูลสายลำดับอินพุตจากที่ได้จากห้องทดลอง และฐานข้อมูล Greengenes ประกอบด้วย รหัสสายลำดับ ชื่อสายลำดับ และลำดับดีเอ็นเอ โดยที่ความยาวของสายลำดับที่ต้องการระบุชื่อ จะต้องมีความยาวไม่เกินความยาวสายลำดับในฐานข้อมูล โดยพัฒนาเทคนิคการพรุนนิ่งด้วยภาษาไพทอน (Python) และใช้โปรแกรม BLAST ระบุชื่อสายลำดับ

วิธีดำเนินการวิจัย

ในการคำนวณหาค่าความเหมาะสมของความคล้ายคลึงของสายลำดับจะต้องทดสอบโดยมี 5 ขั้นตอนหลัก ได้แก่ การสุ่มเลือกสายลำดับอินพุต การสร้างโพรไฟล์ให้แก่สายลำดับอินพุตและสายลำดับในฐานข้อมูล การพรุนนิ่ง การระบุชื่อสายลำดับด้วยโปรแกรม BLAST และการประเมินผล โดยมีขั้นตอนดังภาพที่ 2



ภาพที่ 2 กรอบแนวคิดงานวิจัย

1. การสุ่มเลือกสายลำดับอินพุต

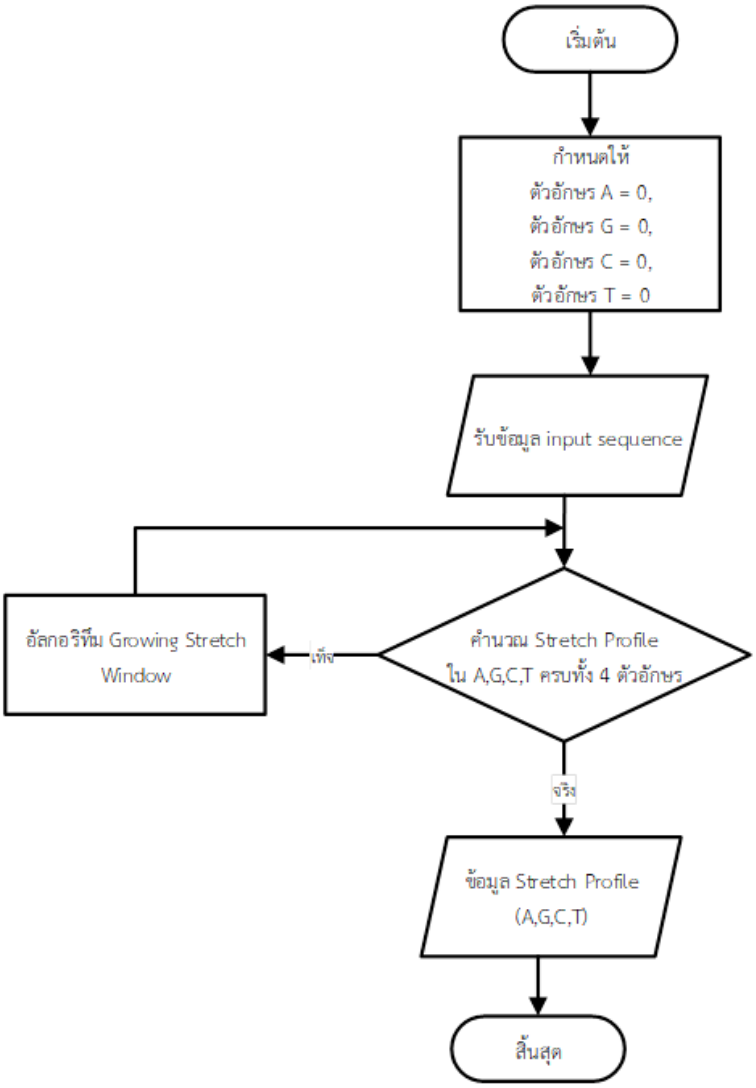
ข้อมูลสายลำดับที่นำมาใช้ในการวิจัยครั้งนี้ เป็นข้อมูลที่ได้รับความอนุเคราะห์จากศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ ซึ่งได้จัดเก็บในรูปแบบ FASTA จำนวน 410,551 สายลำดับ สายลำดับที่สั้นที่สุดมี 365 ตัวอักษร สายลำดับที่ยาวที่สุดมี 456 ตัวอักษร ค่าเฉลี่ยของความยาวอยู่ที่ 412 ตัวอักษรและเป็นสายลำดับที่ยังไม่ได้รับการระบุชื่อ (Unknown Sequence) ตัวอย่างดังภาพที่ 2 โดยผู้วิจัยทำการสุ่มเลือกจำนวน 100 สายลำดับ แต่ละสายลำดับจะมีความยาวประมาณ 400 ตัวอักษร (ในการใช้งานจริง นักวิจัยจะใช้สายลำดับอินพุตที่มีความยาวของสายลำดับประมาณ 400 ตัวอักษร) ส่วนฐานข้อมูล ผู้วิจัยเลือกใช้ฐานข้อมูล Greengenes ที่ประกอบด้วยสายลำดับจำนวน 1,075,173 สายลำดับ สายลำดับที่สั้นที่สุดมี 1,253 ตัวอักษร สายลำดับที่ยาวที่สุดมี 2,368 ตัวอักษร ค่าเฉลี่ยของความยาวอยู่ที่ 1,396 ตัวอักษร

```
>M00218_59_000000000-A923H_1_1101_22879_9217
TGGGGAATATTGGACAATGGGCGCAAGCCTGATCCAGCCATGCCGCGTGAGTGATGAAGGCCTTAGGGTTGTAAAGCTCTTTTGTCCGGGAAGATAATGACTGTACCGGAAGAATAAGCCCCGG
CTAACTTCGTGCCAGCAGCCGCGGTAATACGAAGGGGGCTAGCGTTGCTCGGAATCACTGGGCGTAAAGGGCGCGTAGGCGGACTCTTAAGTCGGGGGTGAAAGCCCAGGGCTCAACCCTGGAA
TTGCCTTCGATACTGAGAGTCTTGAGTTCGGAAGAGGTTGGTGGAACTGCGAGTGTAGAGGTGAAATTCGTAGATATTCGCAAGAACACCAGTGGCGAAGGCGGCCAACTGGTCCGATACTGAC
GCTGAGGCGCGAAAGCGTGGGGAGCAAACAG
>M00218_59_000000000-A923H_1_1101_17529_9699
TGGGGAATATTGGACAATGGGCGAAAGCCTGATCCAGCAATGCCGCGTGAGTGATGAAGGCCTTAGGGTTGTAAAGCTCTTTTACCCGGGATGATAATGACAGTACCGGGAGAATAAGCTCCGG
CTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGAGCTAGCGTTGTTCGGAATTACTGGGCGTAAAGCGCACGTAGGCGGCTTTGTAAGTCAGAGGTGAAAGCCTGGAGCTCAACTCCAGAA
CTGCCTTTGAGACTGCATCGCTTGAATCCAGGAGAGGTGAGTGGAATTCCGAGTGTAGAGGTGAAATTCGTAGATATTCGGAAGAACACCAGTGGCGAAGGCGGCTCACTGGACTGGTATTGAC
GCTGAGGTGCGAAAGCGTGGGGAGCAAACAG
>M00218_59_000000000-A923H_1_1101_20161_9146
TAGGGAATCTTCCGCAATGGGCGAAAGCCTGACGGAGCAACGCCGCGTGAGTGAAGAAGGTCTTCGGATCGTAAAACTCTGTTATTAGGGAAGAACAAATGTGTAAGTAACTATGCACGTCTTG
ACGGTACCTAATCAGAAAGCCACGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGTGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGCGCGCGTAGGCGGTTTTTTAAGTCTGATGT
GAAAGCCCACGGCTCAACCGTGGAGGGTCATTGGAAACTGGAAAACTTGAGTGCAGAAGAGGAAAGTGGAATTCCATGTGTAGCGGTGAAATGCGCAGAGATATGGAGGAACACCAGTGGCGAA
GGCGACTTTCTGGTCTGTAACTGACGCTGATGTGCGAAAGCGTGGGGATCAAACAG
>M00218_59_000000000-A923H_1_1105_16240_8149
TGGGGAATATTGCACAATGGGCGCAAGCCTGATGCAGCCATGCCGCGTGTGTGAAGAAGGCCTTCGGGTTGTAAAGCACTTTCAGTCAGGAGGAAAGGTTAGTAGTTAATACCTGCTAGCTGTG
ACGTTACTGACAGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTACGCAGGCGGTTTGTTAAGCGAGATGT
GAAAGCCCCGGGCTCAACCTGGGAACTGCATTTCGAACTGGCAAACTAGAGTGTGATAGAGGGTGGTAGAATTTCAGGTGTAGCGGTGAAATGCGTAGAGATCTGAAGGAATACCGATGGCGAA
GGCAGCCACCTGGGTCAACACTGACGCTCATGTACGAAAGCGTGGGGAGCAAACAG
```

ภาพที่ 2 ตัวอย่างลำดับที่ยังไม่ได้รับการระบุชื่อ (Input Sequence)

2. การสร้างโพรไฟล์ให้แก่สายลำดับอินพุตและสายลำดับในฐานข้อมูล

เป็นขั้นตอนการสร้างโพรไฟล์ให้แก่สายลำดับด้วยอัลกอริทึม Growing Stretch Window เพื่อค้นหาช่วงตัวอักษรที่เหมือนกันตามอัตราค่าความคล้ายคลึงที่ได้กำหนด ซึ่งผู้วิจัยได้กำหนดเงื่อนไขในการทดสอบไว้ 3 เงื่อนไข คือ กำหนดอัตราค่าความคล้ายคลึงที่ >60% >70% และ >80% ตามลำดับ เช่น กรณีกำหนดค่าความคล้ายคลึงที่ >80% จะต้องทำการสร้างโพรไฟล์ให้แก่สายลำดับอินพุตและสายลำดับในฐานข้อมูลที่มีค่าความคลึงที่ >80% เหมือนกัน จะทำให้ได้สายลำดับอินพุตและฐานข้อมูลที่มีโพรไฟล์เดียวกัน โดยมีขั้นตอนการทำงานที่สามารถแสดงภาพรวมกระบวนการสร้างโพรไฟล์ได้ดังภาพที่ 3



ภาพที่ 3 ขั้นตอนการสร้างโพรไฟล์ให้แก่สายลำดับอินพุตและสายลำดับในฐานข้อมูล

3. การพรุนนิ่ง

ขั้นตอนนี้เป็นการนำโพรไฟล์ของสายลำดับอินพุตที่ได้จากการกำหนดค่าความคล้ายคลึงที่ >60% >70% และ >80% เทียบกับโพรไฟล์ของฐานข้อมูลที่สร้างจากเงื่อนไขเดียวกัน เพื่อคัดกรองสายลำดับที่ไม่เกี่ยวข้องออก ขั้นตอนนี้จะทำให้ได้ฐานข้อมูลที่ประกอบด้วยสายลำดับที่มีความเกี่ยวข้องกับสายลำดับอินพุต (Pruned Database)

4. การระบุชื่อสายลำดับด้วยโปรแกรม BLAST

หลังจากที่ได้ฐานข้อมูลที่ประกอบด้วยสายลำดับที่มีความเกี่ยวข้องกับสายลำดับอินพุตแล้วนั้น ในขั้นตอนนี้เป็นการระบุชื่อให้แก่สายลำดับอินพุตด้วยโปรแกรม BLAST ซึ่งเป็นโปรแกรมที่ได้รับความนิยมและมี

มาตรฐาน โดยใช้ข้อมูลสายลำดับอินพุตจำนวน 100 สายลำดับและฐานข้อมูลที่ได้ผ่านการคัดกรองแล้วเพื่อทำการระบุชื่อ และผลลัพธ์ที่ได้จะอยู่ในรูปแบบ .txt ดังภาพที่ 4

```
# BLASTN 2.2.25 [Feb-01-2011]
# Query: 1575175
# Database: Database_1575175.fasta
# Fields: Query id, Subject id, % identity
1575175 1575175;k__Archaea;p__Euryarchaeota;c__Thermoplasmata;o__Thermoplasmatales;f__Picrophilaceae;g__Thermogymnomonas;s__    100.00
1575175 368917;k__Archaea;p__Euryarchaeota;c__Thermoplasmata;o__Thermoplasmatales;f__Picrophilaceae;g__Thermogymnomonas;s__ 99.50
1575175 809650;k__Archaea;p__Euryarchaeota;c__Thermoplasmata;o__Thermoplasmatales;f__Picrophilaceae;g__Thermogymnomonas;s__ 99.25
1575175 56630;k__Archaea;p__Euryarchaeota;c__Thermoplasmata;o__Thermoplasmatales;f__Picrophilaceae;g__Thermogymnomonas;s__  98.22
1575175 806353;k__Archaea;p__Euryarchaeota;c__Thermoplasmata;o__Thermoplasmatales;f__Picrophilaceae;g__Thermogymnomonas;s__ 98.22
1575175 205614;k__Archaea;p__Euryarchaeota;c__Thermoplasmata;o__Thermoplasmatales;f__Picrophilaceae;g__Thermogymnomonas;s__ 97.72
1575175 63281;k__Archaea;p__Euryarchaeota;c__Thermoplasmata;o__Thermoplasmatales;f__Picrophilaceae;g__Thermogymnomonas;s__  96.98
```

ภาพที่ 4 ผลลัพธ์ที่ได้จากการระบุชื่อสายลำดับ

5. การประเมินผล

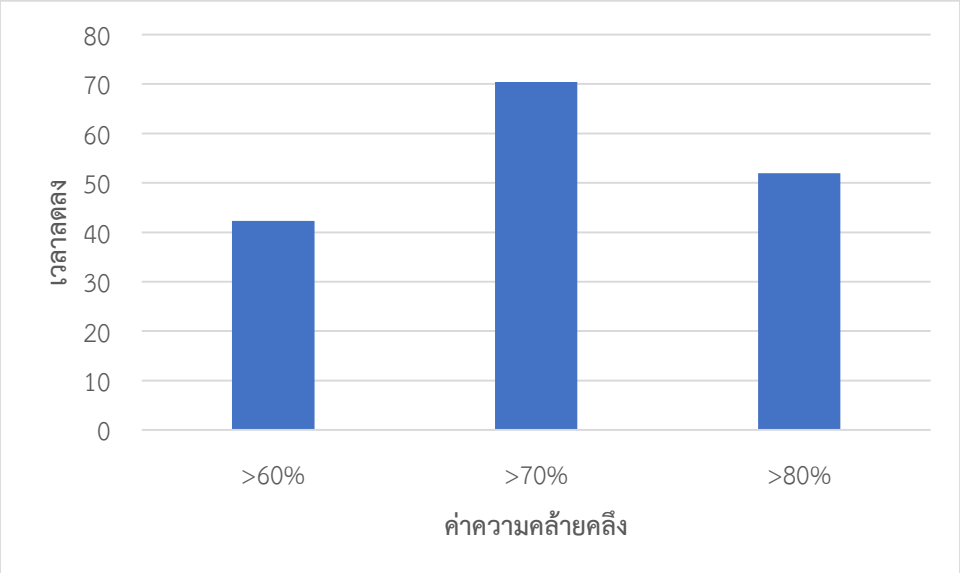
การวิจัยนี้ผู้วิจัยได้วัดประสิทธิภาพจากจำนวนที่ลดลงของสายลำดับในฐานข้อมูลที่ผ่านการคัดกรองความไว (Sensitivity) ของแต่ละเงื่อนไข (>60% >70% และ >80%) โดยจะพิจารณาสายลำดับ 5 อันดับแรกที่ปรากฏในไฟล์ผลลัพธ์ และค่าเฉลี่ยเวลาที่ใช้ในการระบุชื่อสายลำดับ ซึ่งจะทำการเปรียบเทียบการระบุชื่อสายลำดับด้วยวิธีการแบบดั้งเดิม และการระบุชื่อสายลำดับด้วยโปรแกรม BLAST ร่วมกับเทคนิคที่นำเสนอ

ผลการวิจัย

จากการเปรียบเทียบการระบุชื่อสายลำดับด้วยวิธีการแบบดั้งเดิม และการระบุชื่อสายลำดับด้วยโปรแกรม BLAST ร่วมกับเทคนิคที่นำเสนอ โดยวัดประสิทธิภาพจากจำนวนที่ลดลงของสายลำดับในฐานข้อมูลความไวของแต่ละเงื่อนไข ซึ่งจะพิจารณาสายลำดับ 5 อันดับแรก และค่าเฉลี่ยเวลาที่ใช้ในการระบุชื่อสายลำดับ เพื่อวิเคราะห์ผลกระทบของอัตราค่าความคล้ายคลึงของสายลำดับ ที่จะนำไปใช้ในกระบวนการสร้างโพรไฟล์ ซึ่งทำการทดลองกับข้อมูลสายลำดับอินพุตที่มีความยาวประมาณ 400 ตัวอักษร จำนวน 100 สายลำดับ และฐานข้อมูล Greengenes โดยกำหนดการทดลอง 3 เงื่อนไข คือ 1) ค่าความคล้ายคลึงที่ >60% 2) ค่าความคล้ายคลึงที่ >70% และ 3) ค่าความคล้ายคลึงที่ >80% โดยใช้ข้อมูลชุดเดียวกันทุกเงื่อนไข สามารถแสดงผลการวิจัยได้ดังตารางที่ 1

ตารางที่ 1 ค่าความเหมาะสมของการกำหนดค่าความคล้ายคลึงที่ >60% >70% และ >80%

ค่าความคล้ายคลึง	ฐานนิยมของสายลำดับอินพุต (A,G,C,T)	ฐานนิยมของฐานข้อมูล (A,G,C,T)	ค่าความไว (Sensitivity)	สายลำดับเฉลี่ยในฐานข้อมูล (Pruned Database)	เวลาเฉลี่ยที่ใช้ในการระบุชื่อ (วินาที/สายลำดับ)		
					Original	Pruned	เวลาลดลง (%)
>60%	(8,12,9,8)	(11,12,11,9)	0.97	599,017	3.31	1.91	42.30
>70%	(6,11,7,4)	(7,8,7,7)	1.00	297,017	3.31	0.98	70.39
>80%	(5,8,5,4)	(6,7,7,6)	1.00	459,327	3.31	1.59	51.96



ภาพที่ 5 การเปรียบเทียบเวลาที่ลดลงในการระบุชื่อที่ค่าความคล้ายคลึง >60% >70% และ >80%

ผลการทดลองในตารางที่ 1 จะเห็นได้ว่าการกำหนดค่าความคล้ายคลึงที่ >60% โดยการระบุชื่อสายลำดับที่นำเสนอ (Pruned Search) ใช้เวลา 1.91 วินาทีต่อสายลำดับ ในขณะที่การระบุชื่อด้วยวิธีการดั้งเดิมใช้เวลา 3.31 วินาทีต่อสายลำดับ เนื่องจากการลดจำนวนของสายลำดับในฐานข้อมูลที่ผ่านการคัดกรอง (เหลือ 599,017 สายลำดับจากทั้งหมด 1,075,170 สายลำดับ) และค่าความไวอยู่ที่ 0.97 โดยการกำหนดความคล้ายคลึงที่ >60% ช่วยให้โพรไฟล์สายลำดับอินพุตมีความยาวตามที่ระบุจากค่าฐานนิยม (8, 12, 9, 8)

การกำหนดค่าความคล้ายคลึงที่ >70% ซึ่งการระบุชื่อสายลำดับที่นำเสนอใช้เวลาค้นหา 0.98 วินาทีต่อสายลำดับ ในขณะที่การระบุชื่อด้วยวิธีการดั้งเดิมใช้เวลา 3.31 วินาทีต่อสายลำดับ เนื่องจากการลดจำนวนของสายลำดับในฐานข้อมูลที่ผ่านการคัดกรอง (เหลือ 297,017 สายลำดับจากทั้งหมด 1,075,170 สายลำดับ)

และค่าความไวอยู่ที่ 1.00 โดยการกำหนดความคล้ายคลึงที่ >70% ทำให้โพรไฟล์สายลำดับอินพุตมีความยาวลดลงสังเกตได้จากค่าฐานนิยม (6, 11, 7, 4)

การกำหนดค่าความคล้ายคลึงที่ >80% ซึ่งการระบุชื่อสายลำดับที่นำเสนอใช้เวลาค้นหา 1.59 วินาทีต่อสายลำดับ ในขณะที่การระบุชื่อด้วยวิธีการดั้งเดิมใช้เวลา 3.31 วินาทีต่อสายลำดับ เนื่องจากการลดจำนวนของสายลำดับในฐานข้อมูลที่ผ่านการคัดกรอง (เหลือ 459,327 สายลำดับจากทั้งหมด 1,075,170 สายลำดับ) และค่าความไวอยู่ที่ 1.00 โดยการกำหนดความคล้ายคลึงที่ >80% ทำให้โพรไฟล์สายลำดับอินพุตมีความยาวลดลงสังเกตได้จากค่าฐานนิยม (5, 8, 5, 4)

สำหรับการเปรียบเทียบประสิทธิภาพด้านเวลาที่ใช้ในการระบุชื่อด้วยวิธีการแบบดั้งเดิม และการระบุชื่อสายลำดับด้วยโปรแกรม BLAST ร่วมกับเทคนิคที่นำเสนอนั้นพบว่า การกำหนดค่าความคล้ายคลึงที่ >60% โดยการระบุชื่อสายลำดับที่นำเสนอจะลดเวลาค้นหาโดยเฉลี่ย 42.30% (1.91 วินาทีต่อสายลำดับจาก 3.31 วินาทีต่อสายลำดับ) การกำหนดค่าความคล้ายคลึงที่ >70% นั้นการระบุชื่อสายลำดับที่นำเสนอจะลดเวลาค้นหาโดยเฉลี่ย 70.39% (0.98 วินาทีต่อสายลำดับจาก 3.31 วินาทีต่อสายลำดับ) และการกำหนดค่าความคล้ายคลึงที่ >80% การระบุชื่อสายลำดับที่นำเสนอจะลดเวลาค้นหาโดยเฉลี่ย 51.96% (1.59 วินาทีต่อสายลำดับจาก 3.31 วินาทีต่อสายลำดับ) ดังภาพที่ 5

อภิปรายผลการวิจัย

งานวิจัยนี้ ผู้วิจัยได้ทำการทดลองเพื่อศึกษาผลกระทบการกำหนดค่าความคล้ายคลึงที่นำไปใช้ในกระบวนการสร้างโพรไฟล์ โดยข้อมูลที่ใช้มีสายลำดับอินพุตที่มีความยาวประมาณ 400 ตัวอักษร จำนวน 100 สายลำดับ และฐานข้อมูล Greengenes จำนวน 1,075,170 สายลำดับ แบ่งการทดลองออกเป็น 3 เงื่อนไข คือ 1) อัตราค่าความคล้ายคลึงที่ >60% 2) อัตราค่าความคล้ายคลึงที่ >70% และ 3) อัตราค่าความคล้ายคลึงที่ >80% เปรียบเทียบการระบุชื่อสายลำดับด้วยวิธีการแบบดั้งเดิม และการระบุชื่อสายลำดับด้วยโปรแกรม BLAST ร่วมกับเทคนิคที่นำเสนอ จากผลการทดลองพบว่า การกำหนดอัตราค่าความคล้ายคลึงของสายลำดับที่ >80% ให้ผลลัพธ์ที่ดีกว่ากำหนดค่าความคล้ายคลึงของสายลำดับที่ >60% และ >70% ทั้งนี้เนื่องจากฐานข้อมูลที่ได้มีความครอบคลุมและเกี่ยวข้องกับสายลำดับอินพุต ซึ่งจะเห็นได้ว่าค่าความไวในการระบุชื่อสายลำดับมีค่าสูงถึง 1.00 และการลดลงของจำนวนสายลำดับในฐานข้อมูล จึงส่งผลให้ลดเวลาค้นหาโดยเฉลี่ย 51.96% (1.59 วินาทีต่อสายลำดับจาก 3.31 วินาทีต่อสายลำดับ) ดังนั้นเมื่อกำหนดค่าความคล้ายคลึงของสายลำดับที่ >80% จะให้ผลลัพธ์ที่เหมาะสมที่สุด อาจเป็นไปได้ว่าปัจจัยที่ได้จากการกำหนดอัตราค่าความคล้ายคลึง และการลดลงของสายลำดับในฐานข้อมูลช่วยส่งเสริมความไวในการระบุชื่อสายลำดับให้ดีขึ้น นอกจากนี้ยังช่วยเพิ่มประสิทธิภาพด้านความเร็วในการค้นคืนสายลำดับให้แก่โปรแกรม BLAST อีกด้วย

เอกสารอ้างอิง

Pareek, C. S., Smoczynski, R., & Tretyn, A. (2011). Sequencing technologies and genome sequencing. Journal of applied genetics, 52(4), 413-435. doi:10.1007/s13353-011-0057-x

Ng, P. C., & Kirkness, E. F. (2010). Whole genome sequencing. Methods Mol Biol, 628, 215-226. doi:10.1007/978-1-60327-367-1_12

National Center for Biotechnology Information., (1988) GenBank and WGS Statistics. Retrieved April 20, 2021, from https://www.ncbi.nlm.nih.gov/genbank/statistics/

Altschul, S. F., et al. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403-410. doi: https://doi.org/10.1016/S0022-2836(05)80360-2

Pearson, W. R. (2014). BLAST and FASTA Similarity Searching for Multiple Sequence Alignment. In D. J. Russell (Ed.), Multiple Sequence Alignment Methods (pp. 75-101). Totowa, NJ: Humana Press.

Pearson, W. R. (1991). Searching protein sequence libraries: Comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. Genomics, 11(3), 635-650. doi:https://doi.org/10.1016/0888-7543(91)90071-L

Williams, H. E., & Zobel, J. (2002). Indexing and retrieval for genomic databases. IEEE Transactions on Knowledge and Data Engineering, 14(1), 63-78. doi:10.1109/69.979973

Albrecht, F. (2015). Genoogle: an indexed and parallelized search engine for similar DNA sequences.