การพัฒนาโปรแกรมสำหรับการรวมและกำจัดความซ้ำซ้อนของเอเอสวี ในชุดข้อมูลไมโครไบโอมด้วยภาษาไพทอน

Development of a Python-Based Program for ASVs Integration and Deduplication in Microbiome Datasets

เทวินทร์ ฟักเอม[1*], ธนพร อึ้งเวชวานิช[2], มณฑล ฟักเอม[3], ปิยะพงษ์ โอฬารทิชาชาต[4] และ อานนท์ จันทร์เกษม[5]

Tawin Fakaim[1*], Tanaporn Uengwetwanit[2], Monthol Fak-Aim[3], Piyapong Olranthichachat[4] and Arnon Jankasem[5]

สาขาวิชาเทคโนโลยีดิจิทัล คณะนวัตกรรม เทคโนโลยี และการสร้างสรรค์ มหาวิทยาลัยฟาร์อีสเทอร์น[1,5]
ทีมวิจัยไมโครอะเรย์แบบครบวงจร, ศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ (ไบโอเทค),
สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ (สวทช.)[2]
สาขาวิศวกรรมไฟฟ้าสื่อสารและอิเล็กทรอนิกส์ คณะเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยราชภัฏพิบูลสงคราม[3,4]
Department of Digital Technology, Faculty of Innovation Technology and Creativity, The Far Eastern University[1,5]
Microarray Research Team, National Center for Genetic Engineering and Biotechnology (BIOTEC), National Science and Technology Development Agency (NSTDA)[2]
Electrical Communication and Electronics Engineering, Faculty of Industrial Technology,
Pibulsongkram Rajabhat University[3,4]
E-Mail : tawin@feu.ac.th[1*], tanaporn.uen@biotec.or.th[2], monthol@psru.ac.th[3],
piyapong@psru.ac.th[4], arnon@feu.ac.th[5]
(Received : July 23, 2025; Revised : November 12, 2025; Accepted : December 9, 2025)

บทคัดย่อ

การรวมชุดข้อมูลแอมพลิคอนซีเควนส์แวเรียนต์ (เอเอสวี) จากหลายการทดลองเพื่อการวิเคราะห์ชุดข้อมูลไมโครไบโอม มักประสบปัญหาความซ้ำซ้อนของข้อมูล นำไปสู่การประมวลผลช้า การระบุอนุกรมวิธานไม่ถูกต้อง และเพิ่มภาระงานในการจัดการข้อมูลด้วยตนเอง งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาและประเมินประสิทธิภาพโปรแกรมภาษาไพทอนสำหรับการรวมและกำจัดความซ้ำซ้อนของเอเอสวี โปรแกรมพัฒนาภายใต้กรอบแนวคิดวงจรชีวิตการพัฒนาระบบ (เอสดีแอลซี) ใช้วิธีการจับคู่ตรงกันทุกตำแหน่ง (เอ็กแซกต์แมตช์) เพื่อรักษาความละเอียดลำดับเอเอสวี รองรับการทำงานที่ยืดหยุ่นผ่านส่วนติดต่อผู้ใช้แบบกราฟฟิก จูปิเตอร์ และคอมมานด์ไลน์

ผลการประเมินประสิทธิภาพเชิงเทคนิคด้วยชุดข้อมูล 2 ถึง 10 ชุด จำนวนเอเอสวี 33,445 ถึง 117,678 สาย พบว่า โปรแกรมมีความถูกต้องในการกำจัดความซ้ำซ้อน 100% และใช้เวลาประมวลผลเฉลี่ยเพียง 4.29 ถึง 35.58 วินาที ตามลำดับ ด้านผลการประเมินความพึงพอใจผู้ใช้งานอยู่ในระดับดี ค่าเฉลี่ย 4.29 โดยผู้ใช้พึงพอใจด้านความถูกต้องของผลลัพธ์สูงสุด โปรแกรมนี้ช่วยลดความซ้ำซ้อนของเอเอสวีและเพิ่มความถูกต้องในการเตรียมชุดข้อมูลไมโครไบโอมสำหรับการวิเคราะห์ปลายทางด้านชีวสารสนเทศ

คำสำคัญ : ไมโครไบโอม, เอเอสวี, จับคู่ตรงกันทุกตำแหน่ง, ความพึงพอใจ

ABSTRACT

Integrating amplicon sequence variant (ASV) datasets from multiple experiments for microbiome analysis often leads to data redundancy, resulting in slow processing, inaccurate taxonomic classification, and an increased burden of manual data management. This study aimed to develop and evaluate the performance of a Python-based program for ASV integration and deduplication. The program was developed under the System Development Life Cycle (SDLC) framework and employs an exact-match method to preserve ASV sequence resolution. It supports flexible operation through a graphical user interface (GUI), Jupyter environment, and command-line interface (CLI).

Technical performance was evaluated using 2 to 10 datasets, comprising 33,445 to 117,678 ASV sequences. The program achieved 100% deduplication accuracy, with average processing times ranging from 4.29 to 35.58 seconds, respectively. User satisfaction with the program was at a good level (mean score = 4.29), with the highest satisfaction reported for the accuracy of the results. Overall, the program effectively reduces ASV redundancy and enhances the reliability of microbiome dataset preparation for downstream bioinformatics analyses.

Keywords : Microbiome, ASVs, Exact match, Satisfaction

บทนำ

ความก้าวหน้าของเทคโนโลยีสำหรับการหาลำดับสารพันธุกรรมด้วยความเร็วสูง High-Throughput Sequencing (HTS) [1] ทำให้การพัฒนาการวิจัยด้านไมโครไบโอม (Microbiome) มีการสร้างชุดข้อมูลขนาดใหญ่อย่างต่อเนื่อง ความเข้าใจเกี่ยวกับไมโครไบโอมเปลี่ยนแปลงไปมากจากการค้นพบข้างต้น หลายประเทศได้ริเริ่มโครงการวิจัยไมโครไบโอมระดับนานาชาติ ซึ่งประสบความสำเร็จ และช่วยผลักดันให้การวิจัยไมโครไบโอมเข้าสู่ยุคทอง [2] การสร้างโปรไฟล์ของชุมชนจุลินทรีย์ (Microbial Community Profiling) โดยการวิเคราะห์ลำดับเอเอสวี (ASVs) ของยีน 16 เอสอาร์อาร์เอ็นเอ (16S rRNA) เป็นวิธีการที่นิยม [3] และมีความแม่นยำสูงในการระบุอนุกรมวิธาน (Taxonomy Assignment) [4] นอกจากนั้น เอเอสวีเป็นลำดับที่ตรงกันแบบสมบูรณ์ หรือเป็นลำดับที่เหมือนกันทุกตำแหน่งแบบไม่มีความคลาดเคลื่อน (Exact Sequence) สามารถนำผลจากงานวิจัยหนึ่งไปเทียบกับอีกงานวิจัยหนึ่ง ช่วยยกระดับความแม่นยำในการเปรียบเทียบ [5] และนำไปสู่การวิเคราะห์เชิงเปรียบเทียบ หรือเมตาอะแนลิสิซ (Meta-Analysis) [6, 7]

ทั้งนี้หลังจากรวบรวมเอเอสวีจากหลายเงื่อนไขการทดลอง หรือระหว่างงานวิจัยพบปัญหาการจัดการชุดข้อมูลจำนวนมากที่มีความซ้ำซ้อน ซึ่งปัจจุบันมีโปรแกรมที่นิยม เช่น ไคม์ทู (Quantitative Insights Into Microbial Ecology 2: QIIME2) แพลตฟอร์มชีวสารสนเทศแบบโอเพนซอร์ส (Open Source) [8] เน้นกระบวนการวิเคราะห์ต้นทางเพื่อการสร้างเอเอสวีจากข้อมูลดิบ มีไพป์ไลน์ (Pipeline) สามารถรวมและกำจัดความซ้ำซ้อนของเอเอสวีได้ แต่ไม่สามารถสรุปความซ้ำซ้อนและเก็บลำดับเอเอสวีซ้ำเพื่อศึกษาเปรียบเทียบ ซึ่งต้องทำงานในสภาพแวดล้อมไคม์ทูบนระบบปฏิบัติการลินุกซ์ (Linux Operating System) หรือการใช้โปรแกรมประเภทตารางการคำนวณ เช่น ไมโครซอฟท์เอ็กซ์เซล (Microsoft Excel) แม้ว่าเป็นโปรแกรมใช้งานง่ายและคุ้นเคย แต่ยังขาดฟังก์ชัน (Function) ในการคำนวณ และการประมวลผลที่ซับซ้อนเฉพาะทาง จากปัญหาการใช้โปรแกรมข้างต้น ทำให้นักวิจัยต้องใช้เวลาจัดการข้อมูลด้วยตนเอง ซึ่งเสี่ยงต่อความผิดพลาด ส่งผลให้การประมวลผลซ้ำซ้อน ใช้เวลานาน และเกิดผลลัพธ์การระบุอนุกรมวิธานลวง [9]

จากการสำรวจสภาพปัญหาข้างต้น ทำให้เกิดการการพัฒนาโปรแกรมให้ใช้งานง่ายและยืดหยุ่นบนสภาพแวดล้อมไพทอน แก้ปัญหาการรวมและกำจัดความซ้ำซ้อนของชุดข้อมูลที่ผ่านกระบวนการต้นทางมาแล้วอย่าง

เป็นระบบ โดยข้อมูลต้องไม่สูญหายในระหว่างการกำจัดความซ้ำซ้อน นักวิจัยสามารถนำผลไปใช้ในกระบวนการวิเคราะห์ปลายทางได้ถูกต้องและรวดเร็ว

1. วัตถุประสงค์การวิจัย

1.1 เพื่อพัฒนาโปรแกรมสำหรับการรวมและกำจัดความซ้ำซ้อน

1.2 เพื่อศึกษาความพึงพอใจของผู้ใช้งานที่มีต่อโปรแกรม

1.3 เพื่อประเมินประสิทธิภาพเชิงเทคนิคของโปรแกรม

2. เอกสารและงานวิจัยที่เกี่ยวข้อง

2.1 แนวคิดการพัฒนา

การพัฒนาโปรแกรมประกอบด้วยขั้นตอนในการปฏิบัติงานหลายขั้นตอน เพื่อให้การปฏิบัติงานมีประสิทธิภาพและสำเร็จตามระยะเวลาที่กำหนด จึงมีการกำหนดขั้นตอนการปฏิบัติงานเป็นลำดับ ตั้งแต่เริ่มโครงการจนสิ้นสุดโครงการ เรียกว่า วงจรการพัฒนาระบบ (System Development life Cycle: SDLC) อรยา ปรีชาพานิช [10] ได้อธิบายว่า การสร้างระบบสารสนเทศใหม่ หรือการปรับปรุงระบบสารสนเทศเดิมให้มีประสิทธิภาพมากยิ่งขึ้น วงจรการพัฒนาระบบประกอบด้วย 6 ขั้นตอนหลักคือ 1) เริ่มต้นจากการวางแผน เพื่อสำรวจข้อเท็จจริงที่เกี่ยวข้องกับประเด็นปัญหาโดยประยุกต์ใช้ระบบคอมพิวเตอร์เป็นเครื่องมือหลัก 2) วิเคราะห์ระบบเป็นการรวบรวมความต้องการใช้งานของผู้ใช้ และนำมาวิเคราะห์เป็นความต้องการของระบบที่จะพัฒนาขึ้นเพื่อใช้งาน 3) ออกแบบระบบเชิงตรรกะเป็นการกำหนดรายละเอียดขององค์ประกอบให้ระบบ กิจกรรมย่อยแบ่งเป็น 4 ส่วน คือการออกแบบในส่วนของรูปแบบผลลัพธ์ที่ได้จากระบบ (Output) การออกแบบในส่วนของรูปแบบการนำเข้าข้อมูล (Input) การออกแบบในส่วนของกระบวนการทำงาน (Process) และการออกแบบในส่วนของส่วนติดต่อผู้ใช้ (User Interface) 4) ออกแบบระบบเชิงกายภาพประกอบด้วย การกำหนดคุณลักษณะเฉพาะของฮาร์ดแวร์และซอฟต์แวร์ การออกแบบฐานข้อมูลของระบบ การออกแบบลักษณะเฉพาะของโปรแกรม เช่น ซูโดโค้ด (Pseudocode) ผังงานโปรแกรม (Program Flowchart) และผังงานระบบ (System Flowchart) 5) การพัฒนาระบบเป็นการทำให้ระบบเกิดผลลัพธ์ที่ใช้ได้จริง ประกอบด้วยการเขียนโปรแกรม การทดสอบ การติดตั้ง การถ่ายโอนข้อมูลจากระบบเดิมสู่ระบบใหม่ การจัดทำเอกสารของระบบ การฝึกอบรม และการประเมินผลระบบ 6) การบำรุงรักษาระบบเป็นการติดตามผลการใช้งานระบบ และให้ความช่วยเหลือแก่ผู้ใช้ระบบเพื่อให้สามารถใช้งานระบบได้อย่างต่อเนื่อง และมีประสิทธิภาพตามที่กำหนดไว้

2.2 เครื่องมือในการพัฒนา

เครื่องมือที่ใช้พัฒนาโปรแกรมและสร้างเอเอสวีเพื่อทดสอบโปรแกรม แบ่งออกเป็น 3 ส่วนหลัก ดังนี้

1) ภาษาที่ใช้ในการพัฒนาโปรแกรม ใช้ภาษาไพทอน (Python) เป็นภาษาโปรแกรมระดับสูง (High level) แปล และประมวลผลทีละคำสั่งทันที (Interpreted) มีลักษณะเชิงวัตถุ (Object Oriented) เหมาะสำหรับการพัฒนาแอปพลิเคชันแบบรวดเร็ว (Rapid Application Development) ไวยากรณ์ของภาษาถูกออกแบบให้เรียนรู้ง่าย ช่วยลดต้นทุนในการดูแลรักษาโปรแกรม แพ็คเกจช่วยให้เกิดการออกแบบโปรแกรมแบบโมดูล เอื้อให้สามารถนำโค้ดกลับมาใช้ใหม่ได้ [11] มีการเขียนโปรแกรมด้วยไพทอนอย่างแพร่หลายในวิทยาศาสตร์ข้อมูล (Data Science) แมชชีนเลิร์นนิง (Machine learning) และสามารถใช้ได้ฟรี [12]

2) สภาพแว้ดลอมและไลบรารี (Library) ประกอบด้วย 1) คอนด้า (Conda) เป็นเครื่องมือจัดการแพ็คเกจและสภาพแวดล้อมแบบโอเพนซอร์ส (Open Source) ถูกพัฒนาขึ้นเพื่อแก้ไขปัญหาที่ซับซ้อนในการจัดการงานคำนวณทางวิทยาศาสตร์ ให้สามารถใช้งานได้ข้ามระบบปฏิบัติการ ไม่ต้องใช้สิทธิ์ผู้ดูแลระบบ และรองรับการใช้งานในสภาพแวดล้อมเสมือนแบบแยกส่วน ทำให้เหมาะสำหรับการใช้งานบนระบบคอมพิวเตอร์ประสิทธิภาพสูง หรือเอชพีซี (High-Performance Computing: HPC) โครงสร้างพื้นฐานบนคลาวด์ โดยการควบคุม

สภาพแวดล้อมทำได้อย่างแม่นยำ ไม่พึ่งพาการตั้งค่าของระบบหลัก [13] 2) พิพ (Pip Installs Packages: Pip) เริ่มโครงการเมื่อปี 2008 พัฒนาโดยไพธอนแพ็คเกจจิ่งออธอริที หรือไพพีเอ (Python Packaging Authority: PyPA) เป็นเครื่องมือมาตรฐานเพื่อจัดการแพ็คเกจของไพทอน เช่นการติดตั้ง ปรับปรุง ถอนการติดตั้งแพ็คเกจด้วยไพพีไอ (Python Package Index: PyPI) ที่เป็นมาตรฐานตามพีอีพี 503ศูนย์สาม (PEP 503) [14] 3) แพนด้า (Pandas) เป็นแพ็คเกจภาษาไพทอนสำหรับจัดการข้อมูลที่รวดเร็ว ยืดหยุ่น และชัดเจน ออกแบบการทำงานกับข้อมูลที่มีโครงสร้างเชิงสัมพันธ์ (Relational) เป็นเครื่องมือจัดการ และวิเคราะห์ข้อมูลแบบโอเพนซอร์สที่ใช้งานได้จริง [15] 4) โครงสร้างข้อมูลของแพนด้า ได้แก่ ซีรีส์ (Series) 1 มิติ และ เดตาเฟรม (DataFrame) 2 มิติ รองรับการใช้งานได้หลายศาสตร์ เช่น สาขาวิทยาศาสตร์ การเงิน สถิติ สังคมศาสตร์ วิทยาศาสตร์ข้อมูลและวิศวกรรม เป็นต้น [16] 5) นัมไพ (Numerical Python: Numpy) เป็นแพ็คเกจที่สำคัญสำหรับการคำนวณเชิงวิทยาศาสตร์ มีโครงสร้างข้อมูลพื้นฐานแบบเอ็นดีอาร์เรย์ (ndarray) เก็บข้อมูลชนิดเดียวกันทั้งหมด [17] 6) แมตพลอตลิบ (MATLAB และ plotting library: Matplotlib) เป็นไลบรารีสำหรับการสร้างภาพกราฟฟิก (Visualizations) ประกอบด้วยการแสดงผลภาพกราฟฟิกทางสถิติที่สามารถเคลื่อนไหวและมีปฏิสัมพันธ์กับผู้ใช้ ซึ่งมีจุดเด่นคล้ายแมตแลบ (Matrix Laboratory: MATLAB) [18] 7) ไบโอไพทอน (Biopython) เป็นความร่วมมือระดับนานาชาติจากความสมัครใจของนักพัฒนาที่มีประสบการณ์สูง จัดทำไลบรารีด้วยภาษาไพทอนให้อยู่ในรูปแบบโอเพนซอร์ส เพื่อแก้ปัญหาด้านชีวสารสนเทศ (Bioinformatics) โดยมุ่งให้เกิดไลบรารีคุณภาพสูง สำหรับการประมวลผลข้อมูลชีวภาพอย่างมีประสิทธิภาพ และลดการเขียนโปรแกรมที่ซ้ำซ้อน [19] 8) ทีเคอินเทอร์เฟซ (Tk interface) เป็นอินเทอร์เฟซมาตรฐานของภาษาไพทอน สำหรับใช้งานกับเครื่องมือสร้างส่วนติดต่อผู้ใช้แบบกราฟฟิก (Tcl/Tk GUI toolkit) [20] 9) อัพเซต (UpSet) เป็นการแสดงภาพชุดข้อมูลที่มีการตัดกันในรูปแบบเมทริกซ์ (Metric) ซึ่งเป็นเครื่องมือที่เหมาะสำหรับการวิเคราะห์เชิงปริมาณของข้อมูลที่มีชุดข้อมูลมากกว่าสามชุด ต่างจากแผนภาพเวนน์ (Venn) ที่ไม่สามารถขยายขนาดได้เมื่อมีชุดข้อมูลเกินสามหรือสี่ชุด [21]

3) การเตรียมข้อมูลเอเอสวี ตารางข้อมูลอะบันเดินซ์ (Abundance) และแทกซอโนมี (Taxonomy) เพื่อทดสอบโปรแกรมนี้ ใช้ไพป์ไลน์ของไคม์ทูเป็นแพลตฟอร์ม (Platform) ชีวสารสนเทศแบบโอเพนซอร์ส สามารถวิเคราะห์ข้อมูลไมโครไบโอมและสร้างเอเอสวีที่มีคุณภาพด้วยดาดาทู (DADA2) [8, 22]

2.3 งานวิจัยที่เกี่ยวข้อง

บทบาทสำคัญของเอเอสวีของยีน 16 เอสอาร์อาร์เอ็นเอ มาใช้เป็นลำดับตัวแทนสำหรับการวิเคราะห์ไมโครไบโอม คือการช่วยลดความซ้ำซ้อนของข้อมูล สอดคล้องกับ Wang และคณะ [22] อธิบายว่าดาดาทูใช้โมเดลการเรียนรู้ข้อผิดพลาดจากข้อมูลการจัดลำดับ (Denoising Algorithm) เพื่อแยกลำดับที่เหมือนกันทุกตำแหน่งเอเอสวี ออกจากข้อผิดพลาดในการจัดลำดับอย่างแม่นยำ โดยไม่อนุญาตให้มีมิสแมตช์ (Mismatch) แม้แต่เพียงตำแหน่งเดียว 1 เบส ถือว่าเป็นเอเอสวีคนละตัว ทำให้เอเอสวีมีความเฉพาะตัว (Unique Sequences) ความละเอียดสูง อย่างไรก็ตามแม้ว่าลำดับเอเอสวีจะมีความแม่นยำสูง แต่อาจเกิดความซ้ำซ้อนได้ระหว่างชุดข้อมูลที่มาจากหลายการศึกษา ซึ่งความแตกต่างเพียงเล็กน้อยสามารถนำไปสู่การแปลผลที่ผิดพลาดและประมวลผลช้า ดังนั้นต้องปรับปรุงข้อมูลก่อนการวิเคราะห์ปลายทาง Özkurt และคณะ [23] ได้พัฒนาโลตัสทู (LotuS2) เป็นไพป์ไลน์สำหรับการวิเคราะห์ข้อมูลแอมพลิคอนซีเควนส์ (Amplicon Sequence) ที่มีประสิทธิภาพด้านความเร็ว ความแม่นยำ และความสามารถในการทำซ้ำผลลัพธ์ได้ (Reproducibility) จุดเด่นของโลตัสทู คือ การรวมขั้นตอนการกรองลำดับ (Filtering) การจัดกลุ่ม (Clustering) การขยายลำดับตัวแทน (Seed Extension) การจัดกลุ่มซ้ำซ้อน การกรองโฮสดีเอ็นเอ (Host DNA) และการจำแนกสายพันธุ์อย่างครอบคลุม เมื่อเปรียบเทียบกับเครื่องมือชั้นนำ เช่น ไคม์ทู ดาดาทู และมอเธอร์ (Mothur) พบว่าโลตัสทูสามารถจัดการข้อมูลได้เร็วกว่าโดยเฉลี่ยถึง 29 เท่า และให้ผลลัพธ์ที่แม่นยำกว่า นอกจากนั้นการพัฒนาเครื่องมือสำหรับจัดการและวิเคราะห์ข้อมูลไมโครไบโอมของ Lu และคณะ [24] ได้พัฒนาไมโครไบโอมอะนาลิสต์สองจุดศูนย์ (MicrobiomeAnalyst 2.0) เครื่องมือออนไลด์ที่ครบวงจร (Web Application) ถูกพัฒนาขึ้นเพื่อใช้กับข้อมูลมัลติโอมิก (Multi-Omics) และการวิเคราะห์ข้อมูลที่ซับซ้อน

สำหรับนักวิจัยที่ไม่ได้เชี่ยวชาญด้านสถิติหรือการเขียนโปรแกรม รองรับการวิเคราะห์จากข้อมูลดิบ (Raw Sequencing data) ที่ใช้เวิร์กโฟลว์ของดาดาทูเพื่อสร้างเอเอสวี

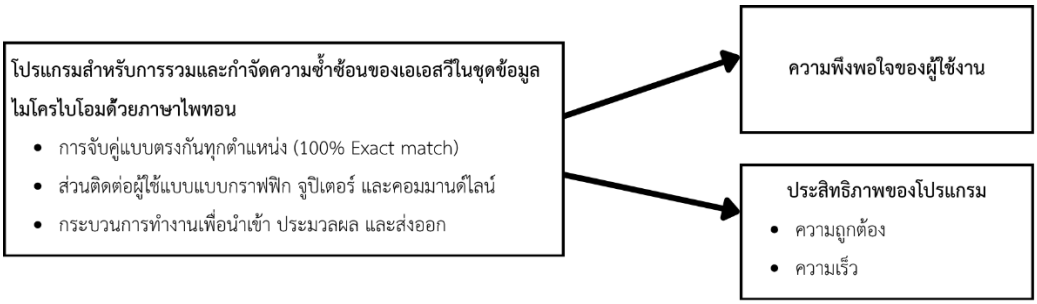
สำหรับงานวิจัยนี้ได้เลือกภาษาไพทอนมาเป็นเครื่องมือในการพัฒนาโปรแกรม สอดคล้องกับทีมพัฒนาไคม์ทูได้พัฒนาเครื่องมือเรียกว่าไคม์ทูอาร์ติแฟกต์เอพีไอ (QIIME 2 Artifact API) เป็นอินเทอร์เฟซภาษาไพทอน ซึ่งถูกออกแบบมาสำหรับนักวิจัย ชีวสารสนเทศ หรือผู้เชี่ยวชาญด้านไมโครไบโอม ที่ต้องการทำการวิเคราะห์ข้อมูลอย่างเป็นระบบและทำซ้ำได้ภายในสภาพแวดล้อมของจูพิเทอะโน้ตบุ๊ค (Jupyter Notebook) ไม่จำเป็นต้องใช้อินเทอร์เฟซแบบคอมมานด์ไลน์ทั้งหมด [25] จากโปรแกรมและงานวิจัยที่เกี่ยวข้องข้างต้น แสดงให้เห็นว่าปัจจุบันมีโปรแกรมหลากหลาย แต่ถูกออกแบบมาเพื่อวัตถุประสงค์ที่แตกต่างกัน ซึ่งยังคงมีช่องว่างของงานวิจัยสำหรับโปรแกรมที่ต้องการความยืดหยุ่นและสามารถทำงานเฉพาะ เพื่อเปรียบเทียบฟังก์ชันของโปรแกรมที่มีในปัจจุบันกับฟังก์ชันของโปรแกรมที่พัฒนา แสดงรายละเอียดดังตารางที่ 1

ตารางที่ 1 เปรียบเทียบฟังก์ชันของโปรแกรมที่มีในปัจจุบันกับฟังก์ชันของโปรแกรมที่พัฒนา

ฟังก์ชันของโปรแกรมที่มีอยู่ในปัจจุบัน	ฟังก์ชันของโปรแกรมที่พัฒนา
ดาดาทู ใช้สร้างเอเอสวีจากข้อมูลดิบ	จัดการข้อมูลเอเอสวีหลังจากที่สร้างเสร็จแล้ว
โลตัสทู ไพป์ไลน์ครบวงจรสำหรับการวิเคราะห์ข้อมูลแอมพลิคอน	เป็นไพทอนสคริปต์ (Script) ที่เบาและยืดหยุ่น เน้นเฉพาะการรวมข้อมูล ไม่ใช่ไพป์ไลน์ทั้งหมด
ไมโครไบโอมอะนาลิสต์สอง เครื่องมือออนไลน์สำหรับสร้างเอเอสวีด้วยดาดาทูและการวิเคราะห์ปลายทาง	ใช้เตรียมข้อมูลนำเข้าเพื่อใช้กับการวิเคราะห์ปลายทาง
ไคม์ทูอาร์ติแฟกต์เอพีไอ ไพทอนอินเทอร์เฟซสำหรับทำงานกับข้อมูลเฉพาะ เช่น คิวซีเอไฟล์ (qza file) ภายในสภาพแวดล้อมของไคม์ทู	โปรแกรมนี้ทำงานโดยตรงกับไฟล์มาตรฐาน เช่น ฟาสต้าไฟล์ (fasta file) และเท็กซ์ไฟล์ (text file) ทำให้มีความยืดหยุ่นและใช้งานภายนอกสภาพแวดล้อมของไคม์ทูได้

2.4 กรอบแนวคิดการวิจัย

การวิจัยนี้เป็นการพัฒนาโปรแกรมสำหรับการรวมและกำจัดความซ้ำซ้อนของชุดข้อมูลไมโครไบโอมด้วยภาษาไพทอน สามารถค้นหาเอเอสวีที่ซ้ำด้วยวิธีการจับคู่แบบตรงกันทุกตำแหน่ง ผู้ใช้สามารถใช้งานได้ผ่านมุมมองกราฟฟิก จูปิเตอร์ และคอมมานด์ไลน์ มีกระบวนการทำงานเพื่อนำเข้าข้อมูล ประมวลผล และแสดงผลสำหรับนำไปวิเคราะห์ต่อ โปรแกรมถูกประเมินประสิทธิภาพด้านความถูกต้องและความเร็ว ประเมินความพึงพอใจการใช้งานจากผู้ใช้ หรือนักชีวสารสนเทศ เพื่อนำไปตีความและผลสรุปทางชีววิทยาต่อไป



ภาพที่ 1 กรอบแนวคิดการวิจัย

วิธีดำเนินการวิจัย

1. เครื่องมือการวิจัย

1.1 การวิจัยครั้งนี้ได้ใช้แบบสอบถามความพึงพอใจโปรแกรมโดยแบบสอบถามเป็นการประเมินความพึงพอใจแบบมาตราส่วน 5 ระดับ ประกอบด้วย 2 ตอน คือ ตอนที่ 1 ข้อมูลทั่วไปของผู้ตอบแบบสอบถาม และตอนที่ 2 ความพึงพอใจของผู้ใช้ระบบ โดยข้อคำถามแบ่งเป็น 5 ด้าน จำนวน 10 ข้อ ซึ่งคำถามทั้ง 5 ด้าน ได้แก่ 1) ความสะดวกในการใช้งาน 2) ความถูกต้องของผลลัพธ์ที่ได้ 3) ความรวดเร็วในการประมวลผล 4) ความเหมาะสมของอินเทอร์เฟซ และ 5) ความพึงพอใจโดยรวม โดยผ่านการหาค่าความสอดคล้องระหว่างข้อคำถามและวัตถุประสงค์จากนักชีวสารสนเทศจำนวน 2 ท่าน มีค่าความสอดคล้องเท่ากับ 1.00

1.2 งานวิจัยนี้มีการประเมินประสิทธิภาพเชิงเทคนิคด้านความถูกต้องและความเร็ว ใช้ชุดข้อมูลจำนวน 4 ชุด จากทีมวิจัยไมโครอะเรย์แบบครบวงจร ศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ (ไบโอเทค) สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ (สวทช.) วิจัยเกี่ยวกับจุลินทรีย์ในลำไส้ของกุ้งกุลาดำ Penaeus monodon โดยนำข้อมูล 16 เอสอาร์อาร์เอ็นเอ จากฐานข้อมูลเอ็นซีบีไอ (NCBI) การเข้าถึงข้อมูลอ้างอิงจากรหัสโครงการดังนี้ PRJNA673004 [26], PRJNA689097 [27], PRJNA553862 [28] และ PRJNA553862 [29]

1.3 เครื่องมือในการพัฒนาโปรแกรม ได้แก่ จูพิเทอะโน้ตบุ๊ค จูปิเตอร์แล็บ วิชวลสตูดิโอโค้ด (Visual Studio Code) คอมมานด์พรอมท์ (Command prompt) และภาษาไพทอน ภายใต้สภาพแวดล้อมการทำงานและไลบรารี่ ประกอบด้วยคอนด้า พิพ แพนด้า นัมไพ แมตพลอตลิบ ไบโอไพทอน ไคม์ทู ติกอินเทอร์ และอัพเซต

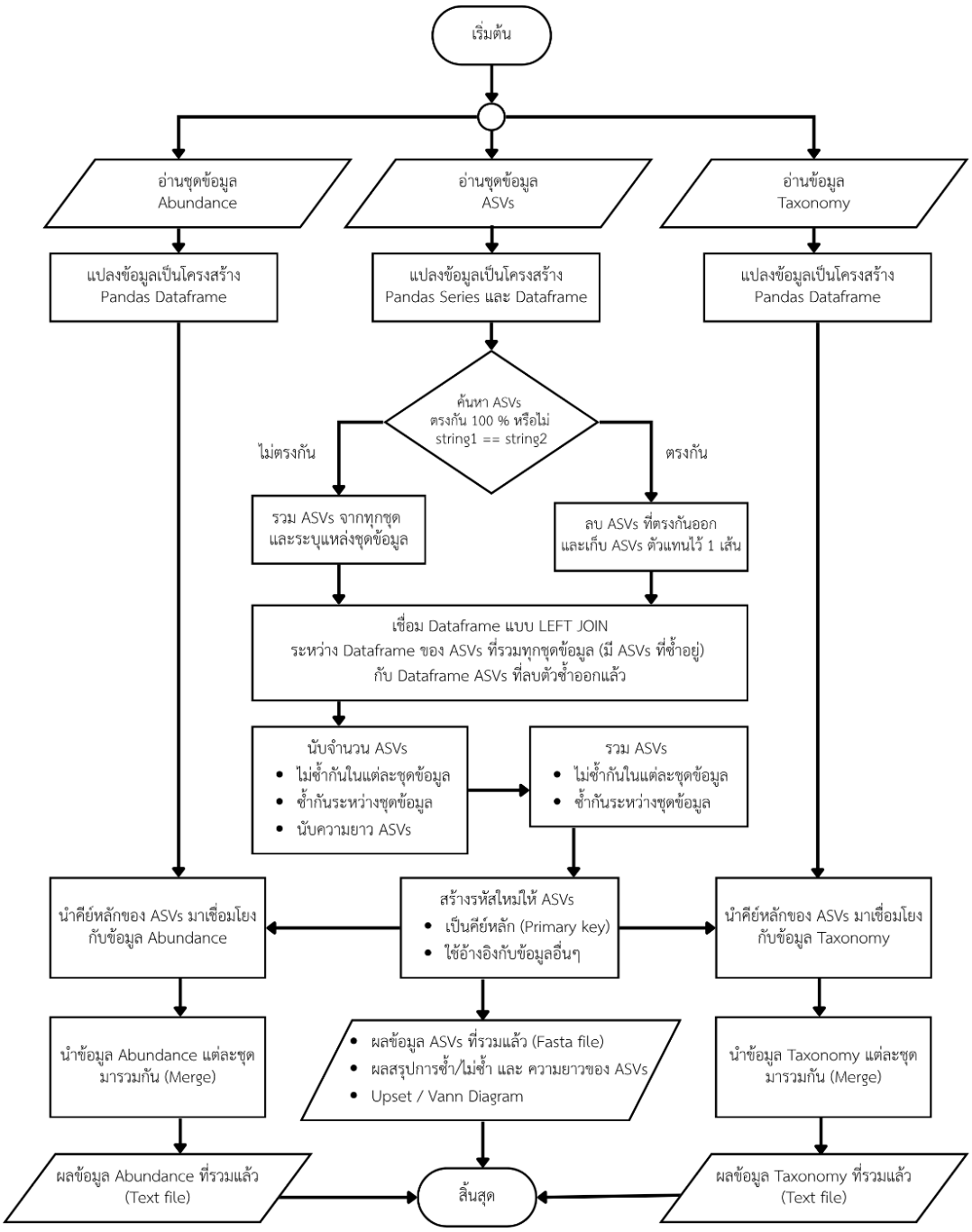
2. กลุ่มเป้าหมาย

กลุ่มเป้าหมายจำนวน 4 คนเท่านั้น สำหรับประเมินความพึงพอใจโปรแกรมนี้ ใช้การคัดเลือกแบบเจาะจงตามคุณสมบัติดังนี้ 1) เป็นผู้ปฏิบัติงานวิจัยด้านไมโครไบโอมโดยตรง 2) มีความจำเป็นต้องรวมชุดข้อมูลเอเอสวีจากหลายการทดลอง 3) เป็นผู้ใช้งานที่ใช้โปรแกรมเฉพาะด้านชีววิทยา จุลชีววิทยา และชีวสารสนเทศ

3. ขั้นตอนการดำเนินการวิจัย

3.1 กระบวนการพัฒนาโปรแกรมเริ่มอย่างเป็นขั้นตอนตั้งแต่การวางแผน วิเคราะห์ ออกแบบ พัฒนา ทดสอบ และนำไปใช้งานจริง เพื่อพัฒนาโปรแกรมให้ตอบสนองความต้องการของผู้ใช้งานมากที่สุด ทั้งนี้การใช้เอสดีแอลซีช่วยให้การพัฒนาโปรแกรมมีทิศทางชัดเจนตามกรอบการดำเนินงาน [10]

3.2 การออกแบบขั้นตอนการรวมและกำจัดความซ้ำซ้อน มีรายละเอียดดังนี้ 1) ใช้ไลบรารีไบโอไพทอนอ่านฟาสต้าไฟล์ของเอเอสวีทั้งหมด 2) แปลงลำดับเอเอสวีเป็นซีรีส์ก่อน (Serie) และแปลงเป็นโครงสร้างข้อมูลแบบดาต้าเฟรมด้วยไลบรารีแพนด้า (Pandas DataFrame) 3) ค้นหาลำดับเอเอสวีที่เหมือนกันทุกตำแหน่ง Sequence1 == Sequence2 เรียกว่าวิธีการจับคู่แบบตรงกันทุกตำแหน่งของเอเอสวี (100% Exact Match Method) 4) จากนั้นแบ่งข้อมูลออกเป็น 2 กลุ่ม กลุ่มแรกรวมชุดลำดับเอเอสวีทั้งหมดเข้าด้วยกัน (มีลำดับเอเอสวีที่ซ้ำรวมอยู่ด้วย) และระบุแหล่งข้อมูล กลุ่มที่ 2 ทำการลบลำดับเอเอสวีซ้ำซ้อน (Remove Duplicates) เพื่อให้ได้ชุดข้อมูลเอเอสวีที่ไม่ซ้ำกันเลย จากนั้นนำชุดข้อมูลทั้ง 2 มาเชื่อมกันแบบใช้ข้อมูลด้านซ้ายเป็นหลัก (Left Join) หรือชุดลำดับเอเอสวีกลุ่มแรกเป็นหลัก ซึ่งวิธีการนี้ทำให้สามารถระบุได้ว่าลำดับเอเอสวีใดตรงกัน หรือไม่ตรงกันแล้วนับจำนวนเอเอสวีรวมทั้งความยาวของลำดับเอเอสวี 5) จากนั้นรวมลำดับเอเอสวีที่ไม่ซ้ำกันเลย มากำหนดรหัสใหม่ เช่น จากรหัสเดิม b692df4926b5a7e01493ac39cffbc3f7 เป็น ASV_0001 เป็นต้น และ 6) ใช้รหัสใหม่เป็นคีย์หลัก (Primary Key) ในการเชื่อมโยงกับตารางข้อมูลอะบันเดินซ์และแทกซอโนมี จากนั้นนำแต่ละชุดข้อมูลมารวมกัน 7) สรุปข้อมูลเอเอสวี สร้างไฟล์และแผนภาพแบบกราฟฟิกอัพเซต กรณีชุดข้อมูลมากกว่าสามชุด หรือ แสดงเป็นแผนภาพเวนน์เมื่อมีชุดข้อมูลไม่เกินสามชุด แสดงดังภาพที่ 2



ภาพที่ 2 แผนภาพขั้นตอนการรวมและกำจัดความซ้ำซ้อน

4. สถิติที่ใช้ในการวิจัย ได้แก่ ค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐาน สามารถแปลความหมายของค่าคะแนน [30] ดังนี้

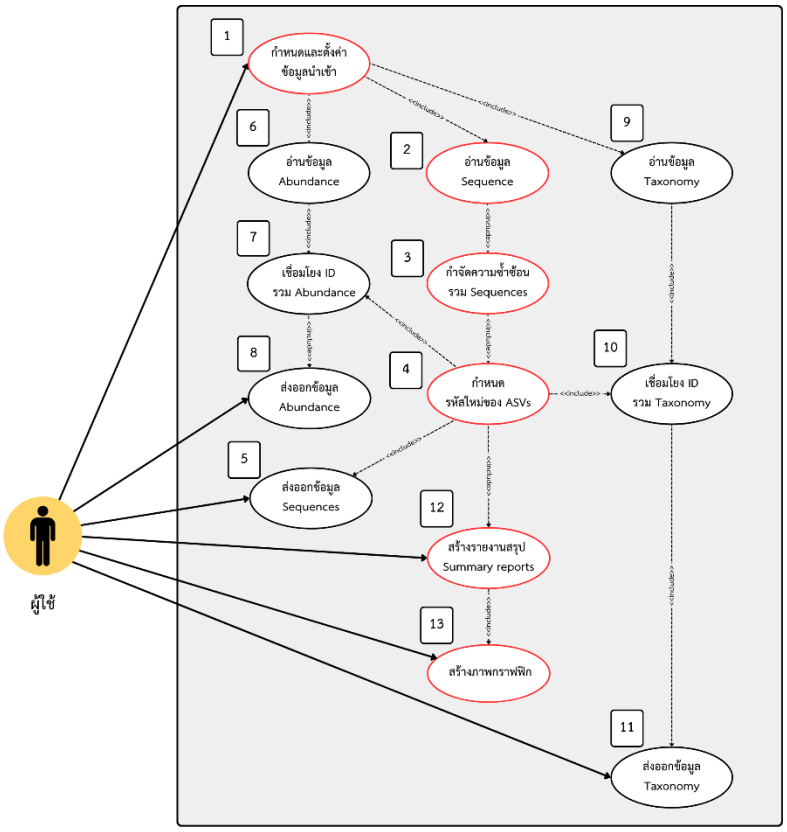
ค่าเฉลี่ยเท่ากับ 4.51 – 5.00 หมายความว่า ระดับดีมาก
ค่าเฉลี่ยเท่ากับ 3.51 – 4.50 หมายความว่า ระดับดี
ค่าเฉลี่ยเท่ากับ 2.51 – 3.50 หมายความว่า ระดับปานกลาง

ค่าเฉลี่ยเท่ากับ 1.51 – 2.50 หมายความว่า ระดับน้อย
ค่าเฉลี่ยเท่ากับ 1.00 – 1.50 หมายความว่า ระดับน้อยมาก

ผลการวิจัย

1. ผลการพัฒนาโปรแกรมสำหรับการรวมและกำจัดความซับซ้อน

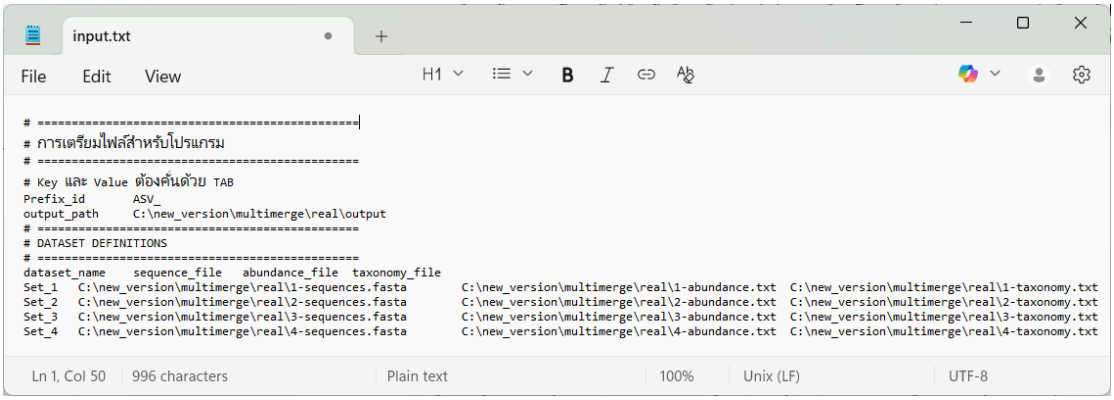
1.1. ผู้ใช้งานหลักของโปรแกรมนี้คือนักวิจัยด้านชีววิทยาและชีวสารสนเทศ (Actor) มีสิทธิ์ในการเข้าถึงฟังก์ชันทั้งหมดของโปรแกรม ผู้ใช้งานมีปฏิสัมพันธ์กับฟังก์ชันดังนี้ 1) กำหนดไฟล์ตั้งค่านำเข้าข้อมูล 2) อ่านลำดับของเอเอสวี 3) รวมและลบลำดับของเอเอสวีที่ซ้ำกัน 4) สร้างรหัสหลักใหม่ให้กับเอเอสวีที่รวมกัน 5) สร้างฟาสต้าไฟล์ของเอเอสวีที่รวมแล้ว 6) อ่านข้อมูลอะบันเดินซ์ 7) รวมและเชื่อมข้อมูลอะบันเดินซ์กับรหัสเอเอสวีใหม่ 8) สร้างเท็กซ์ไฟล์ของข้อมูลอะบันเดินซ์ที่รวมแล้ว 9) อ่านข้อมูลแทกซอโนมี 10) รวมและเชื่อมข้อมูลแทกซอโนมีกับรหัสเอเอสวีใหม่ 11) สร้างเท็กซ์ไฟล์ของข้อมูลแทกซอโนมีที่รวมแล้ว 12) สร้างไฟล์รายงานสรุปจำนวนเอเอสวีที่ซ้ำและไม่ซ้ำกัน และ 13) สร้างภาพกราฟฟิกจากจำนวนเอเอสวีที่ซ้ำและไม่ซ้ำกัน แสดงดังภาพที่ 3



ภาพที่ 3 แบบจำลองความสัมพันธ์ระหว่างผู้ใช้ที่มีสิทธิ์ในการเข้าถึงฟังก์ชันทั้งหมดของโปรแกรม

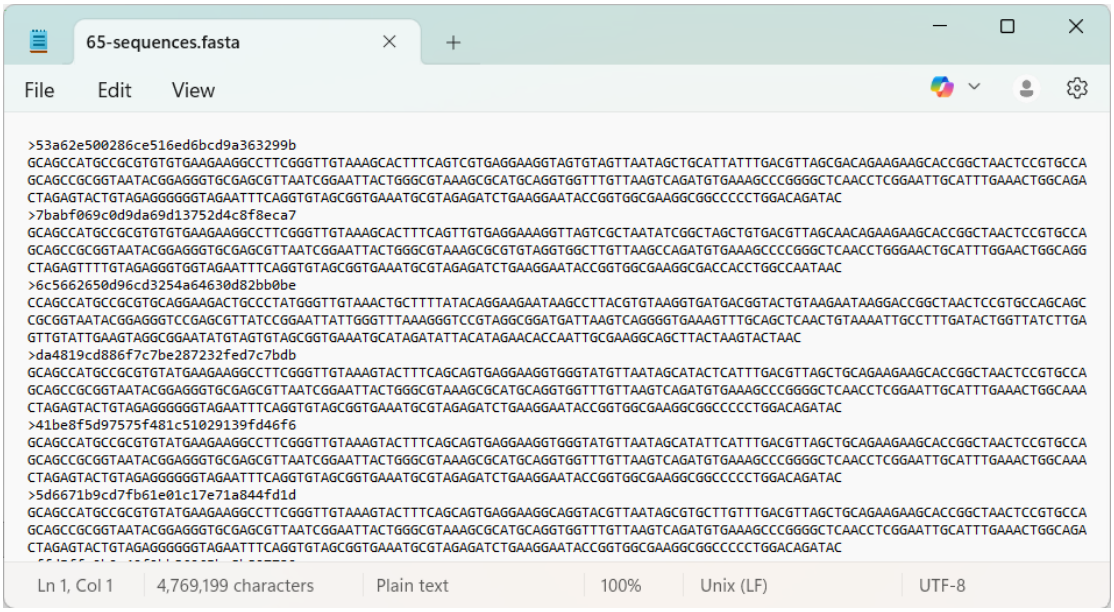
1.2 การเตรียมไฟล์นำเข้า ก่อนพิมพ์คำสั่งผ่านคอมแมนด์พรอมพ์เพื่อเรียกใช้โปรแกรม ต้องเตรียมเท็กซ์ไฟล์สำหรับข้อมูลนำเข้าก่อน การตั้งค่าไฟล์ประกอบด้วย 3 ส่วน 1) การกำหนดอักขระนำหน้า (Prefix_id) ของรหัสเอเอสวี เช่น “ASV_” 2) การระบุเส้นทางแฟ้มที่ต้องการเก็บผลจากโปรแกรม (output_path) เช่น C:\output 3) การระบุชุดข้อมูลหัวตารางหรือคอลัมน์ต้องคั่นด้วยแท็บ (Tab) ประกอบด้วย ชื่อชุดข้อมูล (dataset_name)

เส้นทางไปยังฟาสต้าไฟล์ (sequence_file) เส้นทางไปยังอะบันเดินซ์ไฟล์ (abundance_file) และเส้นทางไปแทกซอโนมีไฟล์ (taxonomy_file) ต้องคั่นด้วยแท็บเช่นกัน แสดงดังภาพที่ 4



ภาพที่ 4 การเตรียมไฟล์นำเข้า

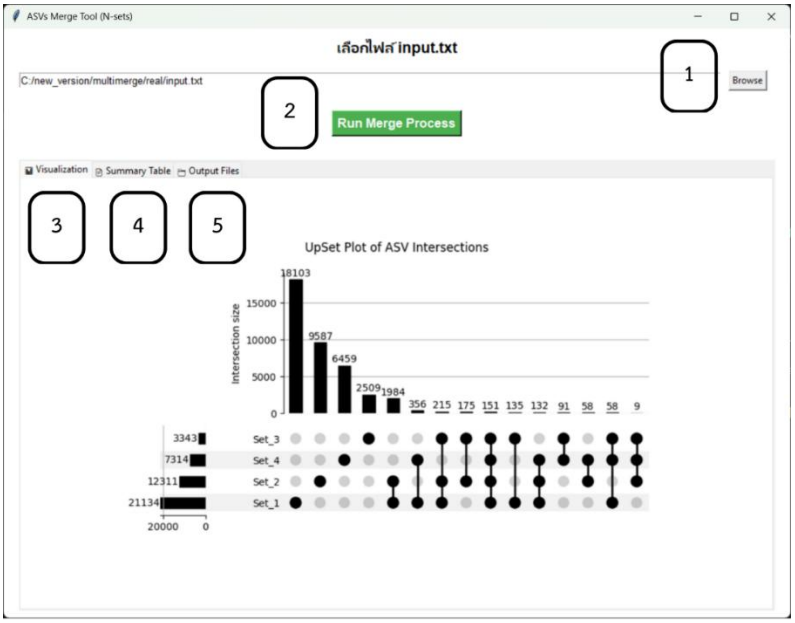
ตัวอย่างไฟล์ข้อมูลเอเอสวีที่ผ่านกระบวนการวิเคราะต้นทางด้วยไคม์ทู โครงสร้างของไฟล์มี 2 ส่วน 1) บรรทัดส่วนหัวขึ้นต้นด้วยเครื่องหมาย > ตามด้วยเป็นชื่อหรือรหัสของลำดับเอเอสวี เช่น >ASV_1 หรือ >53a62e500286ce516ed6bcd9a363299b 2) บรรทัดถัดมาต่อบรรทัดส่วนหัวใช้เก็บลำดับเอเอสวีของยีน 16 เอสอาร์อาร์เอ็นเอ ประกอบด้วยอักษร A, G, C และ T แสดงดังภาพที่ 5



ภาพที่ 5 ตัวอย่างลำดับเอเอสวีที่ผ่านกระบวนการวิเคราะต้นทางด้วยไคม์ทู

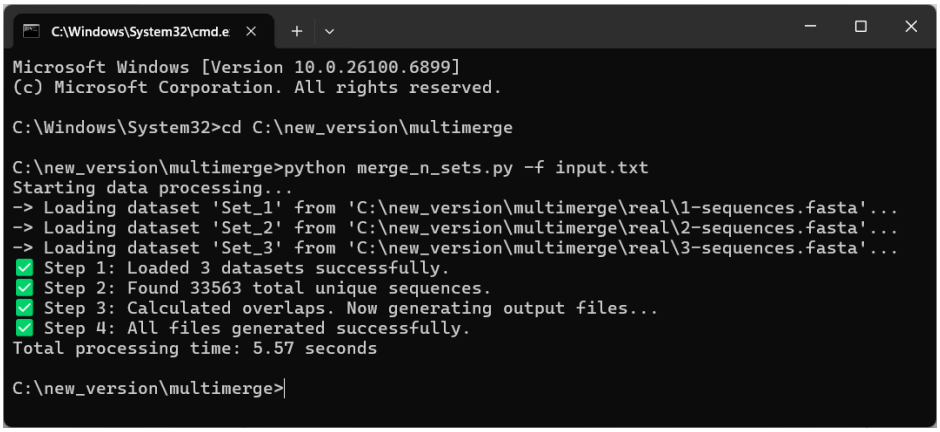
การแสดงผลสำหรับส่วนติดต่อผู้ใช้แบบกราฟฟิก หลังจากเตรียมไฟล์นำเข้าแล้ว ต้องพิมพ์คำสั่งผ่านคอมแมนด์พรอมพ์เพื่อเรียกใช้โปรแกรมด้วยคำสั่ง “python gui_merge_app.py” ขั้นตอนการใช้งานดังนี้ 1) กดปุ่ม “Browse” เลือกไฟล์นำเข้า 2) ก่อนกดปุ่ม “Run Merge Process” ส่วนแสดงผลประกอบด้วย 3 แท็บคือ

3) แท็บแผนภาพกราฟฟิก “Visualization” เพื่อแสดงความสัมพันธ์ระหว่างชุดข้อมูล 4) แท็บรายงานสรุป “Summary Table” เพื่อแสดงความยาวของเอเอสวี จำนวนเอเอสวีแต่ละชุดข้อมูล จำนวนเอเอสวีที่ซ้ำและไม่ซ้ำระหว่างชุดข้อมูล 5) สุดท้ายแท็บเชื่อมโยงยังไฟล์ผลต่าง ๆ “Output Files” ประกอบด้วยไฟล์รวบรวมของเอเอสวีที่ซ้ำและไม่ซ้ำ (ASVs_overlapping.txt) ฟาสต้าไฟล์ของเอเอสวีที่กำหนดรหัสใหม่และกำจัดความซ้ำซ้อนแล้ว (ASVs_sequences.fasta) ไฟล์ตารางสรุปจำนวนเอเอสวีและความยาวของเอเอสวี (summary_ASVs.txt) ไฟล์อะบันเดินซ์ที่รวมและเชื่อมกับรหัสเอเอสวีใหม่แล้ว (merge_abundances.txt) ไฟล์แทกซอโนมีที่รวมและเชื่อมกับรหัสเอเอสวีไอดีใหม่แล้ว (merge_taxonomy.txt) แสดงดังภาพที่ 6



ภาพที่ 6 การใช้งานและการแสดงผลสำหรับส่วนติดต่อผู้ใช้แบบกราฟฟิก

โปรแกรมที่พัฒนานอกจากสามารถทำงานได้บนส่วนติดต่อผู้ใช้แบบกราฟฟิกที่ทำงานบนระบบปฏิบัติวินโดวส์ (Windows) ผู้ใช้ยังสามารถใช้งานได้บนคอมมานด์ไลน์ เพื่อให้สามารถทำงานได้บนระบบปฏิบัติการลินุกซ์ เช่น อูบันตู (Ubuntu) โดยผลลัพธ์ที่ได้เหมือนกันกับส่วนติดต่อผู้ใช้แบบกราฟฟิก แสดงดังภาพที่ 7



ภาพที่ 7 การใช้งานส่วนติดต่อผู้ใช้แบบคอมมานด์ไลน์

2. ผลการศึกษาความพึงพอใจของผู้ใช้งานที่มีต่อโปรแกรม

ผลการประเมินความพึงพอใจ ได้จากแบบศึกษาความพึงพอใจของผู้ใช้ จำนวน 4 คน ในการทดลองใช้โปรแกรม โดยความพึงพอใจเฉลี่ย 4.29 ระดับดี โดยมีรายละเอียดดังตารางที่ 2

ตารางที่ 2 ความความพึงพอใจของผู้ใช้งานที่มีต่อโปรแกรม

รายการประเมิน	$\bar{X}$	S.D.	แปลผล
1. ความสะดวกในการใช้งาน	3.75	0.99	ดี
2. ความถูกต้องของผลลัพธ์ที่ได้	5.00	0.00	ดีมาก
3. ความรวดเร็วในการประมวลผล	4.25	0.46	ดี
4. ความเหมาะสมของอินเทอร์เฟซ	3.94	0.84	ดี
5. ความพึงพอใจโดยรวม	4.50	0.58	ดี
รวมเฉลี่ย	**4.29**	**0.69**	**ดี**

จากตารางที่ 2 พบว่าโปรแกรมให้ผลลัพธ์ที่มีความถูกต้อง ได้คะแนนความพึงพอใจ 5.00 ระดับดีมาก รองลงมาเป็นความพึงพอใจโดยรวมต่อการใช้งาน ได้คะแนนความพึงพอใจ 4.50 ระดับดี รองลงมาเป็นความรวดเร็วในการประมวลผล ได้คะแนนความพึงพอใจ 4.25 ระดับดี ส่วนความเหมาะสมของอินเทอร์เฟซต่อผู้ใช้ ได้คะแนนความพึงพอใจ 3.94 ระดับดี และความสะดวกในการใช้งาน ได้คะแนนความพึงพอใจ 3.75 ระดับดี

3. ผลการประเมินประสิทธิภาพเชิงเทคนิคของโปรแกรม

การทดสอบประสิทธิภาพด้านความถูกต้องและความเร็ว ใช้ชุดข้อมูลจากทีมวิจัยไมโครอะเรย์แบบครบวงจร ศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ (ไบโอเทค) สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ (สวทช.) วิจัยเกี่ยวกับจุลินทรีย์ในลำไส้ของกุ้งกุลาดำ Penaeus monodon โดยนำข้อมูล 16 เอสอาร์อาร์เอ็นเอจากฐานข้อมูลเอ็นซีบีไอ ศูนย์ข้อมูลเทคโนโลยีชีวภาพแห่งชาติ ประเทศสหรัฐอเมริกา สำหรับการค้นหาข้อมูลดิบดังกล่าว อ้างอิงจากรหัสโครงการ (Bioproject number) จากนั้นข้อมูลดิบมาวิเคราะห์ต้นทางด้วยไพป์ไลน์ของโปรแกรมไคม์ทูโดยผ่านกระบวนการกรองคุณภาพข้อมูลเอเอสวีก่อน (Denoise > 50%) ได้ลำดับเอเอสวีจำนวนรวม 44,102 สาย เพื่อนำมาทดสอบประสิทธิภาพเชิงเทคนิคในโปรแกรมนี้ มีรายละเอียดดังตารางที่ 3

ตารางที่ 3 แสดงรายการชุดข้อมูลเอเอสวีที่นำมาทดสอบประสิทธิภาพเชิงเทคนิค

ชุดที่	รหัสโครงการ	รายละเอียดงานวิจัย	จำนวนเอเอสวี
1	PRJNA673004	การนำกุ้งกุลาดำวัยอ่อนมาปรับตัวเข้ากับระดับความเค็มที่แตกต่าง เพื่อศึกษาจุลินทรีย์ในลำไส้และการแสดงออกของยีนของกุ้งที่ตอบสนองต่อการเปลี่ยนแปลงความเค็ม [27]	21,134
2	PRJNA689097	การวิเคราะห์เพื่อทำความเข้าใจปฏิสัมพันธ์ระหว่างกุ้งกับจุลินทรีย์ในลำไส้เมื่อสัมผัสกับเชื้อแบคทีเรียก่อโรค *Vibrio harveyi* [28]	12,311
3	PRJNA553862	การวิเคราะห์หลายมิติในการตรวจสอบจุลินทรีย์ในลำไส้ เมแทบอลิซึมทรานสคริปโตมิกส์ของกุ้งและความสัมพันธ์ที่มีต่อการเติบโตของกุ้ง [29]	3,343
4	PRJNA553862	การวิเคราะห์แบคทีเรียในระยะการพัฒนาช่วงต้นของกุ้งกุลาดำ [30]	7,314
รวมจำนวนเอเอสวีทั้งหมด			**44,102**

การทดสอบประสิทธิภาพใช้วิธีการเพิ่มชุดข้อมูลตั้งแต่ 2 ชุด จนถึง 10 ชุด โดยดำเนินการทดสอบซ้ำ 3 ครั้ง เพื่อหาความเร็วเฉลี่ยด้วยเครื่องคอมพิวเตอร์โน๊ตบุ๊คบนระบบปฏิบัติการวินโดวส์ 11 (AMD Ryzen 5 5600H) หน่วยความจำขนาด 16 กิกะไบต์ (RAM 16 GB) และทดสอบการเพิ่มชุดข้อมูลซ้ำ เพื่อประเมินการจำกัดความซ้ำซ้อนระหว่างชุดข้อมูล ผลการทดสอบแสดงดังตารางที่ 4

ตารางที่ 4 แสดงผลการทดสอบเพื่อประเมินความถูกต้องการจำกัดความซ้ำซ้อนและความเร็ว

ลำดับที่	ชุดข้อมูลที่ใช้ทดสอบ	เอเอสวีทั้งหมด	เอเอสวีไม่ซ้ำ	เวลาประมวลผล (วินาที)			
				ครั้งที่ 1	ครั้งที่ 2	ครั้งที่ 3	เฉลี่ย
1.	1, 2	33,445	30,963	4.37	4.25	4.26	**4.29**
2.	1, 2, 3	36,788	33,563	5.42	5.42	5.30	**5.38**
3.	1, 2, 3, 4	44,102	40,022	7.36	6.52	6.43	**6.77**
4.	1, 2, 3, 1, 2, 3	73,576	33,563	9.92	9.43	10.02	**9.79**
5.	1, 2, 3, 1, 2, 3, 1, 2	107,021	33,563	15.98	15.68	15.89	**15.85**
6.	1, 2, 3, 1, 2, 3, 1, 2, 3, 4	117,678	40,022	35.04	36.04	35.68	**35.58**

จากตารางที่ 4 พบว่า การทดสอบลำดับที่ 4 และ 5 เมื่อทำการเพิ่มชุดข้อมูลเดิม ได้แก่ ชุดที่ 1, 2 และ 3 สังเกตได้ว่า จำนวนเอเอสวีทั้งหมดเพิ่มขึ้น แต่จำนวนเอเอสวีที่ไม่ซ้ำยังคงเท่าเดิมที่จำนวน 33,563 สาย เท่ากับการทดสอบลำดับที่ 2 เป็นการพิสูจน์ว่าการกำจัดความซ้ำซ้อนถูกต้อง 100% โดยสามารถลบเอเอสวีที่ซ้ำกันออกไปได้แม้ว่าจะนำเข้าข้อมูลชุดเดิมซ้ำกี่ครั้งก็ตาม ในทางกลับกันเมื่อเพิ่มข้อมูลใหม่ชุดที่ 4 เข้าไปในการทดสอบในลำดับที่ 6 จำนวนเอเอสวีที่ไม่ซ้ำได้เพิ่มขึ้นเป็น 40,022 สาย เท่ากับการทดสอบลำดับที่ 3 ซึ่งแสดงให้เห็นว่าโปรแกรมสามารถแยกแยะระหว่างข้อมูลซ้ำหรือที่ต้องลบ และข้อมูลใหม่หรือที่ต้องเก็บได้ถูกต้อง สำหรับผลการทดสอบความเร็วเมื่อเพิ่มชุดข้อมูล พบว่า เวลาในการประมวลผลเฉลี่ยของโปรแกรมไม่ช้าลงแบบก้าวกระโดด โดยโปรแกรมสามารถประมวลผลข้อมูลจำนวน 10 ชุด ซึ่งมีจำนวนเอเอสวีทั้งหมดมากกว่า 117,678 สาย ได้ภายในเวลาเฉลี่ย 35.58 วินาที

อภิปรายผลการวิจัย

บทความวิจัยนี้ได้นำเสนอการพัฒนาโปรแกรมเพื่อการรวมและกำจัดเอเอสวีที่ซ้ำซ้อน จากการประเมินความพึงพอใจโดยผู้ใช้ โปรแกรมได้รับความพึงพอใจเฉลี่ย 4.29 ระดับดี เนื่องจากโปรแกรมถูกพัฒนาด้วยภาษาไพทอน ไม่สามารถใช้งานได้ทันที ต้องติดตั้งไลบรารีและสภาพแวดล้อมให้เหมาะสมก่อน อย่างไรก็ตามเมื่อผู้ใช้ติดตั้งโปรแกรมนี้สำเร็จ ผู้ใช้ไม่จำเป็นต้องมีทักษะการเขียนโปรแกรม ไม่ต้องใช้โปรแกรมผ่านส่วนติดต่อผู้ใช้แบบคอมมานด์ไลน์ทั้งหมด สอดคล้องกับทีมพัฒนาของไคม์ทู [25] ได้พัฒนาเครื่องมือเรียกว่าไคม์ทูอาร์ติแฟกต์เอพีไอ เป็นอินเทอร์เฟซภาษาไพทอน ถูกออกแบบมาสำหรับนักวิจัย ชีวสารสนเทศ หรือผู้เชี่ยวชาญด้านไมโครไบโอม ที่ต้องการทำการวิเคราะห์ข้อมูลอย่างเป็นระบบด้วยจูพิเทอะโน้ตบุ๊ค โดยไม่จำเป็นต้องใช้งานแบบคอมมานด์ไลน์ทั้งหมด นอกจากนั้น Lu และคณะ [24] ได้พัฒนาได้พัฒนาไมโครไบโอมอะนาลิสต์สองจุดศูนย์ โปรแกรมครบวงจรรองรับทั้งการวิเคราะห์ความหลากหลายของชุมชนจุลชีพ เครื่องมือนี้ช่วยให้นักวิจัยสามารถทำการวิเคราะห์ข้อมูลที่ซับซ้อนโดยไม่ต้องเขียนโปรแกรม

ผลการประเมินผลด้านความถูกต้อง โปรแกรมสามารถการกำจัดความซ้ำซ้อนของเอเอสวีระหว่างชุดข้อมูลได้ถูกต้อง 100% ซึ่งเกิดจากการเลือกใช้วิธีการจับคู่แบบตรงกันทุกตำแหน่ง ทั้งนี้เพื่อรักษาความละเอียดของข้อมูลและรักษาจำนวนเอเอสวีไม่ให้สูญหาย ตามแนวคิดของ Wang และคณะ [22] อธิบายว่าข้อมูลเอเอสวีแม้แต่เพียงตำแหน่งเดียว (1 เบส) จะถือว่าเป็นเอเอสวีคนละตัว (เอเอสวี 1 ตัว คือตัวแทนของจุลินทรีย์ 1 สายพันธุ์) ทำให้เอ

เอสวีมีข้อดีคือ สามารถเปรียบเทียบข้ามงานวิจัยได้ อย่างไรก็ตาม ข้อเสียของวิธีการนี้คือ เมื่อเปรียบเทียบระหว่างงานวิจัย อาจจะไม่ได้ผลข้อมูลใด ๆ เลย เนื่องจากวิธีการนี้ไม่สามารถจัดกลุ่มเอเอสวีที่มีความคล้ายกันได้

ผลการประเมินผลด้านความเร็ว เมื่อทดสอบโปรแกรมด้วยวิธีการเพิ่มชุดข้อมูลจำนวน 10 ชุด จำนวนเอเอสวีมากกว่า 117,678 สาย โปรแกรมสามารถประมวลผลภายในเวลาเฉลี่ย 35.58 วินาที แสดงให้เห็นว่าการประมวลผลไม่ช้าลงแบบก้าวกระโดด และเวลาในการประมวลผลเฉลี่ยเพิ่มขึ้นในลักษณะเกือบเป็นเชิงเส้นเมื่อเทียบกับเอเอสวีทั้งหมดที่นำเข้า อย่างไรก็ตามเมื่อชุดข้อมูลที่มากขึ้นจนกระทั่งข้อมูลมีขนาดใหญ่มาก (Big data) ความเร็วในการประมวลผลของโปรแกรมจะขึ้นอยู่กับทรัพยากรคอมพิวเตอร์และหน่วยความจำที่ใช้ ซึ่งอาจต้องพิจารณาใช้เทคโนโลยีอื่นในอนาคต เช่น ระบบคอมพิวเตอร์สมรรถนะสูงที่ใช้คอมพิวเตอร์หลายเครื่องทำงานร่วมกันเพื่อประมวลผลข้อมูล [13]

ข้อเสนอแนะ

จากผลการวิจัย นำไปสู่ข้อเสนอแนะทั่วไปและข้อเสนอแนะในการวิจัยครั้งต่อไป ดังนี้

1. ข้อเสนอแนะทั่วไป

1) จัดทำคู่มือการใช้งานและเอกสารประกอบการติดตั้งอย่างละเอียด รองรับผู้ใช้งานที่ไม่เชี่ยวชาญภาษาไพทอน

2) พัฒนาโปรแกรมให้อยู่ในรูปแบบเว็บแอปพลิเคชัน

3) เพิ่มฟังก์ชันวิเคราะห์ปลายทางภายในโปรแกรม เช่น การจำแนกประเภททางอนุกรมวิธานหรือการจำแนกสายพันธุ์จุลินทรีย์แบบวีเสิร์ช (VSEARCH)

2. ข้อเสนอแนะในการวิจัยในครั้งต่อไป

เป็นการรวมรวมชุดข้อมูลเอเอสวีจากบทความวิจัยเกี่ยวสัตว์น้ำเศรฐกิจมากกว่า 20 บทความ เพื่อรวมและกำจัดความซ้ำซ้อนด้วยโปรแกรมนี้ จากนั้นนำข้อมูลไปจำแนกสายพันธุ์จุลินทรีย์และนำผลสรุปทั้งหมดไปวิเคราะห์และออกแบบฐานข้อมูลเชิงสัมพันธ์ สร้างรายงานสรุปเป็นกราฟฟิกบนแดชบอร์ด (Dashboard) สำหรับนักวิจัยด้านชีววิทยาและชีวสารสนเทศ

เอกสารอ้างอิง

[1] Lee, J.-Y. (2023). *The principles and applications of high-throughput sequencing technologies*. *Development & Reproduction, 27*(1), 9–24. https://doi.org/10.12717/DR.2023.27.1.9

[2] Liu, Y.-X., Qin, Y., Chen, T., Lu, M., Qian, X., Guo, X., & Bai, Y. (2021). *A practical guide to amplicon and metagenomic analysis of microbiome data*. *Protein & Cell, 12*(5), 315–330. https://doi.org/10.1007/s13238-020-00724-8

[3] Dacey, D. P., & Chain, F. J. J. (2021). *Concatenation of paired-end reads improves taxonomic classification of amplicons for profiling microbial communities*. *BMC Bioinformatics, 22*, Article 493. https://doi.org/10.1186/s12859-021-04410-2

[4] Fasolo, A., Deb, S., Stevanato, P., Concheri, G., & Squartini, A. (2024). *ASV vs OTUs clustering: Effects on alpha, beta, and gamma diversities in microbiome metabarcoding studies*. *PLOS ONE, 19*(10). https://doi.org/10.1371/journal.pone.0309065

[5] Chiarello, M., McCauley, M., Villéger, S., & Jackson, C. R. (2022). *Ranking the biases: The choice of OTUs vs. ASVs in 16S rRNA amplicon data analysis has stronger effects on diversity measures than rarefaction and OTU identity threshold*. *PLOS ONE, 17*(2), e0264443. https://doi.org/10.1371/journal.pone.0264443

[6] Lin, Q., Dorsett, Y., Mirza, A., Tremlett, H., Piccio, L., Longbrake, E. E., Ni Choileain, S., Hafler, D. A., Cox, L. M., Weiner, H. L., Yamamura, T., Chen, K., Wu, Y., & Zhou, Y. (2024). *Meta-analysis identifies common gut microbiota associated with multiple sclerosis*. *Genome Medicine, 16*(1), Article 94. https://doi.org/10.1186/s13073-024-01364-x

[7] Muller, E., Algavi, Y. M., & Borenstein, E. (2021). *A meta-analysis study of the robustness and universality of gut microbiome–metabolome associations*. *Microbiome, 9*(1), Article 203. https://doi.org/10.1186/s40168-021-01149-z

[8] Estaki, M., Jiang, L., Bokulich, N. A., McDonald, D., González, A., Kosciolek, T., Martino, C., Zhu, Q., Birmingham, A., Vázquez-Baeza, Y., Dillon, M. R., Bolyen, E., Caporaso, J. G., & Knight, R. (2020). *QIIME 2 enables comprehensive end-to-end analysis of diverse microbiome data and comparative studies with publicly available data*. *Current Protocols in Bioinformatics, 70*(1), Article e100. https://doi.org/10.1002/cpbi.100

[9] Xiao, L., Zhang, F., & Zhao, F. (2022). *Large-scale microbiome data integration enables robust biomarker identification*. *Nature Computational Science, 2*(5), 307–316. https://doi.org/10.1038/s43588-022-00247-8

[10] อรยา ปรีชาพานิช. (2557). *คู่มือเรียน การวิเคราะห์และออกแบบระบบ (System Analysis and Design) ฉบับสมบูรณ์*. ไอดีซีฯ.

[11] Python Software Foundation. (n.d.). *What is Python? Executive summary*. Retrieved from https://www.python.org/doc/

[12] Amazon Web Services. (n.d.). *What is Python?* Retrieved from https://aws.amazon.com/th/what-is/python/

[13] Köster, J., & Rahmann, S. (2021). *Sustainable data analysis with Snakemake*. *F1000Research, 10*, 33. https://doi.org/10.12688/f1000research.29032.2

[14] The Python Packaging Authority. (n.d.). *Python Packaging User Guide: Tool recommendations*. Retrieved from https://packaging.python.org/en/latest/guides/tool-recommendations/

[15] Python Package Index. (n.d.). *pandas: Powerful Python data analysis toolkit*. Retrieved from https://pypi.org/project/pandas

[16] Pandas via NumFOCUS. (n.d.). *Package overview: Pandas*. Retrieved from https://pandas.pydata.org/docs/getting_started/overview.html

[17] NumPy Developers. (n.d.). *What is NumPy?* Retrieved from https://numpy.org/devdocs/user/whatisnumpy.html

[18] The Matplotlib Development Team. (n.d.). *Pyplot tutorial*. Retrieved from https://matplotlib.org/stable/tutorials/pyplot.html

[19] International Collaboration of Volunteer Developers. (n.d.). *Python tools for computational molecular biology (Biopython)*. Retrieved from https://biopython.org/

[20] Python Software Foundation. (n.d.). *tkinter: Python interface to Tcl/Tk*. Retrieved from https://docs.python.org/3/library/tkinter.html

[21] Lex, A. (n.d.). *UpSet*. Retrieved from https://upset.app/

[22] Wang, C., Liu, C., Zhang, Y., & Wei, R. (2023). *An independent evaluation in a CRC patient cohort of microbiome 16S rRNA sequence analysis methods: OTU clustering, DADA2, and Deblur*. *Frontiers in Microbiology, 14*, 1178744. https://doi.org/10.3389/fmicb.2023.1178744

[23] Özkurt, E., Fritscher, J., Soranzo, N., Ng, D. Y. K., Davey, R. P., Bahram, M., & Hildebrand, F. (2022). *LotuS2: An ultrafast and highly accurate tool for amplicon sequencing analysis*. *Microbiome, 10*, Article 176. https://doi.org/10.1186/s40168-022-01365-1

[24] Lu, Y., Zhou, G., Ewald, J., Pang, Z., Shiri, T., & Xia, J. (2023). *MicrobiomeAnalyst 2.0: Comprehensive statistical, functional and integrative analysis of microbiome data*. *Nucleic Acids Research, 51*(W1), W310–W318. https://doi.org/10.1093/nar/gkad407

[25] QIIME 2 Development Team. (2024). *Artifact API (Using QIIME 2 with Python)*. Retrieved from https://docs.qiime2.org/2024.10/interfaces/artifact-api/

[26] Chaiyapechara, S., Uengwetwanit, T., Arayamethakorn, S., Bunphimpapha, P., Phromson, M., Jangsutthivorawat, W., Tala, S., Karoonuthaisiri, N., & Rungrassamee, W. (2022). *Understanding the host-microbe-environment interactions: Intestinal microbiota and transcriptomes of black tiger shrimp Penaeus monodon at different salinity levels*. *Aquaculture, 546*, Article 737371. https://doi.org/10.1016/j.aquaculture.2021.737371

[27] Angthong, P., Uengwetwanit, T., Uawisetwathana, U., Koehorst, J. J., Arayamethakorn, S., Schaap, P. J., Martins Dos Santos, V., Phromson, M., Karoonuthaisiri, N., Chaiyapechara, S., & Rungrassamee, W. (2023). *Investigating host-gut microbial relationship in Penaeus monodon upon exposure to Vibrio harveyi*. *Aquaculture, 567*, Article 739252. https://doi.org/10.1016/j.aquaculture.2023.739252

[28] Uengwetwanit, T., Uawisetwathana, U., Arayamethakorn, S., Khudet, J., Chaiyapechara, S., Karoonuthaisiri, N., & Rungrassamee, W. (2020). *Multi-omics analysis to examine microbiota, host gene expression and metabolites in the intestine of black tiger shrimp (Penaeus monodon) with different growth performance*. *PeerJ, 8*, Article e9646. https://doi.org/10.7717/peerj.9646

[29] Angthong, P., Uengwetwanit, T., Arayamethakorn, S., Chaitongsakul, P., Karoonuthaisiri, N., & Rungrassamee, W. (2020). *Bacterial analysis in the early developmental stages of the black tiger shrimp (Penaeus monodon)*. *Scientific Reports*. https://doi.org/10.1038/s41598-020-61559-1

[30] กาญจนา รูปต่ำ. (2565). *การพัฒนาระบบทดสอบออนไลน์สำหรับวิทยาลัยเทคโนโลยีประจวบคีรีขันธ์* (วิทยานิพนธ์ปริญญามหาบัณฑิต). มหาวิทยาลัยธุรกิจบัณฑิตย์. https://libdoc.dpu.ac.th/thesis/Kanjana.Rupt.pdf