

การประยุกต์ใช้โปรแกรม R สำหรับการสอนและการวิจัยทางชีววิทยา

Application of R Programming for Biological Teaching and Research

กาญจน์ คัมทรัพย์*

คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏเพชรบูรณ์

ตำบลสะเดียง อำเภอเมือง จังหวัดเพชรบูรณ์ 67000

Kan Khoomsab*

Faculty of Science and Technology, Phetchabun Rajabhat University,

Sadiang, Muang, Phetchabun 67000

บทคัดย่อ

ศาสตร์ทางด้านการวิเคราะห์ข้อมูลเข้ามามีบทบาทอย่างมากต่อการสอนและการทำงานวิจัย เนื่องจากปัจจุบันนี้เป็นยุคของการสร้างข้อมูลขนาดใหญ่ เครื่องมือที่จะนำมาใช้ในการวิเคราะห์ข้อมูลนั้น ต้องมีประสิทธิภาพ และเป็นระบบที่สามารถวิเคราะห์ข้อมูลได้อย่างถูกต้อง สำหรับนักชีววิทยาเป็นนักวิทยาศาสตร์แขนงหนึ่งที่ต้องมีการรวบรวมข้อมูลสิ่งมีชีวิตจำนวนมาก ในการศึกษาทางความหลากหลายของสิ่งมีชีวิต นิเวศวิทยา หรืออนุกรมวิธาน อีกทั้งงานวิเคราะห์ข้อมูลทางด้านการวางแผนการทดลองเพื่อทดสอบสมมติฐาน ตลอดจนการศึกษาความสัมพันธ์ และการทำนายปรากฏการณ์ที่เกิดขึ้นจากตัวแปรต่าง ๆ ในการศึกษาวิจัย ดังนั้นการใช้โปรแกรม R จึงเป็นทางเลือกที่น่าสนใจ เนื่องจากโปรแกรม R เป็นโอเพนซอร์สที่มีประสิทธิภาพสูงในการวิเคราะห์ข้อมูลในศาสตร์หลายแขนง โปรแกรม R มีแพ็คเกจด้านชีววิทยาจำนวนมากให้เลือกใช้ได้อย่างสะดวก เป็นโปรแกรมที่มีลิขสิทธิ์ แต่ไม่เสียค่าใช้จ่ายในการใช้งาน สามารถดาวน์โหลดมาใช้ได้ทุกที่ทั่วโลก จากเว็บไซต์ <http://www.r-project.org> และเป็นโปรแกรมที่ได้รับการยอมรับจากนักวิชาการทั่วโลกว่าเป็นโปรแกรมประยุกต์ทางสถิติที่มีประสิทธิภาพสูง ดังนั้นครูผู้สอนหรือนักวิจัยจึงสามารถประยุกต์ใช้โปรแกรม R มาใช้เพื่อการสอนและการวิจัย ซึ่งไม่เฉพาะแค่สาขาชีววิทยาเท่านั้นยังรวมไปถึงศาสตร์สาขาอื่น ๆ อีกด้วย

คำสำคัญ : โปรแกรม R; ชีววิทยา; การสอน; การวิจัย

Abstract

Data analytics has played an important role in education and research. Since today is the age of big data generation, a tool used to analyze data must be systemically effective and can accurately analyze data. Biologist is a data scientist who has to gather a lot of information of

organism in order to study the diversity, ecology or taxonomy, analyze experimental design to test the hypothesis, study the relationship and predict the occurrence from variables in the research. Therefore, the use of R may be an interesting choice for biologists since R is open source software which can offer highly effective analysis of data in many fields. R is conveniently available in numbers of biology packages. Though the software is copyrighted, it is free to use. The software can be downloaded from <http://www.r-project.org/> and used worldwide. R is recognized by academics across the world as statistical software with high efficiency. Thus, teachers or researchers can use R for teaching and researching purposes not only in biology but also in other fields.

Keywords: R programming; biology; teaching; research

1. บทนำ

โปรแกรม R เป็นทั้งภาษาคอมพิวเตอร์ (R language) และโปรแกรมสำเร็จรูปที่เริ่มจากการนำภาษา R มาเขียนและพัฒนาเป็นโปรแกรมคอมพิวเตอร์สำหรับประมวลข้อมูลทางสถิติและการสร้างกราฟิกเพื่อแสดงผลทางสถิติ จุดเริ่มต้นของ R คือ โปรแกรม S ซึ่งพัฒนาโดย John Chambers AT & T Bell Laboratories ในนิวเจอร์ซีย์ในปี ค.ศ.1980 (พ.ศ. 2523) และพัฒนาต่อเป็น S-PLUS ซึ่งเป็นโปรแกรมทางการค้าและมีลิขสิทธิ์ ต่อมาในปี ค.ศ. 1996 (พ.ศ. 2539) Ross Ihaka อาจารย์ภาควิชาสถิติ จากมหาวิทยาลัยโอคแลนด์ ประเทศนิวซีแลนด์ และ Robert Gentleman ผู้เชี่ยวชาญ bioinformatics และ Bioconductor ได้เป็นผู้เริ่มต้นเขียนโปรแกรม R ซึ่งมีลักษณะการทำงานเหมือนกับโปรแกรม S หรือ S-PLUS [1] แต่แตกต่างกันตรงที่โปรแกรม R เป็นโปรแกรมที่ใช้ฟรี (free software) กล่าวคือ เป็นโปรแกรมที่มีลักษณะเป็นโอเพนซอร์ส (open source) หมายถึง โปรแกรมที่สามารถนำไป ใช้งาน ศึกษา แก้ไข และ เผยแพร่ได้อย่างเสรี ปราศจากเงื่อนไขเพิ่มเติม หรือการคิดค่าลิขสิทธิ์ของโปรแกรม โดยสามารถเปิดเผยซอร์สโค้ด (source code) ให้สาธารณะนำไป

พัฒนาต่อยอดได้หรือเขียนโปรแกรมเพิ่มเติมเอง โดยเขียนรูปแบบไวยากรณ์ของภาษา R ซึ่งมีความคล้ายคลึงกับการเขียนภาษา C ดังนั้นจึงสามารถนำภาษา C C⁺⁺ หรือ Fortran มาใช้งานร่วมกันได้ [2] และจากปี ค.ศ. 1997 มาจนถึงปัจจุบัน โปรแกรม R อยู่ภายใต้การดูแลของทีมพัฒนาซึ่งเป็นอาสาสมัครจากทั่วโลกที่มีชื่อว่า “R Development Core Team” ดังนั้น R language จึงกลายเป็นภาษาคอมพิวเตอร์มาตรฐานที่ได้รับการยอมรับกันอย่างแพร่หลายจากนักสถิติทั่วโลก [3-5]

ภาษา R เป็นภาษาคอมพิวเตอร์ที่ถูกสร้างขึ้นภายใต้ GNU project ซึ่งมีการแจกซอร์สโค้ดฟรีภายใต้ลิขสิทธิ์ของ GNU General Public License ซึ่งซอร์สโค้ดนั้นสามารถทำงานได้บนระบบปฏิบัติการเกือบทุกระบบ เช่น ระบบปฏิบัติการ Unix และระบบปฏิบัติการ Windows การใช้งานโปรแกรม R อาจมีข้อจำกัดที่ต้องใช้การทำงานในรูปแบบการเขียนคำสั่ง (command line) ซึ่งจะสร้างความไม่คุ้นเคยกับผู้เคยใช้งานแบบชี้แล้วคลิก (point and click) เหมือนกับโปรแกรม SPSS อย่างไรก็ตาม หากต้องการวิเคราะห์สถิติเบื้องต้น รวมทั้งการทดสอบสมมติฐานในงานวิจัย สามารถใช้แพ็คเกจ (package) ที่มีชื่อว่า

Rcmdr (R commander) ซึ่งสะดวกต่อการใช้งาน [6]
[7] โดยรายละเอียดจะได้กล่าวต่อไป

ปัจจุบันโปรแกรม R มีมูลนิธิ R เพื่อการคำนวณทางสถิติ (R foundation for statistical computing) ตั้งอยู่ที่กรุงเวียนนา ประเทศออสเตรีย ซึ่งทำหน้าที่ในการจัดการประชุมความก้าวหน้าและพัฒนาโปรแกรม R ระดับนานาชาติ สำหรับประเทศไทยนั้น มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ เป็นแหล่งเซิร์ฟเวอร์ (server) เพื่อบริการดาวน์โหลดแพ็คเกจต่าง ๆ ให้ผู้ใช้ในประเทศไทยสามารถเรียกใช้ได้อย่างสะดวกและรวดเร็ว อีกทั้งยังมีการพัฒนาโปรแกรม R โดย ศาสตราจารย์ นพ.วีระศักดิ์ จงสูวิวัฒน์วงศ์ หน่วยระบาดวิทยา คณะแพทยศาสตร์ มหาวิทยาลัยสงขลานครินทร์ ซึ่งได้รับรางวัลนักวิจัยดีเด่นแห่งชาติ ประจำปี พ.ศ. 2552 สาขาวิทยาศาสตร์การแพทย์ จากสมาชิยแห่งชาติ ผู้พัฒนา Package: Epicalc เพื่อใช้ในงานด้านระบาดวิทยา [8] [9] สำหรับศาสตร์อื่น ๆ โปรแกรม R ยังสามารถนำไปประยุกต์ใช้งานด้านการเงิน เศรษฐศาสตร์ เกษษศาสตร์ การแพทย์ ภูมิศาสตร์ การตลาด การพัฒนาผลิตภัณฑ์อาหาร [10] การวิเคราะห์ข้อมูลในงานวิจัยด้านพืช และการวางแผนการทดลองทางการเกษตร [11,12] อาจกล่าวได้ว่าโปรแกรม R สามารถนำไปใช้ในทุกสาขาวิชาที่มีการคำนวณหรือการวิเคราะห์ข้อมูล ในขณะที่งานทางด้านชีววิทยาที่โปรแกรม R สามารถวิเคราะห์ได้ คือ ทาง morphometric [13] และงานทางด้านชีวสารสนเทศศาสตร์ (bioinformatics) โดยเฉพาะอย่างยิ่งการวิเคราะห์ DNA sequencing และสร้าง phylogenetic tree [14,15] การทำงานของโปรแกรม R ยังมีความโดดเด่นและนิยมใช้คือการทำ data visualization คือ การทำให้ข้อมูลในเชิงปริมาณดูน่าสนใจ เข้าใจง่าย เห็นภาพรวมได้ชัดเจนง่ายต่อการจดจำ และนิยมนำมาใช้ประกอบในการ

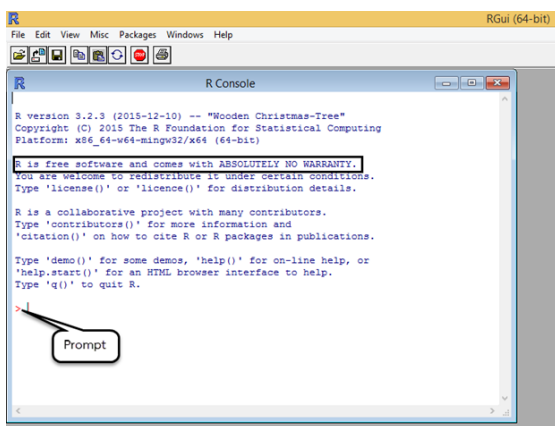
สร้างแผนภูมิและกราฟที่สวยงามสำหรับการนำเสนอข้อมูลหรือใช้เป็นภาพประกอบการเขียนบทความวิจัย [16]

2. แนะนำโปรแกรม R เบื้องต้น

เมื่อเปิดโปรแกรม R ขึ้นมาจะพบหน้าต่างเมนูหลักและหน้าต่าง R console ซึ่งข้อความภายใน R console จะแสดงข้อความการต้อนรับเข้าสู่การใช้งานโปรแกรม R แต่สิ่งที่เราจำเป็นต้องทราบก็คือในย่อหน้าที่ 2 ที่เขียนว่า “R is free software and comes with ABSOLUTELY NO WARRANTY” (รูปที่ 1) หมายความว่า ทางกลุ่มผู้สร้างโปรแกรม R จะไม่รับประกันความผิดพลาดจากระบบของโปรแกรมที่ขัดข้อง การเกิดอาการแฉกของโปรแกรม โปรแกรมหยุดการทำงาน ข้อมูลที่ทําไว้สูญหายหรือได้รับความเสียหาย ซึ่งเหตุการณ์ดังกล่าวเป็นเรื่องปกติของการใช้โปรแกรมคอมพิวเตอร์ทั่ว ๆ ไป ที่สามารถเกิดขึ้นได้ ไม่ใช่เฉพาะโปรแกรม R อย่างเดียว อย่างไรก็ตาม หากเกิดปัญหาระหว่างการใช้งานให้ ปิดโปรแกรมแล้วเปิดใหม่ การบันทึกข้อมูลหรือเตรียมข้อมูลนำเข้าสามารถใช้ Notepad และบันทึกไฟล์เป็นนามสกุล .txt หรือไมโครซอฟท์เอกเซลโดยบันทึกไฟล์เป็นนามสกุล .CSV การฝึกหรือเริ่มทำความคุ้นเคยกับโปรแกรม R เริ่มต้นด้วยการพิมพ์คำสั่งหลังเครื่องหมาย > เรียกว่าเครื่องหมาย prompt (อ่านว่าพร้อมท์) [17] ที่ปรากฏที่หน้า R console (รูปที่ 1)

โปรแกรม R สามารถแสดงผลการคำนวณทางคณิตศาสตร์ได้ โดยใช้สัญลักษณ์ บวก (+) ลบ (-) คูณ (*) หาร (/) ได้ โดยการพิมพ์ตัวเลขเพื่อคำนวณใน R console [18] และเพื่อให้หน้า R console ดูสะอาดตา สามารถลบข้อมูลเก่าออกให้กดปุ่ม ctrl ค้างไว้แล้วกดปุ่ม L ในกรณีที่ต้องการดูคำสั่งที่ถูกลบไป สามารถใช้ปุ่มลูกศร ขึ้น ลง ที่แป้นพิมพ์เพื่อเรียกดูคำสั่งเก่าที่ได้

เขียนและสำรวจวิเคราะห์ข้อมูลก่อนหน้านี้ เพื่อที่จะไม่ต้องพิมพ์คำสั่งใหม่ [19,20]



รูปที่ 1 หน้าต่างโปรแกรม R และ R console

3. โปรแกรม R กับชีวสารสนเทศศาสตร์

ชีวสารสนเทศศาสตร์ เป็นศาสตร์ที่วัดด้วยการนำข้อมูลทางชีววิทยามาใช้อย่างเป็นระบบ รวมทั้งการจัดเก็บ การพัฒนาโปรแกรมประยุกต์แบบต่าง ๆ สำหรับเปรียบเทียบ วิเคราะห์ คำนวณ ประเมินผลข้อมูลที่ได้จากการสืบค้นผ่านเครือข่ายสารสนเทศต่าง ๆ ทั้งในประเทศและต่างประเทศ โดยยุคของชีวสารสนเทศศาสตร์ได้เริ่มขึ้นมาพร้อม ๆ กับโครงการจีโนมมนุษย์ในปี ค.ศ. 1985 โดย Dr. Charles DeLisi ได้เสนอโครงการ Human Genome Project ซึ่งเป็นการถอดรหัสสารพันธุกรรมดีเอ็นเอของมนุษย์ และเป็นการค้นพบศาสตร์สารสนเทศที่เป็นการก้าวกระโดดครั้งสำคัญของวงการชีววิทยา ดังนั้นในยุคดังกล่าวเครื่องมือในการวิเคราะห์ข้อมูลด้านสารสนเทศจะยังมีข้อจำกัดในการใช้งานซึ่งสามารถแบ่งออกเป็น 3 ประเภท คือ ประเภทแรก ประเภทที่มีราคาแพง และมีใช้เฉพาะในศูนย์หรือสถาบันการศึกษาวิจัยเท่านั้น ประเภทที่สอง ประเภทราคาถูก ไม่เสียค่าใช้จ่าย เช่นซอฟต์แวร์ที่ให้บริการผ่านเว็บไซต์ หรือสามารถดาวน์โหลด

โหลดมาใช้งานกับคอมพิวเตอร์ส่วนบุคคลได้ แต่มักประสบปัญหาคือใช้งานยาก ไม่มีคำอธิบายที่ชัดเจน และประเภทที่สาม คือ ประเภทต้องสมัครเป็นสมาชิกและใช้ระบบผ่านอินเทอร์เน็ต [21] จนกระทั่งมาถึงปัจจุบัน การวิเคราะห์ข้อมูลด้านชีวสารสนเทศศาสตร์กลายเป็นเรื่องง่ายซึ่งมีซอฟต์แวร์ที่มีประสิทธิภาพสูงให้เลือกใช้อย่างมากมาย หนึ่งในนั้น คือ โปรแกรม R ซึ่งสามารถใช้วิเคราะห์ข้อมูลทางชีววิทยาโมเลกุล (molecular biology) กันอย่างแพร่หลาย โดยเฉพาะอย่างยิ่งการเรียนการสอนและการวิจัยทางวิทยาศาสตร์ชีวภาพด้านการวิเคราะห์ข้อมูลทางพันธุกรรมจากการจัดเรียงลำดับนิวคลีโอไทด์ (nucleotide) การทำงานทางชีวโมเลกุลด้านการวิเคราะห์หาอินโครงสร้างของยีน เช่น การควบคุมการทำงานของยีน (regulatory elements) การแปลรหัสเพื่อให้ได้โปรตีน (coding sequence) รวมถึงการจัดจำแนกและชี้เฉพาะสปีชีส์ของสิ่งมีชีวิตโดยเทียบความคล้ายคลึงจากลำดับเบส และสร้าง phylogenetic tree แสดงความสัมพันธ์ของสิ่งมีชีวิต [22] ปัจจุบันโปรแกรม R นิยมใช้วิเคราะห์ข้อมูลทางชีวสารสนเทศศาสตร์ ได้แก่ Microarrays, Classification Methods, DNA Sequences Analysis, Gene Ontology, Annotation, Network Analysis, Cluster Analysis and Trees, Decision Tree, Support Vector Machines, Principle Component Analysis [23-25] ซึ่งในปัจจุบันนี้เราจะได้อินสัฟท์ใหม่ ๆ ในวงการของวิทยาศาสตร์ข้อมูล (data science) และชีวสารสนเทศศาสตร์ คือ big data analytics และ machine learning ซึ่งเป็นการใช้ศาสตร์ทางด้านวิทยาการคอมพิวเตอร์ เทคโนโลยีสารสนเทศ คณิตศาสตร์ และสถิติมาใช้ในการบริหารจัดการข้อมูลจำนวนมาก โดยการจัดเก็บข้อมูลอย่างเป็นระบบและวิเคราะห์ข้อมูลเพื่อใช้ในการตัดสินใจ

ตามวัตถุประสงค์ [26-29] จากศักยภาพของโปรแกรม R ในทางชีวสารสนเทศศาสตร์ จะเห็นได้ว่าโปรแกรม R มีความน่าสนใจที่จะนำมาใช้เป็นเครื่องมือเพื่อการเรียนการสอนและวิเคราะห์ข้อมูลขนาดใหญ่ อีกทั้งยังได้รับการยอมรับจากการวิจัยและการตีพิมพ์จากบทความวิจัยต่าง ๆ ที่แสดงให้เห็นว่าโปรแกรม R เป็นโปรแกรมที่สามารถนำมาประยุกต์ใช้เพื่อการเรียนการสอนและการวิจัยทางชีววิทยาได้อย่างเหมาะสมและมีประสิทธิภาพ

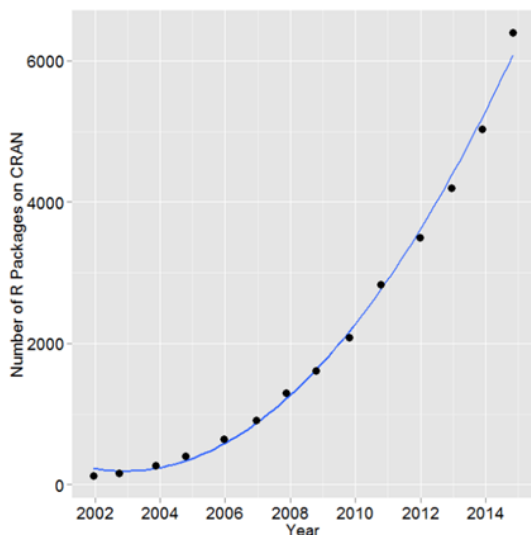
สำหรับผู้สนใจที่จะใช้โปรแกรม R เพื่อวิเคราะห์ข้อมูลทางชีวสารสนเทศ สามารถศึกษาข้อมูลเพิ่มเติมได้ที่ <https://www.bioconductor.org> ซึ่งมีแพ็คเกจต่าง ๆ และเนื้อหาเกี่ยวกับการวิเคราะห์ข้อมูลที่สามารถเลือกนำมาใช้ได้ ตัวอย่าง เช่น งานวิจัยของกุหลาบและคณะ [30] ใช้แพ็คเกจ SIGGENE (<http://www.bioconductor.org>) และ SAM procedure ใน R language software เพื่อตรวจหา SFP และวิเคราะห์ข้อมูลเพื่อค้นหาความแตกต่างในระดับจีโนมของข้าวเจ้าหอมนิลและข้าวเจ้าหอมนิลพันธุ์กลายที่ทนทานต่อสภาวะอุณหภูมิสูงโดยใช้เทคนิค single feature polymorphism (SFP) หรือการค้นหาข้อมูลเกี่ยวกับการใช้โปรแกรม R สามารถเข้าไปที่เว็บไซต์ <http://www.rseek.org> ซึ่งเป็น search engine ตัวเดียวกับ Google [31] ทำให้ค้นหาข้อมูลเกี่ยวกับการใช้งานหรือแหล่งเรียนรู้ของโปรแกรม R ได้อย่างรวดเร็ว อย่างไรก็ตาม การฝึกฝนการใช้โปรแกรม R โดยการทบทวนภาษาและคำสั่งของภาษา R เป็นประจำจะยังสามารถทำให้ผู้ใช้งานวิเคราะห์ข้อมูลต่าง ๆ ได้อย่างคล่องแคล่วและรวดเร็วมากยิ่งขึ้น

4. โปรแกรม R กับแพ็คเกจทางชีววิทยา

คำว่า แพ็คเกจ (package) เป็นคำศัพท์ที่ผู้ใช้โปรแกรม R ค้นเคยและเข้าใจตรงกันว่าเป็นชุดคำสั่ง

หรือโปรแกรมที่เขียนด้วยภาษา R หรือเป็นการออกแบบส่วนของโปรแกรมคอมพิวเตอร์ให้มีการโต้ตอบกับผู้ใช้ (graphical user interface, GUI) ที่ผู้ใช้สามารถเลือกใช้ได้ตามความต้องการเพื่อที่จะนำไปใช้งานให้เหมาะสมกับข้อมูลหรือวัตถุประสงค์การวิเคราะห์ข้อมูลของผู้ใช้งานเอง เริ่มต้นจากเข้าไปยังเว็บไซต์ของ <http://www.r-project.org> จากนั้นเลือกเมนู CRAN จะเข้าสู่หน้าต่าง CRAN mirrors เลื่อนเมาส์ลงมาเพื่อหาเซิร์ฟเวอร์ของประเทศไทย <http://mirrors.psu.ac.th/pub/cran> ซึ่งจะเห็นข้อความ Thailand เป็นที่สังเกต จะเห็นได้ว่าในลิงค์จะมีชื่อเว็บไซต์ psu.ac.th ซึ่งเป็นชื่อเว็บไซต์ของมหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ ซึ่งเป็นผู้บริการจัดเก็บข้อมูลต่าง ๆ ของโปรแกรม R ให้ผู้ใช้ในประเทศไทยได้ดาวน์โหลดได้สะดวกและรวดเร็วมากยิ่งขึ้น ขั้นตอนต่อไปเมื่อคลิกเข้ามาแล้วจะเข้าสู่หน้าต่าง The Comprehensive R Archive Network (CRAN) ให้สังเกตหัวข้อ Source Code for all Platforms ซึ่งเป็นส่วนที่สองต่อจากส่วนแรกซึ่งเป็นการดาวน์โหลดโปรแกรม จากนั้นสังเกตหัวข้อ Contributed Extension Packages แล้วคลิกคำว่า Packages จะเข้าสู่หน้าต่าง Contributed Packages จากนั้นจะปรากฏ Available Packages และจะปรากฏข้อความว่าจะเลือกดูแพ็คเกจตามลำดับของการอัปเดตเพื่อการใช้งานหรือเลือกแพ็คเกจเรียงลำดับตามรายชื่อ โดยปกติผู้เขียนจะเลือกดูตามลำดับรายชื่อซึ่งชื่อของแพ็คเกจจะค่อนข้างสอดคล้องกับงานที่ต้องการวิเคราะห์ข้อมูล ตัวอย่าง เช่น ต้องการวิเคราะห์ข้อมูลเกี่ยวกับความหลากหลายทางชีวภาพ ก็ให้ไปเลือกหาแพ็คเกจตามชื่อภาษาอังกฤษ โดยหาคำว่า Diversity ผู้ใช้อาจทดลองเลือกใช้แพ็คเกจ Diverse เพื่อใช้ในการคำนวณหาดัชนีความหลากหลายของสิ่งมีชีวิตได้ เป็นต้น จากรูปที่ 2 จะพบว่าจำนวนแพ็คเกจของโปรแกรม R มีจำนวน

แพ็คเกจเพิ่มขึ้นอย่างต่อเนื่อง ซึ่งขณะที่ผู้เขียนกำลังเขียนบทความนี้พบว่ามีจำนวน 7,861 แพ็คเกจให้เลือกใช้



รูปที่ 2 จำนวนแพ็คเกจสำหรับโปรแกรม R ที่เพิ่มขึ้นอย่างต่อเนื่อง [32]

สำหรับนักชีววิทยาสามารถใช้งานโปรแกรม R ได้ 2 วิธี คือ การใช้วิธีพิมพ์บรรทัดคำสั่งหรือเลือกใช้แพ็คเกจสำเร็จรูปเพื่อใช้ในการวิเคราะห์ข้อมูล สำหรับบทความนี้จะเสนอตัวอย่างการใช้งานโปรแกรม R ด้วยวิธีการใช้แพ็คเกจทั้ง 2 แบบ จำนวน 3 แพ็คเกจที่น่าสนใจสำหรับงานวิจัยทางชีววิทยา ดังต่อไปนี้

4.1 แพ็คเกจที่ 1 ชื่อแพ็คเกจ Rcmdr

Rcmdr เป็นแพ็คเกจที่ใช้ในการวิเคราะห์ข้อมูลทางสถิติเชิงพรรณนาและสถิติอ้างอิง ประกอบด้วยสถิติขั้นพื้นฐาน ขั้นกลาง และขั้นสูง โดยการเริ่มการติดตั้งแพ็คเกจ ให้เริ่มจากไปที่เมนู Packages จากนั้นเลือก Install package(s)... เลือก (HTTP mirrors) คลิก OK และเลือก Thailand และคลิก OK ค้นหาและเลือกแพ็คเกจ Rcmdr จากนั้นคลิก OK

โปรแกรมจะดำเนินการติดตั้งแพ็คเกจ Rcmdr เมื่อติดตั้งเสร็จเรียบร้อยแล้ว ต้องดำเนินการโหลดโปรแกรมโดยไปที่เมนู Packages จากนั้นเลือก Load package... เลือก Rcmdr คลิก OK จากนั้นหน้าต่างการทำงานของแพ็คเกจ Rcmdr จะปรากฏขึ้น สำหรับงานวิจัยทางชีววิทยาสามารถเลือกใช้ใช้งานได้จากแพ็คเกจ Rcmdr [33,34] ได้ดังต่อไปนี้

4.1.1 สถิติพื้นฐาน ค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน ความถี่ ค่าต่ำสุด ค่าสูงสุด ค่ามัธยฐาน ค่าสัมประสิทธิ์ความแปรปรวน ค่าพิสัยควอไทล์ ความเบ้ ความโด่ง

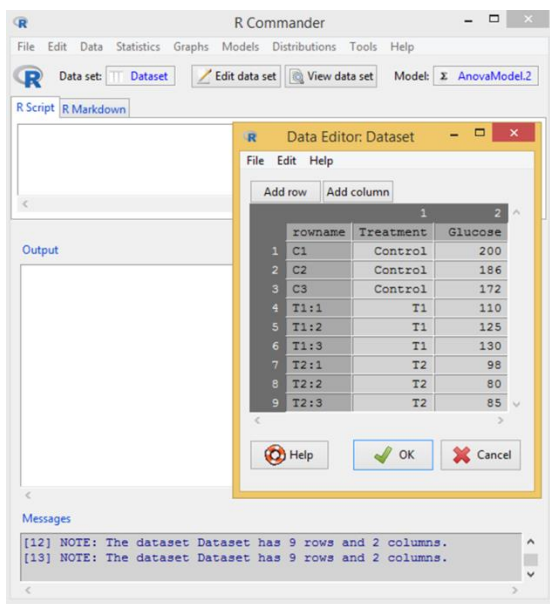
4.1.2 สถิติทดสอบสมมติฐาน การประมาณค่าสัดส่วน การทดสอบความแปรปรวนของข้อมูล Shapiro-Wilk test, Bartlett's test, Levene's test, single sample t-test, Independent sample t-test, Paired sample t-test, One way ANOVA, Multi-way ANOVA, Correlation, Fit model, Regression, Chi-square และสถิติไม่ใช้พารามิเตอร์

4.1.3 สถิติขั้นสูงหรือสถิติหลายตัวแปร ได้แก่ การวิเคราะห์ Scale reliability, Principal components analysis (PCA), Factor analysis, Cluster analysis, k-means cluster analysis และ Hierachicak cluster analysis

4.1.4 การสร้างแผนภูมิ ได้แก่ Index plot, Dot plot, Histogram, Density estimate, Stem and leaf display, Boxplot, Quantile comparison plot, Scatter plot, Line graph, XY conditioning plot, strip chart, Bar graph, Pie chart และ 3D graph

รูปที่ 3 แสดง Interface ของ Rcmdr (R commander) และตัวอย่างการบันทึกข้อมูลลงใน Data Editor โดยแสดงตัวอย่างการวางแผนการทดลองแบบสุ่มตลอด (completely randomized

design) เปรียบเทียบระดับน้ำตาลในเลือดหนูเมื่อทดสอบให้หนูทดลองกินน้ำบอระเพ็ดสกัด ประกอบด้วยกลุ่มควบคุม 1 กลุ่ม (control) และกลุ่มทดลอง 2 กลุ่ม (T1 และ T2) กลุ่มละ 3 ซ้ำ การใช้งานมีลักษณะคล้ายกับโปรแกรม SPSS โดยเลือกที่เมนู statistics จากนั้นเลือก means แล้วเลือกสถิติวิเคราะห์ด้วย one way ANOVA เมื่อวิเคราะห์ข้อมูลแล้วพบความแตกต่างระหว่างค่าเฉลี่ย แพ็คเกจ Rcmdr (R commander) สามารถวิเคราะห์ความแตกต่างระหว่างค่าเฉลี่ย (multiple comparison) ด้วยวิธี Tukey's test โดยการเลือกการวิเคราะห์ที่กล่อง pairwise comparisons of means จะได้ผลการวิเคราะห์ ANOVA และกราฟแสดงผลการวิเคราะห์ความแตกต่างระหว่างค่าเฉลี่ยดังรูปที่ 4



รูปที่ 3 การกรอกข้อมูลใน Data Editor: Dataset ในแพ็คเกจ Rcmdr (R commander)

4.2 แพ็คเกจที่ 2 ชื่อแพ็คเกจ vegan

vegan เป็นแพ็คเกจที่ใช้ในการวิเคราะห์ข้อมูลทางด้านนิเวศวิทยา (community ecology

package) เพื่อใช้วิเคราะห์ดัชนีความหลากหลายทางชีวภาพ (diversity index) การวิเคราะห์สังคมของสิ่งมีชีวิต (community analysis) การจำแนกความเหมือนและความต่างของสังคมสิ่งมีชีวิต (dissimilarity analysis) การวิเคราะห์ข้อมูลหลายตัวแปร (multivariate analysis) การจัดกลุ่ม (cluster) และการจัดอันดับ (ordination) รวมถึงการสร้างกราฟแสดงผล [35, 36] ขั้นตอนแรกให้ไปดาวโหลดแพ็คเกจ vegan ที่ CRAN หรือ <http://mirrors.psu.ac.th/pub/cran> ค้นหาแพ็คเกจ vegan แล้วคลิกเลือก สังเกตบรรทัดข้อความ Package source: vegan_2.3-3.tar.gz คลิกโหลดแพ็คเกจลงเครื่อง จากนั้นเปิดโปรแกรม R แล้วให้ไปที่เมนู Packages เลือก Install packages(s)... จากนั้นคลิกเลือก (HTTP mirrors) แล้วคลิก OK จะเข้าสู่หน้าต่าง HTTP CRAN mirror จากนั้นเลือก Thailand แล้วคลิก OK จะเข้าสู่หน้าต่าง Packages แล้วค้นหาแพ็คเกจ ชื่อ vegan แล้วคลิก OK จากนั้นจะเข้าสู่ขั้นตอนที่ 2 คือการดาวน์โหลดแพ็คเกจ โดยเริ่มต้นที่ เมนู Packages เลือก Load package... จะเข้าสู่หน้าต่าง Select one ค้นหาแพ็คเกจ vegan จากนั้นคลิก OK เสร็จสิ้นกระบวนการติดตั้งแพ็คเกจ

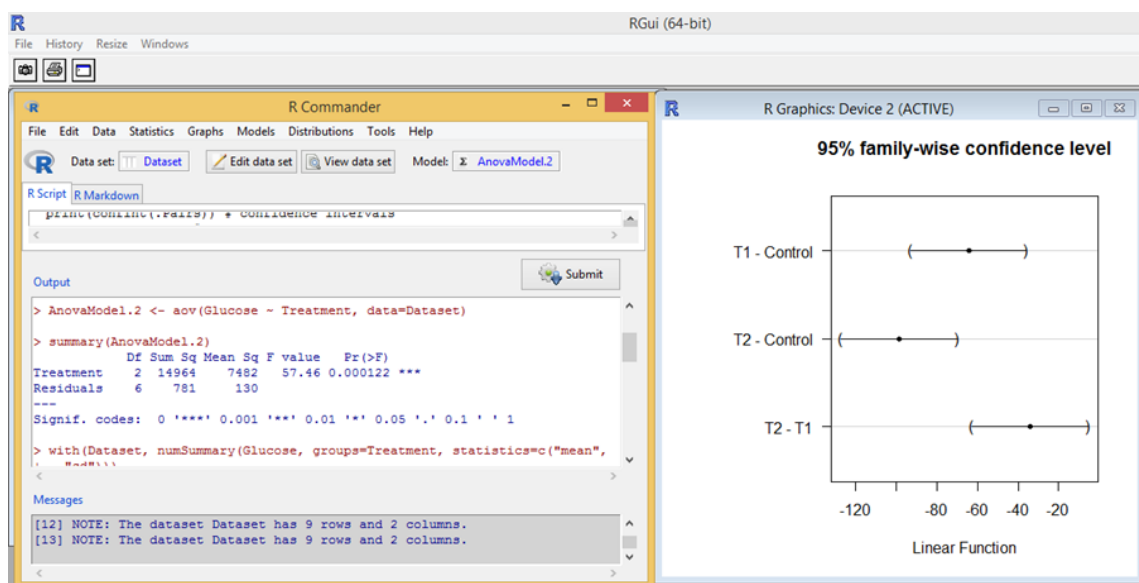
ตัวอย่างการศึกษาวิจัยในแพ็คเกจ vegan โดยการวิเคราะห์ข้อมูลหลายตัวแปร ด้วยวิธี principal component analysis (PCA) วิธีการวิเคราะห์ข้อมูลดำเนินการด้วยวิธีการเขียนบรรทัดคำสั่งดังนี้

```
> library(vegan) # เรียกแพ็คเกจ vegan
ขึ้นมาใช้งาน
```

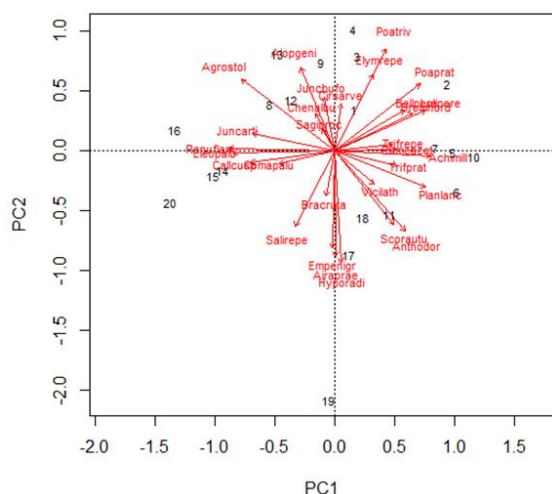
```
> data(dune) # นำเข้าข้อมูล dune เป็น
ข้อมูลตัวอย่างในแพ็คเกจ
```

```
> mod <- rda(dune, scale = TRUE) #
กำหนดให้ mod เป็นตัวแปรที่จะวิเคราะห์ข้อมูล
```

```
> biplot(mod, scaling = "symmetric")
# พล็อตกราฟ PCA
```



รูปที่ 4 ผลการวิเคราะห์ข้อมูลด้วย one way ANOVA



รูปที่ 5 ผลการวิเคราะห์ข้อมูลด้วยวิธี PCA

รูปที่ 5 แสดงผลการวิเคราะห์ PCA ด้วยโปรแกรม R ซึ่งเป็นการจัดลำดับ (ordination) โดยการสร้างแผนที่แสดงตำแหน่งของข้อมูล 2 มิติ โดยใช้กราฟ 2 องค์ประกอบ ซึ่งการวางตำแหน่งของจุดตัวอย่างจะแสดงถึงความเหมือนหรือความต่างของโครงสร้างประชาคมของสิ่งที่ศึกษา โดยแสดงลักษณะ

ของ Biplot ที่แสดงลักษณะของตัวอย่าง ซึ่งจุดที่อยู่ใกล้กันจะหมายถึงการมีโครงสร้างประชาคมคล้ายกันมากกว่าจุดที่ห่างออกไป การวิเคราะห์องค์ประกอบหรือ PCA นี้เป็นวิธีที่นิยมใช้ในการวิเคราะห์ข้อมูลทางนิเวศวิทยาหลายตัวแปรที่น่าสนใจโดยการใช้โปรแกรม R วิเคราะห์ข้อมูล

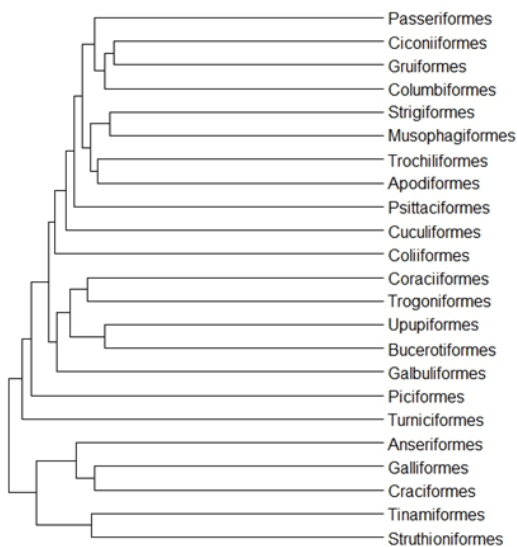
4.3 แพ็คเกจที่ 3 ชื่อแพ็คเกจ ape

ape เป็นแพ็คเกจที่ใช้ในการวิเคราะห์ข้อมูลด้านการสร้างวิวัฒนาการ (analyses of phylogenetics and evolution) การวิเคราะห์ลำดับดีเอ็นเอ คำนวณระยะห่างทางพันธุกรรม การเปรียบเทียบวิเคราะห์ความสัมพันธ์ทางวิวัฒนาการของสิ่งมีชีวิต รวมถึงการสร้างแผนภูมิแสดงผลการวิเคราะห์ข้อมูล [37]

ตัวอย่างการจัดลำดับวิเคราะห์ความสัมพันธ์ของอันดับ (order) ของนก โดยการดาวน์โหลดแพ็คเกจดำเนินการตามขั้นตอนเหมือนกับแพ็คเกจ Rcmdr และ vegan เพียงแต่เลือกการใช้งานแพ็คเกจ ape เข้ามาใช้งาน ผู้ใช้สามารถดาวน์โหลด

คู่มือการใช้งานที่เป็นฉบับ pdf ไฟล์ มาศึกษาเพื่อความเข้าใจในการทำงานของแต่ละแพ็คเกจได้ สำหรับแพ็คเกจ ape มีความเหมาะสมกับงานทางด้านการวิเคราะห์โครงสร้างสายวิวัฒนาการของสิ่งมีชีวิตทุกชนิด ผู้ใช้งานควรอ่านคู่มือการพิมพ์บรรทัดคำสั่งเพื่อฝึกฝนการใช้งาน เนื่องจากแพ็คเกจนี้จะใช้วิธีการพิมพ์บรรทัดคำสั่งในการวิเคราะห์ผลข้อมูล ตัวอย่างต่อไปนี้เป็นตัวอย่างผลการวิเคราะห์ดังรูปที่ 6

```
> data(bird.orders) # นำเข้าข้อมูล
bird.orders เป็นข้อมูลตัวอย่างในแพ็คเกจ
> plot(bird.orders, font = 1) # สร้าง
Phylogenetic tree
```



รูปที่ 6 ผลการวิเคราะห์สายวิวัฒนาการของนกในระดับอันดับด้วยแพ็คเกจ ape

5. การอ้างอิงโปรแกรม R

กรณีผู้ใช้งานต้องการเขียนอ้างอิงการใช้โปรแกรม R ในเอกสารอ้างอิงหรือบรรณานุกรม สามารถเขียนคำสั่งลงใน R console ดังนี้

>citation() แล้วกด Enter จะปรากฏข้อความดังนี้ To cite R in publications use: R Core Team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

ผู้ใช้งานสามารถคัดลอกข้อความที่แสดงด้านบนไปใช้ในเอกสารอ้างอิง หรือบรรณานุกรมได้ หรือปรับเปลี่ยนรูปแบบการเขียนการอ้างอิงตามรูปแบบของวารสารที่กำหนดไว้

6. สรุป

การใช้โปรแกรม R เพื่อการประมวลผลข้อมูลทางชีววิทยาสามารถเป็นทางเลือกหนึ่งสำหรับผู้สนใจด้านการใช้โปรแกรมสำเร็จรูปในการวิเคราะห์ผลข้อมูล หรือเพื่อใช้ในการเรียนการสอนและวิจัยทางชีววิทยา สำหรับบทความนี้ผู้เขียนได้ทบทวนและนำเสนอตัวอย่างแพ็คเกจทางชีววิทยาจำนวน 3 แพ็คเกจ ที่นิยมใช้สำหรับการเรียน การสอนและวิจัยทางชีววิทยา คือ แพ็คเกจ Rcmdr ใช้สำหรับวิเคราะห์ข้อมูลทางสถิติโดยเฉพาะการทดสอบสมมติฐาน แพ็คเกจ vegan ใช้สำหรับการวิเคราะห์ข้อมูลทางนิเวศวิทยาและแพ็คเกจ ape สำหรับวิเคราะห์ข้อมูลด้านการสร้างวงศวานวิวัฒนาการของสิ่งมีชีวิต ซึ่งเป็นแพ็คเกจที่นักชีววิทยาส่วนใหญ่ใช้กันอย่างแพร่หลาย อย่างไรก็ตาม ยังมีแพ็คเกจทางชีววิทยาอีกเป็นจำนวนมากที่สามารถให้ครูผู้สอนและนักวิจัยได้เลือกใช้ให้เหมาะสมกับงานของตนเองได้อย่างมีประสิทธิภาพมากยิ่งขึ้น ข้อดีของโปรแกรม R คือ ฟรี ไม่มีค่าใช้จ่ายในการใช้งาน มีประสิทธิภาพการวิเคราะห์สูงและมีความหลากหลายในการวิเคราะห์ข้อมูลและการสร้างกราฟที่สวยงาม ในขณะที่ข้อเสียของโปรแกรม R คือยังมีข้อจำกัดและความยุ่งยากในการเขียนบรรทัดคำสั่ง หรือการเขียน

ภาษา R ซึ่งเป็นปัญหาสำหรับผู้เริ่มต้นศึกษาและไม่นัดในการเขียนคำสั่งเพื่อประมวลผลข้อมูล อย่างไรก็ตาม ในปัจจุบันนี้ยังมีผู้ใช้โปรแกรม R อยู่ในวงจำกัด แต่ในขณะที่เดียวกันก็มีผู้สนใจทดลองใช้โปรแกรม R เพื่อใช้ในการสอนและการวิจัยเพิ่มมากขึ้น

7. รายการอ้างอิง

- [1] Crawley, M.J., 2015, Statistics an introduction using R, John Wiley & Sons, Ltd., United Kingdom, 339 p.
- [2] Dalgaard, P., 2008, Introductory statistics with R, Springer Science + Business Media, LLC, 363 p.
- [3] Field, A., Miles, J., and Field, Z., 2012, Discovering statistics using R, SAGE Publication Ltd., 957 p.
- [4] AHO, K.A., 2014, Foundational and applied statistics for biologists using R, CRC Press Taylor & Francis group, LLC, 596 p.
- [5] Radziwill, N.M., 2015, Statistics (the easier way) with R, Lapis Lucera, San Francisco, California, 504 p.
- [6] ศิริชัย พงษ์วิชัย, 2553, สถิติเพื่อการวิจัยด้วยโปรแกรม R เล่ม 1 การทดสอบความแตกต่าง, พิมพ์ครั้งที่ 2, บริษัท สุพีเรียพรีนติ้งเฮาส์ จำกัด, กรุงเทพฯ, 348 น.
- [7] Fox, J., 2005, The R commander: A basic-statistics graphical user interface to R, J. Stat. Softw. 14(9): 1-42.
- [8] หัชชา ศรีปลั่ง, 2550, การวิเคราะห์ทางสถิติพื้นฐานด้วย R-ICE, พิมพ์ครั้งที่ 1, ชานเมืองการพิมพ์, กรุงเทพฯ, 240 น.
- [9] ขนิษฐา อัครชัยณรงค์, 2552, แนวทางการเลือกใช้โปรแกรมประยุกต์ทางสถิติ, ว.สำนักบริหารและรับรองห้องปฏิบัติการ 5(14): 20-24.
- [10] กัลยาณี เต็งพงศธร, 2554, การประยุกต์ใช้โปรแกรม R เพื่อการวิเคราะห์ข้อมูลทางสถิติในงานวิจัยด้านพัฒนาผลิตภัณฑ์อาหาร, ว.เกษตรพระจอมเกล้า 29(1): 104-113.
- [11] นราทิพย์ จันสกุล, 2552, แผนแบบการทดลองและการวิเคราะห์ข้อมูลโดยใช้โปรแกรม R, สำนักพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ, 330 น.
- [12] ชูศักดิ์ จอมพุก, 2555, สถิติ : การวางแผนการทดลองและการวิเคราะห์ข้อมูลในงานวิจัยด้านพืชด้วย R, สำนักพิมพ์มหาวิทยาลัยเกษตรศาสตร์, กรุงเทพฯ, 335 น.
- [13] Claude, J., 2008, Morphometrics with R, Springer Science + Business Media, LLC, 317 p.
- [14] Diaconis, P.W. and Holmes, S.P., 1998, Matchings and phylogenetic trees, Proceedings of the National Academy of Sciences USA, 95: 14600-14602.
- [15] Aldous, D.J., 2001, Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today, Stat. Sci. 16: 23-34.
- [16] Sarkar, D., 2008, Lattice multivariate data visualization with R, Springer Science + Business Media, LLC, 258 p.
- [17] Albert, J. and Rizzo, M., 2012, R by example, Springer Science + Business Media, LLC, 359 p.
- [18] Pruim, R., 2011, Computational statistics

- using R and R studio an introduction for scientists, SC 11 Education Program, Available Source: <http://www.unifr.ch/appecon/assets/files/Bachelorcourses/pruim-2011-computational-statistics-using-r-and-r-studio.pdf>, March 9, 2016.
- [19] Schumacker, R.E. and Tomek, S., 2013, Understanding statistics using R, Springer Science + Business Media, New York, 292 p.
- [20] Schumacker, R.E., 2015, Learning statistics using R, SAGE Publications, Inc., USA 623 p.
- [21] วสันต์ จัทราทิตย์ และวีระพงศ์ ลูลิตานนท์, 2544, ชีวสารสนเทศศาสตร์, สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ, กรุงเทพฯ, 194 น.
- [22] ชมกฤษ วิรุณานนท์ และวรุณี จุฬาลักษณ์นกุล, 2553, ชีวสารสนเทศ : การประยุกต์ใช้ในงานวิจัยวิทยาศาสตร์ชีวภาพ, ว.วิทยาศาสตร์บูรพา 15(2): 99-106.
- [23] อภิชาติ ศุภธณี และกิตติพร พลายมาศ, 2553, แนะนำ R Statistical Programing, Thai Bioinformatics E-Magazine, แหล่งที่มา : <http://thaibioinfonetwork.blogspot.com/2010/10/r-programing.html#>, 9 มีนาคม 2559.
- [24] Krijnen, W.P., 2009, Applied statistics for bioinformatics using R, Available Source: <https://cran.r-project.org/doc/contrib/Krijnen-IntroBioInfStatistics.pdf>, March 9, 2016.
- [25] Wheeler, D., 2015, Introduction to R for biologists, Available Source: https://bioinformatics.cancer.gov/sites/default/files/course_material/R-intro.pdf, March 9, 2016.
- [26] Nemade, P. and Kharche, H., 2013, Big data in bioinformatics & the era of cloud computing, IOSR-JCE 14(2): 53-56.
- [27] Shahzad, M. and Ahsan, K., 2014, Comparison of big data analytics tools: A bioinformatics case study, FUUAST J. Biol. 4: 113-118.
- [28] Tarca, A.L., Carey, V.J., Chen, X., Romero, R., and Draghici, S., 2007, Machine learning and its applications to biology, PLoS Comput. Biol. 3: 953-963.
- [29] Kashyap, H., Ahmed, H.A., Hoque, N., Roy, S., and Bhattacharyya, D.K., 2014, Big data analytics in bioinformatics: A machine learning perspective, J. Latex Class Files 13(9):1-20.
- [30] กุหลาบ สีสุขกลาง, สุไลมาน เจ๊ะอาบู, วินิตชาญ รื่นใจชน, อภิชาติ วรณวิจิตร, พูนพิภพ เกษมทรัพย์ และชนเชษฐ์ ม้าลำพอง, 2556, การค้นหาคความแตกต่างในระดับจีโนมของข้าวเจ้าหอมนิลและข้าวเจ้าหอมนิลพันธุ์กลายที่ทนทานต่อสภาวะอุณหภูมิสูงโดยใช้เทคนิค single feature polymorphism (SFP), Thai J. Genet. S(1): 107-110.
- [31] Eglen, S.J., 2009, A quick guide to teaching R programming to computational biology students, PLoS Comput. Biol. 5(8): 1-4.
- [32] Muenchen, R.A., 2016, The popularity of data analysis software, Available Source: <http://r4stats.com/articles/popularity>, March 9, 2016.

- [33] Seefeld, K. and Linder, E., 2007, Statistics using R with biological examples, University of New Hampshire, Durham, NH Department of Mathematics & Statistics, 324 p.
- [34] Shahbaba, B., 2012, Biostatistics with R an introduction to statistics through biological data, Springer Science + Business Media, LLC, 352 p.
- [35] Borcard, D., Gillet, F., and Legendre, P., 2011, Numerical ecology with R, Springer Science + Business Media, LLC, 306 p.
- [36] Oksanen, J., 2016, Package 'vegan', Available Source: <https://cran.r-project.org/web/packages/vegan/vegan.pdf>, March 9, 2016.
- [37] Paradis, E., 2015, Package 'ape', Available Source: <http://ape-package.ird.fr>, March 9, 2016.