ขั้นตอนง่าย ๆ ในการสร้างไฟโลจีเนติกทรี

สมชาย แสงอำนาจเดช

Simple Steps in Reconstruction of Phylogenetic Trees

Somchai Saengamnatdej

ภาควิชาจุลชีววิทยาและปรสิตวิทยา คณะวิทยาศาสตร์การแพทย์ มหาวิทยาลัยนเรศวร อ.เมือง จ.พิษณุโลก 65000
Department of Microbiology and Parasitology, Faculty of Medical Science, Naresuan University, Phitsanulok 65000, Thailand.
Corresponding author. E-mail address: somchais@nu.ac.th (S. Saengamnatdej)

Received 24 July 2007; accepted 26 February 2008

บทสรุป

การวิเคราะห์ไฟโลจีเนติกทรี นอกจากจะใช้เพื่อศึกษาวิวัฒนาการและความหลากหลายของสิ่งมีชีวิตแล้ว ยังใช้เป็นเครื่องมือที่มีประโยชน์มากในด้านอื่น ๆ เช่น ระบาดวิทยา การควบคุมโรค การออกแบบวัคซีน ยา และเวคเตอร์ การค้นหาสปีชีส์ทดแทนที่มียีนที่ต้องการ การค้นหาสารออกฤทธิ์ เป็นต้น บทความปริทัศน์นี้เสนอขั้นตอนง่าย ๆ ที่จะนำไปสู่ความเข้าใจในการสร้าง การอ่าน การทดสอบ และการนำเสนอ โดยใช้ชุดโปรแกรมไฟลิพเป็นแนวทาง พร้อมทั้งแนะนำโปรแกรมชีวสารสนเทศที่เกี่ยวข้อง บทความนี้เหมาะสำหรับผู้ต้องการทำความรู้จักกับไฟโลจีเนติกทรีเป็นครั้งแรก

คำสำคัญ: ไฟโลจีเนติก ทรี; ไฟลิพ; การสร้างทรี; ชีวสารสนเทศ

Summary

Not only is analysis of phylogenetic trees used to study the evolution and diversity of organisms, but also it makes subtle contribution to many other fields or applications, such as epidemiology, control of infectious diseases, design of vaccines, drugs and vectors, search of similar natural products from closely related species, and bioprospecting of bioactive compounds, etc. This review presents simple steps guiding to the understanding of the reconstruction, reading, test, and presentation of phylogenetic trees by using some programs in PHYLIP package as guidance. Other relevant bioinformatic programs are also introduced. This article is suitable for the novice readers who encounter phylogenetic trees for the first time.

Keywords: Phylogenetic tree; PHYLIP; Tree reconstruction; Bioinformatics

บทนำ

การสร้างไฟโลจีเนติกทรีนอกจากจะใช้เพื่อศึกษาวิวัฒนาการและความหลากหลายของสิ่งมีชีวิตแล้ว ยังมีประโยชน์ในด้านอื่น เช่น (1) ประโยชน์ทางระบาดวิทยา การควบคุมโรคติดเชื้อและการออกแบบวัคซีน เช่น การวิเคราะห์ไฟโลจีเนติกทรีของแบคทีเรีย enterotoxigenic *Escherichia coli* (ETEC) ซึ่งเป็นสาเหตุของโรคท้องร่วงสายพันธุ์ต่าง ๆ ที่มาจากหลายแหล่งภูมิศาสตร์และชนิดของโฮสต์ ช่วยให้เข้าใจการได้รับยีนของสารพิษที่ก่อความรุนแรงในการเกิดอาการท้องเสียจากแบคทีเรียชนิดนี้ (Turner et al., 2006) หรือการวิเคราะห์ไฟโลจีเนติกทรีของไวรัสโปลิโอ (poliovirus) และไวรัสค็อกซ์สากี (coxsackie virus) ที่จัดรวมอยู่ในเอ็นเทอร์โรไวรัสคลัสเตอร์ C ของคน (HEV-C) ช่วยให้สามารถทำนายความเป็นไปได้ที่ไวรัสโปลิโออาจจะกำเนิดมาจากการกลายพันธุ์ของไวรัสใกล้เคียงที่ก่อโรคไม่รุนแรง ทำให้ได้ข้อมูลที่เป็นประโยชน์ในโครงการรณรงค์ทำลายล้างไวรัสโปลิโอทั่วโลก (Jiang et al., 2007) (2) ใช้เป็นแผนที่นำทางเพื่อช่วยทำนายคุณสมบัติของสปีชีส์ที่ยังไม่มีการศึกษา เพื่อค้นหาสารที่มีฤทธิ์ทางเภสัชวิทยาเพื่อใช้เป็นยา หรือเครื่องมือวิจัยทางสรีรวิทยา เช่น การใช้ไฟโลจีเนติกทรีช่วยทำนายพบว่า ปลามากกว่า 1,200 ชนิด ที่อาจมีพิษและจะนำไปสู่การศึกษาเพื่อใช้ประโยชน์หรือการพบสปีชีส์ใกล้เคียง อาจใช้เป็นแหล่งของสารตั้งต้นทดแทนสารที่ต้องการ เช่น การใช้ใบของ *Taxus baccata* เป็น แหล่งของสารแทนการใช้เปลือกของ *T. brevifolia* ในการผลิตยาต้านมะเร็ง paclitaxel (Smith & Wheeler, 2006) (3) นำไปสู่ความรู้ความหลากหลายในโครงสร้างสำหรับเป็นแหล่งศึกษาโครงสร้างและหน้าที่ และ การออกแบบยารักษา เช่น การวิเคราะห์ไฟโลจีเนติกทรีของยีนของแฟคเตอร์ eIF4E (Joshi et al., 2005) หรือ การวิเคราะห์ทรี พบว่า งูและกิ้งก่ามีระบบพิษที่วิวัฒนาการจากจุดเริ่มเดียวกันที่จะนำไปสู่การวิจัยทางชีวการแพทย์และการออกแบบยา (Fry et al., 2006) (4) เป็นพื้นฐานสำหรับการออกแบบเวคเตอร์อะดีโนไวรัสให้มีความจำเพาะต่อเนื้อเยื่อ (Madisch et al., 2007)

บทความนี้ขยายความและเสริมส่วนที่ผู้เขียนเห็นว่าสำคัญต่อความเข้าใจที่ขาดไปจากบทความที่ตีพิมพ์เผยแพร่ก่อนหน้านี้ (Baldauf, 2003; Harrison & Langdale, 2006) จะช่วยให้ผู้ที่ยังไม่เคยสร้างหรือเข้าใจการอ่านไฟโลจีเนติกทรี และ ผู้ที่อยู่ในสาขาอื่นที่ไม่ได้ศึกษาด้านวิวัฒนาการและความหลากหลายของสิ่งมีชีวิต แต่สนใจจะนำไปใช้เป็นเครื่องมือในสาขาของตน สามารถเข้าใจการสร้างการอ่านทรี รู้จักโปรแกรมที่สำคัญ หลักการ และวิธีการทดสอบความถูกต้องของไฟโลจีเนติกทรีและพื้นฐานที่

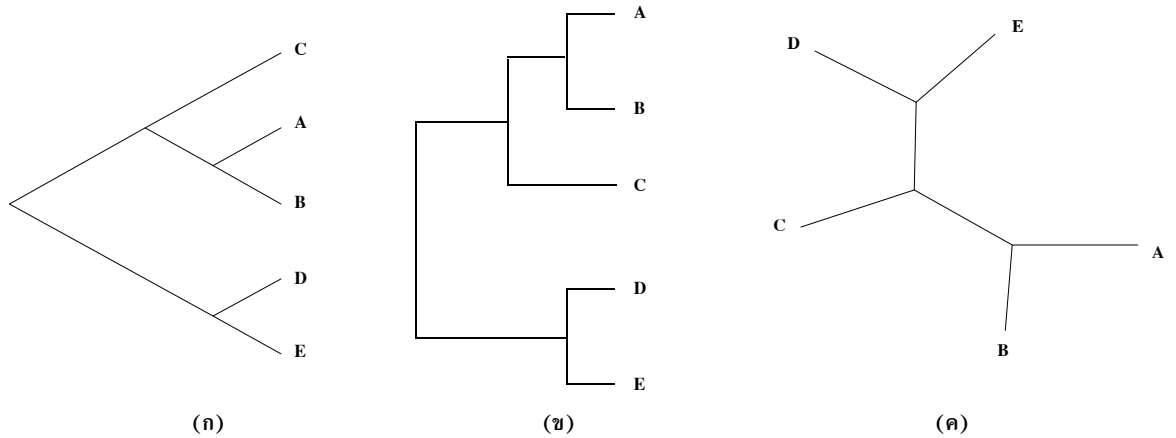
เกียวข้องได้ในระดับหนึง

การสร้างไฟโลจีเนติกทรี

ไฟโลจีเนติกทรีสามารถสร้างโดยอาศัยข้อมูลลักษณะสัณฐานวิทยา ซากดึกดำบรรพ์หรือข้อมูลระดับโมเลกุล โดยเฉพาะลำดับนิวคลีโอไทด์และลำดับโปรตีนของยีนต่าง ๆ ทีเป็นออร์โธล็อก (orthologue) ในบทความนีจะใช้ข้อมูลระดับโมเลกุล ก่อนทีจะกล่าวถึงขันตอนต่าง ๆ ควรจะต้องทำความเข้าใจองค์ประกอบทีสำคัญของไฟโลจีเนติกทรี และคำศัพท์ที่เกียวข้อง

ไฟโลจีเนติกทรี มีองค์ประกอบคล้ายต้นไม้ คือ ประกอบด้วยกิ่งก้านหรือแขนง (branch) ก้านอาจแตกเป็นกิงย่อยแบบสองทาง (bifurcation) หรือหลายทาง (multifurcation) ซึงแบบหลังนีพบน้อยมาก ตำแหน่งทีก้านแตกเป็นกิงย่อยเป็นจุดต่อ เรียกว่า โนด (node) ทีปลายสุดของกิงจะเป็นลำดับเบส หรือโปรตีนสายพันธุ์หรือสปีชีส์ของสิงมีชีวิต เรียกว่า ใบ (leaf) หรือแทกซอน (taxon) หรือหน่วยอนุกรมวิธาน เชิงปฏิบัติการ (OTU; operational taxonomic unit) อาจจะ พิจารณาทรีเป็นกราฟชนิดหนึงซึงมีส่วนยอด (vertex) เป็นโนด และแขนงทีต่อโนดเป็นเส้นขอบ (edge) แต่ทรีอาจมีจุดกำเนิด ร่วมหรือราก (rooted tree) ซึงเป็นส่วนยอดทีเป็นตำแหน่งของบรรพบุรุษร่วมของแทกซา (taxa) ทังหมดหรือไม่มีราก (unrooted tree) ก็ได้ ทรีทีมีรากจะทำให้ทราบทิศทาง เวลา และบรรพบุรุษในการวิวัฒนาการความยาวของก้านทีแตกต่างกัน แสดงระยะห่างของวิวัฒนาการ ซึงแต่ละก้านอาจมีอัตราเร็วของวิวัฒนาการแตกต่างกัน ระยะห่างอาจแสดงเป็นตัวเลขบนแต่ละก้านหรืออาจเขียนเป็นสเกลความยาวบอกขนาดในภาพทรีอาจเขียนในลักษณะเส้นตรงทีแยกออกจากกัน (รูปที 1ก.) หรืออาจเขียนด้วยเส้นแนวตังและแนวนอน (รูปที 1ข.) เพือให้เปรียบเทียบความยาวของก้านได้ง่าย ในกรณีเส้นในแนวตังเป็นเส้นเชือมต่อไม่มีความหมายอืนใด ทรีอาจจะเขียนในลักษณะทีส่วนของใบเสมอกันและโดยความยาวของก้านไม่มีความหมาย ทรีในลักษณะนี เรียกว่า เคลดโดแกรม (cladogram) ถ้าพิจารณาทรีโดยสนใจเพียงโครงสร้างการแตกกิงก้าน ไม่สนใจความยาวของขอบ สามารถจะกล่าวว่า ทรีนันมีโทโพโลยี (topology) เดียวกันหรือเรียก topological trees เมือสามารถบิดหรือยืดขอบต่าง ๆ แล้วทำให้ทรีทีมีลักษณะโครงสร้างของกิงก้านแบบเดียวกัน ทรีอาจเรียกว่าเป็น additive tree ถ้าระยะห่างของวิวัฒนาการของสองแทกซาทีแยกมาจากโนดเดียวกันเท่ากับผลบวกของระยะห่างทังสองจากโนดนัน ถ้าตังสมมุติฐานให้ทรีมีอัตราการกลายพันธุ์ (mutation rate) คงทีสำหรับทุกสายทาง (lineages) สำหรับการวิวัฒนาการของลำดับเบสหรือโปรตีน จะเรียกสมมุติฐานนีว่า สมมุติฐานนาฬิกาของโมเลกุล (molecular clock assumption) เนืองจากอัตราการกลายพันธุ์คงที ดังนัน ปริมาณการกลายพันธุ์ของแต่ละเส้นขอบจะเป็นสัดส่วนกับเวลาทีผ่านไป ดังนัน ไม่ว่าจะใช้ความยาวของเส้นแทนปริมาณการกลายพันธุ์หรือเวลาทีผ่านไปจะได้ตัวเลขเดียวกัน แต่ถ้าไม่ใช้สมมุติฐานนาฬิกาของโมเลกุลทีขอบทีมีระยะเวลาเท่ากัน แต่มีอัตราเร็วของการกลายพันธุ์แตกต่างกัน ขอบทีมีอัตราเร็วการกลายพันธุ์มากกว่าจะมีปริมาณการกลายพันธุ์มากกว่าในช่วงเวลาทีผ่านไปเท่ากันนัน (Allman & Rhodes, 2004) กลุ่มของแทกซาทีกำเนิดมาจากบรรพบุรุษร่วมกันเรียกว่า เคลด (clad) ทรีอาจมีเพียงเคลดเดียว เรียกว่า monophyletic tree หรือ มีหลายเคลดเรียกว่า polyphyletic tree และแต่ละคู่ของสปีชีส์หรือแทกซาทีแยกจากกันและกันหนึงช่วงโนดทีอยู่ภายในทรี (internal node) หรือ กล่าวอีกอย่างก็คือ แทกซาคูใด ๆ ทีเส้นขอบวิงไปบรรจบกัน เรียกว่า เป็นเพือนบ้านกัน (neighbors) สำหรับสปีชีส์หรือแทกซาทีแยกออกจากสปีชีส์อืนทีกำลังศึกษาในระยะแรก ๆ ของการวิวัฒนาการ เรียกว่า พวกนอกกลุ่ม (outgroup) ซึงสปีชีส์เหล่านีสามารถนำมาใช้เพือสร้างทรีแบบมีรากได้ (Deodier et al., 2005; Krane & Raymer, 2003) องค์ประกอบทีสำคัญของไฟโลจีเนติกทรี ดังแสดงในรูปที 1

จำนวนของทรีทังชนิดมีรากและไม่มีราก สามารถคำนวณหาได้เมือทราบจำนวนของแทกซา ซึงจะมีจำนวนมากมาย ดังนัน โปรแกรมสำหรับสร้างไฟโลจีเนติกทรีจึงมุ่งไปสู่การสร้างหรือค้นหาทรีทีเหมาะสมหรือดีทีสุด ซึงบางวิธีจะใช้เวลาและกำลังคอมพิวเตอร์ในการวิเคราะห์มาก การสร้างไฟโลจีเนติกทรีมี 5 ขันตอน คือ การรวบรวมชุดข้อมูล การจัดเรียงลำดับ การพิจารณาเลือกวิธีวิเคราะห์ เลือกโมเดลวิวัฒนาการ และโปรแกรมการทดสอบ และ การนำเสนอ



รูปที 1 องค์ประกอบทีสำคัญของไฟโลจีเนติกทรี ทรีในรูป ก และ ข เป็นแบบชนิดมีราก โดยรูป ข ทรีอยู่ในลักษณะเส้นนอน และเส้นตัง ส่วนรูป ค เป็นแบบไม่มีราก ทังสามทรีนีมีโทโพโลยี เหมือนกัน A–E คือ แทกซา โดย A, B แยกมาจากโนดหนึง และ D, E จากอีกโนดหนึง และแทกซา A, B, C อยู่ในเคลด หนึงส่วนแทกซา D, E อยู่อีกเคลดหนึงต่างเคลดกัน

ขั้นที่ 1 การรวบรวมชุดข้อมูล

การค้นหาและเรียกลำดับจากฐานข้อมูลนิวคลีโอไทด์ทำได้ 2 วิธี เพื่อเพิ่มโอกาสในการได้ลำดับที่คล้ายกับยีนที่สนใจ ควรใช้ทั้งสองวิธี คือ การเรียกข้อมูลโดยใช้คำสำคัญ อาจใช้โปรแกรมเรียกข้อมูล เช่น SฐS (httprl l srs.eซi.ac.uk) หรือ Entrez (httprl l www.ncซi.nlm.nih .gov) จากฐานข้อมูลทั่วไป คือ ว ว BJ (httprl l www.ddซV nig.ac.Vp) EMBL (httprl l www.eซi.ac.ukl ว ataซasesl) FenBank (httprl l www. ncซi. nlm.nih.gov) หรือจากฐานข้อมูลยีโนมจำเพาะ เช่น TIFฐ (httprl l www.tigr.orgl tdซl mdซl mdซcomplete. html) JFI (httpr l l www.Vgi.doe.govl JFIถmicroซiall htmll index.html) Sanger (httprl l www.sanger.ac.ukl ro Vctsl Microซesl) และ 5CBI (httprl l www.ncซi.nlm.nih.govl Fenomesl index.html) และการเรียกข้อมูลโดยการค้นความคล้ายกัน โดยใช้โปรแกรมบลาส (BLำ ST) ซึ่งมีในทุกฐานข้อมูลและเว็บไซต์ยีโนมส่วนใหญ่ เช่น 5CBI BLำ ST (httprl l www.ncซi.nlm.nih.govl BLำ STl)

ขั้นที่ 2 การนำลำดับทั้งหมดมาจัดเรียง

เนื่องจากการจัดเรียงลำดับโดยใช้โปรแกรม เช่น Clustalu หรือ ClustalW (ftprl l ftp. eซi.ac.ukl puซl softwarel clustalw2l) ผลที่ได้อาจเกิดช่องว่าง (gaps) ในตำแหน่งที่ไม่เหมาะสม ดังนั้น สมควรที่จะนำมาปรับแต่งโดยการตรวจสอบด้วยตนเองอีกครั้ง โดยใช้โปรแกรม BioEdit (httprl l www.mซio.ncsu.edul BioEditl ซioedit.html)

ขั้นที่ 3 การเลือกใช้วิธีวิเคราะห์โมเดลและโปรแกรมสำหรับสร้างทรี

3.1 พิจารณาเลือกว่าจะสร้างทรีจากลำดับของดีเอ็นเอหรือลำดับของโปรตีน

ชนิดของกรดอะมิโนมีจำนวนหลากหลายกว่าชนิดของนิวคลีโอไทด์ (20r4) แต่ถ้าลำดับโปรตีนมีความใกล้ชิดกันมากจะมีการเปลี่ยนแปลงมากกว่าที่ระดับของดีเอ็นเอ ดังนั้น ควรใช้ลำดับของดีเอ็นเอ แต่ถ้าลำดับมีความสัมพันธ์ที่ห่างกันมาก ลำดับของกรดอะมิโนจะเก็บข้อมูลมากกว่า อย่างไรก็ดีก็ยังสามารถใช้ลำดับดีเอ็นเอ สำหรับลำดับที่มีความสัมพันธ์ห่างกันได้การตัดเอาโคดอนตำแหน่งที่สามออก เพราะจะรบกวนหรือทำให้ผลแย่ลง (Baldauf, 2003)

3.2 พิจารณาเลือกว่าจะสร้างทรีเป็นแบบมีราก หรือไม่มีราก

ในบางกรณีที่พิจารณาทรีเพื่อตรวจสอบว่ากลุ่มยีนเป็นออร์โธลอกกันหรือไม่ อาจใช้ทรีแบบไม่มีราก แต่ส่วนมากแล้วการสร้างทรีแบบมีราก คือทรีที่แสดงบรรพบุรุษร่วมของทุกแทกซาที่ศึกษาจะเป็นพื้นฐานความเข้าใจกระบวนการวิวัฒนาการ ดังนั้น จึงต้องสร้างทรีแบบมีราก วิธีการทำได้สองวิธี คือ การใส่ลำดับของพวกนอกกลุ่มเข้าไปในการสร้างทรี โดยเลือกหาพวกนอกกลุ่มจากข้อมูลหรือจากลำดับของการจัดเรียงหรือโดยใช้ยีนที่ถูกจำลองมา (duplicated gene) โดยใช้ลำดับจากเคลดหนึ่งไปใช้เป็นฐานของอีกเคลดหนึ่ง

3.3 ตรวจดูการจัดเรียงแล้วตัดสินใจเลือกว่าจะรวมส่วนไหนและตัดส่วนไหนออกจากการวิเคราะห์

โดยพื้นฐาน คือ แต่ละคอลัมน์ของการจัดเรียงหรือแต่ละเรสิดิวซ์จากแต่ละลำดับนั้นเป็นโฮโมโลกัสกัน นั่นคือ วิวัฒนาการจากตำแหน่งเดียวกันบนลำดับโปรตีนของบรรพบุรุษร่วมกันโดยไม่มีการแทรกหรือตัด จำไว้ว่าไฟโลจีเนระดับโมเลกุลจะมีคุณภาพดีได้เท่า ๆ กับคุณภาพการจัดเรียงลำดับที่นำมาใช้สร้าง หลักการทั่วไป คือ ให้ตัดส่วนใดก็ตามที่มีช่องว่าง (gaps) รวมทั้งคอลัมน์ที่ยังคลุมเครือออกไป เพราะไม่สามารถเชื่อมั่นได้เลยว่าบริเวณเหล่านี้จัดเรียงกันอย่างถูกต้อง และถ้าการจัดเรียงไม่ถูกต้องแล้วก็จะไม่ได้ข้อมูลไฟโลจีเนที่ถูกต้อง นอกจากนี้แม้ว่าจะทราบแน่นอนว่าบริเวณที่มีช่องว่างนั้นเป็นการจัดเรียงอย่างถูกต้องบริเวณนั้นก็ยังคงมีผลต่อทรีต่อไป เพราะว่าช่องว่างยิ่งมีขนาดใหญ่ก็จะมีจำนวนตัวอักษรที่รวมเป็นกลุ่มร่วมกันมากขึ้น (เนื่องจากในความเป็นจริงช่องว่างหนึ่ง ๆ เป็นเพียงเหตุการณ์ทางวิวัฒนาการครั้งเดียว โดยไม่ขึ้นกับขนาด) ความสำคัญของช่องว่างช่องหนึ่งจะไม่เป็นสัดส่วนเดียวกับขนาดของช่อง

3.4 เลือกวิธีการสำหรับใช้คำนวณไฟโลจีเนติกทรี

วิธีที่ใช้คำนวณแบ่งเป็นสองกลุ่ม (ำ llman & ฐhodes, 2004; ITC-ำ cademic Computing Health Science, 1WW&)

3.4.1 วิธีที่ใช้แมทริกซ์ของระยะห่าง (Distance-matrix methods)

วิธีนี้จะสรุปความแตกต่างระหว่างลำดับโดยการคำนวณค่าระยะห่างเป็นคู่ ๆ ระหว่างลำดับที่มาจัดเรียงทั้งหมด อาจเรียกว่าวิธีจัดรวมกลุ่ม หรือวิธีเลขคณิต (clustering or algorithmic methods) วิธีนี้จะมีการสร้างตารางแมทริกซ์ ซึ่งบอกระยะห่างหรือความแตกต่างเป็นค่าตัวเลข ระหว่างลำดับเบสหรือโปรตีนต่าง ๆ ที่เปรียบเทียบ โดยเปรียบเทียบลำดับแต่ละคู่ และหาความแตกต่างระหว่างลำดับ จากนั้นจับกลุ่มให้ลำดับวิธีต่อไปนี้จัดอยู่ในกลุ่มนี้

7 FMำ (unweighted pair-group method with arithmetic mean) วิธีนี้จะสร้างทรีชนิดมีราก และใช้สมมุติฐานนาฬิกาโมเลกุล ทรีที่ได้จะวางแทกซาทั้งหมดอยู่ในระยะห่างจากรากเท่ากัน ดังนั้นปริมาณการกลายพันธุ์จากรากจนถึงแต่ละแทกซอนจะเท่ากัน

Fitch-Margoliash วิธีนี้อาศัยพื้นฐานเดียวกับวิธี 7 FMำ แต่จะใช้เทคนิคการรวมแทกซาที่นำมาวิเคราะห์บางแทกซาเข้าเป็นกลุ่มเดียวกันในลักษณะที่จะทำให้เหลือ 3 กลุ่ม ที่จะเปรียบเทียบ แล้วใช้สูตรคำนวณระยะห่างแบบสามจุด (three-point formula) ในแต่ละครั้งเมื่อจับคู่แทกซาใด ๆ ได้แล้ว ก็จะแยกกระจายกลุ่มของแทกซาที่รวมไว้ในครั้งแรกออก แล้วทำการคำนวณใหม่ ทำซ้ำเช่นนี้ต่อไปจนได้ผลสมบูรณ์ วิธีนี้มักจะได้ทรีแบบเดียวกับวิธี 7 FMำ แต่ทรีที่ได้จะเป็นทรีที่ไม่มีราก

5eighซor-Vining วิธีนี้มาจากการพิจารณาทรีไม่มีรากที่ประกอบด้วย 4 แทกซา พบว่า ผลบวกระยะห่างระหว่างคู่ของเพื่อนบ้านทั้งสองคู่จะน้อยกว่าผลบวกระยะห่างระหว่างคู่ที่ไม่ใช่เพื่อนบ้านเสมอ เรียกว่า เป็นเงื่อนไขสี่จุด (4-point condition) และใช้เป็นพื้นฐานในการคำนวณในการทดสอบความถูกต้องของวิธีที่ใช้สร้างทรี ทั้งการจำลองการกลายพันธุ์ของดีเอ็นเอและการใช้แทกซาจริงที่ทราบทรีแน่นอน พบว่า วิธีนี้เป็นวิธีที่ให้ผลดีกว่าวิธี 7 FMำ และ Fitch-Margoliash แม้ว่าสองวิธีหลังจะให้ผลเชื่อถือได้ในบางสถานการณ์ แต่วิธี neighซor-Vining จะใช้ได้ดีกับข้อมูลในช่วงกว้างมากกว่าและถ้าข้อมูลไม่อยู่บนพื้นฐานนาฬิกาโมเลกุล ซึ่งปัจจุบัน

ข้อมูลส่วนมากไม่อยู่บนสมมุติฐานนี้ วิธี neighbor-joining จะดีกว่าเนื่องจากไม่ได้ตั้งบนสมมุติฐานนั้น (ดังนั้น สองวิธีแรกในทางปฏิบัติจึงไม่ค่อยได้ใช้จริง)

ในชุดโปรแกรมไฟลิพ ในขั้นตอนการสร้างทรีด้วยวิธีสร้างแมทริกซ์ระยะห่างจะเริ่มด้วยโปรแกรมสำหรับสร้างแมทริกซ์ (dnadist สำหรับดีเอ็นเอ และ protdist สำหรับ โปรตีน) แล้วนำผลไปสร้างทรีโดยใช้โปรแกรม Fitch (วิธี Fitch-Margolish) หรือ kitsch (วิธี kitsch-joining) หรือ neighbor itself อย่างใดอย่างหนึ่ง (Tuimala, 2006)

3.4.2 วิธีที่ใช้ข้อมูลที่ไม่เกี่ยวข้องกัน เช่น ตัวอักษร (Discrete data methods)

วิธีนี้อาจเรียกว่าวิธีการค้นหาทรี (tree searching methods) โดยใช้ข้อมูลในลักษณะตัวอักษรตลอดทั้งลำดับ วิธีที่จัดอยู่ในกลุ่มนี้ คือ

พาร์สิโมนี (parsimony) วิธีนี้ยึดถือหลักว่าทรีที่เหมาะสม คือ ทรีที่มีการกลายพันธุ์หรือการเปลี่ยนแปลงน้อยที่สุด ดังนั้น จะนับจำนวนการกลายพันธุ์ที่น้อยที่สุดในแต่ละทรีเรียกเป็นคะแนนพาร์สิโมนี (parsimony score) แล้วเลือกทรีที่ค่าคะแนนน้อยสุดเป็นคำตอบ (Crane & Raymer, 2003)

แมกซิมัมไลค์ลิฮูด (maximum likelihood) วิธีนี้เป็นการหาค่าไลค์ลิฮูดของพารามิเตอร์โมเดลที่เป็นพื้นฐานของข้อมูล คือ การนำเอาข้อมูลที่มีอยู่มาใช้ประมาณค่าพารามิเตอร์ โดยใช้วิธีการแทนค่าพารามิเตอร์ต่าง ๆ ลงในโมเดลของความน่าจะเป็น แล้วพิจารณาดูค่าพารามิเตอร์ที่มีค่าไลค์ลิฮูดสูงสุดหรือทำให้ข้อมูลนั้นน่าจะเป็นไปได้ที่สุด (Purcell, 2007) ในโปรแกรมไฟลิพจะใช้ตารางสถิติในการคำนวณค่าไลค์ลิฮูดของการเปลี่ยนแปลงที่จะเกิดขึ้นที่ตำแหน่งหนึ่ง ๆ วิธีนี้จะคำนวณหาความน่าจะเป็นที่ชุดข้อมูลหนึ่ง ๆ จะเหมาะกับทรีที่สร้างมาจากชุดข้อมูลนั้น โดยใช้โมเดลของวิวัฒนาการของลำดับที่กำหนดให้ ดังนั้น ในขั้นแรกจะเปรียบเทียบข้อมูลกับชุดของโมเดลของการวิวัฒนาการของลำดับ จากนั้นจะเลือกโมเดลที่สามารถอธิบายรูปแบบของการเปลี่ยนแปลงลำดับได้ดีที่สุด (แต่ผู้ใช้โปรแกรมอาจจะเลือกโมเดลใดก็ได้ตามต้องการ) จากนั้นจึงนำโมเดลการวิวัฒนาการของลำดับที่ได้นั้นไปใช้ในการวิเคราะห์ค่าไลค์ลิฮูด การวิเคราะห์จะเริ่มต้นจากการสร้างทรีจากชุดข้อมูลนั้น จากนั้นย้ายส่วนของแขนงไปส่วนต่าง ๆ จนทรีที่ได้ค่าคะแนน likelihood (ค่าความน่าจะเป็นของความเหมาะสมกับข้อมูลสูงสุด) ซึ่งค่าคะแนนแปรตามโทโพโลยี และความยาวของแขนงของทรี (Tuimala, 2006)

วิธีของเบส์ (Bayesian method) เป็นวิธีหาค่าไลค์ลิฮูดอีกชนิดหนึ่งที่ได้รับความนิยม มีสมมุติฐานมากขึ้นและใช้อัลกอลิธึมที่แตกต่างกัน (Mau et al., 1999)

ในชุดโปรแกรมไฟลิพ จะมีโปรแกรมหลายชนิดให้เลือกใช้ตามวิธีคำนวณที่ต้องการ (แสดงในตารางที่ 1) เมื่อเปรียบเทียบทั้งสองวิธี วิธีหาระยะห่างมักจะใช้ในการสร้างทรีในขั้นต้น ถ้านำไปใช้เป็นคำตอบทันทีต้องระมัดระวัง ส่วนวิธีพาร์สิโมนีและไลค์ลิฮูดเป็นวิธีที่ดีกว่า เนื่องจากเป็นวิธีที่สำรวจความพันธ์ระหว่างทรีและองค์ประกอบอื่น ๆ อย่างละเอียด แต่โดยทั่วไปแล้วส่วนมากให้ผลเหมือน ๆ กัน ความแตกต่างของทั้งสองกลุ่มที่สำคัญ คือ ความยากง่ายความรวดเร็วและข้อมูลที่ได้รับบางครั้ง นอกจากนี้ไม่ว่าจะด้วยเหตุผลใดก็ตาม การใช้วิธีใดวิธีหนึ่งอาจให้ผลไม่ถูกต้อง ดังนั้น โดยทั่วไปทรีที่ได้ควรจะถูกทดสอบโดยใช้มากกว่า 1 วิธี

3.5 พิจารณาเลือกโมเดลของการวิวัฒนาการ

วิธีคำนวณระยะห่างของลำดับอาจใช้โมเดลการวิวัฒนาการที่แตกต่างกันได้ โดยโมเดลนี้ คือ สูตรทางคณิตศาสตร์ที่จะชดเชยปัญหาการแทนที่ที่เกิดซ้ำ ๆ กันหลายครั้ง (multiple substitutions) และการแทนที่ที่มีน้ำหนักไม่เท่ากันระหว่างการแทนที่แบบทรานสิชัน คือ ระหว่างเบสภายในกลุ่มพัยริมิดีนหรือภายในกลุ่มพิวรีน (transition; Ts) และการแทนที่แบบทรานสเวอร์ชัน คือ ข้ามกลุ่มพัยริมิดีนหรือพิวรีน (transversion; Tv) ชนิดของโมเดลที่สำคัญ โดยเฉพาะที่สามารถเลือกใช้จากโปรแกรมไฟลิพและลักษณะของโมเดลโดยสรุปดังแสดงในตารางที่ 2

โมเดลหลายชนิดพัฒนาขึ้นเพื่อใช้ประเมินความแตกต่างที่แท้จริงระหว่างลำดับ โดยอาศัยการเปรียบเทียบชนิดของเบสในลำดับ ในปัจจุบันโมเดลหลายชนิดจะให้คะแนนแตกต่างกันระหว่างลำดับ โมเดลจะแตกต่างกันตามน้ำหนักที่ให้ในการแทนที่ที่ต่างกัน เช่น จะให้น้ำหนักการเกิดทรานสเวอร์ชันมากกว่าการเกิดทรานสิชัน (อัตราส่วน Tsl Tv มักใช้ที่ค่า 1-2) หรือการปรับค่าแบบแกมมา (gamma corrections) เนื่องจากการนับความแตกต่างธรรมดาระหว่างสองลำดับ อาจจะประเมินการวิวัฒนาการที่เกิดขึ้นจริงต่ำไป โดยเฉพาะที่ตำแหน่งที่มีวิวัฒนาการรวดเร็วมากและเกิดการกลายพันธุ์ซ้ำ ๆ (multiple mutations) คือมีการเปลี่ยนแปลงไปจากสิ่งที่เปลี่ยนแปลงไปแล้ว วิธีสร้างไฟโลจีเนติกทรีมักจะสันนิษฐานว่าทุกตำแหน่งมีวิวัฒนาการในอัตราเดียวกัน ซึ่งในความเป็นจริงไม่ถูกต้อง เช่น โคดอนตำแหน่งที่สาม จะมีวิวัฒนาการเร็วกว่าตำแหน่งที่สอง เนื่องจากตำแหน่งที่สามไม่มีผลเปลี่ยนชนิดกรดอะมิโนเหมือนตำแหน่งที่สอง อัตราที่แตกต่างกันนี้สามารถแก้โดยการใช้การกระจายแกมมา (gamma distribution) โดยรูปร่างการกระจายแบบแกมมาถูกกำหนดโดยพารามิเตอร์ที่ เรียกว่า อัลฟา (α) ซึ่งจะกำหนดสัมประสิทธิ์ของการผันแปร (coefficient of variation; CV) ของอัตราการแทนที่ระหว่างตำแหน่ง ซึ่งค่า CV $0.11\sqrt{\alpha}$ และค่าโดยประมาณของอัลฟา สำหรับยีนที่เก็บรหัสโปรตีนส่วนมาก คือ 0.5 ซึ่งสามารถหาจาก โปรแกรม Tree-Puzzle (Schmidt, et al., 2002) สำหรับลำดับโปรตีนมีโมเดลวิวัฒนาการให้เลือก 5 ชนิด คือ JTT (Jones-Taylor-Thornton), Dayhoff (Dayhoff), mtREV (Blosum), Kimura, และ Categories (Tuimala, 2006)

3.6 เลือกโปรแกรมที่จะใช้

โปรแกรมที่มีความสมบูรณ์มากที่สุดและใช้กันอย่างกว้างขวาง เช่น PAUP, Mega, และ PHYLIP ซึ่งต่างรวมโมเดลและวิธีการคำนวณหลากหลาย โดยโปรแกรม Mega3.1 ใช้ง่ายที่สุด เนื่องจากเป็นลักษณะของวินโดวส์ และให้เลือกรายการจากแถบรายการ ส่วน PHYLIP มีความซับซ้อนมากสุดที่อยู่เว็บไซต์ของโปรแกรมต่าง ๆ ดังแสดงในตารางที่ 3

ตารางที่ 1 วิธีคำนวณของโปรแกรมต่างๆในชุดโปรแกรมไฟลิพ เวอร์ชัน 3.66

โปรแกรม	พื้นฐานวิธีคำนวณ
Dnadist	Distance matrix (สำหรับดีเอ็นเอ)
Protdist	Distance matrix (สำหรับโปรตีน)
Dnapars	Parsimony (สำหรับดีเอ็นเอ) ใช้อัลกอริธึมชนิดฮิวริสติก (heuristic)
Dnapenny	Parsimony (สำหรับดีเอ็นเอ) (branch-and-bound algorithm, วิธีนี้รับรองว่าจะได้ทรีสั้นที่สุด)
Protpars	Parsimony (สำหรับโปรตีน)
Fitch	Fitch-Margoliash (สำหรับสร้าง ทรี จากระยะห่าง)
Neighbor	Neighbor-joining (สำหรับสร้าง ทรี จากระยะห่าง)
Kitsch	สำหรับสร้าง ทรี จากระยะห่าง
Modeltest	Maximum likelihood
MrModeltest	Maximum likelihood
MrBayes	Bayesian

ตารางที่ 2 โมเดลวิวัฒนาการชนิดต่างๆ (Allman & Rhodes, 2004)

โมเดล	คำอธิบายลักษณะโมเดล
Jukes-Cantor distance	เป็นโมเดลที่ใช้พารามิเตอร์ค่าเดียว (α) ในการกำหนดอัตราการกลายพันธุ์ คือ ใช้สมมุติฐานว่า การแทนที่ทั้งหมดสามารถเกิดได้เท่าๆ กัน (และที่สมดุลเบสทุกชนิดมีความถี่เท่ากับ 0.25)
Kimura distance	เป็นโมเดลที่กำหนดอัตราการเปลี่ยนแปลงแตกต่างกัน 2 ชนิด คือ อัตราหนึ่งสำหรับการแทนที่แบบทรานสิชัน (อัตราเบตา) และ อีกอัตราสำหรับ การแทนที่แบบทรานสเวอร์ชัน (อัตราแกมมา) (ที่สมดุลเบสทุกชนิดมีความถี่ 0.25) โมเดลนี้เหมาะกับข้อมูลที่เห็นว่าการแทนที่ทั้งสองแบบเกิดในอัตราที่ไม่เท่ากัน
F84 distance	โมเดลนี้กำหนดพารามิเตอร์ของอัตราเป็นสองอัตรา คือ อัตราสำหรับการแทนที่แบบทรานสิชัน และอัตราสำหรับการแทนที่แบบทรานสเวอร์ชัน (แต่ความถี่ที่สมดุลของแต่ละเบสจะมีค่าแตกต่างกันได้)
LotDet distance	โมเดลนี้ใช้การแปลงค่าด้วยการหาดีเทอร์มิแนนท์ของแมทริกซ์ ตามด้วยการหาค่าล็อกการิธึมฐานธรรมชาติ (Lockhart et al., 1994) โมเดลนี้ใช้เมื่อมีความแตกต่างความถี่ของเบสสูงมากระหว่างลำดับในทรี แต่โมเดลนี้ไม่สามารถใช้กับอักษรที่ยังคลุมเครือ เช่น เบส P, R, หรือ N เป็นต้น ดังนั้น ลำดับจะต้องทราบแน่นอนทั้งหมด (Tuimala, 2006)

ตารางที่ 3 โปรแกรมสำหรับสร้างไฟโลจีเนติกทรี

โปรแกรม	ที่อยู่เว็บไซต์
PHYLIP	http://evolution.genetics.washinton.edu/phylip.html
PAML	http://abacus.gene.ucl.ac.uk/software/paml.html
TREE-PUZZLE	http://www.tree_puzzle.de
Mega3.1 (latest version: 4.0)	http://www.megasoftware.net/
PAUP*	http://paup.csit.fsu.edu/index.html
Phylowin	http://pbil.univ-lyon1.fr/software/phylowin.html
Treeview	http://taxonomy.zoology.gla.ac.uk/rod/treeview.html
Seaview	http://pbil.univ-lyon1.fr/software/seaview.html
ClustalW	http://www.bioinformatics-toolkit.org/Help/Topics/clustalw.html

การสร้างทรีอย่างคร่าว ๆ อาจใช้โปรแกรมคลัสตัล (Clustal) ซึ่งคำนวณทรี โดยวิธีหาระยะห่างวิวัฒนาการ (evolutionary distances) โดยใช้คำสั่งสำคัญในเมนู ขทรีข มี 4 ขั้นตอน คือ 1. ตัดส่วนของการจัดเรียงลำดับที่มีช่องว่างออกโดยเลือกคำสั่ง ขdelete positions with gapsข 2. ถ้าลำดับเหมือนกันน้อยกว่า W5 เปอร์เซ็นต์ ควรเลือกที่จะปรับการวัดระยะห่างสำหรับการเกิดการแทนที่หลายครั้ง (multiple suชstitutions) โดยเลือกคำสั่ง ขcorrect for multiple suชstitutionsข แล้ว 3. ตรวจดูว่าได้เลือกใช้ฟอร์แมทที่ถูกต้องสำหรับผล โดยเลือกที่ ขoutput formatข แล้วเลือก ขlinkข จากนั้นเปลี่ยนที่คำสั่ง ขbootstrap laชel optionsข จาก ขbranchesข ไปเป็น ขnodesข แล้วปิด 4. คำนวณทรี โดยเลือกคำสั่ง ขbootstrap 5J treeข ขั้นตอนสุดท้ายนี้จริง ๆ แล้วประกอบด้วยหลายขั้นตอนย่อย คือ มีโปรแกรมคำนวณระยะห่าง สร้างทรีด้วยวิธี neighชour-Vining จากนั้น ประเมินทรีที่ได้จากการทดสอบสถิติที่ เรียกว่า บูทสแตร็พพิง (รายละเอียดในขั้นที่ 4) จากนั้นจึงดึงข้อมูลทั้งหมดมารวมกันลงในไฟล์ผลลัพธ์อันเดียว ซึ่งจะประกอบด้วยทรีที่มีค่าบุทแสตร็พที่เหมาะสม เมื่อต้องการดูผลจะต้องใช้โปรแกรมสำหรับใช้ดูทรี เช่น ทรีวิว (Treeview) วิธีการสร้างทรีจากโปรแกรมคลัสตัลนี้มักเป็นวิธีที่ดีสำหรับการเริ่มต้น เช่น การจัดเรียงลำดับจำนวนมาก ๆ อย่างรวดเร็ว เพื่อเลือกเอาชุดตัวแทนมาใช้หรือเพื่อตัดสินใจต่อไปว่า ข้อมูลที่ได้มานั้นมีความน่าสนใจเพียงพอที่จะศึกษาต่อในรายละเอียดหรือไม่

ขั้นที่ 4 การทดสอบ

4.1 บูทสแตร็พพิง (Bootstrapping)

เป็นวิธีทดสอบที่ง่ายที่สุด สำหรับตรวจความถูกต้องของไฟโลจีเนติกทรี หลักการสำคัญของการวิเคราะห์อยู่บนพื้นฐานง่าย ๆ คือ ถ้าความเกี่ยวข้องระหว่างลำดับสองลำดับมีมาก ๆ การเปลี่ยนแปลงเบสเพียงบางเบสแบบไม่จำเพาะเจาะจงไม่น่าจะมีผลต่อความเกี่ยวข้องนี้ ตรงกันข้ามถ้าทั้งสองลำดับเกี่ยวข้องกันน้อยมาก การเปลี่ยนแปลงบางเบสจะมีผลอย่างมาก คือ เป็นการทดสอบว่าชุดของข้อมูลทั้งหมดสอดคล้องกับทรีหรือไม่ หรือว่าทรีที่ได้เป็นเพียงทรีหนึ่งในบรรดาทรีอื่น ๆ ทั้งหมดที่ดีเท่าเทียมกันจำนวนมาก โปรแกรมที่ใช้ คือ เซคบูท (Seqชoot)

ขั้นตอนการทดสอบนี้ประกอบด้วย 3 ขั้นตอน คือ การสุ่มตัวอย่างย่อย ๆ ในแต่ละคอลัมน์มาจากคอลัมน์ใด ๆ ของชุดข้อมูล (ตัวอย่างย่อยนี้มีขนาดความยาวเท่าลำดับข้อมูลเดิม) จากนั้นนำมาสร้างทรีต่าง ๆ สำหรับแต่ละตัวอย่างย่อย เช่น จำนวน 1,000 ทรี แล้วคำนวณความถี่ที่ส่วนต่าง ๆ ของทรีนั้นถูกสร้างขึ้นในแต่ละตัวอย่างย่อยที่สุ่มขึ้นมาเหล่านั้น เช่น ถ้าพบการจับกลุ่มของลำดับแบบหนึ่ง ๆ ในทุกทรีของทุกตัวอย่างย่อย ก็หมายความว่า ค่าสนับสนุนบูทสแตร็พ (ชootstrap support) เป็น 100 เปอร์เซ็นต์ แต่ถ้าพบเพียง 2 ใน 3 ของทรีของตัวอย่างย่อย ก็แสดงว่าค่าสนับสนุนบูทสแตร็พเท่ากับ 67 เปอร์เซ็นต์เท่านั้น ค่าจะเขียนไว้ที่โนดของทรีเรียก ความน่าจะเป็นโพสทีเรียร์ (posterior proชaชility)

4.2 แจ็คไนฟ (Jack-knife)

เป็นเปอร์เซ็นต์จำนวนครั้งที่เคลดปรากฏ เมื่อจำนวนเปอร์เซ็นต์ของตัวอักษรถูกเอาออกอย่างสุ่มจากชุดข้อมูลแล้วทำการวิเคราะห์อีกครั้ง

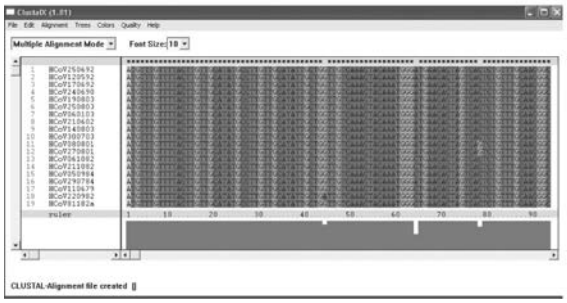
4.3 แขนงยาว (Long branches)

ปัญหาแขนงยาว คือ การที่มีการโน้มเอียงของลำดับที่แยกออกมาก ๆ จะจับกลุ่มกันในทรี โดยไม่เป็นความสัมพันธ์แท้จริง อย่างน้อยส่วนหนึ่งเกิดเนื่องจากลำดับวิวัฒนาการอย่างรวดเร็ว หรือ ลำดับที่ไม่มีสายสัมพันธ์ใกล้ชิดจะมีการกลายพันธุ์จำเพาะมากมาย แต่เนื่องจากความจำกัดที่จำนวนชนิดกรดอะมิโนมี 20 ชนิด หรือนิวคลีโอไทด์มี 4 ชนิด ทำให้การเปลี่ยนแปลงจำนวนมากเริ่มที่จะเกิดความคล้ายกัน ถ้าแขนงมีความยาวมาก ความคล้ายกันเหล่านี้จะกลบลักษณะของไฟโลจีนีที่แท้จริง แล้วลำดับเหล่านั้นจะดึงดูดกันเอง ซึ่งจะเป็นปัญหาหลายอย่าง ปัญหาหนึ่ง คือ ทำให้ค่าบูทสแตร็พของทรีไม่ดี วิธีการทดสอบว่ามีลักษณะของแขนงยาวเกิดขึ้นในชุดข้อมูลหรือไม่ คือ ตัดลำดับเหล่านี้ออกไปจากชุดข้อมูล แล้วหาค่าบูทสแตร็พใหม่ดูว่าค่าจะเพิ่มขึ้นหรือไม่ การใช้วิธีอื่น เช่น แมกซิมัมไลค์ลิฮูดจะถูกผลกระทบจากปัญหานี้น้อยกว่า วิธีที่ดีที่สุดในการแยกส่วนของแขนงยาว ทำโดยการหาลำดับที่อยู่ระหว่างกลางมาใส่รวมในทรี แล้วตรวจดูว่าลำดับนั้นเข้ากันดีกับลำดับที่เหลือหรือไม่

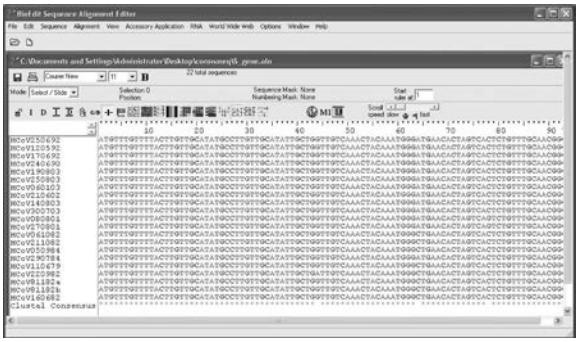
ขั้นที่ 5 การนำเสนอข้อมูล

การนำเสนอไฟโลจีเนติกทรีมีหลักปฏิบัติที่ยอมรับ คือ 1. ความยาวของแขนงจะเขียนตามสเกล (คือ เป็นสัดส่วนกับปริมาณวิวัฒนาการ) ทุกครั้ง แม้ว่าความสัมพันธ์ระหว่างความยาวของแขนงและช่วงเวลาจริงจะห่างไกลกัน ถ้าเขียนตรงไปตรงมาอาจไม่น่าเชื่อถือสำหรับยีน ๆ เดียว แต่ความยาวจะยังคงให้ภาพที่ดีสำหรับแสดงอัตราการเปลี่ยนแปลงสัมพัทธ์ตลอดทรี 2. ให้บอกค่าบูทสแตร็พเป็นเปอร์เซ็นต์ ซึ่งจะทำให้ทรีอ่านและเปรียบเทียบกันได้ง่ายและรายงานผลเพียงถ้าค่าบูทสแตร็พนั้นสูงกว่าหรือเท่ากับ 50 เปอร์เซ็นต์ โดยการแก้ไขที่ไฟล์ของทรี 3. เลือกใช้ชื่อ ฒT7 ที่มีความหมาย

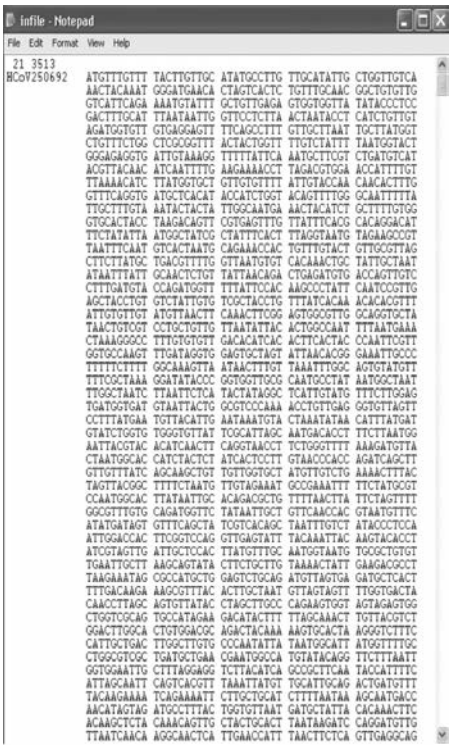
ตัวอย่างการสร้างไฟโลจีเนติกทรีของไวรัสโคโรนา สายพันธุ์ 22WE จากลำดับดีเอ็นเอของยีนสไปค์ (spike) ซึ่งมีความยาว 3513 นิวคลีโอไทด์ จำนวน 21 ลำดับ แสดงในรูปที่ 2 ลำดับนิวคลีโอไทด์ถูกเรียกมาจากฐานข้อมูล FenBank แล้วนำมาจัดเรียงในฟอร์แมทของ fasta จากนั้นใช้โปรแกรม Clustalu ในการจัดเรียงลำดับทั้งหมดดังแสดงในรูปที่ 2ก จากนั้นใช้โปรแกรม Bioedit ในการตรวจดูความเหมาะสมของการจัดเรียง ในกรณีนี้ไม่มีช่องว่างและการจัดเรียงจากโปรแกรมดีแล้วจึงไม่มีการแก้ไข (รูปที่ 2ข) จึงบันทึกไฟล์ด้วยนามสกุล phy เพื่อให้ไฟล์อยู่ในฟอร์แมทไฟลิพ แสดงในรูปที่ 2ค เปลี่ยนชื่อเป็น infile แล้วนำไปวิเคราะห์ด้วยโปรแกรมเซคบูท นำผลลัพธ์ที่ได้มาวิเคราะห์ด้วยวิธีพาร์สิโมนี โดยใช้โปรแกรม ว napars และวิเคราะห์หาทรีที่เป็นตัวแทนทรีทั้งหมดที่ได้ด้วยโปรแกรม Consense ตามลำดับ จากนั้นจึงเปิดดูทรีด้วยโปรแกรมสำหรับใช้ดูภาพและแก้ไขทรี ซึ่งมีมากมาย ในที่นี้ใช้ทรีวิว (age, 1WW6) และ เลือกแสดงทรีแบบไม่มีราก บันทึกไฟล์แบบภาพกราฟิค แล้วใช้โปรแกรมแต่งภาพต่าง ๆ ตกแต่งเพื่อนำเสนอ ในกรณีนี้มีการใช้สีพื้นหลังเพื่อชี้ให้ชัดเจนว่าไวรัสทั้งหมดนี้สามารถจัดจำแนกเป็นสี่กลุ่ม (รูปที่ 2ง) ทรีที่ได้จากการวิเคราะห์วิธีนี้มีความคล้ายคลึงกับทรีที่มีการตีพิมพ์ก่อนหน้านี้ (Chiชo & Birch, 2006) ซึ่งวิเคราะห์โดยใช้โปรแกรม ว nadist ด้วยโมเดล kimura ค่า transitionl transversion ratio เท่ากับ 1.62 ค่า CV เท่ากับ 2.5 (โดยแทนค่าอัลฟาเท่ากับ 0.16)



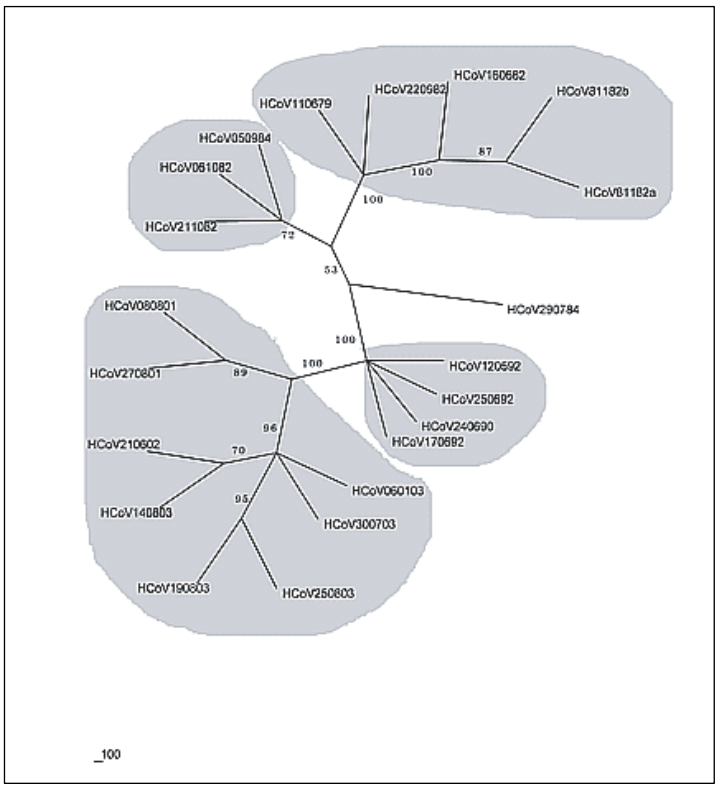
(ก)



(ข)



(ค)



(ง)

รูปที่ 2 ตัวอย่างการสร้างไฟโลจีเนติกทรีของไวรัสโคโรนาสายพันธ์ 229E ไฟโลจีเนติกทรีสร้างจากลำดับดีเอ็นเอของยีนสไปค์ จำนวน 21 ไวรัส ที่แยกได้ในเวลาที่แตกต่างกัน เลขเรียกข้อมูล (accession number) ตั้งแต่ DQ243964-DQ243987 ก.) ผลจากการจัดเรียงด้วยโปรแกรม clustalX ข.) การตรวจการจัดเรียงด้วยโปรแกรม Bioedit ค.) การตรวจดูไฟล์ในฟอร์แมท์ไฟลิพก่อนใช้วิเคราะห์ด้วย PHYLIP3.66 ง.) ไฟโลจีนีติกทรีแบบไม่มีราก ตัวเลขค่าบูทสแตร็พแสดงที่ตำแหน่งโนด สเกลระยะห่างวิวัฒนาการ (evolutionary distance) แสดงด้วยเส้นขีดใต้ภาพ แทกซอนแสดงด้วยชื่อย่อไวรัสพร้อมวันที่ไวรัสถูกแยกได้ เพื่อให้การนำเสนอชัดเจนได้ใช้สีพื้นหลังเพื่อแสดงกลุ่มของไวรัส

บทสรุป

การสร้างไฟโลจีเนติกทรีสามารถนำไปใช้เชิงประยุกต์โดยนำมาวิเคราะห์เพื่อหาข้อมูลที่เป็นประโยชน์ในด้านต่าง ๆ เช่น การควบคุมโรคติดเชื้อ การออกแบบยาวัคซีนและการค้นหาสารมีฤทธิ์ทางชีวภาพ ขั้นตอนในการสร้างที่สำคัญมี 5 ขั้นตอน คือ การรวบรวมลำดับการจัดเรียงลำดับการเลือกใช้วิธีวิเคราะห์โมเดลและโปรแกรมสำหรับสร้างทรี การทดสอบ และ การนำเสนอข้อมูล

เอกสารอ้างอิง

Allman, E. S., & Rhodes, J. A. (2004). *Mathematical models in Biology: An introduction*. New York: Cambridge University Press.

Baldauf, S. L. (2003). Phylogeny for the faint of heart: A tutorial. *Trends in Genetics*, *19*, 345-351.

Chibo, D., & Birch, C. (2006). Analysis of human coronavirus 229E spike and nucleoprotein genes demonstrates genetic drift between chronologically distinct strains. *Journal of General Virology*, *87*, 1203-1208.

Deonier, R. C., Tavaré, S., & Waterman, M. S. (2005). *Computational genome analysis: An introduction.* New York: Springer Science Bussiness Media, Inc.

Fry, B. G., Vidal, N., Norman, J. A., Vonk, F. J., Scheib, H., Ramjan, S. F., et al. (2006). Early evolution of the venom system in lizards and snakes. *Nature*, *439*, 584-588.

Harrison, C. J., & Langdale, J. A. (2006). A step by step guide to phylogeny reconstruction. *The Plant Journal*, *45*, 561-572.

ITC-Academic Computing Health Science. (1998). *Phylip: Phylogenetic analysis workshop*. Virginia: University of Virginia.

Jiang, P., Faase, J. A. J., Toyoda, H., Paul, A., Wimmer, E., & Gorbalenya, A. E. (2007). Evidence for emergence of diverse polioviruses from C-cluster coxackie A viruses and implications for global poliovirus eradication. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 9457-9462.

Joshi, B., Lee, K., Maeder, D. L., & Jagus, R. (2005). Phylogenetic analysis of eIF4E-family members. *BMC Evolutionary Biology*, *5*, 48.

Krane, D. E., & Raymer, M. L. (2003). *Fundamental concepts of bioinformatics*. San Francisco: Benjamin Commings.

Lockhart, P. J., Steel, M. A., Hendy, M. D., & Penny, D. (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. *Molecular Biology and Evolution*, *11*, 605-612.

Madisch, I., Hofmayer, S., Moritz, C., Grintzalis, A., Hainmueller, J., Pring-Akerblom, P., et al. (2007). Phylogenetic analysis and structural predictions of human adenovirus penton proteins as a basis for tissue specific adenoviral vector design. *Journal of Virology*, *81*, 8270-8281.

Mau, B., Newton, M. A., & Larget, B. (1999). Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics*, *55*, 1-12.

Page, R. D. M. (1996). TREEVIEW: An application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences*, *12*, 357-358.

Purcell, S. (2007). BGIM: Maximum likelihood estimation primer. Retrieved July 16, 2007, from http://statgen.iop.kcl.ac.uk/bgim/mle/sslike_1.html

Schmidt, H. A., Strimmer, K., Vongron, M., & von Haeseler, A. (2002). TREE_PUZZLE: Maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics Applications Note*, *18*, 502-504.

Smith, W. L., & Wheeler, W. C. (2006). Venom evolution widespread in fishes: A phylogenetic road map for the bioprospecting of piscine venoms. *Journal of Heredity*, *97*, 206-217.

Tuimala, J. (2006). *A primer to phylogenetic analysis using the PHYLIP package*. Espoo: CSC-Scientific Computing Ltd.

Turner, S. M., Chaudhuri, R. R., Jiang, Z. D., DuPont, H., Gyles, C., Penn, C. W., et al. (2006). Phylogenetic comparisons reveal multiple acquisitions of the toxin genes by enterotoxicgenic *Escherichia coli* strains of different evolutionary lineages. *Journal of Clinical Microbiology*, *44*, 4528-4536.