การจำลองแบบการประกอบของจีโนมไวรัสกึ่งสปีชีส์หลายเส้นด้วยเทคโนโลยีการอ่านลำดับนิวคลีโอไทด์แบบขนานจำนวนมาก

นางสาวพุธิตา สุมานนท์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2551

ASSEMBLING SIMULATION OF MULTIPLE VIRAL QUASISPECIES GENOMES FROM MASSIVELY PARALLEL SEQUENCING TECHNIQUE

Miss Puthita Sumanon

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Engineering Program in Computer Engineering

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2008

หัวข้อวิทยานิพนธ์	การจำลองแบบการประกอบของจีโนมไวรัสกึ่งสปีชีส์หลายเส้นด้วยเทคโนโลยีการอ่านลำดับนิวคลีโอไทด์แบบขนานจำนวนมาก
โดย	นางสาว พุธิตา สุมานนท์
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร.บุญสม เลิศหิรัญวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(ศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา)

.....กรรมการภายนอกมหาวิทยาลัย
(อาจารย์ ดร.ประพัฒน์ สุริยผล)

พุธิตา สุมานนท์ : การจำลองแบบการประกอบของจีโนมไวรัสกึ่งสปีชีส์หลายเส้น ด้วยเทคโนโลยีการอ่านลำดับนิวคลีโอไทด์แบบขนานจำนวนมาก. (ASSEMBLING SIMULATION OF MULTIPLE VIRAL QUASISPECIES GENOMES FROM MASSIVELY PARALLEL SEQUENCING TECHNIQUE) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ศ.ดร.ประภาส จงสถิตย์วัฒนา, 70 หน้า.

วิทยานิพนธ์นี้เสนอวิธีการประกอบชุดของแฮปโพลไทป์และประมาณค่าความถี่แฮปโพลไทป์ของสิ่งมีชีวิตกึ่งสปีชีส์ด้วยข้อมูลที่อ่านได้จากเทคโนโลยีอ่านลำดับนิวคลีโอไทด์แบบขนานจำนวนมาก โดยมุ่งศึกษาแฮปโพลไทป์ของไวรัสเด็งกี่ จากข้อมูลที่ได้จากเครื่องอ่านลำดับเบส Roche GS FLX โดยในงานวิจัยนี้ได้เสนอวิธีประกอบแฮปโพลไทป์สายหลักซึ่งเป็นแฮปโพลไทป์ที่มีความถี่สูงสุด โดยวิธีที่ให้ประสิทธิภาพสูงคือ วิธีประกอบแฮปโพลไทป์ด้วยอัลลีลที่มีความถี่สูงสุดในแต่ละตำแหน่งและวิธีประกอบแฮปโพลไทป์ด้วยสายลำดับที่อ่านได้ที่มีความถี่สูงสุดแบบสุ่มตำแหน่ง แฮปโพลไทป์ที่ประกอบขึ้นมีความแม่นเฉลี่ยร้อยละ 92.07 และ 90.05 ตามลำดับ โดยมีความผิดพลาดสัมบูรณ์ของความถี่แฮปโพลไทป์เป็นร้อยละ 7.19 และ 1.54 ตามลำดับ

วิธีประกอบแฮปโพลไทป์สายหลักนี้ นำไปประยุกต์ใช้กับการประกอบแฮปโพลไทป์ในลำดับถัดไป โดยประกอบแฮปโพลไทป์สายหลักทีละเส้น แล้วกรองข้อมูลที่คาดว่ามาจากแฮปโพลไทป์สายหลักทิ้ง นำข้อมูลที่เหลือมาประกอบแฮปโพลไทป์ลำดับถัดไป ทำซ้ำเช่นนี้จนกว่าจะได้แฮปโพลท์ตามที่กำหนด วิธีประกอบชุดของแฮปโพลไทป์นี้ให้ความแม่นเฉลี่ยร้อยละ 69.79 และมีความแม่นเฉลี่ยสูงสุดในชุดข้อมูลที่ประกอบด้วยสายลำดับ 100,000 เส้น และความถี่ของสายลำดับหลักอยู่ในช่วงร้อยละ 90-99 ชุดของแฮปโพลไทป์ที่ประกอบขึ้นจากชุดข้อมูลนี้มีความแม่นเฉลี่ยร้อยละ 94.99 เปรียบเทียบความแม่นของวิธีที่นำเสนอกับวิธีที่ใช้ในโปรแกรมสำเร็จ ShoRAH พบว่าวิธีที่นำเสนอให้ความแม่นสูงกว่าสำหรับชุดข้อมูลที่จำลองขึ้นจากจีโนมของไวรัสเด็งกี่นี้

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต พุธิตา สุมานนท์

สาขาวิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่ออ.ที่ปรึกษาวิทยานิพนธ์หลัก ประภาส

ปีการศึกษา 2551

5070382621 : MAJOR COMPUTER ENGINEERING

KEYWORDS : Massively Parallel Sequencing Technique / Dengue / Haplotype

PUTHITA SUMANON : ASSEMBLING SIMULATION OF MULTIPLE VIRAL QUASISPECIES GENOMES FROM MASSIVELY PARALLEL SEQUENCING TECHNIQUE. ADVISOR : PROF.PRABHAS CHONGSTITVATANA, Ph.D., 70 pp.

The goal of this study is to reconstruct haplotypes of dengue virus and estimate haplotype frequency of each haplotype from simulated DNA fragments obtained from massively parallel sequencing technique, focused on Roche GS FLX sequencer. Firstly, we try to reconstruct the major haplotype of dengue population and propose two main methods, reconstruction using major alleles and reconstruction using the highest frequency read of each position. These methods provide averages of 92.07% and 90.05% accuracy respectively for major sequences and provide average of 7.19% and 1.54% absolute error for estimated frequencies.

After that, we apply the method of major haplotype reconstruction to reconstruct the whole set of haplotypes. After reconstructing the major haplotype, we discard data obtained from this major haplotype and use the remains as input for reconstructing the next haplotype. This method provides an average of 69.79% accuracy of whole sets of haplotypes and a maximum of 94.99%. When looking at the performance of this proposed method in comparison with that of the method used by software package ShoRAH, the proposed method provides higher accuracy on this test data simulated from dengue genome.

Department : Computer Engineering Student's Signature P. Sumanon

Field of Study : Computer Engineering Advisor's Signature

Academic Year : 2008

กิตติกรรมประกาศ

ขอขอบพระคุณอาจารย์ที่ปรึกษาวิทยานิพนธ์ ศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนาผู้คอยให้คำปรึกษาและเสนอแนะแนวทางอันเป็นประโยชน์ต่องานวิจัยนี้ ขอขอบพระคุณคณะกรรมการสอบวิทยานิพนธ์ ผู้ให้คำแนะนำและชี้จุดบกพร่องที่ควรแก้ไข ทำให้วิทยานิพนธ์นี้เสร็จสมบูรณ์ได้

ขอขอบพระคุณ ดร.ประพัฒน์ สุริยผล หัวหน้าหน่วยชีวสารสนเทศและจัดการข้อมูลวิจัย สถานส่งเสริมการวิจัย คณะแพทยศาสตร์ศิริราชพยาบาล ตลอดจนสมาชิกในหน่วยชีวสารสนเทศ ผู้ให้คำปรึกษา คำแนะนำ ให้ความช่วยเหลือด้วยดีเสมอมา ตลอดจนอนุญาตให้ใช้ข้อมูลจากห้องปฏิบัติการเพื่อการศึกษาวิจัยในวิทยานิพนธ์นี้

ขอขอบคุณสมาชิกในห้องปฏิบัติการวิจัยระบบอัจฉริยะ และพี่ๆ เพื่อนๆ ในระดับบัณฑิตศึกษาทุกคน ที่ช่วยเสนอความคิดเห็น ให้คำปรึกษา และช่วยสร้างบรรยากาศที่ดีในการศึกษาค้นคว้า วิจัยตลอดมา

ขอขอบพระคุณคณาจารย์ทุกท่านในภาควิชาวิศวกรรมคอมพิวเตอร์ที่ช่วยวางรากฐานความรู้อันเป็นประโยชน์ต่อการทำวิจัยและการศึกษาในระดับบัณฑิตศึกษานี้ ตลอดจนให้คำชี้แนะที่เป็นประโยชน์เสมอมา

ขอขอบพระคุณทุน อุดหนุนการศึกษาระดับบัณฑิตศึกษา จุฬาลงกรณ์มหาวิทยาลัย เพื่อเฉลิมฉลองวโรกาสที่พระบาทสมเด็จพระเจ้าอยู่หัวทรงเจริญพระชนมายุครบ 72 พรรษา ที่ได้สนับสนุนค่าใช้จ่ายตลอดระยะเวลาการศึกษา

สุดท้ายนี้ขอกราบขอบพระคุณคุณพ่อ คุณแม่และญาติ ผู้คอยให้กำลังใจ คอยช่วยเหลือ สนับสนุนอย่างเต็มที่เสมอมาและเป็นส่วนสำคัญที่ทำให้ความสำเร็จนี้เกิดขึ้นได้

สารบัญ

สารบัญตาราง

สารบัญภาพ

หน้า

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ไวรัสเด็งกี่เป็นต้นเหตุของโรคไข้เลือดออก โดยมียุงลาย (*Aedes aegypti)* เป็นพาหะ โรคไข้เลือดออกได้กลายเป็นปัญหาสาธารณสุขระดับโลก เนื่องจากโรคได้แพร่กระจายอย่างกว้างขวาง และมีจำนวนผู้ป่วยเพิ่มขึ้นอย่างมากในช่วง 30 ปีที่ผ่านมา จนกลายเป็นโรคประจำถิ่นในประเทศต่างๆ มากกว่า 100 ประเทศ โดยเฉพาะประเทศในเขตร้อนและเขตอบอุ่น องค์การอนามัยโลก (WHO) ประมาณการว่าประชากรโลกมากกว่าร้อยละ 40 (2,500 ล้านคน) เสี่ยงต่อการติดเชื้อไวรัสเด็งกี่ [1]

โรคไข้เลือดออกจำแนกตามกลุ่มอาการได้ 3 กลุ่ม คือ ไข้เด็งกี่ (Dengue fever: DF), ไข้เลือดออก (Dengue haemorrhagic fever: DHF) และไข้เลือดออกช็อก (Dengue shock syndrome: DSS) ลักษณะสำคัญของโรคไข้เลือดออก (DHF) คือ มีไข้สูงเฉียบพลัน ไข้สูงลอยอยู่ 2 - 7 วัน มีอาการเลือดออกส่วนใหญ่จะพบที่ผิวหนัง ตับโต ในรายที่มีอาการรุนแรง อาจมีภาวะช็อก (DSS) เป็นเหตุให้เสียชีวิตได้ แตกต่างจากไข้เด็งกี่ (DF) ซึ่งมีอาการไม่รุนแรง โดยทั่วไปไม่ทำให้เสียชีวิต [2] ปัจจุบันการวินิจฉัยโรคไม่สามารถจำแนกได้ว่าเป็นไข้เด็งกี่หรือไข้เลือดออกได้ภายในระยะแรกของการติดโรค การทดสอบมักจะเชื่อถือได้ในผู้ป่วยที่มีไข้แล้วหลายวันซึ่งส่วนใหญ่จะมีอาการของโรคชัดเจนแล้ว [3] ดังนั้น การวินิจฉัยโรคจากความแตกต่างทางพันธุกรรมจึงเป็นทางเลือกที่อาจช่วยให้สามารถวินิจฉัยโรคได้เร็วและแม่นยำยิ่งขึ้น ซึ่งจะช่วยลดความเสี่ยงที่จะเสียชีวิต และประหยัดทรัพยากรในการรักษาพยาบาล

สนิปส์ หรือ SNPs (Single Nucleotide Polymorphisms) เป็นความแตกต่างทางพันธุกรรมที่เกิดจากเบสบนสาย ลำดับนิวคลีโอไทด์ ณ ตำแหน่งหนึ่งๆ แตกต่างกัน เป็นรูปแบบความแตกต่างทางพันธุกรรมพื้นฐานที่พบบ่อยที่สุด [4] ใช้เป็นเครื่องหมายทางพันธุกรรม (genetic marker) เพื่อการวินิจฉัยโรค ทำนายความเสี่ยงต่อการเกิดโรค ใช้ในการพัฒนายาและวิธีการรักษาโรคที่เหมาะสมสำหรับแต่ละบุคคล[5] โดยจะมีประสิทธิภาพมากยิ่งขึ้นเมื่อศึกษาสายลำดับของสนิปส์ตำแหน่งต่างๆ บนดีเอ็นเอ ซึ่งเรียกว่า แฮปโพลไทป์ (haplotype) แทนที่จะดูจากสนิปส์ที่ตำแหน่งใดตำแหน่งหนึ่งเพียงตำแหน่งเดียว [6]

จากความสำคัญของสนิปส์และแฮปโพลไทป์ดังกล่าว จึงมีงานวิจัยจำนวนมากที่เสนอขั้นตอนวิธีสำหรับประกอบแฮปโพลไทป์ เช่นในงาน [5, 6, 7, 8, 9, 10, 11] งานวิจัยเหล่านี้ศึกษาเฉพาะแฮปโพลไทป์ของมนุษย์และสิ่งมีชีวิตที่มีโครโมโซม 2 ชุด (diploid organism) เท่านั้น

สำหรับสิ่งมีชีวิตที่มีโครโมโซมชุดเดียว (haploid organism) เช่น ไวรัสเด็งกี่ ข้อมูลประชากรของไวรัส คือ ชุดของสายแฮปโพลไทป์ และความถี่แฮปโพลไทป์ (haplotype frequency) แต่ละเส้น มีผลต่อการเกิดโรค ดังนั้นเราจึงต้องการประกอบสายแฮปโพลไทป์ และประมาณค่าความถี่แฮปโพลไทป์ เพื่อใช้เป็นเครื่องหมายทางพันธุกรรมและเพื่อศึกษากลไกการเกิดโรค

อย่างไรก็ตาม ไวรัสเด็งกี่เป็นอาร์เอ็นเอไวรัส (RNA virus) มีการถ่ายแบบ (replication) ด้วยอาร์เอ็นเอพอลิเมอเรส (RNA polymerase) ซึ่งมีข้อผิดพลาดสูง ทำให้มีการกลายพันธุ์สูง มีความหลากหลายสูง เชื้อไวรัสเด็งกี่นี้สามารถแยกออกเป็น 4 ซีโรไทป์ (serotype) ได้แก่ ซีโรไทป์ 1 – 4 แต่ละซีโรไทป์ ก็แยกย่อยออกไปได้อีก สาเหตุหลักที่ทำให้เชื้อไวรัสเด็งกี่มีความหลากหลายสูงเกิดจากการเปลี่ยนชนิดของเบส โดยไวรัสเด็งกี่ที่อยู่ในซีโรไทป์เดียวกันจะมี สายลำดับนิวคลีโอไทด์ที่คล้ายคลึงกันมากกว่าไวรัสเด็งกี่ที่อยู่ต่างซีโรไทป์กันแต่ก็ไม่เหมือนกันทั้งหมด กล่าวได้ว่า ไวรัสเด็งกี่มีลักษณะกึ่งสปีชีส์ (quasispecies) [12] ดังนั้นในผู้ป่วยแต่ละคนจึงมีประชากรของอาร์เอ็นเอไวรัสที่มียีนแตกต่างกันแต่คล้ายกันมาก ซึ่งการศึกษาความหลากหลายของสายลำดับนิวคลีโอไทด์ของสิ่งมีชีวิตที่มีลักษณะกึ่งสปีชีส์ ไม่สามารถทำได้ด้วยเทคโนโลยีอ่านลำดับเบส (sequencing technology) ที่มีในอดีต แต่สามารถทำได้ด้วยเทคโนโลยี พิโกไทเตอร์เพลตไพโรซีเควนซิง (picotiter plate pyrosequencing) ซึ่งอ่านสายลำดับ นิวคลีโอไทด์จำนวนมากพร้อมๆ กันแบบขนาน [13] จึงสามารถอ่านสายลำดับนิวคลีโอไทด์ได้จำนวนมากภายในเวลาอันสั้น

เทคโนโลยีอ่านลำดับ เบสแบบขนานจำนวนมาก (massively parallel sequencing) นี้ มีหลักการพื้นฐานเช่นเดียวกับเทคโนโลยี อ่านสายลำดับเบสแบบชอตกัน (shotgun sequencing) ซึ่งเป็นวิธีที่เป็นที่นิยมที่สุดในปัจจุบัน แต่มีข้อแตกต่างบางประการที่สำคัญ คือ สายลำดับ นิวคลีโอไทด์ที่อ่านได้มีความยาวสั้น กว่าเดิมมาก แต่สามารถอ่านสายลำดับ นิวคลีโอไทด์ได้จำนวนมากต่อการอ่านหนึ่งครั้ง สำหรับเครื่องอ่านลำดับเบส Roche GS FLX อ่านสายลำดับ นิวคลีโอไทด์ได้มากถึง 400,000 เส้นต่อการอ่านหนึ่งครั้ง โดยมีความยาวเฉลี่ย 200 -300 คู่เบส (bp) ซึ่งสั้นกว่าเดิมที่มีความยาว 500-1000 คู่เบส [14]

จากการที่ไวรัสเด็งกี่มีรูปแบบกึ่งสปีชีส์ ไม่ได้เป็นสิ่งมีชีวิตที่มีโครโมโซมสองคู่ (diploid organism) และเทคโนโลยีอ่านลำดับ เบสที่ใช้ซึ่งแตกต่างจากเทคโนโลยีเดิมที่นิยมใช้กันทั่วไป ทำให้ต้องพัฒนาวิธีการประกอบชุดของแฮปโพลไทป์ขึ้นใหม่ เพื่อ ให้เหมาะสมกับลักษณะของไวรัสเด็งกี่และรูปแบบข้อมูลที่ได้จากเทคโนโลยีใหม่นี้

1.2 วัตถุประสงค์

งานวิจัยนี้มีวัตถุประสงค์เพื่อนำเสนอวิธีการประกอบแฮปโพลไทป์และประมาณความถี่แฮปโพลไทป์ของประชากรไวรัสเด็งกี่ จากข้อมูลสายลำดับ นิวคลีโอไทด์ที่ได้จากเทคโนโลยีอ่าน

ลำดับเบสแบบขนานจำนวนมาก โดยมุ่งเน้นที่ข้อมูลจากเครื่องอ่านลำดับเบส Roche GS FLX เป็นหลัก

1.3 ขอบเขตการวิจัย

1. ข้อมูลขาเข้า (input data) สำหรับวิธีการที่นำเสนออยู่ในรูปแบบเดียวกับข้อมูลที่ได้จากเครื่องอ่านลำดับเบส Roche GS FLX คือ มีความยาวเฉลี่ย 250 คู่เบส และเป็นเส้นเดี่ยว (ไม่มีข้อมูลของเส้นคอมพลีเมนต์)
2. ข้อมูลขาเข้า (input data) คือข้อมูลสายลำดับนิวคลีโอไทด์ที่อ่านได้จากเทคโนโลยีอ่านลำดับเบสแบบขนานจำนวนมาก ที่ผ่านการจัดการความผิดพลาด (Error) แล้ว
3. สายลำดับที่อ่านได้นี้ มีจีโนมอ้างอิง และสามารถจัดเรียง (align) สายลำดับเหล่านี้บนจีโนมอ้างอิงได้ถูกตำแหน่ง
4. วิธีการที่นำเสนอมุ่งเน้นสำหรับไวรัสเด็งกี่ ซึ่งมีรูปแบบ กึ่งสปีชีส์ ความยาวของจีโนมทั้งเส้นไม่มากนัก ประมาณ 11,000 คู่เบส (11 Kbp) และไม่มีส่วนที่ซ้ำกัน (repeat) อันเป็นปัญหาหลักในการประกอบจีโนมของมนุษย์

1.4 ประโยชน์ที่คาดว่าจะได้รับ

สามารถนำชุดของสาย แฮปโพลไทป์ ที่ได้ไปหาเครื่องหมายทางพันธุกรรม (genomic marker) เพื่อใช้ในการวินิจฉัยโรคไข้เลือดออกได้ หรือช่วยให้สามารถอธิบายการเกิดโรคและการระบาดของโรคได้ดียิ่งขึ้น และสามารถประยุกต์ขั้นตอนวิธีการประกอบชุดของสาย แฮปโพลไทป์นี้กับจีโนมของสิ่งมีชีวิตอื่นที่มีรูปแบบกึ่งสปีชีส์ได้

1.5 ขั้นตอนดำเนินงานวิจัย

1. ศึกษาข้อมูลพื้นฐานทางชีววิทยาที่เกี่ยวข้อง เช่น ดีเอ็นเอ, อาร์เอ็นเอ, สนิปส์, ความหลากหลายทางพันธุกรรม, แฮปโพลไทป์ , เทคโนโลยีการอ่านสายลำดับนิวคลีโอไทด์ , ไวรัสเด็งกี่ เป็นต้น

2. ศึกษารายละเอียดของเทคโนโลยี พิโกไทเตอร์เพลตไพโรซีเควนซิง (picotiter plate pyrosequencing) โดยเฉพาะรูปแบบของข้อมูลที่ได้ ซึ่งจะเป็นข้อมูลเข้า (input) สำหรับงานวิจัยนี้

3. ทดสอบความเป็นไปได้ ในการใช้เทคโนโลยี พิโกไทเตอร์เพลตไพโรซีเควนซิง ศึกษาความหลากหลายทางพันธุกรรมของไวรัสเด็งกี่

4. ศึกษาวิธีการประกอบแฮปโพลไทป์ที่ใช้ในปัจจุบัน

5. ออกแบบขั้นตอนวิธีในการประกอบชุดของแฮปโพลไทป์สำหรับไวรัสเด็งกี่

6. ทดสอบวิธีการที่นำเสนอโดยใช้ฐานข้อมูลจีโนมเชื้อไวรัสเด็งกี่ ในการจำลองข้อมูลให้มีลักษณะเช่นเดียวกับข้อมูลที่จะได้จากเครื่องอ่านลำดับเบส Roche GS FLX

7. เปรียบเทียบวิธีที่นำเสนอกับวิธีที่เสนอในงานวิจัยเรื่องการประมาณประชากรไวรัสโดยใช้ไพโรซีเควนซิง (Viral population estimation using pyrosequencing)

8. วิเคราะห์ผลการทดลอง

9. สรุปผลและเรียบเรียงวิทยานิพนธ์

1.6 ผลงานที่ตีพิมพ์จากวิทยานิพนธ์

ส่วนหนึ่งของวิทยานิพนธ์นี้ได้รับการตอบรับให้ตีพิมพ์เป็นบทความทางวิชาการในหัวข้อเรื่อง *"การจำลองแบบการประกอบสายลำดับหลักของสนิปส์ใน ไวรัสเด็งกี่ด้วยเทคโนโลยีการอ่านลำดับนิวคลีโอไทด์แบบขนานจำนวนมาก : Major SNPs Sequence Assembling Simulation of Dengue Virus Genome from Massively Parallel Sequencing Technique"* โดย พุธิตา สุมานนท์, ศ.ดร. ประภาส จงสถิตย์วัฒนา และ ดร.ประพัฒน์ สุริยผล ในงานประชุมวิชาการ "วิทยาการคอมพิวเตอร์และวิศวกรรมคอมพิวเตอร์แห่งชาติ ครั้งที่ 12" (The 12th National Computer Science and Engineering Conference: NCSEC2008) ณ โรงแรมลองบีชการ์เด้นโฮเทลแอนด์สปา จังหวัดพัทยา ระหว่างวันที่ 20-21 พฤศจิกายน 2551

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 เทคโนโลยีการอ่านสายลำดับ (Sequencing technique)

ระเบียบวิธีที่มีประสิทธิภาพในการอ่านสายลำดับนิวคลีโอไทด์ถูกเสนอครั้งแรกในปี ค.ศ. 1977 โดยมี 2 วิธีที่ถูกตีพิมพ์ในเวลาใกล้เคียงกัน คือ

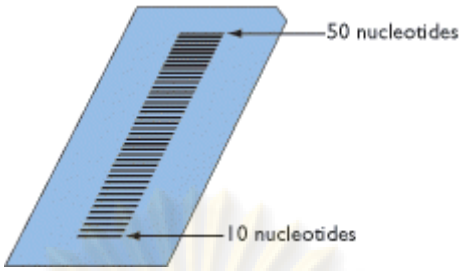
1. ระเบียบวิธีสิ้นสุดสาย (chain termination method) เสนอโดย Sanger และคณะ ในวิธีนี้ สายลำดับ นิวคลีโอไทด์ ของดีเอ็นเอเส้นเดี่ยวจะถูกอ่านโดยการสังเคราะห์สายนิวคลีโอไทด์ที่เข้าคู่กัน (complementary polynucleotide chain) ซึ่งการสังเคราะห์สายลำดับเหล่านี้จะสิ้นสุด ณ ตำแหน่งนิวคลีโอไทด์ที่เฉพาะเจาะจง
2. ระเบียบวิธีลดความซับซ้อนทางเคมี (chemical degradation method) เสนอโดย Maxam และ Gilbert ในวิธีนี้ สายลำดับ นิวคลีโอไทด์ของดีเอ็นเอเส้นคู่จะถูกอ่านด้วยปฏิบัติการทางเคมีซึ่งตัดโมเลกุลดีเอ็นเอ ณ ตำแหน่งนิวคลีโอไทด์ที่เฉพาะเจาะจง

ทั้ง 2 วิธีนี้เป็นที่นิยมใกล้เคียงกันในระยะแรก แต่ในปัจจุบันนิยมใช้ระเบียบวิธีสิ้นสุดสายมากกว่า เนื่องจากระเบียบวิธีลดความซับซ้อนทางเคมีต้องใช้สารที่เป็นพิษและอันตรายต่อผู้วิจัย และเหตุผล ที่สำคัญคือสามารถปรับปรุงระเบียบวิธีสิ้นสุดสายให้เป็นระบบอัตโนมัติได้ง่ายกว่า ดังนั้นในที่นี้จะกล่าวถึงเฉพาะวิธีการอ่านสายลำดับดีเอ็นเอด้วยวิธีสิ้นสุดสายเท่านั้น

2.1.1.1 การอ่านสายลำดับดีเอ็นเอด้วยวิธีสิ้นสุดสาย (chain termination DNA sequencing) [15]

การอ่านสายลำดับดีเอ็นเอด้วยวิธีสิ้นสุดสายอาศัยหลักการพื้นฐานคือ สายดีเอ็นเอที่มีความยาวแตกต่างกันจะถูกแยกออกจากกันได้ ด้วยการเคลื่อนที่ในเจลพอลิอะไครลาไมด์ (polyacrylamide) ภายใต้อิทธิพลของสนามไฟฟ้า ซึ่งจะทำให้สามารถอ่านสายลำดับนิวคลีโอไทด์ซึ่งมีความยาวระหว่าง 10 – 1000 คู่เบส ได้จากชุดของแถบในเจล ดังรูปที่ 2.1

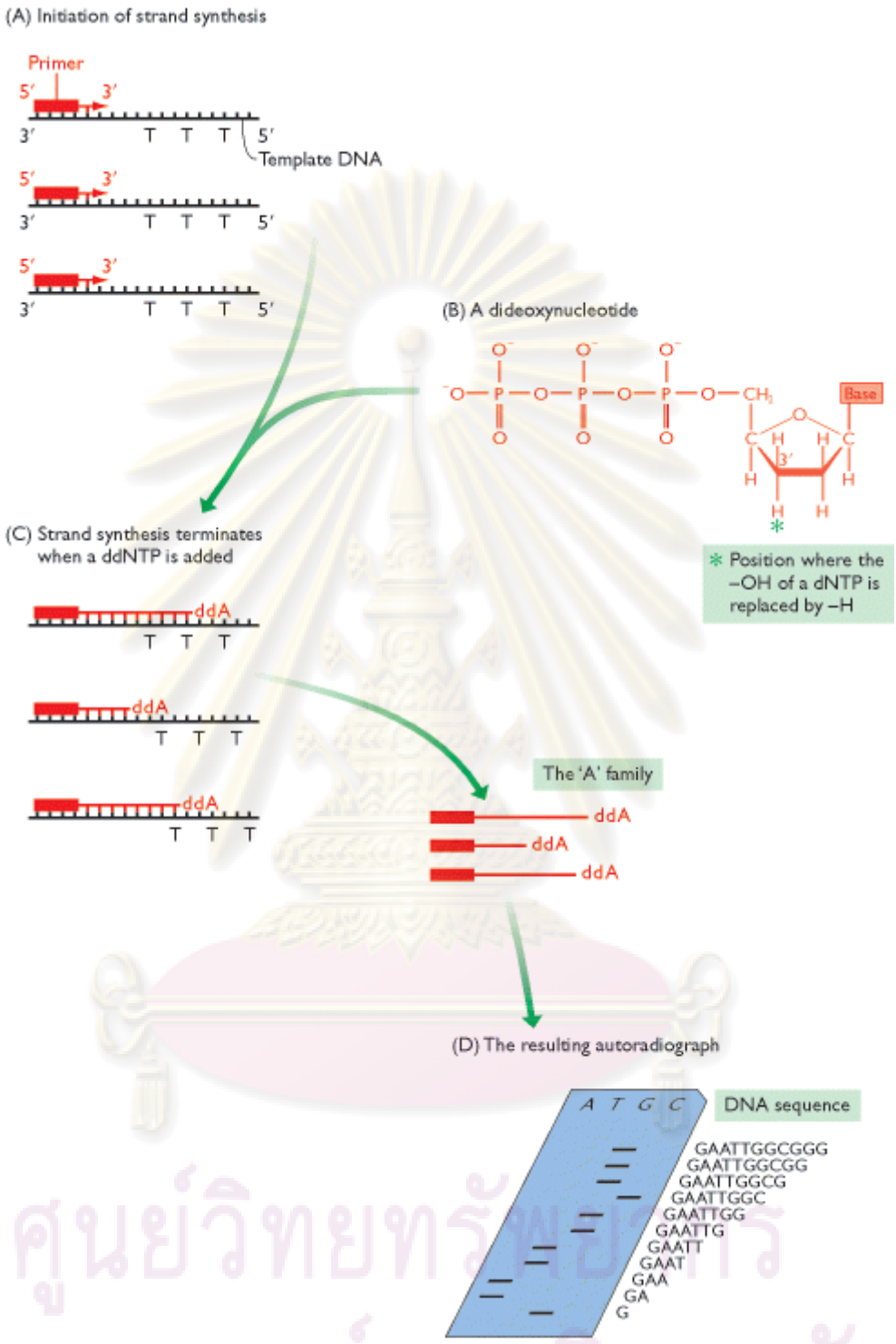
ขั้นตอนแรกคือการเตรียมไพรเมอร์ (primer) หรือ สายลำดับเส้นเดี่ยวท่อนสั้นๆ ที่เป็นคู่เบส (complementary) ของสายดีเอ็นเอที่ต้องการอ่าน นำไพรเมอร์นี้ไปประกบคู่กับสายที่ต้องการอ่าน ดังรูปที่ 2.2(A) จากนั้นเติมเอนไซม์ดีเอ็นเอพอลิเมอเรส (DNA polymerase) ซึ่งเป็นเอนไซม์ที่ทำให้เกิดการต่อสายลำดับด้วยนิวคลีโอไทด์ที่เป็นคู่เบสกับสายต้นแบบ เติมดีออกซีไรโบนิวคลีโอไทด์ไตรฟอสเฟตทั้ง 4 ชนิด คือ dATP, dCTP, dGTP และ dTTP และไดดีออกซีไรโบนิวคลีโอไทด์



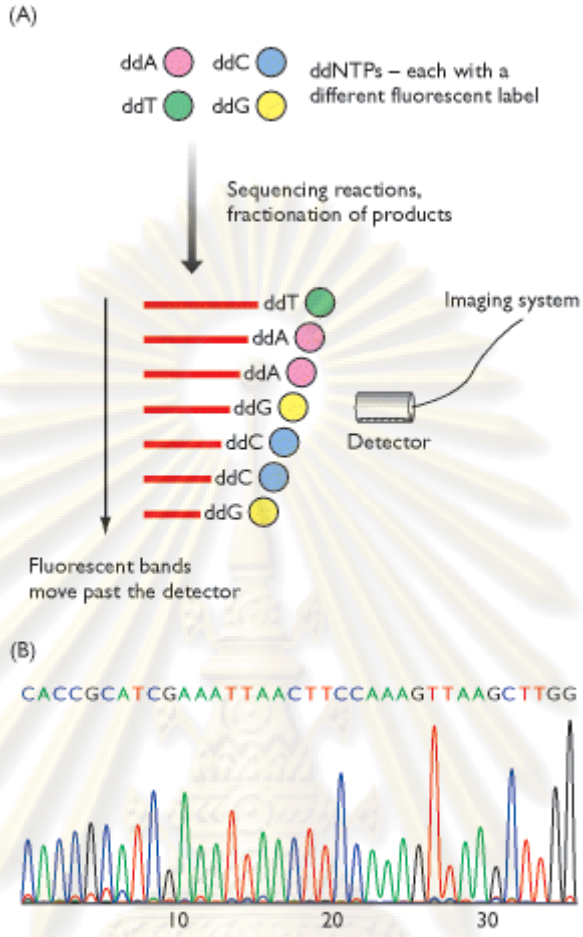
รูปที่ 2.1 แถบที่เกิดจากการเคลื่อนที่ของสายดีเอ็นเอที่มีความยาวแตกต่างกันในเจลพอลิอะไครลาไมด์ (polyacrylamide) ภายใต้อิทธิพลของสนามไฟฟ้า [15]

ไตรฟอสเฟต 1 ชนิด เช่น ddATP จำนวนเล็กน้อย ซึ่งไดดีออกซีไรโบนิวคลีโอไทด์ไตรฟอสเฟต (ddNTPs) นี้ มีโครงสร้างแตกต่างจากดีออกซีไรโบนิวคลีโอไทด์ไตรฟอสเฟต (dNTPs) โดยหมู่ไฮดรอกซิล (hydroxyl) ที่ปลายด้าน 3′ ถูกแทนที่ ดังรูปที่ 2.2(B) ซึ่งหมู่ไฮดรอกซิลนี้จำเป็นสำหรับการต่อสายกับนิวคลีโอไทด์ถัดไป ดังนั้นสายที่ถูกต่อด้วยไดดีออกซีไรโบนิวคลีโอไทด์ไตรฟอสเฟต (ddNTPs) การสังเคราะห์นิวคลีโอไทด์จะสิ้นสุดลง ดังรูปที่ 2.2(C) เนื่องจากในแต่ละหลอดทดลองมีสายต้นแบบที่ต้องการอ่านจำนวนมาก ดังนั้นแต่ละสายจะถูกหยุดที่ตำแหน่งต่างๆ กัน ทำเช่นนี้กับไดดีออกซีไรโบนิวคลีโอไทด์ไตรฟอสเฟตทุกชนิดที่เหลือทีละชนิดจนครบ จากนั้นนำแต่ละหลอดทดลองไปใส่ในเจลโดยแยกช่องทาง (lane) กันรวมทั้งสิ้น 4 ช่องทาง หลังจากให้สนามไฟฟ้าซึ่งทำให้เกิดการเคลื่อนที่ของสายลำดับแล้ว จะสามารถอ่านสายลำดับดีเอ็นเอได้โดยตรงจากแถบที่ปรากฏบนเจลดังรูปที่ 2.2(D)

วิธีมาตรฐานของการอ่านลำดับด้วยวิธีสิ้นสุดสายนี้ใช้กัมมันตภาพรังสีในการทำป้าย (label) และมองเห็นแถบบนเจลได้ด้วยเครื่องอ่านกัมมันตภาพรังสี (autoradiography) ซึ่งเราสามารถแทนที่ป้ายกัมมันตภาพนี้ได้ด้วยป้ายเรืองแสง (fluorolabel) ซึ่งช่วยการตรวจจับเอื้ออำนวยต่อการทำให้เป็นอัตโนมัติมากกว่า ป้ายเรืองแสงนี้จะติดกับไดดีออกซีไรโบนิวคลีโอไทด์ไตรฟอสเฟต (ddNTPs) โดยใช้ป้ายเรืองแสงที่แตกต่างกันสำหรับนิวคลีโอไทด์แต่ละชนิด ดังรูปที่ 2.3(A) การทำเช่นนี้จะทำให้สามารถสังเคราะห์สายลำดับได้ในหลอดทดลองเดียวกันและอ่านได้ด้วยเจลเพียง 1 ช่องทาง (lane) เนื่องจากเครื่องตรวจจับเห็นความแตกต่างระหว่างป้ายของแต่ละนิวคลีโอไทด์ สามารถบอกได้ว่าแต่ละแถบแสดงถึงนิวคลีโอไทด์ใด เราสามารถอ่านผลจากเครื่องโดยตรง พิมพ์ออกมาในรูปของกราฟ ดังรูปที่ 2.3(B) หรือให้เก็บผลไว้ในคอมพิวเตอร์โดยตรงก็ได้ เมื่อนำวิธีนี้ไปรวมกับอุปกรณ์หุ่นยนต์ซึ่งทำหน้าที่เตรียมสารตั้งต้นและบรรจุเจลแล้ว ระบบการอ่านด้วยป้ายเรืองแสงนี้ให้จำนวนผลลัพธ์เพิ่มขึ้น และลดความผิดพลาดซึ่งอาจเกิดขึ้นจากการอ่านด้วยตาแล้วเก็บค่าลงในคอมพิวเตอร์ด้วยการป้อนข้อมูลผ่านทางแป้นพิมพ์ ระบบอัตโนมัตินี้ช่วยให้การอ่านสายลำดับของทั้งจีโนมสามารถทำได้ในเวลาที่เหมาะสม



รูปที่ 2.2 ขั้นตอนการอ่านสายลำดับดีเอ็นเอด้วยวิธีสิ้นสุดสาย A) เตรียมไพรเมอร์ (primer) และจับคู่กับสายต้นแบบ B) ในไดดีออกซิล (ddNTPs) หมู่ไฮดรอกซิล (-OH) ที่ตำแหน่ง 3′ ใน dNTPs ถูกแทนที่ ทำให้ไม่สามารถสร้างสายลำดับต่อได้ C) การสังเคราะห์สายลำดับจะหยุด ณ ตำแหน่งที่ต่อกับ ddATP ทำให้สายลำดับที่สังเคราะห์ได้มีความยาวที่แตกต่างกันออกไป D) นำสายลำดับไปใส่ในเจลภายใต้สนามไฟฟ้า ทำให้เกิดแถบที่ความยาวต่างๆ กัน ซึ่งทำให้สามารถอ่านสายลำดับดีเอ็นเอได้จากแถบเหล่านี้โดยตรง [15]

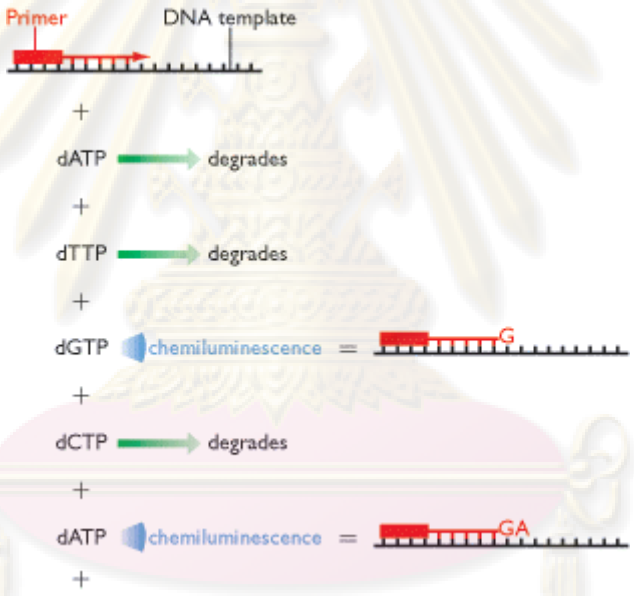


รูปที่ 2.3 การอ่านสายลำดับดีเอ็นเอแบบอัตโนมัติโดยใช้ป้ายเรืองแสง A) ปฏิกิริยาสิ้นสุดสายเกิดขึ้นในหลอดทดลองเดียวด้วยป้ายเรืองแสงที่แตกต่างกันออกไปสำหรับแต่ละไดดีออกซีนิวคลีโอไทด์ในเครื่องอ่านลำดับแบบอัตโนมัติ แถบในเจลจะเคลื่อนที่ผ่านเครื่องตรวจจับแสงซึ่งจะระบุไดดีออกซีนิวคลีโอไทด์ที่แสดงในแต่ละแถบ B) ผลที่พิมพ์จากเครื่องอ่านสายลำดับ แสดงด้วยชุดของจุดยอด แต่ละจุดยอดแทนนิวคลีโอไทด์แต่ละตำแหน่งบนสายลำดับ [15]

ถึงแม้ว่าการอ่านสายลำดับด้วยวิธีสิ้นสุดสายจะถูกพัฒนาเป็นระบบอัตโนมัติ แต่ยังมีข้อจำกัดที่สามารถอ่านได้จำนวนไม่มาก ไม่ถึง 1000 คู่เบส ต่อการทดลอง 1 ครั้ง ซึ่งถ้าใช้ในโครงการจีโนมมนุษย์ (Human Genome Project) การอ่าน 1 ครั้งจะอ่านได้เพียงหนึ่งในห้าล้านของความยาวจีโนมทั้งหมด จึงมีความพยายามพัฒนาเทคโนโลยีให้เร็วขึ้น เช่น วิธีของ Mulikan และ McMurray ใช้หลอดคาปิลลารี (capillary) แทนเจล มีทั้งหมด 96 ช่องทาง ทำให้สามารถอ่านลำดับพร้อมๆ กันได้ถึง 96 สาย โดยใช้เวลาน้อยกว่า 2 ชั่วโมง และสามารถอ่านสายลำดับได้ถึง 1000 สายต่อวัน หรือวิธีที่พัฒนาโดย Rogers พัฒนาให้สามารถอ่านได้พร้อมๆ กัน 384 -1024 สายลำดับต่อครั้ง

2.1.1.2 เทคโนโลยีการอ่านสายลำดับแบบขนาน (parallel sequencing)

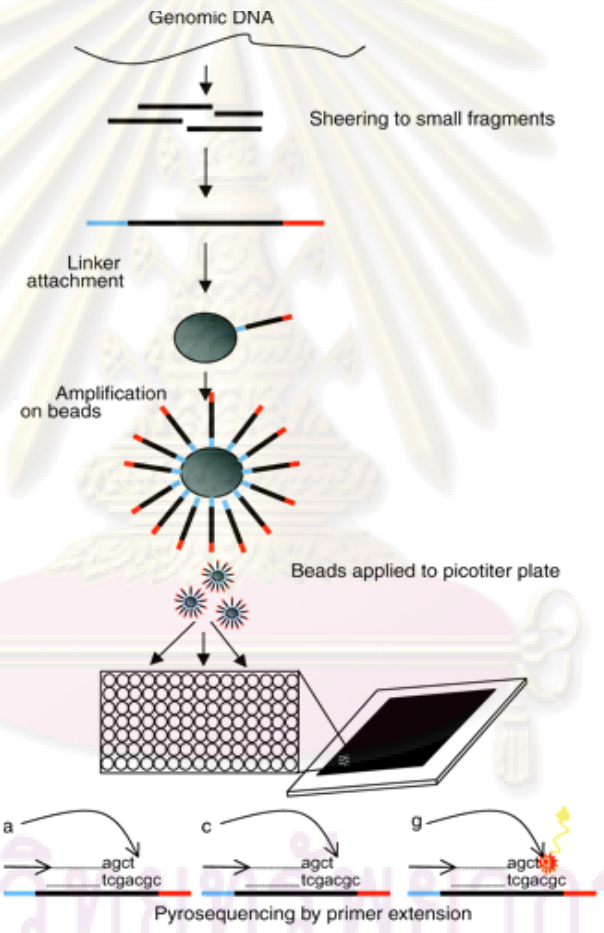
จากความพยายามในการพัฒนาให้สามารถอ่านลำดับนิวคลีโอไทด์ได้รวดเร็วยิ่งขึ้น จึงมีงานวิจัยจำนวนมากที่เสนอวิธีการ อ่านสายลำดับแบบใหม่ ดังเช่นวิธี ไพโรซีเควนซิง (pyrosequencing) ซึ่งสายต้นแบบจะถูกคัดลอกโดยไม่ต้องใช้ไดดีออกซีไรโบนิวคลีโอไทด์ไตรฟอสเฟต (ddNTPs) ในกระบวนการ เมื่อสายใหม่ถูกสร้างขึ้น ลำดับของดีออกซีไรโบนิวคลีโอไทด์ไตรฟอสเฟต (dNTPs) ที่ถูกใช้ในการสังเคราะห์สายใหม่นี้จะถูกตรวจจับ โดยดีออกซีไรโบนิวคลีโอไทด์ไตรฟอสเฟตที่รวมกับสายต้นแบบจะปล่อยโมเลกุลไพโรฟอสเฟต (pyrophosphate) ซึ่งสามารถเปล่งแสงได้โดยใช้เอนไซม์ซัลเฟอริเลส (sulferylase) แต่ละดีออกซีไรโบนิวคลีโอไทด์ไตรฟอสเฟตจะถูกใส่เข้ามาในหลอดทดลองทีละชนิด ถ้าเกิดแสงแสดงว่าชนิดนั้นต่อได้ แต่ถ้าต่อไม่ได้ก็จะไม่เกิดแสง ดังรูปที่ 2.4 [15]



รูปที่ 2.4 การอ่านสายลำดับดีเอ็นเอแบบขนาน ดีออกซีนิวคลีโอไทด์ (dNTP) จะถูกเติมเข้าไปทีละชนิด ชนิดใดถูกนำไปใช้ในการต่อสายจะเปล่งแสงออกมา ทำให้สามารถอ่านสายลำดับนิวคลีโอไทด์ได้ตามลำดับชนิดของนิวคลีโอไทด์ที่เติมเข้าไป [15]

วิธีไพโรซีเควนซิง (pyrosequencing) นี้ง่ายต่อการทำให้เป็นแบบอัตโนมัติ และสามารถทำได้พร้อมๆ กันจำนวนมาก จึงนำมาใช้ในเครื่องอ่านลำดับเบส Roche GS20 หรือ Roche GS FLX ซึ่งพัฒนาโดยกลุ่มวิจัย 454 Life Sciences ซึ่งสามารถอ่านได้ถึง 25 ล้านเบสต่อครั้งสำหรับเครื่องอ่านลำดับเบส GS20 [13] และเพิ่มขึ้นเป็น 80 ล้านเบสต่อครั้ง ในเครื่องอ่านลำดับเบส GS FLX [14]

ระบบของ 454 Life Sciences นี้เริ่มตั้งแต่การเพิ่มจำนวนสายดีเอ็นเอโดยสายต้นแบบจะถูกตัดเป็นชิ้นเล็กๆ แต่ละชิ้นจะถูกเชื่อมกับเม็ดกลมๆ เล็กๆ 1 ชิ้นส่วนดีเอ็นเอต่อเม็ด และเข้าสู่กระบวนการเพิ่มจำนวน (amplification) แยกแต่ละเม็ด หลังจากนั้นนำไปใส่ใน พิโกไทเตอร์เพลต (picotiter plate) ซึ่งเป็นแผ่นที่ประกอบด้วยหลุมเล็กๆ จำนวนมาก ขนาดหลุมพอดีสำหรับ 1 เม็ดเท่านั้น จากนั้นอ่าน สายลำดับนิวคลีโอไทด์ตามหลักการของการอ่านสายลำดับแบบขนานสำหรับแต่ละหลุม ดังนั้นในการทดลอง 1 ครั้งจะมีการอ่านลำดับเบสพร้อมกันจำนวนมากแบบขนาน (400,000 เส้นสำหรับเครื่องอ่านลำดับเบส GS FLX) [16] ดังรูปที่ 2.5



รูปที่ 2.5 กระบวนการอ่านสายลำดับด้วยวิธีของ 454 Life Sciences สายดีเอ็นเอถูกตัดเป็นชิ้นเล็กๆ เข้าสู่กระบวนการเพิ่มจำนวนโดย 1 ชิ้นจะติดกับเม็ดกลม 1 เม็ด นำผลที่ได้จากการเพิ่มจำนวนไปใส่ในพิโกไทเตอร์เพลต หลุมละไม่เกิน 1 เม็ด จากนั้นอ่านลำดับเบสในแต่ละหลุม [16]

2.2 งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้องโดยตรงกับงานวิจัยนี้ คือ งานวิจัยเรื่อง การประมาณประชากรไวรัสโดยใช้ไพโรซีเควนซิง (Viral population estimation using pyrosequencing) [17] เสนอโดย Eriksson และคณะ ในปี 2008 มุ่งศึกษาโครงสร้างของประชากรของไวรัสจากข้อมูลสายลำดับที่ได้จาก เครื่องอ่านสายลำดับ Roche GS20 ซึ่งใช้เทคโนโลยีอ่านสายลำดับแบบขนานจำนวนมาก และโครงการ คิว แอสเซมเบลอร์ (Q Assembler) [18] ที่มีวัตถุประสงค์หลักเพื่อพัฒนาซอฟ ต์แวร์สำหรับประกอบกลุ่มจีโนมของไวรัสจากข้อมูลสายลำดับที่อ่านได้จากเทคโนโลยีอ่านสายลำดับ เบสแบบขนานจำนวนมาก นอกจากนี้ยังมีงานวิจัยที่เกี่ยวข้อง คือ งานวิจัยเกี่ยวกับการประกอบแฮปโพลไทป์ และงานวิจัยที่เกี่ยวข้องกับเทคโนโลยีอ่านสายลำดับเบสแบบขนานจำนวนมาก

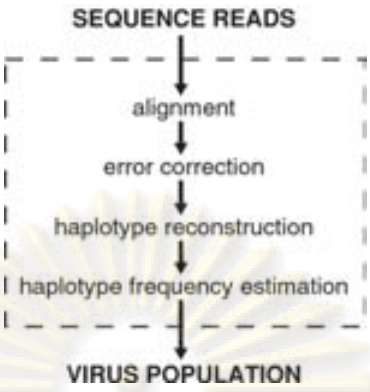
2.2.1 งานวิจัยเรื่องการประมาณประชากรไวรัสโดยใช้ไพโรซีเควนซิง (Viral population estimation using pyrosequencing)

Eriksson และคณะ [17] ได้เสนอขั้นตอนการประมาณโครงสร้างประชากรของไวรัสจากการอ่านสายลำดับเบสแบบขนาน โดยผลลัพธ์หลักที่ต้องการคือ เซตที่เล็กที่สุดของแฮปโพลไทป์ที่สามารถอธิบายประชากรของไวรัสทั้งหมดในกลุ่มตัวอย่าง และความถี่ของแฮปโพลไทป์เหล่านั้น โดยใช้ข้อมูลสายลำดับที่อ่านได้จากเทคโนโลยี อ่านสายลำดับเบสแบบขนาน ขั้นตอนที่นำเสนอนี้ใช้ได้ เมื่อ สามารถจัดเรียงสายลำดับที่อ่านได้เทียบกับจีโนมอ้างอิงได้ การอ่านสายลำดับมีความครอบคลุม (coverage) เพียงพอ และระยะห่างทางพันธุกรรม (genetic distance) ระหว่างแฮปโพลไทป์มากเพียงพอ

งานวิจัยนี้ได้แบ่งขั้นตอนการประมาณโครงสร้างประชากรของไวรัสนี้ ออกเป็น 4 ขั้นตอนหลัก ดังรูปที่ 2.6 คือ จัดเรียงสายลำดับที่อ่านได้เทียบกับเส้นต้นแบบ (alignment) แก้ไขข้อผิดพลาด (error correction) ประกอบแฮปโพลไทป์ (haplotype reconstruction) และประมาณความถี่แฮปโพลไทป์ (haplotype frequency estimation) โดยได้นำเสนอขั้นตอนวิธีสำหรับ 3 ขั้นตอนหลัง คือ ขั้นตอนการแก้ไขข้อผิดพลาด การประกอบแฮปโพลไทป์และการประมาณความถี่แฮปโพลไทป์ ส่วนการจัดเรียงใช้วิธีการเทียบเคียงเข้าคู่ (pairwise alignment) ที่มีนำเสนออยู่แล้ว

2.2.1.1 ขั้นตอนแก้ไขข้อผิดพลาด (error correction)

เนื่องจาก การอ่านลำดับเบสแบบขนาน มีอัตราความผิดพลาดสูง ส่วนใหญ่เป็นการเติมหรือการลบออก (indel) ในบริเวณ โฮโมโพลิเมอร์ริก (homopolymeric) ซึ่งประกอบด้วยเบสชนิดเดียวกันเรียงต่อกัน แต่ละเบสที่อ่านได้จะถูกกำกับด้วยคะแนนคุณภาพ (quality score) แสดงความน่าเชื่อถือ

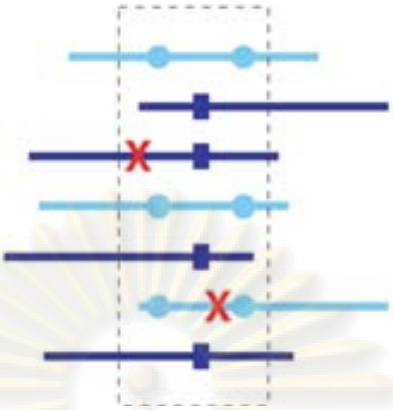


รูปที่ 2.6 ขั้นตอนหลักของการประมาณ ประชากร ไวรัส โดยใช้ไพโรซีเควนซิง (pyrosequencing) สายลำดับที่อ่านได้จะถูกจัดเรียงเทียบกับสายอ้างอิง แก้ไขข้อผิดพลาด ประกอบแฮปโพลไทป์ และประมาณค่าความถี่ของแฮปโพลไทป์ที่ประกอบขึ้น ใช้ชุดของแฮปโพลไทป์และความถี่ที่ประมาณได้นี้อนุมานประชากรของไวรัส [17]

ของเบสนั้นๆ อัตราความผิดพลาดของเครื่องอ่านลำดับเบส Roche GS20 มีค่าประมาณ 5-10 ตำแหน่งต่อพันเบส (errors/kb) เมื่อกำจัดสายลำดับที่ความยาวผิดปกติหรือมีเบสกำกวมโดยพิจารณาจากคะแนนคุณภาพจะเหลือ 1-3 ตำแหน่งต่อพันเบส ส่วนที่เหลือนี้ประกอบด้วยการเติม (insertion) ร้อยละ 50 ที่เหลือเป็นการลบออก (deletion) ร้อยละ 25 และการแทนที่ (substitution) อีกร้อยละ 25 จากสมมติฐานของงานวิจัยนี้ที่ไม่มีการเติมหรือการลบออก จึงแก้ไขโดยจัดเรียงแล้วแก้ส่วนที่เป็น การเติมหรือการลบออกเทียบกับเส้นอ้างอิง ซึ่งจะทำให้เหลือข้อผิดพลาดประมาณ 1 ตำแหน่งต่อพันเบส

จากนั้นเสนอวิธีแก้ไขข้อผิดพลาดสำหรับข้อผิดพลาดที่เหลือ โดยพิจารณาทีละหน้าต่าง (window) ทดสอบแต่ละ ตำแหน่ง ในหน้าต่างด้วยการทดสอบทวินาม (binomial test) เพื่อหาเบสที่กลายพันธุ์ (mutation) ซึ่งปรากฏมากอย่างมีนัยสำคัญ จากนั้น ใช้การทดสอบความแม่นตรงของฟิชเชอร์ (Fisher's exact test) ทดสอบทีละสองตำแหน่งเพื่อหาคู่ของเบสที่กลายพันธุ์ ซึ่งไปด้วยกัน นับจำนวนการกลายพันธุ์และคู่ของการกลายพันธุ์ที่ปรากฏอย่างมีนัยสำคัญและไม่ซ้ำกัน แบ่งสายลำดับที่อ่านได้ในหน้าต่างออกเป็นกลุ่มตามจำนวนการกลายพันธุ์และคู่ของการกลายพันธุ์ที่นับได้ด้วยวิธีเคมีนส์คลัสเตอร์ริง (k-means clustering) หาสายลำดับที่เป็นตัวแทนของแต่ละกลุ่ม (consensus sequence) และแก้ไขสายลำดับในแต่ละกลุ่มตามตัวแทนสายลำดับนั้น ดังรูปที่ 2.7

ทดสอบวิธีแก้ไขข้อผิดพลาด นี้โดยใช้โปรแกรม ReadSim จำลองข้อผิดพลาดจากกระบวนการ อ่านสายลำดับนิวคลีโอไทด์ด้วยเครื่องอ่านลำดับเบส Roche GS20 พบว่าเมื่อนำสายลำดับที่จำลองได้มาแก้ไขด้วยวิธีข้างต้น ข้อผิดพลาดลดลง 30 เท่า จากอัตราความผิดพลาดหลังจัดการกับการเติมหรือการลบออก (indel) เป็น 1-3 ตำแหน่งต่อพันเบส เมื่อผ่านขั้นตอนนี้แล้วลดลงเหลือ 0.1 ตำแหน่งต่อพันเบส



รูปที่ 2.7 การแก้ไขข้อผิดพลาด (error correction) จากรูปแสดงสายลำดับที่ต่างกันสองแบบคือเส้นสีเข้มและสีอ่อน แต่ละแบบมาจากแฮปโพลไทป์สองเส้นที่ต่างกัน ความแตกต่างทางพันธุกรรมแสดงด้วยจุดวงกลมและสี่เหลี่ยม พิจารณาทีละหน้าต่าง (แสดงด้วยกล่องเส้นประ) แบ่งกลุ่มสายลำดับในหน้าต่างตามขั้นตอนวิธีที่นำเสนอได้เป็นสองกลุ่ม จากนั้นแก้ไขข้อผิดพลาดซึ่งแสดงด้วยกากบาท [17]

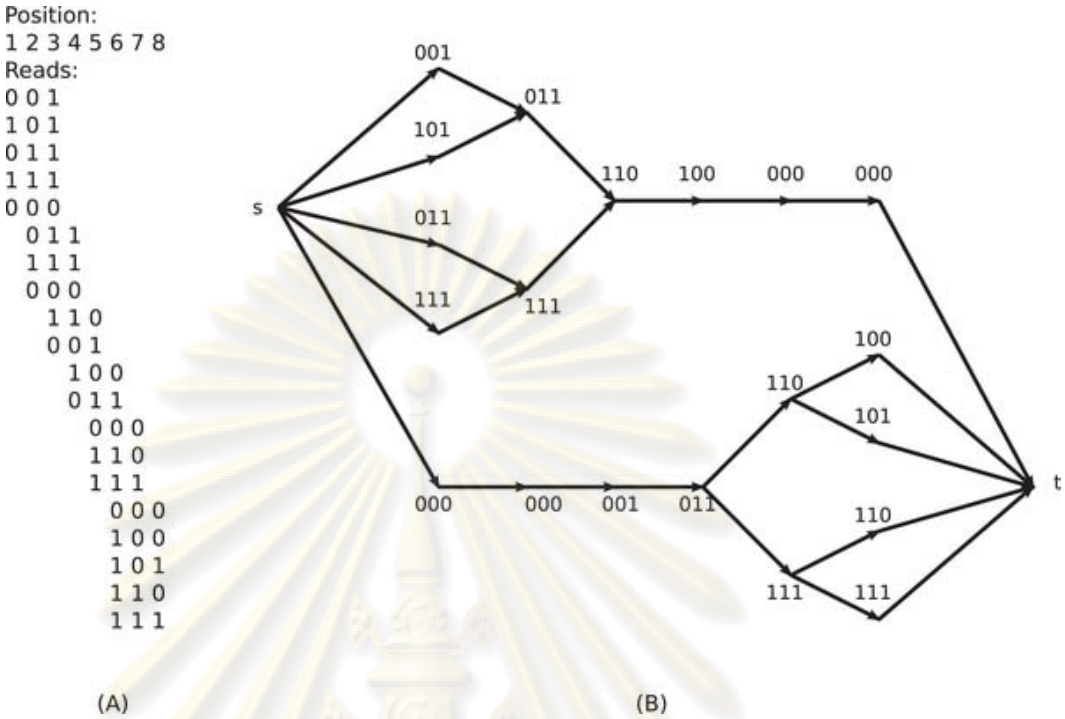
2.2.1.2 ขั้นตอนประกอบแฮปโพลไทป์

ในงานวิจัยนี้ต้องกา รชุดของแฮปโพลไทป์ที่ ประกอบด้วยแฮปโพลไทป์จำนวนน้อย ที่สุดที่สามารถอธิบายสายลำดับทั้งหมดได้ โดยสร้างกราฟของสายลำดับที่อ่านได้ (read graph) ซึ่งเป็นกราฟระบุทิศทางไม่มีวง (acyclic directed graph) ประกอบด้วยจุดยอด 3 แบบ คือ จุดเริ่มต้น จุดสิ้นสุด และจุดยอดสำหรับสายลำดับที่อ่านได้ทุกสายที่ไม่ซ้ำกัน ลากเส้นเชื่อมระหว่างจุดยอด โดยลากจากจุดยอด r_1 ไป r_2 เมื่อ จุดยอด r_1 มีตำแหน่งเริ่มต้นบนจีโนมก่อน r_2 ทั้งจุดยอด r_1 และ r_2 มีส่วนซ้อนทับ (overlap) ที่สอดคล้องกัน และยังไม่มีเส้นทางจาก r_1 ไป r_2 ยกเว้นเส้นเชื่อมนี้ สุดท้ายลากเส้นเชื่อมจากจุดเริ่มต้นไปยังจุดยอดซึ่งอยู่ตำแหน่งแรกสุด และลากเส้นเชื่อมจากจุดยอดที่อยู่ท้ายสุดไปยังจุดสิ้นสุด ดังตัวอย่างในรูปที่ 2.8

เมื่อได้กราฟของสายลำดับที่อ่านได้แล้ว หาเซตของเส้นทางจากจุดเริ่มต้นไป ยังจุดสิ้นสุด ซึ่งมีจำนวนเส้นทางน้อยที่สุดที่ยังครอบคลุมทุกจุดยอดในกราฟ ด้วย ขั้นตอนวิธีการจับคู่สูงสุด (maximum matching algorithm) ซึ่งเส้นทางที่ได้จะไม่เหมือนกันในแต่ละครั้ง แต่จำนวนสมาชิกของเซตที่เล็กที่สุดนี้จะเท่ากันทุกครั้ง แต่ละเส้นทางในเซตคือแฮปโพลไทป์ที่ประกอบได้

2.2.1.3 ขั้นตอนประมาณค่าความถี่แฮปโพลไทป์

สร้างแบบจำลองทางสถิติสำหรับกระบวนการอ่านสายลำดับ โดยสมมติว่าแต่ละสายลำดับที่อ่านได้ถูกสุ่มอ่านดังนี้ แฮปโพลไทป์ h ถูกสุ่มขึ้นมาด้วยการแจกแจงความน่าจะเป็น (probability



รูปที่ 2.8 ตัวอย่างกราฟของสายลำดับที่อ่านได้ ของจีโนมที่มีความยาว 8 คู่เบส แสดงสายลำดับด้วยเลขฐานสอง คือ 0, 1 สายลำดับที่อ่านได้จำนวน 20 เส้นแต่ละเส้นยาว 3 คู่เบส A) สายลำดับที่อ่านได้ถูกจัดเรียงเทียบกับสายลำดับอ้างอิง B) นำสายลำดับที่จัดเรียงแล้วไปสร้างกราฟ ซึ่งมี 22 จุดยอด (20 จุดยอดสำหรับสายลำดับที่อ่านได้ 20 เส้น, จุดเริ่มต้น s และจุดสิ้นสุด t) และ 28 เส้นเชื่อม [17]

distribution) ที่ไม่ทราบรูปแบบ p_h จากนั้นสายลำดับจะถูกสุ่มอ่านจากแฮปโพลไทป์นี้ด้วยความน่าจะเป็นแบบสม่ำเสมอ (uniform probability) จากแบบจำลองนี้จะได้ความน่าจะเป็นที่จะสุ่มอ่านได้สายลำดับ r ดังสมการ (1)

$$P(R = r) = \sum_{h \in H} p_h P(R = r | H = h) \text{.....(1)}$$

ประมาณค่าความถี่แฮปโพลไทป์ p_h โดยใช้ฟังก์ชันความควรจะเป็นลอการิทึม (log-likelihood function) ในสมการ (2) ด้วยขั้นตอนวิธีอีเอ็ม (EM algorithm) เมื่อ u_r คือจำนวนครั้งที่พบสายลำดับ r

$$l(p_1, \dots, p_{|H|}) = \sum_{r \in R} u_r \log P(R = r) \text{.....(2)}$$

งานวิจัยนี้ทดสอบวิธีที่นำเสนอโดยใช้สายลำดับที่ได้จากการจำลองการอ่าน จากนั้นทดสอบด้วยการประมาณค่าประชากรไวรัส เอชไอวี (HIV) 4 กลุ่มที่อิสระต่อกัน จากสายลำดับที่อ่านได้จาก

เทคนิค ไพโรซีเควนซิง (pyrosequencing) เทียบกับผลที่ได้จากการอ่านสายลำดับเบสด้วยวิธีของ Sanger จากกลุ่มตัวอย่างเดียวกัน จากการทดสอบปรากฏว่าสามารถใช้วิธีที่นำเสนอประมาณโครงสร้างประชากรของไวรัสได้

งานวิจัยนี้ใช้เฉพาะส่วนที่ซ้อนทับกัน (overlap) ในการประกอบแฮปโพลไทป์ ไม่ใช้ข้อมูลความถี่อัลลีลในการประกอบแฮปโพลไทป์ ต่างจากงานวิจัยที่จะนำเสนอซึ่งใช้ข้อมูลความถี่อัลลีลและส่วนที่ซ้อนทับกันเป็นข้อมูลหลักในการประกอบแฮปโพลไทป์

2.2.2 โครงการคิวแอสเซมเบลอร์ (Q Assembler)

โครงกา รคิวแอสเซมเบลอร์ [18] เกิดขึ้นเนื่องจากความสำคัญ ของการศึกษาจีโนมไวรัสซึ่งมีรูปแบบกึ่งสปีชีส์ (quasispecies) และความเป็นไปได้ในการใช้เทคโนโลยีอ่านสายลำดับเบสแบบขนานจำนวนมากศึกษาจีโนมของประชากรไวรัส ซึ่งโปรแกรมประกอบจีโนม (genome assembly) ที่มีในปัจจุบัน เช่น Phred/Phrap, TIGR Assembler หรือ 454's Newbler Assembler ถูกออกแบบเพื่อประกอบสายลำดับที่อ่านได้เข้าเป็นจีโนมเพียงเส้นเดียว จึงไม่เหมาะที่จะใช้ประกอบจีโนมของกลุ่มประชากรไวรัสซึ่งประกอบด้วยจีโนมหลายๆ เส้นที่คล้ายคลึงกัน โครงการนี้จึงมีวัตถุประสงค์หลักเพื่อพัฒนาซอฟต์แวร์สำหรับประกอบกลุ่มจีโนมของไวรัสจากสายลำดับที่ได้จากเทคโนโลยีอ่านสายลำดับเบสแบบขนานนี้

แนวคิดหลักของโครงการนี้คือ ควรใช้การประกอบแบบเทียบเคียง (comparative assembly) ในการประกอบกลุ่มจีโนมของไวรัส โดยสายลำดับจะถูกจัดเรียงเทียบกับสายอ้างอิงแทนที่จะใช้การประกอบจีโนมแบบ โอเวอร์แลปเลย์เอาท์คอนเซนซัส (overlap-layout–consensus) ซึ่งเป็นวิธีมาตรฐานในการประกอบจีโนมทั่วไป คณะผู้วิจัยเสนอให้ดัดแปลงวิธีการประกอบแบบเทียบเคียง ในขั้นตอนการแบ่งส่วนตามวิวัฒนาการชาติพันธุ์ (phylogenetic partitioning) โดยให้สายลำดับที่อ่านได้ซึ่งเป็นอินพุต จัดเรียงเทียบกับสายลำดับที่เป็นตัวแทนจากแต่ละบรรพบุรุษหลัก จากนั้น จัดเรียงแต่ละกลุ่มของสายลำดับที่อ่านได้ใหม่เทียบกับสายลำดับรองของบรรพบุรุษหลักนั้น จากการศึกษาเบื้องต้นพบว่าวิธีนี้ประสบความสำเร็จในการแยกสายลำดับที่อ่านได้เป็นกลุ่มๆ ซึ่งสามารถเป็นตัวแทนของจีโนมต้นแบบได้ และคาดว่าการประกอบจีโนมในขั้นสุดท้ายสามารถเกิดขึ้นได้

ระเบียบวิธีที่พัฒนานี้ถูกทดสอบ โดยใช้อินพุตเป็นสายลำดับจากเครื่องอ่านสายลำดับเบส GS20 ของจีโนมไวรัสเอชไอวี ทั้งเส้น 2 ตัว เทียบจีโนมที่ประกอบได้กับโคลน (clone) ที่ได้จากวิธีอ่านสายลำดับเบสของ Sanger

โครงการนี้ยังไม่มีผลงานตีพิมพ์ และ ซอฟต์แวร์ที่พัฒนาขึ้น ยังอยู่ในขั้นก่อนรุ่นทดสอบรุ่นแรก (pre-alpha) ยังไม่เผยแพร่

2.2.3 งานวิจัยด้านการประกอบแฮปโพลไทป์

มีงานวิจัยด้าน ชีวสารสนเทศ (bioinformatics) จำนวนมาก ที่เกี่ยวข้องกับการประกอบแฮปโพลไทป์ โดยงานวิจัยส่วนใหญ่ได้นำเสนอขั้นตอนวิธีและวิธีเชิงสถิติต่างๆ เพื่อใช้ในการประกอบแฮปโพลไทป์ แต่งานวิจัยเหล่านี้ได้นำเสนอวิธีสำหรับการประกอบแฮปโพลไทป์ของสิ่งมีชีวิตที่มีโครโมโซมสองคู่ โดยมุ่งเน้นที่แฮปโพลไทป์ของคนเป็นหลัก นอกจากนี้ยังมีงานวิจัยที่แสดงให้เห็นถึงการใช้ประโยชน์จากแฮปโพลไทป์ในการจำแนกความแตกต่างและทำนายความเสี่ยงต่อการเกิดโรคหรือความผิดปกติต่างๆ [19, 20, 21, 22, 23]

Pierre-Yves Boëlle [9] รวบรวมรายละเอียดวิธีการเชิงสถิติพื้นฐานสำหรับการประกอบแฮปโพลไทป์ ได้แก่ วิธีความควรจะเป็นสูงสุด (maximum likelihood method) โดยใช้ขั้นตอนวิธีอีเอ็ม (expectation maximization algorithm) วิธีแบบเบส์ (Bayesian method) และยกตัวอย่างเครื่องมือในการประกอบแฮปโพลไทป์ที่มีอยู่ในปัจจุบันที่ใช้วิธีแบบเบส์คือ PHASE และ Haplotyper

Matthew Stephens และคณะ[11] ได้เสนอวิธีการทางสถิติแบบใหม่ซึ่งให้ข้อมูลที่สมบูรณ์ยิ่งขึ้น และลดอัตราความผิดพลาดลงกว่า 50% โดยสนใจที่ข้อมูลของประชากรและ พิจารณาแต่ละตำแหน่งที่เชื่อมโยงกันบนจีโนไทป์ด้วย ต่อมาได้ปรับปรุงวิธีการนี้และใช้ใน โปรแกรม สำเร็จ (software package) ที่ชื่อ PHASE และในปี 2003 Matthew Stephens และ Peter Donnelly [10] ได้เปรียบเทียบขั้นตอนวิธีที่พัฒนาขึ้นซึ่งได้บรรจุไว้ในโปรแกรมสำเร็จ PHASE เวอร์ชัน 2.0 กับวิธีอื่นที่มีพื้นฐานจากวิธีแบบเบส์ซึ่งมีอยู่ในขณะนั้น 3 วิธี ผลปรากฏว่าวิธีที่พัฒนาขึ้นเหนือกว่าทั้งจากการทดสอบด้วยข้อมูลจริงและข้อมูลที่จำลองขึ้น

นอกจากวิธีการเชิงสถิติที่มีพื้นฐานจากวิธีแบบเบส์แล้วยังมีการนำเสนอแบบจำลองของมาร์คอฟ (Markov model) ในการประกอบแฮปโพลไทป์ด้วย เช่น L. Eronen และคณะ [24] ได้เสนอแบบจำลองมาร์คอฟ อย่างง่ายสำหรับประกอบแฮปโพลไทป์โดยมุ่งเน้นสำหรับสายแฮปโพลไทป์ที่ยาว โดยอาศัยความไม่สมดุลของความเชื่อมโยง (linkage disequilibrium : LD) ระหว่างสนิปส์ที่อยู่ใกล้กัน วิธีการที่นำเสนอนี้มีความยืดหยุ่น ทนทาน สามารถใช้ได้กับทั้งข้อมูลจริงและข้อมูลที่จำลองขึ้น นอกจาก L. Eronen แล้ว Pasi Rastas และคณะ [25]ได้เสนอเทคนิคฮิดเดนมาร์คอฟ (Hidden Markov) สำหรับการประกอบแฮปโพลไทป์ โดยมองแฮปโพลไทป์เป็นผลลัพธ์ของการรวมกันเป็นรอบๆ (iterated recombination) จากแฮปโพลไทป์ตั้งต้น สร้างแบบจำลองความควรจะเป็นสูงสุดด้วยขั้นตอนวิธีอีเอ็ม จากนั้นเปรียบเทียบผลกับ PHASE, HAP และ GERBIL ภายใต้มาตรฐานที่กำหนดและชุดข้อมูลใหม่ ซึ่งขั้นตอนวิธีที่เสนอเร็วกว่าและให้ผลลัพธ์ที่ดีเป็นอันดับ 1 หรือ 2

Yu-Ying Zhao และคณะ [5] ได้เสนอขั้นตอนวิธีสำหรับการประกอบแฮปโพลไทป์โดยใช้เทคนิคการจับกลุ่มแบบพลวัต (dynamic clustering) และเสนอวิธีการแก้ปัญหาใหม่สำหรับการประกอบแฮปโพลไทป์ตั้งชื่อว่า CWMLF โดยเทคนิคและวิธีที่นำเสนอนี้ตั้งอยู่บนข้อมูลและระเบียบวิธีของการประกอบสายลำดับแบบชอตกัน (shotgun sequence assembly)

งานวิจัยที่นำเสนอนี้แตกต่างจากงานวิจัยในการประกอบแฮปโพลไทป์ข้างต้นเนื่องจากสาเหตุหลัก 2 ประการ คือ 1) เทคโนโลยีที่ใช้ ซึ่งให้รูปแบบของข้อมูลสายลำดับเบสที่อ่านได้แตกต่างไปจากเดิม และ 2) รูปแบบจีโนไทป์ของไวรัสเด็งกี่ที่แตกต่างจากมนุษย์และสิ่งมีชีวิตที่มีโครโมโซมสองชุด อย่างไรก็ตาม การประยุกต์ใช้วิธีการประกอบแฮปโพลไทป์เหล่านี้กับสิ่งมีชีวิตกึ่งสปีชีส์ทำได้ยาก เนื่องจากเราไม่รู้จำนวนแฮปโพลไทป์ของประชากรสิ่งมีชีวิตกึ่งสปีชีส์

2.2.4 งานวิจัยด้านการประยุกต์ใช้เทคโนโลยีการอ่านสายลำดับนิวคลีโอไทด์แบบขนานจำนวนมาก

เนื่องจากเทคโนโลยีพิโกไทเตอร์เพลตไพโรซีเควนซิง (picotiter plate pyrosequencing) ซึ่งเป็น รูปแบบหนึ่งของ เทคโนโลยีอ่านสายลำดับ นิวคลีโอไทด์แบบขนานจำนวนมากและเป็นเทคโนโลยี ที่ใช้ในงาน วิจัยนี้ เป็นเทคโนโลยีใหม่ สามารถอ่านสายลำดับนิวคลีโอไทด์ได้จำนวนมาก ในเวลาอันสั้น มีต้นทุนต่อเบสที่อ่านได้ต่ำ แต่มีรูปแบบแตกต่างไปจากเทคโนโลยีอ่านสายลำดับนิวคลีโอไทด์แบบเดิม จึงมีงานวิจัยเพื่อศึกษารูปแบบการทำงาน รูปแบบของสายลำดับที่อ่านได้ ประสิทธิภาพและการประยุกต์ใช้เทคโนโลยีนี้ เช่น งานวิจัย [13] ได้เสนอรายละเอียดเทคโนโลยีนี้ โปรโตคอลที่ใช้ รูปแบบผลลัพธ์ที่ได้ รวมถึงการนำไปใช้งาน โดยเสนอการประกอบจีโนมสำหรับ ไมโคพลาสมาเจนิทาเลียม (Mycoplasma genitalium) ซึ่งมีความ ครอบคลุม (coverage) ถึง 96% และมีความแม่น (accuracy) สูงถึง 99.96%

ต่อมาในงานวิจัย [14] ได้ใช้เทคโนโลยี พิโกไทเตอร์เพลตไพโรซีเควนซิง (picotiter plate pyrosequencing) ที่ปรับปรุงล่าสุด ในเครื่องอ่านสายลำดับ Roche GS FLX ทำให้อ่านสายลำดับได้ยาว 250 คู่เบส โดยเฉลี่ย จำนวน 400,000 เส้นต่อครั้ง ในงานวิจัยได้กล่าวถึงประสิทธิภาพของเทคโนโลยีใหม่ล่าสุด ความแตกต่าง ข้อดี ข้อเสีย เมื่อเทียบกับวิธี อ่านสายลำดับเบสแบบชอตกัน (shotgun sequencing) และเสนอ SHARP (Short Read Assembly Protocol) ซึ่งเป็นโปรโตคอลของสายลำดับและระเบียบวิธีในการประกอบจีโนมเพื่อให้สามารถใช้ประโยชน์จากเทคโนโลยีนี้ได้เต็มประสิทธิภาพที่สุด ในตอนท้ายได้ทดสอบ SHARP โดยใช้ข้อมูล D. melanogaster และโครโมโซมมนุษย์คู่ที่ 1, 11 และ 21 ซึ่งให้ผลลัพธ์ที่ถูกต้อง

Iman และคณะ [26] ยังได้นำเสนอความเป็นไปได้ในการอ่านสายลำดับซ้ำ (resequencing) จากชิ้นส่วนดีเอ็นเอที่นำมารวมกันแล้วส่งไปอ่านด้วยเทคโนโลยีการอ่านลำดับนิวคลีโอไทด์แบบขนานจำนวนมาก พร้อมทั้งเสนอวิธีที่เหมาะสมที่สุดในการรวมชิ้นส่วนดีเอ็นเอ จากหลายๆ จีโนม ที่

ต้องการหาสายลำดับนิวคลีโอไทด์ เพื่อส่งเข้าไปอ่านด้วยเครื่องอ่านสายลำดับนิวคลีโอไทด์พร้อมกันในการเดินเครื่องครั้งเดียว

บทที่ 3
วิธีการดำเนินงาน

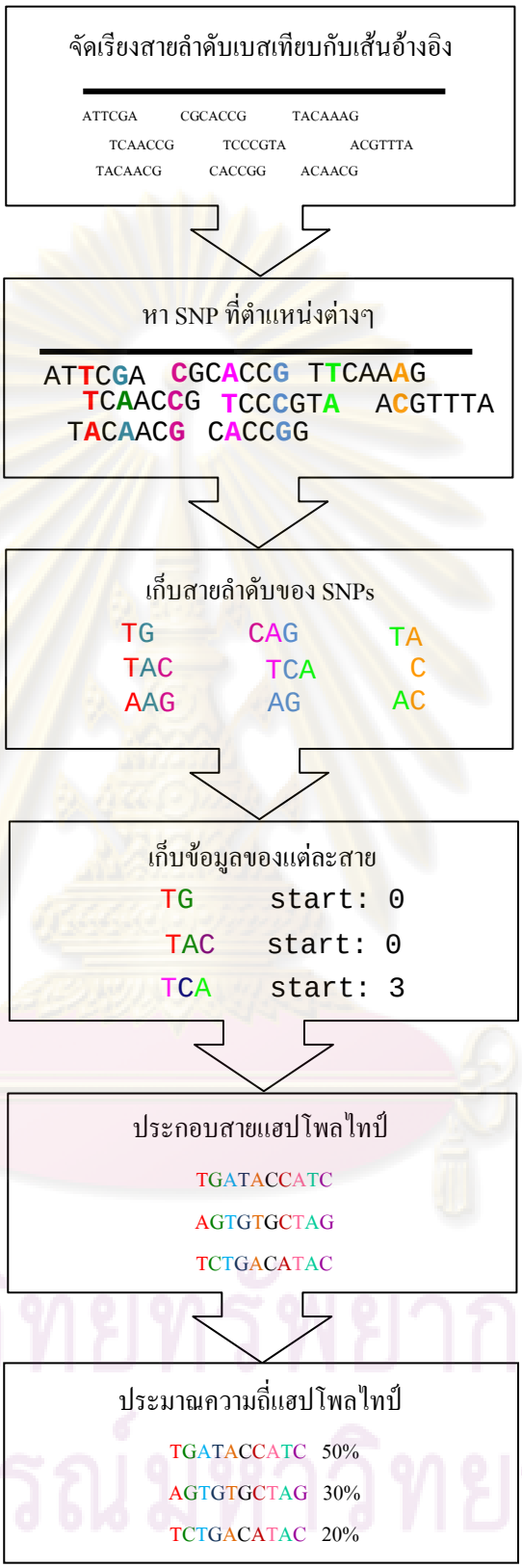
3.1 ภาพรวมของงานวิจัย

งานวิจัยนี้ต้องการหาขั้นตอนวิธีประกอบชุดของสายแฮปโพลไทป์ของเชื้อไวรัสเด็งกี่และความถี่ของแต่ละสาย จากข้อมูลชิ้นส่วนของสายลำดับที่อ่านได้ (read) จากเครื่องอ่านสายลำดับ Roche GS FLX โดยตัวอย่างที่ส่งเข้าไปอ่านนั้นคือกลุ่มประชากรของไวรัสเด็งกี่ซึ่งประกอบด้วยจีโนมที่มีรูปแบบแตกต่างกันรวมกันอ่านในครั้งเดียว เพื่อศึกษาความแปรผัน (variation) ของไวรัสเด็งกี่ โดยสมมติฐานของงานวิจัยนี้คือ อินพุตของขั้นตอนวิธีที่นำเสนอ คือ สายลำดับที่อ่านได้ที่ผ่านการจัดการความผิดพลาด (error correction) แล้ว ไม่ปรากฏความผิดพลาด และสามารถจัดเรียงกับสายต้นแบบได้ถูกต้อง นั่นคือ สามารถทราบตำแหน่งของสายลำดับนั้นบนจีโนมต้นแบบได้ โดยมีขั้นตอนหลักของขั้นตอนวิธีประกอบแฮปโพลไทป์และประมาณความถี่ของแฮปโพลไทป์ คือ จัดเรียงสายลำดับที่เป็นอินพุตเทียบกับเส้นอ้างอิง หาสนิปส์ที่ตำแหน่งต่างๆ สร้างสายลำดับของสนิปส์จากสนิปส์ที่อยู่บนสายลำดับที่อ่านได้เส้นเดียวกัน เก็บข้อมูลของแต่ละสาย ได้แก่ รูปแบบสายลำดับเบส และตำแหน่ง ของเบสตัวแรกของสายลำดับนั้น จากนั้นนำไปประกอบแฮปโพลไทป์และประมาณความถี่แฮปโพลไทป์ ดังรูปที่ 3.1

อย่างไรก็ตาม การประกอบชุดของแฮปโพลไทป์เป็นปัญหาที่ซับซ้อน และในช่วงแรกของการวิจัยยังไม่มีงานวิจัยเกี่ยวกับการศึกษาความหลากหลายทางพันธุกรรมของสิ่งมีชีวิตที่มีรูปแบบกึ่งสปีชีส์ด้วยเทคโนโลยีอ่านลำดับนิวคลีโอไทด์แบบขนานจำนวนมากมาก่อน ในงานวิจัยนี้จึงแบ่งการดำเนินงานเป็น 4 ขั้นตอน คือ

1. ศึกษาความเป็นไปได้ในการใช้ เทคโนโลยีอ่านสายลำดับนิวคลีโอไทด์แบบขนานจำนวนมาก ศึกษาความหลากหลายทางพันธุกรรมของเชื้อไวรัสเด็งกี่
2. ประกอบสายแฮปโพลไทป์หลักและประมาณความถี่แฮปโพลไทป์สายหลัก
3. ประกอบชุดของสายแฮปโพลไทป์และประมาณความถี่ของแต่ละสาย
4. เปรียบเทียบประสิทธิภาพของวิธีที่นำเสนอกับวิธีที่เสนอโดย Eriksson และคณะ [17]

ในขั้นตอนแรกได้ศึกษาความเป็นไปได้ในการใช้เทคโนโลยีอ่านสายลำดับนิวคลีโอไทด์แบบขนานจำนวนมากศึกษาความหลากหลายทางพันธุกรรมของเชื้อไวรัสเด็งกี่ โดยมุ่งศึกษาประสิทธิภาพของเครื่องอ่านสายลำดับเบส Roche GS20 และ GS FLX ซึ่งเป็นผู้นำของเทคโนโลยีนี้และเป็นเครื่องที่ใช้ในศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ (BIOTEC) โดยพิจารณา



รูปที่ 3.1 ขั้นตอนหลักในการประกอบแฮปโพลไทป์และประมาณความถี่แฮปโพลไทป์ เริ่มจากจัดเรียงสายลำดับ หาสนิปส์ สร้างสายลำดับของสนิปส์ เก็บข้อมูล ประกอบแฮปโพลไทป์และประมาณความถี่แฮปโพลไทป์

ความเป็นไปได้ของเทคโนโลยีใน 4 ประเด็นหลักคือ ความถูกต้องในการจัดเรียงสายลำดับที่อ่านได้เทียบกับสายต้นแบบ ความครอบคลุมของสายลำดับที่อ่านได้ ความไวในการอ่านส่วนที่มีการผันแปร และความสัมพันธ์ระหว่างความถี่ของสายลำดับที่อ่านได้กับความถี่ของสายลำดับตั้งต้น ดังรายละเอียดในบทที่ 4

ออกแบบวิธีการประกอบแฮปโพลไทป์โดยพิจารณาเฉพาะการประกอบแฮปโพลไทป์สายหลักก่อนเพื่อลดความซับซ้อนของปัญหา จากนั้นนำวิธีที่ได้จากกการประกอบแฮปโพลไทป์สายหลักนี้ไปใช้ในการประกอบชุดของแฮปโพลไทป์ ซึ่งได้กล่าวถึงรายละเอียดของวิธีประกอบแฮปโพลไทป์สายหลัก การทดสอบและผลการทดสอบไว้ในบทที่ 5 ส่วนรายละเอียดการประกอบชุดของแฮปโพลไทป์ ได้แก่ ขั้นตอนวิธี การทดสอบ และผลการทดสอบได้กล่าวถึงในบทที่ 6 จากนั้นทดสอบวิธีที่นำเสนอในงานวิจัยนี้เปรียบเทียบกับวิธีที่นำเสนอโดย Eriksson และคณะ ในบทความตีพิมพ์เรื่องการประมาณประชากรไวรัสโดยใช้ไพโรซีเควนซิง (Viral population estimation using pyrosequencing) ซึ่งได้แสดงวิธีการทดสอบและผลการทดสอบไว้ในบทที่ 7

3.2 การจำลองข้อมูลสำหรับทดสอบวิธีประกอบแฮปโพลไทป์และประมาณความถี่แฮปโพลไทป์

สำหรับการทดสอบวิธีประกอบแฮปโพลไทป์ที่นำเสนอและวิธีที่เสนอโดย Eriksson และคณะนั้น [17] จะทดสอบด้วยข้อมูลที่จำลองขึ้น โดยแบ่งเป็น 2 ขั้นตอนหลักคือ 1) จำลองประชากรของไวรัสเด็งกี่ หรือชุดจีโนมของไวรัสเด็งกี่ ซึ่งประกอบด้วยจีโนมของไวรัสเด็งกี่ที่ต่างกันหลายๆ เส้น และความถี่ของแต่ละเส้น 2) จำลองการอ่านสายลำดับด้วยเครื่องอ่านสายลำดับ Roche GS FLX นำสายลำดับที่อ่านได้มาเป็นอินพุตของวิธีที่นำเสนอ มีรายละเอียดดังนี้

3.2.1 จำลองประชากรของไวรัสเด็งกี่ มีขั้นตอนดังนี้

1) สุ่มเลือกจีโนมไวรัสเด็งกี่จากฐานข้อมูล GenBank มา 1 เส้น

2) กำหนดความถี่ของสายลำดับหลักอย่างสุ่มตามช่วงที่กำหนด โดยการทดลองนี้แบ่งความถี่ของสายลำดับหลักเป็น 9 ช่วง คือ ร้อยละ 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89 และ 90-99

3) จำลองสายลำดับอื่นโดยสุ่มเลือกตำแหน่งที่เกิดการกลายพันธุ์ร้อยละ 2 ของความยาวจีโนมทั้งหมด (ประมาณ 200 – 220 คู่เบส จาก 10,000 -11,000 คู่เบส) ตามร้อยละการแปรผันของไวรัสเด็งกี่ที่มีการรายงานไว้

4) แต่ละตำแหน่งเลือกอัลลีลแบบสุ่ม โดยการทดลองนี้กำหนดให้แต่ละตำแหน่งมีเพียง 2 อัลลีล ได้แก่อัลลีลเดียวกับสายลำดับหลักและอัลลีลที่ได้จากการสุ่มเลือก ซึ่งเป็นรูปแบบการแปรผันที่พบบ่อยที่สุด

5) กำหนดความถี่ของสายลำดับที่จำลองขึ้น

6) ทำซ้ำข้อ 3) – 5) จนความถี่ของสายลำดับรวมครบ 100%

3.2.2 จำลองการอ่านข้อมูลด้วยเครื่องอ่านลำดับเบส Roche GS FLX ดังขั้นตอนต่อไปนี้

1) สุ่มเลือกสายลำดับโดยถ่วงน้ำหนักตามความถี่ของแต่ละสายลำดับ

2) สำหรับสายลำดับที่เลือก สุ่มตำแหน่งเริ่มต้นและอ่านจนครบความยาว 200-300 คู่เบส

3) ทำซ้ำจนกว่าจะได้จำนวนสายลำดับครบจำนวน ในที่นี้จำลองข้อมูลให้มีจำนวนสายลำดับที่อ่านได้แตกต่างกันไปทั้งสิ้น 8 ค่า คือ 500, 1000, 3000, 5000, 10000, 30000, 50000 และ 100000 เส้น

ข้อมูลที่จำลองขึ้นสำหรับใช้ทดสอบวิธีที่นำเสนอนี้จะมีความถี่ของสายลำดับหลักแตกต่างกันออกไปทั้งสิ้น 9 ช่วง และในแต่ละช่วงจะอ่านสายลำดับโดยกำหนดจำนวนสายลำดับที่อ่านได้แตกต่างกันไป 8 ค่า โดยจำลองข้อมูลทั้งสิ้น 10 รอบ จึงได้ชุดของสายลำดับที่แตกต่างกัน 720 ชุด สำหรับทดสอบวิธีที่นำเสนอ

3.3 ขั้นตอนการทดสอบวิธีประกอบแฮปโพลไทป์และประมาณความถี่แฮปโพลไทป์

จากสายลำดับที่ได้จากการจำลองการอ่านสายลำดับด้วยเครื่อง Roche GS FLX นำมาหาตำแหน่งที่มีสนิปส์ จากนั้นเก็บเฉพาะสายลำดับของสนิปส์ และตำแหน่งของสนิปส์ตัวแรกบนสายลำดับนั้น นำไปเป็นอินพุตสำหรับขั้นตอนวิธีประกอบแฮปโพลไทป์และประมาณความถี่แฮปโพลไทป์ที่นำเสนอ จากนั้นรายงานผลการทดสอบประสิทธิภาพเป็นความแม่นของสายลำดับแฮปโพลไทป์ที่ประกอบขึ้น เทียบกับประชากรของไวรัสเด็งกี่ที่จำลองขึ้น แสดงขั้นตอนการทดสอบได้ดังรูปที่ 3.2

รูปที่ 3.2 ขั้นตอนการทดสอบประสิทธิภาพของขั้นตอนวิธีที่นำเสนอ

แต่ในการทดสอบวิธีที่นำเสนอโดย Eriksson และคณะจะนำข้อมูลสายลำดับที่จำลองขึ้นและตำแหน่งของสายลำดับนั้นบนจีโนมไปแปลงให้อยู่ในรูปแบบที่เหมาะสมก่อน ดังรายละเอียดในบทที่ 7 ข้อ 7.2 นำอินพุตที่ได้จากการแปลงชุดของสายลำดับที่จำลองขึ้นทั้ง 720 ชุดไปประกอบแฮปโพลไทป์และประมาณความถี่แฮปโพลไทป์ จากนั้นรายงานความแม่นของขั้นตอนวิธีเช่นเดียวกับการทดสอบวิธีที่นำเสนอ

บทที่ 4
การศึกษาความเป็นไปได้ในการใช้เทคโนโลยีอ่านสายลำดับนิวคลีโอไทด์แบบขนานจำนวนมากศึกษาความหลากหลายทางพันธุกรรมของไวรัสเด็งกี่

ในกลุ่มของเทคโนโลยีอ่านสายลำดับนิวคลีโอไทด์แบบขนานจำนวนมาก เทคโนโลยีที่เป็นผู้นำในด้านนี้คือ เทคโนโลยีพิโกไทเตอร์เพลตไพโรซีเควนซิง (picotiter plate pyrosequencing) ในเครื่องอ่านลำดับเบส Roche GS 20 และ Roche GS FLX ซึ่งพัฒนาโดยกลุ่ม 454 Life Sciences เนื่องจากเทคโนโลยี นี้เป็นเทคโนโลยีใหม่และยังไม่ถูกใช้ในการศึกษาความหลากหลายของไวรัสเด็งกี่ซึ่งมีลักษณะกึ่งสปีชีส์ จึงต้องศึกษาความเป็นไปได้ในการนำเทคโนโลยีนี้มาใ ช้กับเชื้อไวรัสเด็งกี่ โดยแบ่งการทดสอบเป็น 2 ส่วน คือ ศึกษาโดยใช้ข้อมูลที่จำลองขึ้นและศึกษาโดยใช้ข้อมูลที่อ่านได้จากเครื่อง อ่านลำดับเบส โดยใช้ข้อมูลที่อ่านได้จากเครื่องอ่านลำดับเบสจากหน่วยชีวสารสนเทศและจัดการข้อมูลวิจัย ซึ่งเป็นผู้ออกแบบและทำการทดลองในห้องปฏิบัติการเพื่อศึกษาความเป็นไปได้นี้

4.1 ศึกษาความเป็นไปได้โดยใช้ข้อมูลสายลำดับที่จำลองขึ้น

ในขณะที่ทดลองเครื่องรุ่นล่าสุดที่ใช้เทคโนโลยี พิโกไทเตอร์เพลตไพโรซีเควนซิง คือ เครื่องอ่านลำดับเบส Roche GS20 ซึ่งอ่านสายลำดับได้ยาว 100 คู่เบส จำนวน 250,000 เส้น และกระบวนการเพิ่มจำนวนดีเอ็นเอด้วยเทคนิคพีซีอาร์ที่ใช้สามารถจำลองสายลำดับได้ที่ความยาว 2,000 - 3,000 คู่เบส จึงจำลองข้อมูลขึ้นตามลักษณะของเครื่อง อ่านลำดับเบส Roche GS20 และเทคนิคพีซีอาร์ที่ใช้ในขณะนั้น โดยทดสอบ 2 ประเด็นหลักคือ

4.1.1 ความถูกต้องในการจัดเรียง (alignment) มีขั้นตอนการทดสอบดังนี้

1) สุ่มจีโนมจากฐานข้อมูล GenBank ขึ้นมา 1 เส้น

2) จำลองการอ่านสายลำดับให้มีความยาวเช่นเดียวกับที่ได้จากเครื่อง Roche GS20 คือประมาณ 100 คู่เบส โดยอ่านให้ทุกๆ ตำแหน่งได้เป็นตำแหน่งเริ่ม ทั้งเส้นบวกและเส้นลบ

3) จัดเรียง (align) สายลำดับที่จำลองได้เทียบกับแม่แบบ (template) ที่มีอยู่ คือ จีโนมของไวรัสเด็งกี่ที่มีในฐานข้อมูล GenBank ทั้งหมดยกเว้นจีโนมที่ถูกสุ่มมาเป็นตัวทดสอบ ด้วย BLAST

4) แต่ละสายลำดับ เลือกแม่แบบที่เหมาะสมที่สุด คือ มีค่าความยาวที่จัดเรียงได้ (alignment length) ยาวที่สุดและเหมือนกับแม่แบบมากที่สุด เลือกตามร้อยละของความเหมือน (% identity) สูงสุด

5) ตรวจสอบว่าแม่แบบที่เหมาะสมที่สุดอยู่ในตำแหน่งเดียวกับสายลำดับที่จำลองขึ้นมาหรือไม่ และเป็นซีโรไทป์ (serotype) เดียวกันหรือไม่

ผลที่ได้ คือ สายลำดับที่จำลองขึ้นมาสามารถจัดเรียงได้ถูกต้องทั้งซีโรไทป์และตำแหน่ง

4.1.2 ความครอบคลุม (coverage) มีขั้นตอนดังนี้

1) เลือกจีโนมจากฐานข้อมูล GenBank ขึ้นมา 1 เส้น

2) แบ่งจีโนมเป็นสายลำดับย่อย 5 สาย แต่ละสายยาว 2,300 คู่เบส ตามการทดลองในห้องวิจัยที่ต้องแบ่งออกเป็นสายลำดับย่อยก่อนเข้าสู่กระบวนการเพิ่มจำนวนดีเอ็นเอด้วยเทคนิคพีซีอาร์ (PCR: Polymerase Chain Reaction) ซึ่งการแบ่งสายลำดับนี้ทำให้มีส่วนที่คาบเกี่ยวกัน (overlap) ระหว่างสายลำดับย่อย

3) อ่านสายลำดับย่อยเป็นท่อนสั้นๆ ท่อนละประมาณ 100 คู่เบส โดยสุ่มตำแหน่งเริ่มต้นและอ่านต่อไปจนได้ความยาว 100 คู่เบส หรือสุดสาย จำนวน 10,000 ท่อน (เครื่อง Roche GS20 อ่านได้ 200,000 เส้นต่อครั้ง แต่เราต้องการอ่านพร้อมกันหลายๆ ตัวใน 1 ครั้ง จึงจำลองการอ่านที่ 10,000 ท่อน)

4) นำสายลำดับที่อ่านได้มาสร้างกราฟของความครอบคลุม โดยแกนนอน คือ ตำแหน่งบนจีโนม และแกนตั้ง คือ จำนวนครั้งที่อ่านตำแหน่งนั้น

ผลที่ได้ สามารถอ่านสายลำดับนิวคลีโอไทด์ได้ ครอบคลุมทุกตำแหน่ง โดย แต่ละตำแหน่งจะถูกอ่านเฉลี่ยเท่าๆ กัน ยกเว้นส่วนที่มีการซ้อนทับ (overlap) จากการแบ่งสายลำดับย่อย จะอ่านได้เป็น 2 เท่า

4.2 ศึกษาความเป็นไปได้โดยใช้ข้อมูลสายลำดับที่อ่านได้จากเครื่องอ่านลำดับเบส Roche GS FLX

เนื่องจากศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ (BIOTEC) ได้ยกระดับเครื่องอ่านลำดับเบสจาก Roche GS20 มาเป็น Roche GS FLX ซึ่งเป็นรุ่นที่พัฒนาต่อมาจาก GS20 ทำให้อ่านสายลำดับเบสได้ยาวขึ้นและจำนวนมากขึ้น คือ ยาวเฉลี่ย 200-300 คู่เบส และอ่านได้มากถึง 400,000 เส้นต่อครั้ง ในการทดลองนี้จึงเปลี่ยนมาใช้เครื่องอ่านลำดับเบส Roche GS FLX

การทดลองนี้ออกแบบมาเพื่อพิสูจน์ความเที่ยงและความไว (sensitivity) ในการอ่านสายลำดับเบส โดยเตรียมสายดีเอ็นเอที่จะส่งเข้าไปอ่านลำดับเบสจากเส้นต้นแบบ (prototype strain) และโคลน (clone) ที่รู้ตำแหน่งการแปรผัน (variation) แล้ว นำรวมกันในอัตราส่วนที่กำหนดไว้ โดยซีโรไทป์ (serotype) 1,3 และ 4 ใช้เฉพาะเส้นต้นแบบเพียง 1 เส้น ในอัตราส่วนซีโรไทป์ละ 18% ส่วนซีโรไทป์ 2 นำเส้นต้นแบบและโคลนมารวมกันให้มีอัตราส่วนรวม 46% ดังตารางที่ 4.1

ตารางที่ 4.1 อัตราส่วนของปริมาณไวรัสเด็งกี่ทั้ง 4 ซีโรไทป์ก่อนส่งไปหาลำดับนิวคลีโอไทด์

dengue type	Strain	Percent mixed RNA (Only serotype 2) (%)	Percent mixed purified DNA (%)
1	Hawaii	-	18
2	16681	10.0	46
	wild type 16681	29.9	
	NGC	5.0	
	N130A	1.0	
	N207A	0.1	
3	H87	-	18
4	H241	-	18
รวม		46.0	100

นำชุดของสายดีเอ็นเอที่เตรียมจากดีเอ็นเอของไวรัสเด็งกี่ทั้ง 4 ซีโรไทป์มารวมกันตามอัตราส่วนตามที่กำหนด ไปอ่านด้วยเครื่อง อ่านลำดับเบส Roche GS FLX ได้ข้อมูลทั้งสิ้น 54,416 สาย นำข้อมูลที่ได้นี้มาตรวจหาความผิดพลาด (Error Detection) เบื้องต้น จากนั้น นำข้อมูลไปจัดเรียง (align) กับสายลำดับของเส้นต้นแบบ (prototype strain) ด้วยโปรแกรม BLAST นำผลที่ได้มาวิเคราะห์ในประเด็นต่างๆ ต่อไปนี้

4.2.1 ความถูกต้อง พิจารณาจาก

- ความสามารถในการจับคู่กับเส้นแม่แบบ จาก BLAST ได้ผลดังตารางที่ 4.2

ตารางที่ 4.2 จำนวนสายลำดับที่อ่านได้แยกตามซีโรไทป์

ซีโรไทป์	จำนวนสายลำดับที่อ่านได้	
	(เส้น)	(%)
1	10028	18.43
2	30035	55.20
3	10591	19.46
4	3724	6.84
ไม่สามารถแยกได้	38	0.07
รวม	54416	100.00

จะเห็นว่า ซีโรไทป์ 1 และ 3 มีอัตราส่วนของสายลำดับเบสที่อ่านได้ใกล้เคียงกับที่เตรียมไว้ คือ ร้อยละ 18 และมีซีโรไทป์ 2 มากที่สุด เช่นเดียวกับตัวอย่างสายลำดับที่เตรียม แต่ซีโรไทป์ 2 มีเปอร์เซ็นต์มากกว่าที่เตรียม ในขณะที่มีซีโรไทป์ 4 น้อยที่สุด คาดว่ามีสาเหตุจากขั้นตอนการเตรียมสารตั้งต้นก่อนเข้าเครื่องอ่านสายลำดับนิวคลีโอไทด์

- เปอร์เซ็นต์ความแปรผัน แม้ว่าเราจะรู้สายลำดับของเส้นต้นแบบและตั้งใจ เตรียมตัวอย่าง ให้เป็นไปตามตาราง 1 แต่ในความเป็นจริงเส้นแม่แบบที่เรานำมาเตรียม นั้นยังมีความแปรผันซึ่งเราไม่ทราบอยู่ เราจึงทดลองและรายงานความแปรผันเหล่านี้

เมื่อนำข้อมูลที่อ่านได้จากเครื่องมาจัดการความผิดพลาดและจัดเรียงด้วย BLAST จากนั้นนำผลที่ได้มาแยกซีโรไทป์ และพิจารณาความแปรผันแยกทีละซีโรไทป์ หาเปอร์เซ็นต์ความแปรผันหรือสัดส่วนของตำแหน่งที่มีความแปรผันเทียบกับตำแหน่งทั้งหมด ได้ผลดังตารางที่ 4.3

จะเห็นว่าความแปรผันโดยเฉลี่ยประมาณร้อยละ 2 ซึ่งไม่ขัดแย้งกับคุณลักษณะของไวรัสเด็งกี่ และ มีความแปรผันมากในซีโรไทป์ 2 ซึ่งเป็นซีโรไทป์เดียวที่ เตรียมตัวอย่างจาก เส้นต้นแบบรวมกับโคลนทั้งหมด 5 เส้น

ตารางที่ 4.3 ความแปรผันในแต่ละซีโรไทป์

ซีโรไทป์	ความแปรผัน (%)
1	1.65
2	6.88
3	1.58
4	0.46
เฉลี่ย	2.64

4.2.2 ความครอบคลุม (coverage)

นำผลจาก BLAST มาพิจารณาแต่ละตำแหน่งบนสายดีเอ็นเอของทุกซีโรไทป์ จากนั้นคำนวณเปอร์เซ็นต์ความครอบคลุม เป็นสัดส่วนระหว่าง ตำแหน่งที่ถูกอ่านต่อตำแหน่งทั้งหมด จากผลการทดลอง สายลำดับที่อ่านได้และผ่านตัวกรองความผิดพลาดแล้ว มีความครอบคลุมสูง โดยเฉลี่ยสูงถึง 99.66% ดังตารางที่ 4.4

ตารางที่ 4.4 ความครอบคลุมตำแหน่งบนสายดีเอ็นเอ

ซีโรไทป์	ความครอบคลุม (%)
1	99.67
2	99.67
3	99.60
4	99.74
เฉลี่ย	99.66

4.2.3 ความสามารถในการอ่านส่วนที่มีการผันแปร

พิจารณาจากซีโรไทป์ 2 ซึ่งเรารู้ตำแหน่งที่มีการผันแปรของสายดีเอ็นเอที่ใช้เตรียมตัวอย่าง ดังตารางที่ 4.5

ตารางที่ 4.5 ตำแหน่งที่มีการผันแปรของสายดีเอ็นเอในไวรัสเด็งกี่ซีโรไทป์ 2 ที่ส่งไปอ่านลำดับนิวคลีโอไทป์

Position	16681 prototype	wild type 16681
403	Insert	C
498	G	A
2943	A	C
4308	C	T
4530	A	G
8155	G	A
8571	C	T
10331	A	G
10561	A	C
Position	16681 prototype	mutant N130A
2809	AAC	GCG
Position	16681 prototype	mutant N207A
3040	AAT	GCT

ในการทดลองนี้พิจารณาเฉพาะความแปรผันที่เกิดจากการเปลี่ยนเบส ดังนั้นจะยังไม่พิจารณาที่ตำแหน่ง 403 เมื่อพิจารณาข้อมูลของซีโรไทป์ 2 จาก BLASTเฉพาะตำแหน่งที่รู้ความแปรผันแล้วได้ผลดังตารางที่ 4.6

ตารางที่ 4.6 ความผันแปรในสายดีเอ็นเอของไวรัสเด็งกี่ซีโรไทป์ 2 ในตำแหน่งที่รู้ความแปรผันของชุดตัวอย่างที่ส่งไปหาสายลำดับ เปรียบเทียบระหว่างตัวอย่างที่ส่งไปอ่านและข้อมูลที่อ่านได้

ตำแหน่ง	ข้อมูลตัวอย่างสายดีเอ็นเอ				ข้อมูลที่อ่านได้			
	อัลลีลหลัก (major allele)		อัลลีลรอง (minor allele)		อัลลีลหลัก (major allele)		อัลลีลรอง (minor allele)	
	เบส	ความถี่ (%)	เบส	ความถี่ (%)	เบส	ความถี่ (%)	เบส	ความถี่ (%)
498	A	78.26	G	21.74	G	53.74	A	46.26
2809	A	97.83	G	2.17	A	98.33	C/G	0.21 / 1.46
2810	A	97.83	C	2.17	A	98.54	C	1.46
2811	C	97.83	G	2.17	C	98.95	G	1.05
2943	C	78.26	A	21.74	C	63.53	A/G	30.50/5.96
3040	A	99.78	G	0.22	A	100.00	-	-
3041	A	99.78	C	0.22	A	100.00	-	-
3042	T	99.78	T	0.22	T	100.00	-	-
4308	T	78.26	C	21.74	T	61.56	C / A	37.40/1.04
4530	G	78.26	A	21.74	G	76.39	A	23.61
8155	A	78.26	G	21.74	A	64.14	G / C	35.74 / 0.12
8571	T	78.26	C	21.74	T	65.67	C	34.33
10331	G	78.26	A	21.74	G	80.43	A	19.57
10561	C	78.26	A	21.74	C	70.00	A	30.00

จากตารางจะเห็นว่าเครื่อง อ่านลำดับเบส Roche GS FLX สามารถรายงานความแปรผันได้ยกเว้นตำแหน่ง 3040 ถึง 3042 ที่ไม่พบความแปรผัน เนื่องจากความแปรผันน้อยมาก เพียงร้อยละ 0.22

4.2.4 ความถี่ชุดตัวอย่างตั้งต้นมีผลต่อความถี่ที่อ่านได้

เนื่องจากในงานวิจัยนี้จะใช้ความถี่อัลลีลในการประกอบสายลำดับ และรายงานความถี่ของแต่ละสายลำดับ ดังนั้น ความถี่ที่อ่านได้จากเครื่องต้องเป็นไปตามที่มีอยู่จริง จากผลการทดลองพบว่าความถี่ที่อ่านได้สอดคล้องกับความถี่ที่มีในชุดตัวอย่าง เห็นได้จากอัตราส่วนสายลำดับแยกตามซีโรไทป์ (ตารางที่ 4.2) และความผันแปรในซีโรไทป์ 2 (ตารางที่ 4.6)

จากการทดลองพบว่า สามารถใช้เทคโนโลยีอ่านสายลำดับนิวคลีโอไทด์แบบพิโคไทเตอร์เพลตไพโรซีเควนซิง ในการศึกษาความหลากหลายทางพันธุกรรมของไวรัสเด็งกี่ได้

บทที่ 5
การประกอบสายแฮปโพลไทป์หลักและประมาณความถี่แฮปโพลไทป์สายหลัก

การประกอบแฮปโพลไทป์ทั้งหมดสำหรับกลุ่มประชากรไวรัสเป็นปัญหาที่ซับซ้อน ดังนั้นในขั้นแรกจึงมุ่งประกอบเฉพาะแฮปโพลไทป์สายหลัก คือ แฮปโพลไทป์ที่พบมากที่สุดในกลุ่มประชากรนั้น หรือกล่าวได้ว่ามีความถี่แฮปโพลไทป์สูงสุด และประมาณค่าความถี่แฮปโพลไทป์สายหลักที่ประกอบขึ้นได้ โดยได้ออกแบบขั้นตอนวิธีและพัฒนาขึ้นเป็น 3 รูปแบบหลักคือ

1. การประกอบแฮปโพลไทป์สายหลักจากสายลำดับที่อ่านได้ที่มีความถี่สูงสุด โดยประกอบไปตามลำดับ
2. การประกอบแฮปโพลไทป์สายหลักจากสายลำดับที่อ่านได้ที่มีความถี่สูงสุด โดยสุ่มตำแหน่ง
3. การประกอบแฮปโพลไทป์จากอัลลีลหลัก (major allele) หรืออัลลีลที่พบบ่อยที่สุดในแต่ละตำแหน่ง

5.1 การประกอบแฮปโพลไทป์สายหลักจากสายลำดับที่อ่านได้ที่มีความถี่สูงสุด โดยประกอบไปตามลำดับ

การประกอบแฮปโพลไทป์สายหลักด้วยวิธีนี้จะเลือกสายลำดับที่อ่านได้ที่มีความถี่สูงสุดในแต่ละตำแหน่งมาประกอบกันไปเรื่อยๆ โดยเริ่มจากตำแหน่งแรกสุดไล่ไปจนถึงตำแหน่งสุดท้าย จากนั้นประมาณค่าความถี่แฮปโพลไทป์สายหลักด้วยค่าเฉลี่ยเลขคณิตและมัธยฐาน ของความถี่อัลลีลแต่ละตำแหน่ง

5.1.1 ขั้นตอนวิธี

ข้อมูลอินพุตสำหรับขั้นตอนวิธี นี้คือ สายลำดับย่อยของสนิปส์ และตำแหน่งของสนิปส์ตัวแรกบนสายแฮปโพลไทป์ ซึ่งได้จากการเก็บข้อมูลสายลำดับสนิปส์จากสายลำดับ นิวคลีโอไทด์ที่อ่านได้ ตามวิธีเก็บสายลำดับสนิปส์ในบทที่ 3 นำอินพุตนี้มาเข้าสู่ขั้นตอนการประกอบแฮปโพลไทป์สายหลัก ดังนี้

1) เรียงสายลำดับสนิปส์ทั้งหมดตามตำแหน่งเริ่มต้น

2) รวมสายลำดับสนิปส์ที่เหมือนกัน คือ มีสายลำดับเบสเหมือนกันและอยู่ในตำแหน่งเดียวกันเข้าด้วยกัน เช่น มีสายลำดับสนิปส์ที่อ่านได้ดังนี้

AAATTT	เริ่มต้นที่ตำแหน่ง	0
AAATTT	เริ่มต้นที่ตำแหน่ง	0

AAATTT เริ่มต้นที่ตำแหน่ง 0

AAATTT เริ่มต้นที่ตำแหน่ง 0

AAATTT เริ่มต้นที่ตำแหน่ง 0

AACTTG เริ่มต้นที่ตำแหน่ง 0

AACTTG เริ่มต้นที่ตำแหน่ง 0

AACTTG เริ่มต้นที่ตำแหน่ง 0

CTTG เริ่มต้นที่ตำแหน่ง 1

AACTTG เริ่มต้นที่ตำแหน่ง 5

สามารถรวมได้เป็น

1. AAATTT เริ่มต้นที่ตำแหน่ง 0 5 เส้น
2. AACTTG เริ่มต้นที่ตำแหน่ง 0 3 เส้น
3. CTTG เริ่มต้นที่ตำแหน่ง 1 1 เส้น
4. AACTTG เริ่มต้นที่ตำแหน่ง 5 1 เส้น

3) ในแต่ละตำแหน่ง รวมสายลำดับสั้นๆ ที่สามารถรวมเข้ากับสายลำดับเส้นยาวได้เข้าด้วยกัน นับจำนวนอัลลีลของแต่ละตำแหน่ง เช่น ตำแหน่ง เริ่มต้นที่ 0 มีสายลำดับของสนิปส์ 3 ชุด คือ

1. AAATTT 5 สาย
2. AACTTG 3 สาย
3. AAAT 1 สาย

สามารถรวมได้เป็น

I. AAATTT จำนวนอัลลีล 6,6,6,6,5,5

II. AACTT จำนวนอัลลีล 3,3,3,3,3

จากข้อมูลสายลำดับสนิปส์นี้ รวมได้เป็น 2 ชุด คือ ชุดที่ I มีสายลำดับเป็น AAATTT ซึ่งมีจำนวนของ A ทั้งสามตัว (ตำแหน่ง 0-2) และ T ตัวแรกซึ่งอยู่ที่ตำแหน่ง 3 ของสายลำดับนี้ เป็น 6 จากการรวมกันของสายลำดับที่ 1 และ 3 และ T สองตัวสุดท้ายมีจำนวนอัลลีลเป็น 5 จากสายลำดับที่ 1 ส่วนอีกชุด (ชุดที่ II) คือ AACTT ได้จากสายลำดับที่ 2 ซึ่งไม่สามารถรวมกับเส้นอื่นได้ มีจำนวนแต่ละอัลลีลในสายลำดับเป็น 3

4) คำนวณความถี่อัลลีล โดยนำจำนวนอัลลีลแต่ละตำแหน่งในแต่ละชุดสายลำดับที่รวมไว้หารด้วยจำนวนอัลลีลทั้งหมดของตำแหน่งนั้นๆ เช่น จากตัวอย่างด้านบน ความถี่อัลลีลของ A ในชุดที่ I คือ 0.67 (จาก 6 ÷ (6+3)) และความถี่อัลลีลของ C ใน II. คือ 0.33 (จาก 3 ÷ (6+3))

5) ประกอบสายแฮปโพลไทป์ โดยเริ่มจากตำแหน่งเริ่มต้นที่น้อยที่สุด เลือกสายลำดับที่มีความถี่อัลลีลเฉลี่ยสูงสุดมา ประกอบเป็นสายลำดับลัพธ์ จากนั้นพิจารณาตำแหน่งเริ่มต้นลำดับถัดไป เลือกสายลำดับที่มีความถี่เฉลี่ยสูงสุดมาพิจารณาว่าสามารถต่อกับสายลำดับลัพธ์ที่มีอยู่ได้หรือไม่ โดยพิจารณา จากส่วนที่ซ้อนทับกัน (overlap) ถ้าได้ คือ ทุกๆ ตำแหน่งที่เหลื่อมกันมีเบสชนิดเดียวกันทั้งหมด ให้นำสายลำดับนั้น ต่อกับสายลำดับ ลัพธ์ที่มีอยู่ พร้อมทั้ง รวมความถี่อัลลีลแต่ละตำแหน่ง ของสายลำดับที่เลือกเข้ากับ สายลำดับลัพธ์ แต่ถ้าไม่ได้ให้เลือกสายลำดับที่มีความถี่รองลงมามาพิจารณา ทำเช่นนี้ไปเรื่อยๆ จนสุดสาย

6) ประมาณค่าความถี่ แฮปโพลไทป์ (haplotype frequency) ด้วยค่าเฉลี่ยเลขคณิต และค่ามัธยฐานจากความถี่อัลลีลแต่ละตำแหน่งบนสายลำดับลัพธ์ที่ได้

5.1.2 การทดสอบ

5.1.2.1 จำลองข้อมูลสำหรับทดสอบขั้นตอนวิธี

ทดสอบความถูกต้องของขั้นตอนวิธีโดยจำลองข้อมูลให้มีลักษณะคล้ายกับสายลำดับของไวรัสเด็งกี่ที่อ่านได้จากเครื่องอ่านลำดับเบส Roche GS FLX โดยใช้ข้อมูลจีโนมจากฐานข้อมูลของ GenBank เป็นต้นแบบ แล้วจำลองประชากรและการอ่านสายลำดับเบสตามขั้นตอนการจำลองสายลำดับที่อ่านได้ในบทที่ 3 โดยให้ความถี่ของแฮปโพลไทป์สายหลักและจำนวนสายลำดับที่อ่านได้แตกต่างกัน คือ จำลองให้มีความถี่ของแฮปโพลไทป์สายหลักทั้งหมด 9 ช่วง คือ ร้อยละ 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89 และ 90-99 ในแต่ละช่วงความถี่กำหนดให้มีจำนวนสายลำดับที่อ่านได้ 8 ค่า คือ 500, 1000, 3000, 5000, 10000, 30000, 50000 และ 100000 เส้น ทำซ้ำทั้งสิ้น 10 รอบ รวมเป็นข้อมูลทั้งสิ้น 720 ชุด

ในแต่ละชุดจะได้ สายลำดับที่อ่านได้ซึ่งจำลองการอ่านเลียนแบบการอ่านด้วยเครื่องอ่านสายลำดับ Roche GS FLX มาเป็นอินพุตของขั้นตอนวิธีที่นำเสนอ และเก็บข้อมูลแฮปโพลไทป์สายหลักและความถี่ของแฮปโพลไทป์สายหลักที่จำลองขึ้นไว้สำหรับเปรียบเทียบผลกับสายลำดับแฮปโพลไทป์ที่ประกอบได้จากขั้นตอนวิธีที่นำเสนอ

5.1.2.2 ทดสอบ

ทดสอบทีละชุด โดยนำสายลำดับที่อ่านได้ในแต่ละชุดมาเป็นอินพุตของขั้นตอนวิธีการประกอบแฮปโพลไทป์สายหลัก และประมาณค่าความถี่ของแฮปโพลไทป์สายหลัก เก็บข้อมูลสายลำดับลัพธ์และความถี่ที่ได้ของแต่ละชุดมาทดสอบประสิทธิภาพของขั้นตอนวิธี

5.1.2.3 รายงานผล

ในแต่ละชุด นำสายลำดับ ลัพธ์ที่ได้ไปเปรียบเทียบกับแฮปโพลไทป์สายหลักที่จำลองขึ้นซึ่งเป็นเสมือนสายลำดับจริงที่คาดหวังจากขั้นตอนวิธี รายงานความแม่น (accuracy) โดยวัดเป็นร้อยละ

ของตำแหน่งที่ถูกต้อง จากนั้นเปรียบเทียบความถี่ ที่ประมาณได้กับความถี่ของ แฮปโพลไทป์ สายหลัก รายงานผลด้วยความผิดพลาดสัมบูรณ์ (absolute error) ซึ่งคำนวณจากความแตกต่างระหว่างความถี่ที่คำนวณได้ กับความถี่หลักที่กำหนด หรือ ค่าสัมบูรณ์ของความแตกต่างระหว่างความถี่ที่คำนวณได้กับความถี่จริง

นำผลที่ได้จากทุกชุดมาสรุปรายงานความแม่นของแฮปโพลไทป์สายหลักและความผิดพลาดสัมบูรณ์ของความถี่แฮปโพลไทป์สายหลัก แยกตามช่วงความถี่ของสายลำดับหลัก

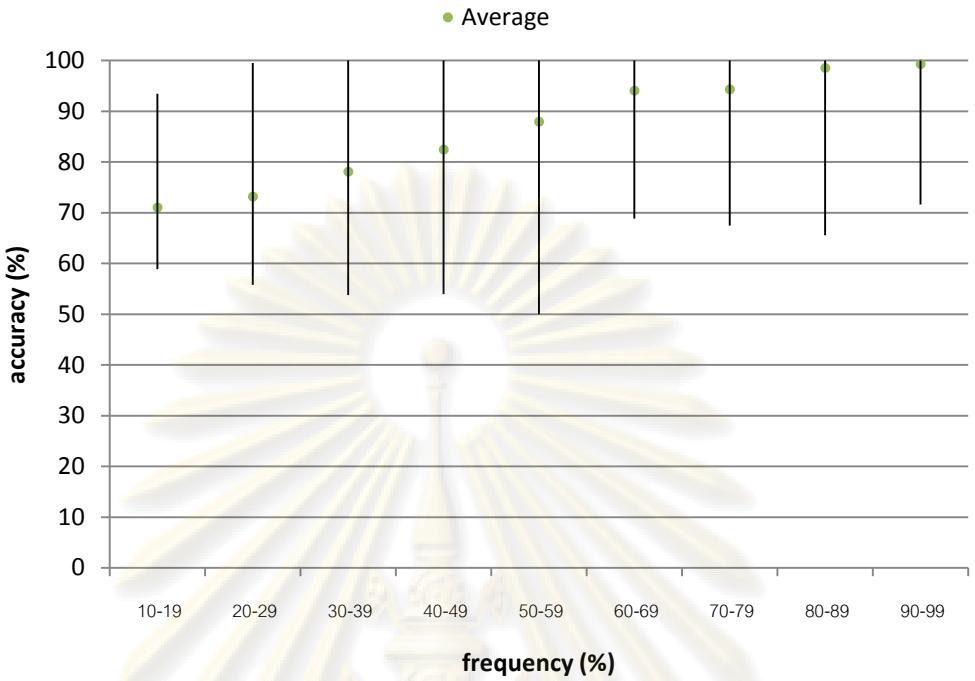
5.1.3 ผลการทดลอง

จากการทดลองประกอบ แฮปโพลไทป์ สายหลักจากชุดของสายลำดับที่มีความแตกต่างกันประมาณร้อยละ 2 และมี 2 อัลลีลในแต่ละตำแหน่ง ได้ความแม่นของสาย แฮปโพลไทป์ ที่ประกอบได้ดังตารางที่ 5.1

ตารางที่ 5.1 ความแม่นของสายแฮปโพลไทป์หลักที่ประกอบจากสายลำดับที่อ่านได้ที่มีความถี่สูงสุด โดยการประกอบไปตามลำดับ

ช่วงความถี่สายลำดับหลัก (%)	ความแม่นสูงสุด (%)	ความแม่นต่ำสุด (%)	ความแม่นเฉลี่ย (%)
10-19	93.43	58.88	71.04
20-29	99.53	55.81	73.19
30-39	100.00	53.74	78.14
40-49	100.00	53.95	82.48
50-59	100.00	50.00	87.96
60-69	100.00	68.84	94.07
70-79	100.00	67.44	94.33
80-89	100.00	65.58	98.55
90-99	100.00	71.63	99.34
ทั้งหมด	**100.00**	**50.00**	**86.57**

นำความแม่นที่ได้ แยกตามช่วงความถี่ของสายลำดับหลัก มาสร้างกราฟได้ดังรูปที่ 5.1 จากผลการทดลองจะเห็นว่า ความแม่นเพิ่มขึ้นเมื่อความถี่สายลำดับหลัก เพิ่มขึ้น โดยความแม่นเฉลี่ยของชุดข้อมูลที่สายลำดับหลักมีความถี่ 90-99 % คือร้อยละ 99.34 และความแม่นเฉลี่ยทั้งหมดคือร้อยละ 86.57



รูปที่ 5.1 กราฟความแม่นของสายแฮปโพลไทป์หลักที่ประกอบได้ ด้วยการประกอบสายลำดับที่อ่านได้ที่มีความถี่สูงสุด แบบประกอบตามลำดับ แยกตามช่วงความถี่ของสายลำดับหลักที่จำลองขึ้น วัดความแม่นจากร้อยละของตำแหน่งที่ มีอัลลีลตรงกับเส้นต้นแบบ จุดในกราฟแสดงความแม่นเฉลี่ยของสายแฮปโพลไทป์และช่วงในกราฟแสดงช่วงความแม่นของสายแฮปโพลไทป์ที่ประกอบได้จากความแม่นต่ำสุดถึงความแม่นสูงสุด

จากการประมาณค่าความถี่ของสาย แฮปโพลไทป์ หลักด้วยค่าเฉลี่ยเลขคณิตและมัธยฐานของความถี่อัลลีลแต่ละตำแหน่งของสายลำดับที่ใช้ในการประกอบแฮปโพลไทป์สายหลัก รายงานผลด้วยความผิดพลาดสัมบูรณ์ของความถี่แฮปโพลไทป์ที่คำนวณได้ แสดงได้ดังตารางที่ 5.2 จาก ตาราง การประมาณความถี่ของแฮปโพลไทป์ด้วยมัธยฐานให้ผลดีกว่าการประมาณด้วยค่าเฉลี่ยเลขคณิตเนื่องจากความถี่แฮปโพลไทป์ที่คำนวณได้ใกล้เคียงกับความถี่แฮปโพลไทป์จริงมากกว่า จากชุดข้อมูลที่ทดลองนี้ ความถี่ที่ได้จากการ ประมาณความถี่ด้วยมัธยฐานแตกต่างจากความถี่แฮปโพลไทป์จริงเพียงร้อยละ 0.58 สำหรับชุดข้อมูลที่สายลำดับหลักมีความถี่ 90-99 % และมีความผิดพลาดสัมบูรณ์เฉลี่ยร้อยละ 2.33 ในขณะที่การประมาณความถี่แฮปโพลไทป์ด้วยค่าเฉลี่ยเลขคณิต มีความผิดพลาดสัมบูรณ์ของความถี่แฮปโพลไทป์ที่คำนวณได้เฉลี่ยร้อยละ 3.48

ตารางที่ 5.2 ความผิดพลาดสัมบูรณ์ของความถี่แฮปโพลไทป์สายหลักที่คำนวณได้ จากการประกอบแฮปโพลไทป์สายหลักด้วยสายลำดับที่อ่านได้ที่มีความถี่สูงสุด โดยการประกอบไปตามลำดับ เปรียบเทียบระหว่างการประมาณค่าความถี่แฮปโพลไทป์ด้วยค่าเฉลี่ยเลขคณิตและมัธยฐาน

ช่วงความถี่สายลำดับหลัก (%)	ความผิดพลาดสัมบูรณ์จากการประมาณความถี่ด้วยค่าเฉลี่ยเลขคณิต (%)			ความผิดพลาดสัมบูรณ์จากการประมาณความถี่ด้วยค่ามัธยฐาน (%)		
	สูงสุด	ต่ำสุด	เฉลี่ย	สูงสุด	ต่ำสุด	เฉลี่ย
10-19	0.04	10.40	2.55	0.01	10.67	2.52
20-29	0.07	8.35	2.53	0.01	9.31	2.28
30-39	0.02	11.73	3.36	0.02	18.25	3.11
40-49	0.19	17.22	4.33	0.01	22.18	3.38
50-59	0.11	21.60	3.99	0.05	28.21	2.59
60-69	0.18	34.55	5.36	0.00	53.54	2.57
70-79	0.01	33.66	5.76	0.00	54.75	2.01
80-89	0.01	46.57	2.31	0.01	80.10	1.71
90-99	0.01	42.21	1.16	0.01	2.82	0.58
ทั้งหมด	**0.01**	**46.57**	**3.48**	**0.00**	**80.10**	**2.33**

5.2 การประกอบแฮปโพลไทป์สายหลักจากสายลำดับที่อ่านได้ที่มีความถี่สูงสุด โดยสุ่มตำแหน่ง

เนื่องจากการประกอบแฮปโพลไทป์สายหลักตามวิธี 5.1 จะมีความลำเอียง (bias) จากสายลำดับในตำแหน่งที่อยู่ก่อน โดยเฉพาะตำแหน่งเริ่มต้นซึ่งในทางปฏิบัติตำแหน่งหัวและตำแหน่งท้ายของสายลำดับเป็นตำแหน่งที่มีโอกาสเกิดความผิดพลาดสูง ในวิธีนี้จึงสุ่มตำแหน่งแล้วเลือกสายลำดับที่อ่านได้ที่มีความถี่สูงสุดในตำแหน่งนั้นมาประกอบเป็นแฮปโพลไทป์สายหลัก โดยมีวิธีพิจารณาเช่นเดียวกับวิธี 5.1 ประกอบจนครบทั้งสาย จากนั้นประกอบซ้ำโดยสุ่มตำแหน่งใหม่ ทำซ้ำเช่นนี้ตามจำนวนรอบที่กำหนด นำผลลัพธ์ที่ได้ทั้งหมดมาเลือกเส้นที่มีความถี่แฮปโพลไทป์สูงที่สุด

5.2.1 ขั้นตอนวิธี

สำหรับวิธีนี้มีอินพุตเช่นเดียวกับวิธี 5.1.1 และมีขั้นตอนวิธีคล้ายคลึงกันคือ การรวมสายลำดับสนิปส์และการคำนวณความถี่อัลลีลในขั้นตอนที่ 1)-4) ทำเช่นเดียวกับวิธี 5.1.1 แต่ในขั้นตอนการประกอบแฮปโพลไทป์

1) – 4) ทำเช่นเดียวกับวิธี 5.1.1 ได้แก่ จัดเรียงสายลำดับสนิปส์ตามตำแหน่งเริ่มต้น รวมสายลำดับสนิปส์ที่เหมือนกันเข้าด้วยกัน รวมสายลำดับสนิปส์เส้นสั้นๆ เข้ากับเส้นยาวที่มีตำแหน่งเริ่มต้นเดียวกัน คำนวณความถี่อัลลีลของสายลำดับสนิปส์แต่ละเส้น

5) ประกอบสาย แฮปโพลไทป์ โดยสุ่มตำแหน่งขึ้นมา เลือกสายลำดับ ในตำแหน่งนั้น ที่มีความถี่อัลลีลเฉลี่ยสูงสุดมาประกอบเป็นสายลำดับลัพธ์ จากนั้นสุ่มตำแหน่งถัดไป เลือกสายลำดับในตำแหน่งนั้น ที่มีความถี่เฉลี่ยสูงสุดมาพิจารณาว่าสามารถต่อกับสายลำดับลัพธ์ที่มีอยู่ได้หรือไม่ โดยพิจารณา จากส่วนที่ซ้อนทับกัน (overlap) ถ้าไม่มีตำแหน่งใดที่ต่างกัน ให้นำสาย ต่อกับสายลำดับลัพธ์ที่มีอยู่ พร้อมทั้ง รวมความถี่อัลลีลแต่ละตำแหน่ง ของสายลำดับที่เลือกเข้ากับ สายลำดับลัพธ์ แต่ถ้าไม่ได้ให้เลือกสายลำดับ ในตำแหน่งนั้น ที่มีความถี่รองลงมามาพิจารณา จากนั้นสุ่มเลือกตำแหน่งถัดไป ทำเช่นนี้ไปเรื่อยๆ จนสุดสาย

6) ประมาณค่าความถี่ แฮปโพลไทป์ (haplotype frequency) ด้วยค่ามัธยฐาน จากความถี่อัลลีลแต่ละตำแหน่งบนสายลำดับลัพธ์ที่ได้

7) ทำซ้ำข้อ 1) ถึง 6) ตามจำนวนรอบที่กำหนด นำสายลำดับทั้งหมดที่ได้มาเลือกสายลำดับที่มีความถี่แฮปโพลไทป์สูงที่สุดเป็นผลลัพธ์ของขั้นตอนวิธีนี้

5.2.2 การทดสอบ

ทดสอบความถูกต้องของขั้นตอนวิธี ด้วยความแม่นของแฮปโพลไทป์สายหลักที่ประกอบได้และความแตกต่างของความถี่แฮปโพลไทป์สายหลักที่คำนวณได้เทียบกับความถี่จริง โดยใช้ข้อมูลเดียวกับการทดสอบ 5.1.2 โดยแต่ละชุดข้อมูลกำหนดจำนวนรอบในการทำซ้ำขั้นตอน 1) – 6) ทั้งสิ้น 10 รอบ

5.2.3 ผลการทดลอง

จากการทดลองประกอบ แฮปโพลไทป์ สายหลักจากชุดของสายลำดับ ชุดเดียวกับที่ใช้ทดสอบขั้นตอนวิธี 5.1 ได้ความแม่นของสายแฮปโพลไทป์ที่ประกอบได้ดังตารางที่ 5.3

ตารางที่ 5.3 ความแม่นของสายแฮปโพลไทป์หลักที่ประกอบจากสายลำดับที่อ่านได้ที่มีความถี่สูงสุด โดยสุ่มตำแหน่ง

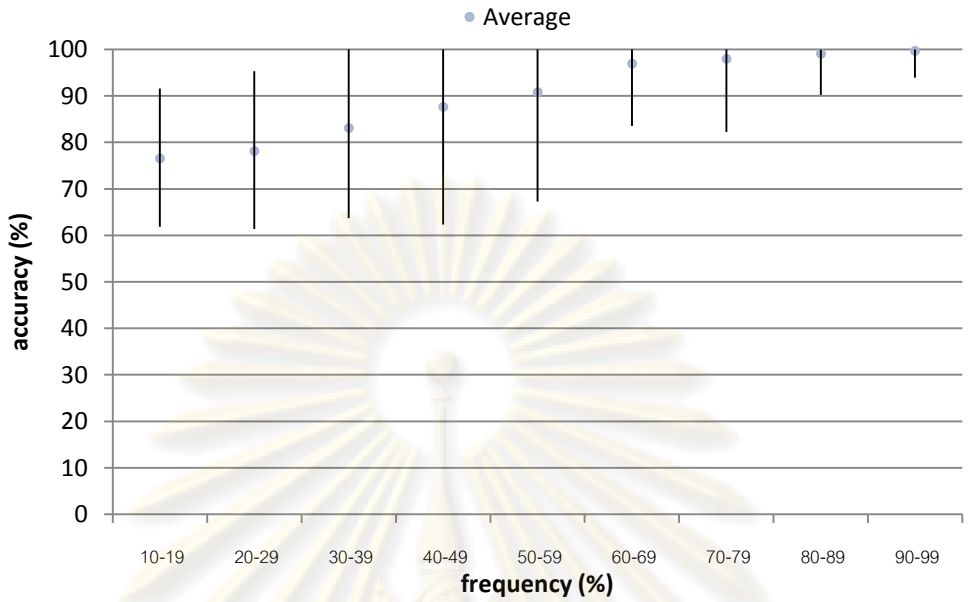
ช่วงความถี่สายลำดับหลัก (%)	ความแม่นสูงสุด (%)	ความแม่นต่ำสุด (%)	ความแม่นเฉลี่ย (%)
10-19	91.63	61.86	76.61
20-29	95.35	61.40	78.18
30-39	100.00	63.72	83.17
40-49	100.00	62.33	87.72
50-59	100.00	67.29	90.90
60-69	100.00	83.57	97.03
70-79	100.00	82.24	98.02
80-89	100.00	90.23	99.14
90-99	100.00	93.95	99.73
ทั้งหมด	**100.00**	**61.40**	**90.05**

สร้างกราฟแสดงความแม่นของแฮปโพลไทป์สายหลักที่ประกอบได้ แยกตามช่วงความถี่ของสายลำดับหลักมาสร้างกราฟได้ดังรูปที่ 5.2

จากผลการทดลองจะเห็นว่า ความแม่นเพิ่มขึ้นเมื่อความถี่สายลำดับหลัก เพิ่มขึ้น โดยความแม่นเฉลี่ยของชุดข้อมูลที่สายลำดับหลักมีความถี่ 90-99 % คือร้อยละ 99.73 และความแม่นเฉลี่ยทั้งหมดคือร้อยละ 90.05

จากนั้นประมาณค่าความถี่ของสาย แฮปโพลไทป์ หลักจากค่ามัธยฐานของ ความถี่อัลลีลแต่ละตำแหน่ง วัดความผิดพลาดสัมบูรณ์ของความถี่ที่คำนวณได้เทียบกับความถี่จริงของแฮปโพลไทป์สายหลักที่เป็นอินพุตของขั้นตอนวิธีนี้ ได้ผลดังตารางที่ 5.4

จาก ตาราง การประมาณความถี่ของแฮปโพลไทป์สายหลักของชุดข้อมูลนี้ด้วยค่ามัธยฐาน มีความผิดพลาดสัมบูรณ์ เฉลี่ยร้อยละ 1.54 สำหรับชุดข้อมูลทั้งหมด และ มีความผิดพลาดสัมบูรณ์เฉลี่ยร้อยละ 0.57 สำหรับชุดข้อมูลที่สายลำดับหลักมีความถี่ 90-99 %



รูปที่ 5.2 กราฟความแม่นของ สายแฮปโพลไทป์ หลักที่ประกอบได้ ด้วยการประกอบสายลำดับที่อ่านได้ที่มีความถี่สูงสุด แบบสุ่มตำแหน่ง แยกตามช่วงความถี่ของสายลำดับหลักที่จำลองขึ้น โดยวัดความแม่นจากร้อยละของตำแหน่งที่ มีอัลลีลตรงกับเส้นต้นแบบ จุดในกราฟแสดงความแม่นเฉลี่ยของสายแฮปโพลไทป์และช่วงในกราฟแสดงช่วงความแม่นของสายแฮปโพลไทป์ที่ประกอบได้จากความแม่นต่ำสุดถึงความแม่นสูงสุด

ตารางที่ 5.4 ความผิดพลาดสัมบูรณ์ของความถี่แฮปโพลไทป์สายหลักที่คำนวณได้ จากการประกอบแฮปโพลไทป์สายหลักด้วยสายลำดับที่อ่านได้ที่มีความถี่สูงสุดแบบสุ่มตำแหน่ง โดยประมาณค่าความถี่แฮปโพลไทป์ด้วยมัธยฐาน

ช่วงความถี่สายลำดับหลัก (%)	ความผิดพลาดสัมบูรณ์ต่ำสุด (%)	ความผิดพลาดสัมบูรณ์สูงสุด (%)	ความผิดพลาดสัมบูรณ์เฉลี่ย (%)
10-19	0.01	10.67	2.22
20-29	0.01	7.50	1.84
30-39	0.02	10.04	2.34
40-49	0.01	6.30	1.85
50-59	0.05	11.04	1.64
60-69	0.01	8.14	1.46
70-79	0.01	9.52	1.16
80-89	0.01	3.37	0.78
90-99	0.01	2.82	0.57
ทั้งหมด	**0.01**	**11.04**	**1.54**

5.3 การประกอบแฮปโพลไทป์สายหลักจากอัลลีลที่มีความถี่สูงสุดในแต่ละตำแหน่ง

เนื่องจากรูปแบบของประชากรไวรัส ที่มีความถี่ของแฮปโพลไทป์สายหลักสูงกว่าความถี่ของแฮปโพลไทป์สายอื่น และสายแฮปโพลไทป์สายอื่นเกิดขึ้นเนื่องจากการกลายพันธุ์ของแฮปโพลไทป์สายหลักนี้ ทำให้โดยส่วนใหญ่อัลลีลบนแฮปโพลไทป์สายหลักจะเป็นอัลลีลที่มีความถี่สูงสุดในตำแหน่งนั้นๆ จึงได้ประกอบแฮปโพลไทป์สายหลักขึ้นจากอัลลีลที่มีความถี่สูงสุดในแต่ละตำแหน่ง

5.3.1 ขั้นตอนวิธี

สำหรับวิธีนี้มีอินพุต คือ สายลำดับย่อยของสนิปส์ และตำแหน่งของสนิปส์ตัวแรกบนสายแฮปโพลไทเช่นเดียวกับ 2 วิธีข้างต้น จากนั้นประกอบแฮปโพลไทป์สายหลักตามขั้นตอนดังนี้

1) อ่านสายลำดับย่อยของสนิปส์ทีละเส้น เก็บอัลลีลและจำนวนแต่ละอัลลีลตามตำแหน่งของอัลลีลนั้นบนสายแฮปโพลไทป์ เช่น

อินพุตประกอบด้วยสายลำดับดังต่อไปนี้

AAATTC	เริ่มต้นที่ตำแหน่ง	0
AAATTT	เริ่มต้นที่ตำแหน่ง	0
AAATTT	เริ่มต้นที่ตำแหน่ง	0
AAATTT	เริ่มต้นที่ตำแหน่ง	0
CAGATT	เริ่มต้นที่ตำแหน่ง	0
TAAGT	เริ่มต้นที่ตำแหน่ง	1
GATC	เริ่มต้นที่ตำแหน่ง	2
GC	เริ่มต้นที่ตำแหน่ง	4

เก็บค่าอัลลีลแต่ละตำแหน่งและจำนวนของแต่ละอัลลีลได้ดังนี้

ตำแหน่ง	0 ประกอบด้วย	A 4 ตัว	C 1 ตัว
ตำแหน่ง	1 ประกอบด้วย	A 5 ตัว	T 1 ตัว
ตำแหน่ง	2 ประกอบด้วย	A 5 ตัว	G 2 ตัว
ตำแหน่ง	3 ประกอบด้วย	T 4 ตัว	A 3 ตัว
ตำแหน่ง	4 ประกอบด้วย	T 6 ตัว	G 2 ตัว
ตำแหน่ง	5 ประกอบด้วย	T 5 ตัว	C 3 ตัว

2) คำนวณความถี่อัลลีล โดยในแต่ละตำแหน่ง นำจำนวนอัลลีลแต่ละ ตัวมาหารด้วยจำนวนอัลลีลทั้งหมดของตำแหน่งนั้น เช่น จากตัวอย่างข้างต้นคำนวณความถี่อัลลีลได้ดังนี้

ตำแหน่ง	0 ประกอบด้วย	A ความถี่ 4 ÷ 5 = 0.80	C ความถี่ 1 ÷ 5 = 0.20
ตำแหน่ง	1 ประกอบด้วย	A ความถี่ 5 ÷ 6 = 0.83	T ความถี่ 1 ÷ 6 = 0.17
ตำแหน่ง	2 ประกอบด้วย	A ความถี่ 5 ÷ 7 = 0.71	G ความถี่ 2 ÷ 7 = 0.29
ตำแหน่ง	3 ประกอบด้วย	T ความถี่ 4 ÷ 7 = 0.57	T ความถี่ 3 ÷ 7 = 0.43
ตำแหน่ง	4 ประกอบด้วย	T ความถี่ 6 ÷ 8 = 0.75	G ความถี่ 2 ÷ 8 = 0.25
ตำแหน่ง	5 ประกอบด้วย	T ความถี่ 5 ÷ 8 = 0.625	T ความถี่ 3 ÷ 8 = 0.375

3) ประกอบสายแฮปโพลไทป์หลัก โดยเลือกอัลลีลที่มีความถี่สูงสุดในแต่ละตำแหน่งมาประกอบเป็นสายแฮปโพลไทป์หลัก จากตัวอย่าง ตำแหน่งที่ 0-2 เลือกอัลลีล A และตำแหน่งที่ 3-5 เลือกอัลลีล T ดังนั้นได้แฮปโพลไทป์สายหลักคือ AAATTT

4) ประมาณค่าความถี่แฮปโพลไทป์สายหลัก (haplotype frequency)

4.1) ประมาณค่าความถี่แฮปโพลไทป์สายหลัก ด้วยค่าเฉลี่ยเลขคณิต ค่าสูงสุด และค่าต่ำสุดโดยใช้ความถี่อัลลีลแต่ละ ตำแหน่งบนสายลำดับลักษณ์ที่ได้ โดยตรง เช่น จากตัวอย่างประมาณค่าความถี่ด้วยค่าเฉลี่ยเลขคณิตได้เป็น (0.80+0.83+0.71+0.57+0.75+0.625) ÷ 6 = 0.714 หรือร้อยละ 71.4 และประมาณค่าความถี่ด้วยมัธยฐานได้เป็น (0.71+0.75) ÷ 2 = 0.73 หรือ ร้อยละ 73

4.2) จากผลการทดลองพบว่าการประมาณค่าความถี่จากความถี่อัลลีลสูงสุดในแต่ละตำแหน่งโดยตรงไม่มีประสิทธิภาพ และการประมาณค่าด้วยวิธีนี้ไม่ได้ใช้ความรู้เกี่ยวกับการไปด้วยกันของแต่ละอัลลีลที่ได้จากอัลลีลที่อยู่บนสายลำดับที่อ่านได้สายเดียวกันเลย ดังนั้นจึงพัฒนาวิธีการประมาณความถี่แฮปโพลไทป์สายหลักขึ้นใหม่ โดยใช้แนวคิดเดียวกับวิธีการประกอบแฮปโพลไทป์สายหลักจากสายลำดับที่อ่านได้ที่มีความถี่สูงสุด ได้ขั้นตอนประมาณความถี่ดังนี้

- นำสายลำดับที่เหมือนกัน คือ มีความยาวเท่ากัน มีตำแหน่งเดียวกัน และแต่ละตำแหน่งมีเบสตัวเดียวกัน รวมเข้าด้วยกัน
- เรียงสายลำดับที่อ่านได้ทั้งหมดตามตำแหน่งจากน้อยไปมาก (จากหัวไปท้าย)
- แต่ละตำแหน่งเรียงสายลำดับตามความยาวจากมากไปน้อย
- แยกสายลำดับเป็นสองกลุ่มคือ กลุ่มที่เป็นส่วนหนึ่งของแฮปโพลไทป์สายหลักที่ประกอบขึ้นคือแต่ละตำแหน่งบนสายลำดับนั้นมีเบสเดียวกับตำแหน่งนั้นๆ บนแฮปโพลไทป์ และกลุ่มที่ไม่เหมือนกับแฮปโพลไทป์สายหลักที่ประกอบขึ้น โดยการแยกสายลำดับนี้ทำทีละตำแหน่งและเรียงลำดับตามความยาว

- สายลำดับเส้นสั้นที่สามารถอยู่ได้ทั้งสองกลุ่ม แบ่งสายลำดับนั้นให้อยู่ทั้งสองกลุ่มด้วยสัดส่วนตามจำนวนอัลลีลในแต่ละกลุ่ม เช่น แบ่งสายลำดับที่อ่านเข้ามาก่อนหน้าได้เป็น 2 กลุ่ม คือ

ก. กลุ่มที่เหมือนกับแฮปโพลไทป์สายหลัก คือ AAATTGGG ทั้งหมด 5 เส้น

ข. กลุ่มที่ไม่เหมือนกับแฮปโพลไทป์สายหลัก ประกอบด้วย AAAGTT จำนวน 1 เส้น และ AAC จำนวน 2 เส้น

สายลำดับที่อ่านได้ลำดับถัดไปคือ AA จำนวน 8 เส้น สายลำดับ AA นี้อยู่ได้ทั้งสองกลุ่ม จึงแบ่งให้ทั้งสองกลุ่มด้วยสัดส่วนตามสายลำดับที่มีอยู่ในแต่ละกลุ่ม ดังนั้น แบ่งให้ AA อยู่กลุ่ม ก. 5 เส้น และอยู่กลุ่ม ข. 3 เส้น

- ทำเช่นนี้จนครบทุกตำแหน่ง จากนั้นหาความถี่อัลลีลแต่ละตำแหน่งในกลุ่มที่เหมือนกับแฮปโพลไทป์สายหลัก

- ประมาณค่าความถี่แฮปโพลไทป์สายหลักด้วยค่าเฉลี่ยเลขคณิตและมัธยฐานของความถี่อัลลีลแต่ละตำแหน่งที่คำนวณได้จากกลุ่มที่เหมือนกับแฮปโพลไทป์สายหลักนี้

5.3.2 การทดสอบ

ทดสอบความถูกต้องของขั้นตอนวิธี ด้วยความแม่นของแฮปโพลไทป์สายหลักที่ประกอบได้และความถี่แฮปโพลไทป์สายหลัก โดยใช้ข้อมูลเดียวกับ 2 วิธีก่อนหน้า ด้วยข้อมูลรวมทั้งสิ้น 720 ชุด

5.3.3 ผลการทดลอง

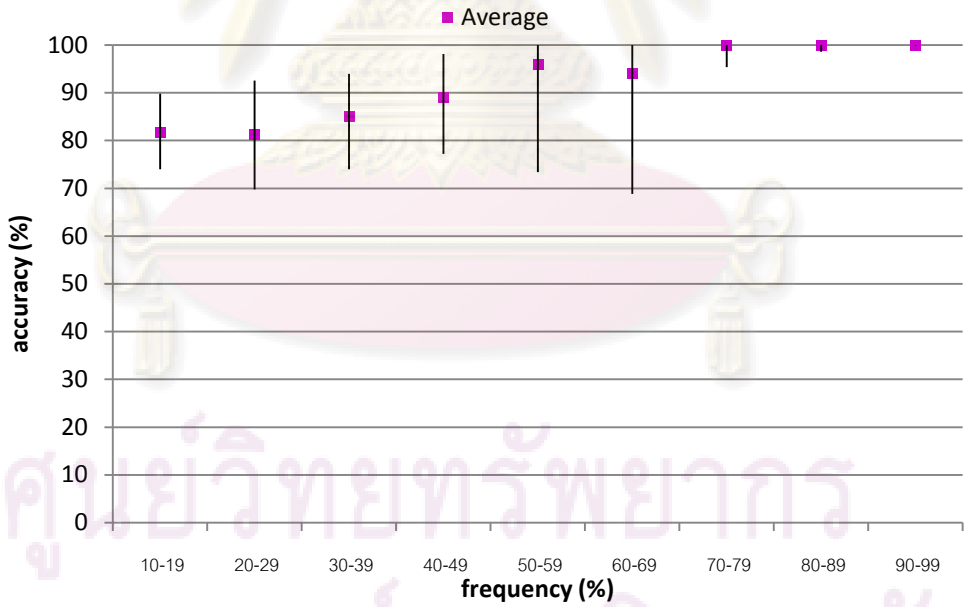
จากการทดลองประกอบ แฮปโพลไทป์สายหลักจากชุดของสายลำดับที่มีความแตกต่างกันประมาณร้อยละ 2 และมี 2 อัลลีลในแต่ละตำแหน่ง ได้ความแม่นของสาย แฮปโพลไทป์ที่ประกอบได้ดังตารางที่ 5.5

จากนั้น นำความแม่นของสายแฮปโพลไทป์ที่ได้จากการประกอบแฮปโพลไทป์สายหลักด้วยอัลลีลที่มีความถี่สูงสุด แยกตามช่วงความถี่ของสายลำดับหลัก มาสร้างกราฟได้ดังรูปที่ 5.3

จากผลการทดลองจะเห็นว่า ความแม่นเพิ่มขึ้นเมื่อความถี่สายลำดับหลัก เพิ่มขึ้น โดยความแม่นเฉลี่ยของชุดข้อมูลที่สายลำดับหลักมีความถี่ 90-99 % คือร้อยละ 100 และความแม่นเฉลี่ยทั้งหมดคือร้อยละ 92.07 ในความเป็นจริง การประกอบแฮปโพลไทป์ด้วยวิธีนี้จากชุดข้อมูลที่สายลำดับหลักมีความถี่ตั้งแต่ 51% ขึ้นไปควรจะได้สายลำดับแฮปโพลไทป์ที่ถูกต้องทุกตำแหน่ง แต่จากผลการทดลองไม่เป็นเช่นนั้น เพราะกรณีที่จำนวนสายลำดับที่อ่านมาเป็นอินพุตไม่มากพอ จะอ่านสายลำดับได้ไม่ครอบคลุมทั้งหมด ทำให้ตำแหน่งที่ได้เกิดข้อผิดพลาด สำหรับชุดข้อมูลนี้ที่ความถี่ของสายลำดับหลักเป็น 60% ขึ้นไป จะได้แฮปโพลไทป์ที่ถูกต้องทั้งเส้นเมื่อสายลำดับที่อ่าน

ตารางที่ 5.5 ความแม่นของสายแฮปโพลไทป์หลักที่ประกอบจากอัลลีลที่มีความถี่สูงสุดในแต่ละตำแหน่ง

ช่วงความถี่สายลำดับหลัก (%)	ความแม่นสูงสุด (%)	ความแม่นต่ำสุด (%)	ความแม่นเฉลี่ย (%)
10-19	89.77	73.95	81.60
20-29	92.56	69.77	81.32
30-39	93.95	73.95	84.96
40-49	98.14	77.21	89.08
50-59	100.00	73.36	96.03
60-69	100.00	68.84	94.07
70-79	100.00	95.35	99.83
80-89	100.00	98.60	99.98
90-99	100.00	100.00	100.00
ทั้งหมด	**100.00**	**69.77**	**92.07**



รูปที่ 5.3 กราฟความแม่นของ สายแฮปโพลไทป์ หลักที่ประกอบได้ ด้วยอัลลีลที่มีความถี่สูงสุด แยกตามช่วงความถี่ของสายลำดับหลักที่จำลองขึ้น โดยวัดความแม่นจากร้อยละของตำแหน่งที่ มีอัลลีล ตรงกับเส้นต้นแบบ จุดในกราฟแสดงความแม่นเฉลี่ยของสายแฮปโพลไทป์และช่วงในกราฟแสดงช่วงความแม่นของสายแฮปโพลไทป์ที่ประกอบได้จากความแม่นต่ำสุดถึงความแม่นสูงสุด

ได้มีจำนวน 3,000 เส้นขึ้นไป และสายลำดับที่อ่านได้ต้องมีจำนวน 10,000 เส้นขึ้นไปสำหรับชุดข้อมูลที่มีสายลำดับหลักอยู่ในช่วง 52-59%

เมื่อได้แฮปโพลไทป์สายหลักแล้ว นำความถี่อัลลีลแต่ละตำแหน่งบนแฮปโพลไทป์ที่ได้ไปประมาณค่าความถี่ด้วยค่าเฉลี่ยเลขคณิต ค่าสูงสุดและค่าต่ำสุด ได้ผลดังตารางที่ 5.6

ตารางที่ 5.6 ความผิดพลาดสัมบูรณ์ของงความถี่แฮปโพลไทป์สายหลักที่คำนวณได้ จากการประกอบแฮปโพลไทป์สายหลักด้วยจากอัลลีลที่มีความถี่สูงสุด เปรียบเทียบระหว่างการประมาณค่าด้วยค่าเฉลี่ยเลขคณิต ค่าสูงสุดและค่าต่ำสุด

ช่วงความถี่สายลำดับหลัก (%)	ความแตกต่างจากการประมาณด้วยค่าเฉลี่ยเลขคณิต (%)			ความแตกต่างจากการประมาณด้วยค่าสูงสุด (%)			ความแตกต่างจากการประมาณด้วยค่าต่ำสุด (%)		
	สูงสุด	ต่ำสุด	เฉลี่ย	สูงสุด	ต่ำสุด	เฉลี่ย	สูงสุด	ต่ำสุด	เฉลี่ย
10-19	50.54	60.85	54.97	79.25	86.63	83.41	30.21	36.73	34.40
20-29	41.69	52.22	46.86	68.21	79.87	75.64	21.09	30.32	26.55
30-39	29.69	44.63	35.98	47.94	68.89	61.58	10.47	19.26	14.71
40-49	24.93	36.27	30.10	47.84	59.50	54.44	2.40	13.29	5.85
50-59	15.50	28.95	22.16	28.71	49.53	43.17	0.05	8.16	3.43
60-69	13.10	21.21	18.32	33.00	37.31	33.68	0.03	19.16	6.23
70-79	7.43	13.50	10.76	18.87	29.37	23.57	0.28	28.45	9.28
80-89	0.02	10.25	6.07	1.92	17.94	12.88	0.27	32.61	7.76
90-99	0.01	5.91	2.22	0.60	9.39	5.53	0.62	33.52	6.98
ทั้งหมด	**0.01**	**60.85**	**26.07**	**0.60**	**86.63**	**44.92**	**0.03**	**36.73**	**13.14**

จากตารางจะเห็นว่าการประมาณค่าความถี่จากความถี่อัลลีลที่มากที่สุดในแต่ละตำแหน่งโดยตรงให้ผลไม่ดีเท่าที่ควร ดังนั้นจึงได้ปรับปรุงการประมาณค่าความถี่แฮปโพลไทป์ใหม่โดยใช้แนวคิดเดียวกับการประมาณค่าความถี่ของแฮปโพลไทป์สายหลักที่ได้จากสายลำดับที่อ่านได้ที่มีความถี่สูงสุด ดังขั้นตอนในขั้นตอนวิธี 5.3.1 ข้อ 4.2) ได้ความผิดพลาดสัมบูรณ์ ของความถี่แฮปโพลไทป์สายหลักที่ประมาณได้เทียบกับความถี่แฮปโพลไทป์จริง ดังแสดงในตารางที่ 5.7

จากตารางที่ 5.7 การประมาณค่าความถี่แฮปโพลไทป์สายหลักโดยแบ่งกลุ่มสายลำดับที่อ่านได้ออกเป็น 2 กลุ่ม คือ กลุ่มที่คาดว่ามาจากแฮปโพลไทป์สายหลักที่ประกอบขึ้น กับกลุ่มที่ไม่ได้มาจากแฮปโพลไทป์สายหลักนี้ แล้วจึงประมาณความถี่แฮปโพลไทป์จากความถี่อัลลีลของสาย

ลำดับเฉพาะที่อยู่ในกลุ่มที่คาดว่ามาจากแฮปโพลไทป์สายหลักนี้ด้วยค่าเฉลี่ยเลขคณิตและมัธยฐาน ให้ผลดีกว่าการประมาณค่าความถี่แฮปโพลไทป์จากความถี่อัลลีลสูงสุดในแต่ละตำแหน่งโดยตรง โดยการประมาณค่าความถี่แฮปโพลไทป์จากความถี่อัลลีลของสายลำดับกลุ่มที่มาจากแฮปโพลไทป์สายหลักที่ประกอบได้ ด้วยค่าเฉลี่ยเลขคณิตและมัธยฐานนั้นให้ประสิทธิภาพใกล้เคียงกัน

ตารางที่ 5.7 ความผิดพลาดสัมบูรณ์ของความถี่แฮปโพลไทป์สายหลักที่คำนวณได้ จากการประกอบแฮปโพลไทป์สายหลักด้วยจากอัลลีลที่มีความถี่สูงสุด เปรียบเทียบระหว่างการประมาณค่าด้วยค่าเฉลี่ยเลขคณิตและมัธยฐานของความถี่อัลลีลบนสายลำดับที่มาจากแฮปโพลไทป์สายหลัก

ช่วงความถี่สายลำดับหลัก (%)	**ความผิดพลาดสัมบูรณ์จากการประมาณความถี่ด้วยค่าเฉลี่ยเลขคณิต (%)**			**ความผิดพลาดสัมบูรณ์จากการประมาณความถี่ด้วยค่ามัธยฐาน (%)**		
	สูงสุด	ต่ำสุด	เฉลี่ย	สูงสุด	ต่ำสุด	เฉลี่ย
10-19	0.02	11.56	4.69	0.28	15.08	7.12
20-29	1.31	16.54	10.79	2.09	20.78	14.29
30-39	0.36	26.21	16.58	3.00	35.23	20.72
40-49	1.74	33.55	15.22	0.03	39.58	13.94
50-59	0.03	24.70	6.16	0.01	29.41	3.05
60-69	0.01	14.62	3.18	0.03	5.93	1.47
70-79	0.08	13.57	2.13	0.01	9.09	0.99
80-89	0.01	13.8	2.99	0.01	4.11	0.81
90-99	0.01	6.54	0.93	0.01	2.82	0.58
ทั้งหมด	**0.01**	**33.55**	**7.19**	**0.01**	**39.58**	**7.36**

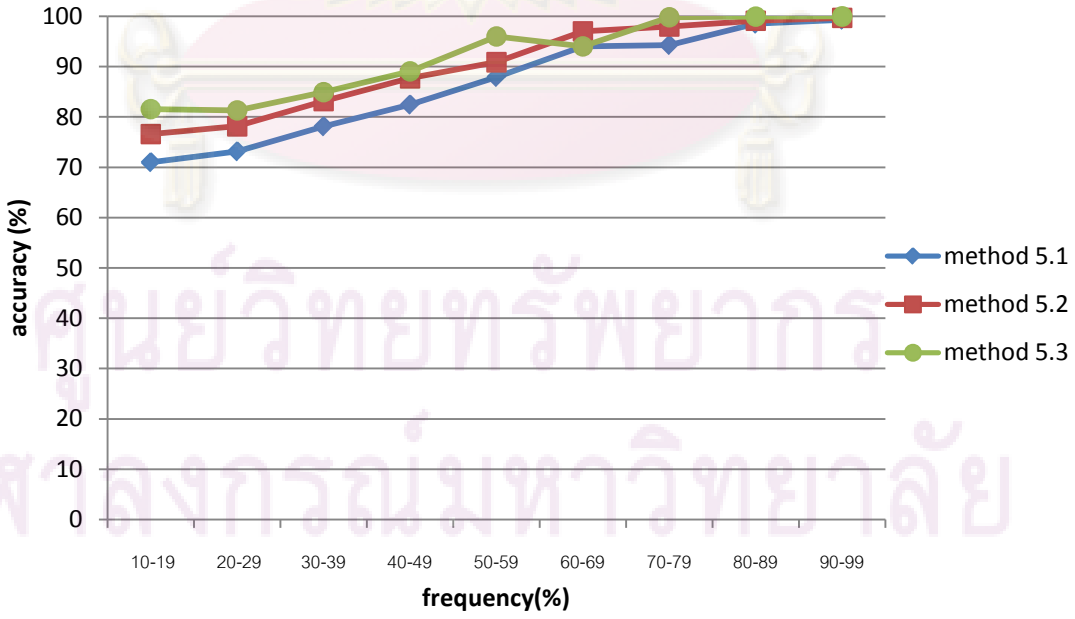
5.4 เปรียบเทียบแต่ละขั้นตอนวิธี

เปรียบเทียบประสิทธิภาพขั้นตอนวิธีประกอบแฮปโพลไทป์สายหลักทั้งสามวิธี คือ ประกอบจากสายลำดับที่อ่านได้ที่มีความถี่สูงสุดโดยเรียงตามลำดับ ประกอบจากสายลำดับที่อ่านได้ที่มีความถี่สูงสุดโดยสุ่มตำแหน่ง และประกอบจากอัลลีลที่มีความถี่สูงสุด โดยใช้ค่าความแม่นได้ผลดังตารางที่ 5.8

ตารางที่ 5.8 ความแม่นของสายแฮปโพลไทป์หลัก เปรียบเทียบระหว่างขั้นตอนวิธี 5.1, 5.2 และ 5.3

ช่วงความถี่สายลำดับหลัก (%)	ความแม่นจากขั้นตอนวิธี 5.1 (%)			ความแม่นจากขั้นตอนวิธี 5.2 (%)			ความแม่นจากขั้นตอนวิธี 5.3 (%)		
	สูงสุด	ต่ำสุด	เฉลี่ย	สูงสุด	ต่ำสุด	เฉลี่ย	สูงสุด	ต่ำสุด	เฉลี่ย
10-19	93.43	58.88	71.04	91.63	61.86	76.61	89.77	73.95	81.60
20-29	99.53	55.81	73.19	95.35	61.40	78.18	92.56	69.77	81.32
30-39	100.00	53.74	78.14	100.00	63.72	83.17	93.95	73.95	84.96
40-49	100.00	53.95	82.48	100.00	62.33	87.72	98.14	77.21	89.08
50-59	100.00	50.00	87.96	100.00	67.29	90.90	100.00	73.36	96.03
60-69	100.00	68.84	94.07	100.00	83.57	97.03	100.00	68.84	94.07
70-79	100.00	67.44	94.33	100.00	82.24	98.02	100.00	95.35	99.83
80-89	100.00	65.58	98.55	100.00	90.23	99.14	100.00	98.60	99.98
90-99	100.00	71.63	99.34	100.00	93.95	99.73	100.00	100.00	100.00
ทั้งหมด	**100.00**	**50.00**	**86.57**	**100.00**	**61.40**	**90.05**	**100.00**	**69.77**	**92.07**

นำความแม่นเฉลี่ยของแต่ละขั้นตอนวิธีมาสร้างกราฟ โดยจำแนกตามช่วงความถี่ของสายลำดับหลักได้ดังรูปที่ 5.4



รูปที่ 5.4 กราฟความแม่นของสายแฮปโพลไทป์หลักที่ประกอบขึ้น เปรียบเทียบระหว่างขั้นตอนวิธี 5.1, 5.2 และ 5.3

5.5 สรุป

จากรูปที่ 5.4 จะเห็นว่าขั้นตอนวิธีที่ให้ความแม่นสูงสุดคือ ขั้นตอนวิธี 5.3 การประกอบแฮปโพลไทป์สายหลักจากอัลลีลที่มีความถี่สูงสุด รองลงมาคือ ขั้นตอนวิธี 5.2 การประกอบแฮปโพลไทป์สายหลักจากสายลำดับที่อ่านได้ โดยสุ่มตำแหน่ง ซึ่งขั้นตอนวิธี 5.2 เป็นวิธีแบบความน่าจะเป็น (probabilistic method) ผลของแต่ละรอบไม่เหมือนกัน ต่างจากขั้นตอนวิธีที่ 5.3 ซึ่งคำตอบขึ้นอยู่กับอินพุตที่ได้เท่านั้น จึงเลือกใช้ทั้ง 2 ขั้นตอนวิธีนี้สำหรับประกอบชุดของแฮปโพลไทป์ของประชากรไวรัสเด็งกี่ในบทถัดไป

สำหรับการประมาณค่าความถี่นั้นเลือกใช้การประมาณค่าความถี่ด้วยค่ามัธยฐานของความถี่อัลลีลบนสายแฮปโพลไทป์หลักที่ประกอบขึ้นสำหรับขั้นตอนวิธี 5.2 และใช้การประมาณค่าความถี่ด้วยค่าเฉลี่ยเลขคณิตของอัลลีลในกลุ่มของสายลำดับที่อ่านได้ที่เหมือนกับสายแฮปโพลไทป์หลักสำหรับขั้นตอนวิธี 5.3

บทที่ 6
การประกอบชุดของสายแฮปโพลไทป์และประมาณความถี่แฮปโพลไทป์

จากวิธีประกอบแฮปโพลไทป์สายหลัก สามารถนำมาประยุกต์ใช้กับการประกอบชุดของแฮปโพลไทป์ของประชากรไวรัสเด็งกี่ได้ โดยประกอบแฮปโพลไทป์สายหลัก จากนั้นกรองเอาสายลำดับที่คาดว่ามาจากแฮปโพลไทป์สายหลักออก นำสายลำดับที่เหลือมาประกอบแฮปโพลไทป์ที่มีความถี่สูงสุดในกลุ่มที่เหลือ วนซ้ำเช่นนี้ไปเรื่อยๆ จนถึงเกณฑ์ที่กำหนด จากนั้นทดสอบความแม่นของผลลัพธ์ที่ได้และรายงานผล

6.1 ขั้นตอนวิธี

อินพุตสำหรับขั้นตอนวิธีนี้เป็นสายลำดับที่อ่านได้เช่นเดียวกับในบทที่ 5 โดยขั้นตอนวิธีนี้ประกอบด้วย 2 ขั้นตอนหลักคือ ขั้นตอนการประกอบแฮปโพลไทป์สายหลัก และขั้นตอนการกรองสายลำดับที่คาดว่าเป็นสายลำดับที่อ่านได้จากแฮปโพลไทป์สายหลักในรอบนั้นๆ ออก จากนั้นนำสายลำดับที่เหลือเป็นอินพุตของรอบถัดไป วนซ้ำเช่นนี้ไปเรื่อยๆ จนกว่าสายลำดับจะหมด หรือถึงขอบเขตที่กำหนด โดยในการทดลองนี้กำหนดให้วนซ้ำจนกว่าสายลำดับแฮปโพลไทป์ที่เหลือมีความถี่น้อยกว่าร้อยละ 1 และให้วนซ้ำทั้งสิ้นไม่เกิน 50 รอบ เมื่อถึงขอบเขตที่กำหนดแล้วให้ประกอบแฮปโพลไทป์สายหลักจากสายลำดับที่เหลือ และกำหนดค่าความถี่ของแฮปโพลไทป์สายนั้นตามความถี่แฮปโพลไทป์ที่ยังขาดอยู่ เป็นอันสิ้นสุดขั้นตอนวิธี แสดงขั้นตอนวิธีหลักได้นี้ดังรูปที่ 6.1 โดยขั้นตอนประกอบแฮปโพลไทป์สายหลักและขั้นตอนกรองสายลำดับมีรายละเอียดดังนี้

6.1.1 ขั้นตอนประกอบแฮปโพลไทป์สายหลัก

ในขั้นตอนนี้ทำเหมือนขั้นตอนวิธีในบทที่ 5 โดยเลือก 2 วิธีที่ให้ความแม่นสูงสุดจากการทดลองในบทที่ 5 คือ ขั้นตอนวิธี 5.2 การประกอบแฮปโพลไทป์สายหลักจากสายลำดับที่อ่านได้ที่มีความถี่สูงสุด แบบสุ่มตำแหน่ง โดยใช้การประมาณค่าความถี่ด้วยค่ามัธยฐานของความถี่อัลลีลบนสายแฮปโพลไทป์หลักที่ประกอบขึ้น และขั้นตอนวิธี 5.3 การประกอบแฮปโพลไทป์สายหลักจากอัลลีลที่มีความถี่สูงสุดและประมาณค่าความถี่ด้วยค่าเฉลี่ยเลขคณิตของอัลลีลในกลุ่มของสายลำดับที่อ่านได้ที่เหมือนกับสายแฮปโพลไทป์หลัก

6.1.2 ขั้นตอนกรองสายลำดับ

ขั้นตอนนี้จะกรองสายลำดับที่คาดว่าได้มาจากการอ่านสายลำดับที่เป็นแฮปโพลไทป์สายหลักออกจากสายลำดับเดิมที่เป็นอินพุตของรอบนั้น โดยมีขั้นตอนดังนี้

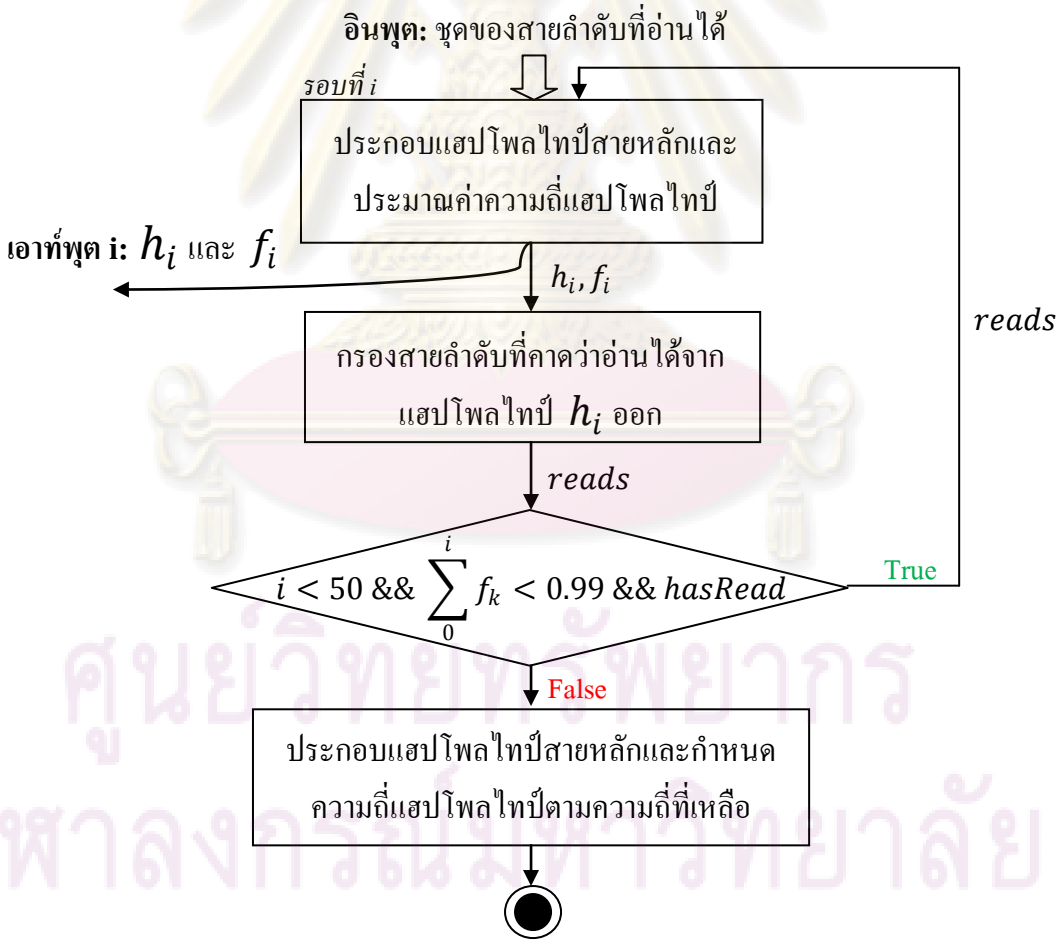
1) ในแต่ละตำแหน่งบนแฮปโพลไทป์สายหลัก คำนวณจำนวนอัลลีลที่มาจากแฮปโพลไทป์สายนี้ โดยนำค่าความถี่แฮปโพลไทป์สายหลักที่คำนวณได้คูณกับจำนวนอัลลีลทั้งหมดที่อ่านได้ในตำแหน่งนั้น

2) นำสายลำดับที่เหมือนกันรวมเข้าด้วยกัน โดยเก็บจำนวนสายลำดับนั้นไว้

3) เรียงสายลำดับตามความยาวของสายลำดับจากมากไปน้อย

4) เรียงสายลำดับที่ความยาวเดียวกันตามจำนวนจากมากไปน้อย

5) ไล่ดูทีละสายลำดับ ถ้าสายลำดับนั้นเป็นส่วนหนึ่งของแฮปโพลไทป์สายหลักที่ประกอบได้ในรอบนั้น และมีจำนวนไม่เกินจำนวนอัลลีลบนแฮปโพลไทป์สายหลักที่เหลืออยู่ของตำแหน่งนั้น (ได้จากการคำนวณในข้อ 1) หรือมากกว่าไม่เกิน 1% ให้กรองสายลำดับนั้นทิ้ง แต่ถ้าสายลำดับนั้นมีจำนวนมากกว่าจำนวนอัลลีลที่เหลืออยู่เกิน 1% ให้กรองออกเท่ากับจำนวนอัลลีลบนแฮปโพลไทป์สายหลักที่เหลืออยู่



รูปที่ 6.1 แผนผังแสดงขั้นตอนประกอบชุดของแฮปโพลไทป์และประมาณค่าความถี่ของแฮปโพลไทป์แต่ละเส้น

6) ลบจำนวนอัลลีลบนแฮปโพลไทป์สายหลักที่เหลืออยู่ออกตามจำนวนสายลำดับที่กรองออกไป

7) ทำซ้ำจนครบทุกสายลำดับ หรือจำนวนอัลลีลที่เหลืออยู่หมด

8) ชุดของสายลำดับที่เหลือ คือชุดของสายลำดับที่คาดว่าไม่อยู่บนแฮปโพลไทป์ที่ได้ในรอบนั้น

6.2 การทดสอบ

ทดสอบด้วยชุดข้อมูลเดียวกับในบทที่ 5 แต่เก็บข้อมูลแฮปโพลไทป์ทั้งหมดที่จำลองขึ้นและความถี่ของแต่ละแฮปโพลไทป์สำหรับใช้เปรียบเทียบผลด้วย แทนที่จะเก็บเฉพาะข้อมูลแฮปโพลไทป์สายหลักเพียงอย่างเดียว ทดสอบผลเปรียบเทียบระหว่างการเลือกใช้วิธีประกอบแฮปโพลไทป์ด้วยขั้นตอนวิธี 5.2 และ 5.3 โดยเปรียบเทียบความแม่นกับสายแฮปโพลไทป์ที่จำลองขึ้น เรียงตามลำดับความถี่ นั่นคือ เปรียบเทียบแฮปโพลไทป์ที่จำลองขึ้นเส้นที่มีความถี่สูงสุด กับแฮปโพลไทป์ที่มีความถี่สูงสุดที่ได้จากการประกอบด้วยวิธีที่นำเสนอ ไล่ไปจนถึงเส้นที่มีความถี่น้อยที่สุด

6.3 ผลการทดลอง

เปรียบเทียบประสิทธิภาพของขั้นตอนวิธีด้วยความแม่นของสายลำดับที่ประกอบขึ้น โดยแยกตามช่วงความถี่ของสายลำดับหลักและลำดับที่ของสายแฮปโพลไทป์ที่ประกอบได้ ดังแสดงในตารางที่ 6.1

ตารางที่ 6.1 ความแม่นของแฮปโพลไทป์ที่ประกอบได้ แยกตามช่วงความถี่ของสายลำดับหลักและลำดับของสายแฮปโพลไทป์ที่ประกอบขึ้น เปรียบเทียบระหว่างการเลือกใช้วิธี 5.2 และ 5.3 ในขั้นตอนประกอบแฮปโพลไทป์สายหลัก

เส้นที่	ความแม่นเฉลี่ยของแฮปโพลไทป์แยกตามช่วงความถี่สายลำดับหลัก (%)								
	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
1	**76.68**	**77.34**	**82.65**	**86.99**	**90.94**	**97.05**	**97.75**	**99.52**	**99.94**
	81.60	81.32	84.96	89.08	96.03	99.63	99.83	99.98	100.00
2	**63.25**	**61.90**	**59.29**	**60.32**	**54.70**	**54.44**	**45.88**	**43.42**	**41.30**
	69.71	68.06	69.08	75.89	88.63	80.16	92.66	86.87	76.99
3	**62.27**	**58.79**	**55.12**	**52.09**	**49.65**	**54.16**	**46.12**	**49.87**	**39.09**
	67.12	67.07	64.89	63.35	69.19	71.63	71.17	62.82	65.75
4	**62.06**	**59.12**	**51.80**	**51.97**	**49.86**	**52.41**	**51.16**	**47.27**	**0**

เส้นที่	ความแม่นเฉลี่ยของแฮปโพลไทป์แยกตามช่วงความถี่สายลำดับหลัก (%)								
	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
	67.40	65.51	57.75	54.56	63.03	61.58	58.93	58.83	60.54
5	**62.96**	**58.91**	**53.09**	**51.93**	**51.64**	**53.00**	**53.00**	**0**	**0**
	68.66	64.77	56.32	54.39	57.29	58.36	56.96	59.70	56.82
6	**61.96**	**57.72**	**52.74**	**51.53**	**57.67**	**55.59**	**0**	**0**	-
	67.30	62.06	53.04	54.02	56.57	55.62	54.02	55.50	-
7	**61.41**	**55.71**	**56.15**	**56.77**	**55.76**	**0**	-	-	-
	66.61	59.48	57.22	55.64	56.50	56.59	-	-	-
8	**61.52**	**56.79**	**56.27**	**55.23**	**53.02**	**0**	-	-	-
	65.79	58.70	56.66	55.35	59.93	60.20	-	-	-
9	**61.19**	**55.46**	-	-	-	-	-	-	-
	66.40	57.39	-	-	-	-	-	-	-
10	**60.21**	**56.33**	-	-	-	-	-	-	-
	64.95	59.73	-	-	-	-	-	-	-
ทั้งหมด	**61.79**	**60.01**	**59.61**	**60.22**	**60.16**	**61.90**	**60.56**	**60.98**	**58.42**
	66.50	64.73	64.90	66.52	75.22	71.82	81.40	80.23	82.17

■ *แสดง ความแม่นเฉลี่ยของแฮปโพลไทป์ที่ประกอบขึ้นโดยใช้วิธี 5.2 ในขั้นตอนประกอบแฮปโพลไทป์สายหลัก*

■ *แสดงความแม่นเฉลี่ยของแฮปโพลไทป์ที่ประกอบขึ้น โดยใช้วิธี 5.3 ในขั้นตอนประกอบแฮปโพลไทป์สายหลัก*

- คือ แฮปโพลไทป์เส้นนั้นไม่มีอยู่จริงในอินพุตของขั้นตอนวิธี

0 คือ ไม่มีสามารถประกอบแฮปโพลไทป์ในลำดับนั้นได้ด้วยวิธีที่นำเสนอ เนื่องจากแฮปโพลไทป์นั้นมีความถี่น้อย

จากตารางจะเห็นว่าความแม่นเฉลี่ยของแฮปโพลไทป์ที่ประกอบขึ้นโดยใช้วิธี 5.3 ในขั้นตอนประกอบแฮปโพลไทป์สายหลักสูงกว่าการเลือกใช้วิธี 5.2 ความแม่นของแฮปโพลไทป์ที่ประกอบขึ้นโดยเลือกใช้วิธี 5.3 ในขั้นตอนประกอบแฮปโพลไทป์ เฉลี่ยทั้งหมดทุกเส้นและทุกช่วงความถี่สายลำดับหลักเป็นร้อยละ 69.79 ในขณะที่การประกอบแฮปโพลไทป์โดยเลือกใช้วิธี 5.2 มีความแม่นเฉลี่ยร้อยละ 66.65

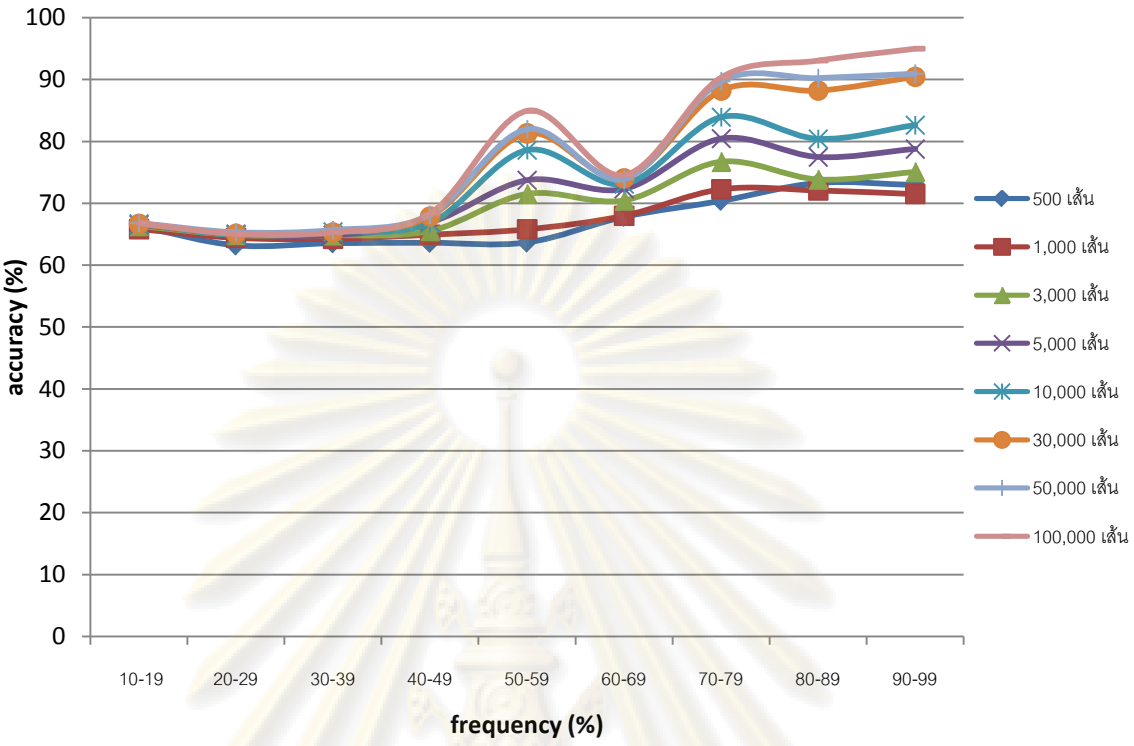
นอกจากนี้ จากตารางพบว่า แฮปโพลไทป์เส้นแรกๆ ซึ่งมีความถี่สูงจะมีความแม่นสูงกว่าแฮปโพลไทป์ลำดับถัดไป และแฮปโพลไทป์ของชุดสายลำดับที่มีความถี่สายลำดับหลักสูงจะมีความแม่นโดยเฉลี่ยสูงกว่าแฮปโพลไทป์ของชุดสายลำดับที่มีความถี่สายลำดับหลักต่ำ

เมื่อพิจารณาวิธีประกอบแฮปโพลไทป์โดยเลือกวิธี 5.3 ในขั้นตอนประกอบแฮปโพลไทป์สายหลัก โดยพิจารณาจำนวนสายลำดับที่อ่านได้ซึ่งเป็นอินพุตของขั้นตอนวิธี เมื่อสายลำดับที่อ่านได้มีจำนวน 500, 1000, 3000, 5000, 10000, 30000, 50000 และ 100000 เส้น โดยใช้ประชากรเดียวกัน คือ มีแฮปโพลไทป์ชุดเดียวกัน แต่ละแฮปโพลไทป์มีสายลำดับเบสและความถี่เดียวกัน พบว่า เมื่อจำนวนสายลำดับที่อ่านได้มากขึ้น ความแม่นของแฮปโพลไทป์ที่ประกอบได้สูงขึ้น เนื่องจากเมื่ออ่านสายลำดับจำนวนมากขึ้น ความครอบคลุมสูงขึ้น แต่ละเส้น แต่ละตำแหน่งถูกอ่านมากครั้งขึ้นทำให้อ่านสายลำดับได้ถูกต้องยิ่งขึ้น แสดงความแม่นของสายแฮปโพลไทป์ที่ประกอบขึ้น จำแนกตามจำนวนสายลำดับที่อ่านได้ซึ่งเป็นอินพุตของขั้นตอนวิธีได้ดังตารางที่ 6.2

ตารางที่ 6.2 ความแม่นเฉลี่ยของแฮปโพลไทป์ที่ประกอบได้จากการประกอบแฮปโพลไทป์ทั้งชุด โดยเลือกใช้วิธี 5.3 ในขั้นตอนประกอบแฮปโพลไทป์สายหลัก จำแนกตามจำนวนสายลำดับที่อ่านเข้ามาเป็นอินพุตของขั้นตอนวิธี

จำนวนสายลำดับ	**ความแม่นเฉลี่ยของแฮปโพลไทป์แยกตามช่วงความถี่สายลำดับหลัก (%)**									
	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99	ทั้งหมด
500	66.28	63.19	63.54	63.60	63.66	67.81	70.39	73.30	72.96	65.95
1,000	65.82	64.40	64.25	64.92	65.78	67.97	72.24	72.05	71.46	66.62
3,000	66.35	64.90	64.84	65.54	71.52	70.46	76.70	73.84	75.05	68.36
5,000	66.53	65.01	64.87	66.91	73.73	72.31	80.45	77.47	78.78	69.58
10,000	66.68	64.90	65.41	66.82	78.59	73.11	83.92	80.41	82.63	70.72
30,000	66.66	65.09	65.30	67.89	81.30	74.05	88.22	88.20	90.43	71.93
50,000	66.79	65.30	65.66	68.15	81.97	73.90	89.73	90.24	90.95	72.17
100,000	66.81	65.08	65.33	68.35	84.96	74.56	90.35	93.07	94.99	72.58

จากตาราง นำความแม่นเฉลี่ยของแฮปโพลไทป์ที่ประกอบขึ้นจากชุดของสายลำดับที่อ่านเข้ามาจำนวนต่างๆ กัน นำมาสร้างกราฟจำแนกตามความถี่ของสายลำดับหลักได้ดังรูปที่ 6.2



รูปที่ 6.2 กราฟความแม่นเฉลี่ยของชุดสายลำดับแฮปโพลไทป์ที่ประกอบขึ้น จากข้อมูลสายลำดับที่อ่านได้จำนวน 500, 1000, 3000, 5000, 10000, 30000, 50000 และ 100000 เส้น

จากผลการทดลองความแม่นเฉลี่ยของแฮปโพลไทป์สูงขึ้นเมื่อจำนวนสายลำดับที่อ่านได้ซึ่งเป็นอินพุตของขั้นตอนวิธีเพิ่มขึ้น โดยความแม่นเฉลี่ยเมื่ออ่านสายลำดับเข้ามาทั้งสิ้น 100,000 เส้นเป็นร้อยละ 72.58 และความแม่นเฉลี่ยเมื่ออ่านสายลำดับ 500 เส้นเป็นร้อยละ 65.95 กรณีที่อินพุตเป็นสายลำดับจำนวน 100,000 เส้น และมีช่วงความถี่ของสายลำดับหลักระหว่าง 90-99% ซึ่งเป็นกรณีที่ให้ความแม่นเฉลี่ยของชุดสายลำดับแฮปโพลไทป์สูงสุด มีความแม่นเฉลี่ยของชุดแฮปโพลไทป์เป็นร้อยละ 94.99

ประมาณค่าความถี่แฮปโพลไทป์ที่ประกอบขึ้นโดยเลือกใช้วิธี 5.3 ในขั้นตอนการประกอบแฮปโพลไทป์ และประมาณค่าความถี่ด้วยค่าเฉลี่ยเลขคณิตของความถี่อัลลีลในกลุ่มของสายลำดับที่อ่านได้ที่เหมือนกับสายแฮปโพลไทป์หลัก ความถี่แฮปโพลไทป์ที่ได้มีความผิดพลาดสัมบูรณ์เฉลี่ยร้อยละ 4.76 และเมื่อสายลำดับหลักมีความถี่ระหว่าง 90-99 % ความถี่แฮปโพลไทป์ที่คำนวณได้มีความผิดพลาดสัมบูรณ์เพียงร้อยละ 0.91 แสดงรายละเอียดดังตารางที่ 6.3

ตารางที่ 6.3 ความผิดพลาดสัมบูรณ์ของความถี่แฮปโพลไทป์ที่คำนวณได้กับความถี่แฮปโพลไทป์จริง จากการประกอบแฮปโพลไทป์ทั้งชุดโดยใช้วิธี 5.3 ในขั้นตอนประกอบแฮปโพลไทป์สายหลัก

เส้นที่	ความผิดพลาดสัมบูรณ์ของความถี่แฮปโพลไทป์แยกตามช่วงความถี่สายลำดับหลัก (%)								
	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
1	4.69	10.79	16.58	15.22	6.16	3.18	2.13	2.99	0.93
2	3.16	10.16	16.73	15.29	7.09	5.01	1.49	1.72	0.92
3	4.14	10.45	14.23	9.30	4.12	5.49	2.04	1.43	0.79
4	4.48	9.19	8.94	3.35	2.70	3.55	1.33	1.00	1.16
5	4.40	7.77	4.31	2.57	2.03	2.39	1.00	1.58	0.79
6	4.08	5.09	1.48	1.46	1.74	1.17	0.61	0.67	-
7	3.79	3.52	1.40	1.13	1.70	0.93	-	-	-
8	3.62	2.34	1.08	0.67	1.80	1.12	-	-	-
9	3.25	1.64	-	-	-	-	-	-	-
10	3.00	1.66	-	-	-	-	-	-	-
ทั้งหมด	**2.97**	**6.49**	**10.52**	**8.49**	**4.45**	**3.51**	**1.73**	**1.92**	**0.91**

6.4 สรุป

ขั้นตอนวิธีประกอบชุดของแฮปโพลป์โดยประกอบแฮปโพลไทป์สายหลักทีละเส้น จากนั้นกรองสายลำดับเส้นที่คาดว่ามาจากแฮปโพลท์สายที่ประกอบได้นั้นออก แล้วนำสายลำดับที่เหลือมาประกอบแฮปโพลไทป์ลำดับถัดไป โดยใช้วิธี 5.3 ในขั้นตอนประกอบแฮปโพลไทป์สายหลักของแต่ละรอบให้ประสิทธิภาพดีกว่าที่เลือกใช้วิธี 5.2 โดยมีความแม่นเฉลี่ยร้อยละ 69.79 และจะให้ความแม่นสูงสุดเมื่อชุดของสายลำดับที่อ่านได้มีจำนวน 100,000 เส้น และความถี่ของสายลำดับหลักอยู่ระหว่าง 90-99 % โดยมีความแม่นเฉลี่ยร้อยละ 94.99

ดังนั้น จึงเลือกใช้วิธีประกอบชุดของแฮปโพลไทป์โดยเลือกใช้วิธี 5.3 คือ ประกอบแฮปโพลไทป์จากอัลลีลที่มีความถี่สูงสุดในแต่ละตำแหน่ง สำหรับการประกอบแฮปโพลไทป์สายหลักในแต่ละรอบ และใช้การประมาณความถี่ด้วยค่าเฉลี่ยเลขคณิตของความถี่อัลลีลบนสายลำดับที่อ่านได้ในกลุ่มที่เหมือนกับสายแฮปโพลไทป์หลักที่ประกอบได้ในรอบนั้น

บทที่ 7
การเปรียบเทียบประสิทธิภาพ

เปรียบเทียบประสิทธิภาพของวิธีประกอบชุดของแฮปโพลไทป์และประมาณความถี่แฮปโพลไทป์ที่นำเสนอ คือ การประกอบชุดแฮปโพลไทป์โดยประกอบแฮปโพลไทป์สายหลักทีละเส้นด้วยอัลลีลที่มีความถี่สูงสุดในแต่ละตำแหน่ง จากนั้นกรองสายลำดับเส้นที่คาดว่ามาจากแฮปโพลไทป์สายที่ประกอบได้นั้นออก แล้วนำสายลำดับที่เหลือมาประกอบแฮปโพลไทป์ลำดับถัดไป และประมาณค่าความถี่แฮปโพลไทป์แต่ละเส้นด้วยค่าเฉลี่ยเลขคณิตของความถี่อัลลีลบนสายลำดับในกลุ่มที่คาดว่ามาจากแฮปโพลไทป์สายนั้น เปรียบเทียบกับวิธีที่นำเสนอในงานวิจัยเรื่องการประมาณประชากรไวรัสโดยใช้ไพโรซีเควนซิง (Viral population estimation using pyrosequencing) ซึ่งเสนอโดย Eriksson และคณะ ในปี 2008 [17]

ในบทนี้จะได้กล่าวถึงภาพรวมของวิธีที่นำเสนอในงานวิจัยเรื่องการประมาณประชากรไวรัสโดยใช้ไพโรซีเควนซิง การเตรียมข้อมูลและขั้นตอนการทดสอบ จากนั้นจะนำเสนอผลการทดลองเปรียบเทียบกับวิธีที่นำเสนอในงานวิจัยนี้ และสรุปผล

7.1 วิธีที่นำเสนอในงานวิจัยเรื่องการประมาณประชากรไวรัสโดยใช้ไพโรซีเควนซิง

ในงานวิจัยนี้ได้แบ่งขั้นตอนการประมาณโครงสร้างประชากรของไวรัสออกเป็น 4 ขั้นตอนหลัก คือ การจัดเรียงสายลำดับที่อ่านได้เทียบกับจีโนมต้นแบบ (alignment) การแก้ไขข้อผิดพลาด (error correction) การประกอบแฮปโพลไทป์ (haplotype reconstruction) และการประมาณความถี่แฮปโพลไทป์ (haplotype frequency estimation) ซึ่งในงานวิจัยนี้ได้นำเสนอขั้นตอนวิธีสำหรับ 3 ขั้นตอนหลัง คือ การแก้ไขข้อผิดพลาด การประกอบแฮปโพลไทป์และการประมาณความถี่แฮปโพลไทป์ โดยมีรายละเอียดในบทที่ 2 ข้อ 2.2.1

สำหรับการประกอบแฮปโพลไทป์และการประมาณค่าความถี่แฮปโพลไทป์ซึ่งเป็นโจทย์เดียวกับงานวิจัยนี้ Eriksson และคณะ ได้เสนอให้นำสายลำดับที่อ่านได้มาสร้างกราฟ โดยแต่ละจุดยอดคือสายลำดับที่อ่านได้ที่ไม่ซ้ำกัน (irredundant read) และลากเส้นเชื่อมระหว่างจุดยอดโดยพิจารณาจากส่วนที่ซ้อนทับกัน (overlap) และตำแหน่งของสายลำดับนั้น โดยจะลากเส้นเชื่อมจากจุดยอด r_1 ไป r_2 เมื่อจุดยอด r_1 อยู่ก่อน r_2 มีส่วนที่ซ้อนทับกันเหมือนกัน และยังไม่มีเส้นทางอื่นจากจุดยอด r_1 ไป r_2 จากนั้นหาค่าครอบคลุมที่มีค่าน้อยสุด (minimal cover) ของกราฟของสายลำดับที่

อ่านได้ ด้วยขั้นตอนวิธีการจับคู่สูงสุด (maximum matching algorithm) แล้วขยายสายโซ่ (chain) ย่อยที่คำนวณได้นี้ให้เป็นเส้นทางจากจุดเริ่มต้นไปถึงจุดสิ้นสุด แต่ละเส้นทางคือแฮปโพลไทป์ที่ประกอบได้ นำชุดของแฮปโพลไทป์ที่ประกอบได้ และชุดของสายลำดับที่อ่านได้ทั้งหมดไปเป็นอินพุตสำหรับประมาณค่าความถี่ แฮปโพลไทป์ทั้งชุดด้วยขั้นตอนวิธีอีเอ็ม (EM algorithm)

Eriksson และคณะได้ทดสอบวิธีที่นำเสนอด้วยข้อมูลที่จำลองขึ้นและข้อมูลสายลำดับนิว คลีโอไทด์ของไวรัส เอชไอวี 1,000 ตำแหน่งเปรียบเทียบระหว่างแฮปโพลไทป์ที่ประกอบขึ้นจากวิธีที่นำเสนอโดยมีอินพุตเป็นสายลำดับที่อ่านด้วยเทคโนโลยีการอ่านสายลำดับแบบขนานจำนวนมาก และแฮปโพลไทป์ที่อ่านได้โดยตรงจากวิธีของ Sanger ซึ่งอ่านได้ความยาวครบทั้ง 1,000 ตำแหน่ง

สำหรับการทดสอบด้วยข้อมูลที่จำลองขึ้นนั้น ได้แยกทดสอบทีละขั้นตอนก่อน จากนั้นจึงทดสอบประสิทธิภาพรวมของขั้นตอนการประกอบแฮปโพลไทป์และการประมาณความถี่แฮปโพลไทป์ โดยให้อินพุตเป็นสายลำดับที่อ่านได้ซึ่งไม่มีข้อผิดพลาดและจัดเรียงตำแหน่งเรียบร้อยแล้ว สุดท้ายจึงทดสอบรวมทั้งสามขั้นตอนตั้งแต่การจัดการข้อผิดพลาดจนถึงการประมาณความถี่แฮปโพลไทป์

ในบทนี้เราจะทดสอบวิธีที่ได้นำเสนอในบทที่ 6 เทียบกับ 2 ขั้นตอนหลัง โดยให้อินพุตเป็นสายลำดับที่อ่านได้ที่ไม่มีข้อผิดพลาด และจัดเรียงเทียบกับจีโนมต้นแบบแล้ว ซึ่งเป็นข้อกำหนดเดียวกับการทดสอบประสิทธิภาพรวมของ 2 ขั้นตอนนี้ที่ได้นำเสนอในงานวิจัย

7.2 การทดสอบ

Eriksson และคณะได้พัฒนาโปรแกรมสำเร็จ (software package) ชื่อ “ShoRAH” (Short Read Assembly into Haplotypes) ขึ้นจากวิธีที่นำเสนอในบทความ และเผยแพร่ภายใต้ สัญญาอนุญาตสาธารณะทั่วไปของกนู (GNU General Public License) สามารถดาวน์โหลดได้จากเว็บไซต์ http://www.bsse.ethz.ch/cbg/software/shorah

ทดสอบขั้นตอนวิธีประกอบแฮปโพลไทป์และประมาณความถี่แฮปโพลไทป์ที่เสนอโดย Eriksson และคณะ โดยใช้อินพุตสำหรับโปรแกรมสำเร็จ ShoRAH เป็นข้อมูลสายลำดับที่อ่านได้ชุดเดียวกับที่ใช้ทดสอบขั้นตอนวิธีที่นำเสนอในงานวิจัยนี้ โดยแปลงให้อยู่ในรูปแบบที่เหมาะสมสำหรับแต่ละขั้นตอน ซึ่งเราจะใช้ชุดคำสั่ง mm.py และ freqEst.cc ในโปรแกรมสำเร็จ ShoRAH

สำหรับประมวลผลขั้นตอนวิธีการจับคู่สูงสุด (maximum matching algorithm) และขั้นตอนวิธีอีเอ็ม ตามลำดับ ซึ่งจะได้เอาท์พุทเป็นชุดของแฮปโพลไทป์และความถี่ของแฮปโพลไทป์แต่ละเส้น โดยมีขั้นตอนดังนี้

1) สร้างไฟล์อินพุต โดยใช้นามสกุล .read เก็บสายลำดับที่อ่านได้ทั้งหมดให้อยู่ในรูปแบบ 1 บรรทัด ต่อ 1 สายลำดับที่อ่านได้ แต่ละบรรทัดประกอบด้วย 2 สดมภ์ สดมภ์แรกเป็นตำแหน่งของเบสตัวแรกบนสายลำดับ และสดมภ์ที่สองเป็นสายลำดับที่อ่านได้ ถ้าสายลำดับใดอ่านได้มากกว่า 1 ครั้ง จะมีสายลำดับนี้ในไฟล์ .read ปรากฏตามจำนวนที่อ่านได้ ไฟล์ .read นี้มีรูปแบบดังแสดงในรูปที่ 7.1

```
0    tgttgg
0    tgttgg
0    tatttg
0    aaacgt
0    taatt
0    tgtt
0    tgtt
0    taat
1    gttgg
1    aacgt
1    aacg
1    aacg
2    acgt
2    acgt
2    tttg
2    ttgg
3    tggt
3    cgta
3    ttgt
4    gtactg
4    ggtacg
5    gtacg
6    tacgt
6    tcttt
```

รูปที่ 7.1 รูปแบบไฟล์ .read ของสายลำดับที่อ่านได้ โดยในสดมภ์แรกเป็นตำแหน่งของเบสตัวแรกบนสายลำดับและสดมภ์ที่สองเป็นสายลำดับที่อ่านได้

2) สร้างไฟล์ .rest เก็บข้อมูลสายลำดับเฉพาะสายลำดับที่ไม่ซ้ำกัน (irredundant read) คือ ไม่มีสายลำดับใดที่เป็นส่วนหนึ่งของสายลำดับอื่นในไฟล์นี้ เช่น จากไฟล์ในรูป 7.1

- สายลำดับที่ 1 และ 2 ซ้ำกัน จึงเก็บแค่สายลำดับเดียวในไฟล์ .rest

- สายลำดับที่ 6 และ 7 (`0 tgtt`) ซ้ำกับสายลำดับแรก (`0 tgttgg`) เนื่องจากเป็นสายอักขระย่อย (substring) ของสายลำดับที่ 1 และมีเบสเดียวกันในตำแหน่งเดียวกัน จึงไม่เก็บในไฟล์ .rest

- สายลำดับที่ 9 (`1 gttgg`) ซ้ำกับสายลำดับแรก (`0 tgttgg`) เนื่องจากเป็นสายอักขระย่อย (substring) ของสายลำดับที่ 1 และมีเบสเดียวกันในตำแหน่งเดียวกัน จึงไม่เก็บในไฟล์ .rest แสดงไฟล์ .rest ของสายลำดับในรูปที่ 7.1 ได้ดังรูปที่ 7.2

```
0    tgttgg
0    tatttg
0    aaacgt
0    taatt
3    tggt
3    cgta
3    ttgt
4    gtactg
4    ggtacg
6    tacgt
6    tcttt
```

รูปที่ 7.2 รูปแบบไฟล์ .rest เก็บสายลำดับที่อ่านได้เฉพาะสายลำดับที่ไม่ซ้ำกัน

3) นำไฟล์ .rest ป้อนเป็นอินพุตของชุดคำสั่ง mm.py เพื่อประมวลผลด้วยขั้นตอนวิธีการจับคู่สูงสุด ได้ผลลัพธ์เป็นไฟล์นามสกุล .geno มีรูปแบบดังรูปที่ 7.3 แต่ละบรรทัดแสดงแฮปโพลไทป์ที่ได้จากขั้นตอนวิธีการจับคู่สูงสุด

```
atgaatacttacaaaatcagtctcacgaatgctgaccagagagtccccgagatccatcct
ttcaatagttataaaatgtgtctcacgattgcagatcactgagttccctgagaaggctct
acccttttgctataactggtgtcagacgaaggctaatctcaaagatactgagatagattat
acgatgtgtaccttaagcaccgtgcatatttaaaattagtatcttatcggagaaggttaa
ttcaagactaatttattgtgcctgcctcttgctagttactaagacattgaaaaacattca
```

รูปที่ 7.3 รูปแบบไฟล์ .geno เก็บแฮปโพลไทป์ที่ได้จากขั้นตอนวิธีการจับคู่สูงสุด ซึ่งเป็นจำนวนแฮปโพลไทป์ที่น้อยที่สุดที่สามารถอธิบายสายลำดับที่อ่านได้ทั้งหมดได้ โดยแต่ละบรรทัดคือแฮปโพลไทป์แต่ละเส้น

4) นำไฟล์ .read และ .geno ป้อนเป็นอินพุตของชุดคำสั่ง freqEst เพื่อประกอบชุดของแฮปโพลไทป์และประมาณค่าความถี่ของแต่ละแฮปโพลไทป์ โดยแฮปโพลไทป์ที่ได้จะมีจำนวนมากกว่าแฮปโพลไทป์ที่ได้จากข้อ 3) ผลที่ได้จะถูกเก็บในไฟล์ .popl โดยแต่ละบรรทัดคือแต่ละแฮปโพลไทป์ เก็บในรูปแบบ

>HAPLOTYPE_NAME FREQUENCY HAPLOTYPE_SEQUENCE

แสดงตัวอย่างของรูปแบบไฟล์ .popl ที่ได้ ดังรูปที่ 7.4

```
> HAP0  0.016093   acccttgctataactggtgtcagacgaaggctaattagtat
> HAP1  0.087744   acgatgtgtaccttaagcaccgtgcatatttaaaattagtat
> HAP2  0.000233   atgaatactaatttattgtgtctcacgattgcaaatcactga
> HAP3  0.003091   atgaatacttacaaaatcagtctcacgattgcagatcactga
> HAP4  0.011951   ttcaagacttacaaaatcagcctgcctcttgctagttactaa
> HAP5  0.001615   ttcaatagttataaaatgtgcctgcctcttgctagttactga
> HAP6  0.008305   ttcaatagttataaaatgtgtctcacgaaggctaatctcaaa
> HAP7  0.002439   ttcaatagttataaaatgtgtctcacgaatgctgaccagaga
> HAP8  0.207654   ttcaatagttataaaatgtgtctcacgattgcaaatcactga
> HAP9  0.008419   ttcaatagttataaaatgtgtctcacgattgcaaatcactga
> HAP10 0.000469   ttcaatagttataaaatgtgtctcacgattgcaaatcactga
```

รูปที่ 7.4 รูปแบบไฟล์ .popl เก็บแฮปโพลไทป์และความถี่ของแฮปโพลไทป์แต่ละเส้นที่ได้ โดยสดมภ์ที่สองแสดงความถี่แฮปโพลไทป์ และสดมภ์ที่สามคือสายลำดับของแฮปโพลไทป์

5) นำผลจากไฟล์ .popl ซึ่งเป็นชุดของแฮปโพลไทป์และความถี่แฮปโพลไทป์แต่ละเส้นที่ประกอบได้จากขั้นตอนวิธีที่เสนอในงานวิจัยเรื่อง การประมาณประชากรไวรัสโดยใช้ไพโรซีเควนซิง มาคำนวณความแม่นเทียบกับชุดของแฮปโพลไทป์ที่จำลองขึ้น โดย เรียงตามลำดับความถี่ คือ เปรียบเทียบแฮปโพลไทป์ที่จำลองขึ้นเส้นที่มีความถี่สูงสุด กับแฮปโพลไทป์ที่มีความถี่สูงสุดที่ได้จากการประกอบด้วยวิธีนี้ไล่ไปจนถึงเส้นที่มีความถี่น้อยที่สุด

6) เปรียบเทียบความแม่นของสายแฮปโพลไทป์และความถี่แฮปโพลไทป์ที่ได้จากวิธีนี้เทียบกับวิธีที่นำเสนอในบทที่ 6

7.3 ผลการทดลอง

ใช้ชุดของสายลำดับที่จำลองขึ้นชุดเดียวกับการทดสอบขั้นตอนวิธีในบทที่ 6 ได้ความแม่นเฉลี่ยของแฮปโพลไทป์ แยกตามช่วงความถี่ของสายลำดับหลักและลำดับของแฮปโพลไทป์ได้ดังตารางที่ 7.1

ตารางที่ 7.1 ความแม่นเฉลี่ยของแฮปโพลไทป์ที่ได้จากการประกอบแฮปโพลไทป์ทั้งชุดด้วยวิธีที่นำเสนอโดย Eriksson และคณะ

เส้นที่	ความแม่นเฉลี่ยของแฮปโพลไทป์แยกตามช่วงความถี่สายลำดับหลัก (%)								
	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
1	70.24	68.15	71.71	72.72	75.52	77.51	80.87	85.29	88.86
2	63.73	62.12	62.35	65.95	61.98	61.37	58.62	59.87	50.82
3	62.39	61.19	60.10	59.83	57.38	58.98	54.54	55.64	47.95
4	62.25	60.71	61.10	56.42	56.59	58.05	55.27	51.87	57.59
5	64.14	60.66	58.11	54.60	57.92	57.84	56.44	56.63	52.22
6	62.19	59.66	57.72	57.22	59.91	56.70	56.80	60.35	-
7	62.29	58.46	59.07	57.84	57.73	58.73	-	-	-
8	62.40	59.62	58.15	59.24	58.14	59.20	-	-	-
9	62.42	58.85	-	-	-	-	-	-	-
10	61.51	58.76	-	-	-	-	-	-	-
ทั้งหมด	62.60	60.99	62.04	61.81	62.10	61.97	62.28	63.08	63.35

จากตารางความแม่นของแฮปโพลไทป์สายแรกๆ สูงกว่าแฮปโพลไทป์ลำดับท้ายๆ โดยความแม่นเฉลี่ยของแฮปโพลไทป์สายแรกของสายลำดับที่มีความถี่ของสายลำดับหลักอยู่ในช่วง 90-99% เป็นร้อยละ 88.86 และความแม่นเฉลี่ยของแฮปโพลไทป์ทั้งหมดเป็นร้อยละ 62.18

นำความถี่ของแฮปโพลไทป์ที่ประมาณได้ไปเปรียบเทียบกับความถี่ของแฮปโพลไทป์ที่จำลองขึ้น แสดง ความผิดพลาดสัมบูรณ์ ของความถี่แฮปโพลไทป์ที่ประมาณได้เทียบกับความถี่แฮปโพลไทป์ที่จำลองขึ้นได้ดังตารางที่ 7.2

ตารางที่ 7.2 ความผิดพลาดสัมบูรณ์ของความถี่แฮปโพลไทป์ที่คำนวณได้กับความถี่แฮปโพลไทป์จริง จากการประกอบแฮปโพลไทป์ทั้งชุดด้วยวิธีที่นำเสนอโดย Eriksson และคณะ

เส้นที่	ความผิดพลาดสัมบูรณ์ของความถี่แฮปโพลไทป์แยกตามช่วงความถี่สายลำดับหลัก (%)								
	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
1	10.15	13.16	20.16	25.78	33.73	44.65	44.06	50.11	39.17
2	4.49	8.62	12.5	14.14	13.59	5.14	6.58	9.84	17.38
3	4.22	7.09	8.9	5.79	5.56	4.54	9.48	10.64	10.67
4	3.96	5.22	4.12	5.68	5.85	5.53	8.2	7.44	7.07
5	3.54	3.83	4.92	5.63	5.26	5.07	6.11	6.11	4.33
6	2.95	2.62	4.53	4.66	4.19	4.88	4.59	5.56	-
7	2.44	2.86	4.33	3.62	3.74	4.21	-	-	-
8	2.30	2.58	3.87	3.94	1.92	2.65	-	-	-
9	1.90	2.78	-	-	-	-	-	-	-
10	1.55	2.49	-	-	-	-	-	-	-
ทั้งหมด	**2.50**	**5.28**	**9.28**	**10.85**	**12.75**	**11.71**	**16.62**	**18.72**	**21.94**

เปรียบเทียบความแม่นของสายลำดับแฮปโพลไทป์ที่ประกอบด้วยวิธีที่เสนอโดย Eriksson และคณะกับวิธี ประกอบชุดแฮปโพลไทป์จากการประกอบแฮปโพลไทป์สายหลักทีละเส้นด้วยอัลลีลที่มีความถี่สูงสุดในแต่ละตำแหน่ง ซึ่งเป็นวิธีที่นำเสนอในงานวิจัยนี้ ได้ดังตารางที่ 7.3

ตารางที่ 7.3 ความแม่นเฉลี่ยของชุดแฮปโพลไทป์ที่ประกอบขึ้น เปรียบเทียบระหว่างวิธีที่เสนอโดย Eriksson และคณะกับวิธีที่นำเสนอในงานวิจัยนี้

เส้นที่	ความแม่นเฉลี่ยของแฮปโพลไทป์แยกตามช่วงความถี่สายลำดับหลัก (%)								
	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
1	**70.24**	**68.15**	**71.71**	**72.72**	**75.52**	**77.51**	**80.87**	**85.29**	**88.86**
	81.60	81.32	84.96	89.08	96.03	99.63	99.83	99.98	100.00
2	**63.73**	**62.12**	**62.35**	**65.95**	**61.98**	**61.37**	**58.62**	**59.87**	**50.82**
	69.71	68.06	69.08	75.89	88.63	80.16	92.66	86.87	76.99
3	**62.39**	**61.19**	**60.10**	**59.83**	**57.38**	**58.98**	**54.54**	**55.64**	**47.95**
	67.12	67.07	64.89	63.35	69.19	71.63	71.17	62.82	65.75
4	**62.25**	**60.71**	**61.10**	**56.42**	**56.59**	**58.05**	**55.27**	**51.87**	**57.59**

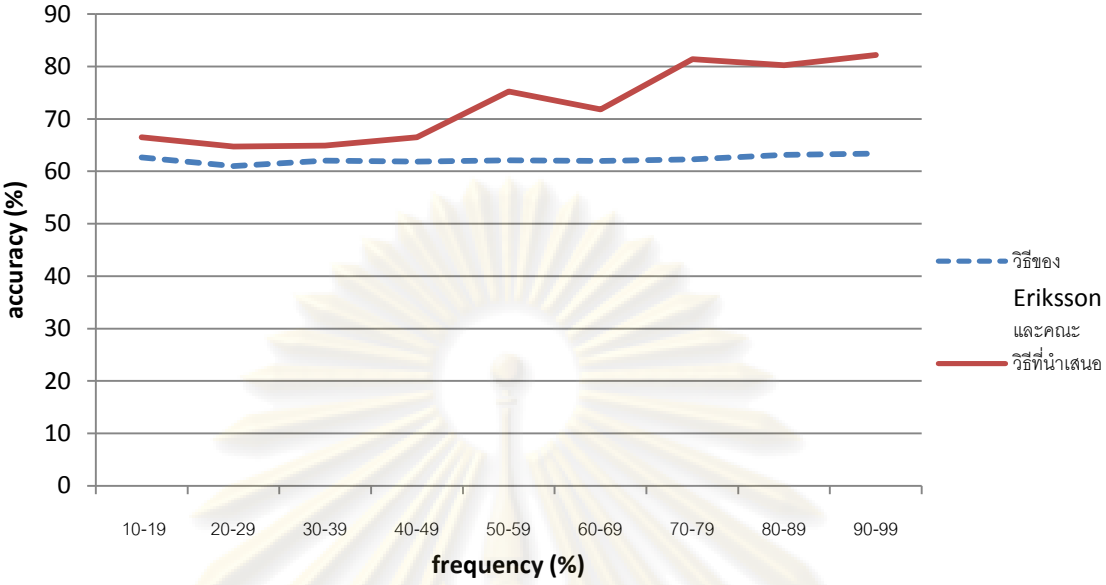
เส้นที่	ความแม่นเฉลี่ยของแฮปโพลไทป์แยกตามช่วงความถี่สายลำดับหลัก (%)								
	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
	67.40	65.51	57.75	54.56	63.03	61.58	58.93	58.83	60.54
5	**64.14**	**60.66**	**58.11**	**54.60**	**57.92**	**57.84**	**56.44**	**56.63**	**52.22**
	68.66	64.77	56.32	54.39	57.29	58.36	56.96	59.70	56.82
6	**62.19**	**59.66**	**57.72**	**57.22**	**59.91**	**56.70**	**56.80**	**60.35**	-
	67.30	62.06	53.04	54.02	56.57	55.62	54.02	55.50	-
7	**62.29**	**58.46**	**59.07**	**57.84**	**57.73**	**58.73**	-	-	-
	66.61	59.48	57.22	55.64	56.50	56.59	-	-	-
8	**62.40**	**59.62**	**58.15**	**59.24**	**58.14**	**59.20**	-	-	-
	65.79	58.70	56.66	55.35	59.93	60.20	-	-	-
9	**62.42**	**58.85**	-	-	-	-	-	-	-
	66.40	57.39	-	-	-	-	-	-	-
10	**61.51**	**58.76**	-	-	-	-	-	-	-
	64.95	59.73	-	-	-	-	-	-	-
ทั้งหมด	**62.60**	**60.99**	**62.04**	**61.81**	**62.10**	**61.97**	**62.28**	**63.08**	**63.35**
	66.50	64.73	64.90	66.52	75.22	71.82	81.40	80.23	82.17

■ *แสดง ความแม่นเฉลี่ยของแฮปโพลไทป์ที่ประกอบขึ้นจากวิธีที่เสนอโดย Eriksson และคณะ*

■ *แสดงความแม่นเฉลี่ยของแฮปโพลไทป์ที่ประกอบขึ้นจากวิธีที่นำเสนอในงานวิจัยนี้*

- *คือ แฮปโพลไทป์เส้นนั้นไม่มีอยู่จริงในอินพุตของขั้นตอนวิธี*

นำความแม่นเฉลี่ยของสายลำดับแฮปโพลไทป์ทั้งหมดมาสร้างกราฟเปรียบเทียบระหว่างวิธีที่เสนอโดย Eriksson และคณะ และวิธีที่เสนอในงานวิจัยนี้ได้ดังรูปที่ 7.5



รูปที่ 7.5 กราฟความแม่นเฉลี่ยของชุดสายลำดับแฮปโพลไทป์ที่ประกอบขึ้นเปรียบเทียบระหว่างวิธีที่เสนอโดย Eriksson และคณะ และวิธีที่นำเสนอในงานวิจัยนี้

จากผลการทดลอง จะเห็นว่าวิธีที่นำเสนอให้ความแม่นของสายลำดับแฮปโพลไทป์สูงกว่าวิธีที่นำเสนอโดย Eriksson และคณะ สำหรับชุดของสายลำดับนี้ ในทุกช่วงความถี่ของสายลำดับหลัก อย่างไรก็ตามสำหรับแฮปโพลไทป์เส้นท้ายๆ ของชุดสายลำดับที่มีความถี่ในช่วง 30-39, 40-49 และ 50-59% วิธีที่เสนอโดย Eriksson และคณะให้ความแม่นเฉลี่ยสูงกว่าวิธีที่เสนอในงานวิจัยนี้

7.4 สรุป

วิธีที่นำเสนอให้ความแม่นของแฮปโพลไทป์ที่ประกอบขึ้นสูงกว่าวิธีที่เสนอในงานวิจัยเรื่องการประมาณประชากรไวรัสโดยใช้ไพโรซีเควนซิง สำหรับชุดข้อมูลทดสอบนี้ ซึ่งเป็นสายลำดับที่จำลองจากจีโนมของไวรัสเด็งกี่ โดยมีอัตราการแปรผัน 2% และสนิปส์แต่ละตำแหน่งมีเพียง 2 อัลลีล และจำลองการอ่านตามแบบเครื่องอ่านสายลำดับเบส Roche GS FLX ทำให้สายลำดับที่อ่านได้มีความยาวเฉลี่ย 250 คู่เบส

บทที่ 8
สรุปผลการวิจัย และข้อเสนอแนะ

8.1 สรุปผลการวิจัย

เทคโนโลยีอ่านสายลำดับนิวคลีโอไทด์แบบขนานจำนวนมากทำให้การศึกษาความหลากหลายของสิ่งมีชีวิตแบบกึ่งสปีชีส์ เช่น ไวรัสเด็งกี่ สามารถทำได้ จากเดิมที่การศึกษาความหลากหลายทางพันธุกรรมของสิ่งมีชีวิตประเภทนี้ทำได้ยากเนื่องจากข้อจำกัดของเทคโนโลยีอ่านสายลำดับนิวคลีโอไทด์ โดยเมื่อศึกษาความเป็นไปได้ในการใช้เครื่องอ่านสายลำดับนิวคลีโอไทด์ของ Roche ซึ่งพัฒนาโดยกลุ่ม 454 Life Sciences ซึ่งเป็นผู้นำของเทคโนโลยีนี้ และมีผลิตภัณฑ์ออกสู่ตลาดเป็นรายแรก พบว่า สามารถอ่านส่วนที่มีการผันแปรได้ในสัดส่วนที่ถูกต้อง สามารถจัดเรียงสายลำดับที่อ่านได้กับเส้นแม่แบบได้ความถูกต้อง สายลำดับที่อ่านได้ครอบคลุมทั้งจีโนมและความถี่ของสารตั้งต้นมีผลต่อความถี่ของสายลำดับที่อ่านได้ แสดงให้เห็นว่าสามารถใช้เทคโนโลยีอ่านสายลำดับนิวคลีโอไทด์ของกลุ่ม 454 Life Sciences ในการศึกษาความหลากหลายทางพันธุกรรมของไวรัสเด็งกี่ซึ่งเป็นสิ่งมีชีวิตที่มีรูปแบบกึ่งสปีชีส์ได้

ในการศึกษาความหลากหลายทางพันธุกรรมของไวรัสเด็งกี่ด้วยเทคโนโลยีอ่านสายลำดับนิวคลีโอไทด์แบบขนานจำนวนมาก จะเตรียมตัวอย่างเป็นประชากรของไวรัสซึ่งประกอบด้วยจีโนมหลายเส้น หลายรูปแบบซึ่งแต่ละแบบมีความแตกต่างกันเพียงเล็กน้อย นำตัวอย่างนี้ไปอ่านด้วยเครื่องอ่านสายลำดับเบส จากนั้นนำสายลำดับที่อ่านได้มาประกอบเป็นชุดของแฮปโพลไทป์ซึ่งเป็นปัญหาที่ซับซ้อน ในงานวิจัยนี้จึงนำเสนอให้แยกประกอบทีละเส้น โดยประกอบแฮปโพลไทป์สายหลัก หรือแฮปโพลไทป์ที่มีความถี่สูงสุดในประชากรนั้นก่อน แล้วกรองเอาสายลำดับที่อ่านได้จากแฮปโพลไทป์สายหลักนี้ออกไป แล้วนำสายลำดับที่เหลือมาประกอบแฮปโพลไทป์เส้นถัดไปซึ่งมีความถี่รองลงมา และประมาณค่าความถี่แฮปโพลไทป์แต่ละเส้นด้วยความถี่อัลลีลที่อยู่บนแฮปโพลไทป์นั้น

เมื่อแยกพิจารณาเฉพาะการประกอบแฮปโพลไทป์สายหลัก พบว่าวิธีประกอบแฮปโพลไทป์สายหลักที่มีประสิทธิภาพสูงคือ การประกอบแฮปโพลไทป์จากสายลำดับที่อ่านได้ที่มีความถี่สูงสุดโดยสุ่มตำแหน่ง และการประกอบแฮปโพลไทป์จากอัลลีลที่มีความถี่สูงสุดในแต่ละตำแหน่ง ซึ่งให้ความแม่นเฉลี่ยร้อยละ 98.46 และ 92.07 ตามลำดับ แต่เมื่อนำวิธีทั้งสองไปใช้ในการ

ประกอบชุดของแฮปโพลไทป์ซึ่งประกอบด้วยขั้นตอนหลักคือการประกอบแฮปโพลไทป์สายหลักและขั้นตอนกรองสายลำดับที่มาจากแฮปโพลไทป์ที่ประกอบได้ทิ้งไป พบว่าการประกอบแฮปโพลไทป์จากอัลลีลที่มีความถี่สูงสุดในแต่ละตำแหน่งให้ประสิทธิภาพสูงกว่า โดยมีความแม่นเฉลี่ยร้อยละ 69.79 จากการทดลองพบว่า ความแม่นของแฮปโพลไทป์ที่ประกอบได้แปรตามจำนวนสายลำดับที่อ่านได้ ความถี่ของสายลำดับหลัก และลำดับของแฮปโพลไทป์ โดยถ้าอ่านสายลำดับเข้ามาเป็นอินพุตมากความแม่นจะเพิ่มขึ้น แฮปโพลไทป์ของสายลำดับที่มีความถี่ของสายลำดับหลักมากจะมีความแม่นสูง และแฮปโพลไทป์ลำดับแรกๆ จะมีความแม่นสูงกว่าแฮปโพลไทป์ลำดับถัดไป กรณีที่อ่านสายลำดับเข้ามาทั้งสิ้น 100,000 เส้น จากประชากรที่มีความถี่ของสายลำดับหลักอยู่ในช่วง 90-99% จะได้ความแม่นเฉลี่ยของแฮปโพลไทป์ทั้งหมดเป็นร้อยละ94.99

เมื่อเปรียบเทียบประสิทธิภาพของวิธีที่นำเสนอกับวิธีที่เสนอโดย Eriksson และคณะ โดยวัดประสิทธิภาพจากความแม่นของสายลำดับแฮปโพลไทป์ที่ประกอบได้เทียบกับชุดของแฮปโพลไทป์ในประชากรที่จำลองขึ้น พบว่าวิธีที่นำเสนอให้ประสิทธิภาพสูงกว่า สำหรับชุดข้อมูลนี้ซึ่งจำลองขึ้นจากจีโนมของไวรัสเด็งกี่ ซึ่งมีความยาวประมาณ 10,000 คู่เบส โดยจำลองให้มีอัตราการแปรผันร้อยละ 2 และมีความยาวของสายลำดับที่อ่านได้เฉลี่ย 250 คู่เบส

จากงานวิจัยนี้จะเห็นว่า สามารถประกอบสายลำดับของแฮปโพลไทป์ทั้งชุดได้โดยการประกอบแฮปโพลไทป์ทีละสาย จากแฮปโพลไทป์สายหลักของแต่ละรอบซึ่งมีอินพุตของรอบนั้นๆ เป็นสายลำดับที่อ่านได้ซึ่งกรองเอาสายลำดับที่คาดว่ามาจากแฮปโพลไทป์สายหลักในรอบก่อนหน้าทิ้งไปแล้ว

8.2 ข้อเสนอแนะ

วิธีประกอบแฮปโพลไทป์ที่ได้นำเสนอนี้ประกอบด้วย 2 ขั้นตอนหลัก คือ ขั้นตอนประกอบแฮปโพลไทป์และขั้นตอนกรองสายลำดับที่คาดว่ามาจากแฮปโพลไทป์ที่ประกอบได้ ดังนั้นจึงสามารถพัฒนาวิธีที่นำเสนอได้โดยพัฒนา 2 ขั้นตอนหลักนี้ โดยปรับปรุงวิธีประกอบแฮปโพลไทป์ให้แม่นยิ่งขึ้นโดยพยายามประกอบแฮปโพลไทป์ให้มีความถี่สูงสุด แทนที่จะเลือกประกอบจากสายลำดับที่อ่านเข้ามาได้สูงสุดที่มีส่วนซ้อนทับยาวสุด หรือเลือกจากอัลลีลที่มีความถี่สูงสุดเพียงอย่างเดียว และอาจปรับปรุงวิธีประกอบแฮปโพลไทป์ให้ทนต่อสัญญาณรบกวนมากยิ่งขึ้น ซึ่งสัญญาณรบกวนนี้จะมีมากขึ้นตามจำนวนรอบของการประกอบแฮปโพลไทป์ หรืออาจ

ปรับปรุงขั้นตอนการกรองสายลำดับให้กรองได้มีประสิทธิภาพยิ่งขึ้น ซึ่งจะส่งผลให้สัญญาณรบกวนสำหรับรอบถัดไปน้อยลง

นอกจากการพัฒนาวิธีที่นำเสนอแล้ว การวิเคราะห์ปัจจัยที่มีผลต่อความแม่นของแฮปโพลไทป์ที่ประกอบขึ้น สำหรับประมาณความเชื่อมั่นของสายลำดับแฮปโพลไทป์นั้น มีประโยชน์ต่อการนำสายแฮปโพลไทป์ที่ประกอบได้ไปใช้ในการศึกษาโครงสร้างของประชากรไวรัส หรือใช้เป็นเครื่องหมายทางพันธุกรรมเพื่อทำนายอัตราเสี่ยงต่อการเป็นโรค หรือเพื่อศึกษากลไกการเกิดโรคต่อไป

นอกจากนี้ควรทดลองนำวิธีที่นำเสนอนี้ไปทดสอบด้วยสายลำดับที่อ่านได้จากเทคโนโลยีอ่านสายลำดับนิวคลีโอไทด์แบบขนานจำนวนมากวิธีอื่นๆ เช่น แพลตฟอร์มของ Illumina หรือ ABI SOLid ซึ่งสายลำดับที่อ่านได้สั้นกว่า แต่มีจำนวนเบสที่อ่านได้มากกว่า

รายการอ้างอิง

[1] Diseases of Environmental and Zoonotic Origin Team. *Dengue worldwide: an overview of the current situation and the implications for Europe*. Euro Surveill 12 (June 2007):

[2] สุจิตรา นิมมานนิตย์. *ไข้เลือดออก* (*Dengue and Dengue Hemorrhogic Fever)*. ใน นลินี อัศวโภคี, สุรภี เทียนกริม, ศศิธร ลิขิตนุกูล และ อัศฎา วิภากุล, ประสบการณ์ด้านโรคติดเชื้อในประเทศไทย, หน้า 13-26. กรุงเทพมหานคร: โฮลิสติก พับลิชชิ่ง, 2542.

[3] ผศ.นท.นพ. ชิษณุ พันธุ์เจริญ. *ฝนมา ไข้เลือดออกก็มา*. นิตยสารรักลูก, 2547.

[4] Chakravarti, A. *It's raining SNPs, hallelujah?*. Nature Genetics 19 (1998): 216-217.

[5] Zhao, Y. Y.; Wu, L. Y.; Zhang, J. H.; Wang, R. S.; and Zhang, X. S. *Haplotype assembly from aligned weighted SNP fragments*. Computational Biology and Chemistry 29 (August 2005): 281-287.

[6] Pe'er, I., and Beckmann, J. S. *Resolution of Haplotype and Haplotype Frequencies from SNP Genotypes of Pooled Samples*. Proceedings of the seventh annual international conference on Research in computational molecular biology, ACM Press, pp. 237-246. Berlin, Germany, 2003.

[7] Li, L.; Kim, J. H.; and Waterman, M. S. *Haplotype reconstruction from SNP alignment.* Proceedings of the seventh annual international conference on Research in computational molecular biology, ACM Press, pp. 207-216. Berlin, Germany, 2003.

[8] Wang, Y.; Feng, E.; Wang, R.; and Zhang, D. *The haplotype assembly model with genotype information and iterative local-exhaustive search algorithm*. Computational Biology and Chemistry 31 (August 2007): 288-293.

[9] Boëlle, P. Y. *Statistical and computational methods for haplotype reconstruction*. Applied Stochastic Models and Data Analysis, ASMDA 2005, pp. 161-166. Brest, France, 2005.

[10] Stephens, M. and Donnelly, P. *A comparison of bayesian methods for haplotype reconstruction from population genotype data*. American journal of human genetics 73 (November 2003): 1162-1169.

[11] Stephens, M.; Smith, N. J.; and Donnelly, P. *A New Statistical Method for Haplotype Reconstruction from Population Data*. American journal of human genetics 68 (April 2004): 978-989.

[12] Wang, W. K.; Lin, S.R.; Lee, C. M.; King, C.C.; and Chang, S.C. *Dengue Type 3 Virus in Plasma Is a Population of Closely Related Genomes: Quasispecies*. Journal of Virology 76 (May 2002): 4662-4665.

[13] Margulies, M., et al. *Genome sequencing in microfabricated high-density picolitre reactors*. Nature 437 (September 2005): 376-380.

[14] Sundquist, A.; Ronaghi, M.; Tang, H.; Pevzner, P.; and Batzoglou, S. *Whole-Genome Sequencing and Assembly with High-Throughput, Short-Read Technologies*. PLoS ONE 2 (May 2007): e484.

[15] Brown, T. A. *Genomes*. 2nd ed. Oxford: BIOS Scientific, 2001.

[16] Hall, N. *Advanced sequencing technologies and their wider impact in microbiology*. Journal of Experimental Biology 209 (2007): 1518-1525.

[17] Eriksson, N., et al. *Viral population estimation using pyrosequencing.* *PLoS Computational Biology* 4 (May 2008): e1000074

[18] Domselaar, G. V., et al. *Q Assembler. Available from: http://www.bioinformatics.org/qassembler/wiki/[2009, March 7]*

[19] Park, H.W., et al. *Association between genetic variations of vascular endothelial growth factor receptor 2 and atopy in the Korean population*. Journal of Allergy and Clinical Immunology 117 (April 2006): 774-779.

[20] Tovar, F.; Chiurillo, M. A.; Borjas, L.; Lander, N.; and J. L. Ramírez. *Chromosome Y haplotypes database in a Venezuelan population.* the 21st International ISFG Congress, pp. 246-248. Ponta Delgada, The Azores, Portugal, 2003

[21] Kim, J. H.; Waterman, M. S.; and Li, L. M. *Diploid genome reconstruction of Ciona intestinalis and comparative analysis with Ciona savignyi*. Genome Research 17 (2007): 1101-1110.

[22] Arrivillaga J. C.; Norris D. E.; Feliciangeli M. D.; and Lanzaro G. C., *Phylogeography of the neotropical sand fly Lutzomyia longipalpis inferred from mitochondrial DNA sequences.* Infection, Genetics and Evolution 2(December 2002): 83-95.

[23] Stevanin, G.; Sousa, P. S.; Cancel, G.; and Dürr, A. *The gene for Machado–Joseph disease maps to the same 3-cM interval as the spinal cerebellar ataxia 3 gene on chromosome 14q*. Neurobiology of Disease 1(November 1994): 79-82.

[24] Eronen, L.; Geerts, F.; and Toivonen, H. *A Markov Chain Approach to Reconstruction of Long Haplotypes*. Pacific Symposium on Biocomputing 9 (2004): 104-115.

[25] Rastas, P.; Koivisto, M.; Mannila, H.; and Ukkonen, E. *A Hidden Markov Technique for Haplotype Reconstruction*. Lecture Notes in Computer Science 3692 (2005): 140-151.

[26] Hajirasouliha, I.; Hormozdiari, F.; Sahinalp, S. C.; and Birol, I. *Optimal pooling for genome re-sequencing with ultra-highthroughput short-read technologies*. Bioinformatics 24 (July 2008): i30-i40.

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวพุธิตา สุมานนท์ เกิดเมื่อวันที่ 19 มีนาคม พ.ศ. 2528 ที่โรงพยาบาลจุฬาลงกรณ์ สำเร็จการศึกษาระดับประถมศึกษาจากโรงเรียนอนุบาลกำแพงเพชร จังหวัดกำแพงเพชร สำเร็จการศึกษาระดับมัธยมศึกษาจากโรงเรียนเตรียมอุดมศึกษาพัฒนาการ จากนั้นเข้าศึกษาต่อ ณ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย และสำเร็จการศึกษาหลักสูตรวิศวกรรมศาสตรบัณฑิต (วิศวกรรมคอมพิวเตอร์) ในปีการศึกษา 2549

มีผลงานทางวิชาการที่ได้รับการตีพิมพ์ คือ บทความทางวิชาการในหัวข้อเรื่อง *"การจำลองแบบการประกอบ สายลำดับหลักของสนิปส์ใน ไวรัส เด็งกี่ ด้วยเทคโนโลยีการอ่านลำดับนิวคลีโอไทด์แบบขนานจำนวนมาก : Major SNPs Sequence Assembling Simulation of Dengue Virus Genome from Massively Parallel Sequencing Technique"* โดย พุธิตา สุมานนท์, ศ.ดร. ประภาส จงสถิตย์วัฒนา และ ดร.ประพัฒน์ สุริยผล ในงานประชุมวิชาการ "The 12th National Computer Science and Engineering Conference (NCSEC2008)" ระหว่างวันที่ 20-21 พฤศจิกายน 2551

ในระหว่างการศึกษาระดับปริญญามหาบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์นี้ ได้รับทุนอุดหนุนการศึกษาระดับบัณฑิตศึกษา จุฬาลงกรณ์มหาวิทยาลัย เพื่อเฉลิมฉลองวโรกาสที่พระบาทสมเด็จพระเจ้าอยู่หัวทรงเจริญพระชนมายุครบ 72 พรรษา ตลอดระยะเวลาศึกษา