การจำลองแบบการประกอบสายลำดับหลักของสนิปส์ในไวรัสเด็งกี่ด้วยเทคโนโลยีการอ่านลำดับนิวคลีโอไทด์แบบขนานจำนวนมาก

Major SNPs Sequence Assembling Simulation of Dengue Virus Genome from Massively Parallel Sequencing Technique

พุธิตา สุมานนท์[1] ประภาส จงสถิตย์วัฒนา[1] และ ประพัฒน์ สุริยผล[2]

[1]ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

254 ถนนพญาไท แขวงวังใหม่ เขตปทุมวัน กทม. 10330 โทรศัพท์ : 0-2719-3872 E-mail: u46psm@cp.eng.chula.ac.th, prabhas.c@chula.ac.th

[2]หน่วยอณูชีววิทยาการแพทย์ คณะแพทยศาสตร์ศิริราชพยาบาล

ตึกอดุลยเดชวิกรม ชั้น 12 รพ.ศิริราช ถ.พรานนก บางกอกน้อย กรุงเทพฯ 10700 โทรศัพท์ : 0-2418-4793 E-mail: sipur@mahidol.ac.th

บทคัดย่อ

สนิปส์เป็นความแตกต่างทางพันธุกรรมพื้นฐานที่พบบ่อยที่สุด ใช้เป็นเครื่องหมายทางพันธุกรรมในการวินิจฉัยโรค ช่วยให้การวินิจฉัยโรคมีประสิทธิภาพ ถูกต้องและรวดเร็วยิ่งขึ้น เครื่องหมายทางพันธุกรรมนี้จะมีประสิทธิภาพเพิ่มขึ้นเมื่อใช้สายลำดับของสนิปส์ตลอดทั้งเส้นแทน ที่จะใช้สนิปส์เพียงตำแหน่งเดียว ในอดีตการศึกษาความแตกต่างทางพันธุกรรมนี้ในสิ่งมีชีวิตที่มีลักษณะกึ่งสปีชีส์ไม่สามารถทำได้ด้วยข้อจำกัดทางเทคนิค แต่ด้วยเทคโนโลยีการอ่านสายลำดับนิวคลีโอไทด์แบบขนานจำนวนมากที่พัฒนาขึ้นในปัจจุบันทำให้การศึกษาความหลากหลายนี้ทำได้ในเวลาอันสั้น งานวิจัยนี้ได้เสนอขั้นตอนวิธีในการประกอบสายลำดับหลักของสนิปส์ของไวรัสเด็งกี่จากการจำลองข้อมูลที่ได้จากเครื่องอ่านสายลำดับเบสรุ่น Roche GS FLX ซึ่งเป็นเครื่องรุ่นล่าสุดของ 454 Life Sciences ผู้พัฒนาเทคโนโลยีนี้จนวางตลาดสำเร็จเป็นรายแรก จากขั้นตอนวิธีที่นำเสนอสามารถประกอบสายลำดับหลักของสนิปส์ได้ที่ความแม่นสูงสุดร้อยละ 98.60 และมีความแม่นของความถี่สูงสุดร้อยละ 99.09

คำสำคัญ : เด็งกี่, เครื่องหมายทางพันธุกรรม, สนิปส์, เทคโนโลยีอ่านสายลำดับนิวคลีโอไทด์

Abstract

Single nucleotide polymorphisms (SNPs) are the major source of genetic variation. These genomic markers are crucial for the analysis of genetic diseases and traits. However, when trying to associate genotypes with phenotypic traits, it is more effective to use complete sequences of markers than single SNP. In the past, there is no method to reconstruct SNPs sequences of quasispecies because of technical difficulties in the sequencing process. However, the massively parallel sequencing technology, recently developed sequencing technique, can be used for studying diversity of RNA virus genome. Thus, the goal of this study is to reconstruct major SNPs sequences of Dengue virus from simulated DNA fragments obtained from Roche GS FLX sequencer. The proposed method provides a maximum of 98.60% accuracy for a major sequence and 99.09% for an estimated frequency.

Keywords: Dengue, genetic marker, SNPs, sequencing technology

1. คำนำ

ไวรัสเด็งกี่เป็นต้นเหตุของโรคไข้เลือดออก ซึ่ง เป็นปัญหาสาธารณสุขระดับโลก เนื่องจากโรคได้แพร่กระจายอย่างกว้างขวางและมีจำนวนผู้ป่วยเพิ่มขึ้นมากในช่วง 30 ปีที่ผ่านมา จนกลายเป็นโรคประจำถิ่นในประเทศต่างๆ มากกว่า 100 ประเทศโดยเฉพาะประเทศในเขตร้อนและเขตอบอุ่น องค์การอนามัยโลก (WHO) ประมาณการว่าประชากรโลกมากกว่าร้อยละ 40 (2,500 ล้านคน) เสี่ยงต่อการติดเชื้อไวรัสเด็งกี่ [1]

การติดเชื้อไวรัสเด็งกี่นี้จำแนกตามกลุ่มอาการได้ 3 กลุ่ม คือ ไข้เด็งกี่ (Dengue fever: DF), ไข้เลือดออก (Dengue haemorrhagic fever: DHF) และไข้เลือดออกช็อก (Dengue shock syndrome: DSS) อาการของไข้เลือดออกและไข้เลือดออกช็อกที่สำคัญคือ ระดับเกล็ดเลือดต่ำ มีการรั่วของพลาสมา อาจเกิดอาการช็อก เป็นเหตุให้เสียชีวิตได้ แตกต่างจากไข้เด็งกี่ (DF) ซึ่งมีอาการไม่รุนแรง โดยทั่วไปไม่ทำให้เสียชีวิต [2] การวินิจฉัยโรคในปัจจุบันไม่สามารถจำแนกได้ว่าเป็นไข้เด็งกี่หรือไข้เลือดออกได้ในระยะแรกของการติดโรค การทดสอบมักจะเชื่อถือได้ในผู้ป่วยที่มีไข้แล้วหลายวันซึ่งส่วนใหญ่มีอาการของโรคชัดเจนแล้ว [3]

ดังนั้น ความแตกต่างทางพันธุกรรมของเชื้อไวรัสเด็งกี่ จึงเป็นทางเลือกที่อาจช่วยให้สามารถวินิจฉัยโรคได้รวดเร็วและแม่นยำยิ่งขึ้น

ความแตกต่างทางพันธุกรรมพื้นฐานที่พบบ่อยที่สุดคือ สนิปส์ หรือ SNPs (Single Nucleotide Polymorphisms) เป็นความแตกต่างทางพันธุกรรมที่เกิดจากเบสบนสายนิวคลีโอไทด์ ณ ตำแหน่งหนึ่งๆ แตกต่างกัน [4] ใช้เป็นเครื่องหมายทางพันธุกรรม (genetic marker) เพื่อการวินิจฉัยโรค ทำนายความเสี่ยงต่อการเกิดโรค ใช้ในการพัฒนายาและวิธีการรักษาโรคที่เหมาะสมสำหรับแต่ละบุคคล[5] โดยจะมีประสิทธิภาพมากยิ่งขึ้นเมื่อศึกษาสายลำดับของสนิปส์ทุกตำแหน่งบนดีเอ็นเอแต่ละเส้นแทนที่จะดูจากสนิปที่ตำแหน่งใดๆ เพียงตำแหน่งเดียว [6] มีรายงานว่าเชื้อไวรัสที่ก่อให้เกิดโรคมีลักษณะการทำงานเป็นแบบกลุ่มประชากร สายลำดับของสนิปส์ (haplotype) และความถี่ของสายลำดับแต่ละเส้น (haplotype frequency) ในกลุ่มประชากรทั้งหมดมีผลต่อการเกิดโรค ไม่ได้อิงสายลำดับสนิปส์สายใดสายหนึ่งเพียงอย่างเดียว

ในอดีตการศึกษาความหลากหลายทางพันธุกรรมของไวรัสเด็งกี่ทำได้ยากและใช้เวลานาน เนื่องจากไวรัสเด็งกี่ มีการกลายพันธุ์สูง มีความหลากหลายสูง นอกจากจะมีการแยกออกเป็น 4 ซีโรไทป์ (serotype) ได้แก่ ซีโรไทป์ 1 – 4 แต่ละซีโรไทป์ยังสามารถแยกย่อยออกไปได้อีก แม้แต่ไวรัสเด็งกี่ซีโรไทป์เดียวกันที่ตรวจพบในคนไข้คนเดียวก็มีลำดับนิวคลีโอไทด์คล้ายคลึงกันแต่ไม่เหมือนกันทั้งหมด กล่าวได้ว่า ไวรัสเด็งกี่มีลักษณะกึ่งสปีชีส์ (quasispecies) [7]

ปัจจุบัน การศึกษาความหลากหลายของไวรัสกึ่งสปีชีส์นี้สามารถทำได้ด้วยเทคโนโลยีอ่านลำดับนิวคลีโอไทด์แบบขนานจำนวนมาก เช่น เครื่องอ่านสายลำดับเบส Roche พัฒนาโดย 454 Life Sciences, เทคโนโลยี Solexa พัฒนาโดย Illumina และระบบ ABI SoLiD ของ Applied Biosystems ซึ่งสามารถอ่านสายลำดับเบสได้จำนวนมากภายในระยะเวลาอันสั้น [8,9] และสามารถอ่านชุดของสายลำดับที่แตกต่างกันที่รวมกันอยู่ในการอ่านเพียงครั้งเดียวได้ [10] แม้เทคโนโลยีอ่านสายลำดับแบบขนานจำนวนมากนี้จะสามารถอ่านสายลำดับได้จำนวนมาก แต่มีข้อด้อยที่สำคัญคือ สายลำดับที่อ่านได้ค่อนข้างสั้น โดยเครื่องอ่านสายลำดับเบส Roche GS FLX อ่านสายลำดับได้มากถึง 400,000 เส้นต่อการอ่านหนึ่งครั้ง แต่มีความยาวเฉลี่ยเพียง 200 -300 bp ซึ่งสั้นกว่าวิธีของ Sanger ที่อ่านได้ยาว 500-1000 bp [11]

ในงานวิจัยนี้ มุ่งศึกษาการประกอบสายลำดับสนิปส์ของไวรัสเด็งกี่สายหลักและความถี่ของสายลำดับ (major haplotype and its frequency) จากข้อมูลที่ได้จากการจำลองการอ่านสายลำดับของเครื่อง Roche GS FLX Sequencer ซึ่งเป็นรุ่นล่าสุดที่พัฒนาโดย 454 Life Sciences

2. งานวิจัยที่เกี่ยวข้อง

เนื่องจากเทคโนโลยีอ่านสายลำดับแบบขนานจำนวนมาก สามารถอ่านสายลำดับได้จำนวนมากในระยะเวลาอันสั้น จึงมีงานวิจัยจำนวนมากที่ศึกษาและนำเทคโนโลยีนี้ไปใช้ เช่น Margulies และคณะ นำเสนอรายละเอียดเทคโนโลยี ของ 454 Life Sciences โปรโตคอลที่ใช้ รูปแบบผลลัพธ์ที่ได้ รวมถึงการนำไปใช้งาน โดยเสนอการประกอบจีโนมสำหรับ Mycoplasma genitalium ซึ่งครอบคลุม (coverage) ถึงร้อยละ 96 และมีความแม่น (accuracy) สูงถึงร้อยละ 99.96 [8] ต่อมาในงานวิจัย [11] ได้ใช้เทคโนโลยี picotiter plate pyrosequencing ที่ปรับปรุงล่าสุด ทำให้อ่าน สายลำดับได้ยาว 250 bp โดยเฉลี่ย จำนวน 400,000 เส้นต่อครั้ง ในงานวิจัยได้กล่าวถึงประสิทธิภาพของเทคโนโลยีใหม่ล่าสุด ความแตกต่าง ข้อดี ข้อเสีย เมื่อเทียบกับวิธี shotgun sequencing และเสนอ SHARP (Short Read Assembly Protocol) ซึ่งเป็นโปรโตคอลของสายลำดับและระเบียบวิธีในการประกอบจีโนมเพื่อให้สามารถใช้ประโยชน์จากเทคโนโลยีนี้ได้ อย่างมีประสิทธิภาพที่สุด และได้ทดสอบ PHARP ด้วยข้อมูลจีโนมของ D. melanogaster และโครโมโซมมนุษย์คู่ที่ 1, 11 และ 21 ซึ่งให้ผลลัพธ์ที่ถูกต้อง

นอกจากนี้ Iman และคณะยังได้นำเสนอความเป็นไปได้ในการอ่านสายลำดับซ้ำ (resequencing) จากชิ้นส่วนดีเอ็นเอที่นำมารวมกันแล้วส่งไปอ่านด้วยเทคโนโลยีการอ่านลำดับนิวคลีโอไทด์แบบขนานจำนวนมาก พร้อมทั้งเสนอวิธีที่เหมาะสมที่สุดในการรวมชิ้นส่วนดีเอ็นเอที่ต้องการหาสายลำดับนิวคลีโอไทด์ [10]

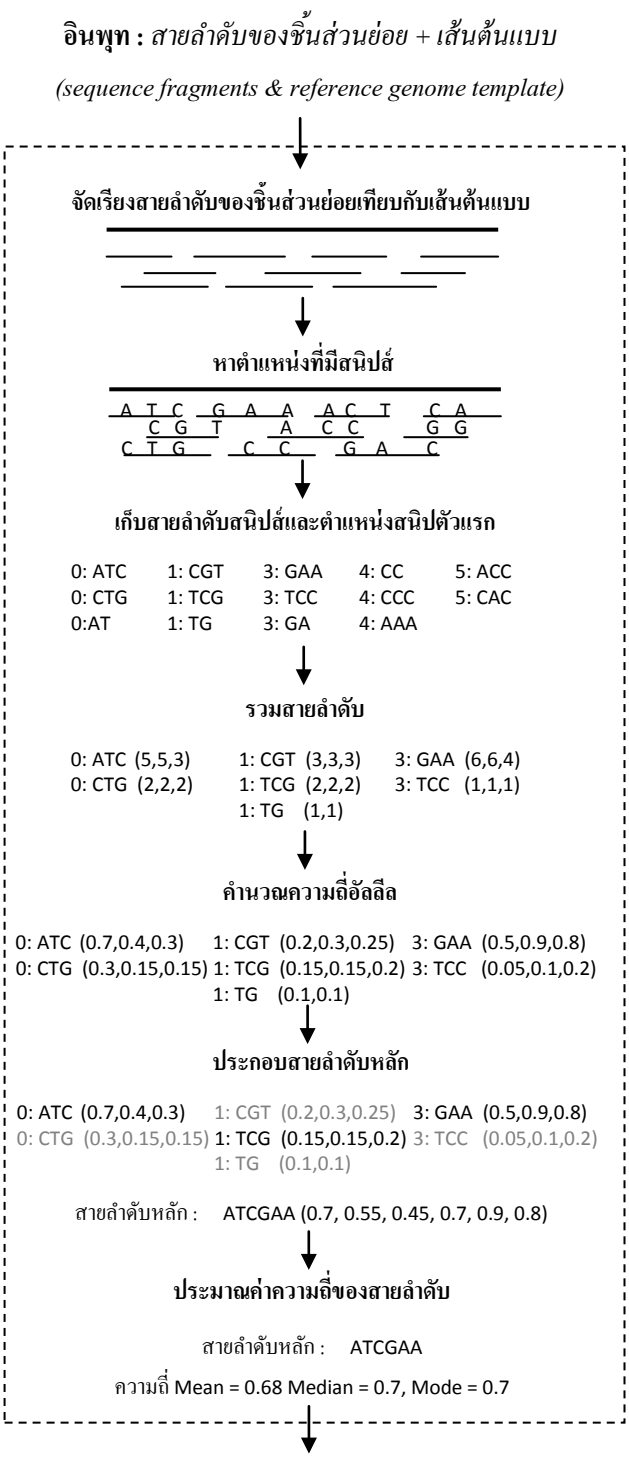
3. ขั้นตอนวิธี

ข้อมูลอินพุทที่ได้สำหรับขั้นตอนวิธีที่นำเสนอคือ สายลำดับนิวคลีโอไทด์ของชิ้นส่วนดีเอ็นเอ (sequence fragment) ที่มีความยาวเฉลี่ย 250 bp และจีโนมต้นแบบ (reference genome template) นำมาเข้าสู่กระบวนการประกอบสายลำดับหลักของสนิปส์ (major haplotype) ดังนี้ (รูปที่ 1)

1) จัดเรียง (align) ชิ้นส่วนดีเอ็นเอเทียบกับจีโนมต้นแบบ

2) เก็บสายลำดับสนิปส์ของแต่ละชิ้นส่วนพร้อมทั้งตำแหน่งของสนิปตัวแรกในชิ้นส่วนนั้นและจำนวนที่พบสายลำดับสนิปส์ที่เหมือนกันคือ มีสายลำดับเบสเหมือนกันและอยู่ในตำแหน่งเดียวกัน

3) เรียงสายลำดับตามตำแหน่งของสนิปส์ตัวแรก รวมสายลำดับสั้นๆ ที่สามารถรวมเข้ากับสายลำดับเส้นยาวได้เข้าด้วยกัน นับ



รูปที่ 1 แผนผังขั้นตอนการประกอบสายลำดับหลักและการประมาณค่าความถี่ของสายลำดับหลัก

จำนวนอัลลีลของแต่ละตำแหน่ง เช่น ตำแหน่งที่ 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 AAATTT ซึ่งมีจำนวนของ A ทั้งสามตัว (ตำแหน่ง 0-2) และ T ตัวแรกซึ่งอยู่ที่ตำแหน่ง 3 ของสายลำดับนี้ เป็น 6 จากการรวมกันของสายลำดับที่ 1 และ 3 ส่วนอีกชุดคือ AACTT ซึ่งไม่สามารถรวมกับเส้นอื่นได้ มีจำนวนแต่ละอัลลีลเป็น 3

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

5) เริ่มประกอบสายโดยเริ่มจากตำแหน่งเริ่มต้นที่น้อยที่สุด เลือกสายลำดับที่มีความถี่อัลลีลเฉลี่ยสูงสุดมาเป็นสายลำดับลัพธ์ จากนั้นพิจารณาตำแหน่งเริ่มต้นลำดับถัดไป เลือกสายลำดับที่มีความถี่เฉลี่ยสูงสุดมาพิจารณาว่าสามารถต่อกับสายลำดับลัพธ์ที่มีอยู่ได้หรือไม่ โดยพิจารณาส่วนที่ซ้อนทับกัน (overlap) ถ้าได้ให้ต่อกับสายลำดับที่มีอยู่พร้อมทั้งเก็บความถี่อัลลีลแต่ละตำแหน่งของสายลำดับลัพธ์ แต่ถ้าไม่ได้ให้เลือกสายลำดับที่มีความถี่รองลงมามาพิจารณา ทำเช่นนี้ไปเรื่อยๆ จนสุดสาย

6) ประมาณค่าความถี่ของสายลำดับ (haplotype frequency) ด้วยค่าเฉลี่ยเลขคณิต ค่ามัธยฐานและฐานนิยมของความถี่อัลลีลในแต่ละสายลำดับลัพธ์ที่ได้

4. การทดลอง

ทดสอบความถูกต้องของขั้นตอนวิธีประกอบสายลำดับและการประมาณความถี่โดยจำลองข้อมูลให้มีลักษณะคล้ายกับสายลำดับของไวรัสเด็งกี่ที่อ่านได้จากเครื่อง Roche GS FLX Sequencer โดยใช้ข้อมูลจีโนมจากฐานข้อมูลของ GenBank ดังนี้ (รูปที่ 2)

4.1 จำลองชุดจีโนมของไวรัสเด็งกี่

1) สุ่มเลือกจีโนมไวรัสเด็งกี่จากฐานข้อมูล GenBank มา 1 เส้น

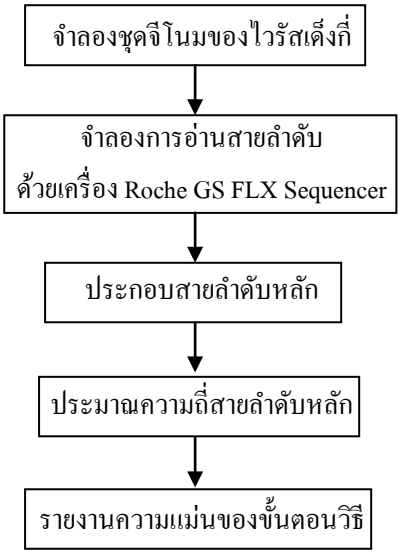
2) กำหนดความถี่ของสายลำดับหลักอย่างสุ่มตามช่วงที่กำหนด โดยการทดลองนี้แบ่งความถี่ของสายลำดับหลักเป็น 9 ช่วง คือ ร้อยละ 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89 และ 90-99

3) จำลองสายลำดับอื่นโดยสุ่มเลือกตำแหน่งที่เกิดการกลายพันธุ์ร้อยละ 2 ของความยาวจีโนมทั้งหมด (ประมาณ 200 – 220 bp จาก 10-11 Kbp) ตามร้อยละการแปรผันของไวรัสเด็งกี่ที่มีการรายงานไว้

4) แต่ละตำแหน่งเลือกอัลลีลแบบสุ่ม โดยการทดลองนี้กำหนดให้แต่ละตำแหน่งมีเพียง 2 อัลลีล ได้แก่อัลลีลเดียวกับสายลำดับหลักและอัลลีลที่ได้จากการสุ่มเลือก ซึ่งเป็นรูปแบบการแปรผันที่พบบ่อยที่สุด

5) กำหนดความถี่ของสายลำดับที่จำลองขึ้น

6) ทำซ้ำข้อ 3) – 5) จนความถี่ของสายลำดับรวมครบ 100%



รูปที่ 2 ขั้นตอนการทดลอง

4.2 จำลองการอ่านข้อมูลด้วยเครื่อง Roche GS FLX Sequencer

1) สุ่มเลือกสายลำดับโดยถ่วงน้ำหนักตามความถี่ของแต่ละสายลำดับ

2) สำหรับสายลำดับที่เลือก สุ่มตำแหน่งเริ่มต้นและอ่านจนครบความยาวระหว่าง 200-300 bp

3) ทำซ้ำจนกว่าจะได้จำนวนชิ้นส่วนของสายลำดับครบจำนวน ในที่นี้อ่านทั้งหมด 10,000 เส้น

4.3 ประกอบสายลำดับหลัก

1) นำชิ้นส่วนสายลำดับที่ได้มาประกอบด้วยขั้นตอนวิธีที่นำเสนอ

2) รายงานสายลำดับหลักที่ประกอบได้และความถี่

4.4 รายงานผลการทดลอง

1) ทำซ้ำตั้งแต่หัวข้อ 4.1 – 4.3 เพื่อเก็บข้อมูล ในการทดลองนี้ทำซ้ำทั้งสิ้น 10 ชุด แต่ละชุดมีข้อมูลความถี่ของสายลำดับหลักทั้ง 9 ช่วง โดยแต่ละช่วงมีชุดของสายลำดับดีเอ็นเอที่แตกต่างกัน นั่นคือมีการสุ่มจำลองชุดของสายลำดับดีเอ็นเอทั้งสิ้น 90 ชุด

2) ในแต่ละชุด นำสายลำดับหลักที่ได้ไปเปรียบเทียบกับสายลำดับหลักที่จำลองขึ้นซึ่งเป็นเสมือนสายลำดับจริงที่คาดหวังจากขั้นตอนวิธี รายงานความแม่น (accuracy) โดยวัดเป็นร้อยละของตำแหน่งที่ถูกต้อง

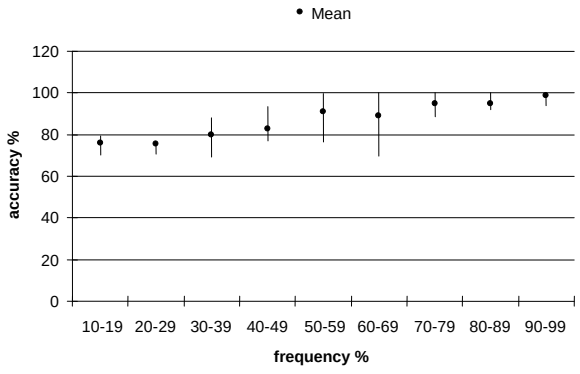
3) ในแต่ละชุด เปรียบเทียบความถี่โดยประมาณที่ได้กับความถี่ของสายลำดับหลักที่กำหนดในหัวข้อ 4.1 ข้อ 2 รายงานความแม่นเป็นร้อยละเทียบกับความถี่หลักที่กำหนด

4) นำข้อมูลความแม่นของสายลำดับไปสร้างกราฟแยกตามช่วงของความถี่ของสายลำดับหลัก

5) หาค่าเฉลี่ยของความแม่นของความถี่โดยประมาณที่ได้จากวิธีต่างๆ คือ ค่าเฉลี่ยเลขคณิต มัธยฐานและฐานนิยมของแต่ละช่วงความถี่นำไปสร้างกราฟ

5. ผลการทดลอง

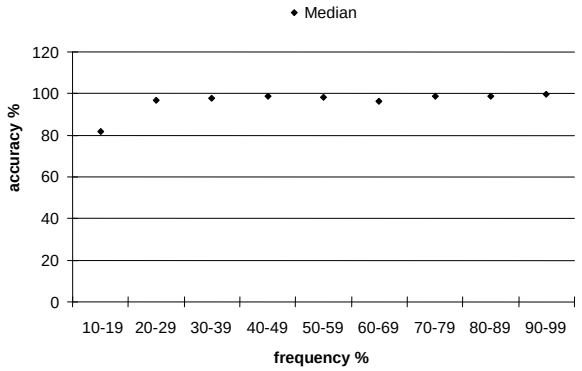
จากการทดลองประกอบสายลำดับหลักจากชุดของสายลำดับที่มีความแตกต่างกันประมาณร้อยละ 2 และมี 2 อัลลีล สำหรับแต่ละตำแหน่ง ได้ความแม่นของสายลำดับที่ประกอบได้ดังรูปที่ 3



รูปที่ 3 กราฟความแม่นของสายลำดับหลักที่ประกอบได้ โดยวัดความแม่นจากร้อยละของตำแหน่งที่ตรงกับเส้นต้นแบบ

จากรูป ความแม่นเพิ่มขึ้นเมื่อความถี่สายลำดับหลักมากขึ้น โดยความแม่นเฉลี่ยของชุดข้อมูลที่สายลำดับหลักมีความถี่ 90-99 % คือร้อยละ 98.60 และความแม่นเฉลี่ยทั้งหมดคือร้อยละ 86.91

ผลที่ได้จากการประมาณค่าความถี่ของสายลำดับหลักจากความถี่อัลลีลแต่ละตำแหน่งด้วยค่ามัธยฐาน ได้ความแม่นดังรูปที่ 4



รูปที่ 4 กราฟความแม่นของความถี่สายลำดับหลักที่ประมาณได้ จากค่ามัธยฐานของความถี่อัลลีลแต่ละตำแหน่ง

จากรูป ความแม่นของความถี่สายลำดับเพิ่มขึ้นเมื่อความถี่สายลำดับหลักเพิ่มขึ้นเช่นเดียวกับความแม่นของสายลำดับ และจากชุดข้อมูลที่ทดลองนี้ค่ามัธยฐานให้ค่าประมาณที่ใกล้เคียงที่สุดโดยมีความแม่นร้อยละ 99.09 สำหรับชุดข้อมูลที่สายลำดับหลักมีความถี่ 90-99 % และมีความแม่นเฉลี่ยร้อยละ 96.40

6. สรุป

ขั้นตอนวิธีที่นำเสนอสามารถประกอบสายลำดับหลักได้ที่ความแม่นสูงสุดร้อยละ 98.60 โดยมีความแม่นของความถี่สูงสุดร้อยละ 99.09 ซึ่งความแม่นอาจสูงขึ้นเมื่ออ่านชิ้นส่วนของสายลำดับมากขึ้น (ในงานวิจัยนี้กำหนดจำนวนการอ่านที่ 10,000 ชิ้น ซึ่งต่ำกว่าประสิทธิภาพของเครื่อง)

นอกจากนี้ยังสามารถออกแบบขั้นตอนวิธีให้มีประสิทธิภาพมากยิ่งขึ้นโดยอาศัยวิธีการเชิงสถิติที่ซับซ้อนยิ่งขึ้นในการประมาณค่าความถี่ของสายลำดับหลัก หรือ ปรับปรุงขั้นตอนการเลือกชิ้นส่วนตั้งต้นในขั้นตอนประกอบสายลำดับหลัก จากที่เลือกตำแหน่งแรกหรือส่วนหัวก่อน อาจปรับเป็นเลือกชิ้นส่วนตั้งต้นจากตำแหน่งที่มีความถี่สูงสุดหรือสุ่มเลือกตำแหน่งตั้งต้นต่างๆ กันหลายๆ ครั้งแล้วเลือกผลลัพธ์ที่พบบ่อยที่สุด เป็นต้น

ขั้นตอนวิธีการประกอบสายลำดับหลักนี้อาจพัฒนาไปใช้กับข้อมูลจากเทคโนโลยีการอ่านสายลำดับแบบขนานจำนวนมากเทคโนโลยีอื่น เช่น Solexa ซึ่งสามารถอ่านได้สั้นกว่า แต่อ่านได้จำนวนมากกว่า

เอกสารอ้างอิง

[1] Diseases of Environmental and Zoonotic Origin Team, “Dengue worldwide: an overview of the current situation and the implications for Europe”, Euro Surveill 2007, 12 (6).

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

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

[4] A. Chakravarti, “It's raining SNPs, hallelujah?”, Nature Genetics, 19, pp. 216-217.

[5] Y.-Y. Zhao, L.-Y. Wu, J.-H. Zhang, R.-S. Wang and X.-S. Zhang, “Haplotype assembly from aligned weighted SNP fragments”, Computational Biology and Chemistry, 29 (4), pp. 281-287.

[6] I. Pe'er and J. S. Beckmann, “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, Berlin, Germany, 2003, pp. 237-246.

[7] W.-K. Wang, S.-R. Lin, C.-M. Lee, C.-C. King and S.-C. Chang, “Dengue Type 3 Virus in Plasma Is a Population of Closely Related Genomes: Quasispecies”, Journal of Virology, 76 (9), pp. 4662-4665.

[8] M. Margulies, M. Egholm, W. E. Altman, A. S, B. JS and B. LA, “Genome sequencing in microfabricated high-density picolitre reactors”, Nature, 437 (7057), pp. 376-380.

[9] K.R. Chi, “The year of sequencing”, Nature Methods, 5 (1), pp. 11-14, 2007.

[10] I. Hajirasouliha, F. Hormozdiari, S. C. Sahinalp, and I. Birol, “Optimal pooling for genome re-sequencing with ultra-high-throughput short-read technologies”, ISMB 2008, 24, pp. i30-i40, 2008.

[11] A. Sundquist, M. Ronaghi, H. Tang, P. Pevzner and S. Batzoglou, “Whole-Genome Sequencing and Assembly with High-Throughput, Short-Read Technologies”, PLoS ONE, 2 (5), pp. e484.