

รหัสโครงการ: TRG5280016
ชื่อโครงการ: การตรวจวัดความหลากหลายทางพันธุกรรมในข้อมูลจีโนมของกลุ่มคน
ชื่อนักวิจัย: ดร. เบญจรัตน์ ภูภักดี สถาบันวิจัยจุฬาภรณ์
Email Address: benjarath@cri.or.th
ระยะเวลาโครงการ: ๑๖ มีนาคม ๒๕๕๒ – ๓๐ กันยายน ๒๕๕๕

บทคัดย่อ

ปัจจุบันนี้เทคโนโลยีการถอดรหัสพันธุกรรมทั้งจีโนมได้มีการพัฒนาก้าวหน้าไปมาก ซึ่งช่วยให้การถอดรหัสข้อมูลพันธุกรรมส่วนบุคคลสามารถเป็นไปได้ในราคาที่ถูกลงและใช้ระยะเวลาสั้นลงเป็นอย่างมาก คาดว่าจะมีการประยุกต์นำเทคโนโลยีดังกล่าวไปใช้ทั่วไปเร็ว ๆ นี้ รวมถึงการใช้ความจำเพาะของรหัสพันธุกรรมส่วนบุคคล ร่วมกับการรักษาของแพทย์ เพื่อมุ่งเน้นวิธีการรักษาโรคที่เหมาะสมในแต่ละบุคคลด้วย การพัฒนาอย่างรวดเร็วของเทคโนโลยีการถอดรหัสพันธุกรรมทั้งจีโนม เป็นสาเหตุหนึ่งที่ทำให้ความหลากหลายทางพันธุกรรมแบบ novel Sequence ได้ถูกนิยามอย่างเป็นทางการและได้รับความสนใจในวงการวิจัยมากขึ้นเมื่อไม่นานมานี้ ความหลากหลายทางพันธุกรรมชนิดนี้ สามารถพบได้ในข้อมูลพันธุกรรมมนุษย์แบบ *de novo* ซึ่งงานวิจัยที่ผ่านมาได้พบว่า ในข้อมูลพันธุกรรมของมนุษย์คนหนึ่ง อาจมี novel sequence ได้มากถึง 3 – 5 ล้านเบส ดังนั้นการพัฒนาวิธีวิจัยเพื่อค้นหา novel sequence ในข้อมูลพันธุกรรมส่วนบุคคลอย่างมีประสิทธิภาพและประสิทธิผล จึงมีความสำคัญเป็นอย่างยิ่ง

รหัสพันธุกรรมทั้งจีโนมแบบ *de novo* มีรูปแบบต่างจากรหัสพันธุกรรมทั้งจีโนมแบบ mapping โดยรหัสแบบ *de novo* ประกอบไปด้วย scaffolds และ contigs ซึ่งมีขนาดเล็กใหญ่ต่างกันมาก อีกทั้งไม่มีข้อมูลเบื้องต้นว่า scaffolds และ contigs เหล่านี้มาจากส่วนใดของโครโมโซม การหา novel sequence ต้องใช้ข้อมูลดังกล่าว ซึ่งสามารถหาได้โดยวิธีเปรียบเทียบความเหมือนของรหัสพันธุกรรมแบบ *de novo* กับรหัสพันธุกรรมอ้างอิง (human reference sequence) การเปรียบเทียบดังกล่าวต้องใช้โปรแกรมคอมพิวเตอร์ (sequence aligner) ปัจจุบันนี้ยังไม่มี sequence aligner ที่พัฒนามาโดยเฉพาะเจาะจงเพื่อการเปรียบเทียบดังกล่าว โปรแกรมที่ก้าวหน้าที่สุดเท่าที่จะหาได้ในปัจจุบันยังต้องใช้ computing resources สูง และใช้เวลานานในการคำนวณ

ในงานวิจัยนี้ ได้มีการพัฒนาโปรแกรม **NSIT (Novel Sequence Identification Tool)** ขึ้นเพื่อการเปรียบเทียบความเหมือนของรหัสพันธุกรรมทั้งจีโนมของมนุษย์ แบบ *de novo* กับรหัสพันธุกรรมทั้งจีโนมของมนุษย์ แบบอ้างอิง โดยเป้าหมายหลักคือเพื่อการค้นหา novel sequence ในรหัสพันธุกรรมแบบ *de novo* โดยเฉพาะ โปรแกรม NSIT สามารถค้นหา novel sequence ได้อย่างแม่นยำและรวดเร็วภายในเวลาไม่กี่ชั่วโมง โดยใช้คอมพิวเตอร์ทั่วไปได้ ผลการทดลองที่ได้มี sensitivity เกือบ 100% และ precision ประมาณ 90%

คำหลัก: รหัสข้อมูลพันธุกรรมส่วนบุคคลของมนุษย์ แบบ *de novo*, ความหลากหลายทางพันธุกรรม novel sequence, โปรแกรมด้าน bioinformatics

Project Code: TRG5280016
Project Title: Detection of Genetic Variation in Multiple Personalized Genomes
Investigator: Dr. Benjarath Pupacdi Chulabhorn Research Institute
Email Address: benjarath@cri.or.th
Project Period: 16 March 2009 – 30 September 2012

Abstract

Motivation: Recent developments in DNA sequencing technology have been paving the road to faster and cheaper personal genome sequencing. Many useful applications have been anticipated, such as, routinely performing human genome comparisons to detect sequence variation when designing personalized medical treatments. Novel sequences have recently been discovered as a new class of sequence variation detectable in a *de novo* assembly, where as much as 3–5 Mb of novel sequences may be present in a *de novo* personal genome. Detecting them effectively and efficiently is therefore an important task.

A typical *de novo* human genome assembly, unlike a mapping assembly, consists of a very large number of scaffolds and contigs whose lengths vary highly, and chromosomal positions and orientations are unknown. Identification of novel sequences in the assembly requires comparing it to the reference sequence. However, the input characteristics make the problem computationally expensive. Existing state-of-the-art large genome aligners could require several days to run, when applied to this problem.

Results: We introduce **NSIT (Novel Sequence Identification Tool)**, a software for aligning a *de novo* genome assembly against the reference genome assembly to identify novel sequences. NSIT is fast, accurate, and requires only modest computing resources. To the best of our knowledge, NSIT is the first algorithm designed specifically for this task. Our algorithm requires only a few hours on a commodity desktop and yields high quality results (near 100% sensitivity and about 90% precision).

Keywords: human genome *de novo* assembly, novel sequences, bioinformatics software tool