Thai Journal of Genetics 2009, 2(2) : 131–144

การวิเคราะห์ลำดับนิวคลีโอไทด์ EST ของอ้อยด้วยคอมพิวเตอร์เพื่อการจัดคลัสเตอร์คุณภาพสูงและการค้นหา SSR

Computational analysis of sugarcane ESTs for high-quality clusters and SSR mining

ปิยรัตน์ พลยะเรศ[1] เทวัญ เริ่มสูงเนิน[2] งามนิจ อาจอินทร์[3] วิชัย ณีรัตนพันธุ์[1] ชุติพงศ์ อรรคแสง[4] และ นภาภรณ์ ตันติสุวิชวงษ์[1]*

Piyarat Ponyared[1], Tawun Remsungnen[2], Ngamnij Arch-int[3], Wichai Neeratanaphan[1], Chutipong Akkasaeng[4], and Napaporn Tantisuwichwong[1]*

[1] ภาควิชาชีววิทยา [2]ภาควิชาคณิตศาสตร์ [3]ภาควิชาคอมพิวเตอร์ คณะวิทยาศาสตร์

[4] ภาควิชาพืชศาสตร์และทรัพยากรการเกษตร คณะเกษตรศาสตร์ มหาวิทยาลัยขอนแก่น ขอนแก่น 40000

[1] Department of Biology, [2]Department of Mathematics, [3]Department of Computer Science, Faculty of Science.

[4] Department of Plant Science and Agricultural Resources, Faculty of Agriculture, Khon Kaen University, Khon Kaen 40000, Thailand

**Corresponding author: naptan@kku.ac.th*

บทคัดย่อ

การพัฒนาเครื่องหมายพันธุกรรมชนิด SSR สามารถทำได้จากคลัสเตอร์ของ expressed sequence tags (ESTs) ที่มีคุณภาพสูง โดยวิธีทั่วไป การจัดคลัสเตอร์ของ EST อาศัยหลักการของความคล้ายคลึงกันของนิวคลีโอไทด์ ช่วยลดความซ้ำซ้อนและเพิ่มคุณภาพของลำดับนิวคลีโอไทด์ ระดับของความคล้ายคลึงเป็นหนึ่งในพารามิเตอร์สำคัญ ที่มีผลต่อคุณภาพของคลัสเตอร์ EST การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อตรวจสอบคุณภาพของคลัสเตอร์ EST ด้วยการเปลี่ยนแปลงค่าความคล้ายคลึง และค้นหาตำแหน่งของ SSR ภายในคลัสเตอร์ EST ที่จัดตามค่าความคล้ายคลึงที่แปรเปลี่ยนไป ข้อมูล EST จำนวน 2,268 โมเลกุล จากลำต้นของอ้อย (*Saccharum* spp.) พันธุ์ลูกผสม CP72-2086 ที่เจริญเติบโตเต็มที่ ที่เก็บอยู่ในฐานข้อมูล dbEST ของ GenBank เข้าสู่ขั้นตอนการเตรียมข้อมูล EST เพื่อกำจัดนิวคลีโอไทด์ที่เกิดจากความผิดพลาดในการอ่านลำดับและนิวคลีโอไทด์ที่ปนเปื้อน EST ที่ผ่านขั้นเตรียมการ จำนวน 2,167 โมเลกุล การจัดคลัสเตอร์ ESTs ด้วยการตั้งค่าความเหมือนของนิวคลีโอไทด์ (P) ที่ 85, 90, 95 และ 100% เพื่อลดความซ้ำซ้อนของชุดข้อมูล ESTs พบว่า การจัดคลัสเตอร์ที่ค่าความเหมือน 85 เปอร์เซ็นต์ ให้จำนวนคลัสเตอร์ของ EST น้อยที่สุด และเมื่อค้นหาตำแหน่งของ SSR ในคลัสเตอร์ EST พบว่าคลัสเตอร์จากค่าความเหมือน 85 เปอร์เซ็นต์ มีจำนวนของ SSR น้อยที่สุด

ABSTRACT

Expressed sequence tags (ESTs) have provided opportunity for development of powerful marker SSR when high-quality EST clusters are available. EST clustering is commonly performed on the basis of nucleotide similarity to reduce redundancy and increase the sequence quality. The degree of similarity is one of the important parameters affecting the EST cluster quality. This work aimed to determine EST cluster quality with various degrees of nucleotide similarity and identifying SSR locus within the defined EST clusters. A collection of 2,268 ESTs from mature stalk of sugarcane (*Saccharum* spp.) hybrid cultivar CP72-2036, available in dbEST of GenBank, was passed into pre-processing steps to eliminate the sequencing errors and contaminant sequences. This resulted in 2,167 clean ESTs. EST clustering with sequence identity P = 85, 90, 95 and 100% reduced the EST data set. The lowest number of clusters was obtained at P = 85%. Exploring of SSR locus also yielded the lowest number of SSR in EST clusters defined at the P value = 85%.

คำสำคัญ: expressed sequence tags (ESTs), อ้อย, การจัดคลัสเตอร์ EST, ลำดับเบสซ้ำ SSR

Keywords : expressed sequence tags (ESTs), sugarcane, EST clustering, SSR

บทนำ

Expressed sequence tags หรือ EST คือ ดีเอ็นเอสายสั้นๆ เป็นผลของการหาลำดับนิวคลีโอไทด์บางส่วนของดีเอ็นเอคู่สม (complementary DNA หรือ cDNA) ที่อยู่ในเวคเตอร์ ปฏิกิริยาการอ่านลำดับดีเอ็นเอ ใช้ไพรเมอร์ที่มีตำแหน่งบนเวคเตอร์เป็นจุดตั้งต้น และอ่านผ่านเข้าสู่ cDNA เพียงครั้งเดียว (Adams *et al.,* 1991) ดังนั้นความยาวของ EST และจำนวนนิวคลีโอไทด์ที่ไม่อาจระบุชนิดของเบส (N) ในแต่ละโมเลกุลของ EST จึงไม่อาจทราบล่วงหน้าได้ ลำดับนิวคลีโอไทด์ของ EST อาจปนเปื้อนด้วยลำดับนิวคลีโอไทด์ของเวคเตอร์หรือลำดับนิวคลีโอไทด์ของ low complexity (Aaronson *et al.,* 1996) นอกจากนี้ cDNA ที่ใช้เพื่อการผลิตข้อมูล EST เลือกจากห้องสมุด cDNA (cDNA library) ที่ประกอบไปด้วยโคลน (clone) จำนวนมาก และเป็นการเลือกแบบสุ่มโคลน ดังนั้นจึงมีโอกาสที่โคลนที่ถูกสุ่มขึ้นมาจะให้ EST ที่มีความซ้ำซ้อนกัน (redundancy) (Miller *et al.,* 1999)

การลดความซ้ำซ้อนของ EST ทำได้ด้วยการจัดคลัสเตอร์ EST อาศัยหลักการเทียบเคียงลำดับนิวคลีโอไทด์ที่มีความคล้ายคลึงระหว่างกัน (sequence similarity) ข้อดีของการจัดคลัสเตอร์ คือ การลดขนาดและเพิ่มคุณภาพของชุดข้อมูล EST (Burke *et al.,* 1999) ซอฟท์แวร์ที่ใช้จัดคลัสเตอร์ EST ตามหลักการนี้มีหลากหลาย เช่น d_2 cluster (Burke *et al.,* 1999) BLASTClust (Altschul *et al.,* 1997) TGICL (Pertea *et al.,* 2003) และ Cd-hit (Li and Godzik, 2006) เป็นต้น ที่ผ่านมาการเปรียบเทียบซอฟท์แวร์ที่ใช้จัดคลัสเตอร์ EST ของอ้อย (*Saccharum* spp.) พันธุ์ลูกผสม Nco 376 จำนวน 243 โมเลกุล (Carson and Botha, 2000) ได้ข้อมูลว่า ซอฟท์แวร์ Cd-hit สามารถลดความซ้ำซ้อนของ EST ได้อย่างมีประสิทธิภาพ (ปิยรัตน์ และ นภาภรณ์, 2552) แต่อย่างไรก็ตาม การประเมินคุณภาพของคลัสเตอร์ที่ได้จำเป็นต้องศึกษาประกอบกัน พารามิเตอร์หนึ่งที่สำคัญของการจัดคลัสเตอร์ และมีผลต่อคุณภาพของคลัสเตอร์ คือค่าเปอร์เซ็นต์ความเหมือนของนิวคลีโอไทด์ระหว่างโมเลกุลของดีเอ็นเอ (percent identity หรือ P) งานวิจัยครั้งนี้มีวัตถุประสงค์ใช้ซอฟท์แวร์ Cd-hit กับข้อมูล EST ของอ้อยจากห้องสมุด cDNA ที่มีจำนวน EST ขนาดใหญ่ขึ้น เพื่อยืนยันความแม่นยำและประสิทธิภาพของซอฟท์แวร์ ในการจัดการชุดข้อมูลขนาดใหญ่ อีกประการหนึ่งคือตรวจสอบคุณภาพของคลัสเตอร์ EST ที่ตั้งค่า P แตกต่างกันและประการ

สุดท้ายคือการประเมินจำนวนลำดับเบสซ้ำ SSR ในคลัสเตอร์ที่จัดตามค่า P ดังกล่าว

อุปกรณ์และวิธีการทดลอง

ระบบปฏิบัติการคอมพิวเตอร์

คอมพิวเตอร์ระบบปฏิบัติการลินุกซ์ UBUNTU 8.10 Kernel 2.6.24-23-generic หน่วยประมวลผลใช้ Intel ® Core™ 2 Duo Processor, 1.73 GHz หน่วยความจำแบบ DDR2 ขนาด 4 จิกะไบต์ (GB) ฮาร์ดไดร์ฟขนาด 20 จิกะไบต์

คอมพิวเตอร์ซอฟท์แวร์

ซอฟท์แวร์

ซอฟท์แวร์ที่ใช้พัฒนาจากทีมวิจัยอื่นที่อนุญาตให้ใช้ซอฟท์แวร์ได้ และเป็นซอฟท์แวร์ที่ทำงานภายในเครื่องคอมพิวเตอร์ส่วนตัวและสามารถประมวลผลได้ (stand alone) จำนวน 4 ซอฟท์แวร์ คือ SeqClean (Chen *et al.,* 2007), RepeatMasker (Smith *et al.,* 1996-2004), Cd-hit (Li and Godzik, 2006) และ MISA (Thiel *et al.,* 2003)

ชุดคำสั่งที่พัฒนาเอง

ในบางขั้นตอนได้เขียนชุดคำสั่งภาษาเพิร์ล (Perl) ขึ้นใช้ประมวลผลเอง จำนวน 4 คำสั่ง (ปิยรัตน์ และ นภาภรณ์. 2552) คือ (1) xtract.pl (2) length_N.pl (3) length_N.xsl และ (4) ชุดคำสั่งที่ดัดแปลงจาก clstr_sort_prot_by.pl (Li and Godzik, 2006)

ฐานข้อมูล

ฐานข้อมูล EST (dbEST) (http://www.ncbi.nlm.nih.gov/dbEST) สืบค้นเมื่อเดือนมกราคม พ.ศ. 2550

ฐานข้อมูลเวคเตอร์ คือ UniVec ของ NCBI (ftp://ftp.ncbi.nih.gov/pub/UniVec/) สืบค้นเมื่อเดือนกรกฎาคม พ.ศ. 2551

ฐานข้อมูลลำดับเบสคุณภาพต่ำชนิด low complexity คือ RapBase (http://www.girinst.org) สืบค้นเมื่อเดือนธันวาคม พ.ศ. 2550

EST ของอ้อย

EST จำนวน 2,268 โมเลกุล จากห้องสมุด cDNA ชื่อ pSS สร้างจากเนื้อเยื่อลำต้นอ้อย พันธุ์ CP 72-2086 ในระยะหลังออกดอก (Schulze *et al.,* 2002) หมายเลขของ EST คือ BQ535320-BQ537587 ความยาวของนิวคลีโอไทด์รวมเท่ากับ 1,057,371 คู่เบส

การจัดข้อมูล EST รูปแบบไฟล์เดี่ยว

EST ของอ้อยจากฐานข้อมูล GenBank อยู่ในไฟล์เดี่ยวในรูปแบบรวมลำดับนิวคลีโอไทด์ EST จำนวนมากกว่า 2 โมเลกุลไว้ด้วยกัน (multi-FASTA) เพื่อให้สามารถวิเคราะห์ได้สะดวกจึงแยก EST เป็นโมเลกุลเดี่ยวต่อ 1 ไฟล์ (single-FASTA) ด้วยชุดคำสั่ง xtract.pl

การเตรียมข้อมูลของ EST

การตรวจสอบเปอร์เซ็นต์ของเบสที่ไม่ทราบชนิดและความยาวของ EST

ตรวจสอบเปอร์เซ็นต์ของเบสที่ไม่ทราบชนิด (N) และความยาวของ EST ด้วยชุดคำสั่ง length_N.pl และใช้ชุดคำสั่ง length_N.xsl จัดการผลการวิเคราะห์ของ length_N.pl ในรูปแบบตาราง คัดกรอง EST ที่มีความยาวของลำดับนิวคลีโอไทด์มากกว่า 100 คู่เบส และมีเปอร์เซ็นต์ของ N น้อยกว่า 5

การระบุลำดับนิวคลีโอไทด์ของเวคเตอร์ที่ปนเปื้อน

ระบุการปนเปื้อนลำดับนิวคลีโอไทด์ของเวคเตอร์ด้วยซอฟท์แวร์ SeqClean และฐานข้อมูล UniVec โดยตั้งค่า E-value ≥ 1e-700 ค่า gap penalty และ extend gap เท่ากับ 3

การระบุตำแหน่งของ low complexity

ระบุตำแหน่งของ low complexity ใน ESTs ด้วยซอฟท์แวร์ RepeatMasker ร่วมกับฐานข้อมูล RepBase โดยเลือกใช้ Wblast (http://blast.wustl.edu/licensing/) เป็น sequence search engine และตั้งค่าคะแนน (score) เท่ากับหรือมากกว่า 225 bits

การจัดคลัสเตอร์ ESTs

การจัดคลัสเตอร์ EST ใช้ซอฟท์แวร์ Cd-hit โดยตั้งค่า P 4 ค่า ดังนี้ 85, 90, 95 และ 100%

การประเมินจำนวน SSR ในคลัสเตอร์ EST

การระบุ SSR ในคลัสเตอร์ EST อาศัยซอฟท์แวร์ MISA โดยกำหนดให้ SSR ชนิดสองเบสซ้ำ (dimer) มีจำนวนซ้ำ ≥ 6 ส่วนสามเบสซ้ำ (trimer) ถึง หกเบสซ้ำ (hexamer) มีจำนวนซ้ำ ≥ 5

ผลการทดลอง

การออกแบบกระบวนการวิเคราะห์ EST

กระบวนการวิเคราะห์ EST *(EST analysis pipeline)* สรุปใน Figure 1 เริ่มจากการดึงข้อมูล EST ของอ้อยจากฐานข้อมูล dbEST ของ Genbank จากนั้น EST เข้าสู่ขั้นตอนการวิเคราะห์ 3 ขั้นตอนหลัก อย่างเป็นลำดับ คือขั้นเตรียมการ ขั้นการจัดคลัสเตอร์ และขั้นการค้นหา SSR ในขั้นเตรียมการ EST ที่อยู่ในรูปแบบรวมไฟล์ 1 ไฟล์ที่มีจำนวน EST 2,268 โมเลกุล จะถูกแยกออกเป็นไฟล์เดี่ยวจำนวน 2,268 ไฟล์ แต่ละไฟล์มี EST เพียง 1 โมเลกุล แต่ละไฟล์เดี่ยวของ EST จะถูกตรวจสอบเปอร์เซ็นต์ N และความยาว ก่อนที่จะเข้าสู่การคัดกรองการปนเปื้อนของเวคเตอร์และลำดับนิวคลีโอไทด์ low complexity เมื่อผ่านขั้นตอนย่อยเหล่านี้แล้ว EST ที่ผ่านเกณฑ์ประเมินคุณภาพต่างๆเหลือจำนวน 2,167 โมเลกุล ซึ่งจะเข้าสู่ขั้นตอนการจัดคลัสเตอร์และการระบุ SSR ต่อไป

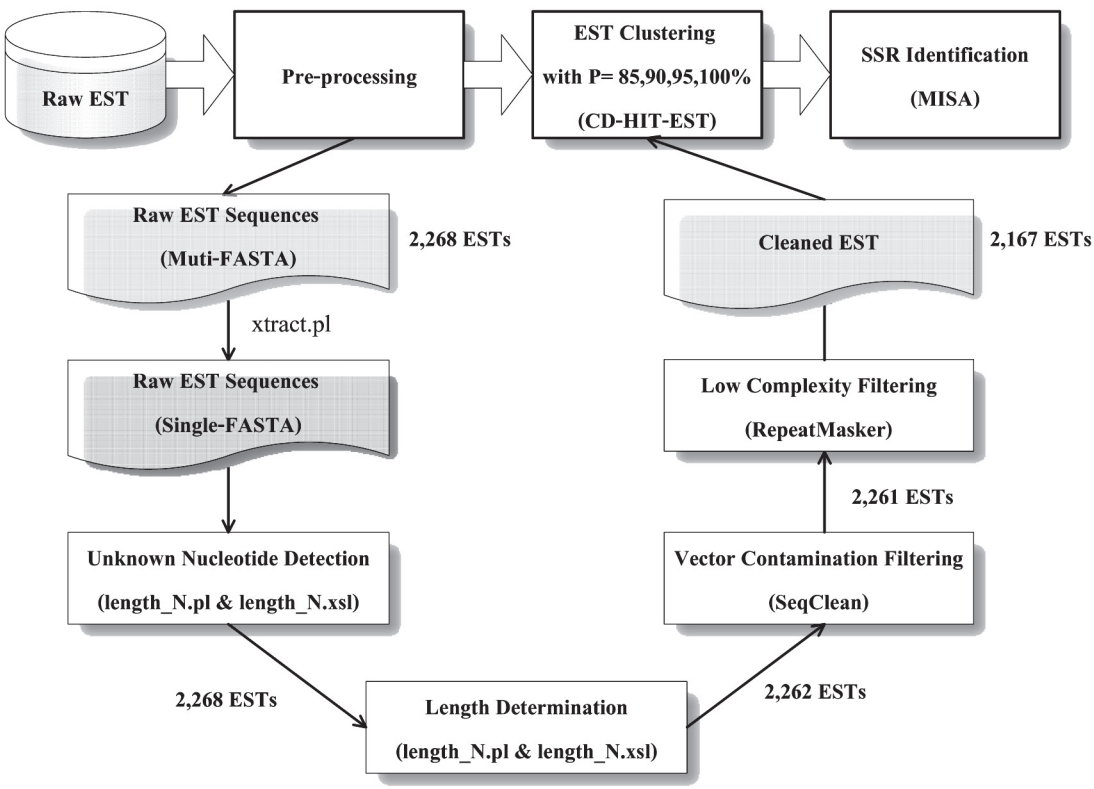

Figure 1 Flowchart depicting EST clustering and analysis pipeline.

การเตรียมข้อมูล EST

การตรวจสอบเปอร์เซ็นต์ของเบสที่ไม่ทราบชนิดและความยาวของ EST

การตรวจสอบ N ใน EST จำนวน 2,268 โมเลกุล พบ EST จำนวน 234 โมเลกุลที่มี N อยู่ในลำดับนิวคลีโอไทด์ เมื่อจำแนกตามเปอร์เซ็นต์ N พบว่ามีค่าอยู่ระหว่าง 0-2.66 เปอร์เซ็นต์

การตรวจสอบความยาวลำดับนิวคลีโอไทด์ของ EST จากห้องสมุด pSS จำนวน 2,268 โมเลกุล พบว่าลำดับนิวคลีโอไทด์ของ EST ที่มีความยาวมากที่สุดคือ 684 คู่เบส และน้อยที่สุด 51 คู่เบส เมื่อแบ่งความยาวของ EST ออกเป็น 7 ช่วง (Fig. 2) พบว่าช่วงความยาว 500-599 คู่เบส มีจำนวนของ EST มากที่สุด คือ 891 โมเลกุล รองลงมาคือช่วง 400-499 คู่เบส จำนวน 435 โมเลกุล และช่วงที่มีจำนวนของ EST น้อยที่สุดคือ 0-99 จำนวน 6 โมเลกุล ได้แก่ หมายเลข BQ535495, BQ535635, BQ535671, BQ 535672, BQ536895 และ BQ537017

ดังนั้นจาก EST ของห้องสมุด pSS จำนวน 2,268 โมเลกุล พบ 6 โมเลกุล คิดเป็น 0.26 เปอร์เซ็นต์ ที่มีความยาวน้อยกว่า 100 คู่เบส ถูกตัดออกจากกระบวนการวิเคราะห์ ดังนั้น จึงเหลือ EST 2,262 โมเลกุล ที่ถูกนำไปวิเคราะห์ในขั้นตอนต่อไป

การระบุลำดับนิวคลีโอไทด์ของเวคเตอร์ที่ปนเปื้อน

ผลการระบุลำดับนิวคลีโอไทด์ของเวคเตอร์ภายใน EST จำนวน 2,262 โมเลกุล พบ EST จำนวน 213 โมเลกุลมีลำดับนิวคลีโอไทด์ของเวคเตอร์ปนเปื้อน ในจำนวนนี้มี 212 โมเลกุล ที่พบเวคเตอร์ปนเปื้อนที่ตำแหน่งปลายสายของ EST ซึ่งในกรณีนี้จะถูกตัดเฉพาะส่วนของเวคเตอร์ออกไป แต่อีก 1 โมเลกุล คือ BQ535762 พบเวคเตอร์อยู่ภายในสาย EST และกำจัดออกจากระบบการตรวจสอบ จึงเหลือ EST จำนวน 2,261 โมเลกุลที่จะเข้าสู่ขั้นตอนต่อไป

การระบุตำแหน่งของ low complexity ใน EST

การระบุตำแหน่งของ low complexity ใน EST จำนวน 2,261 โมเลกุล ด้วยซอฟท์แวร์ RepeatMasker พบว่ามี EST จำนวน 94 โมเลกุล ที่มีตำแหน่ง low complexity ภายในลำดับของนิวคลีโอไทด์ แสดงตัวอย่างตำแหน่งของ low complexity ใน EST หมายเลข BQ535752 พบ C-rich ที่ตำแหน่ง

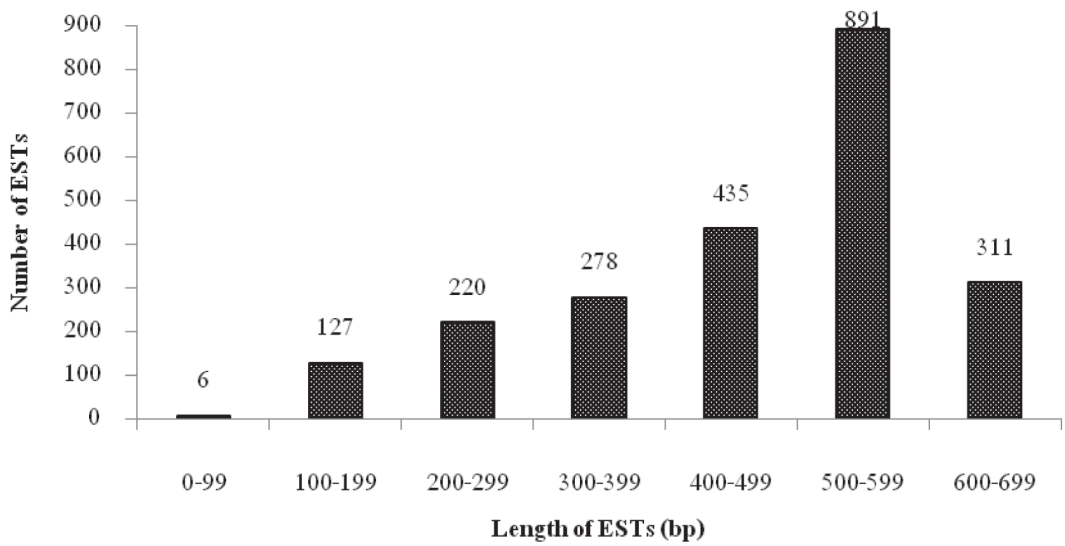

Figure 2 Distribution of the EST sequence length.

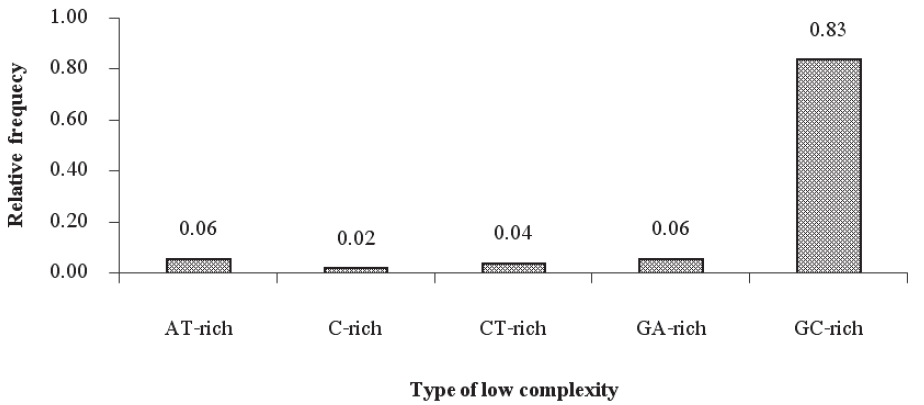

Figure 3 Detection of low complexity region in EST**.** Various types of low complexity were detected within 94 of 2,261 ESTs (used in this study). For instance, C-rich low complexity was found in the BQ 535752 sequence and AT-rich low complexity located in the BQ535650 sequence.

ของนิวคลีโอไทด์ 212-288 และ BQ535650 พบ AT-rich ที่ตำแหน่งของนิวคลีโอไทด์ 390-412 ใน Figure 3

ในจำนวน EST 94 โมเลกุล ที่พบ low complexity มีการกระจายตัวของ low complexity 110 ตำแหน่ง และเมื่อคิดเป็นความถี่สัมพัทธ์พบ GC-rich มีค่าความถี่มากที่สุดเท่ากับ 0.83 รองลงมาคือ GA-rich และ AT-rich เท่ากับ 0.06 ส่วน CT-rich เท่ากับ 0.04 และ C-rich น้อยที่สุดเท่ากับ 0.02 (Fig. 4) EST โมเลกุลที่พบ low complexity จะถูกลบทิ้งโมเลกุลออกจากกระบวนการวิเคราะห์ จึงเหลือ EST จำนวน 2,167 โมเลกุล

การตรวจสอบคุณภาพของคลัสเตอร์ EST เมื่อกำหนดค่า P แตกต่างกัน 4 ค่า

เมื่อจัดคลัสเตอร์ของ EST จำนวน 2,167 โมเลกุล ตามค่า P คือ 85, 90, 95 และ 100% พบว่าที่ค่า P เท่ากับ 100, 95, 90 และ 85% จัดได้จำนวน 2,087, 1,441, 1,347 และ 1,300 คลัสเตอร์ตามลำดับ (Fig. 5)

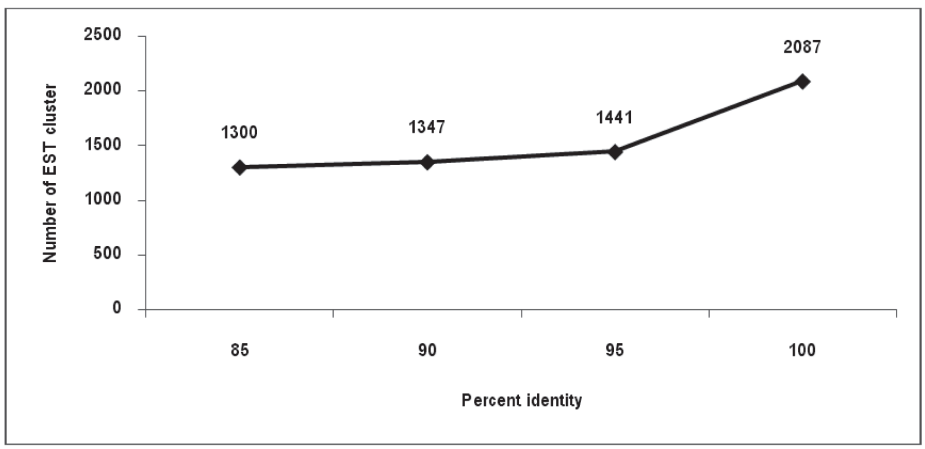

Figure 4 Frequency of low complexity in EST.

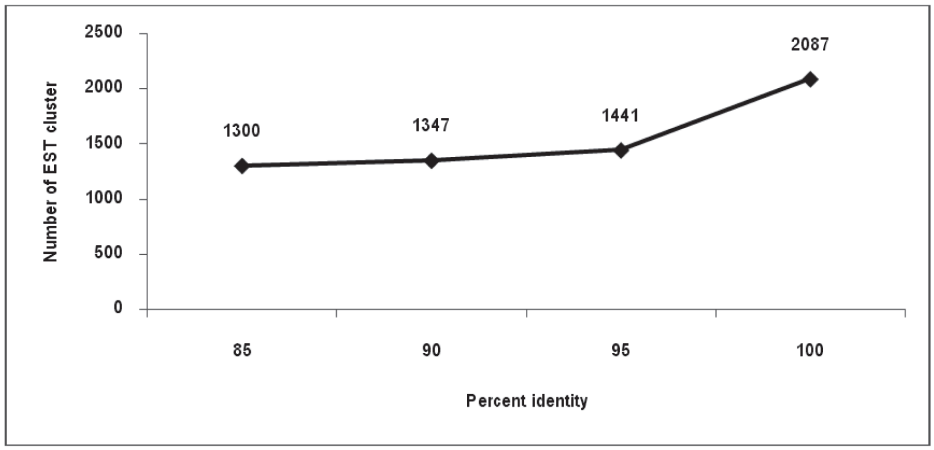

Figure 5 The decreased number of EST clusters following the reduction of P value.

จำนวนสมาชิกของคลัสเตอร์ หรือขนาดของคลัสเตอร์ ที่ค่า P เท่ากับ 85, 90, 95, 100% แสดงใน Table 1 โดยภาพรวม ทุกค่า P ที่ทดสอบ คลัสเตอร์ขนาดเล็กที่สุด คือคลัสเตอร์ที่มี EST 1 โมเลกุลเป็นสมาชิกหรือคลัสเตอร์ชนิด singleton ซึ่งเป็นคลัสเตอร์ที่มีจำนวนสูงสุด ค่า P ที่ลดต่ำลงทำให้ขนาดของคลัสเตอร์ใหญ่ขึ้น ค่า P ที่ 100% ให้คลัสเตอร์ที่มี EST 1-5 โมเลกุลเป็นสมาชิก แต่ค่า P ที่ 95% จัดคลัสเตอร์ที่มี EST 1-40 โมเลกุล เป็นสมาชิก และค่า P ที่ 85 และ 90% จัดคลัสเตอร์ที่มี EST 1-48 โมเลกุลเป็นสมาชิก

ความคล้ายคลึงของลำดับนิวคลีโอไทด์ในแต่ละคลัสเตอร์ ยืนยันได้ด้วยการเปรียบเทียบลำดับนิวคลีโอไทด์ระหว่าง EST ที่เป็นสมาชิกของ

Table 1 Number of EST clusters at 4 percent identities.

Size of cluster	Number of EST in cluster	Number of cluster			
		P=85%	**P=90%**	**P=95%**	**P=100%**
small	1	1,058	1,117	1,214	2,022
	2	119	109	120	55
	3	44	45	40	7
	4	18	20	14	1
	5	18	15	18	2
	6	10	5	5	0
	7	9	10	5	0
	8	3	3	4	0
	9	3	6	4	0
	10	0	1	2	0
	11-20	11	8	10	0
	21-30	4	5	3	0
	31-40	1	2	2	0
big	41-48	2	1	0	0
	total	**1,300**	**1,347**	**1,441**	**2,087**

Figure 6 Sequence alignment of an EST cluster at P = 100% (A) and P = 95% (B). Asterisks indicate identical nucleotide.

คลัสเตอร์ ที่ค่า P เท่ากับ 100% EST ที่เป็นสมาชิกของแต่ละคลัสเตอร์มีความยาวคู่เบสใกล้เคียงกัน และชนิดของนิวคลีโอไทด์เป็นชนิดเดียวกันตลอดความยาวของ EST ดังแสดงใน Figure 6 ที่ค่า P เท่ากับ 100% คลัสเตอร์หนึ่ง ประกอบด้วย EST เพียง 2 โมเลกุล คือ หมายเลข BQ536979 และ BQ537484 ซึ่งมีความยาว 558 และ 563 คู่เบสตามลำดับ (Fig. 6A) ลำดับนิวคลีโอไทด์เปรียบเทียบของ EST ทั้งสองแสดง consensus sequence เมื่อค่า P ลดลง ที่ 95% BQ 536979 และ BQ537484 ถูกจัดคลัสเตอร์ร่วมกับ EST อื่นเพิ่มอีก 3 โมเลกุล คือ BQ536950 (462 คู่เบส), BQ 535815 (482 คู่เบส) และ BQ536741 (527 คู่เบส) ลำดับนิวคลีโอไทด์เปรียบเทียบของ EST 5 โมเลกุลนี้ แสดงให้เห็น (1) บริเวณ consensus sequence (2) บริเวณ (gap) ขนาดเล็ก และ (3) ลำดับนิวคลีโอไทด์ของ EST ที่มีความยาวมากที่สุดในคลัสเตอร์ (Fig. 6B)

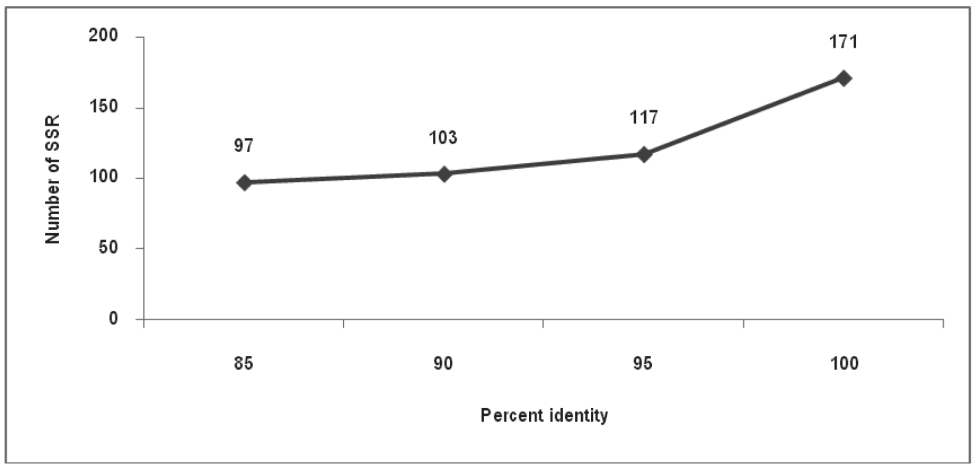

Figure 7 The decreased number of SSR following the reduction of P value used in EST clustering.

การประเมินจำนวน SSR ภายในคลัสเตอร์ของค่า P ต่างๆ

เมื่อนำคลัสเตอร์ EST มาระบุ SSR ด้วยซอฟท์แวร์ MISA พบว่า คลัสเตอร์ที่จัดตามค่า P 4 ค่า ให้จำนวนของ SSR แตกต่างกัน โดยคลัสเตอร์ของ EST ที่ค่า P เท่ากับ 100% พบจำนวน SSR เท่ากับ 171 ตำแหน่ง ที่ค่า P เท่ากับ 95% พบ 117 ตำแหน่ง ที่ค่า P เท่ากับ 90% พบ 103 ตำแหน่ง และที่ค่า P เท่ากับ 85% พบจำนวน SSR น้อยที่สุด จำนวน 97 ตำแหน่ง (Fig. 7)

การจัดคลัสเตอร์มีผลกระทบต่อจำนวนของ SSR ตามตัวอย่างแสดงใน Table 2 ร่วมกับลำดับนิวคลีโอไทด์เปรียบเทียบของ Figure 8 ในที่นี้ยกตัวอย่าง EST จำนวน 8 โมเลกุล ที่มีความยาวของลำดับนิวคลีโอไทด์แตกต่างกัน ได้แก่ BQ535513, BQ536965, BQ535537, BQ536131, BQ535386, BQ 53552, BQ537263 และ BQ536412 มาจัดคลัสเตอร์ตามค่า P ที่แตกต่างกัน ที่ค่า P เท่ากับ 100% พบว่า EST ทั้ง 8 โมเลกุล เป็นคลัสเตอร์ชนิด singleton (Fig. 8A) และให้จำนวน SSR เท่ากับ 5 ตำแหน่ง ที่ค่า P เท่ากับ 95% จัด EST ออกเป็น 3 คลัสเตอร์ คลัสเตอร์ที่ 1 ประกอบด้วยสมาชิก 4 โมเลกุล คลัสเตอร์ที่ 2 มีสมาชิก 3 โมเลกุล และคลัสเตอร์ที่ 3 มีสมาชิก 1 โมเลกุล ให้จำนวน SSR รวมจาก 3 คลัสเตอร์คือ 2 ตำแหน่ง (Fig. 8B) และที่ค่า P เท่ากับ 85 และ 90% จัดได้ 1 คลัสเตอร์ ที่มีสมาชิก 8 โมเลกุล ให้จำนวน SSR 1 ตำแหน่ง (Fig. 8C)

วิจารณ์ผลการทดลอง

คุณภาพของ EST

ลำดับนิวคลีโอไทด์ EST ของอ้อย จากห้องสมุด pSS จำนวน 2,268 โมเลกุล เข้าสู่กระบวนการคัดกรองและวิเคราะห์ ESTs ที่พัฒนาขึ้นในกลุ่มวิจัย (ปิยรัตน์ และนภาภรณ์, 2552) พบว่า ESTs จำนวน 6 โมเลกุล ไม่ผ่านเกณฑ์การประเมินเปอร์เซ็นต์ N และความยาว ซึ่งแสดงถึงปัญหาของปฏิกิริยาเคมีในการอ่านลำดับดีเอ็นเอ (Aaronson *et al.,* 1996) และเมื่อเข้าสู่การตรวจสอบการปนเปื้อนเวคเตอร์ภายใน EST พบว่ามีการปนเปื้อนลำดับเวคเตอร์จำนวน 213 โมเลกุล และ EST ที่มี low complexity อยู่ภายในจำนวน 94 โมเลกุล โดยชนิดที่พบมากที่สุด คือ GC-rich สาเหตุที่พบ low complexity ชนิด GC-rich เนื่องจากอ้อยจัดเป็นพืชใบเลี้ยงเดี่ยว วงศ์เดียวกับข้าว (*Oryza sativa*) ซึ่งมีรายงานว่าจีโนมของพืชกลุ่มนี้มีปริมาณ GC สูง (GC rich) (Kawabe and Miyashita, 2003) จีโนมที่มีปริมาณ GC สูง จะพบ

Table 2 Effect of P value on number of EST clusters and number of SSR

P value	Cluster of EST			Number of SSR
	Total cluster	No. of cluster	ESTs in cluster	
100%	8	1	BQ535513	5
		2	BQ536965	
		3	BQ535537	
		4	BQ536131	
		5	BQ537263	
		6	BQ535386	
		7	BQ535552	
		8	BQ536412	
95%	3	1	BQ535513	2
			BQ535537	
			BQ535552	
			BQ535386	
		2	BQ536965	
			BQ536131	
			BQ537263	
		3	BQ536412	
90%	1	1	BQ535513	1
			BQ536965	
			BQ535537	
			BQ536131	
			BQ537263	
			BQ535386	
			BQ535552	
			BQ536412	
85%	1	1	BQ535513	1
			BQ536965	
			BQ535537	
			BQ536131	
			BQ537263	
			BQ535386	
			BQ535552	
			BQ536412	

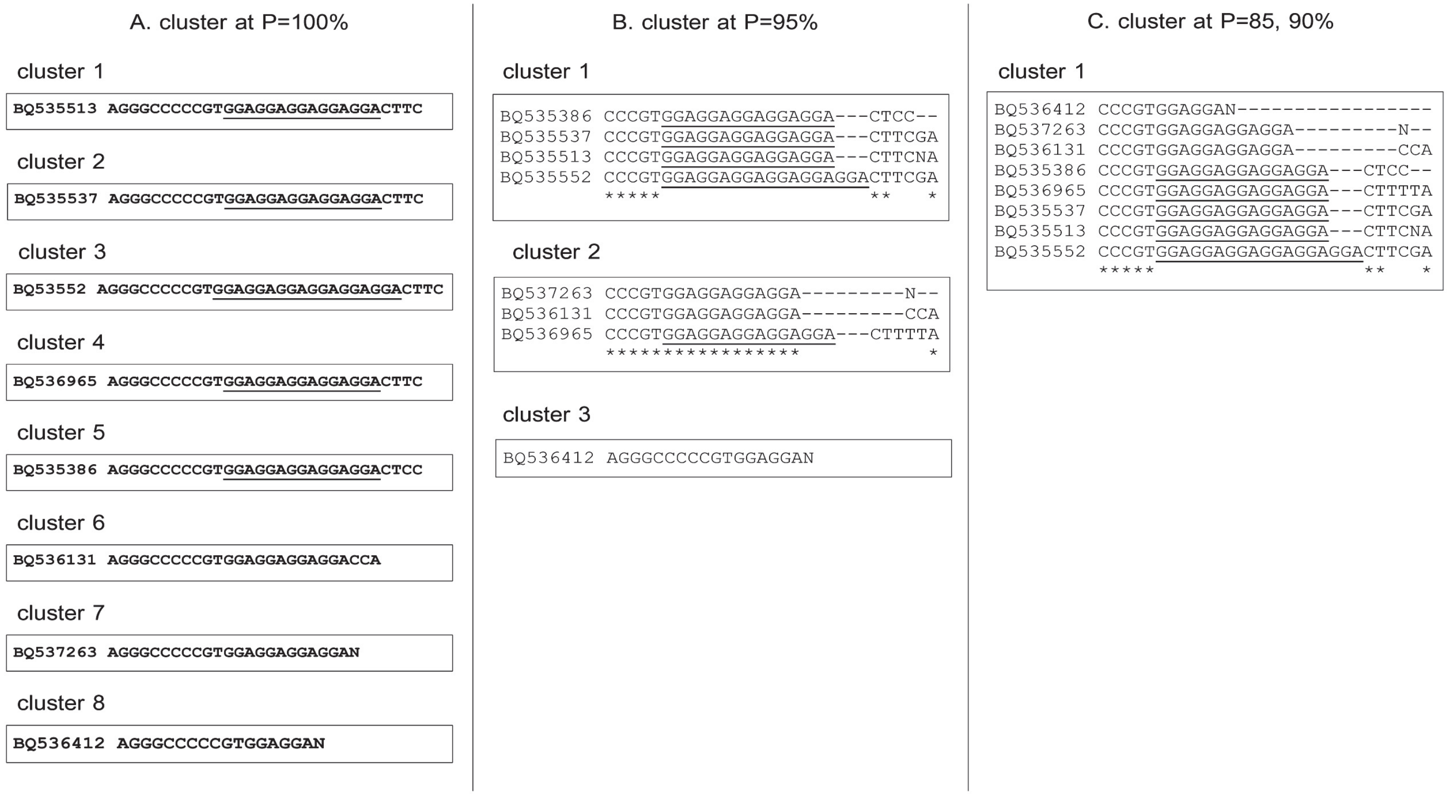

Figure 8 Comparison of loci and number of SSR in EST clusters at different P values. The clusters were partially displayed sequence alignment to highlight the SSR locus and flanking sequences. Asterisks represent identical nucleotide.

low complexity ชนิด GC-rich วางตัวอยู่ในบริเวณเอกซอนมากกว่าบริเวณอินทรอน (Wan and Wootton, 2000) การที่ EST ของ pSS มี low complexity ชนิด GC-rich ในสัดส่วนที่สูงสุด จึงบ่งชี้ว่าข้อมูล EST ชุดนี้มาจากทรานส์คริปต์ที่ผ่านกระบวนการตัดอินทรอนออกไปแล้ว

การจัดคลัสเตอร์ ESTs ตามค่า P 4 ระดับ

การจัดคลัสเตอร์จำเป็นต่อการลดขนาดและความซ้ำซ้อนของข้อมูล EST สำหรับข้อมูล EST ชุดนี้ ที่ค่า P เท่ากับ 100% มีผลลดจำนวน EST จาก 2,167 โมเลกุล เหลือ 2,087 คลัสเตอร์ แสดงให้เห็นว่า EST 80 โมเลกุลมีลำดับนิวคลีโอไทด์เหมือนกับ EST โมเลกุลอื่นจึงถูกจัดเข้าไปอยู่ตามกลุ่มคลัสเตอร์ต่างๆ ได้ ทั้งนี้ค่า P เท่ากับ 100% เป็นค่าสูงสุดของการจัดคลัสเตอร์ และ EST อย่างน้อย 2 โมเลกุลต้องมีลำดับนิวคลีโอไทด์เหมือนกันตลอดสาย แสดงว่าข้อมูล EST ใน pSS นี้มีส่วนหนึ่งผลิตมาจากยีนเดียวกัน (Uchimiya *et al.,* 1992)

การลดขนาดและความซ้ำซ้อนของข้อมูล EST มีความสัมพันธ์กับค่า P การปรับค่า P ลดลงจาก 100% ไปที่ 95% มีผลให้จำนวนคลัสเตอร์ลดลงอย่างก้าวกระโดดจาก 2,087 เหลือ 1,441 คลัสเตอร์ หรือข้อมูลลดลงไป 646 คลัสเตอร์ และการปรับลดค่า P จาก 95% ไปที่ 90% และจาก 90% ไปที่ 85% มีผลให้จำนวนคลัสเตอร์ลดลงเหลือ 1,347 และ 1,300 คลัสเตอร์ ตามลำดับ ทั้งนี้เป็นเพราะความเข้มงวดของการจับคู่ของนิวคลีโอไทด์ชนิดเดียวกัน (clustering stringency) ลดลง ทำให้จำนวนคลัสเตอร์ลดลง แต่ขนาดของคลัสเตอร์ใหญ่ขึ้น หรือจำนวนสมาชิกภายในคลัสเตอร์เพิ่มขึ้น (Miller *et al.,* 1999; Reed, 2001)

ค่า clustering stringency สูงสุดคือค่า P ที่ 100% ให้คลัสเตอร์ขนาดเล็ก ประกอบจาก EST จำนวน 1-5 โมเลกุล เป็นเพราะอัลกอริทึมของ Cd-hit คำนวณจำนวนนิวคลีโอไทด์ชนิดเดียวกัน ระหว่างคู่ของสาย EST และหารด้วยความยาวของ EST สายที่สั้นกว่า (Li *et al.,* 2002; Li and Godzik, 2006) ดังนั้น EST ที่จะถูกจัดอยู่ในคลัสเตอร์ที่ P เท่ากับ 100% จึงต้องมีจำนวนของนิวคลีโอไทด์ชนิดเดียวกันสูงกระจายตลอดสายของ EST ทั้ง 2 โมเลกุล รวมทั้ง EST ทั้ง 2 โมเลกุลต้องมีความยาวใกล้เคียงกัน การลดค่า clustering stringency ด้วยค่า P ที่ลดลง จึงเท่ากับเปิดโอกาสให้ EST ที่มีจำนวนนิวคลีโอไทด์ชนิดเดียวกันน้อยลง และความยาวของ EST ที่เข้ามาในคลัสเตอร์มีความแตกต่างกันได้มากขึ้น (Ptitsyn and Hide, 2005)

การตัดสินใจเลือกค่า P ที่เป็นเกณฑ์ตัดสินของการจัดคลัสเตอร์ (stringency threshold) ควรคำนึงว่าการใช้เปอร์เซ็นต์ของนิวคลีโอไทด์ชนิดเดียวกัน เป็นดัชนีความคล้ายคลึงระหว่างคู่ของ EST (similarity index) ที่ต้องมีค่าสูงถึง stringency threshold จะทำให้เกิดผลเสีย ในกรณีที่ตั้งค่า stringency threshold ไว้สูงมาก จนทำให้ EST ที่มีบางลำดับนิวคลีโอไทด์แตกต่างกันบ้างให้ค่า similarity index ต่ำกว่า threshold ทำให้จัด EST เหล่านี้เป็นคลัสเตอร์ชนิด singleton ดังเช่นในการทดลองนี้ P ที่ 100% จัด EST จำนวน 2,167 โมเลกุล ลงเหลือ 2,087 คลัสเตอร์ และเป็นคลัสเตอร์ชนิด singleton ในจำนวนสูงถึง 2,022 คลัสเตอร์

ลำดับนิวคลีโอไทด์เปรียบเทียบของ EST ที่เป็นสมาชิกของคลัสเตอร์ที่ค่า P ต่างๆ ดีต่อการประเมินความคล้ายคลึงของนิวคลีโอไทด์ระหว่าง EST ที่เป็นสมาชิกภายในคลัสเตอร์ และระบุบริเวณที่เป็น consensus sequence ภายในคลัสเตอร์ เป็นประโยชน์ต่อการเลือกตัวแทนของคลัสเตอร์ (Burke *et al.,* 1999; Miller *et al.,* 1999)

การจัดคลัสเตอร์ของ EST มีผลต่อการระบุ SSR

คุณภาพของคลัสเตอร์มีผลต่อการสืบค้น SSR จากเหมืองข้อมูล EST คลัสเตอร์ที่จัดตามค่า P เท่ากับ 100% ให้จำนวน SSR มากที่สุด 171 ตำแหน่ง ค่า P ที่

95% ให้จำนวน SSR 117 ตำแหน่ง ลดลงไป 54 ตำแหน่ง เมื่อ P มีค่าเท่ากับ 90% ให้จำนวน SSR 103 ตำแหน่ง ลดลงต่ออีก 14 ตำแหน่ง และคลัสเตอร์ที่จัดตามค่า P ที่ 80% ให้จำนวน SSR 97 ตำแหน่ง ลดลงต่อจากเดิมอีก 6 ตำแหน่ง

ลำดับนิวคลีโอไทด์เปรียบเทียบของแต่ละคลัสเตอร์ พิสูจน์ว่า SSR ส่วนใหญ่ของคลัสเตอร์จัดตามค่า P เท่ากับ 100% มาจากคลัสเตอร์ชนิด singleton ซึ่งเมื่อลดค่า P ลง EST ที่เคยเป็นสมาชิกของคลัสเตอร์ singleton ถูกจัดเข้ากลุ่มใหม่ที่มีเพื่อนสมาชิก EST อยู่ภายในคลัสเตอร์ ดังนั้นจำนวน SSR จึงลดลงตามไปด้วย การปรากฏของ consensus sequence ในคลัสเตอร์ขนาดใหญ่ เสริมสร้างความมั่นใจในลำดับนิวคลีโอไทด์ที่อยู่ด้านข้างของ SSR ซึ่งจำเป็นต่อการออกแบบไพรเมอร์ ข้อดีประการสุดท้ายคือบางคลัสเตอร์แสดงความหลากหลายของจำนวนชุดซ้ำ อันเป็นคุณสมบัติสำคัญของการใช้ SSR เป็นเครื่องหมายพันธุกรรมในปฏิบัติการที่อาศัยสารเคมี

เอกสารอ้างอิง

ปิยรัตน์ พลยะเรศ และ นภาภรณ์ ตันติสุวิชวงษ์. 2552. การตรวจสอบคุณภาพและความซ้ำซ้อนของลำดับนิวคลีโอไทด์ใน ESTs ที่ได้จากลำอ้อยที่เติบโตเต็มที่. ในการประชุมวิชาการเสนอผลงานวิจัยระดับบัณฑิตศึกษา ครั้งที่ 2. หน้า 161-171. บัณฑิตวิทยาลัย มหาวิทยาลัยราชภัฏจันทรเกษม กรุงเทพฯ.

Aaronson, J.S., Eckman, B., Blevins, R.A., Borkowski, J.A., Myerson, J., Imran, S. and Elliston, K.O. 1996. Toward the development of a gene index to the human genome: an assessment of the nature of high throughput EST sequence data. *Genome Res* 6: 829-845.

Adams, M.D., Kelly, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moremo, R.F., Kerlavage, A.R., McCombie, W.R. and Venter, J.C. 1991. Complementary DNA sequence tags and human genome project. *Science* 252: 1651-1656.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. 1997. Gapped BLAST and PSIBLAST: a new generation of protein database search programs. *Nucl Acids Res* 25: 3389-3402.

Burke, J., Davison, D. and Hide, W. 1999. d2_ cluster: a validated method for clustering EST and full-length cDNA sequences. *Genome Res* 9: 1135-1142.

Carson, D.L. and Botha, F.C. 2000. Preliminary analysis of expressed sequence tags for sugarcane. *Crop Sci* 40: 1769-1779.

Chen, Y.A., Lin, C.C., Wang, C.D., Wu, H.B. and Hwang, P.I. 2007 An optimized procedure greatly improves EST vector contamination removal. *BMC Bioinformatics* 8: 416.

Kawabe, A. and Miyashita, N.T. 2003. Patterns of codon usage bias in three dicot and four monocot plant species. *Gen Gene Systems* 78**:** 343-352.

Li, W. and Godzik, A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences**.** *Bioinformatics* 22: 1658-1659.

Li, W., Jaroszewski, L. and Godzik, A. 2002 Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics* 18: 77-82.

Miller, R.T., Christoffels, A.G., Gopalakrishnan, C., Burke, J., Ptitsyn, A.A., Broveak, T.R. and Hide, W.A. 1999. A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Res* 9: 1143-1155.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., Tsai, J. and Quackenbush, J. 2003. TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* 19: 651-652.

Ptitsyn, A. and Hide, W. 2005. CLU: A new algorithm for EST clustering. *BMC Bioinformatics* 6 (Suppl2): S3.

Reed, G. 2001. StackPACK clustering system. *Brief Bioinform* 2: 388-404.

Schulze, S.R., Ma, H.M., Meizhu Yang, J., Bowers, J.E., Mirkov, E. and Paterson, A.H. 2002. An EST survey of the sugarcane transcriptome. GenBank.

Smith, A.F.A., Hubley, R. and Green, P. 1996-2004. RepeatMasker Open-3.0. http://www.repeatmasker.org.

Thiel, T., Michalek, W., Varshney, R.K. and Graner, A. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor Appl Genet* 106: 411-422.

Uchimiya, H., Kidou, S., Shimazaki, T., Aotsuka, S., Takamatsu, S., Nishi, R., Hashimoto, H., Matsubayashi, Y., Kidou, N., Umeda, M. and Kata, A. 1992. Random sequenceing of cDNA libraries reveals a variety of expressed genes in cultured-cells of rice (*Oryza sativa* L.). *Plant J* 2: 1005-1009.

Wan, H. and Wootton, J.C. 2000. A global compositional complexity measure for biological sequences: AT-rich and GC-rich genomes encode less complex proteins. *Comp Chem* 24: 71-94.