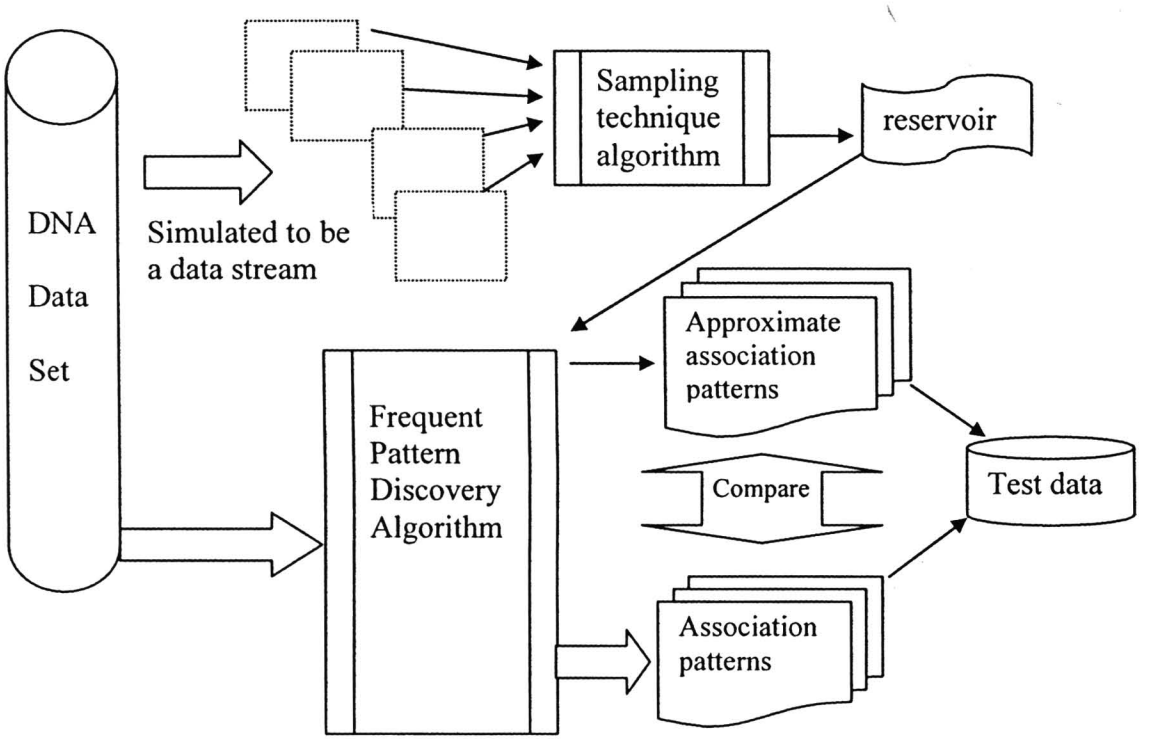
บทที่ 3
การทดสอบโปรแกรมและผลการทดสอบ

3.1 วิธีการทดสอบโปรแกรม

การทดสอบประสิทธิภาพของวิธีการค้นพบโดยประมาณของรูปแบบที่ปรากฏบ่อยในข้อมูลสตรีม ใช้วิธีการจำลองข้อมูลรหัสพันธุกรรมให้แบ่งเป็นข้อมูลหลายไฟล์ในลักษณะของสตรีม แต่ละส่วนของข้อมูลจะถูกสุ่มเก็บไว้ใน reservoir จากนั้นนำข้อมูลที่รวบรวมเก็บไว้ใน reservoir ประมวลผลด้วยโปรแกรมค้นพบรูปแบบที่ปรากฏบ่อย ผลลัพธ์ที่ได้จากโปรแกรมจะเป็น association rules ที่แสดงความสัมพันธ์ที่ปรากฏขึ้นบ่อยที่สุดในกลุ่มข้อมูล (top-k frequent patterns) ความสมบูรณ์ของผลลัพธ์จะถูกนำไปเปรียบเทียบกับกรณีที่ข้อมูลไม่ได้ถูกจำลองเป็นสตรีม ความถูกต้องของผลลัพธ์จะทดสอบด้วยการนำ top-k frequent patterns ไปทำนายรูปแบบในชุดข้อมูลทดสอบ แนวคิดของวิธีการทดสอบนี้แสดงเป็นแผนภาพได้ดังรูปที่ 3.1



รูปที่ 3.1 วิธีการทดสอบความสมบูรณ์และถูกต้องของการค้นพบรูปแบบที่ปรากฏบ่อยในข้อมูลสตรีม

3.2 ข้อมูลที่ใช้และผลการทดสอบ

ข้อมูลที่ใช้ในการทดสอบโปรแกรมเป็นข้อมูลรหัสพันธุกรรม หรือข้อมูล DNA ประกอบด้วยข้อมูลสองชุด ข้อมูลชุดแรกมีจำนวน 2,000 รายการ ใช้ในการค้นพบรูปแบบที่ปรากฏบ่อยทั้งในลักษณะค้นจากข้อมูลทั้งหมด และค้นโดยประมาณจากข้อมูลที่สุ่มเก็บไว้ใน reservoir ข้อมูลชุดที่สองมีจำนวน 1,186 รายการ ข้อมูลทั้งสองชุดมีโครงสร้างเป็นแบบเดียวกันดังแสดงตัวอย่างในรูปที่ 3.2 แต่ละรายการข้อมูลจะประกอบด้วยรหัสพันธุกรรมจำนวน 60 ตำแหน่ง ในแต่ละตำแหน่งอาจปรากฏเป็นสัญลักษณ์ A (adenine) หรือ T (thymine) หรือ C (cytosine) หรือ G (guanine) ตามค่าเบสที่เป็นไปได้ของสายดีเอ็นเอ ค่าในตำแหน่งที่ 61 ระบุเป็น exon/intron หมายถึง สายดีเอ็นเอนั้นเป็นช่วงรอยต่อจาก exon เป็น intron และในทำนองเดียวกันถ้าค่าในตำแหน่งที่ 61 ระบุเป็น intron/exon หมายถึง สายดีเอ็นเอนั้นเป็นช่วงรอยต่อจาก intron เป็น exon โดยที่ exon จะเป็นช่วงของสายดีเอ็นเอที่ถูกใช้ในการคัดลอกรหัสพันธุกรรม รวมถึงใช้ในการสร้างโปรตีน ในกรณีตำแหน่งที่ 61 ระบุเป็น none หมายถึงสายดีเอ็นเอรายการนั้นไม่ใช่ช่วงรอยต่อ intron และ exon แผนภาพของรหัสพันธุกรรมแสดงได้ดังรูปที่ 3.3

T,T,C,T,A,T,G,A,G,A,A,A,C,G,T,G,G,C,A,T,T,G,T,G,C,G,C,A,A,G,G,T,G,G,G,C,
C,C,C,G,C,G,G,G,A,C,G,G,G,G,C,A,G,C,T,C,C,G,G,G,exon/intron

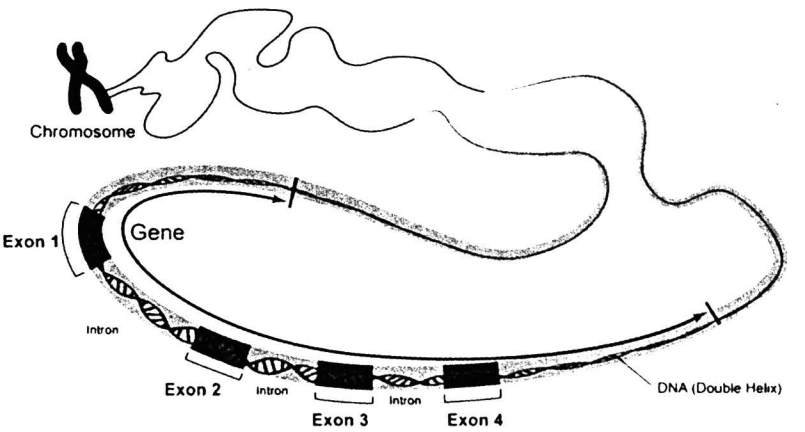
C,T,C,C,C,C,A,C,C,C,A,C,C,T,G,T,C,C,A,C,C,C,G,C,C,C,G,C,A,G,A,T,C,G,C,T,T,
C,C,T,G,G,A,G,C,C,A,G,G,C,A,A,G,A,A,C,T,C,C,A,intron/exon

C,T,G,A,C,T,A,A,G,C,C,G,C,C,C,C,T,T,G,T,C,C,C,T,T,C,T,C,A,G,A,T,T,A,T,G,T,
T,T,G,A,G,A,C,C,T,T,C,A,A,C,A,C,C,C,C,C,G,G,C,C,intron/exon

G,A,G,G,A,G,C,T,A,G,A,C,A,A,G,T,A,C,T,G,G,T,C,T,C,A,G,C,A,G,G,T,G,C,G,T,
G,A,G,G,G,G,A,G,G,G,G,A,T,G,G,C,T,G,C,C,A,A,G,G,exon/intron

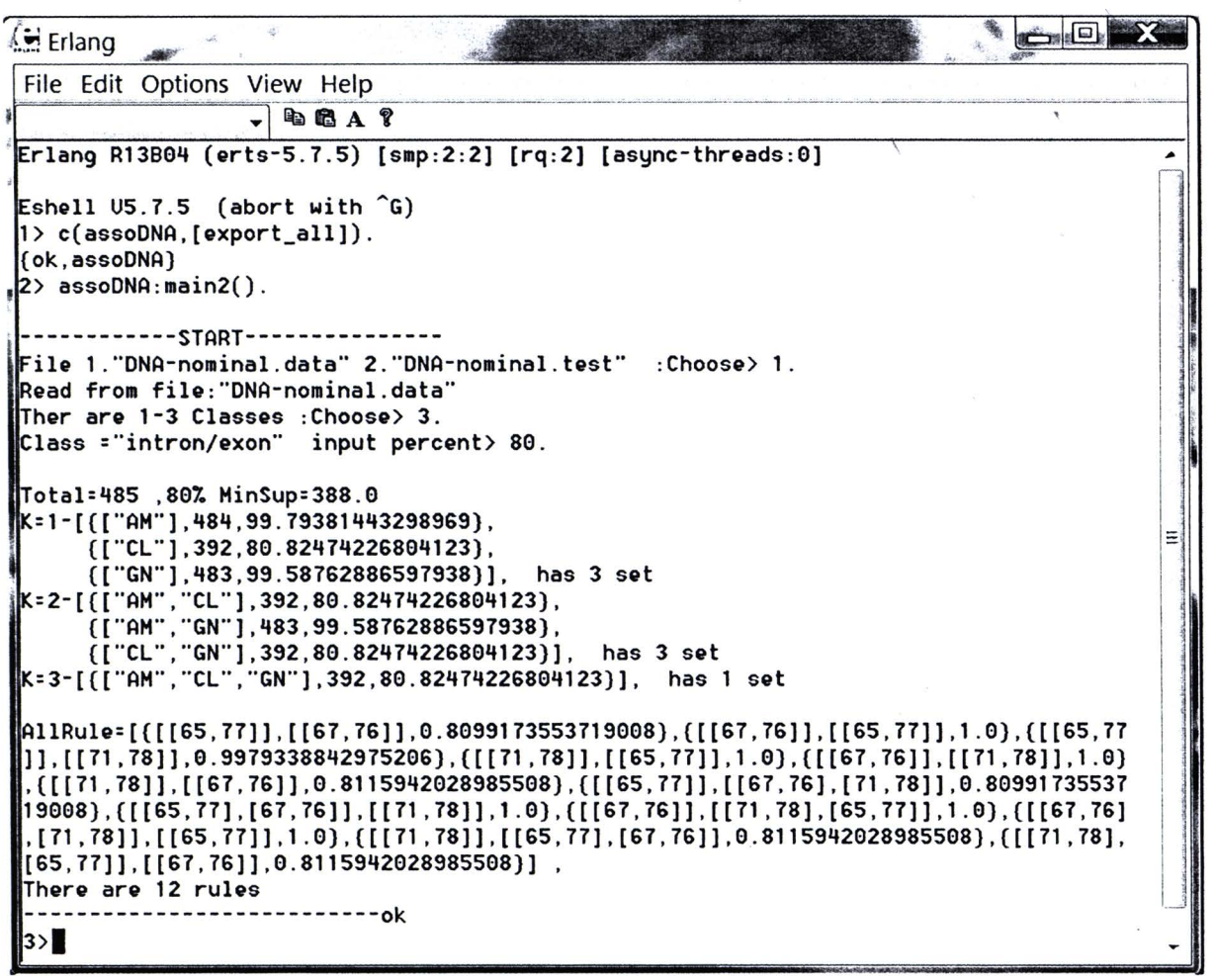
A,A,G,G,C,T,C,A,G,G,A,G,G,A,G,G,G,A,G,A,T,C,A,A,C,A,T,C,A,A,C,C,T,G,C,C,
C,C,G,C,C,C,C,C,T,C,C,C,C,A,G,C,C,T,G,A,T,A,A,A,none

รูปที่ 3.2 ตัวอย่างข้อมูลรหัสพันธุกรรม



รูปที่ 3.3 โครงสร้างของสายรหัสพันธุกรรมที่ประกอบด้วยส่วน exon และ intron (http://genome.gov/Glossary/)

การทดสอบโปรแกรมค้นพบรูปแบบที่ปรากฏบ่อย ใช้ข้อมูลดีเอ็นเอหรือข้อมูลรหัสพันธุกรรม ชุดที่ประกอบด้วยข้อมูล 2,000 รายการ วิธีการรันโปรแกรมแสดงตัวอย่างได้ดังรูปที่ 3.4 คำสั่งแรกที่แสดงในรูปเป็นการคอมไพล์โปรแกรมด้วยคำสั่ง "c(assoDNA, [export_all])." โดย assoDNA เป็นชื่อโมดูล และ export_all เป็นการระบุให้คอมไพเลอร์ export ทุกฟังก์ชันในโมดูลมายัง Erlang shell คำสั่งที่สองเป็นการเริ่มรันโปรแกรมด้วยการเรียกคำสั่ง "assoDNA:main2()." ฟังก์ชัน main2 จะเริ่มทำงานโดยการโต้ตอบกับผู้ใช้ให้เลือกไฟล์ข้อมูล (ในที่นี้ใช้ไฟล์ DNA-nominal.dat) ในข้อมูลประกอบด้วยข้อมูล 3 คลาสคือ none, exon/intron, intron/exon โปรแกรมจะแยกข้อมูลตามคลาสแล้วจึงค้นหารูปแบบที่ปรากฏบ่อยในคลาสที่ผู้ใช้ระบุ ข้อมูลสุดท้ายที่ผู้ใช้ต้องระบุคือค่า minimum support ซึ่งในรูปที่ 3.4 ผู้ใช้ระบุค่านี้เป็น 80% และด้วยค่าเกณฑ์ขั้นต่ำนี้โปรแกรมสามารถหา 1-itemset ได้จำนวน 3 เซต, 2-itemset จำนวน 3 เซต, และ 3-itemset ได้จำนวน 1 เซต ในแต่ละเซตจะระบุจำนวนข้อมูลและค่า % support ในช่วงสุดท้ายของผลลัพธ์จะเป็นการระบุ association rules พร้อมค่า confidence ของ rule



รูปที่ 3.4 ตัวอย่างการใช้โปรแกรมค้นพบรูปแบบที่ปรากฏบ่อยในข้อมูลรหัสพันธุกรรม

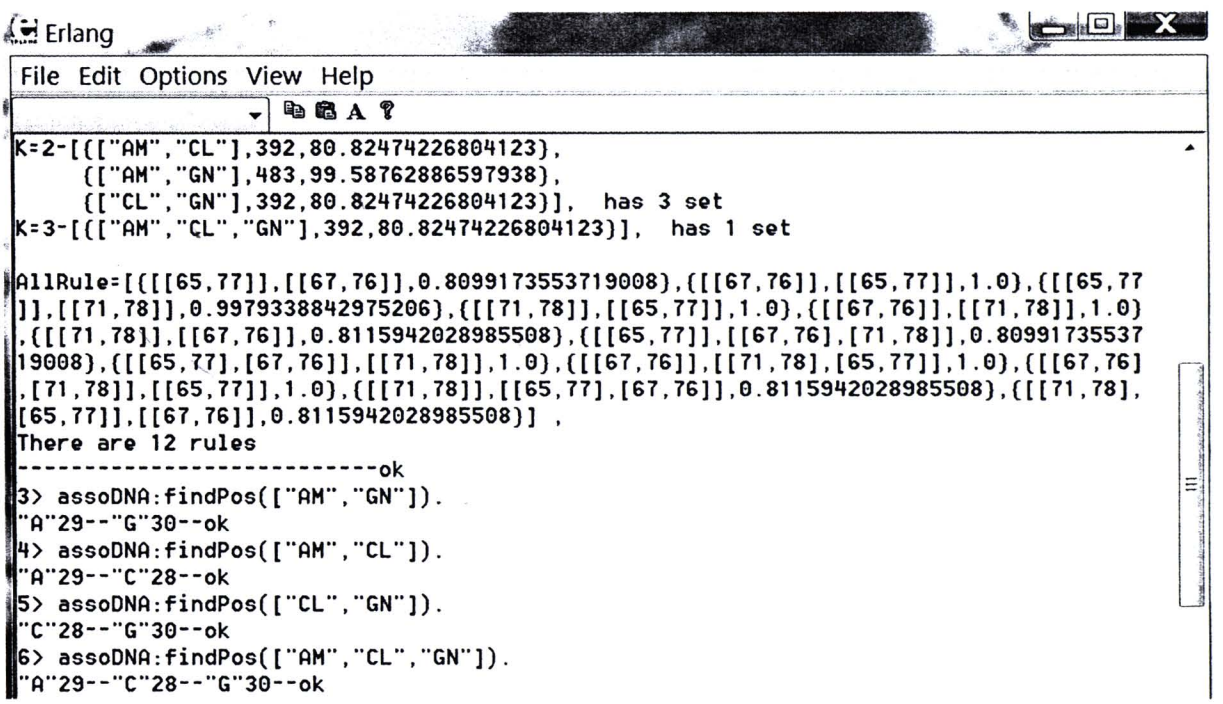
ภาษาเออแลงใช้ลิสต์ของรหัส ASCII แทนรูปแบบข้อมูลสตริง ทำให้การอ่านผลลัพธ์ทำได้ค่อนข้างยาก เพื่อให้การแปลความรูปแบบที่ปรากฏบ่อยทำได้ง่ายขึ้น ผู้วิจัยได้เพิ่มเติมฟังก์ชันการแปลความหมายรูปแบบที่ปรากฏบ่อย หรือ association patterns โดยผู้ใช้สามารถเรียกใช้คำสั่ง "assoDNA:findPos([Patterns])" ดังตัวอย่างในรูปที่ 3.5 ในกรณีที่รหัสพันธุกรรมนั้นเป็นช่วงเปลี่ยน intron/exon สามารถแปลความรูปแบบต่างๆได้ดังนี้

แพทเทิร์น ["AM", "GN"] หมายถึง การปรากฏเบส A ในตำแหน่งที่ 29 ของสายดีเอ็นเอ จะเกิดร่วมกันกับการปรากฏเบส G ที่ตำแหน่งที่ 30

แพทเทิร์น ["AM", "CL"] หมายถึง การปรากฏเบส A ในตำแหน่งที่ 29 ของสายดีเอ็นเอ จะเกิดร่วมกันกับการปรากฏเบส C ที่ตำแหน่งที่ 28

แพทเทิร์น ["CL", "GN"] หมายถึง การปรากฏเบส C ในตำแหน่งที่ 28 ของสายดีเอ็นเอ จะเกิดร่วมกันกับการปรากฏเบส G ที่ตำแหน่งที่ 30

แพทเทิร์น ["AM", "CL", "GN"] หมายถึง การปรากฏเบส A ในตำแหน่งที่ 29 ของสายดีเอ็นเอ จะเกิดร่วมกันกับการปรากฏเบส C ที่ตำแหน่งที่ 28 และเบส G ที่ตำแหน่งที่ 30

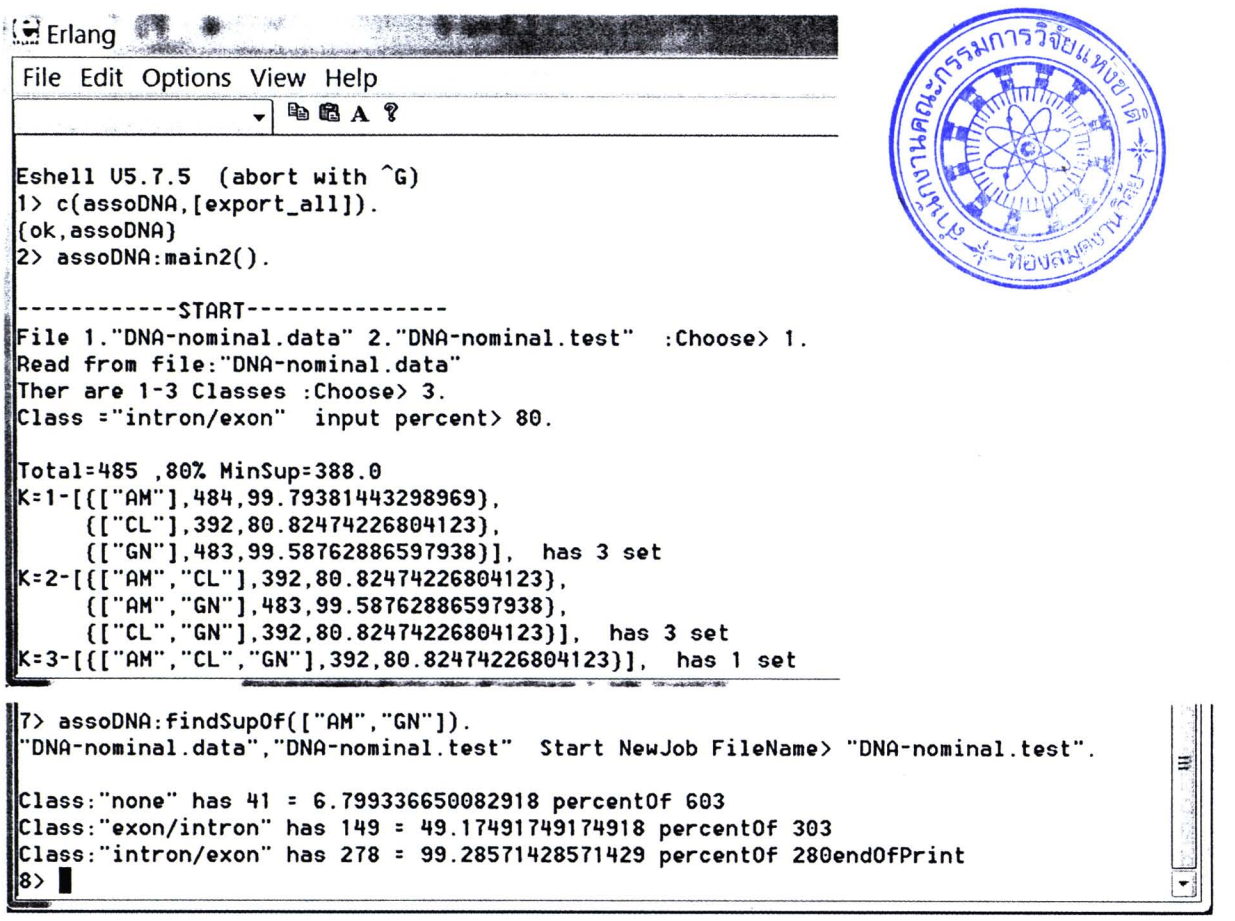


รูปที่ 3.5 ตัวอย่างการใช้ฟังก์ชัน findPos แปลความหมายรูปแบบรหัสพันธุกรรม

การทดสอบความถูกต้องของรูปแบบที่ปรากฏบ่อย ใช้วิธีการทดสอบค้นหารูปแบบในชุดข้อมูลทดสอบที่ประกอบด้วยข้อมูล 1,186 รายการ การค้นหาใช้ฟังก์ชัน findSupOf ดังแสดงในรูปที่ 3.6 จากภาพแสดงการค้นหาแพทเทิร์น ["AM", "GN"] ซึ่งหมายถึง การปรากฏเบส A ในตำแหน่งที่ 29 ของสายดีเอ็นเอ จะเกิดร่วมกันกับการปรากฏเบส G ที่ตำแหน่งที่ 30 (แพทเทิร์นนี้เป็นกรณี intron/exon)

ผลการค้นหาในชุดข้อมูลทดสอบปรากฏว่าแพทเทิร์น ["AM", "GN"] ปรากฏในกลุ่มข้อมูล none เพียง 6.79% และปรากฏในกลุ่มข้อมูล exon/intron เพียง 49.17% แต่ในกลุ่มของ intron/exon ซึ่งเป็นกลุ่มประเภทเดียวกันกับแพทเทิร์นที่นำไปค้นหา ปรากฏว่าพบแพทเทิร์น ["AM", "GN"] มากถึง 99.28%

การปรากฏผลเช่นนี้แสดงว่าเทคนิคการค้นหาแพทเทิร์นหรือรูปแบบที่ปรากฏบ่อย สามารถใช้ระบุความสัมพันธ์ของลำดับเบสในสายดีเอ็นเอได้ค่อนข้างชัดเจน



รูปที่ 3.6 การใช้ฟังก์ชัน findSupOf เพื่อตรวจสอบรูปแบบที่ปรากฏบ่อยในชุดข้อมูลทดสอบ

การทดสอบเพื่อเปรียบเทียบความสมบูรณ์และถูกต้องของวิธีการค้นพบรูปแบบที่ปรากฏบ่อยจากข้อมูลสตรีมแบบใช้ข้อมูลทั้งหมด เพื่อเปรียบเทียบกับแบบโดยประมาณแสดงผลการทดสอบได้ดังตารางที่ 3.1 และ 3.2

ตารางที่ 3.1 เปรียบเทียบจำนวนแพทเทิร์นที่ค้นพบจากข้อมูลสตรีม

Min_sup	การค้นพบแบบใช้ข้อมูลทั้งหมด				การค้นพบโดยประมาณ(ใช้ข้อมูลตัวแทน)				จำนวนแพทเทิร์นที่ตรงกัน
	# 1-item	# 2-item	# 3-item	# 4-item	# 1-item	# 2-item	# 3-item	# 4-item	
Class = "none"									
50%	0	0	0	0	0	0	0	0	0
45%	0	0	0	0	0	0	0	0	0
40%	0	0	0	0	0	0	0	0	0
35%	0	0	0	0	0	0	0	0	0
30%	1	0	0	0	1	0	0	0	1
25%	117	0	0	0	111	0	0	0	111
Class = "exon/intron"									
85%	3	2	0	0	3	2	0	0	5
80%	4	5	2	0	4	5	2	0	11
75%	4	5	2	0	4	5	2	0	11
70%	4	6	3	0	4	6	3	0	13
65%	5	8	5	1	5	8	5	1	19
60%	5	9	7	2	5	8	5	1	19
Class = "intron/exon"									
85%	2	1	0	0	2	1	0	0	3
80%	3	3	1	0	3	3	1	0	7
75%	3	3	1	0	3	3	1	0	7
70%	3	3	1	0	3	3	1	0	7
65%	3	3	1	0	3	3	1	0	7
60%	3	3	1	0	3	3	1	0	7

ตารางที่ 3.2 เปรียบเทียบสัดส่วนการค้นพบแพทเทิร์นจากข้อมูลทดสอบ

แพทเทิร์นจากคลาสที่ตรงกัน	แพทเทิร์นจากการค้นพบในข้อมูลทั้งหมด		แพทเทิร์นจากการค้นพบโดยประมาณด้วยข้อมูลบางส่วน	
	2-item pattern	3-item pattern	2-item pattern	3-item pattern
None	--	--	--	--
Exon/intron	91.24%	84.35%	90.98%	83.27%
Intron/exon	90.11%	87.62%	90.09%	86.89%

การทดสอบนับจำนวนแพทเทิร์นของวิธีการค้นพบรูปแบบที่ปรากฏบ่อย โดยการใช้ข้อมูลทั้งหมดจากสตรีม เมื่อเทียบกับวิธีการค้นพบโดยประมาณของรูปแบบที่ปรากฏบ่อยโดยการใช้ข้อมูลตัวแทนบางส่วน พบว่าให้จำนวนแพทเทิร์นที่ใกล้เคียงกันมาก โดยมีข้อสังเกตในกรณีของข้อมูลในกลุ่มที่รหัสพันธุกรรมเป็น "none" ไม่ปรากฏรูปแบบหรือแพทเทิร์นที่มีขนาด 2-item แสดงว่าข้อมูลไม่มีความสัมพันธ์ที่ชัดเจน

ในขั้นตอนของการทดสอบการค้นหาแพทเทิร์นแบบเดียวกัน แต่เป็นการค้นหาจากข้อมูลชุดทดสอบพบว่าสัดส่วนของการค้นหาพบ (หรือค่า recall) ค่อนข้างสูง และให้ค่าใกล้เคียงกันทั้งในกรณีที่แพทเทิร์นนั้นได้จากวิธีการใช้ข้อมูลทั้งหมด และจากวิธีการใช้ข้อมูลตัวแทน จึงทำให้สามารถสรุปได้ว่าเทคนิคการค้นหารูปแบบที่ปรากฏบ่อยโดยประมาณด้วยข้อมูลตัวแทนที่ดี จะให้ผลลัพธ์ที่ถูกต้องมีความน่าเชื่อถือค่อนข้างสูง