การรู้จำยีนของข้าวโดยแบบจำลองมาร์คอฟ

นางสาว ปวีณา เลิศอำไพพร

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2545
ISBN 974-17-0852-1

RICE GENE RECOGNITION BY MARKOV MODELS

Miss Paveena Lertampaiporn

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science in Computer Science

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2002

ISBN 974-17-0852-1

หัวข้อวิทยานิพนธ์	การรู้จำยีนของข้าวโดยแบบจำลองมาร์คอฟ
โดย	นางสาวปวีณา เลิศอำไพพร
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร. บุญเสริม กิจศิริกุล

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร. สมศักดิ์ ปัญญาแก้ว)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(รองศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา)

..... อาจารย์ที่ปรึกษา
(ผู้ช่วยศาสตราจารย์ ดร. บุญเสริม กิจศิริกุล)

..... กรรมการ
(อาจารย์ ดร. เศรษฐา ปานงาม)

..... กรรมการ
(อาจารย์ ธงชัย โรจน์กังสดาล)

บทคัดย่อวิทยานิพนธ์

ปวีณา เลิศอำไพพร : การรู้จำยีนของข้าวโดยแบบจำลองมาร์คอฟ. (RICE GENE RECOGNITION BY MARKOV MODELS) อ.ที่ปรึกษา : ผศ. ดร.บุญเสริม กิจศิริกุล, จำนวนหน้า 78 หน้า. ISBN 974-17-0852-1.

ปัจจุบันเทคโนโลยีที่เกี่ยวข้องกับงานด้านดีเอ็นเอได้รับการพัฒนาให้มีประสิทธิภาพสูงขึ้น ทำให้สามารถทำงานที่เกี่ยวข้องกับสายดีเอ็นเอได้รวดเร็วมากขึ้น ฐานข้อมูลที่เก็บข้อมูลเหล่านี้ก็มีจำนวนมากขึ้นตามไปด้วย ทำให้เกิดงานวิจัยในสาขาใหม่ที่เรียกว่า ชีวสารสนเทศศาสตร์ ซึ่งเป็นสาขาที่เกี่ยวข้องกับการตีความหมายและการหาข้อสรุปจากข้อมูลดีเอ็นเอที่มีจำนวนมาก ตัวอย่างของงานทางด้านชีวสารสนเทศศาสตร์ เช่น การทำนายหารูปแบบของโปรตีน การทำนายหารูปแบบของยีน เป็นต้น

วัตถุประสงค์ของวิทยานิพนธ์นี้ เพื่อพัฒนาโปรแกรมค้นหายีนข้าวโดยใช้มาร์คอฟโมเดลระดับ 5 โดยในขั้นตอนแรกต้องมีการเก็บรวบรวมข้อมูลสายนิวคลีโอไทด์ข้าวที่มีการหายีนไว้เรียบร้อยแล้ว จากนั้นแบ่งข้อมูลออกเป็น 2 ส่วน โดยแบ่งเป็นข้อมูลสอน และข้อมูลทดสอบ ขั้นต่อมานำข้อมูลสอนมาหาค่าพารามิเตอร์ของเมตริกซ์น้ำหนักสำหรับ จุดเริ่มต้นของยีน จุดสิ้นสุดของยีน ดอร์เนอร์ และแอคเซพเตอร์ จากนั้นหาค่าความน่าจะเป็นเริ่มต้น และความน่าจะเป็นในการเปลี่ยนสถานะของมาร์คอฟโมเดล และสุดท้าย ทดสอบความถูกต้องในการค้นหายีนข้าวของโปรแกรมโดยหาค่า sensitivity และ ค่า specificity ในขั้นตอนการทดสอบนั้นจะใช้ข้อมูลทดสอบจำนวน 557 ยีน จะได้ค่า Sensitivity เท่ากับ 0.775 และค่า Specificity เท่ากับ 0.81

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต ปวีณา เลิศอำไพพร
สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่ออาจารย์ที่ปรึกษา
ปีการศึกษา 2545

AN ABSTRACT

##4370385521 : MAJOR COMPUTER SCIENCE

KEY WORD: BIOINFORMATICS / GENE FINDING / MARKOV MODEL / RICE / WEIGHT MATRIX

PAVEENA LERTAMPAIPORN : RICE GENE RECOGNITION BY MARKOV MODELS, THESIS ADVISOR: BOONSERM KIJSIRIKUL ,PhD., 78 pp. ISBN 974-17-0852-1.

In recent years, development of the technology for efficient, automated DNA sequencing has led to the accumulation of large databases of DNA and protein sequences and a new field of study known as "Bioinformatics" has begun to take shape as researchers work to interpret and draw conclusions from this wealth of new information. Some of the researches in this area include prediction of protein structure and prediction of gene structure.

The objective of this research is to develop a rice gene-finding program by using a 5th order Markov model. First, we collect datasets of annotated rice gene sequences and divide them into training and test sets. We then calculate weight matrices for start sites, stop sites, donors and acceptors. After that, we train the Markov model for finding initial probabilities and transition probabilities. Finally, we test the performance of the rice gene-finding program based on the values of sensitivity and specificity. The program was tested on the testing set containing 557 genes. The result of testing indicates that the program has 0.775 sensitivity and 0.81 specificity.

Department Computer Engineering
Field of study Computer Science
Academic year 2002

Student's signature.....
Advisor's signature.....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลืออย่างดียิ่งจาก ผู้ช่วยศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งท่านได้สละเวลาในการให้คำปรึกษาและคำแนะนำ ตลอดจนช่วยตรวจแก้ไขวิทยานิพนธ์ด้วยความเอาใจใส่อย่างดียิ่ง ผู้วิจัยขอกราบขอบพระคุณในความกรุณาเป็นอย่างสูง

ขอขอบพระคุณ รองศาสตราจารย์ นพ. ประสิทธิ์ ผลิตผลการพิมพ์ ศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ ที่ท่านได้ให้คำแนะนำและคำปรึกษา

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร. สมจารี ปรียานนท์ คณะพาณิชยศาสตร์และการบัญชี ภาควิชาสถิติ ที่ท่านได้ให้คำแนะนำและ ความช่วยเหลือในการทำวิทยานิพนธ์

ขอขอบพระคุณ ดร. ชิเน ธำมรงค์ธรรม ศูนย์พันธุวิศวกรรมและเทคโนโลยีชีวภาพแห่งชาติ ที่ท่านได้ให้ข้อมูล และเอกสารที่ใช้ในการทำวิทยานิพนธ์

ท้ายนี้ผู้วิจัยใคร่ขอกราบขอบพระคุณ บิดามารดา รวมถึงทุกคนในครอบครัวซึ่งให้การสนับสนุนและกำลังใจแก่ผู้วิจัยเสมอมาจนสำเร็จการศึกษา

สารบัญ

หน้า

สารบัญ (ต่อ)

บทที่ หน้า

สารบัญตาราง

ตาราง หน้า

สารบัญภาพ

ภาพประกอบ หน้า

สารบัญภาพ (ต่อ)

ภาพประกอบ หน้า

บทที่ 1

บทนำ

ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันพัฒนาการทางคอมพิวเตอร์และอินเตอร์เน็ตได้รับการพัฒนาจนก้าวหน้าไปไกลมาก ทำให้มีขีดความสามารถในการประมวลผลข้อมูลสูงขึ้นกว่าเดิม จากประสิทธิภาพที่ดีขึ้นมากของคอมพิวเตอร์ ประจวบกับการสร้างเครื่องมือเพื่อศึกษาสิ่งมีชีวิตในระดับโมเลกุลได้มากขึ้น การผสมผสานศาสตร์ทางด้านคอมพิวเตอร์กับชีววิทยาจึงเกิดขึ้น โดยเฉพาะอย่างยิ่งการถอดรหัสพันธุกรรม ซึ่งมีข้อมูลทางพันธุกรรมมากมายที่จะต้องถอดรหัส และต้องใช้ขีดความสามารถเชิงการคำนวณและเทคนิคที่สูงตามด้วย ข้อมูลทางพันธุกรรมและการขยายวิธีการที่เกี่ยวกับการประมวลผล จึงเป็นที่มาของศาสตร์แขนงใหม่ที่กำลังได้รับความสนใจ และจะเป็นประโยชน์ต่อมนุษย์อย่างมาก คือ ชีวสารสนเทศศาสตร์ (Bioinformatics)

ชีวสารสนเทศศาสตร์เป็นการนำสารสนเทศมาประยุกต์ใช้เพื่อจัดการกับข้อมูลทางชีวภาพ และเนื่องจากฐานข้อมูลทางชีวภาพมีการขยายตัวอย่างต่อเนื่องและรวดเร็ว ดังนั้นนักวิทยาศาสตร์และนักวิจัยจึงหนีไม่พ้นที่จะต้องพึ่งพาคอมพิวเตอร์และเทคโนโลยีสารสนเทศเพื่อให้ตามทันข้อมูลที่มีการเติบโตและเปลี่ยนแปลงอย่างรวดเร็ว การวิจัยด้านไบโออินฟอร์เมติกส์มีหลายส่วน นับตั้งแต่การค้นหาข้อมูล การใช้ฐานข้อมูลชีวภาพที่มีให้บริการทางอินเตอร์เน็ต การวิเคราะห์ลำดับนิวคลีโอไทด์ (Nucleotide) ของดีเอ็นเอ (DNA) หรืออาร์เอ็นเอ (RNA) หรือลำดับกรดอะมิโนของเปปไทด์หรือโปรตีนซึ่งได้จากการทดลอง การพยากรณ์ลำดับนิวคลีโอไทด์หรือกรดอะมิโนนั้นว่าจะเป็นยีนอะไรมีการแสดงออกอย่างไร หรือมีบทบาทหน้าที่ในสิ่งมีชีวิตอย่างไร การศึกษาโครงสร้าง 3 มิติ ซึ่งรวมไปถึงการศึกษาถึงคุณสมบัติทางด้านกายภาพและชีวภาพของโครงสร้างต่างๆ ด้วย

นักวิทยาศาสตร์ในหลายๆประเทศให้ความสำคัญกับการทำวิจัยที่เกี่ยวข้องกับโครงการทางด้านการถอดรหัสพันธุกรรมของสิ่งมีชีวิตต่างๆ ตัวอย่างที่เห็นได้ชัด เช่น โครงการฮิวแมนจีโนม (Human Genome Project - HGP) เป็นโครงการระหว่างประเทศที่ร่วมมือกันทำงานหลายองค์กร โดยใช้เงินลงทุนจำนวนมาก ใช้นักวิทยาศาสตร์ นักชีววิทยา นักคอมพิวเตอร์ มาร่วมกันศึกษาและวิเคราะห์ โดยช่วยกันจัดทำข้อมูลสายรหัสพันธุกรรม และรวบรวมเก็บเป็นฐานข้อมูลกลางที่เรียกเข้าถึงหรือดาวน์โหลดมาปรับปรุงข้อมูลและใช้ร่วมกัน การถอดรหัสพันธุกรรมนั้น เป็นเรื่องที่สำคัญและจะให้ประโยชน์แก่มวลมนุษย์ชาติมหาศาล เพราะรหัสพันธุกรรมที่ได้จะเกี่ยวข้องกับ

บทที่ 1

บทนำ

ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันพัฒนาการทางคอมพิวเตอร์และอินเตอร์เน็ตได้รับการพัฒนาจนก้าวหน้าไปไกลมาก ทำให้มีขีดความสามารถในการประมวลผลข้อมูลสูงขึ้นกว่าเดิม จากประสิทธิภาพที่ดีขึ้นมากของคอมพิวเตอร์ ประจวบกับการสร้างเครื่องมือเพื่อศึกษาสิ่งมีชีวิตในระดับโมเลกุลได้มากขึ้น การผสมผสานศาสตร์ทางด้านคอมพิวเตอร์กับชีววิทยาจึงเกิดขึ้น โดยเฉพาะอย่างยิ่งการถอดรหัสพันธุกรรม ซึ่งมีข้อมูลทางพันธุกรรมมากมายที่จะต้องถอดรหัส และต้องใช้ขีดความสามารถเชิงการคำนวณและเทคนิคที่สูงตามด้วย ข้อมูลทางพันธุกรรมและการขยายวิธีการที่เกี่ยวกับการประมวลผล จึงเป็นที่มาของศาสตร์แขนงใหม่ที่กำลังได้รับความสนใจ และจะเป็นประโยชน์ต่อมนุษย์อย่างมาก คือ ชีวสารสนเทศศาสตร์ (Bioinformatics)

ชีวสารสนเทศศาสตร์เป็นการนำสารสนเทศมาประยุกต์ใช้เพื่อจัดการกับข้อมูลทางชีวภาพ และเนื่องจากฐานข้อมูลทางชีวภาพมีการขยายตัวอย่างต่อเนื่องและรวดเร็ว ดังนั้นนักวิทยาศาสตร์และนักวิจัยจึงหนีไม่พ้นที่จะต้องพึ่งพาคอมพิวเตอร์และเทคโนโลยีสารสนเทศเพื่อให้ตามทันข้อมูลที่มีการเติบโตและเปลี่ยนแปลงอย่างรวดเร็ว การวิจัยด้านไบโออินฟอร์เมติกส์มีหลายส่วน นับตั้งแต่การค้นหาข้อมูล การใช้ฐานข้อมูลชีวภาพที่มีให้บริการทางอินเตอร์เน็ต การวิเคราะห์ลำดับนิวคลีโอไทด์ (Nucleotide) ของดีเอ็นเอ (DNA) หรืออาร์เอ็นเอ (RNA) หรือลำดับกรดอะมิโนของเปปไทด์หรือโปรตีนซึ่งได้จากการทดลอง การพยากรณ์ลำดับนิวคลีโอไทด์หรือกรดอะมิโนนั้นว่าจะเป็นยีนอะไรมีการแสดงออกอย่างไร หรือมีบทบาทหน้าที่ในสิ่งมีชีวิตอย่างไร การศึกษาโครงสร้าง 3 มิติ ซึ่งรวมไปถึงการศึกษาถึงคุณสมบัติทางด้านกายภาพและชีวภาพของโครงสร้างต่างๆ ด้วย

นักวิทยาศาสตร์ในหลายๆประเทศให้ความสำคัญกับการทำวิจัยที่เกี่ยวข้องกับโครงการทางด้านการถอดรหัสพันธุกรรมของสิ่งมีชีวิตต่างๆ ตัวอย่างที่เห็นได้ชัด เช่น โครงการฮิวแมนจีโนม (Human Genome Project - HGP) เป็นโครงการระหว่างประเทศที่ร่วมมือกันทำงานหลายองค์กร โดยใช้เงินลงทุนจำนวนมาก ใช้นักวิทยาศาสตร์ นักชีววิทยา นักคอมพิวเตอร์ มาร่วมกันศึกษาและวิเคราะห์ โดยช่วยกันจัดทำข้อมูลสายรหัสพันธุกรรม และรวบรวมเก็บเป็นฐานข้อมูลกลางที่เรียกเข้าถึงหรือดาวน์โหลดมาปรับปรุงข้อมูลและใช้ร่วมกัน การถอดรหัสพันธุกรรมนั้น เป็นเรื่องที่สำคัญและจะให้ประโยชน์แก่มวลมนุษย์ชาติมหาศาล เพราะรหัสพันธุกรรมที่ได้จะเกี่ยวข้องกับ

การรู้ถึงการเกิดโรคต่างๆ เกี่ยวพันกับการผลิตยารักษาโรค การปรับปรุงพันธุ์ การโคลนนิ่ง ตลอดจนการใช้ประโยชน์ในเรื่องการสร้างอาหารและปัจจัยสี่อื่นๆ แต่เนื่องจากในงานถอดรหัสพันธุกรรมจะต้องเกี่ยวข้องกับการอ่านสายรหัสที่มีแต่ลำดับเบสที่แทนด้วยตัวอักษร A T G C จำนวนยาวมากๆ และการแปลความหมายของสายรหัส ซึ่งเป็นงานที่เกี่ยวข้องกับข้อมูลที่มีขนาดใหญ่มาก และต้องใช้เวลานาน จึงเลี่ยงไม่ได้ที่นักวิจัยจะต้องนำเทคนิคทางด้านการจัดการและระบบคอมพิวเตอร์ที่ใช้ในการจัดการข้อมูลเข้ามาช่วย เพื่อให้การดำเนินการต่าง ๆ เป็นไปแบบอัตโนมัติมากขึ้น และมีความน่าเชื่อถือได้สูงขึ้น

ในการวิจัยทางด้านการถอดรหัสพันธุกรรมนั้น นักวิจัยจะเริ่มจากการนำสายรหัสมาหาขอบเขตของยีนที่มีอยู่ในสายรหัสนั้นๆ แล้วจึงนำยีนที่ได้ไปหาขอบเขตของเอ็กซอน (Exon) และอินทรอน (intron) อีกทั้งจำนวนของเอ็กซอน และ อินทรอน ที่มีในขอบเขตของยีนนั้น ๆ แล้วจึงนำยีนที่ตัดอินทรอนออกแล้วไปอ่านแล้วแปลงออกมาเป็นรหัสโปรตีน ซึ่งจากสายโปรตีนที่ได้นี้นักวิจัยจะนำไปหาว่าสายโปรตีนเส้นนี้มีการแสดงออกอย่างไร หรือมีบทบาทหน้าที่อย่างไรในสิ่งมีชีวิต

สำหรับขั้นตอนที่ยุ่งยากมากสำหรับนักวิจัยในการถอดรหัสพันธุกรรม ก็คือ การค้นหาขอบเขตยีน (Gene Finding) สาเหตุที่การค้นหายีนในสายนิวคลีโอไทด์ยุ่งยากก็เนื่องมาจากลักษณะสำคัญ 2 ประการของยีน ประการแรก ก็เนื่องมาจากการที่ในสายนิวคลีโอไทด์นั้นสามารถมียีนอยู่ในสายนั้นจำนวนเท่าไรก็ได้ และไม่รู้ว่า ณ ตำแหน่งใดจะเป็นตำแหน่งเริ่มต้นของยีน และ ณ ตำแหน่งใดจะเป็นตำแหน่งสิ้นสุดของยีน และไม่รู้จำนวนที่แน่นอนของยีนในสายนิวคลีไทด์อีกด้วย ความยุ่งยากประการที่สอง เกิดมาจากลักษณะเฉพาะของยีนยูคาริโอต (Eukaryote) หรือยีนในสัตว์หรือพืชชั้นสูงนั้น ภายในยีนยังประกอบไปด้วยเอ็กซอน และอินทรอน ซึ่งก็ไม่รู้จำนวนว่ามีจำนวนเอ็กซอน และ อินทรอนจำนวนเท่าไร และไม่รู้ขนาดความยาวที่แน่นอนอีกเช่นกัน จากปัญหาความยุ่งยากที่กล่าวมาแล้วจึงทำให้เกิดการพัฒนาโปรแกรมคอมพิวเตอร์ที่จะช่วยอำนวยความสะดวกให้กับนักวิจัยในค้นหาขอบเขตของยีน และขอบเขตของเอ็กซอน และอินทรอน เนื่องจากคอมพิวเตอร์สามารถจะทำงานที่ยุ่งยากและน่าเบื่อเหล่านี้แทนมนุษย์ได้ และอาจสามารถทำได้รวดเร็วกว่ามนุษย์อีกด้วย ตัวอย่างของโปรแกรมที่ถูกพัฒนาขึ้นเพื่อใช้ในการหาขอบเขตของยีน เช่น GeneMark (Borodovsky and McIninch,1993) Glimmer (Salzberg, Pertea, Delcher, Gardner, and Tettelin, 1999) Pombe (Chan and Zhang,1998) และโปรแกรมอื่น ๆ โปรแกรมเหล่านี้พัฒนาขึ้นโดยมีวัตถุประสงค์หลักในการค้นหาขอบเขตยีนของสิ่งมีชีวิตที่แตกต่างกัน เช่น Glimmer จะใช้ค้นหายีนในสายนิวคลีโอไทด์ของสิ่งมีชีวิตจำพวกแบคทีเรีย Genie จะใช้ค้นหายีนในสายนิวคลีโอไทด์มนุษย์ หรือสัตว์มีกระดูกสันหลัง GRAIL จะ

ใช้ค้นหายีนในสายนิวคลีโอไทด์มนุษย์เท่านั้น ส่วน Pombe ถูกพัฒนาขึ้นเพื่อใช้ค้นหายีนในสายนิวคลีโอไทด์ของยีสต์โดยเฉพาะ โปรแกรมที่ช่วยค้นหาขอบเขตของยีนนี้ส่วนมากถูกพัฒนาขึ้นโดยใช้ข้อมูลนิวคลีโอไทด์ของมนุษย์และสัตว์มีกระดูกสันหลังมาเป็นข้อมูลในการที่จะให้โปรแกรมเรียนรู้ในการค้นหายีน สำหรับวิธีการหรือเทคนิคที่ใช้ในการพัฒนาโปรแกรมจะแตกต่างกันไป ตัวอย่างของเทคนิคที่ใช้กัน เช่น นิวรอลเน็ตเวิร์ก (Neural Network) มาร์คอฟโมเดล (Markov chain model) ฮิดเดนมาร์คอฟโมเดล (Hidden Markov Model-HMM) ต้นไม้ตัดสินใจ (Decision Tree) การแก้ปัญหาโดยอาศัยกฎ (Rule-Based) การวิเคราะห์จำแนกประเภทแบบเส้นตรง (Linear Discriminant Analysis) และฟังก์ชันแยกประเภทแบบยกกำลังสอง (Quadratic Discriminant Function)

จากปัญหาและความจำเป็นที่กล่าวมาแล้วจึงทำให้เกิดวิทยานิพนธ์นี้ ซึ่งมีวัตถุประสงค์เพื่อพัฒนาโปรแกรมที่ช่วยในการค้นหาขอบเขตยีนในสายนิวคลีโอไทด์ของข้าว ขอบเขตของเอ็กซอน และ อินทรอน ที่อยู่ภายในยีนข้าว โดยจะใช้ข้อมูลสายนิวคลีโอไทด์ข้าวที่มีการหายีนไว้แล้ว มาเป็นข้อมูลสอน (Training Set) ที่ใช้ในการให้โปรแกรมเรียนรู้ สำหรับสาเหตุที่ต้องการพัฒนาโปรแกรมที่เน้นหายีนเฉพาะในสายนิวคลีโอไทด์ของข้าวก็เนื่องมาจากประเทศไทยเป็นประเทศเกษตรกรรม มีการส่งออกข้าวเป็นสินค้าหลัก ดังนั้นข้าวจึงเป็นพืชเศรษฐกิจที่สำคัญมากของประเทศและข้าวก็ยังเป็นอาหารหลักของมนุษย์เกือบ 3 พันล้านคนทั่วโลก

วัตถุประสงค์

เพื่อพัฒนาโปรแกรมที่ใช้ในการค้นหายีนในสายนิวคลีโอไทด์ข้าว

ขอบเขตการวิจัย

1. ใช้ข้อมูลสายนิวคลีโอไทด์ข้าวในฐานข้อมูลทางชีวภาพที่มีให้บริการในอินเตอร์เน็ตเป็นข้อมูลสอน
2. ค้นหาขอบเขตของยีนข้าว เอ็กซอน และอินทรอนที่มีอยู่ภายในยีนข้าว
3. พัฒนาโปรแกรมที่ใช้ค้นหายีนข้าวในสายนิวคลีโอไทด์ข้าว ที่สามารถใช้บนเครื่องคอมพิวเตอร์ส่วนบุคคล

ขั้นตอนการดำเนินการวิจัย

1. ศึกษาและหาข้อมูลเกี่ยวกับเรื่องยีนและโครโมโซม
2. หาข้อมูลเกี่ยวกับยีนข้าว
3. ศึกษาวิธีการอ่านข้อมูลสายนิวคลีโอไทด์จากฐานข้อมูลที่มีในอินเตอร์เน็ตและเตรียมข้อมูลสายนิวคลีโอไทด์ข้าวที่จะใช้
4. ศึกษาวิธีการหรือทฤษฎีที่จำเป็นต้องรู้ในการพัฒนาโปรแกรม
5. กำหนดขอบเขตของการวิจัย
6. ศึกษางานวิจัยที่เกี่ยวข้อง
7. ออกแบบและพัฒนาโปรแกรม
8. ทดสอบและแก้ไขโปรแกรม
9. สรุปผลการวิจัย และ เรียบเรียงวิทยานิพนธ์

ประโยชน์ที่คาดว่าจะได้รับ

1. สามารถช่วยนักวิจัยที่ทำงานด้านยีนข้าวในการหายีนข้าวและเอ็กซอนอินทรอนในสายนิวคลีโอไทด์ของข้าว
2. ได้โปรแกรมที่ทำงานในการค้นหายีนที่เฉพาะเจาะจงกับยีนข้าวเท่านั้น
3. ได้โปรแกรมที่ผ่านการเรียนรู้ข้อมูลที่เป็นข้อมูลนิวคลีโอไทด์ของยีนข้าวโดยเฉพาะ

บทที่ 2

ความรู้พื้นฐานเกี่ยวกับพันธุศาสตร์โมเลกุลและมาร์คอฟโมเดล

พันธุศาสตร์โมเลกุล

โครโมโซม (Chromosome) คือส่วนประกอบในนิวเคลียส (nucleus) ของเซลล์ ซึ่งมียีน (Gene) เป็นจำนวนมากกระจายอยู่ โครโมโซมนั้นประกอบด้วยเส้นใยบางๆ ของสารนิวคลีโอโปรตีน หรือที่เรียกว่า เส้นใยโครมาติน โครมาตินนี้ประกอบด้วยโมเลกุลของดีเอ็นเอและโปรตีน ดีเอ็นเอเป็นกรดนิวคลีอิกชนิดหนึ่งพบมากในนิวเคลียส ต่างกับอาร์เอ็นเอซึ่งเป็นกรดนิวคลีอิกอีกชนิดที่ใกล้เคียงดีเอ็นเอมาก แต่จะพบมากในไซโตพลาสซึม

1. จีโนม (Genome)

จีโนม หมายถึง ส่วนประกอบที่เกี่ยวข้องกับการถ่ายถอดทางพันธุกรรมที่มีอยู่ในโครโมโซมทั้งหมดของสิ่งมีชีวิตชนิดใดชนิดหนึ่ง และโดยหน้าที่ยังหมายถึงชุดคำสั่งหรือวิธีการทั้งหมดที่ใช้ในการสร้างและดำเนินชีวิต เปรียบเสมือนพิมพ์เขียวสำหรับการสร้างโครงสร้างของเซลล์ทุกเซลล์ พร้อมด้วยกิจกรรมที่เซลล์แต่ละเซลล์ต้องทำตั้งแต่เกิดจนตาย

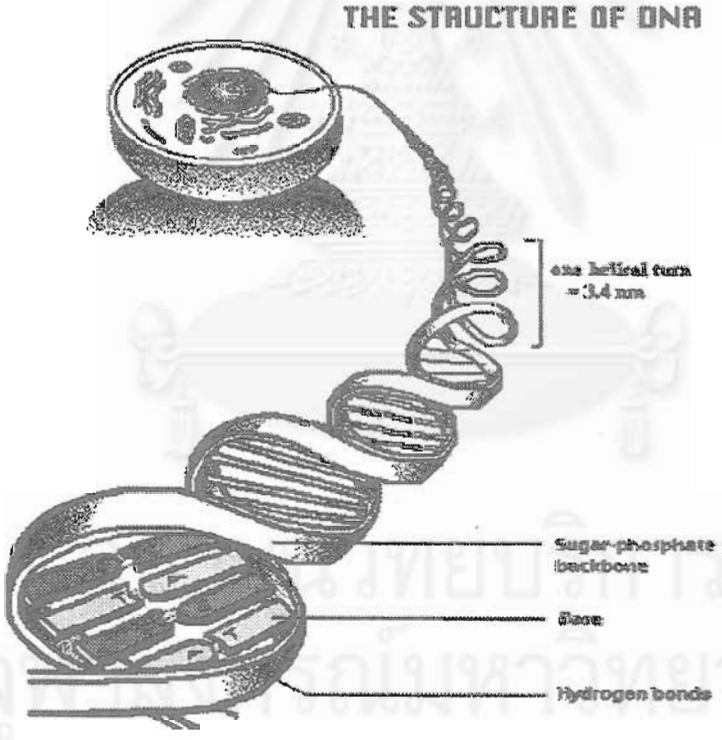
จีโนมอยู่บนดีเอ็นเอและโมเลกุลของดีเอ็นเอจับตัวอยู่กับโมเลกุลของโปรตีนในอัตราส่วนเท่าๆ กัน ประกอบกันขึ้นเป็น โครโมโซม ซึ่งโครโมโซมจะอยู่ภายในนิวเคลียสของเซลล์

ในมนุษย์นอกจากเซลล์เม็ดเลือดแดงที่ไม่มีนิวเคลียสแล้ว เซลล์อื่นๆจะมีโครโมโซมอยู่ภายในทั้งสิ้น โดยปกติมีจำนวน 46 โครโมโซม จำนวนโครโมโซมของสิ่งมีชีวิตจะเป็นจำนวนเลขคู่เสมอ เพราะได้จากการผสมพันธุ์ของพ่อและแม่อย่างละครึ่ง เราสามารถจัดโครโมโซมที่มีลักษณะเหมือนกันแต่มาจากพ่อและแม่เป็นคู่ๆได้ จึงนิยมกล่าวถึงจำนวนโครโมโซมเป็นจำนวนคู่ โครโมโซมของมนุษย์มีจำนวน 23 คู่ สิ่งมีชีวิตแต่ละชนิดมีจำนวนโครโมโซมที่แน่นอน สิ่งมีชีวิตต่างชนิดกันมักจะมีจำนวนโครโมโซมไม่เท่ากัน

2. โมเลกุลดีเอ็นเอ (DNA)

โมเลกุลของดีเอ็นเอที่อยู่บนโครโมโซมมีความยาวมาก เมื่ออยู่ในโครโมโซมจึงอยู่ในลักษณะที่ขมวดเข้าด้วยกัน แต่ละโมเลกุลของดีเอ็นเอแบ่งออกเป็นหน่วยย่อย ๆ เรียกว่า นิวคลีโอไทด์ยูนิต (nucleotide unit) ซึ่งประกอบไปด้วยหมู่ฟอสเฟต น้ำตาลดีออกซีไรโบส

(deoxyribose) และสารประกอบไนโตรเจนที่เรียกว่านิวคลีโอไทด์ (nucleotides) นิวคลีโอไทด์ที่จับตัวกันอยู่บนดีเอ็นเอนี้เรียกว่า เบส (base) เบสในดีเอ็นเอมีเพียง 4 ชนิดได้แก่ กัวนีน (guanine) ไซโตซีน (cytosine) ไทมีน (thymine) และอะดีนีน (adenine) นิยมเรียกย่อ ๆ ว่า G C T และ A ตามลำดับ ดีเอ็นเอมีลักษณะเป็นเส้นนิวคลีโอไทด์ 2 เส้นพันกันแบบเชือกฟั่น บางครั้งจึงนิยมเรียกโมเลกุลของดีเอ็นเอว่า เชือกดีเอ็นเอ (DNA strand) เส้นนิวคลีโอไทด์เกิดจากการจับตัวกันของเบสกับน้ำตาลดีออกซีไรโบส น้ำตาลดีออกซีไรโบสจับกับหมู่ฟอสเฟต หมู่ฟอสเฟตจับกับน้ำตาลดีออกซีไรโบส แล้วน้ำตาลดีออกซีไรโบสก็จับกับเบสอีก เป็นเช่นนี้เรื่อยไปแบบลูกโซ่ ส่วนการเกิดเป็นเส้นคู่นั้น เนื่องจากเบสของเส้นหนึ่งจับกับเบสของอีกเส้นหนึ่งด้วยไฮโดรเจนบอนด์ โดยที่กัวนีน (หรือ G) จับคู่กับ ไซโตซีน (หรือ C) และ อะดีนีน (หรือ A) จับคู่กับ ไทมีน (หรือ T) แต่ละคู่นี้เรียกว่าคู่เบส (base pair) เป็นคู่ที่แน่นอน ไม่มีการสับคู่ ดังแสดงในรูปที่ 2.1



รูปที่ 2.1 โครงสร้างของดีเอ็นเอ

การเรียงลำดับของคู่เบสมีความสำคัญอย่างยิ่ง เพราะเป็นรหัสที่สื่อถึงข้อมูลอันจำเป็นแห่งชีวิต ในคนมีคู่เบสประมาณ 3,000 ล้านคู่ ประกอบขึ้นเป็นจีโนมของคน ถ้าหากนำโมเลกุลดีเอ็นเอที่ขมวดอยู่ในโครโมโซมมายืดออก จะมีความยาวถึง 1.5 เมตร แต่กว้างเพียงเศษ 1 ส่วน

ล้านนิ้ว ซึ่งสายดีเอ็นเอที่เล็กจนมองไม่เห็นนี้มีรหัสบันทึกข้อมูลที่จำเป็นในการสร้างและการดำรงชีวิตสำหรับสิ่งมีชีวิตทุกชีวิตอยู่ ตั้งแต่สิ่งมีชีวิตเซลล์เดียวอย่างแบคทีเรียไปจนกระทั่งถึงสัตว์หลายเซลล์ที่ซับซ้อนอย่างมนุษย์

นอกจากที่ดีเอ็นเอจะมีการเก็บข้อมูลที่จำเป็นในการสร้างและการดำรงชีวิตแล้ว ดีเอ็นเอยังมีประโยชน์ในการใช้เป็นแม่แบบในการสร้างโมเลกุลอาร์เอ็นเออีกด้วย โดยการคลายเกลียวสายโพลีนิวคลีโอไทด์ของดีเอ็นเอออก เพื่อเป็นแม่แบบให้นิวคลีโอไทด์โมเลกุลเดี่ยวเข้ามาจับคู่กับเบสบนสายโพลีนิวคลีโอไทด์เดิม แต่ในขบวนการนี้ อะดินีน (A) จะจับคู่กับ ยูราซีล (U) และกัวนีน (G) จับคู่กับ ไซโตซีน (C) และจะไม่ใช้สายนิวคลีโอไทด์ของดีเอ็นเอทั้งสองเส้นพร้อมกัน แต่จะใช้สายดีเอ็นเอต้นแบบเพียงเส้นเดียวเป็นช่วงๆ ตามความจำเป็นของการใช้อาร์เอ็นเอชนิดนั้น ๆ และเนื่องจากโมเลกุลของอาร์เอ็นเอถูกสร้างขึ้นมาจากสายดีเอ็นเอเพียงข้างเดียว อาร์เอ็นเอจึงเป็นสายโพลีนิวคลีโอไทด์สายเดียว ซึ่งอาร์เอ็นเอนี้จะนำเอาไปใช้ประโยชน์ในการสังเคราะห์โปรตีนต่อไป

3. อาร์เอ็นเอ

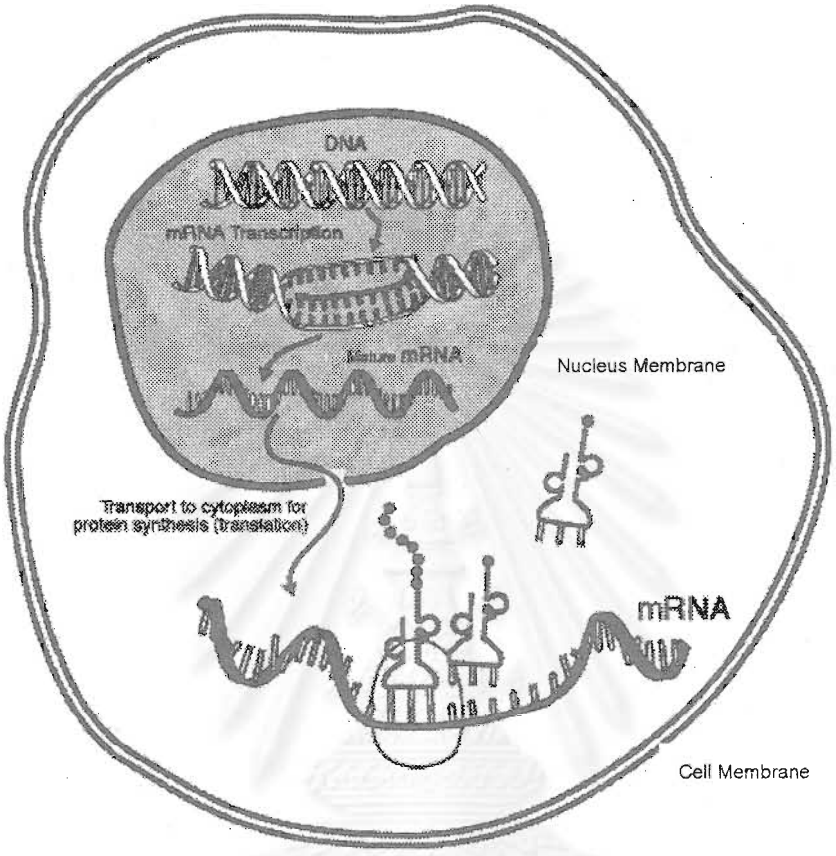
อาร์เอ็นเอ แบ่งออกได้เป็น 3 ชนิด ตามลักษณะโครงสร้างและหน้าที่การทำงาน ดังนี้ คือ tRNA rRNA และ mRNA (Rastogi and Sharma,1995)

3.1 tRNA (transfer RNA) มีขนาดเล็กมาก ถึงแม้ tRNA จะถูกสร้างขึ้นมาโดยลอกลำดับเบสจากดีเอ็นเอเช่นเดียวกับอาร์เอ็นเอชนิดอื่นๆ แต่เบสของมันมีการเปลี่ยนแปลงเพิ่มเติมมากมาย เพื่อให้มีหน้าที่เฉพาะเจาะจงในขบวนการสังเคราะห์โปรตีน กล่าวคือ ปลายด้านหนึ่งของ tRNA จะมีแอนติโคดอน (anticodon) ซึ่งเป็นเบสที่จับคู่กันได้กับเบสที่เป็นรหัส หรือโคดอน (Codon) บนสาย mRNA ส่วนปลายอีกด้านของ tRNA ทำหน้าที่พากรดอะมิโนเฉพาะสำหรับรหัสดังกล่าวเพื่อนำมาเชื่อมต่อเป็นสายโพลีเปปไทด์ ในการสังเคราะห์โปรตีนนั่นเอง

3.2 rRNA (ribosomal RNA) รวมกันเข้ากับโปรตีนประกอบเป็นไรโบโซม (ribosome)

3.3 mRNA (messenger RNA) เป็นอาร์เอ็นเอ ที่มีบทบาทสำคัญที่สุดในการกำหนดลำดับของกรดอะมิโนที่มาเรียงตัวกันเป็นโพลีเปปไทด์ ทั้งนี้เพราะ mRNA เป็นตัวรับคำสั่งจากดีเอ็นเอ โดยถอดรหัสในรูปของลำดับเบสที่เรียงตัวกันอยู่ในดีเอ็นเอ ขบวนการถ่ายทอดรหัสนี้เรียกว่า ทรานสคริปชัน (transcription) ดังจะเห็นในรูปที่ 2.2 รหัสหรือโคดอนบน mRNA นี้เองจะ

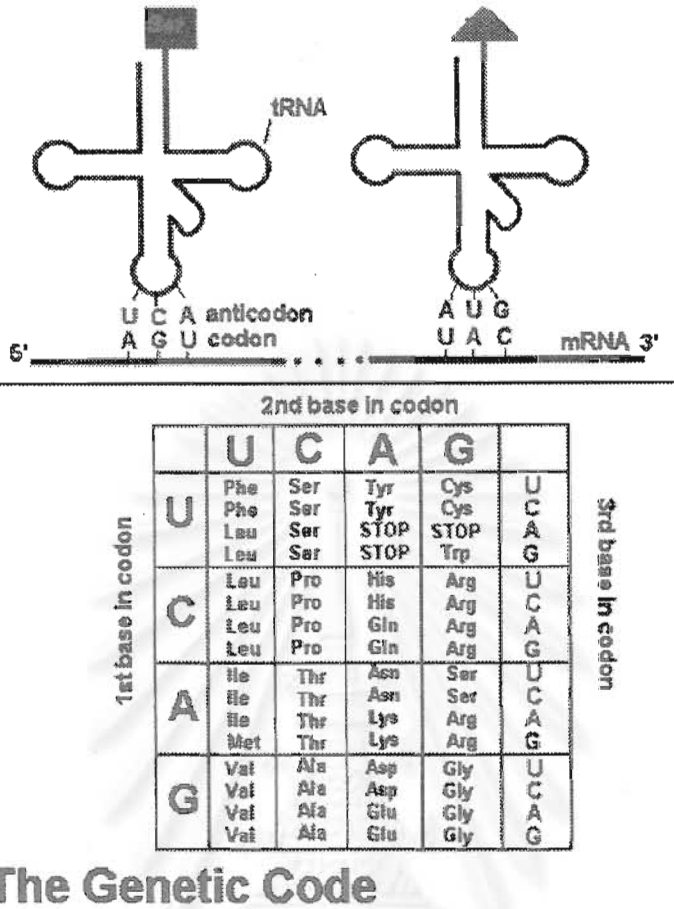
ถูกแปลงรหัส (translation) ที่ไรโบโซม เพื่อสร้างโพลีเปปไทด์ที่เฉพาะเจาะจงตามต้องการ mRNA ถูกสร้างขึ้นเมื่อเซลล์หรือสิ่งมีชีวิตนั้นๆ ต้องการสังเคราะห์โปรตีนที่จำเป็น



รูปที่ 2.2 ลักษณะของอาร์เอ็นเอ

4. รหัสทางพันธุกรรม และการสังเคราะห์โปรตีน

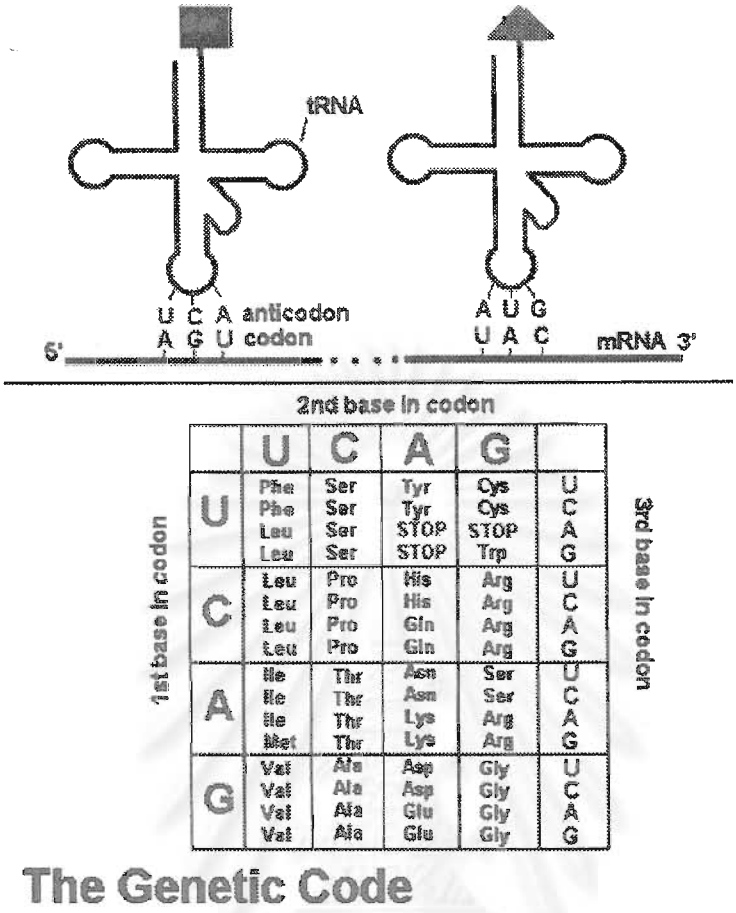
รหัสพันธุกรรมจะอยู่บนดีเอ็นเอ และถูกถ่ายทอดไปยัง mRNA เพื่อนำไปสร้างโปรตีนอีกทอดหนึ่ง เมื่อเราพิจารณาถึงโมเลกุลดีเอ็นเอและอาร์เอ็นเอ จะพบสิ่งที่เหมือนกันคือหน่วยของเบสเท่านั้นที่มีรูปร่างทั้งหมด 4 แบบ คือ A T (หรือ U ในอาร์เอ็นเอ) C และ G ซึ่งเบสเหล่านี้ก็คือ รหัสทางพันธุกรรมนั่นเอง การผสมกันของเบสทั้ง 4 แบบเข้าด้วยกันสามารถใช้เป็นภาษาสื่อความหมายในการนำกรดอะมิโน 20 ชนิดมาเรียงตัวกัน ซึ่งสามารถทำได้วิธีเดียว คือ ต้องใช้เบสอย่างน้อย 3 ตัวผสมกัน หรือที่เรียกว่า รหัสสามตัว หรือ โคดอน ดังรูปที่ 2.3



รูปที่ 2.3 รหัสทางพันธุกรรมหรือโคดอน

จะเห็นได้ว่ากรดอะมิโนส่วนใหญ่มีรหัสได้มากกว่า 1 แบบ เช่น Ser สามารถมีรหัสได้เป็น UCU หรือ UCC หรือ UCA หรือ UCG หรือ AGU หรือ AGC และพบว่าบางรหัสไม่ได้ใช้เป็นตัวกำหนดกรดอะมิโน แต่เป็นตัวบอกจุดเริ่มต้น หรือจุดสุดท้ายของการสร้างโพลีเปปไทด์

เนื่องจากดีเอ็นเอเป็นตัวควบคุมกำหนดการสร้างโปรตีน ขบวนการสังเคราะห์โปรตีนจึงต้องเริ่มจากการรับคำสั่งจากดีเอ็นเอ ซึ่งกระทำโดยการถ่ายทอดลำดับเบสจากดีเอ็นเอมายัง mRNA หรือการทำทรานสคริปชันนั่นเอง ขบวนการนี้เกิดขึ้นในนิวเคลียส จากนั้น mRNA ก็จะเดินทางออกจากนิวเคลียสเข้ามาสู่ไซโตพลาสซึม โดยมารวมเข้ากับไรโบโซมซึ่งจะใช้เป็นสถานที่ในการสร้างโปรตีน ในขณะเดียวกัน tRNA ก็ทำหน้าที่นำพาเอากรดอะมิโนเข้ามาในบริเวณไรโบโซม จากนั้นการแปลรหัสหรือทรานสเลชันก็เกิดขึ้น กล่าวคือ กรดอะมิโนก็จะถูกนำมาเรียงตัวตามลำดับของรหัสบนmRNAโดยมีเอนไซม์ พลังงาน และแฟคเตอร์ต่างๆ มาร่วมในทุกขั้นตอน สายของกรดอะมิโนดังกล่าวจะเชื่อมต่อโยงเป็นสายโพลีเปปไทด์ เมื่อการแปลรหัสในmRNAหมด



รูปที่ 2.3 รหัสทางพันธุกรรมหรือโคดอน

จะเห็นได้ว่ากรดอะมิโนส่วนใหญ่มีรหัสได้มากกว่า 1 แบบ เช่น Ser สามารถมีรหัสได้เป็น UCU หรือ UCC หรือ UCA หรือ UCG หรือ AGU หรือ AGC และพบว่าบางรหัสไม่ได้ใช้เป็นตัวกำหนดกรดอะมิโน แต่เป็นตัวบอกจุดเริ่มต้น หรือจุดสุดท้ายของการสร้างโพลีเปปไทด์

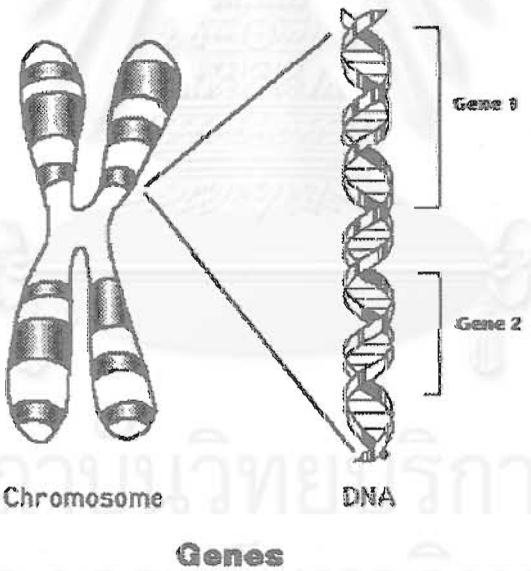
เนื่องจากดีเอ็นเอเป็นตัวควบคุมกำหนดการสร้างโปรตีน ขบวนการสังเคราะห์โปรตีนจึงต้องเริ่มจากการรับคำสั่งจากดีเอ็นเอ ซึ่งกระทำโดยการถ่ายทอดลำดับเบสจากดีเอ็นเอมายัง mRNA หรือการทำทรานสคริปชันนั่นเอง ขบวนการนี้เกิดขึ้นในนิวเคลียส จากนั้น mRNA ก็จะเดินทางออกจากนิวเคลียสเข้ามาสู่ไซโตพลาสซึม โดยมารวมเข้ากับไรโบโซมซึ่งจะใช้เป็นสถานที่ในการสร้างโปรตีน ในขณะเดียวกัน tRNA ก็ทำหน้าที่นำพาเอากรดอะมิโนเข้ามาในบริเวณไรโบโซม จากนั้นการแปลรหัสหรือทรานสเลชันก็เกิดขึ้น กล่าวคือ กรดอะมิโนก็จะถูกนำมาเรียงตัวตามลำดับของรหัสบนmRNAโดยมีเอนไซม์ พลังงาน และแฟคเตอร์ต่างๆ มาร่วมในทุกขั้นตอน สายของกรดอะมิโนดังกล่าวจะเชื่อมต่อโยงเป็นสายโพลีเปปไทด์ เมื่อการแปลรหัสในmRNAหมด

สิ้นลงทั้งสายแล้ว สายโพลีเปปไทด์ที่สร้างขึ้นก็จะถูกปล่อยออกมาจากไรโบโซมกลายเป็นโมเลกุลของโปรตีนตามที่ต้องการ

สิ่งมีชีวิตทั้งหมดมีโปรตีนเป็นองค์ประกอบส่วนใหญ่ และมีบทบาทสำคัญยิ่งเนื่องจากเป็นตัวกำหนดหน้าที่ของเซลล์ เราทราบว่าเซลล์แต่ละส่วนในร่างกายไม่เหมือนกัน เช่น เซลล์ฝ่ามือ เซลล์หนังศีรษะ เซลล์ตับ มีความแตกต่างกันออกไปทั้งหน้าตาและหน้าที่ สิ่งที่ทำหน้าที่กำหนดความแตกต่างเหล่านี้ก็คือโปรตีนที่อยู่ในเซลล์นั่นเอง

5. ยีน (Gene)

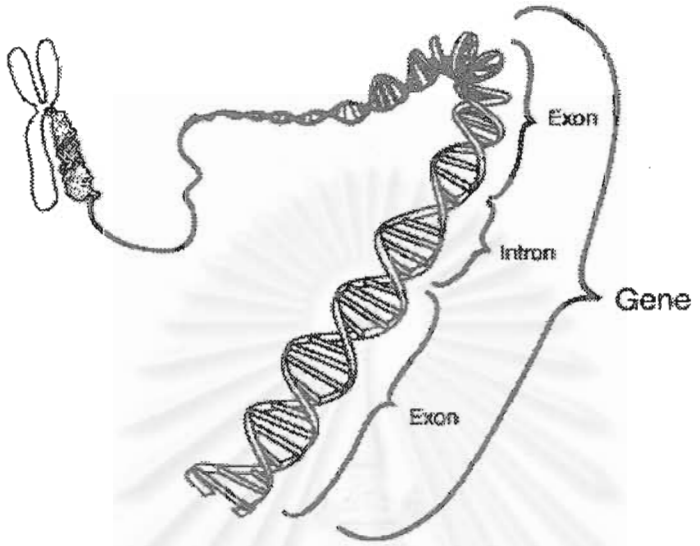
ยีน หมายถึง หน่วยถ่ายทอดพันธุกรรมหน่วยหนึ่ง ซึ่งที่จริงก็คือ ชุดของคู่เบสหลายคู่ที่มาเรียงลำดับกันนั่นเอง แต่ถูกจัดเป็นชุดหรือเป็นยีนหนึ่งๆ เนื่องจากการเรียงลำดับคู่เบสในส่วนนั้นประกอบขึ้นเป็นรหัสที่สื่อถึงข้อมูลได้ประการหนึ่ง



รูปที่ 2.4 ยีน

รูปที่ 2.4 แสดงให้เห็นว่าในโครโมโซมหนึ่ง ๆ จะประกอบด้วยยีนเรียงต่อกันไปโดยจะมีช่วงที่เป็นลำดับเบสที่ไม่ได้เป็นยีนคั่นระหว่างยีนด้วย ยีนแต่ละยีนมีขนาดไม่เท่ากัน ถึงแม้จะอยู่ในจีโนมเดียวกัน ดังนั้นยีนบางยีนอาจมีขนาดใหญ่เป็นหลายพันคู่เบส แต่บางยีนอาจจะมีขนาดเล็กกว่ามาก และในจีโนมหนึ่ง ๆ จะมีคู่เบสที่เป็นรหัสข้อมูลสำหรับสร้างโปรตีนเพียงแค่ร้อยละ 10

เท่านั้น และลำดับคู่เบสเหล่านี้ที่สามารถเป็นรหัสใช้สร้างโปรตีนมีชื่อเรียกว่า เอ็กซอน (exon หรือ exon sequence) ดังแสดงในรูปที่ 2.5



รูปที่ 2.5 เอ็กซอน

ส่วนคู่เบสที่เหลืออีกร้อยละ 90 เป็นคู่เบสที่ไม่เป็นรหัสข้อมูลและอยู่กระจัดกระจายทั่วไปในยีน ซึ่งจะไม่ถูกแปลงเป็นโปรตีน เรียกว่า อินทรอน (intron หรือ intron sequence) คาดว่าดุลยภาพหรือสัดส่วนที่ลงตัวของจีโนมนั้นต้องอาศัยคู่เบสอื่นๆ ที่ไม่แสดงว่าเป็นรหัสเลยด้วย แต่คู่เบสเหล่านี้ทำหน้าที่อะไร ขณะนี้ยังไม่ทราบแต่สันนิษฐานกันว่าอาจจะทำหน้าที่ควบคุมลำดับคู่เบส หรือเป็นตัวกั้นระหว่างส่วนต่างๆ ในยีนก็เป็นได้

6. กรรมวิธีการหารหัสพันธุกรรม

วิธีการหารหัสพันธุกรรมประกอบด้วย 3 ขั้นตอนหลัก คือ

6.1 หาการเรียงตัวของเบสบนเส้นดีเอ็นเอ (sequencing) ทำได้โดยตัดเส้นดีเอ็นเอ เป็นชิ้น ๆ หลังจากนั้นนำชิ้นดีเอ็นเอที่ตัดแล้วเข้าเครื่องอ่านอัตโนมัติ เพื่อหาการเรียงตัวของเบสในแต่ละชิ้นของดีเอ็นเอ

6.2 ประกอบเข้าด้วยกันใหม่ (assembly) นำชิ้นของดีเอ็นเอที่ได้รับการอ่านจากเครื่องแล้วนำมาเรียงกันใหม่ โดยนำชิ้นดีเอ็นเอแต่ละชิ้นที่มีมาร์คเกอร์ (marker) เหมือนกันมาซ้อนกัน (overlap) จนกระทั่งได้ลำดับการเรียงตัวของเบสบนเส้นดีเอ็นเอทั้งหมด

6.3 ระบุตำแหน่งของยีน (annotation) เมื่อทราบการเรียงตัวของเบสบนเส้นดีเอ็นเอทั้งหมดแล้ว จึงค้นหาตำแหน่งของยีนทั้งหมด ซึ่งมีอยู่เพียงประมาณ 3 % ในข้อมูลทั้งหมดบนเส้นดีเอ็นเอเส้นนั้น ทั้งนี้อาจทำโดยการค้นหารหัสบ่งชี้การเริ่มต้นและสิ้นสุดของยีน หรือโดยการเปรียบเทียบกับยีนที่รู้จักแล้ว

จีโนมข้าว (Rice Genomic)

จีโนมข้าวมีขนาดประมาณ 400 เมกกะเบส (400 ล้านคู่เบส) ซึ่งเป็นค่าประมาณที่หามาจากวิธีการทางพันธุศาสตร์ ขนาดที่แน่นอนของจีโนมข้าวจะทราบได้เมื่อทำการลำดับเบสจีโนมข้าวเสร็จเรียบร้อยแล้ว โดย ณ ปัจจุบันยังไม่เสร็จ แต่บริษัท Syngenta (USA) ประกาศว่าได้ทำการลำดับเบสของจีโนมข้าวเสร็จเรียบร้อยแล้วแต่ไม่เปิดเผยข้อมูลสู่สาธารณะ

ข้าวเป็นพืชใบเลี้ยงเดี่ยวมีชื่อทางจีนัส (genus) ว่า Oryza (เขียนแทนด้วย O) ซึ่งจากจีนัสยังแบ่งเป็นอีก 2 สปีชีย์ (Species) ได้แก่ sativa และ glaberrima ข้าวในสปีชีย์ O.sativa เป็นพันธุ์ข้าวที่เพาะปลูกกันในเกือบทั่วโลก แต่ข้าวในสปีชีย์ O.glaberrima จะเพาะปลูกกันเฉพาะในประเทศแถบอาฟริกา ดังนั้นพันธุ์ข้าวเจ้าที่พบเห็นเพาะปลูกกันโดยทั่วไปจะอยู่ในสปีชีย์ O.sativa ซึ่งในข้าวเจ้าสปีชีย์ O.sativa ก็ยังมีอีกหลายสายพันธุ์ในทางวิทยาศาสตร์จะเรียกว่าซับสปีชีย์(Subspecies) ซึ่งมี 2 ซับสปีชีย์ที่สำคัญ คือ indica (มีอีกชื่อหนึ่งว่า kasalath) และ japonica (มีอีกชื่อหนึ่งว่า nipponbare) ซึ่งจะแบ่งคร่าวๆตามพื้นที่ที่เพาะปลูก โดย japonica จะเพาะปลูกกันในแถบเอเชียตะวันออก เช่นญี่ปุ่น และจีนตอนเหนือ เป็นต้น ซึ่งสายพันธุ์ japonica นี้เป็นสายพันธุ์ที่ใช้ในการทำโครงการจีโนมข้าว โดยทุกประเทศที่เข้าร่วมโครงการนี้จะใช้ข้าวสายพันธุ์ japonica เหมือนกัน ซึ่งประเทศต่าง ๆ ที่เข้าร่วมโครงการจีโนมข้าวจะแบ่งโครโมโซมเพื่อไปวิจัย โดยในข้าวนั้นจะมีโครโมโซมทั้งสิ้น 12 คู่ โดยโครโมโซมคู่ที่ 1 จะถูกวิจัยโดยญี่ปุ่นและเกาหลี โครโมโซมคู่ที่2 โดยประเทศแถบยุโรป โครโมโซมคู่ที่ 3 และ 10 โดยอเมริกา คู่ที่ 4 โดยประเทศจีน คู่ที่ 5 โดยไต้หวัน โครโมโซมคู่ที่ 6 7 และ 8 โดยญี่ปุ่น โครโมโซมคู่ที่ 9 โดยประเทศไทย โครโมโซมคู่ที่ 11 โดยอเมริกา แคนาดา และ อินเดีย และสุดท้ายโครโมโซมคู่ที่ 12 โดยประเทศฝรั่งเศสและบราซิล และผลการวิจัยที่ได้จากประเทศต่างๆเหล่านี้จะถูกส่งไปรวบรวมเก็บไว้ที่ฐานข้อมูลนิวคลีโอไทด์ของข้าวที่อยู่ในอินเตอร์เน็ต เช่น NCBI (National Center for

Biotechnology Information) EBI (European Bioinformatics Institute) DDBJ (Genome Net-Japan)

เนื่องจากข้าวจัดเป็นสิ่งมีชีวิตประเภทยูคาริโอต (Eukaryote) ดังนั้นในยีนของข้าวจึงจะประกอบไปด้วย เอ็กซอน และอินทรอน ในขณะที่ยีนของสิ่งมีชีวิตประเภทโปรคาริโอต (Prokaryote) เช่น แบคทีเรีย จะไม่มีอินทรอนในยีนทำให้สิ่งมีชีวิตประเภทยูคาริโอตมีขั้นตอนการหายีนที่ยากกว่าและซับซ้อนกว่าโปรคาริโอต

ทฤษฎีของเบย์ (Bays' Theorem)

สมมติว่า A_1 , A_2,...,A_n เป็นเหตุการณ์ในเซมเปิลสเปซ (Sample Space) S เดียวกัน และเป็นส่วนประกอบภายใน S โดยที่ $A_1 \cup A_2 \cup \ldots \cup A_n = S$ และ $A_i \cap A_j \neq \varnothing$ เมื่อ $i \neq j$ และ $i , j = 1,2, \ldots ,n$

ถ้า B เป็นเหตุการณ์อีกเหตุการณ์หนึ่งใน S เหตุการณ์ B จะหาได้จากยูเนียนของเหตุการณ์ต่างที่ไม่เกิดขึ้นร่วมกันดังนี้

$$B = (A_1 \cap B) \cup (A_2 \cap B) \cup \ldots \cup (A_n \cap B)$$

โดย $(A_i \cap B) \cap (A_j \cap B) = \varnothing$ เมื่อ $i \neq j$

โดยกฎความน่าจะเป็น จะได้

$$\begin{aligned} P(B) &= (A_1 \cap B) + (A_2 \cap B) + \ldots + (A_n \cap B) \\ &= \sum P(A_i \cap B) \\ &= \sum P(A_i) \; P(B / A_i) \end{aligned}$$

จากนิยามความน่าจะเป็นแบบมีเงื่อนไข ถ้าหา $P(A_k / B)$ จะได้ว่า

$$P(A_k / B) = \frac{P(A_k \cap B)}{P(B)} \qquad \text{เมื่อ } k = 1,2,\ldots,n$$

$$= \frac{P(A_k) \; P(B / A_k)}{P(B)}$$

แทนค่า P(B) ด้วย $\sum P(A_i)\ P(B/A_i)$ จะได้

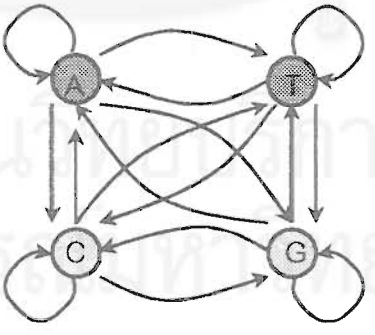
$$P(A_k/B) = \frac{P(A_k)\ P(B/A_k)}{\sum P(A_i)\ P(B/A_i)}$$

ดังนั้น ถ้าให้เงื่อนไขของเหตุการณ์ B แล้วจะหาความน่าจะเป็นของเหตุการณ์ A_k ได้ ซึ่งก็คือ Bays' Theorem

มาร์คอฟโมเดล (Markov Model)

มาร์คอฟโมเดล หมายถึง แบบจำลองของลำดับของเหตุการณ์ที่เกิดขึ้น โดยที่โอกาสของการเกิดเหตุการณ์ต่างๆ ณ จุดใดๆ ในลำดับ ขึ้นกับเหตุการณ์ที่เกิดขึ้นก่อนหน้านั้นเท่านั้น

การวิเคราะห์ลำดับเบสหรือกรดอะมิโนก็สามารถทำได้โดยการสมมติว่าลำดับเบสจะเป็น A T C หรือ G ขึ้นอยู่กับว่าเบสที่อยู่ด้านซ้ายเป็นอะไรเท่านั้น หรือ โอกาสที่จะพบกรดอะมิโนอะไร ณ ตำแหน่งใดตำแหน่งหนึ่งขึ้นอยู่กับว่ากรดอะมิโนที่อยู่ด้านซ้ายเป็นอะไร ตัวอย่างของมาร์คอฟโมเดล สำหรับสายดีเอ็นเอ (Durbin, Eddy, Krogh, and Mitchison, 1998) สามารถเขียนได้ดังรูปที่ 2.6



รูปที่ 2.6 มาร์คอฟโมเดลของสายดีเอ็นเอ

จะเห็นว่ามีสถานะที่ใช้สื่อแทนตัวอักษร 4 ตัว คือ A C G และ T ซึ่งเป็นตัวอักษรแทนเบสในดีเอ็นเอ ตัวแปรความน่าจะเป็นของ A จะเกี่ยวเนื่องกับลูกศรแต่ละอันในรูป ซึ่งการจะ

พิจารณาค่าความน่าจะเป็นจะคิดเทียบกับค่าอื่นที่เป็นค่าที่เกิดขึ้นก่อนจะเกิด A ค่าความน่าจะเป็นนี้จะเรียกว่าค่าความน่าจะเป็นในการเปลี่ยนสถานะ (Transition probabilities) a_{st} ซึ่งสามารถเขียนได้ดังนี้

$$a_{st} = P(x_i = t \mid x_{i-1} = s, x_{i-2} = k, \ldots) = P(x_i = t \mid x_{i-1} = s)$$

สำหรับทุก ๆ โมเดลความน่าจะเป็นของซีเควน จะได้ความน่าจะเป็นของซีเควนเท่ากับ

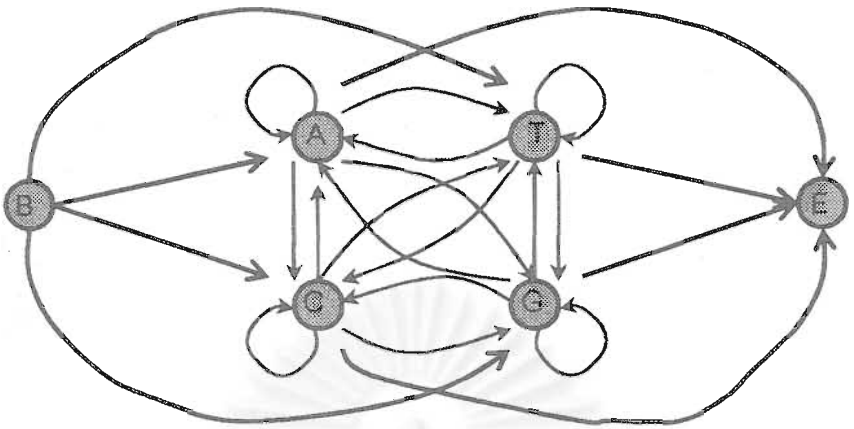
$$\begin{aligned} P(x) &= P(x_L, x_{L-1}, \ldots, x_1) \\ &= P(x_L \mid x_{L-1}, \ldots, x_1) P(x_{L-1} \mid x_{L-2}, \ldots, x_1) \ldots P(x_1) \end{aligned} \qquad (2.1)$$

แต่จากคุณสมบัติของมาร์คอฟโมเดล คือ ค่าความน่าจะเป็นของสัญลักษณ์ X_i จะขึ้นเฉพาะกับค่าที่มาก่อนมัน ซึ่งก็คือ X_{i-1} เช่น $P(x_i, x_{i-1}, \ldots, x_1) = P(x_i \mid x_{i-1}) = a_{x_{i-1} x_i}$ ดังนั้นจากสมการที่ (2.1) จะได้ว่า

$$P(x) = P(x_L \mid x_{L-1}) P(x_{L-1} \mid x_{L-2}) \ldots P(x_2 \mid x_1) P(x_1)$$

$$= P(x_1) \prod_{i=2}^{L} a_{x_{i-1} x_i} \qquad (2.2)$$

เพื่อหลีกเลี่ยงความไม่สอดคล้องกันของคุณสมบัติซึ่งอาจเกิดในสมการ (2.2) ดังนั้นจึงควรเพิ่ม state เริ่มต้น (B) และ สุดท้าย (E) ลงไปด้วย โดยกำหนด x_0 = B เป็นจุดเริ่มต้นซึ่งต้องเพิ่มในสมการ (2.2) ดังนั้นค่าความน่าจะเป็นของตัวอักษรแรกในซีเควนจะเป็น $P(x_1 = S) = a_{BS}$ และค่าความน่าจะเป็นของตัวอักษรสุดท้ายจะเป็น $P(E \mid x_L = t) = a_{tE}$ เพื่อให้สอดคล้องกับสัญลักษณ์ใหม่ที่เพิ่มเข้าไป จะได้โมเดลใหม่ดังรูปที่ 2.7



รูปที่ 2.7 มาร์คอฟโมเดล ของสายดีเอ็นเอหลังจากเพิ่มจุดเริ่มต้นและจุดสุดท้าย

สำหรับตัวอย่างต่อไปนี้แสดงการนำเอามาร์คอฟโมเดลมาใช้แยกแยะว่าบริเวณใดเป็น CpG Island (บริเวณที่พบเบส C และ G จำนวนมากก่อนที่จะพบจุดเริ่มต้นของยีน) โดยการนำข้อมูลเบสที่อยู่ในช่วง CpG Island มาสร้างเป็นโมเดลบวก และ ส่วนเบสที่ไม่ได้อยู่ในบริเวณดังกล่าวมาสร้างเป็นโมเดลลบ

ค่าความน่าจะเป็นในการเปลี่ยนสถานะ (Transition Probabilities) สำหรับ โมเดลบวก หรือบริเวณที่เป็น CpG Island หาได้ดังนี้

$$a^{+}_{st} = \frac{C^{+}_{st}}{\sum_{t'} C^{+}_{st'}}$$

โดย C^{+}_{st} คือ จำนวนครั้งที่ t ตามหลัง s ในบริเวณ CpG Island ซึ่งค่านี้จะนำไปเป็นตัวประมาณ Maximum Likelihood สำหรับค่าความน่าจะเป็นในการเปลี่ยนสถานะ ส่วน a^{-}_{st} หรือค่าความน่าจะเป็นในการเปลี่ยนสถานะสำหรับโมเดลลบ ซึ่งคิดในทำนองเดียวกัน

ผลที่ได้จากการคำนวณค่าความน่าจะเป็นในการเปลี่ยนสถานะ จะได้ดังตารางที่ 2.1 และตารางที่ 2.2

ตารางที่ 2.1 ค่าความน่าจะเป็นในการเปลี่ยนสถานะของโมเดลบวก

+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

ตารางที่ 2.2 ค่าความน่าจะเป็นในการเปลี่ยนสถานะของโมเดลลบ

-	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

โดยที่แต่ละแถวแสดงค่าความถี่ เช่น ในแถวแรกสุดของตารางบน เป็นกรณีที่ A ถูกตามด้วยเบสทั้ง 4 เบส โดย A ที่ตามด้วย A จะมีความถี่ 0.180 A ตามด้วย C มีความถี่ 0.274 A ตามด้วย G มีความถี่ 0.426 A ตามด้วย T มีความถี่ 0.120

ในการที่จะใช้โมเดลนี้ในการแยกแยะหรือจัดประเภท เราจะใช้ Log-odds ratio ซึ่งนิยามดังนี้

$$S(x) = \log\left[\frac{P(x \mid Model +)}{P(x \mid Model -)}\right] = \sum_{i=1}^{L} \log\left[\frac{a^{+}_{x_{i-1}\,x_i}}{a^{-}_{x_{i-1}\,x_i}}\right]$$

$$= \sum_{i=1}^{L} \beta x_{i-1}\, x_i$$

โดย x คือ ซีเควน และ $\beta_{x_{i-1} x_i}$ คือ ค่าลอการิทึมที่ใช้ในการแยกแยะหรือจัดประเภทข้อมูลที่ได้มาจากค่าความน่าจะเป็นในการเปลี่ยนสถานะของโมเดลบวกและโมเดลลบ ซึ่งผลของค่า β แสดงดังตารางที่ 2.3 ข้างล่าง

ตารางที่ 2.3 ค่าลอการิทึมที่ใช้ในการแยกแยะหรือจัดประเภทข้อมูล

β	A	C	G	T
A	-0.740	0.419	0.580	-0.803
C	-0.913	0.302	1.812	-0.685
G	-0.624	0.461	0.331	-0.730
T	-1.169	0.537	0.939	-0.679

งานวิจัยที่เกี่ยวข้อง

1. Pombe: A Gene Finding and Exon-Intron Structure Prediction System for Fission Yeast โดย Ting Chen และ Michael Q.Zhang (1998)

งานวิจัยนี้เกี่ยวข้องกับการพัฒนาโปรแกรมที่ชื่อว่า Pombe ซึ่งเป็นโปรแกรมที่ใช้หายีนที่มีอยู่ภายในสายนิวคลีโอไทด์ของฟิชชันยีสต์ (Fission Yeast) โดยเฉพาะ โปรแกรมนี้พัฒนาขึ้นโดยใช้วิธีการวิเคราะห์จำแนกประเภทแบบเส้นตรงในการพิจารณาเพื่อหาขอบเขตยีน ขอบเขตและจำนวนของเอ็กซอน และ อินทรอนที่มีอยู่ภายในสายนิวคลีโอไทด์ของฟิชชันยีสต์ ส่วนข้อมูลสายนิวคลีโอไทด์ที่งานวิจัยนี้ใช้จะเป็นข้อมูลจาก GenBank จำนวน 131 สายนิวคลีโอไทด์

2. Modeling DNA Splice Regions by Learning Bayesian Networks โดย Denver Dash และ Vanathi Gopalakrishnan (2001)

งานวิจัยนี้เกี่ยวข้องกับการนำ Bayesian Networks มาใช้เรียนรู้ข้อมูลสายนิวคลีโอไทด์เพื่อแยกหา จุดเริ่มต้นของการทำทรานสคริปชัน (Transcription) ในสายนิวคลีโอไทด์ และ หาดอร์เนอร์ (เบส GT) และแอคเซพเตอร์ (เบส AG) ของอินทรอนภายในแมสเซนเจอร์อาร์เอ็นเอ (mRNA) ของยีนในสิ่งมีชีวิตประเภทยูคาริโอต

3. GeneID in Drosophila โดย Genis Parra, Enrique Blanco and Roderic Guigo (2000)

งานวิจัยนี้เกี่ยวกับการนำวิธีแก้ปัญหาโดยอาศัยกฎ (Hierarchical rule-based system) มาใช้ในการค้นหายีนในสิ่งมีชีวิตประเภทแมลงวันผลไม้ (Drosphila melanogaster) โดยใช้ข้อมูลสอนที่เป็นสายนิวคลีโอไทด์ของแมลงวันผลไม้จำนวน 416 สายนิวคลีโอไทด์ซึ่งเป็นชุดข้อมูลจาก Martin Reese และคณะ (Reese, 2000) และทดสอบด้วยข้อมูลทดสอบขนาด 5,689,206 คู่เบส ที่มียีนจำนวน 416 ยีนอยู่ภายในสายนิวคลีโอไทด์นั้น

4. Interpolated Markov models for eukaryotic gene finding โดย Salzberg SL, Pertea M, Delcher AL, Gardner MJ, Tettelin H (1999)

งานวิจัยนี้เกี่ยวกับการพัฒนาโปรแกรมค้นหายีนที่ชื่อ Glimmer ที่พัฒนาขึ้นเพื่อสิ่งมีชีวิตประเภทแบคทีเรียและยูคาริโอตที่มีขนาดเล็ก เช่น Abiradopsis thaliana เป็นต้น ข้อมูลที่ใช้เป็นข้อมูลสอนมาจาก GenBank เป็นข้อมูลของสิ่งมีชีวิต Plasmodium falciparum จำนวน 108 ข้อมูลจากทั้งหมด 14 โครโมโซม และใช้มาร์คอฟโมเดลลำดับที่ 2 ในการหาบริเวณที่เกี่ยวข้องกับการควบคุมการแสดงออกของยีน (Search by signal) สำหรับการทดสอบโปรแกรมจะใช้ข้อมูลทดสอบซึ่งเป็นยีนจำนวน 209 ยีนจากโครโมโซมที่ 2 ของ P. falciparum

5 GeneMark Parallel Gene recognition for Both DNA strands โดย Mark Borodovsky and James McIninch (1993)

งานวิจัยนี้เกี่ยวกับการนำมาร์คอฟโมเดลลำดับที่ 5 มาใช้ในการค้นหายีนในสิ่งมีชีวิตประเภทแบคทีเรียจำนวน 10 ชนิด หนู และมนุษย์

6. A Plasmodium falciparum Genefinder โดย Anthony Ian Wirth (2000)

งานวิจัยนี้เกี่ยวกับการพัฒนาโปรแกรมเพื่อค้นหายีนในสิ่งมีชีวิตสปีชีย์ Plasmodium falciparum ซึ่งเป็นสิ่งมีชีวิตประเภทเชื้อโรคที่ทำให้เกิดโรคไข้มาลาเรีย ซึ่งภายในยีนของสิ่งมีชีวิตชนิดนี้ส่วนใหญ่จะไม่มีอินทรอนแทรกอยู่ รวมทั้งทดสอบโปรแกรม และนำผลการทดสอบโปรแกรมไปเปรียบเทียบกับโปรแกรมที่ใช้สายนิวคลีโอไทด์ของมนุษย์เป็นข้อมูลสอน

บทที่ 3

ขั้นตอนวิธีในการดำเนินการวิจัย

โดยส่วนมากงานที่เกี่ยวข้องกับการหายีนมักจะเน้นออกแบบเพื่อวัตถุประสงค์หลักในการค้นหายีนในสายนิวคลีโอไทด์ของมนุษย์ จึงทำให้ซอฟท์แวร์เหล่านี้มีประสิทธิภาพน้อยในการค้นหายีนในสิ่งมีชีวิตที่มีสปีชีย์ต่างจากมนุษย์ จากงานวิจัยของ Reese และคณะ (Reese, 2000) ได้ทดลองโดยนำยีน *Drosophila melanogaster* ซึ่งเป็นยีนของสัตว์ประเภทแมลงวันผลไม้ชนิดหนึ่งมาใช้กับโปรแกรมค้นหายีนที่ถูกสอนโดยข้อมูลมนุษย์ จากผลการทดลองแสดงถึงประสิทธิภาพของการนำโปรแกรมหาค้นหายีนที่ใช้ข้อมูลมนุษย์เป็นข้อมูลสอน ซึ่งมีประสิทธิภาพไม่ดีเท่ากับโปรแกรมที่ใช้ข้อมูลของแมลงวันผลไม้เป็นข้อมูลสอน ทั้งนี้จากงานวิจัยนั้นก็ได้สรุปว่าโปรแกรมค้นหายีนจะทำงานได้ดีกว่า ถ้ามีการเตรียมข้อมูลสอนที่มีความเฉพาะเจาะจงกับสิ่งมีชีวิตชนิดนั้นๆโดยเฉพาะ เพราะแม้กระทั่งสิ่งมีชีวิตในกลุ่มเดียวกัน เช่น พืช ก็ยังมีลักษณะเฉพาะที่ต่างกัน เช่น ความแตกต่างของปริมาณเบส CG (CG content หรือ Isoshore) และความแตกต่างของพืชกลุ่มใบเลี้ยงเดี่ยว (Monocot plants) และใบเลี้ยงคู่ (Dicot plants) เป็นต้น ดังนั้นจึงทำให้เกิดแนวความคิดในการพัฒนาโปรแกรมค้นหายีนที่เฉพาะเจาะจงกับยีนข้าว

การเก็บข้อมูลสอนและข้อมูลทดสอบ

เริ่มต้นจากการเก็บรวบรวมข้อมูลสายนิวคลีโอไทด์ของข้าวจากฐานข้อมูลขององค์กร National Center for Biotechnology Information (NCBI) หรือ GenBank ที่เว็บไซต์ www.ncbi.nlm.nih.gov โดยใช้เฉพาะข้อมูลข้าวที่ได้ทำการค้นหายีนไว้เรียบร้อย (annotated) แล้วเพื่อนำมาใช้เป็นข้อมูลสอน และทดสอบโปรแกรม การเก็บข้อมูลสายนิวคลีโอไทด์จะเก็บในรูปแบบ FASTA Format ซึ่งเป็นรูปแบบของการเก็บข้อมูลสายดีเอ็นเอที่นิยมกันเนื่องจากอ่านเข้าใจง่าย หลังจากได้ข้อมูลมาแล้วจะต้องมีการทำความสะอาดข้อมูลเนื่องจากเป็นข้อมูลที่เก็บไว้ในฐานข้อมูลทางอินเตอร์เน็ตอาจมีข้อผิดพลาดได้ จึงต้องมีการตรวจสอบดังนี้ คือ เริ่มต้นจากตรวจสอบว่ายีนแต่ละยีนเริ่มต้นด้วยจุดเริ่มต้นของยีน (Start Codons) หรือไม่ และตรวจสอบว่าในแต่ละยีนต้องมีอย่างน้อย 1 เอ็กซอน และขนาดของยีนต้องหารด้วย 3 ลงตัว หลังจากนั้นดูว่าอินทรอนแต่ละ อินทรอนขึ้นต้นด้วย GT หรือ ดอร์เนอร์ (donor site) และจบด้วย AG หรือ แอคเซพเตอร์ (acceptor site) ซึ่งดอร์เนอร์และแอคเซพเตอร์นี้เป็นลักษณะเฉพาะของยีนข้าวและยีนสิ่งมีชีวิตประเภทยูคาริโอตที่ใช้บอกขอบเขตของอินทรอนแต่ละอินทรอน ต่อมาตรวจสอบว่ายีนสิ้นสุดด้วย TAA หรือ TAG หรือ TGA (Stop Codon) และสุดท้ายตรวจสอบว่า ไม่มีจุดสิ้นสุด

ของยีนแทรกอยู่ภายในส่วนของเอ็กซอน หลังจากทำความสะอาดข้อมูลแล้วจะได้ข้อมูลสายนิวคลีโอไทด์ข้าวจำนวน 237 สายนิวคลีโอไทด์ จากข้อมูลจำนวน 237 สายนี้จะมาจากโครโมโซมที่ 1 – 8 โครโมโซมที่ 10 และจากกลุ่มที่ยังไม่รู้ว่าจะจัดอยู่ในโครโมโซมใด (Unknown) ในการแบ่งกลุ่มข้อมูลนั้นจะแบ่งเป็น ข้อมูลสอน (Training set) จำนวน 217 สายซึ่งจะมาจากทุกโครโมโซม สำหรับชื่อสายนิวคลีโอไทด์ที่ใช้เป็นข้อมูลสอนและขนาดสามารถดูได้ที่ภาคผนวก ก และข้อมูลทดสอบ (Testing set) จำนวน 20 สาย ซึ่งก็มาจากทุกโครโมโซมเช่นกัน

วิธีที่ใช้ในการค้นหายีน

หลังจากได้ข้อมูลที่จะใช้สอนและทดสอบโปรแกรมแล้วจะต้องเลือกวิธีที่ใช้ในการค้นหายีน สำหรับวิธีโดยทั่วไปที่ใช้ในการค้นหายีนมีทั้งหมด 3 วิธี คือ

1. วิธีการเทียบสายนิวคลีโอไทด์ใหม่กับสายนิวคลีโอไทด์ที่รู้จักแล้วหรือสายนิวคลีโอไทด์ที่หาโปรตีนไว้แล้วซึ่งจะมาจากฐานข้อมูลโปรตีน (Search by Sequence Similarity) สำหรับโปรแกรมที่ใช้วิธีนี้ คือ BLASTX

2. วิธีหาบริเวณที่เกี่ยวข้องกับการควบคุมการแสดงออกของยีน (Search by Signal) วิธีนี้จะหาส่วนต่าง ๆ ที่สำคัญในยีน เช่น ส่วนโปรโมเตอร์ (Promoter) บริเวณที่มีไรโบโซมมาจับ (Ribosome Binding site) บริเวณที่ควบคุมการสร้างโปรตีน (regulatory elements) จุดเริ่มต้นของยีน (Start Codon) ลำดับเบสซ้ำ (repetitive sequence) บริเวณที่เป็น CpG Islands บริเวณที่มีการเติม poly-A เป็นต้น เพื่อที่จะใช้ลักษณะพิเศษเหล่านี้ช่วยในการค้นหายีน

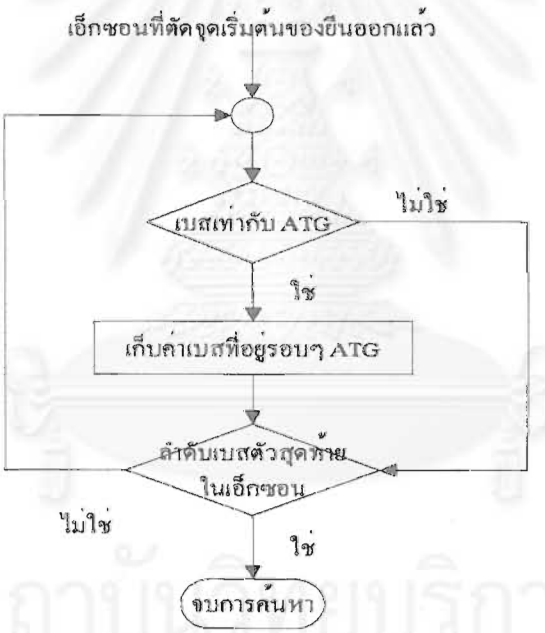
3. วิธีหาคุณสมบัติทางสถิติของลำดับเบสในสายดีเอ็นเอ (Search by Content) วิธีนี้จะใช้ค่าสถิติเกี่ยวกับคุณสมบัติของดีเอ็นเอ เช่น หาความถี่ในการเกิดเบสที่ตำแหน่งของโคดอนต่างๆ ความแตกต่างเกี่ยวกับความถี่ในการเกิดกรดอะมิโนชนิดต่างๆ สัดส่วนการใช้โคดอน (Codon preference) เป็นต้น ซึ่งค่าเหล่านี้จะแตกต่างกันไปในสิ่งมีชีวิตที่ต่างชนิดกัน

สำหรับในงานวิจัยนี้จะเลือกใช้ 2 วิธี คือ วิธีการหาบริเวณที่เกี่ยวข้องกับการควบคุมการแสดงออกของยีน (Search by Signal) และ วิธีการหาคุณสมบัติทางสถิติของลำดับเบสในสายดีเอ็นเอ (Search by Content)

การหาบริเวณที่เกี่ยวข้องกับการควบคุมการแสดงออกของยีน (Search by Signal)

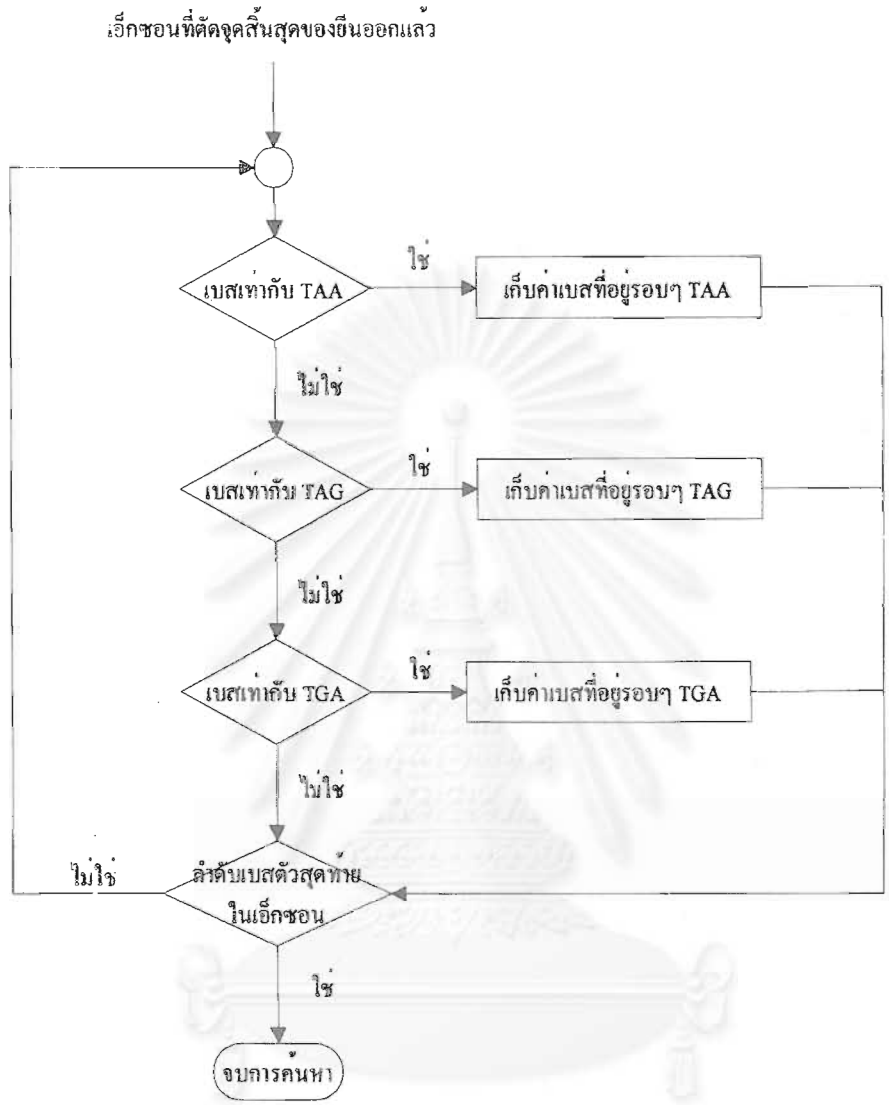
การหาบริเวณที่เกี่ยวข้องกับการควบคุมการแสดงออกของยีน จะใช้เมตริกซ์น้ำหนัก (Weight Matrix) ซึ่งเป็นการหาความถี่ของเบสแต่ละชนิด ที่แต่ละตำแหน่งของลำดับเบสในบริเวณควบคุม จากนั้นหารด้วยความถี่ของการพบเบสที่แต่ละตำแหน่งของบริเวณที่ไม่ใช่บริเวณควบคุม ตัวเลขที่ได้จะแสดงโอกาสของการพบเบสหนึ่งๆ มากน้อยกว่าปกติเป็นกี่เท่าในแต่ละตำแหน่ง จากนั้นเอาค่าความน่าจะเป็นที่ได้ไปหาลอการิทึมของตัวเลขนั้นๆ สำหรับบริเวณการควบคุมการแสดงออกของยีนที่เลือกใช้ในการหายีนจะมี 4 บริเวณ ดังนี้

1. จุดเริ่มต้นของยีน (ATG) โดยจะเก็บข้อมูลความถี่ของเบสระหว่างตำแหน่งที่ -8 ถึง 5 รอบๆ ATG ที่ทำหน้าที่เป็นจุดเริ่มต้นของยีน และ เก็บข้อมูลความถี่ของเบสระหว่างตำแหน่งที่ -8 ถึง 5 รอบๆ ATG ใด ๆ ที่ไม่ได้ทำหน้าที่เป็นจุดเริ่มต้นของยีน ดังแสดงในรูปที่ 3.1



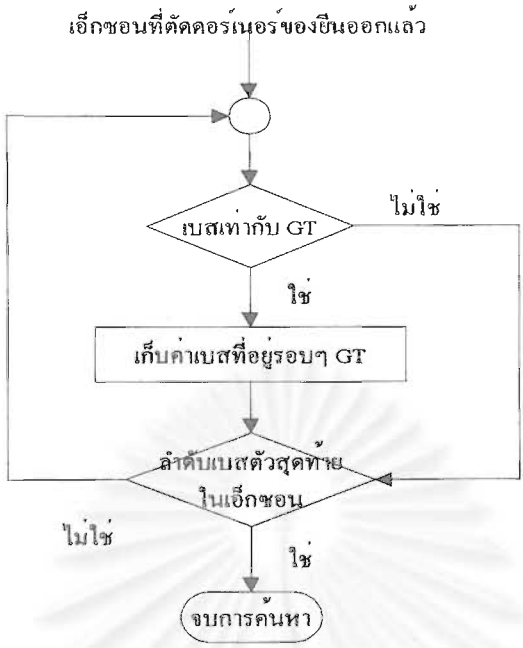
รูปที่ 3.1 วิธีการค้นหาบริเวณที่ไม่ใช่จุดเริ่มต้นของยีน

2. จุดสิ้นสุดของยีน (TAA TGA และ TAG) โดยจะเก็บข้อมูลความถี่ของเบสระหว่างตำแหน่งที่ -5 ถึง 3 รอบๆ TAA TGA หรือ TAG ที่ทำหน้าที่เป็นจุดสิ้นสุดของยีน และ เก็บข้อมูลความถี่ของเบสระหว่างตำแหน่งที่ -5 ถึง 3 รอบๆ TAA TGA หรือ TAG ที่ไม่ได้ทำหน้าที่เป็นจุดสิ้นสุดของยีน ดังแสดงในรูปที่ 3.2



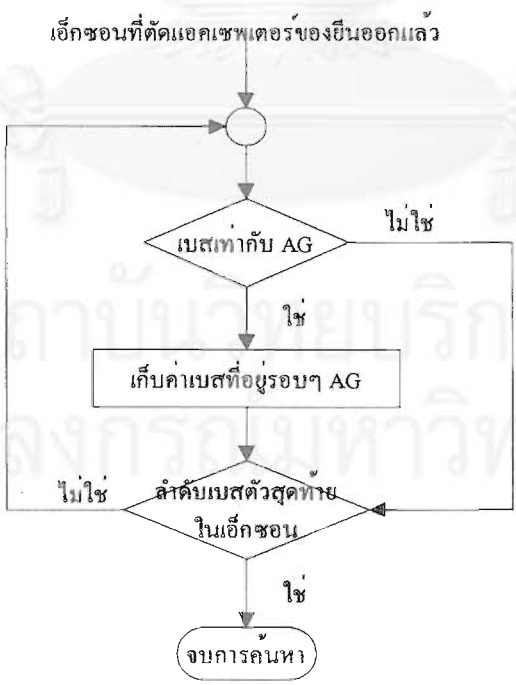
รูปที่ 3.2 วิธีการค้นหาบริเวณที่ไม่ใช่จุดสิ้นสุดของยีน

3. ดอร์เนอร์ (GT) โดยจะเก็บข้อมูลของเบสระหว่างตำแหน่งที่ -3 ถึง 5 รอบๆ GT ที่ทำหน้าที่เป็นดอร์เนอร์ของยีน และ เก็บข้อมูลความถี่ของเบสระหว่างตำแหน่งที่ -3 ถึง 5 รอบๆ GT ใดๆที่ไม่ได้ทำหน้าที่เป็นดอร์เนอร์ของยีน ดังแสดงในรูปที่ 3.3



รูปที่ 3.3 วิธีการค้นหาบริเวณที่ไม่ใช่ดอร์เนอร์ของยีน

4. แอคเซพเตอร์ (AG) โดยเก็บข้อมูลของเบสระหว่างตำแหน่งที่ -22 ถึง 1 รอบๆ AG ที่ทำหน้าที่แอคเซพเตอร์ของยีน และ เก็บข้อมูลความถี่ของเบสระหว่างตำแหน่งที่ -22 ถึง 1 รอบๆ AG ใด ๆ ที่ไม่ได้ทำหน้าที่เป็นแอคเซพเตอร์ของยีน ดังแสดงในรูปที่ 3.4

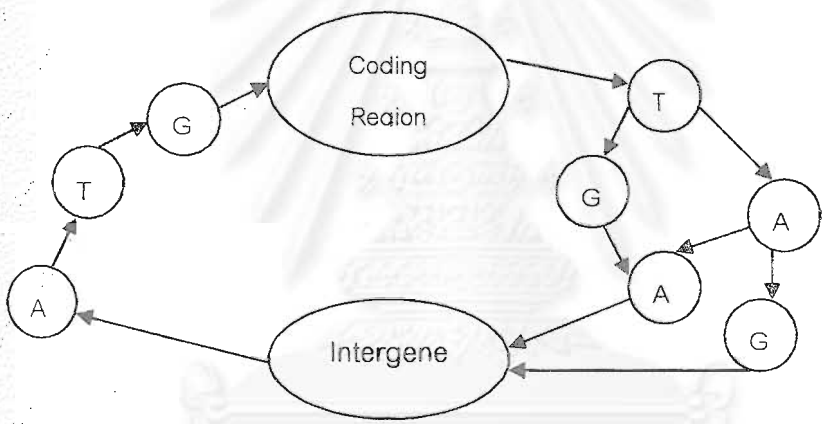


รูปที่ 3.4 วิธีการค้นหาบริเวณที่ไม่ใช่แอคเซพเตอร์ของยีน

จากนั้นนำข้อมูลทั้งหมดที่ได้จากบริเวณดังกล่าวทั้ง 4 บริเวณไปหาค่าเมตริกซ์น้ำหนักเพื่อนำไปใช้เป็นค่าพารามิเตอร์ในการค้นหายีนต่อไป

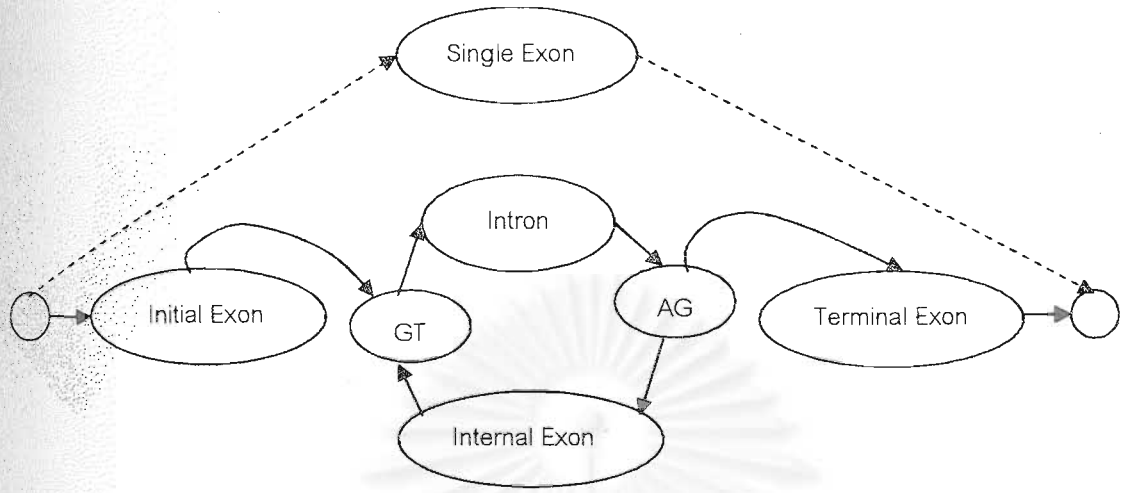
การหาคุณสมบัติทางสถิติของลำดับเบสในสายดีเอ็นเอ (Search by Content)

การหาคุณสมบัติทางสถิติของลำดับเบสในสายดีเอ็นเอ เป็นวิธีการที่ใช้ในการวิเคราะห์คุณสมบัติของลำดับเบสที่ใช้สร้างโปรตีนในยีน (Coding region) ซึ่งจะมีลักษณะเฉพาะแตกต่างกันในสิ่งมีชีวิตแต่ละชนิด เพราะแนวโน้มการใช้โคดอนในยีนจะมีส่วนในการกำหนดลำดับเบสของดีเอ็นเอ ในการวิเคราะห์คุณสมบัติดังกล่าวของดีเอ็นเอจะใช้วิธีมาร์คอฟโมเดลลำดับที่ 5 (5 th Order Markov Model)



รูปที่ 3.5 แบบจำลองภาพรวมทั้งหมด

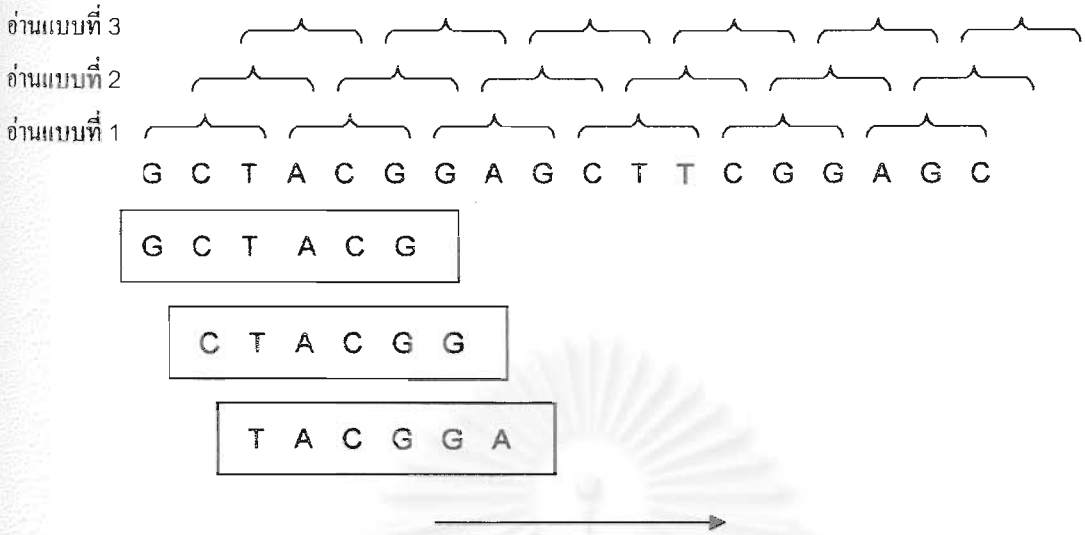
สำหรับการออกแบบโมเดลภาพรวมของการค้นหายีนจะมีลักษณะดังรูปที่ 3.5 เริ่มต้นจาก ATG ที่เป็นจุดเริ่มต้นของยีน ไปยังบริเวณที่จะถูกแปลงเป็นโปรตีน (Coding region) ซึ่งภายในจะมีเอ็กซอนและอินทรอนอยู่ และจบด้วยจุดสิ้นสุดของยีนซึ่งเป็นไปได้สามแบบ คือ TAA TAG และ TGA ต่อจากนั้นก็จะไปที่เบสที่ไม่ใช่ยีนแต่จะคั่นอยู่ระหว่างยีนสองยีน (Intergene)



รูปที่ 3.6 โมเดลส่วนที่เป็นยีน (coding region model)

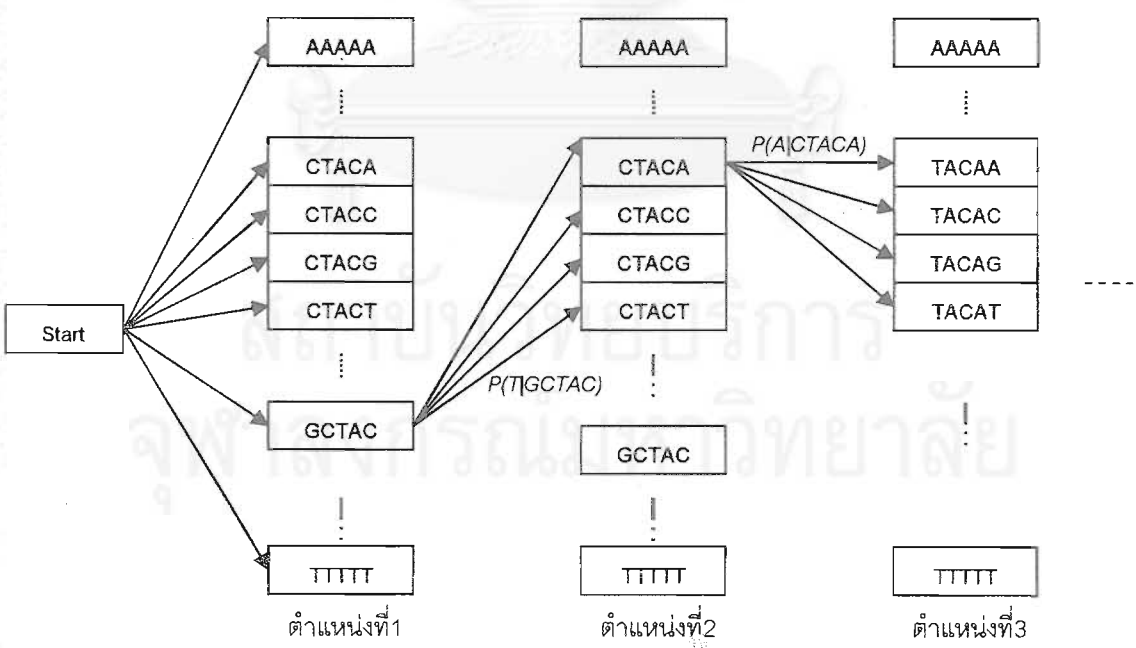
โมเดลในส่วนของยีน แสดงในรูปที่ 3.6 นั้นจะประกอบไปด้วย เอ็กซอน อินทรอน ดอร์เนอร์ (เบส GT) และ แอคเซพเตอร์ (เบส AG) โดยดอร์เนอร์จะเป็นจุดเริ่มต้นของอินทรอน และ แอคเซพเตอร์จะเป็นจุดสิ้นสุดของอินทรอน สำหรับเอ็กซอนจะมีทั้งหมด 4 แบบ ขึ้นกับว่าอยู่ที่ตำแหน่งใด ในกรณีที่ยีนมีเพียง 1 เอ็กซอน (Single Exon) ก็จะไปตามเส้นประของโมเดลซึ่งจะเป็นเอ็กซอนที่อยู่ระหว่างจุดเริ่มต้นกับจุดสิ้นสุดของยีน ส่วนเส้นเข้มในโมเดลนั้นแสดงแทนกรณีที่ยีนมีมากกว่า 1 เอ็กซอน โดยในยีนจะต้องประกอบไปด้วย เอ็กซอนเริ่มต้น (Initial Exon) และเอ็กซอนสุดท้าย (Terminal Exon) ส่วนเอ็กซอนกลาง (Internal Exon) นั้นจะมีหรือไม่มีก็ได้ หรืออาจจะมีมากกว่า 1 ก็ได้

สำหรับโมเดลในส่วนของเอ็กซอนนั้น เนื่องจากสายดีเอ็นเอแต่ละสาย (Forward Strand และ Complementary Strand) สามารถอ่านได้ 3 แบบ (3 Reading Frames) ซึ่งคุณสมบัตินี้เกิดขึ้นจากการที่โคดอนเกิดจากการเรียงตัวกันของเบสสามตัวดังนั้นทำให้สายนิวคลีโอไทด์ 1 สาย สามารถอ่านได้ 3 แบบขึ้นอยู่กับว่าเริ่มอ่านที่ตำแหน่งใดของโคดอน ดังรูป 3.7



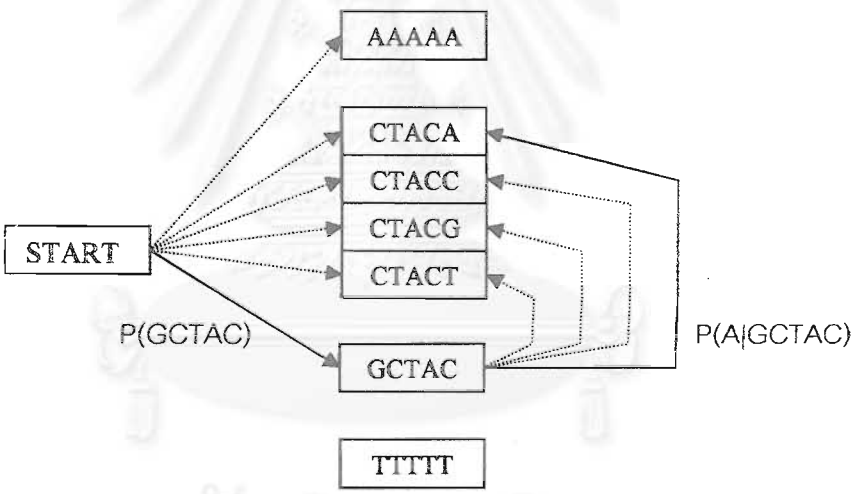
รูปที่ 3.7 การอ่านสายนิวคลีโอไทด์ 3 แบบ

การที่สายนิวคลีโอไทด์สามารถอ่านได้ 3 แบบนี้ ทำให้ในการสร้างโมเดลในส่วนของเอ็กซอนต้องหาค่าพารามิเตอร์ของมาร์คอฟโมเดลตามตำแหน่งของโคดอนซึ่งมี 3 ตำแหน่ง ดังรูปที่ 3.8 เช่น ค่าพารามิเตอร์ AAAAA ก็จะต้องหาทั้งหมด 3 ค่า คือ AAAAA ที่เริ่มต้นจากโคดอนตำแหน่งที่ 1 2 และ 3



รูปที่ 3.8 โมเดลเอ็กซอนที่มีการกำหนดตามตำแหน่งของโคดอน

เนื่องจากเราใช้มาร์คอฟโมเดลลำดับที่ 5 ดังนั้นจึงต้องมีพารามิเตอร์ที่เป็นค่าความน่าจะเป็นเริ่มต้น (initial probabilities) คือตั้งแต่สถานะ AAAAA จนกระทั่งถึงสถานะ TTTTT ทั้งหมดเท่ากับ 1024 ค่า (4*4*4*4*4) และเนื่องจากความแตกต่างกันในการเริ่มอ่านค่าที่ตำแหน่งโคดอนต่าง ๆ ดังนั้นจึงต้องคิดค่าพารามิเตอร์ที่เป็นค่าความน่าจะเป็นเริ่มต้นของทั้งตำแหน่งโคดอนที่ 1 2 และ 3 ตามลำดับ ดังนั้นจึงมีค่าความน่าจะเป็นเริ่มต้นทั้งสิ้น 3072 ค่า สำหรับวิธีการคำนวณค่าความน่าจะเป็นในการเกิดซีเควนแบบต่างๆ สามารถคิดได้ดังรูปที่ 3.9 ซึ่งแสดงวิธีการคำนวณหาค่าความน่าจะเป็นในการเกิดซีเควน GCTACA โดยค่าความน่าจะเป็นในการเกิด GCTACA จะเท่ากับ P(GCTAC) * P(A|GCTAC) ดังนั้นค่าความน่าจะเป็นในการเปลี่ยนสถานะ (Transition probabilities) จึงมีจำนวนทั้งสิ้น 4096 ค่า (4*4*4*4*4*4) และเช่นเดียวกับค่าความน่าจะเป็นเริ่มต้นที่จะต้องหาตามตำแหน่งโคดอนด้วย ดังนั้นจึงมีค่าความน่าจะเป็นในการเปลี่ยนสถานะทั้งสิ้น 12288 ค่า



รูปที่ 3.9 วิธีคำนวณหาความน่าจะเป็นของมาร์คอฟโมเดลลำดับที่ 5

ในการคำนวณของโปรแกรมนั้นจะนำค่าพารามิเตอร์ไปหาลอการิทึมอีกครั้ง เนื่องมาจากค่าพารามิเตอร์ที่ได้มาเมื่อนำมาคูณกันเข้าจะได้เลขที่เป็นจำนวนเล็กมาก อาจทำให้เกิดปัญหาในการคำนวณของเครื่องคอมพิวเตอร์ขึ้นได้ (Underflow)

ขั้นตอนของการค้นหายีน

หลังจากรับข้อมูลสายนิวคลีโอไทด์ที่ต้องการค้นหายีนแล้ว โปรแกรมจะมีขั้นตอนการทำงานดังแสดงในรูปที่ 3.14 ซึ่งมีรายละเอียดดังนี้ คือ

1. หาจุดเริ่มต้น(ATG) จุดสิ้นสุด (TGA TAG หรือ TAA) ดอร์เนอร์ (GT) และแอคเซพเตอร์ (AG) ที่เป็นไปได้ทั้งหมดของสายนิวคลีโอไทด์โดยใช้ค่าพารามิเตอร์จากเมตริกซ์น้ำหนัก

1.1 จุดเริ่มต้น (ATG) คำนวณค่า Log-odds Ratio ที่ตำแหน่งที่ -8 ถึง 5 รอบๆ เบส ATG ในสายนิวคลีโอไทด์

1.2 จุดสิ้นสุด (TGA TAG หรือ TAA) คำนวณค่า Log-odds Ratio ระหว่างตำแหน่งที่ -5 ถึง 3 รอบๆเบส TAA TGA หรือ TAG ในสายนิวคลีโอไทด์

1.3 ดอร์เนอร์ (GT) คำนวณค่า Log-odds Ratio ระหว่างตำแหน่งที่ -3 ถึง 5 รอบๆเบส GT ในสายนิวคลีโอไทด์

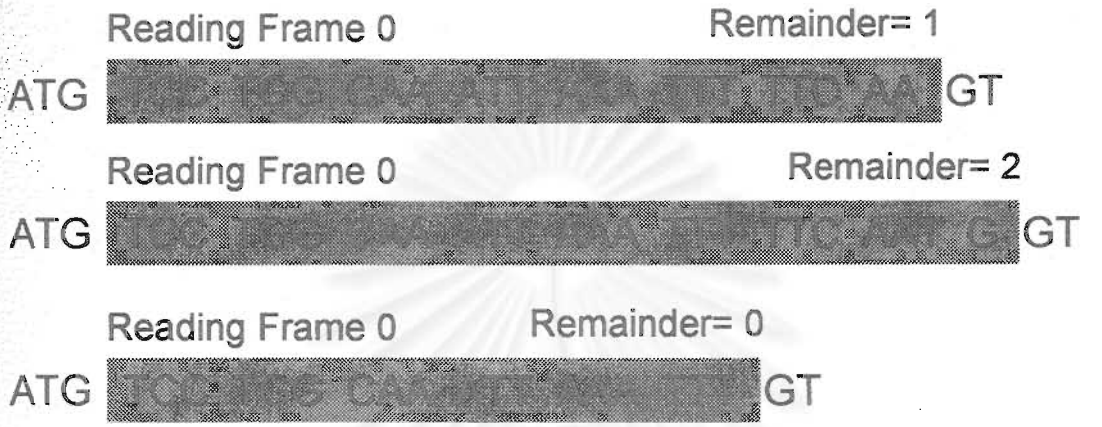
1.4 แอคเซพเตอร์ (AG) คำนวณค่า Log-odds Ratio ระหว่างตำแหน่งที่ -22 ถึง 1 รอบๆเบส AG ในสายนิวคลีโอไทด์

จากจุดเริ่มต้น (ATG) จุดสิ้นสุด (TGA TAG หรือ TAA) ดอร์เนอร์ (GT) และแอคเซพเตอร์ (AG) ที่เป็นไปได้ทั้งหมดนั้นจะคัดแต่เฉพาะที่มีค่าเกินค่าที่กำหนดไว้เท่านั้น โดยค่าที่กำหนดจะคิดจากค่าต่ำที่สุดที่สามารถเป็นไปได้ของบริเวณต่าง ๆ

2. หาเอ็กซอนที่เป็นได้ทั้งหมดโดยใช้ตำแหน่งของ จุดเริ่มต้นของยีน ดอร์เนอร์ แอคเซพเตอร์ และจุดสิ้นสุดของยีน

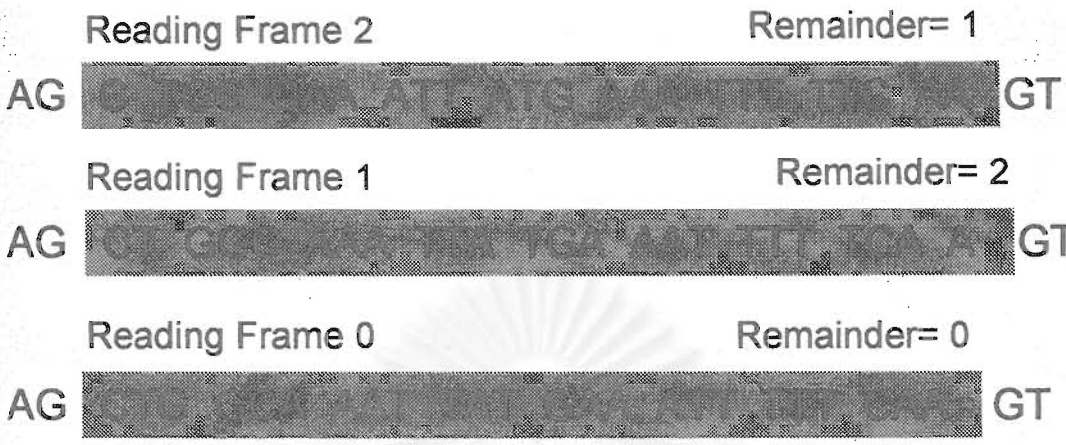
2.1 หาเอ็กซอนที่เป็นเอ็กซอนเริ่มต้น (Initial Exon) ซึ่งจะอยู่ระหว่างจุดเริ่มต้นของยีน และ ดอร์เนอร์ โดยเริ่มหาจากจุดเริ่มต้นของยีน (ATG) ไปยังทุก ดอร์เนอร์ (GT) ที่อยู่ถัดจากจุดเริ่มต้นของยีนโดยจะต้องไม่มีจุดสิ้นสุดของยีน (TAA ,TAG หรือ TGA) แทรกอยู่ ในกรณีที่จากจุดเริ่มต้น (ATG) เดียวกันสามารถหาเอ็กซอนเริ่มต้นได้มากกว่าหนึ่งให้เลือกเฉพาะ เอ็กซอนที่มีค่าความน่าจะเป็นของดอร์เนอร์สูงๆ และให้ค่า Reading Frame เท่ากับ 0 จากนั้นหาค่าเศษที่เหลือเมื่อนำความยาวหารด้วย 3 เพื่อให้ได้ว่าขาดเบสอีกกี่ตัวจึงจะสามารถแปลงเป็น โคดอนได้

ครบ (หารด้วยสามลงตัว) และเก็บไว้เป็นค่า Remainder ซึ่งมีค่าได้ตั้งแต่ 0 1 และ 2 (ค่า Remainder และ Reading Frame นี้ไว้ใช้ในการรวมยีน) ตัวอย่างการคิด Reading Frame และ Remainder ของเอ็กซอนแรกดังแสดงในรูปที่ 3.10



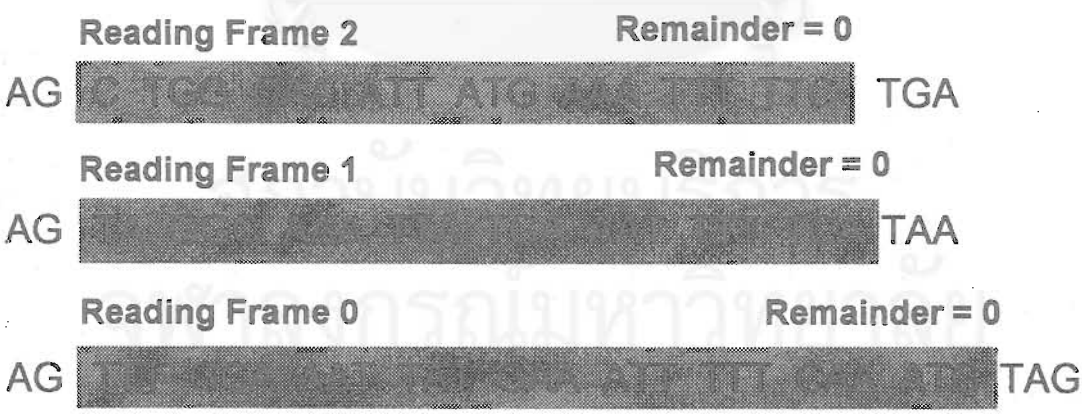
รูปที่ 3.10 การคิด Reading Frame และ Remainder ของเอ็กซอนแรก

2.2 หาเอ็กซอนที่เป็นเอ็กซอนกลาง (Internal Exon) ซึ่งจะอยู่ระหว่างดอร์เนอร์ และ แอคเซพเตอร์ เริ่มหาจากแอคเซพเตอร์ (AG) ไปยังทุกๆ ดอร์เนอร์ (GT)ที่อยู่ถัดไปจากแอคเซพเตอร์โดยจะต้องไม่มีจุดสิ้นสุดของยีน (TAA, TAG หรือ TGA) แทรกอยู่ ในกรณีที่จากแอคเซพเตอร์ (AG) เดียวกันสามารถหา เอ็กซอนกลางได้มากกว่าหนึ่งให้เลือกเฉพาะเอ็กซอนที่มีค่าความน่าจะเป็นของ ดอร์เนอร์สูงที่สุดจำนวน 5 ดอร์เนอร์ ในการหาเอ็กซอนกลางจะต้องหาจากทุก Reading Frame (0 1 และ 2) จากนั้นก็จะหาจุดสิ้นสุดของยีนในแต่ละ Reading Frame และจากนั้นจึงหาดอร์เนอร์ที่อยู่ภายในจุดสิ้นสุดของยีน และนำความยาวของเอ็กซอนที่ได้ บวกกับค่า Reading Frame และหารด้วยสาม เพื่อหาเศษเหลือ เพื่อหาว่ายังขาดเบสอีกกี่ตัวจึงจะครบตำแหน่งของโคดอนจากนั้นนำค่านั้นไปเก็บไว้เป็นค่า Remainder ตัวอย่างการคิดค่า Reading Frame และ Remainder ของเอ็กซอนกลางจะแสดงดังรูปที่ 3.11



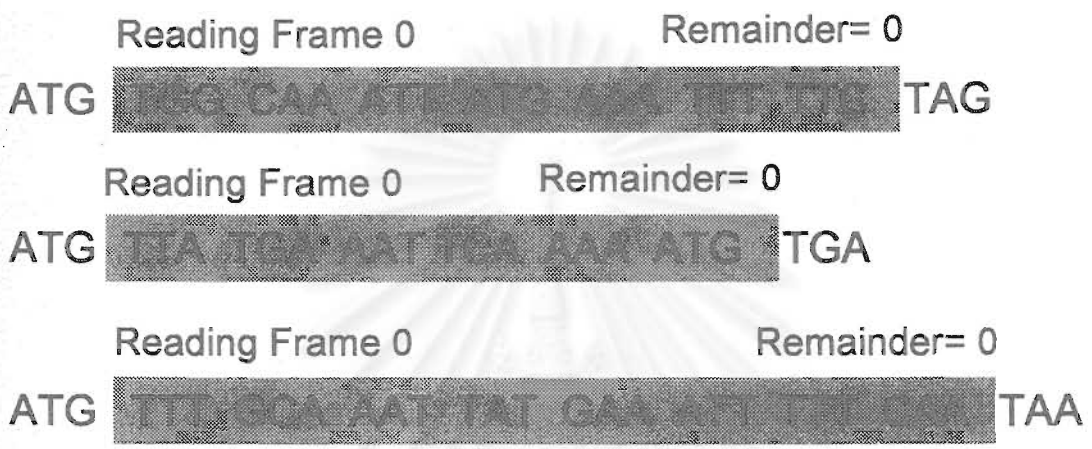
รูปที่ 3.11 การคิด Reading Frame และ Remainder ของเอ็กซอนกลาง

2.3 หาเอ็กซอนที่เป็นเอ็กซอนสุดท้ายในยีน (Terminal Exon) ซึ่งจะอยู่ระหว่างแอคเซพเตอร์ และ จุดสิ้นสุดของยีน เริ่มหาจากแอคเซพเตอร์ของยีนไปยังจุดสิ้นสุดของยีนที่อยู่ถัดไปโดยหาในทุก Reading Frame (0 1 และ 2) และให้ค่า Remainder เท่ากับ 0 ตัวอย่างการคิด Reading Frame และ Remainder ของเอ็กซอนสุดท้าย ดังแสดงในรูปที่ 3.12



รูปที่ 3.12 การคิด Reading Frame และ Remainder ของเอ็กซอนสุดท้าย

2.4 หาเอ็กซอนที่เป็นเอ็กซอนเดี่ยว ๆ ในยีน (Single Exon) ซึ่งจะอยู่ระหว่างจุดเริ่มต้นของยีน กับ จุดสิ้นสุดของยีน เริ่มหาจุดเริ่มต้นของยีนที่อยู่ใกล้ๆกับจุดสิ้นสุดของยีน โดยต้องมีความยาวจากจุดเริ่มต้นของยีนไปยังจุดสิ้นสุดของยีนที่หารด้วยสามลงตัว (แปลงเป็นโคดอนได้ทุกตัว) ตัวอย่างการคิด Reading Frame และ Remainder ของเอ็กซอนเดี่ยว ดังแสดงในรูปที่ 3.13



รูปที่ 3.13 การคิด Reading Frame และ Remainder ของเอ็กซอนเดี่ยว

3 คำนวณ Log-odds Ratio เป็นของเอ็กซอนโดยใช้ค่าพารามิเตอร์ของมาร์คอฟโมเดลลำดับที่ 5 จากค่า Log-odds Ratio ของเอ็กซอนทั้งหมด จะเลือกแต่เฉพาะเอ็กซอนที่มีค่า Log-odds Ratio มากกว่าค่าที่กำหนดไว้เท่านั้น โดยคำนวณดังนี้

เอ็กซอนแรก = ค่า Log-odds Ratio ของจุดเริ่มต้นของยีน + ค่า Log-odds Ratio ของมาร์คอฟโมเดล + ค่า Log-odds Ratio ของดอร์เนอร์

เอ็กซอนกลาง = ค่า Log-odds Ratio ของแอคเซพเตอร์ + ค่า Log-odds Ratio ของมาร์คอฟโมเดล + ค่า Log-odds Ratio ของดอร์เนอร์

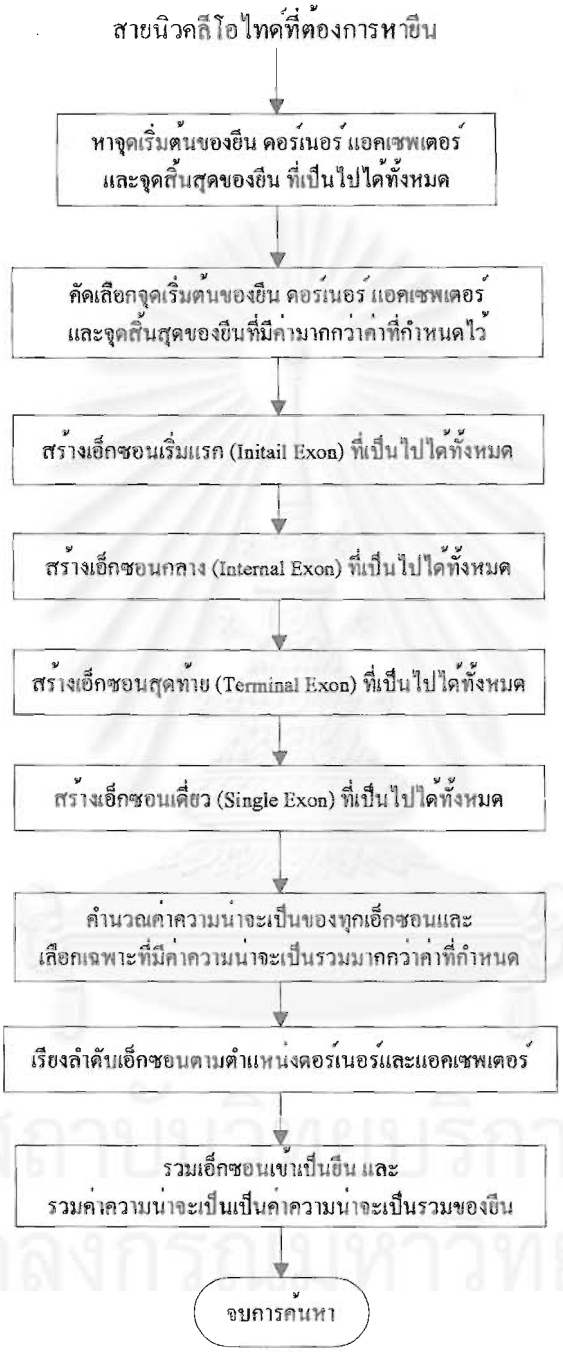
เอ็กซอนสุดท้าย = ค่า Log-odds Ratio ของแอคเซพเตอร์ + ค่า Log-odds Ratio ของมาร์คอฟโมเดล + ค่า Log-odds Ratio ของจุดสิ้นสุดของยีน

เอ็กซอนเดี่ยว = ค่า Log-odds Ratio ของจุดเริ่มต้นของยีน + ค่า Log-odds Ratio ของมาร์คอฟโมเดล + ค่า Log-odds Ratio ของจุดสิ้นสุดของยีน

โดยค่า Log-odds Ratio ของมาร์คอฟโมเดลนั้นจะหาจากตำแหน่งของโคดอน (3 Reading frames) หรือ ค่า Reading Frame (0 1 และ 2)

4 จากเอ็กซอนเริ่มต้น เอ็กซอนกลาง และ เอ็กซอนสุดท้ายจะต้องนำมาเรียงลำดับตามแอคเซพเตอร์ และ ดอร์เนอร์ เพื่อหาว่าเอ็กซอนใดจัดอยู่ในยีนเดียวกันบ้าง

5 รวมเอ็กซอนเข้าเป็นยีน และรวมค่า Log-odds Ratio ของเอ็กซอนเข้าเป็นค่า Log-odds Ratio รวมของยีน ในการรวมยีนจะใช้ค่า Reading Frame และ Remainder ในการรวมยีน โดยเอ็กซอนแรกต้องมี Reading frame เท่ากับ 0 จะมีRemainder เท่าใดก็ได้ แต่ Remainder ของเอ็กซอนแรก เมื่อรวมกับค่า Reading Frame ของเอ็กซอนถัดไปแล้วจะต้องมีค่ารวมที่หารด้วยสามแล้วจะต้องไม่เหลือเศษ (Remainder 0 รวมกับ Reading Frame 0 Remainder 1 รวมกับ Reading Frame 2 และ Remainder 2 รวมกับ Reading Frame 1) การทำเช่นนี้จะทำให้ความยาวรวมทั้งหมดของทุกเอ็กซอนในยีนจะต้องหารด้วยสามลงตัว (แปลงเป็นโคดอนได้หมดทุกเบสในยีน)



รูปที่ 3.14 ขั้นตอนวิธีการค้นหายีน

การพัฒนาโปรแกรม

เราได้พัฒนาโปรแกรมค้นหายีนข้าวโดยใช้ภาษาซี Borland C++ builder มีชื่อว่า RGF (Rice Gene Finding program) โปรแกรม RGF จะทำงานบนระบบปฏิบัติการวินโดว์บนเครื่องคอมพิวเตอร์ส่วนบุคคล ข้อมูลนำเข้าของโปรแกรมจะมี 2 แฟ้มข้อมูล คือ ข้อมูลสายนิวคลีโอไทด์ที่ต้องการหายีนที่อยู่ในรูปแบบ FASTA และ แฟ้มข้อมูลพารามิเตอร์ซึ่งประกอบด้วยความน่าจะเป็นเริ่มต้น ความน่าจะเป็นในการเปลี่ยนสถานะ และเมตริกซ์น้ำหนักของบริเวณต่างๆ

สำหรับการทำงานของโปรแกรม RGF จะเริ่มต้นจากค้นหาบริเวณต่างๆ เช่น จุดเริ่มต้นของยีน จุดสิ้นสุดของยีน ดอร์เนอร์ และแอคเซพเตอร์ โดยใช้เมตริกซ์น้ำหนักเพื่อหาบริเวณที่เป็นไปได้ทั้งหมด โดยค่าเมตริกซ์น้ำหนักจะคำนวณเป็นค่าลอการิทึม จากนั้นคำนวณหาเอ็กซอนที่เป็นไปได้ทั้งหมดโดยใช้ค่าความน่าจะเป็นเริ่มต้น และค่าความน่าจะเป็นในการเปลี่ยนสถานะของมาร์คอฟโมเดลลำดับที่ 5 โดยคำนวณค่าเป็นลอการิทึมเช่นกัน จากนั้นนำเอ็กซอนที่ได้มาเรียงตามค่าความน่าจะเป็น และ รวมเอ็กซอนที่น่าจะเป็นยีนเดียวกันมารวมกัน โดยเอ็กซอนแรกจะต้องอยู่ระหว่างจุดเริ่มต้นของยีน กับดอร์เนอร์ เอ็กซอนกลางจะอยู่ระหว่างดอร์เนอร์และแอคเซพเตอร์ เอ็กซอนสุดท้ายจะอยู่ระหว่างดอร์เนอร์กับจุดสิ้นสุดของยีน ส่วนเอ็กซอนเดี่ยวจะอยู่ระหว่างจุดเริ่มต้นของยีนกับจุดสิ้นสุดของยีน จากนั้นนำค่าความน่าจะเป็นของเอ็กซอนมารวมกันเพื่อเป็นค่าความน่าจะเป็นรวมของยีน จากนั้นนำยีนที่ได้มาแสดงผล ซึ่งผลที่ได้จากโปรแกรม RGF จะมีลักษณะดังนี้

```
พุธ 19 มิถุนายน 2002 10:46:20
Parameter File: d:\RiceProject\include\Oryza.txt   Sequence File: C:\My Documents\Data\AP002867.fasta
Length of Input DNA sequence = 146480 bytes
                Max score Genes 32 genes. Score = 2089.145287
Gene 1(Reverse). 4 exons. Score = 23.762019
    Terminal exon        From  218    To  735  19.71    -
    Internal exon        From  925    To  1090 -1.29    -
    Internal exon        From  2892   To  2987  1.88    -
     First exon          From  3394   To  3433  3.46    -
Gene 2(Forward). 1 exons. Score = 20.067993
     Single exon         From  5700   To  6098 20.07    +
```

รูปที่ 3.15 ผลที่ได้จากโปรแกรม RGF

ผลลัพธ์จากโปรแกรมในรูปที่ 3.15 จะบอกรายละเอียดดังนี้ คือ วันที่และเวลาที่สั่งให้โปรแกรมทำงาน ชื่อแฟ้มข้อมูลพารามิเตอร์ ชื่อสายนิวคลีโอไทด์ที่เป็นข้อมูลนำเข้า ขนาดความยาวของสายข้อมูลนำเข้า จำนวนยีนที่หาได้ทั้งหมด และรายละเอียดของแต่ละยีนรวมทั้งรายละเอียดของ เอ็กซอนภายในยีนนั้นๆ สำหรับรายละเอียดของเอ็กซอนจะมีดังนี้ คือ ตำแหน่งเริ่มต้นของ เอ็กซอน ตำแหน่งสุดท้ายของเอ็กซอน ค่าความน่าจะเป็นของเอ็กซอน และบอกว่าเอ็กซอนนั้นอยู่ที่สายใด

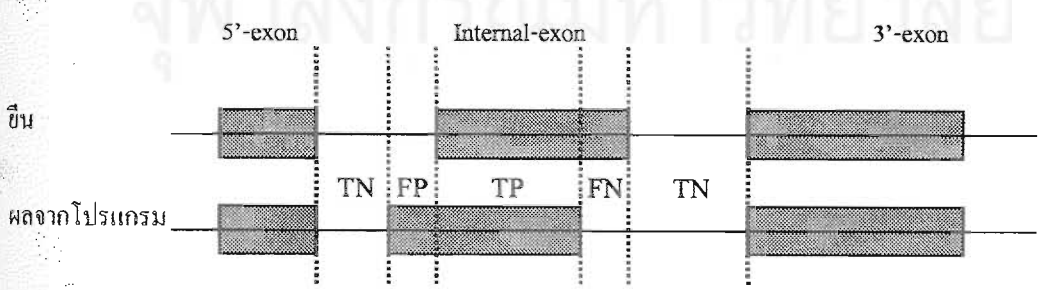
วิธีการทดสอบผลของโปรแกรม

ในขั้นตอนการทดสอบผลของโปรแกรม RGF จะนำข้อมูลที่แบ่งไว้เป็นข้อมูลทดสอบจำนวน 20 สายนิวคลีโอไทด์ หรือ 547 ยีน มาทดสอบผลของโปรแกรม โดยนำผลที่ได้จากโปรแกรมมาเปรียบเทียบกับยีนจริงๆ ของข้อมูลทดสอบ จากนั้นนำมาคำนวณค่าทางสถิติที่ใช้วัดประสิทธิภาพของโปรแกรมในการค้นหายีน ซึ่งจะวัดผลของค่า 2 ค่า ดังนี้ คือ Sensitivity (Sn) และ Specificity (Sp) จากสูตรตามสมการที่ (3.1) และสมการที่ (3.2) ตามลำดับ

$$\text{Sensitivity} = \frac{TP}{TP + FN} \qquad (3.1)$$

$$\text{Specificity} = \frac{TP}{TP + FP} \qquad (3.2)$$

สำหรับค่า TP TN FP และ FN สามารถหาได้โดยเทียบผลลัพธ์ที่ได้จากโปรแกรมหายีนกับสายนิวคลีโอไทด์ที่หายีนไว้แล้ว ดังรูปที่ 3.16



รูปที่ 3.16 บริเวณที่เป็น TP TN FP และ FN

โดยที่ TP แทน true positive ซึ่งเป็นบริเวณที่เป็นเอ็กซอนและสามารถบอกได้ถูกต้อง

TN แทน true negative เป็นบริเวณที่ไม่ใช่เอ็กซอนและสามารถบอกได้ถูกต้อง

FP แทน false positive บริเวณที่ระบุผิดพลาดโดยบอกว่าเป็นเอ็กซอนแต่จริงๆ แล้วบริเวณนั้นไม่ใช่เอ็กซอน

FN แทน false negative บริเวณที่ระบุผิดพลาดโดยบอกว่าไม่ใช่เอ็กซอนแต่จริงๆ แล้ว บริเวณนั้นเป็นเอ็กซอน

Sn แทน อัตราส่วนของ ส่วนของเอ็กซอนที่สามารถระบุได้ถูกต้องโดยโปรแกรมเทียบกับ เอ็กซอนจริงๆในยีน

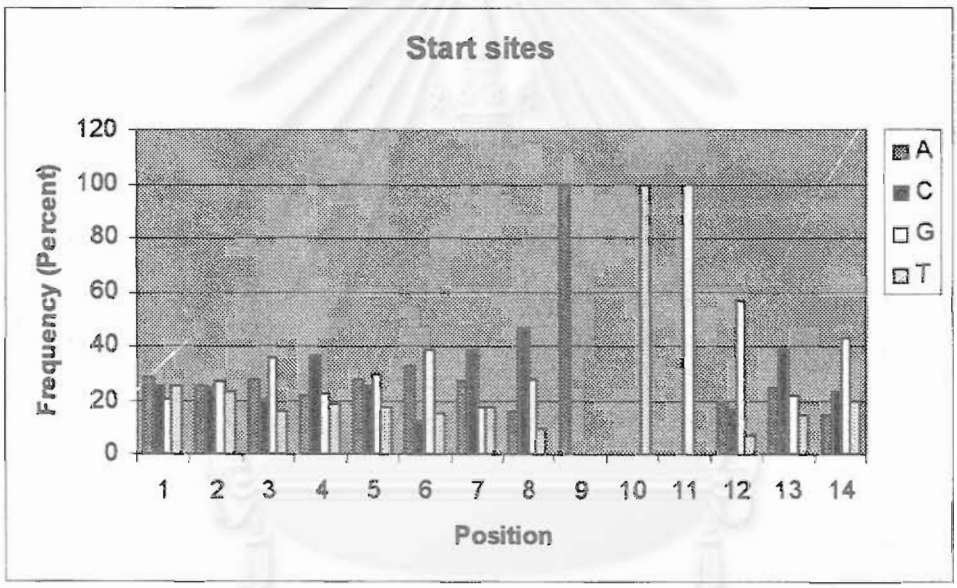
Sp แทน อัตราส่วนของ ส่วนของเอ็กซอนที่สามารถระบุได้ถูกต้องโดยโปรแกรมเทียบกับเอ็กซอนที่ทำนายได้จากโปรแกรม

บทที่ 4

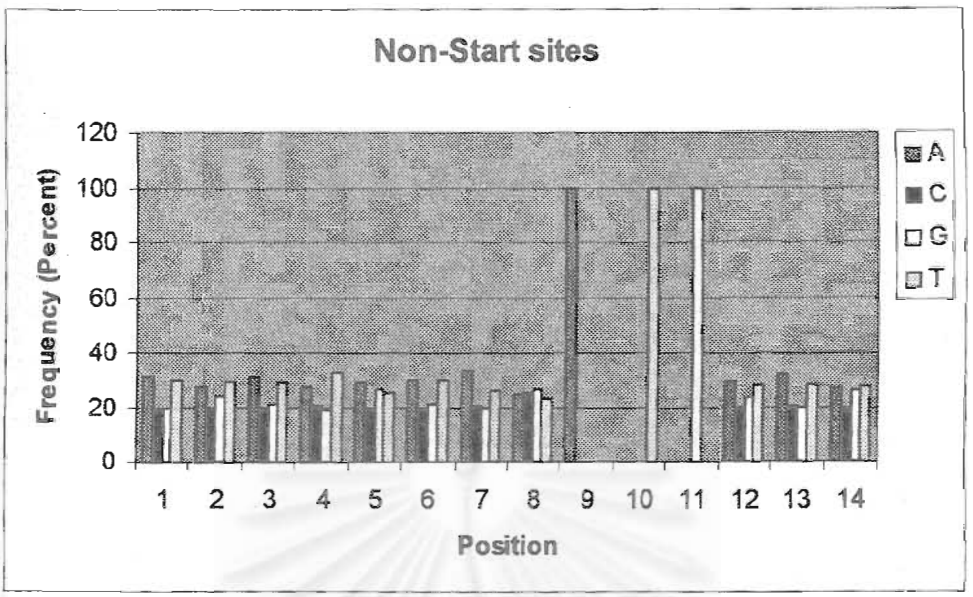
ผลการทดลอง

ผลที่ได้จากการหาค่าพารามิเตอร์ของโปรแกรม

จากการหาพารามิเตอร์แบบเมตริกซ์น้ำหนักสำหรับบริเวณที่เป็นจุดเริ่มต้นของยีน จุดสิ้นสุดของยีน ดอร์เนอร์ และแอคเซพเตอร์ ทำให้ได้ความถี่ของการพบเบส ณ ตำแหน่งต่างๆ ดังต่อไปนี้

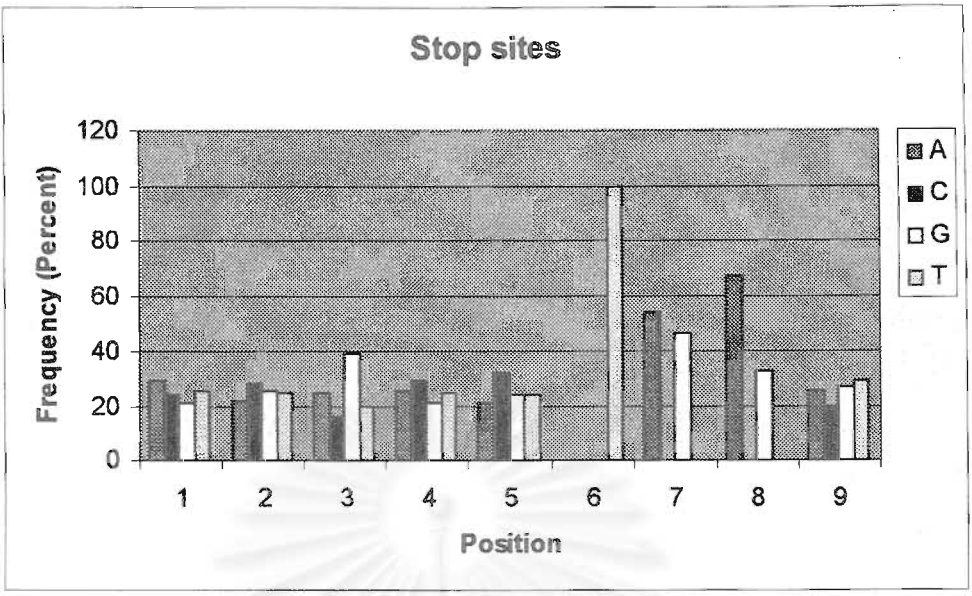


รูปที่ 4.1 ความถี่ของเบสรอบ ๆ ATG ที่เป็นจุดเริ่มต้นของยีน

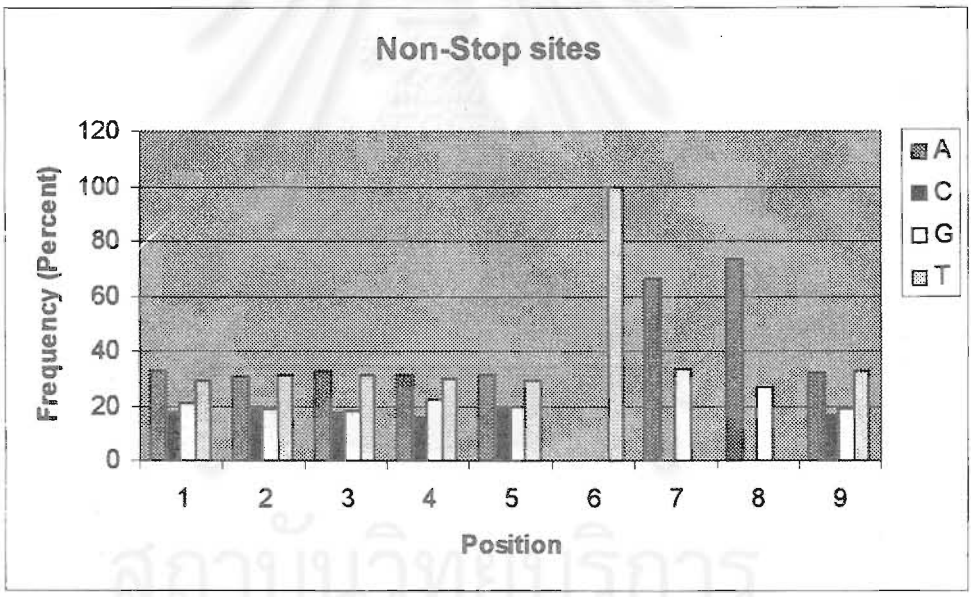


รูปที่ 4.2 ความถี่ของเบสรอบ ๆ ATG ที่ไม่ใช่จุดเริ่มต้นของยีน

กราฟในรูปที่ 4.1 แสดงความถี่ของเบสระหว่างตำแหน่งที่ -8 ถึง 5 รอบๆ ATG ที่เป็นจุดเริ่มต้นของยีน และ กราฟในรูปที่ 4.2 แสดงความถี่ของเบสระหว่างตำแหน่งที่ -8 ถึง 5 รอบๆ ATG ที่ไม่ใช่จุดเริ่มต้นของยีน จะเห็นว่าเบสที่ตำแหน่งรอบๆ ATG ที่ไม่ใช่จุดเริ่มต้นของยีน จะมีการเฉลี่ยความถี่ในการพบเบสต่างๆ ซึ่งต่างกับกราฟที่แสดงความถี่ของเบสรอบๆ จุดเริ่มต้นของยีน ซึ่งจะมีเบส C กับเบส G มากกว่าเบส A และ T เนื่องมาจากลักษณะสำคัญของยีนประเภทยูคาริโอตที่จะการพบบริเวณที่เรียกว่า CpG Island อยู่บริเวณใกล้ ๆ ก่อนที่จะพบจุดเริ่มต้นของยีน

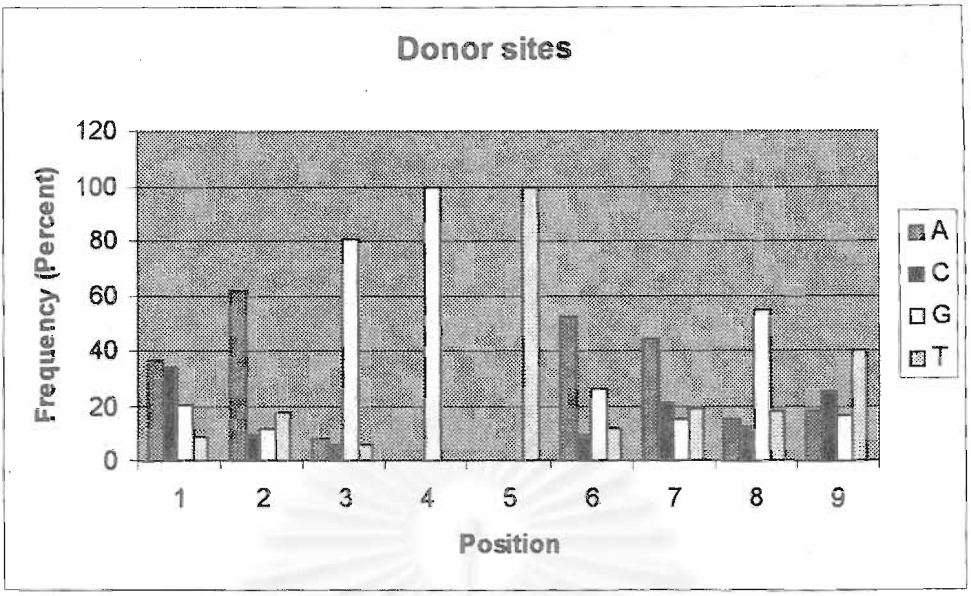


รูปที่ 4.3 ความถี่ของเบสรอบๆ TAA TGA หรือ TAG ที่เป็นจุดสิ้นสุดของยีน

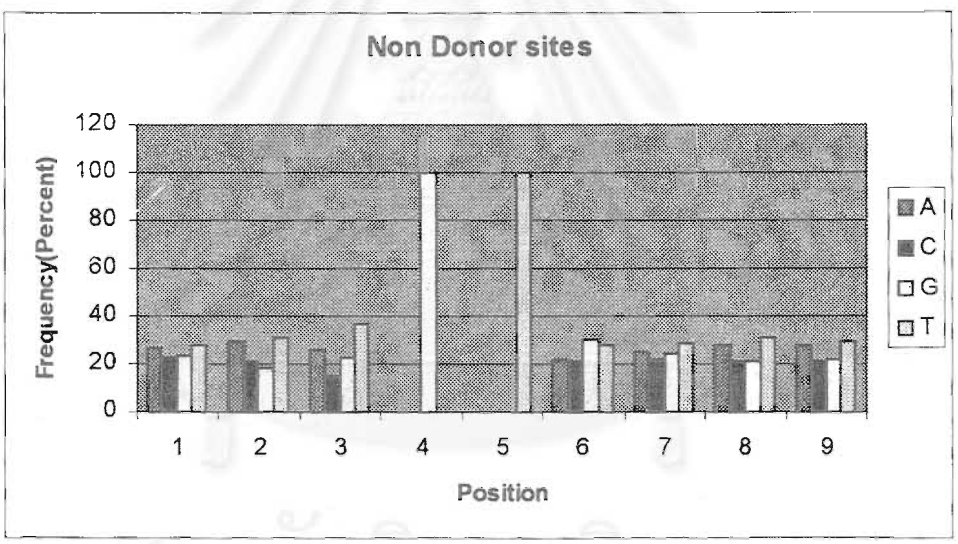


รูปที่ 4.4 ความถี่ของเบสรอบๆ TAA TGA หรือ TAG ที่ไม่ใช่จุดสิ้นสุดของยีน

กราฟรูปที่ 4.3 แสดงความถี่ของเบสระหว่างตำแหน่งที่ -8 ถึง 5 รอบๆ TAA TGA หรือ TAG ซึ่งเป็นจุดสิ้นสุดของยีน และ กราฟรูปที่ 4.4 แสดงความถี่ของเบสระหว่างตำแหน่งที่ -5 ถึง 3 รอบๆ TAA TGA หรือ TAG ที่ไม่ใช่จุดสิ้นสุดของยีน จะเห็นว่าความถี่ของเบสรอบๆ TAA TGA หรือ TAG ที่ไม่ใช่จุดสิ้นสุดของยีนจะมีความถี่ในการพบเบส A และเบส T มากกว่า

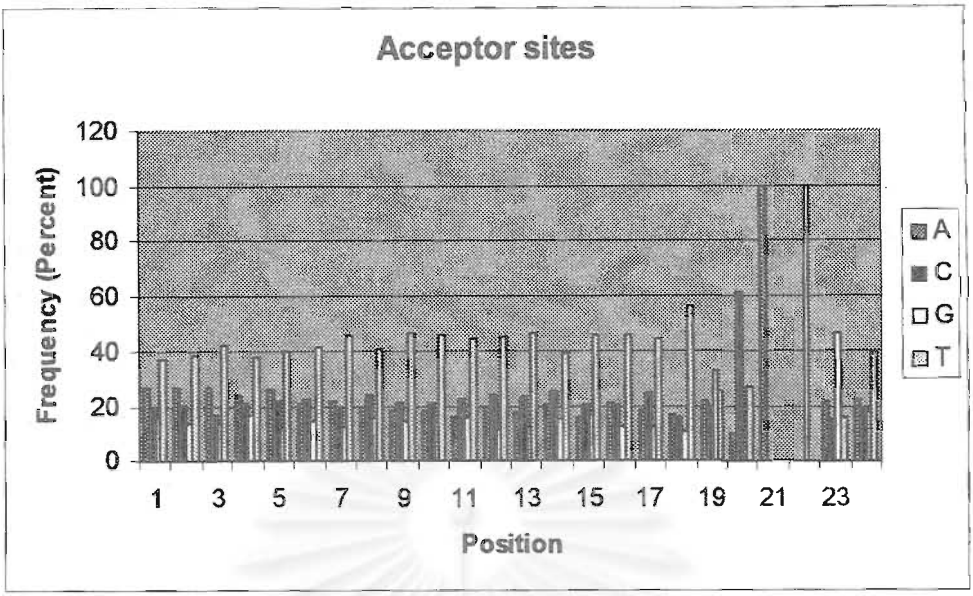


รูปที่ 4.5 ความถี่ของเบสระหว่างตำแหน่งที่ -3 ถึง 5 รอบๆ GT ที่เป็นดอร์เนอร์ของยีน

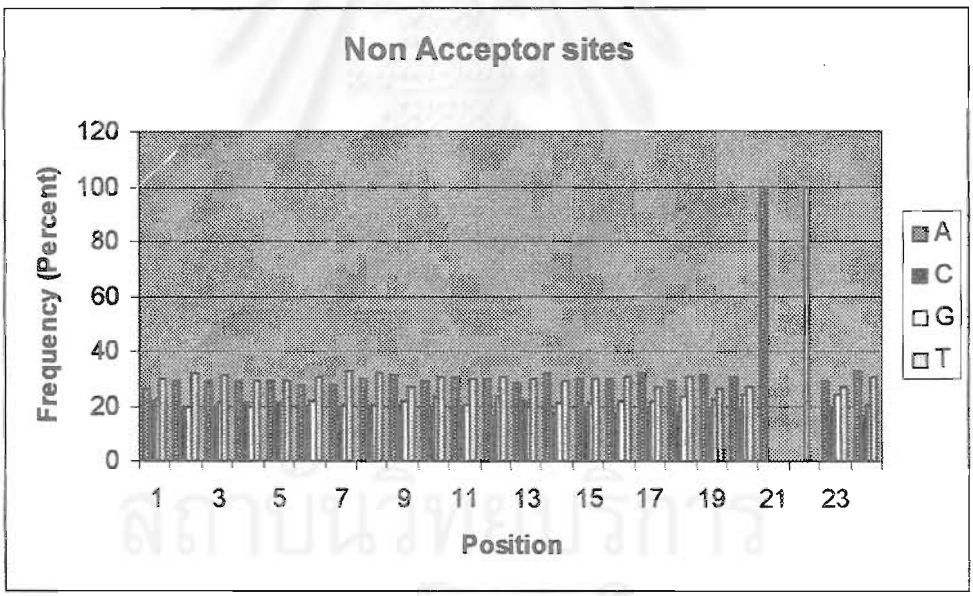


รูปที่ 4.6 ความถี่ของเบสระหว่างตำแหน่งที่ -3 ถึง 5 รอบๆ GT ที่ไม่ใช่ดอร์เนอร์ของยีน

กราฟรูปที่ 4.5 แสดงความถี่ของการพบเบสระหว่างตำแหน่งที่ -3 ถึง 5 รอบๆ GT ที่เป็นทำหน้าที่เป็นดอร์เนอร์ของยีน และ กราฟรูปที่ 4.6 แสดงความถี่ของเบสระหว่างตำแหน่งที่ -3 ถึง 5 รอบๆ GT ที่ไม่ใช่ดอร์เนอร์ของยีน จะพบว่าจากความถี่ที่แสดงในกราฟทำให้สามารถหารูปแบบของดอร์เนอร์ได้ โดยพบว่าส่วนมากดอร์เนอร์ที่พบในยีนข้าวจะมีรูปแบบดังนี้คือ AAGGTAAGT หรือ CAGGTAAGT



รูปที่ 4.7 ความถี่ของเบสระหว่างตำแหน่งที่ -22 ถึง 1 รอบๆ แอคเซพเตอร์ของยีน

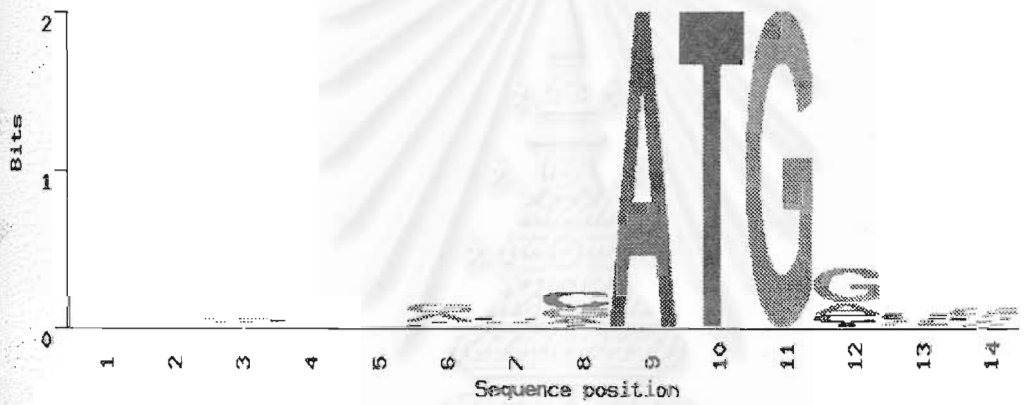


รูปที่ 4.8 ความถี่ของเบสระหว่างตำแหน่งที่ -22 ถึง 1 รอบๆ AG ที่ไม่ใช่แอคเซพเตอร์ของยีน

กราฟรูปที่ 4.7 แสดงความถี่ของเบสระหว่างตำแหน่งที่ -22 ถึง 1 รอบๆ AG ที่เป็นแอคเซพเตอร์ และ กราฟรูปที่ 4.8 แสดงความถี่ของเบสระหว่างตำแหน่งที่ -22 ถึง 1 รอบๆ AG ที่ไม่ใช่แอคเซพเตอร์ จะเห็นว่าความถี่ในการพบเบสต่างๆรอบ AG ที่ไม่ใช่แอคเซพเตอร์จะมีโอกาส

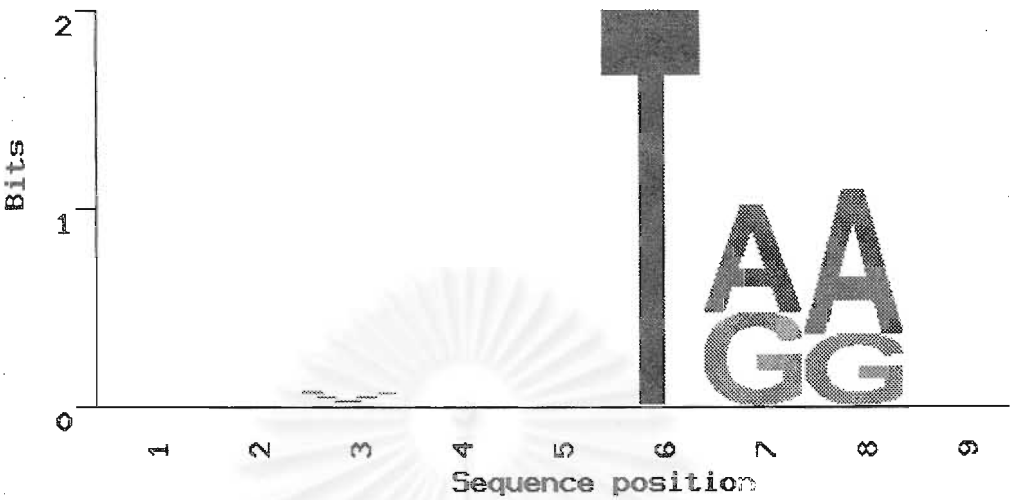
พบเบสต่างๆ เฉลี่ยกันไป แต่สำหรับความถี่ของการพบเบสรอบ ๆ แอคเซพเตอร์จะพบว่าโดยส่วนมากแอคเซพเตอร์ที่พบในยีนข้าวจะมีรูปแบบดังนี้คือ TTTTTTTTTTTTTTTTTTTCAGGT

เราสามารถนำผลจากกราฟที่แสดงความถี่ของการพบเบสที่ตำแหน่งต่างๆรอบจุดเริ่มต้นของยีน จุดสิ้นสุดของยีน ดอร์เนอร์ และแอคเซพเตอร์ ในรูปที่ 4.1 4.3 4.5 และ 4.7 มาแสดงให้เห็นภาพของบริเวณต่างๆได้ชัดเจนยิ่งขึ้นโดยนำมาสร้างเป็นเป็นซีเควนโลโก้ (sequence Logo) ซึ่งซีเควนโลโก้นี้เป็นวิธีที่นิยมใช้ในงานที่เกี่ยวข้องกับการวิเคราะห์สายนิวคลีโอไทด์โดยเป็นการสร้างภาพแสดงการพบเบสที่ตำแหน่งต่างๆเพื่อช่วยให้สามารถเห็นรูปแบบการเรียงตัวของเบสณ ตำแหน่งต่างๆ ได้ชัดเจนยิ่งขึ้น



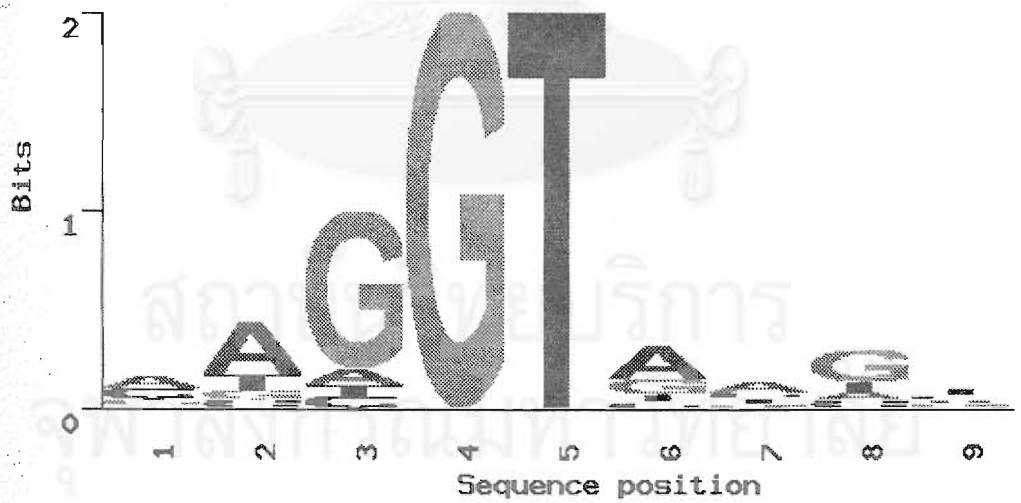
รูปที่ 4.9 ซีเควนโลโก้แสดงจุดเริ่มต้นของยีนข้าว (ATG)

รูปที่ 4.9 แสดงให้เห็นรูปแบบของบริเวณรอบ ๆ จุดเริ่มต้นของยีน จะพบเบส C กับเบส G มากกว่าที่จะพบเบส A และเบส T ส่วนในตำแหน่งที่ไม่มีสัญลักษณ์ใดเลยในซีเควนโลโก้ เช่น ตำแหน่งที่ 1 2 4 และ 5 จะหมายถึงการที่เบสต่างๆมีอัตราการเกิดพอ ๆ กัน



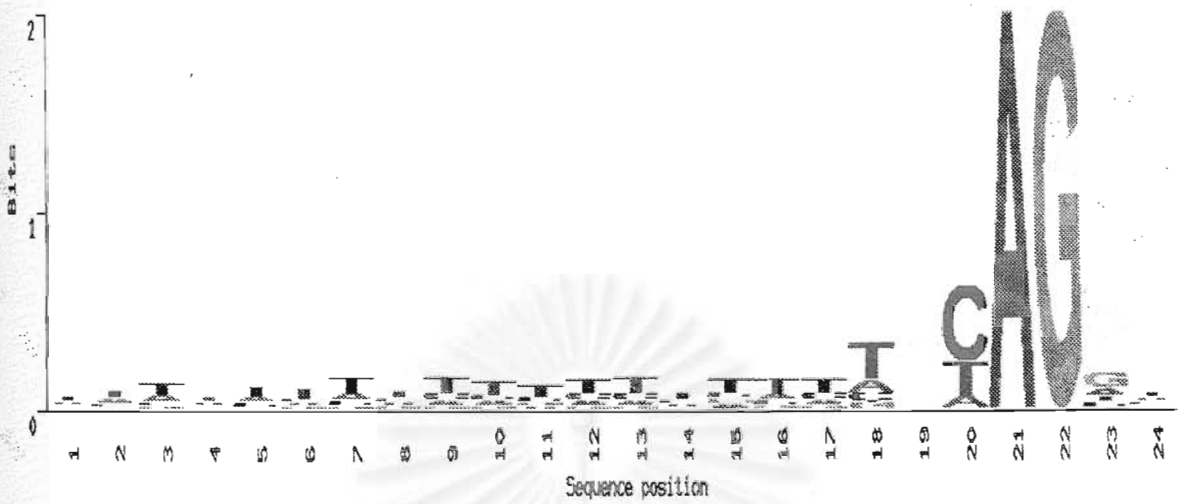
รูปที่ 4.10 ซีเควนโลโก้แสดงจุดสิ้นสุดของยีนข้าว (TAG TAA และ TGA)

รูปที่ 4.10 แสดงว่าบริเวณรอบ ๆ จุดสิ้นสุดของยีน มีอัตราการพบเบสต่างๆ พอกัน



รูปที่ 4.11 ซีเควนโลโก้แสดงดอร์เนอร์ของยีนข้าว (GT)

รูปที่ 4.11 แสดงบริเวณรอบ ๆ ดอร์เนอร์ของยีนข้าวพบว่ามีรูปแบบที่พบบ่อย 2 รูปแบบ คือ AAGGTAAGT หรือ CAGGTAAGT



รูปที่ 4.12 ซีเควนโลโก้แสดงแอคเซพเตอร์ของยีนข้าว (AG)

รูปที่ 4.12 แสดงให้เห็นว่าบริเวณก่อนที่จะพบแอคเซพเตอร์จะมีการพบเบส T มากกว่าเบสชนิดอื่น ๆ

สำหรับข้อมูลที่นำมาใช้ในการสร้างกราฟและซีเควนโลโก้นั้นสามารถดูได้ที่ภาคผนวก ข

ผลที่ได้จากการหาค่าทางสถิติของโคดอนในสายนิวคลีโอไทด์ข้าว

จากการนำข้อมูลสายนิวคลีโอไทด์จำนวน 217 สาย มาใช้เป็นข้อมูลสอนทำให้สามารถได้เรียนรู้ลักษณะเฉพาะบางประการของข้าว ดังนี้

1. ขนาดความยาวเฉลี่ยของเอ็กซอนในยีนข้าวจะมีความยาวประมาณ 219.9 คู่เบส (bp)
2. ขนาดความยาวเฉลี่ยของอินทรอนในยีนข้าวจะมีความยาวประมาณ 296.6 คู่เบส
3. อินทรอนที่มีขนาดสั้นที่สุดมีความยาวประมาณ 57 คู่เบส
4. ในอินทรอนโดยเฉลี่ยจะมี CG content ประมาณ 36.8 %
5. ในเอ็กซอนโดยเฉลี่ยจะมี CG content ประมาณ 55.5 %

เนื่องจากยีนของสิ่งมีชีวิตต่างชนิดกันจะประกอบไปด้วยโคดอนในปริมาณที่แตกต่างกัน ดังนั้น จากข้อมูลยีนข้าวทั้งหมดจึงได้นำมาหาความถี่ในการเกิดโคดอนแบบต่าง ๆ ของยีนข้าว ซึ่งจะได้ค่าตามตารางที่ 4.1

ตารางที่ 4.1 ความถี่ของโคดอนทั้ง 64 โคดอน ในสายนิวคลีโอไทด์ข้าวที่ใช้เป็นข้อมูลสอน

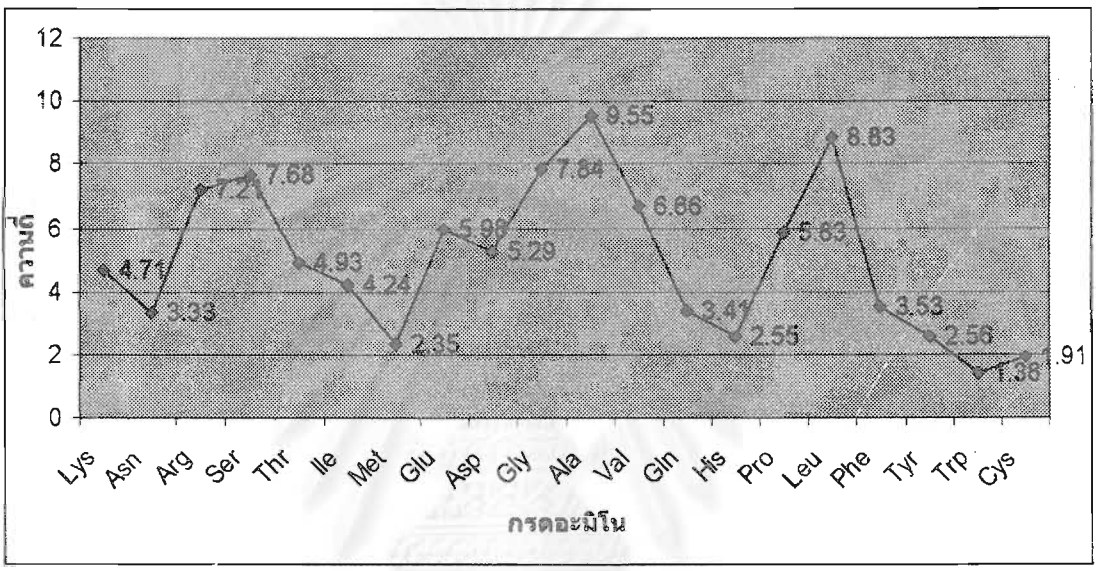
Codon	Aa	%	Codon	Aa	%	Codon	Aa	%	Codon	Aa	%
AAA	Lys	1.53	GAA	Glu	2.05	CAA	Gln	1.31	TAA	*	0.07
AAG	Lys	3.18	GAG	Glu	3.93	CAG	Gln	2.10	TAG	*	0.08
AAC	Asn	1.91	GAC	Asp	2.90	CAC	His	1.45	TAC	Tyr	1.59
AAT	Asn	1.42	GAT	Asp	2.39	CAT	His	1.10	TAT	Tyr	0.97
AGA	Arg	1.01	GGA	Gly	1.57	CGA	Arg	0.71	TGA	*	0.12
AGG	Arg	1.56	GGG	Gly	1.73	CGG	Arg	1.42	TGG	Trp	1.38
AGC	Ser	1.61	GGC	Gly	3.08	CGC	Arg	1.74	TGC	Cys	1.30
AGT	Ser	0.83	GGT	Gly	1.46	CGT	Arg	0.77	TGT	Cys	0.61
ACA	Thr	1.11	GCA	Ala	1.69	CCA	Pro	1.38	TCA	Ser	1.14
ACG	Thr	1.19	GCG	Ala	2.72	CCG	Pro	1.84	TCG	Ser	1.24
ACC	Thr	1.59	GCC	Ala	3.21	CCC	Pro	1.29	TCC	Ser	1.67
ACT	Thr	1.04	GCT	Ala	1.93	CCT	Pro	1.32	TCT	Ser	1.19
ATA	Ile	0.83	GTA	Val	0.66	CTA	Leu	0.74	TTA	Leu	0.57
ATG	Met	2.35	GTG	Val	2.45	CTG	Leu	2.09	TTG	Leu	1.39
ATC	Ile	2.03	GTC	Val	2.06	CTC	Leu	2.59	TTC	Phe	2.27
ATT	Ile	1.38	GTT	Val	1.49	CTT	Leu	1.45	TTT	Phe	1.26

โดยข้อมูลในตารางจะแสดงรหัสโคดอนทั้ง 64 แบบ (Codon) กรดอะมิโนที่โคดอนตัวนั้นๆใช้สื่อความหมายแทน (Aa) และเปอร์เซ็นต์ของการพบโคดอนนั้นในยีนข้าว (%) สำหรับในช่อง Aa ที่มี * อยู่จะหมายถึงโคดอนตัวนั้นๆ ไม่ได้ใช้สื่อแทนกรดอะมิโนตัวใด ๆเลย แต่จะใช้สื่อความหมายแทนตัวสิ้นสุดของยีน หรือ Stop Codon

โคดอนที่พบมากในยีนข้าว คือ GAG ซึ่งสามารถแปลงเป็นกรดอะมิโนกลูตามีน (Glutamine) โดยมีความถี่ที่พบประมาณร้อยละ 3.93 ส่วนโคดอนที่พบน้อยที่สุดในยีนข้าว คือ

TGA TAG และ TAA ซึ่งเป็นโคดอนที่ใช้ในการบอกจุดสิ้นสุดของยีนเท่านั้น ไม่สามารถแปลงเป็นกรดอะมิโนได้โดยมีค่าความถี่ที่พบตามลำดับดังนี้ คือ 0.12 0.08 และ 0.07

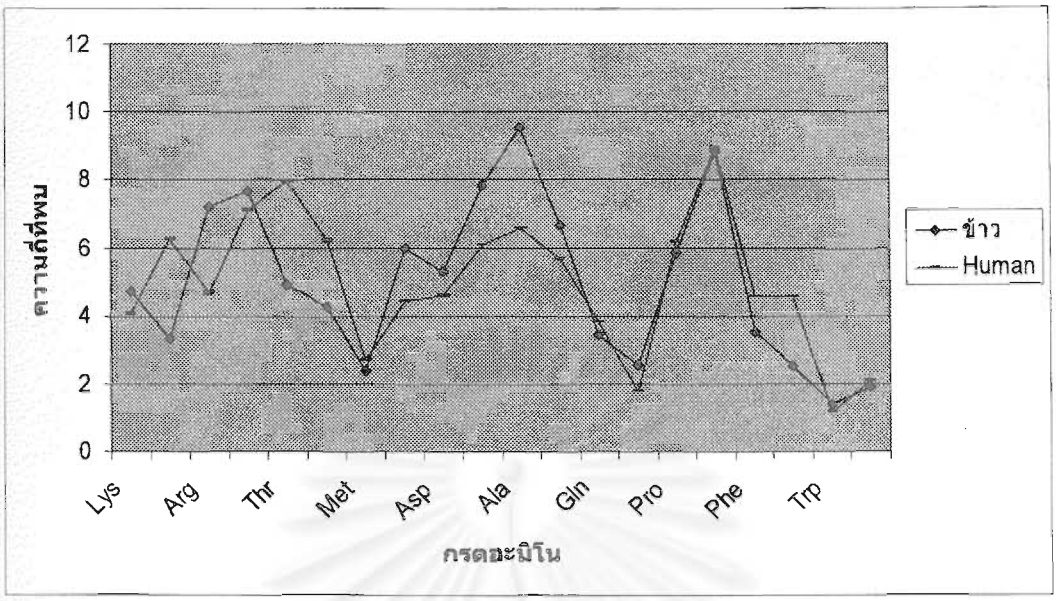
ข้อมูลในตารางที่ 4.1 สามารถหาความถี่ในการพบกรดอะมิโนชนิดต่างๆ ในยีนข้าวและนำมาสร้างกราฟแสดงความถี่ในการพบกรดอะมิโนทั้ง 20 ชนิดในยีนข้าวได้ดังรูปที่ 4.13



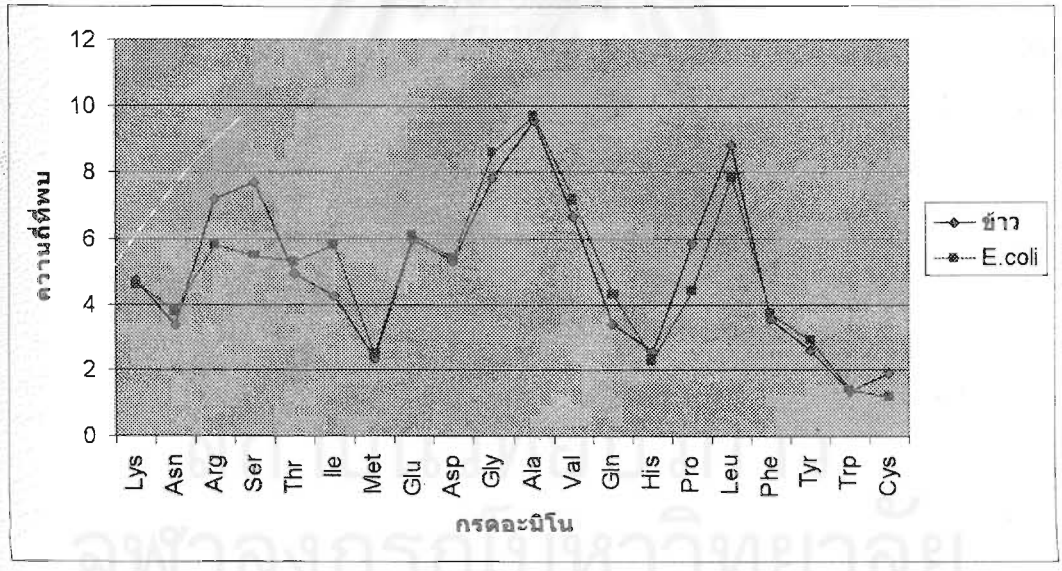
รูปที่ 4.13 ความถี่ของกรดอะมิโนชนิดต่างๆในยีนข้าว

กราฟในรูปที่ 4.13 แสดงให้เห็นว่ากรดอะมิโนที่พบมากในยีนข้าว คือ กรดอะมิโนอลานิน (Alanine) โดยมีความถี่ 9.55 เปอร์เซ็นต์ ซึ่งจะเกิดจากโคดอน GCC ประมาณ 3.21 เปอร์เซ็นต์ GCG ประมาณ 2.72 เปอร์เซ็นต์ GCT ประมาณ 1.93 เปอร์เซ็นต์ และ GCA ประมาณ 1.69 เปอร์เซ็นต์

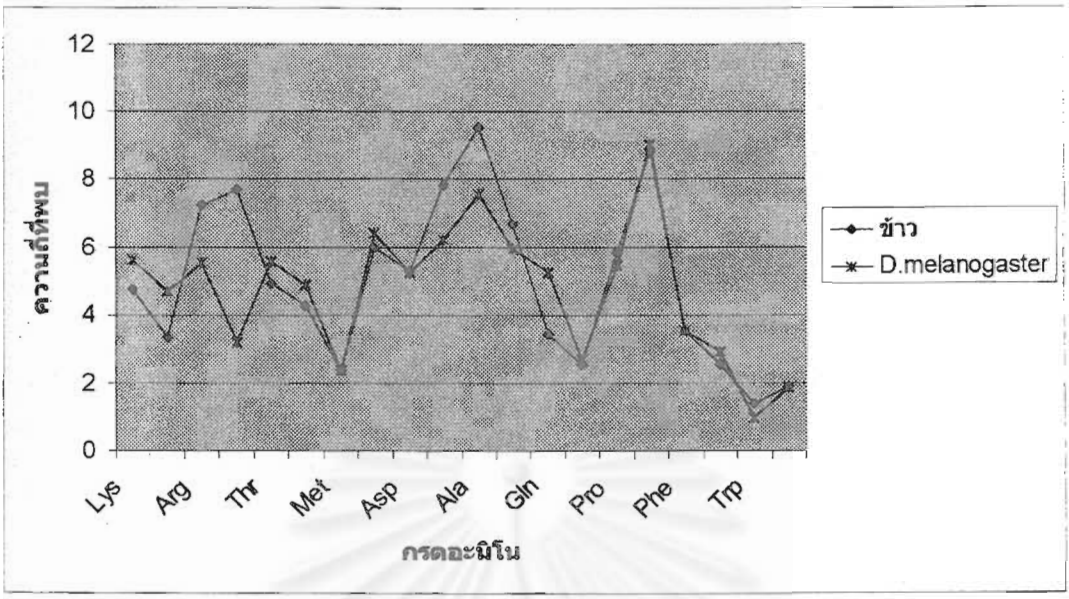
เมื่อนำข้อมูลความถี่ของการพบกรดอะมิโนในยีนข้าวมาเปรียบเทียบกับการพบกรดอะมิโนในสิ่งมีชีวิตชนิดต่างๆ จะได้ผลดังรูปที่ 4.14 – 4.19



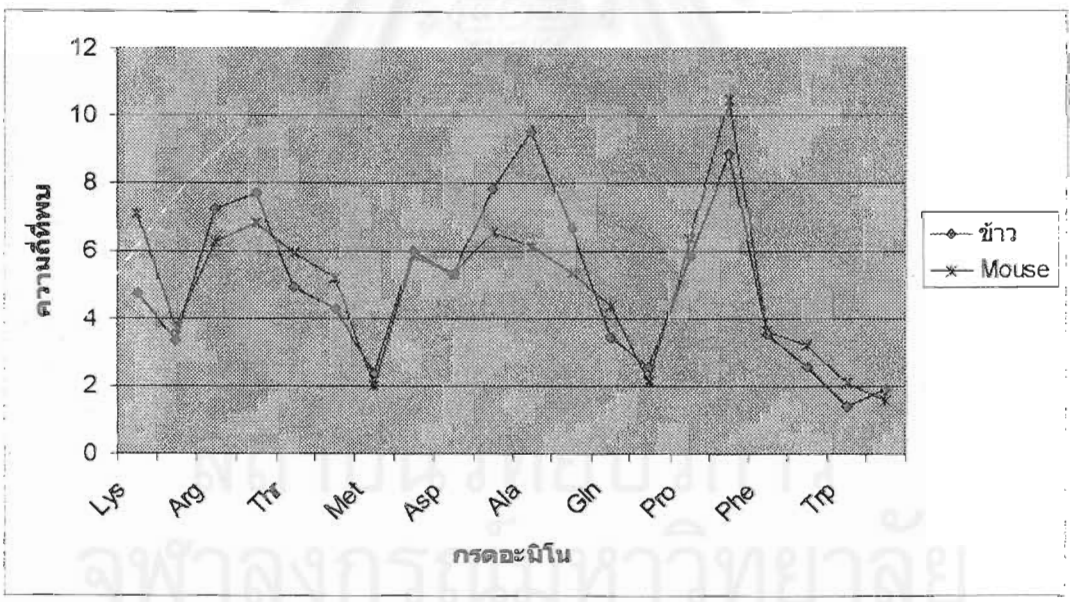
รูปที่ 4.14 เปรียบเทียบการพบกรดอะมิโนในยีนข้าวกับยีนมนุษย์



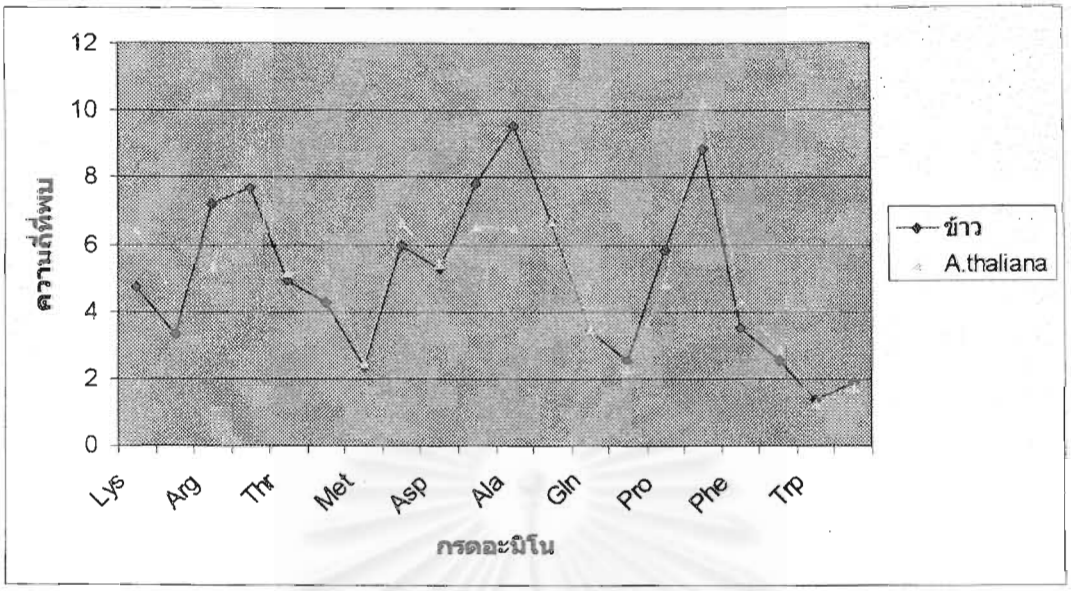
รูปที่ 4.15 เปรียบเทียบการพบกรดอะมิโนในยีนข้าวกับยีนแบคทีเรีย E.coli



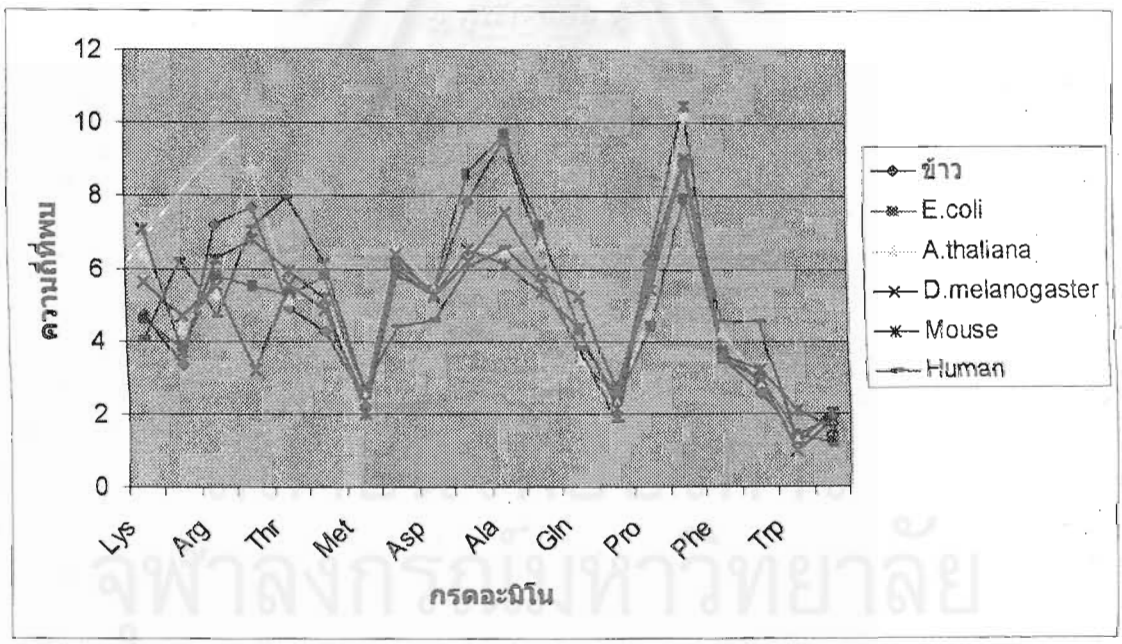
รูปที่ 4.16 เปรียบเทียบการพบกรดอะมิโนในยีนข้าวกับยีนแมลงวันผลไม้



รูปที่ 4.17 เปรียบเทียบการพบกรดอะมิโนในยีนข้าวกับยีนหนู



รูปที่ 4.18 เปรียบเทียบการพบกรดอะมิโนในยีนข้าวกับยีนพืชขนาดเล็กชนิดหนึ่ง



รูปที่ 4.19 เปรียบเทียบการพบกรดอะมิโนในยีนข้าวกับสิ่งมีชีวิตชนิดต่างๆ

กราฟในรูปที่ 4.19 แสดงการเปรียบเทียบการพบกรดอะมิโนในยีนของสิ่งมีชีวิตชนิดต่างๆเทียบกับที่พบในยีนข้าว จะพบว่ามีกรดอะมิโนบางตัวที่สิ่งมีชีวิตส่วนใหญ่มีปริมาณการพบใกล้เคียงกัน เช่น กรดอะมิโนฮิสทีดีน (Histidine) โพรลีน (Proline) ลิวซีน (Leucine) และ

ทริปโตเฟน (Tryptophan) และ ซิสทีน (Cystine) แต่เมื่อพิจารณาเข้าไปในส่วนของโคดอนที่ทำให้เกิดกรดอะมิโนเหล่านี้ พบว่าในสิ่งมีชีวิตต่างชนิดกันก็จะมีอัตราการเกิดกรดอะมิโนโดยใช้โคดอนไม่เหมือนกันเลย เช่น กรดอะมิโนที่ชื่อโพรลีน ในยีนข้าวกรดอะมิโนชนิดนี้ส่วนใหญ่จะเกิดจากโคดอน CCG (1.83 %) ส่วนในยีนมนุษย์นั้นกรดอะมิโนโพรลีนส่วนใหญ่เกิดจากโคดอน CCA (2.06 %) ส่วนในยีนแมลงวันผลไม้กรดอะมิโนโพรลีนส่วนใหญ่มาจากโคดอน CCC (1.8 %) และในยีนของพืชชนิด A.thaliana ส่วนใหญ่ได้มาจากโคดอน CCT (1.9 %) จะเห็นว่ากรดอะมิโนชนิดเดียวกันแต่เกิดมาจากโคดอนไม่เหมือนกันเลย

สำหรับแบคทีเรีย E.coli นั้นถึงแม้ว่าจะมีลักษณะการพบกรดอะมิโนต่างๆ ใกล้เคียงกับยีนข้าว แต่จริงๆแล้วสิ่งมีชีวิตในประเภทแบคทีเรียนั้นมีความแตกต่างอย่างมากเมื่อเทียบกับสิ่งมีชีวิตอื่นๆ ที่นำมาเปรียบเทียบด้วย เนื่องจากจากสิ่งมีชีวิตที่นำมาเปรียบเทียบทั้งหมดนี้เป็นประเภทยูคาริโอตทั้งหมดจึงมีเอ็กซอนและอินทรอนอยู่ภายในยีน ในขณะที่แบคทีเรียนั้นเป็นสิ่งมีชีวิตประเภทโปรคาริโอต ดังนั้นภายในยีนของแบคทีเรียจึงไม่มีการแบ่งแยกเป็นเอ็กซอนและอินทรอน

ผลการทดสอบโปรแกรม

ในการทดสอบโปรแกรม RGF จะใช้ข้อมูลทดสอบจำนวน 20 สายนิวคลีโอไทด์ ซึ่งรวมแล้วทั้งหมด 547 ยีน ที่เก็บจากฐานข้อมูล GenBank ซึ่งเป็นสายนิวคลีโอไทด์ข้าวที่ได้หายีนไว้แล้วนำมาทดสอบหาค่า Sensitivity และค่า Specificity ผลที่ได้จากการทดสอบแสดงดังตารางที่ 4.2

ตารางที่ 4.2 ผลการทดสอบโปรแกรม

สายนิวคลีโอไทด์ข้าว	ขนาด	จำนวนยีน	ค่า Sensitivity	ค่า Specificity
AP002865	139399	31	0.817	0.876
AP002866	169232	36	0.685	0.864
AP002867	146480	31	0.774	0.827
AP002870	122412	23	0.620	0.715
AP002818	146807	26	0.794	0.769
AP002537	142405	30	0.747	0.751
AP002863	193666	31	0.769	0.813
AP002843	150984	29	0.648	0.707
AP002862	148117	28	0.805	0.851
AP002913	143985	25	0.706	0.712
AP002744	156160	36	0.741	0.795
AP002881	111135	20	0.836	0.800
AP003053	58769	15	0.833	0.862
AP002871	145481	31	0.811	0.846
AP002524	158750	27	0.802	0.819
AP002872	159453	32	0.795	0.806
AP002899	145859	31	0.772	0.886
AP002909	143646	29	0.809	0.832
AP002953	164586	24	0.855	0.896
AP003315	148266	22	0.873	0.771
เฉลี่ย	**144779.6**	**27.35**	**0.775**	**0.810**

โดยสรุปผลทดสอบโดยเฉลี่ยของโปรแกรม RGF จะได้ค่า Sensitivity ประมาณ 0.775 และ ค่า Specificity ประมาณ 0.81 สำหรับการเปรียบเทียบโปรแกรมค้นหายีนอื่นมาใช้ในการหายีนในสายนิวคลีโอไทด์ข้าวจะได้ค่า Sensitivity และ ค่า Specificity ดังแสดงในตารางที่ 4.3

ตารางที่ 4.3 ผลการนำโปรแกรมหายีนตัวอื่นมาทดสอบหายีนในสายนิวคลีโอไทด์ข้าว

โปรแกรม	ค่า Sensitivity	ค่า Specificity
Glimmer	0.624	0.622
GeneMark	0.534	0.701
Geneld	0.492	0.691
RGF	0.775	0.810

ตารางที่ 4.3 เป็นการนำเอาโปรแกรมค้นหายีนโปรแกรมอื่นๆ มาทดสอบหายีนในสายนิวคลีโอไทด์ข้าว ซึ่งโปรแกรมที่นำมาทดสอบนั้นเป็นโปรแกรมที่นำข้อมูลสายนิวคลีโอไทด์ที่ไม่ใช่สายนิวคลีโอไทด์ข้าวมาเป็นข้อมูลในการสอนโปรแกรม เช่น Glimmer นำข้อมูลของแบคทีเรีย E.coli และ ข้อมูลสายนิวคลีโอไทด์ของสิ่งมีชีวิตขนาดเล็ก เช่น หนู มาเป็นข้อมูลสอน โปรแกรม Geneld ใช้สายนิวคลีโอไทด์ของ แมลงวันผลไม้ชนิดหนึ่ง (สปีชีย์ Drosophila *melanogaster*) เป็นข้อมูลสอน เป็นต้น

จากผลการทดสอบกับสายนิวคลีโอไทด์ข้าวพบว่า โปรแกรม Glimmer ค้นหายีนในสายนิวคลีโอไทด์ข้าวโดยหาจำนวนยีนได้มากกว่าจำนวนยีนจริงๆ เช่น ในสายนิวคลีโอไทด์ AP002865 มียีนจำนวน 31 ยีน แต่จากโปรแกรม Glimmer สามารถค้นหายีนได้ทั้งหมด 59 ยีนทำให้ค่าทดสอบโปรแกรมมีผลต่ำ ส่วนโปรแกรม GeneMark และ Geneld นั้นหาจำนวนยีนได้น้อยกว่าจำนวนยีนจริงๆ อีกทั้งยังหาเอ็กซอนได้น้อยกว่าความเป็นจริงด้วย ซึ่งส่งผลให้ค่าทดสอบต่ำเช่นกัน ในขณะที่โปรแกรมค้นหายีนข้าว RGF นั้นจะสามารถหายีนได้ใกล้เคียงกับยีนจริง เช่น ในสายนิวคลีโอไทด์ AP002865 มียีนจำนวน 31 ยีน โปรแกรมจะหาได้ประมาณ 34 ยีน แต่ในส่วนของเอ็กซอนภายในยีนนั้นจะหาได้ขนาดสั้นไปบ้าง หรือบางทีก็ยาวไปกว่าในยีนจริงๆ

บทที่ 5

สรุปผลการวิจัย

สรุปผลการวิจัย

งานวิจัยนี้มีจุดประสงค์เพื่อพัฒนาโปรแกรมที่ใช้ในการค้นหายีนในสายนิวคลีโอไทด์ข้าว โดยใช้วิธีมาร์คอฟโมเดลลำดับที่ 5 โดยใช้เทคนิคของการค้นยีน 2 เทคนิค คือ วิธีหาบริเวณที่เกี่ยวข้องกับการควบคุมการแสดงออกของยีน (Search by Signal) และ วิธีหาคุณสมบัติทางสถิติของลำดับเบสในสายดีเอ็นเอ (Search by Content) จากการวิจัยสามารถสรุปผลการวิจัยได้ดังนี้

1. จากการหาค่าพารามิเตอร์ของบริเวณต่างๆที่เกี่ยวข้องกับการควบคุมการแสดงออกของยีน ทำให้ต้องเก็บข้อมูลความถี่ของการเกิดเบสต่างๆ รอบบริเวณจุดเริ่มต้น จุดสิ้นสุดของยีน ดอร์เนอร์ และแอคเซพเตอร์ โดยการเก็บข้อมูลสายนิวคลีโอไทด์ที่ใช้เป็นข้อมูลสอน จากการเก็บข้อมูลเหล่านี้ทำให้สามารถหารูปแบบที่พบมาก(Consensus sequences)ของบริเวณต่างๆ ได้ดังนี้

1.1 บริเวณดอร์เนอร์ของยีนข้าว ส่วนมากจะมีลำดับเบสรอบๆ GT คือ AAGGTAAGT หรือ CAGGTAAGT

1.2 บริเวณแอคเซพเตอร์ของยีนข้าวส่วนมากจะมีลำดับเบสรอบแอคเซพเตอร์ดังนี้ คือ TTTTTTTTTTTTTTTTTTTCAGGT

2. จากข้อมูลนิวคลีโอไทด์ข้าวที่ใช้เป็นข้อมูลสอนทำให้เรียนรู้ลักษณะเฉพาะบางประการของข้าว ดังนี้

2.1 ขนาดความยาวเฉลี่ยของเอ็กซอนในยีนข้าวประมาณ 219.9 คู่เบส

2.2 ขนาดความยาวเฉลี่ยของอินทรอนในยีนข้าวประมาณ 296.6 คู่เบส

2.3 อินทรอนที่มีขนาดสั้นที่สุดมีความยาวประมาณ 57 คู่เบส

2.4 ในอินทรอนโดยเฉลี่ยจะมี CG content ประมาณ 36.8 %

2.5 ในเอ็กซอนโดยเฉลี่ยจะมี CG content ประมาณ 55.5 %

3. โคดอนที่พบมากที่สุดในยีนข้าว คือ GAG ซึ่งใช้แปลงเป็นกรดอะมิโนกลูตามีน (Glutamine) โดยพบเป็นอัตราส่วนร้อยละ 3.93 จากยีนข้อมูลสอนทั้งหมด ส่วนโคดอนที่พบน้อยสุด มี 3 ตัว คือ TGA TAA และ TGA ซึ่งเป็นโคดอนที่ใช้ในการบอกจุดสิ้นสุดของยีนข้าว

4. ผลการทดสอบโปรแกรมค้นหายีนข้าว RGF โดยนำมาทดสอบหาค่า Sensitivity และค่า Specificity ผลที่ได้จากการทดสอบได้ค่าโดยเฉลี่ยดังนี้ ค่า Sensitivity เท่ากับ 0.775 และค่า Specificity เท่ากับ 0.810 ซึ่งค่าทั้งสองนี้ใช้ในการทดสอบโปรแกรมค้นหายีน โดยทั่วไปค่าทั้งสองนี้ยิ่งมีใกล้เคียง 1 ยิ่งแสดงให้เห็นถึงประสิทธิภาพของโปรแกรมค้นหายีน

5. เมื่อเปรียบเทียบกับโปรแกรมค้นหายีนโปรแกรมอื่นๆ แล้ว โปรแกรมค้นหายีนข้าว RGF มีประสิทธิภาพในการค้นหายีนในสายนิวคลีโอไทด์ข้าวได้ดีกว่าโปรแกรมอื่น

ปัญหาและข้อจำกัดที่ได้พบจากการวิจัย

1. เนื่องจากปัจจุบันข้อมูลสายนิวคลีโอไทด์ข้าวที่ทำการค้นหายีนไว้แล้วมีจำนวนไม่มากอีกทั้งข้อมูลที่มีอยู่ในฐานข้อมูล GenBank ก็ไม่สามารถนำมาใช้ได้ทั้งหมดเนื่องจากมีข้อมูลบางสายนิวคลีโอไทด์ที่มีข้อผิดพลาดจึงไม่สามารถมาใช้เป็นข้อมูลสอนได้ ดังนั้นถ้าหากมีข้อมูลสอนจำนวนมากเท่าใด ก็ยิ่งทำให้สามารถหายีนได้ใกล้เคียงมากขึ้น

2. ในการหาค่าพารามิเตอร์ของบริเวณที่เกี่ยวข้องกับการควบคุมการแสดงออกของยีน (Search by signal) นั้น ในงานวิจัยนี้ได้หาเพียง 4 บริเวณพื้นฐาน เช่น จุดเริ่มต้นของยีน จุดสิ้นสุดของยีน ดอร์เนอร์ และแอคเซพเตอร์ เนื่องมาจากความจำกัดในความรู้ทางด้านชีวะโมเลกุลของผู้วิจัย แต่ในความเป็นจริงแล้วในยีนยังมีบริเวณที่มีความสำคัญที่สามารถช่วยให้การค้นหายีนสามารถทำได้แม่นยำมากขึ้น เช่น บริเวณโปรโมเตอร์ บริเวณที่มีการเติม poly-A บริเวณที่มีลำดับเบสซ้ำ (Repetitive sequence) และ บริเวณCpG Islands

3. การทำนายนี้ยังมีข้อผิดพลาดอยู่ค่อนข้างมาก เนื่องจากมาร์คอฟโมเดลสามารถทำนายโดยอาศัยข้อมูลลำดับกรดอะมิโนในบริเวณใกล้เคียงเท่านั้น แต่ในธรรมชาติโครงสร้างของยีนอาจถูกรบกวนโดยปฏิสัมพันธ์ระหว่างกรดอะมิโนที่อยู่ห่างไกลกันมากในสายโปรตีน หรือแม้แต่กรดอะมิโนที่อยู่ในยีนอื่นด้วย

4. มาร์คอฟโมเดลยิ่งมีลำดับสูงขึ้นเท่าใดความแม่นยำจะมากขึ้น แต่ถ้ายิ่งเป็นโมเดลลำดับสูงขึ้นไปจำนวนของพารามิเตอร์ก็จะต้องเพิ่มมากขึ้น เช่น พารามิเตอร์ของดีเอ็นเอที่โมเดลลำดับที่ n ก็จะต้องมีพารามิเตอร์ทั้งหมด 4^{n+1} เป็นต้น ซึ่งทำให้จำเป็นต้องมีข้อมูลสอนจำนวน

มากๆ แต่เนื่องจากความจำกัดในเรื่องของจำนวนของข้อมูลสอนดังนั้นจึงเลือกใช้มาร์คอฟโมเดลลำดับที่ 5 ซึ่งถือเป็นโมเดลลำดับที่ให้ผลลัพธ์น่าเชื่อถือ และใช้ปริมาณข้อมูลไม่ต้องมากเท่าใดนักเมื่อเทียบกับมาร์คอฟโมเดลลำดับที่สูงขึ้นไป

ข้อเสนอแนะ

1. หาค่าพารามิเตอร์ของบริเวณที่เกี่ยวข้องกับการควบคุมการแสดงออกของยีน (Search by signal) เพิ่ม เนื่องจากยังมีบริเวณที่สำคัญที่สามารถช่วยให้การค้นหายีนสามารถทำได้แม่นยำมากขึ้น เช่น บริเวณโปรโมเตอร์ บริเวณที่มีการเติม poly-A บริเวณที่มีลำดับเบสซ้ำ (Repetitive sequence) หรือ บริเวณ CpG Islands

2. นำวิธีค้นหายีนอีกวิธีมาช่วยการค้นหายีน ซึ่งก็คือ วิธีการเทียบสายนิวคลีโอไทด์ใหม่กับสายนิวคลีโอไทด์ที่รู้จักแล้วหรือสายนิวคลีโอไทด์ที่หาโปรตีนไว้แล้วซึ่งจะมาจากฐานข้อมูลโปรตีน (Search by Sequence Similarity)

3. ทดลองนำหลักการอื่นๆ ทางคอมพิวเตอร์มาใช้ค้นหายีนข้าว เพื่อหาหลักการที่จะทำให้การค้นหายีนมีความถูกต้องยิ่งขึ้น

รายการอ้างอิง

ภาษาอังกฤษ

Borodovsky, M. and McIninch, J. Genemark: Parallel gene recognition for both DNA strands. Compute. Chem. 17(1993) : 123-113.

Chen, T. and Zhang, M.Q. Pombe: A Gene-Finding and Exon-Intron Structure Prediction System For Fission Yeast. Yeast.14 (1998) :701-710.

Dash, D. and Gopalakrishnan, V. Modeling DNA Splice Regions by Learing Bayesian Networks. the Ninth International Conference on Intelligent Systems for Molecular Biology, February. 2001.

Durbin, R., Eddy S.R., Krogh, A. and Mitchison, G. Biological Sequence Analysis Probabilistic Models of proteins and nucleic acids. Cambridge university Press , 1998.

Parra, G., Blanco, E. and Guigo, R. GeneID in Drosophila. Genome Research. University Pompeu Fabra, 2000.

Rastogi, S.C. and Sharma, V.N. Concepts in Molecular Biology. Rivised Edition. (n.p.) : New Age Internatinal, December 1995.

Reese, M.G., Hartzell, G., Harris, N.L., Ohler, U. and Lewis, S.E. Genome annotation assessment in Drosophila melanogaster. Genome Research, 2000.

Salzberg, S. L., Pertea, M., Delcher, A.L., Gardner, M.J. and Tettelin, H. Interpolated Markov models for eukaryotic gene finding. Genomics : 24-31,1999.

Wirth, A.I. A Plasmodium falciparum Genefinder. Department of Mathematics and Statistics, University of Melbourne, Parkville, 2000.

ภาคผนวก

ภาคผนวก ก

ข้อมูลสอน (Training set)

สายนิวคลีโอไทด์	ความยาว
AP003018	142268 bp
AP003046	140952 bp
AP003074	150379 bp
AP003210	144404 bp
AP003311	175565 bp
AP003504	154298 bp
AP003256	170020 bp
AP003235	148892 bp
AP002540	167029 bp
AP003045	168253 bp
AP003103	132713 bp
AP003233	137879 bp
AP003075	155031 bp
AP003140	133242 bp
AP003076	161563 bp
AP003244	149610 bp
AP003263	190721 bp
AP003213	150022 bp
AP003793	151303 bp
AP003199	142320 bp
AP003204	156393 bp
AP003560	157987 bp
AP003337	112664 bp
AP003215	154137 bp
AP003255	161004 bp

สายนิวคลีโอไทด์	ความยาว
AP003237	162776 bp
AP000815	142418 bp
AP000836	190014 bp
AP000837	151294 bp
AP000969	168655 bp
AP001072	145684 bp
AP001278	161266 bp
AP001366	146081 bp
AP001383	154378 bp
AP001551	151773 bp
AP001550	143209 bp
AP001859	150594 bp
AP001800	176935 bp
AP002092	176349 bp
AP002093	151806 bp
AP002094	148985 bp
AP002480	160769 bp
AP002483	167405 bp
AP002521	146335 bp
AP002522	163526 bp
AP002523	163095 bp
AP002525	139152 bp
AP002804	50487 bp
AP002538	137462 bp
AP002835	147803 bp
AP002539	150150 bp
AP002860	134673 bp
AP002868	141079 bp
AP002869	158723 bp

สายนิวคลีโอไทด์	ความยาว
AP002861	134967 bp
AP002747	147857 bp
AP003106	166421 bp
AP003144	136351 bp
AP002902	159242 bp
AP003047	141983 bp
AP002970	139201 bp
AP003105	158815 bp
AP002897	156425 bp
AP002971	152396 bp
AP002914	167587 bp
AP003020	159749 bp
AP003021	168258 bp
AP003275	167230 bp
AP003277	173555 bp
AP003023	132470 bp
AP003578	138376 bp
AP002541	145576 bp
AP002972	161290 bp
AP003294	141476 bp
AP003143	140229 bp
AP003417	146812 bp
AP003282	135295 bp
AP003285	162853 bp
AP000816	59843 bp
AP002746	168372 bp
AP001073	152237 bp
AP001080	151369 bp
AP002839	143969 bp

สายนิวคลีโอไทด์	ความยาว
AP001081	140812 bp
AP001633	175072 bp
AP002070	170402 bp
AP002836	146921 bp
AP002484	137174 bp
AP002819	154561 bp
AP002817	140825 bp
AP002486	149699 bp
AP002487	78600 bp
AP002816	158133 bp
AP002526	143515 bp
AP002910	172003 bp
AP003073	172541 bp
AP003048	157201 bp
AP003492	158084 bp
10A19I	99587 bp
AP000492	147174 bp
AP000570	157903 bp
AP001539	168976 bp
AP002481	141111 bp
AP002482	187835 bp
AP002743	179714 bp
AP002820	137332 bp
AC069158	155154 bp
AP000366	35064 bp
AP000367	126038 bp
AC079830	129655 bp
AC082644	158714 bp
AY013245	50000 bp

สายนิวคลีโอไทด์	ความยาว
AC084319	152883 bp
AC079736	135932 bp
AC084320	178161 bp
AC087851	111882 bp
AC079887	168761 bp
AC082645	143681 bp
AC084406	140044 bp
AC079853	152423 bp
AC087181	142711 bp
AC084404	156929 bp
AC087797	122497 bp
AC084831	167288 bp
AC084295	89172 bp
AC084282	128017 bp
AC087852	149375 bp
AC084380	126164 bp
AC084767	163580 bp
AC091123	129778 bp
AC091247	115393 bp
AP000615	154128 bp
AC090485	159636 bp
H0212B02	118969 bp
H0423H10	76529 bp
H0806H05	102515 bp
H0811E11	59348 bp
AC051634	135789 bp
AC016781	151269 bp
AC026758	144798 bp
AC068924	152172 bp

สายนิวคลีโอไทด์	ความยาว
AC023240	131983 bp
AC073867	82909 bp
AC037425	134058 bp
AC051633	138825 bp
AC078840	130059 bp
AC069145	135216 bp
AC026815	147452 bp
AC074283	156654 bp
AC069324	174752 bp
AC073166	142114 bp
AC018727	139999 bp
AC060755	172427 bp
AC021891	144593 bp
AC079685	131599 bp
AC084763	141307 bp
AC079936	143163 bp
AC078829	163278 bp
AC037426	135509 bp
AC022352	139398 bp
AC024594	178692 bp
AC025783	178024 bp
AC022457	162614 bp
AC079128	151163 bp
AC078891	109198 bp
AC079890	132927 bp
AC025296	165394 bp
AC079029	137613 bp
AC020666	158550 bp
AC034258	117157 bp

สายนิวคลีโอไทด์	ความยาว
AC078944	164679 bp
AC079634	161250 bp
AC090487	63206 bp
AC069300	147117 bp
AC051624	148263 bp
AC074354	146391 bp
AC074105	146836 bp
AC084884	134553 bp
AC079632	142381 bp
AC079037	157450 bp
AC091238	127068 bp
AC021893	151343 bp
AC091734	155843 bp
AC087192	113606 bp
AC090483	145417 bp
AC090441	139468 bp
AC091665	118101 bp
AF128457	70311 bp
AC025098	151359 bp
AC037197	141017 bp
H0711G06	148608 bp
OSA243961	74658 bp
OSA245900	124321 bp
OSB6015	131287 bp
OST17804	123337 bp
AP003044	152941 bp
AP002542	156266 bp
AB023482	156054 bp
AB026295	170371 bp

สายนิวคลีโอไทด์	ความยาว
AP000391	155634 bp
AP000399	154180 bp
AF149806	33178 bp
AC084218	164899 bp
AC079022	151589 bp
AC087551	145012 bp
AF111709	52684 bp
AF111710	43422 bp
AC079038	129838 bp
AP000364	137354 bp
AC078839	168192 bp
AC083945	147706 bp
AC080019	149654 bp
AF119222	77605 bp
AP003414	158174 bp
AP003610	141966 bp
AP002069	151978 bp
AP001552	150120 bp
AP001389	157519 bp

ภาคผนวก ข

ข้อมูลความถี่ของเบสรอบบริเวณต่างๆ และเมตริกซ์น้ำหนัก

ตำแหน่ง	A	C	G	T
-8	28.27	25.74	20.67	25.32
-7	25.32	24.47	26.58	23.63
-6	27.85	20.25	35.86	16.03
-5	21.94	36.71	22.78	18.56
-4	27.43	25.31	29.53	17.72
-3	32.91	12.65	38.82	15.61
-2	26.58	38.81	17.30	17.30
-1	16.03	46.84	27.43	9.70
0	100	0	0	0
1	0	0	0	100
2	0	0	100	0
3	18.99	16.88	56.96	7.17
4	24.89	38.82	21.52	14.77
5	14.77	23.20	42.62	19.41

ตารางที่ ข.1 ความถี่ของการพบเบสรอบจุดเริ่มต้นของยีน

ตำแหน่ง	A	C	G	T
-8	-0.134	0.446	0.0512	-0.2504
-7	-0.128	0.3383	0.1529	-0.298
-6	-0.158	0.1	0.7525	-0.8417
-5	-0.339	0.836	0.249	-0.808
-4	-0.0718	0.443	0.1424	-0.5415
-3	0.1525	-0.6024	0.8855	-0.95
-2	-0.34	0.9239	-0.165	-0.614
-1	-0.63	0.87	0.043	-1.24
0	0	-9999	-9999	-9999
1	-9999	-9999	-9999	0
2	-9999	-9999	0	-9999
3	-0.597	-0.2315	1.2918	-1.974
4	-0.358	0.9435	0.1222	-0.93
5	-0.869	0.2999	0.7014	-0.526

ตารางที่ ข.2 เมตริกซ์น้ำหนักของจุดเริ่มต้นของยีน

ตำแหน่ง	A	C	G	T
-5	29.38	24.29	20.90	25.42
-4	21.47	28.25	25.42	24.86
-3	24.86	15.82	39.55	19.77
-2	25.42	28.81	20.90	24.86
-1	20.90	31.64	23.73	23.73
0	0	0	0	100
1	53.67	0	46.33	0
2	67.23	0	32.77	0
3	25.42	19.21	26.55	28.81

ตารางที่ ข.3 ความถี่ของการพบเบสรอบจุดสิ้นสุดของยีน

ตำแหน่ง	A	C	G	T
-5	-0.1448	0.475	0.0109	-0.2045
-4	-0.5248	0.5511	0.4437	-0.325
-3	-0.395	-0.147	1.1138	-0.6725
-2	-0.306	0.8477	-0.1146	-0.267
-1	-0.59	0.6928	0.2716	-0.304
0	-9999	-9999	-9999	0
1	-0.303	-9999	0.4556	-9999
2	-0.123	-9999	0.291	-9999
3	-0.315	0.211	0.4808	-0.182

ตารางที่ ข.4 เมตริกซ์น้ำหนักของจุดสิ้นสุดของยีน

ตำแหน่ง	A	C	G	T
-3	36.73	34.07	20.35	8.85
-2	61.50	9.73	11.50	17.26
-1	7.96	5.75	80.53	5.75
0	0	0	100	0
1	0	0	0	100
2	52.65	9.73	26.11	11.50
3	44.69	20.80	15.49	19.03
4	15.04	12.39	54.42	18.14
5	18.14	25.66	16.37	39.82

ตารางที่ ข.5 ความถี่ของการพบเบสรอบดอร์เนอร์

ตำแหน่ง	A	C	G	T
-3	0.4746	0.5787	-0.1973	-1.6316
-2	1.0703	-1.097	-0.698	-0.855
-1	-1.708	-1.382	1.835	-2.6619
0	-9999	-9999	0	-9999
1	-9999	-9999	-9999	0
2	1.294	-1.105	-0.2	-1.262
3	0.829	-0.0846	-0.651	-0.5815
4	-0.8537	-0.7774	1.4019	-0.772
5	-0.6203	0.277	-0.3832	0.4289

ตารางที่ ข.6 เมตริกซ์น้ำหนักของดอร์เนอร์

ตำแหน่ง	A	C	G	T
-22	26.97	20	16.06	36.97
-21	26.97	20.61	13.94	38.48
-20	27.27	16.67	13.94	42.12
-19	24.24	21.21	16.67	37.88
-18	26.36	21.82	12.12	39.69
-17	20.91	22.73	14.85	41.52
-16	21.82	20	12.42	45.76
-15	19.39	23.94	15.76	40.91
-14	18.18	20.91	14.24	46.67
-13	18.48	20.91	14.85	45.76
-12	16.36	22.73	16.36	44.55
-11	19.69	23.94	11.52	44.85
-10	17.58	23.33	12.73	46.36
-9	20.30	25.45	15.15	39.09
-8	16.36	20.61	16.97	46.06
-7	20.91	20.61	12.42	46.06
-6	18.18	24.85	12.42	44.55
-5	16.67	16.36	10.61	56.36
-4	22.12	19.39	32.73	25.76
-3	10.30	60.91	1.52	27.27
-2	100	0	0	0
-1	0	0	100	0
0	22.12	15.45	46.67	15.76
1	22.73	20	17.88	39.40

ตารางที่ ข.7 ความถี่ของการพบเบสรอบแอคเซพเตอร์

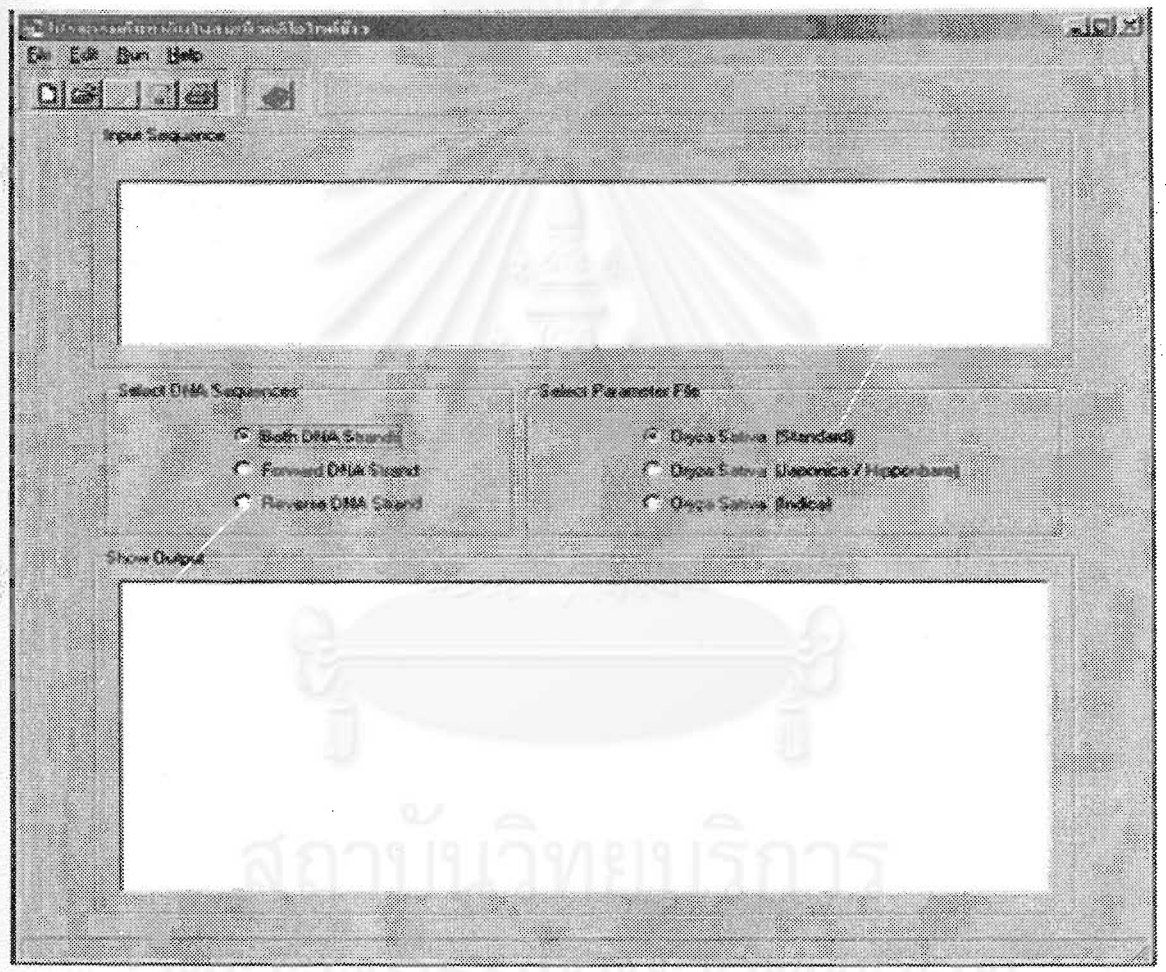
ตำแหน่ง	A	C	G	T
-22	0.0504	-0.1458	-0.4734	0.324
-21	-0.114	0.1186	-0.5208	0.274
-20	-0.0812	-0.1744	-0.6159	0.4433
-19	-0.2806	-0.0159	0.263	0.38
-18	-0.1386	0.0536	-0.7648	0.4351
-17	-0.4213	0.228	-0.5588	0.4343
-16	-0.3333	0.01233	-0.7051	0.4968
-15	-0.6067	0.4011	-0.3683	0.3543
-14	-0.7963	0.058	-0.5963	0.7997
-13	-0.638	0.2264	-0.635	0.595
-12	-0.8928	0.21545	-0.3259	0.593
-11	-0.5884	0.5434	-1.0074	0.5417
-10	-0.7022	0.11591	-0.6521	0.6342
-9	-0.6678	0.51708	-0.4607	0.4255
-8	-0.8724	0.1122	-0.3321	0.6371
-7	-0.5106	0.2401	-0.8216	0.5802
-6	-0.8232	0.4017	-0.8103	0.7098
-5	-0.7959	-0.0712	-1.1471	0.8915
-4	-0.5094	0.0183	0.5155	-0.0393
-3	-1.5602	1.6953	-3.9956	0.0339
-2	0	-9999	-9999	-9999
-1	-9999	-9999	0	-9999
0	-0.4043	-0.3283	0.9543	-0.7894
1	-0.524	0.2766	-0.1679	0.3586

ตารางที่ ข.8 เมตริกซ์น้ำหนักของแอคเซพเตอร์

ภาคผนวก ค

โปรแกรมค้นหายีนข้าว

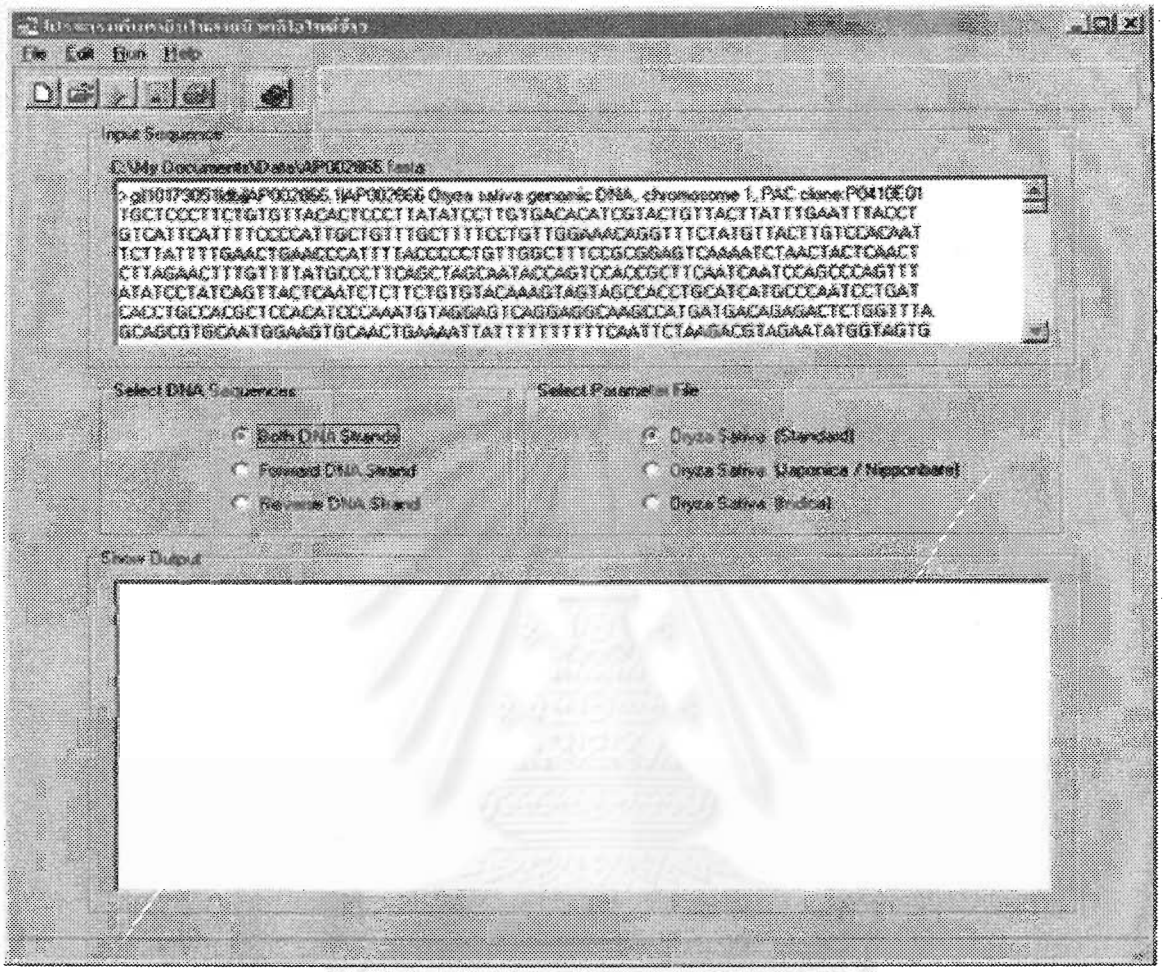
เมื่อเปิดโปรแกรมค้นหายีนข้าวขึ้นมาจะมีลักษณะส่วนติดต่อกับผู้ใช้ ดังรูปที่ ค.1 โดยโปรแกรมจะประกอบด้วยส่วนหลักๆ 3 ส่วน คือ ส่วนแสดงข้อมูลนำเข้า ส่วนแสดงผลลัพธ์ และส่วนเมนูต่างๆ



รูปที่ ค.1 โปรแกรมค้นหายีนข้าว

โปรแกรมนี้จะรับไฟล์ข้อมูลนำเข้าที่อยู่ในรูปแบบ FASTA เท่านั้น (*.FASTA, *.FAS) เมื่อต้องการค้นหายีนในสายนิวคลีโอไทด์ใดให้เลือกที่เมนู File และเลือก Open Sequence File หรือเลือกรูปOpen File ที่ ทูลบาร์ จากนั้นจะมีไดอะล็อกบ๊อกซ์ (Dialog Box) ขึ้นมาให้เลือกไฟล์ข้อมูลสายนิวคลีโอไทด์ เมื่อเลือกข้อมูลสายนิวคลีโอไทด์ที่ต้องการหายีนแล้ว ข้อมูลจะถูกแสดงทางส่วน

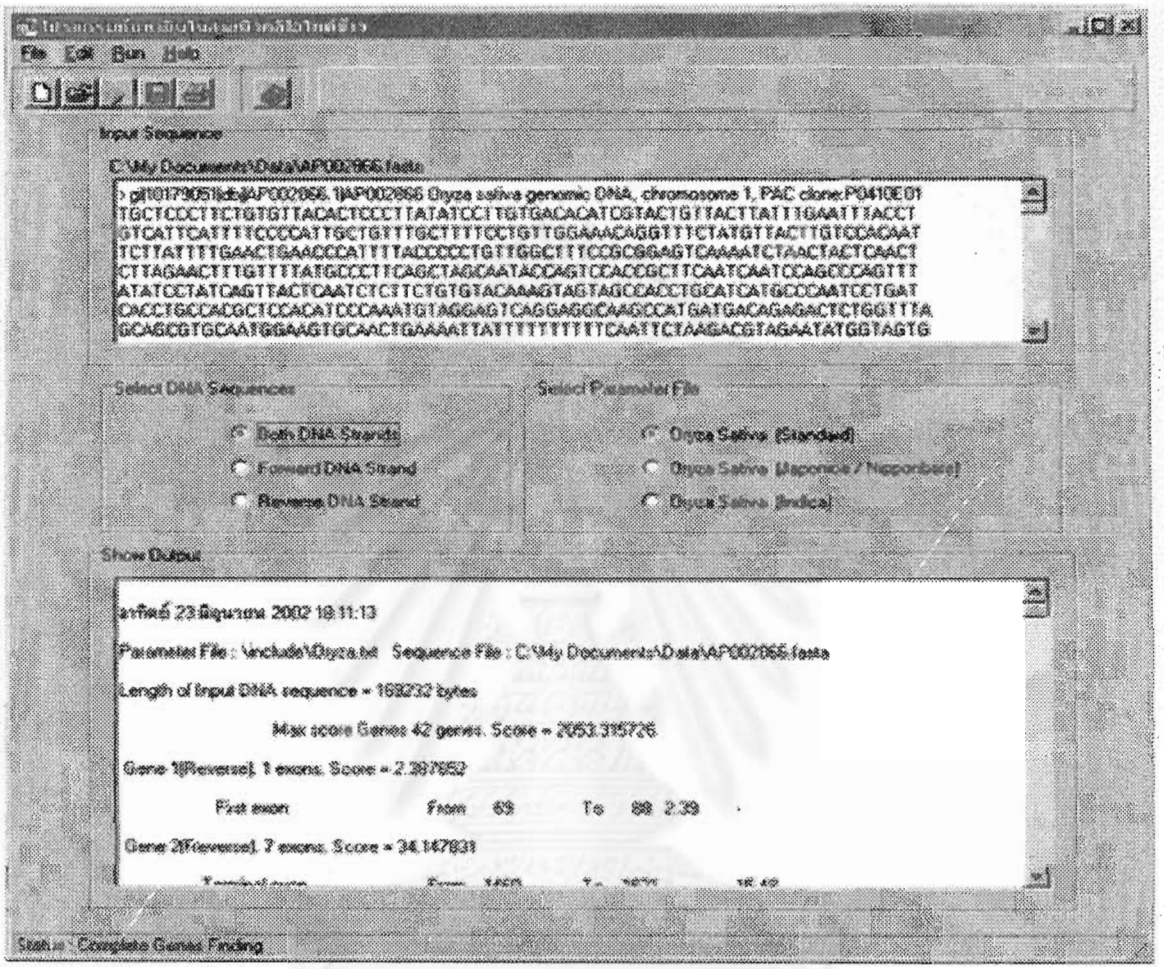
Input Sequence ดังรูป ค.2 และเมื่อเลือกสายนิวคลีโอไทด์แล้วโปรแกรมจึงจะสามารถสั่งรันโปรแกรมได้



รูปที่ ค.2 การนำข้อมูลเข้าสู่โปรแกรม

เมื่อเลือกข้อมูลเข้าได้แล้ว จากนั้นให้เลือกสายดีเอ็นเอที่ต้องการหายีน โดยเลือกที่ส่วน Select DNA sequence ซึ่งโดยปกติโปรแกรมจะเลือกไว้ให้แล้วทั้งสองสายคือ สายปกติ (Forward Strand) กับ สายตรงข้าม (Complementary strand) ด้วย แต่หากต้องการเปลี่ยนก็สามารถเลือกได้ว่าต้องการหายีนในสายดีเอ็นเอสายใด จากนั้นให้เลือกพารามิเตอร์ที่ต้องการใช้ โดยจะมีให้เลือกสามไฟล์พารามิเตอร์ แต่เมื่อเริ่มเปิดโปรแกรมขึ้นมา โปรแกรมจะเลือกพารามิเตอร์มาตรฐาน(Standard)ไว้ให้แล้ว ไฟล์พารามิเตอร์มาตรฐานนั้นเกิดจากการสอนด้วยข้อมูลจากทุกสายนิวคลีโอไทด์ในกลุ่มข้อมูลสอน แต่หากต้องการเปลี่ยนไฟล์พารามิเตอร์ก็สามารถทำได้ และสั่งให้โปรแกรมทำงานโดยเลือก เมนู Run และเลือก Run program ผลที่ได้จากการหายีนจะแสดงที่ส่วน Show output ดังรูปที่ ค.3 ซึ่งหากต้องการเซฟข้อมูลเก็บไว้เป็นไฟล์ก็ให้

เลือกที่ file และเลือก save output หรือ เลือกที่ทูลบาร์ แต่หากต้องการพิมพ์ข้อมูลออกทางเครื่องพิมพ์ให้เลือก File และ เลือก print output หรือเลือกรูปเครื่องพิมพ์บนทูลบาร์แทนได้



รูปที่ ค.3 การแสดงผลลัพธ์ของโปรแกรม

เมื่อต้องการเริ่มต้นการทำงานของโปรแกรมใหม่ โดยลบข้อมูลทั้งหมดที่แสดงอยู่ในโปรแกรมให้อยู่ในรูปแบบเหมือนเพิ่งเปิดโปรแกรมขึ้นมาใหม่ ทำได้โดยการเลือกที่ File และ Clear all หรือเลือกที่ทูลบาร์โดยเลือก New บนทูลบาร์

รูปแบบ FASTA (FASTA format)

รูปแบบ FASTA มีรูปแบบเป็นดังนี้

```
>gi|14090356|dbj|AP003233.3|AP003233 Oryza sativa genomic DNA, chromosome 1, PAC clone:P0037C04
TTACTTTTGTTTTCCTTTCGCCAATACTCTTCTGGACACAGGGCAAGTTTAGACCCTATTTAGATGGGAC
TAAATCCCTATCATATCGAATGTTTGGATACTAATTATAAATATTAAACGTGGACTATTAATAAACCCAT
TCTATAACCCAGAACTAATTCGCGAGACGAATCTATTGAGCCTAATTAATCAATGATTAGCCCATGTGAT
GCTACAGTAAACATACGCTAGTTATGGATTAATTAGGCTTAAAAAATTTATCACGCGAATTAGCTTCTTA
TTTATGTAATTAGTTTTATAAATAGTCTATGTTTAATACTCCAAATTCATCCGTTATGACATGGACTAAA
GTTTAGTCTTTGGATCCAAACTTATTCTTCAAACTTTCAACTTTTCCATCACATCAAAACTATCCTACAC
ACACAAACTTTCAATTTTTCCTTTCATATCGTTCTAACTTCAACCAAACTTCTAATTTTAACGTGAACTA
AACACACCCACAGTACACATGGGAAAGAAGCTGGTTCCAGGGTTAAAGGTCCTGCACTGCAGGCGTTGGC
TGCGCTGGACAGAAAAAAAAAATGACCAAACCAAAAACCGAACTGAATTAACAGAGACCGAGAAATTCGAT
CATCAGTTGTGACTAATTGAATTTAACACTGTCTTTTAAAACGAAATAACTGAGCTGACCGAATCGATGT
```

รูปที่ ง.1 ตัวอย่าง FASTA format

FASTA format เริ่มต้นโดยบรรทัดแรกจะมีเครื่องหมายมากกว่า (>) ตามด้วยคำบรรยาย จำนวน 1 บรรทัด เรียกว่า FASTA definition line ซึ่งจะบอกเกี่ยวกับ Accession number หรือ GenBank Identifier และ Locus number ส่วนบรรทัดต่อไปเป็นลำดับซึ่งมีความกว้างไม่เกิน 80 คอลัมน์ (คือรวมตัวอักษรและช่องว่าง)

สำหรับอักษรย่อที่ใช้แทนนิวคลีโอไทด์หรือกรดอะมิโนต้องเป็นไปตามมาตรฐาน IUB/IUPAC ดังนี้

นิวคลีโอไทด์ :

A แทน adenosine	M แทน A C (amino)
C แทน cytidine	S แทน G C (strong)
G แทน guanine	W แทน A T (weak)
T แทน thymidine	B แทน G T C
U แทน uridine	D แทน G A T
R แทน purine	H แทน A C T

Y แทน pyrimidine

K แทน GT (keto)

- แทน gap

V แทน G C A

N แทน A G C T (any)

กรดอะมิโน :

A แทน alanine

B แทน aspartate or asparagine

C แทน cystine

D แทน aspartate

E แทน glutamate

F แทน phenylalanine

G แทน glycine

H แทน histidine

I แทน isoleucine

K แทน lysine

L แทน leucine

M แทน methionine

N แทน asparagine

P แทน proline

Q แทน glutamine

R แทน arginine

S แทน serine

T แทน threonine

U แทน selenocysteine

V แทน valine

W แทน tryptophan

Y แทน tyrosine

Z แทน glutamate or glutamine

X แทน any

* แทน translation stop

- แทน gap of indeterminate

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวปวีณา เลิศอำไพพร เกิดวันที่ 12 เมษายน พ.ศ. 2522 ที่จังหวัดกรุงเทพมหานคร สำเร็จการศึกษาระดับปริญญาตรีสถิติศาสตรบัณฑิต สาขาเทคโนโลยีสารสนเทศเพื่อธุรกิจ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2542 และ เข้าศึกษาต่อในหลักสูตรวิทยาศาสตร์มหาบัณฑิต สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย เมื่อ พ.ศ. 2543