การทำนายตำแหน่งสไปลไซต์โดยใช้ต้นไม้การตัดสินใจและแบบจำลองมาร์คอฟ

นายสืบกุล กาญจนสุกร์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาการคณนา ภาควิชาคณิตศาสตร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2550

SPLICE SITE PREDICTION USING A DECISION TREE AND MARKOV MODELS

Mr.Suebkul Kanchanasuk

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science Program in Computational Science

Department of Mathematics

Faculty of Science

Chulalongkorn University

Academic Year 2007

หัวข้อวิทยานิพนธ์	การทำนายตำแหน่งสไปลไซต์โดยใช้ต้นไม้การตัดสินใจและแบบจำลองมาร์คอฟ
โดย	นายสืบกุล กาญจนสุกร์
สาขาวิชา	วิทยาการคณนา
อาจารย์ที่ปรึกษา	รองศาสตราจารย์ ดร. ไพศาล นาคมหาชลาสินธุ์
อาจารย์ที่ปรึกษาร่วม (ถ้ามี)	ผู้ช่วยศาสตราจารย์ ศิริสรรพ เหล่าหะเกียรติ

คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิทยาศาสตร์

(ศาสตราจารย์ ดร.สุพจน์ หารหนองบัว)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(รองศาสตราจารย์ ดร. พีระพนธ์ โสพัศสถิตย์)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(รองศาสตราจารย์ ดร.ไพศาล นาคมหาชลาสินธุ์)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม

(ผู้ช่วยศาสตราจารย์ ศิริสรรพ เหล่าหะเกียรติ)

.....กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.รัฐ พิชญางกูร)

สืบกุล กาญจนสุกร์ : การทำนายตำแหน่งสไปลไซต์โดยใช้ต้นไม้การตัดสินใจและแบบจำลองมาร์คอฟ. (SPLICE SITE PREDICTION USING A DECISION TREE AND MARKOV MODELS) อ. ที่ปรึกษา : รองศาสตราจารย์ ดร.ไพศาล นาคมหาชลาสินธุ์, อ.ที่ปรึกษาร่วม : ผู้ช่วยศาสตราจารย์ ศิริสรรพ เหล่าทะเกียรติ 63 หน้า.

ในงานวิทยานิพนธ์ฉบับนี้ เราได้พัฒนาโปรแกรมทำนายตำแหน่งสไปลไซต์บนยีนของมนุษย์ โดยใช้ต้นไม้การตัดสินใจและแบบจำลองมาร์คอฟเพื่อคำนวณคะแนนที่จะใช้ตัดสินว่าลำดับนิวคลีโอไทด์ใดๆ ที่กำหนดให้มีแนวโน้มเป็นสไปลไซต์มากเพียงใด เราใช้ต้นไม้การตัดสินใจเพื่อแบ่งกลุ่มลำดับนิวคลีโอไทด์จากความขึ้นแก่กันแบบ χ^2 และยังใช้แบบจำลองมาร์คอฟอันดับหนึ่งเพื่อคำนวณคะแนนที่ระบุความน่าจะเป็นว่าสไปลไซต์นั้นเป็นจริงหรือเท็จ โปรแกรมนี้มีชื่อว่า "Enhanced GeneSplicer" ซึ่งได้ขยายแนวคิดของโปรแกรม GeneSplicer ด้วยการให้โอกาสแก่กลุ่มสไปลไซต์เท็จอีกครั้ง โดยจะนำมาจำแนกใหม่ และเราจะหาสิ่งที่เหมาะที่สุดของกระบวนการทั้งหมด แม้ว่าเวลาที่ใช้ในการคำนวณจะมากขึ้น แต่เราได้ความแม่นยำในการทำนายที่สูงขึ้น สำหรับค่า false negative 0.2% ในโดเนอร์ไซต์ โปรแกรมสามารถลดค่า false positive จาก 25.5% เหลือ 18.48% ในขณะที่แอกเซพเตอร์ไซต์ลดลงจาก 38.30% เหลือ 34.51%

ภาควิชา คณิตศาสตร์ ลายมือชื่อนิสิต.....

สาขาวิชา วิทยาการคณนา ลายมือชื่ออาจารย์ที่ปรึกษา.....

ปีการศึกษา 2550 ลายมือชื่ออาจารย์ที่ปรึกษาร่วม.....

4772518423 : MAJOR COMPUTATIONAL SCIENCE

KEY WORD: SPLICE SITE PREDICTION / MARKOV MODELS / DECISION TREE

SUEBKUL KANCHANASUK : SPLICE STIE PREDICTION USING DECISION TREE AND MARKOV MODELS. THESIS ADVISOR : ASSOC. PROF. PAISAN NAKMAHACHALASINT, Ph.D., THESIS COADVISOR : ASST.PROF. SIRISUP LAOHAKIAT, 63 pp.

In this thesis, we will develop a splice site prediction program on human genes. The program will use decision trees and Markov models to calculate scores that can be used decide how likely a given portion on a nucleotide sequence is a splice site. Decision trees will be used to classify nucleotide sequences by the χ^2 dependence for each position, while the first-order Markov models compute scores that signify the probabilities of a splice site being true or false. The program is named "Enhanced GeneSplicer" as it extends the concept of the GeneSplicer program by giving a second chance to the false sites – they will be reclassified and we seek for the optimality of the whole process. Despite the increased computational time of Enhanced GeneSplicer, we obtained an improvement on the accuracy of the prediction. With 0.2% of false negatives, the percentage of false positives in donor sites drops from 25.5% to 18.48%, while that of the acceptor sites decreases from 38.30% to 34.51%.

Department	Mathematics	Student's signature	Suebkul Kanchanasuk
Field of study	Computational Science	Advisor's signature	Paisan Nakmahachalasint
Academic year	2007	Co-advisor's signature	Sirisup Laohakiat

กิตติกรรมประกาศ

งานวิจัยชิ้นนี้สำเร็จลุล่วงไปได้ด้วยดีโดยความกรุณาอย่างยิ่งของ รองศาสตราจารย์ ดร. ไพศาล นาคมหาชลาสินธุ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้คอยให้ความรู้และให้คำแนะนำในการปรับปรุงแก้ไขข้อบกพร่องต่าง ๆ อันเป็นประโยชน์และมีคุณค่ายิ่งต่อการทำวิทยานิพนธ์ ตลอดจนได้เสียสละเวลาให้คำชี้แนะในทุก ๆ เรื่อง รวมถึงคอยให้กำลังใจเสมอ ผู้วิจัยขอกราบขอบพระคุณไว้ ณ โอกาสนี้ด้วย

ขอกราบขอบพระคุณ ผู้ช่วยศาสตราจารย์ ศิริสรรพ เหล่าหะเกียรติ อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ที่ได้กรุณาเสียสละเวลา ให้ความรู้และคำแนะนำต่าง ๆ ในการแก้ไขข้อบกพร่องต่าง ๆ ในการทำวิจัย ทำให้วิทยานิพนธ์ฉบับนี้มีความสมบูรณ์มากยิ่งขึ้น

ขอกราบขอบพระคุณ รองศาสตราจารย์ ดร.พีระพนธ์ โสภัศสถิตย์ ประธานกรรมการสอบ ผู้ช่วยศาสตราจารย์ ดร. รัฐ พิชญางกูร กรรมการสอบวิทยานิพนธ์ ที่กรุณาสละเวลาอันมีค่าร่วมสอบวิทยานิพนธ์ รวมถึงคณาจารย์สาขาวิชาวิทยาการคณนาทุกท่านที่มอบความรู้และประสบการณ์ต่าง ๆ อันมีค่าอย่างยิ่งแก่ผู้วิจัย

ขอขอบพระคุณโครงการพัฒนาอาจารย์สาขาขาดแคลน สาขาวิชาคณิตศาสตร์ ในสังกัดของมหาวิทยาลัยนเรศวร ที่สนับสนุนทุนการศึกษาและทุนการวิจัย รวมถึงวิทยาเขตสารสนเทศพะเยาที่สนับสนุนการศึกษาต่อในครั้งนี้

ขอกราบขอบพระคุณ อาจารย์อารยา วิวัฒน์วานิช อาจารย์ประจำภาควิชาคณิตศาสตร์ มหาวิทยาลัยบูรพา ที่ช่วยตรวจสอบและแก้ไขข้อบกพร่องต่าง ๆ รวมทั้งชี้แนะแนวทางการทำวิทยานิพนธ์ฉบับนี้ให้มีความถูกต้องและสมบูรณ์มากยิ่งขึ้น และขอขอบคุณนาย วัชรศักดิ์ ศิริเสรีวรรณ รวมถึงพี่ ๆ เพื่อน ๆ และน้อง ๆ สาขาวิชาวิทยาการคณนาทุกท่าน รวมถึงเจ้าหน้าที่ประจำสาขา ที่ได้ช่วยให้วิทยานิพนธ์ฉบับนี้สำเร็จลงได้

ความสำเร็จในทุกประการของผู้วิจัยจะมีขึ้นไม่ได้หากไม่ได้รับกำลังใจ ความช่วยเหลือและการสนับสนุนส่งเสริมจากครอบครัวอันเป็นที่รักยิ่ง ขอกราบขอบพระคุณคุณพ่อสมเกียรติ กาญจนสุกร์ และคุณแม่อรวรรณ กาญจนสุกร์ รวมถึงครอบครัววีระชัยที่เป็นกำลังใจ และสนับสนุนช่วยเหลือ และนางสาวจันทนี วีระชัย สำหรับความรักอันเปี่ยมล้นด้วยความปรารถนาดีและคอยดูแลเอาใจใส่ด้วยดีเสมอมา

สารบัญ

หน้า

หน้า

สารบัญตาราง

ตารางที่ หน้า

ตารางที่ หน้า

สารบัญภาพประกอบ

ภาพประกอบที่ หน้า

บทที่ 1
บทนำ

ที่มาและความสำคัญของปัญหา

ชีวสารสนเทศศาสตร์ (Bioinformatics) [1] เป็นศาสตร์ที่ว่าด้วยการจัดเก็บข้อมูลทางชีววิทยาจำนวนมากให้เป็นระบบโดยอาศัยเทคโนโลยีด้านสารสนเทศหรือ IT (Information Technology) เข้ามาช่วยจัดการเพื่อให้สามารถวิเคราะห์ข้อมูลที่มีจำนวนมากได้พร้อมกันเพื่อตอบคำถามทางชีววิทยาซึ่งกระทำได้ยากหรือไม่อาจทำได้ในอดีตอันเนื่องจากเกินขีดความสามารถของมนุษย์ที่จะจดจำหรือวิเคราะห์เพื่อหาความสัมพันธ์ของข้อมูลจำนวนมาก ช่วยพัฒนา วิเคราะห์และสร้างแบบจำลองต่าง ๆ ในงานวิจัยทางชีววิทยา

การนำสารสนเทศมาประยุกต์ใช้เพื่อจัดการกับข้อมูลทางชีวภาพนั้น มีความจำเป็นอย่างยิ่ง เนื่องจากข้อมูลทางชีววิทยาในปัจจุบันได้มีการเผยแพร่ทางอินเตอร์เน็ต เพื่อให้นักวิจัยได้ค้นคว้า และพัฒนาข้อมูลเหล่านั้นให้มีประโยชน์ในวงกว้างมากขึ้น อีกทั้งเทคโนโลยีคอมพิวเตอร์ได้มีการพัฒนาให้มีศักยภาพสูงขึ้น ก็ได้เข้ามามีบทบาทในการพัฒนา ความสามารถ ประสิทธิภาพและประสิทธิผลในการค้นคว้าวิจัยมากขึ้น จึงเหมาะกับการนำมาเพื่อจัดการกับข้อมูลทางชีวภาพที่มีปริมาณมากได้เป็นอย่างดี

ชีวสารสนเทศศาสตร์ หรือ ชีววิทยาเชิงคอมพิวเตอร์ (Computational biology) [2] เกี่ยวข้องกับการใช้เทคนิคอันประกอบด้วย คณิตศาสตร์ประยุกต์ สนเทศศาสตร์ สถิติ วิทยาศาสตร์คอมพิวเตอร์ ปัญญาประดิษฐ์ เคมี และชีวเคมี เพื่อแก้ไขปัญหาทางชีววิทยาในระดับโมเลกุล งานวิจัยในด้านชีววิทยาเชิงคอมพิวเตอร์มีความคาบเกี่ยวกันกับงานวิจัยในทางชีววิทยา ซึ่งมีหลายแขนงให้ศึกษาประกอบด้วย การปรับแนวของลำดับ (Sequence alignment) การค้นหายีน (Gene finding) การรวบรวมยีน (Genome assembly) การปรับแนวของโครงสร้างโปรตีน (Protein structure alignment) การทำนายโครงสร้างของโปรตีน (Protein structure prediction) การทำนายการแสดงออกของยีน (Prediction of gene expression) ความสัมพันธ์ระหว่างโปรตีนกับโปรตีน (Protein-protein interaction) และ การสร้างแบบจำลองของวิวัฒนาการ (The modeling of evolution)

งานทางชีวสารสนเทศศาสตร์นั้น การค้นหายีนเป็นแขนงหนึ่งที่มีความสำคัญ เนื่องจากเป็นแหล่งข้อมูลที่ความสำคัญในการที่จะเรียนรู้และศึกษาพฤติกรรมต่าง ๆ ของเซลล์และสิ่งมีชีวิตว่ามีกระบวนการทำงานภายในเป็นอย่างไรและทำงานได้อย่างไร จึงนับเป็นสาขาที่น่าสนใจในการศึกษาค้นคว้าแขนงหนึ่ง

การค้นหายีนนั้นก็ยังมีอีกหลายแขนงให้ศึกษา ทั้งในส่วนของการค้นหายีนทั้งยีน คือ ค้นหาตำแหน่งเริ่มต้นและตำแหน่งสิ้นสุดของยีน หรือการค้นหาองค์ประกอบของยีน โดยในยีนของสิ่งมีชีวิตประเภทยูคาริโอต จะมีลำดับนิวคลีโอไทด์ที่ไม่ได้ใช้ในการสังเคราะห์โปรตีนแทรกอยู่ในยีนด้วย โดยจะมี เอนไซม์จำพวกหนึ่งที่ทำหน้าที่จดจำตำแหน่งของการตัดเอาส่วนเหล่านี้ออกไป ซึ่งโปรแกรมที่ใช้ในการค้นหายีนนั้นมีอยู่หลายโปรแกรม มีเทคนิคและวิธีในการค้นหาที่แตกต่างกันออกไป อาทิเช่น NNSplice, Genie, SpliceView, HSPL และ GeneSplicer [3] สำหรับ การค้นหาส่วนที่ใช้ในการสังเคราะห์โปรตีนของยีนนั้นในทางชีววิทยาสามารถค้นหาตำแหน่งได้อย่างแน่นอน แต่ติดที่เวลาในการทดลองนั้นนานมาก ดังนั้นหากมีโปรแกรมคอมพิวเตอร์ที่ช่วยในการลดจำนวนส่วนที่เป็นไปได้ที่จะทำการตัดและมาต่อของยีนนั้น ย่อมช่วยลดภาระในการหาของนักวิทยาศาสตร์ลงได้อย่างมากเป็นการประหยัดเวลาในการทดลองได้ทางหนึ่ง อีกทั้งการค้นหาส่วนที่ใช้ในการสังเคราะห์โปรตีนของยีนนั้นมีความสำคัญในทางวิทยาศาสตร์เป็นอย่างมากเพราะจะสามารถศึกษาพฤติกรรมของยีนและกลไกต่าง ๆ ที่เกิดขึ้นในร่างกายของมนุษย์ได้ดี และวิเคราะห์ได้ละเอียดมากขึ้นด้วยข้อมูลที่มีความถูกต้องและแม่นยำมากขึ้น ซึ่งจะช่วยให้มีความกว้างขวางในแขนงวิชานี้เพิ่มมากขึ้นไปอีก

วัตถุประสงค์ของการวิจัย

พัฒนาความสามารถในการทำนายของโปรแกรม GeneSplicer ในการลดความผิดพลาดในการทำนายตำแหน่งของสไปลไซต์ในยีนของมนุษย์ให้ผิดพลาดน้อยลง

ขอบเขตของการวิจัย

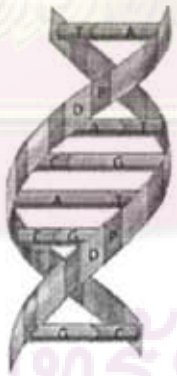
ในการทดลองครั้งนี้ เป็นการศึกษาสไปลไซต์ในยีนของมนุษย์ ดังนั้นผู้วิจัยจึงเลือกใช้ยีนที่มีสไปลไซต์อยู่จริงในทางชีววิทยาสำหรับเป็นข้อมูลในการเรียนรู้และทดสอบ ของการทดลอง

บทที่ 2
ทฤษฎีที่เกี่ยวข้องกับงานวิจัย

ในบทนี้จะกล่าวถึง ลักษณะโครงสร้างของดีเอ็นเอ (The structure of DNA) จีโนม และยีน (Genome and gene) กระบวนการสังเคราะห์โปรตีน (Protein synthesis) สไปลไซต์ (Splice sites) ต้นไม้การตัดสินใจ (Decision tree) MDD (Maximal dependence decomposition) และแบบจำลองมาร์คอฟ (Markov models) ซึ่งเป็นความรู้พื้นฐานที่ใช้ในงานวิจัย

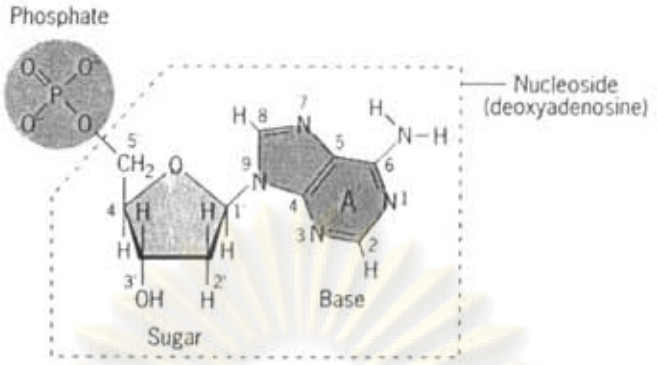
ลักษณะโครงสร้างของดีเอ็นเอ (The structure of DNA)

ในช่วงทศวรรษที่ 1950 ได้มีนักวิทยาศาสตร์จำนวนหนึ่ง จากประเทศสหรัฐอเมริกาและอังกฤษ ได้พยายามค้นคว้าเพื่อไขความลับเกี่ยวกับดีเอ็นเอ จนกระทั่งในปี ค.ศ. 1953 เจมส์ วัตสัน และฟรานซิส คริกค์ ของมหาวิทยาลัยแคมบริด ได้เริ่มอธิบายถึงโครงสร้างของดีเอ็นเอ (deoxyribonucleic acid) ว่ามีลักษณะเป็นสายเกลียวคู่ (double helix) ขมวดกันอยู่ ดังรูปที่ 2.1 โดยโมเลกุลของดีเอ็นเอแต่ละโมเลกุลประกอบไปด้วยหน่วยย่อย ๆ เรียกว่า หน่วยนิวคลีโอไทด์ หรือนิวคลีโอไทด์ (nucleotide unit) ซึ่งประกอบไปด้วย หมู่ฟอสเฟต (phosphate group) น้ำตาลดีออกซีไรโบส (deoxyribose sugar) ซึ่งประกอบด้วยน้ำตาลและสารประกอบไนโตรเจนที่เรียกว่า กรดนิวคลีอิก (nucleic acid) [4] ดังรูปที่ 2.2



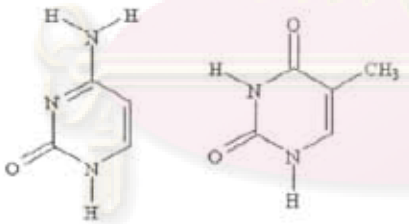
รูปที่ 2.1 โครงสร้างเกลียวคู่ของดีเอ็นเอ

(http://www.bartleby.com/61/indexillus10.html)

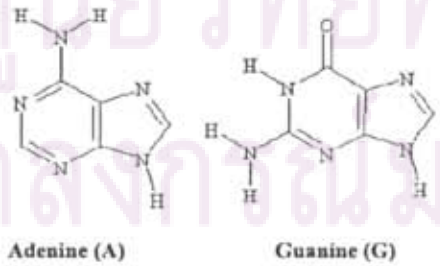


รูปที่ 2.2 โครงสร้างโมเลกุลของดีเอ็นเอ

กรดนิวคลีอิก (nucleic acid) ที่จับตัวอยู่บนดีเอ็นเอนี้ มี 2 แบบ คือแบบไพริมีดีน (Pyrimidine) ซึ่งมีโครงสร้างเป็นแบบ 1 วง มีอยู่ด้วยกัน 2 ชนิด คือ ไซโตซีน (Cytosine: C, c) และไธมีน (Thymine: T, t) และอีกแบบหนึ่งคือแบบพิวรีน (Purine) ซึ่งมีโครงสร้างเป็นแบบ 2 วง มีอยู่ด้วยกัน 2 ชนิด คือ อดีนีน (Adenine: A, a) และกัวนีน (Guanine: G, g) ดังในรูปที่ 2.3 โดยดีเอ็นเอจะเชื่อมต่อกันด้วยพันธะฟอสเฟต (phosphodiester bonds) ระหว่างน้ำตาลกับหมู่ฟอสเฟตจาก 3' ไป 5' การเชื่อมกันด้วยลักษณะนี้เมื่อมี**การอ่านนิวคลีโอไทด์ตามลำดับที่เกิดขึ้นนั้นจะเรียกว่าการอ่านจาก 5' ไป 3' (forwad strand)** สายนิวคลีโอไทด์ที่เชื่อมต่อกันเป็นสายยาวจะมีการสร้างสายนิวคลีโอไทด์อีกสายหนึ่ง โดยมีลักษณะสำคัญในการเชื่อมต่อสายนิวคลีโอไทด์ระหว่างกันคือ อดีนีนจะจับคู่กับไทมีน และกัวนีนจะจับคู่กับไซโตซีนเสมอ ซึ่งสายนิวคลีโอไทด์ 2 สายนี้เชื่อมกันด้วยพันธะโคเวเลนซ์ระหว่างกันเกิดเป็นเกลียวคู่ (double helix) [5]



โครงสร้างแบบไพริมีดีน (Pyrimidine)



โครงสร้างแบบพิวรีน (Purine)

รูปที่ 2.3 ชนิดของกรดนิวคลีอิกในดีเอ็นเอ

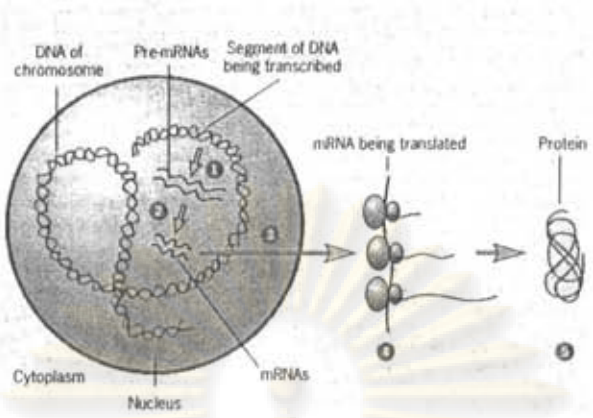
จีโนมและยีน (Genome and gene)

จีโนม คือ ข้อมูลทางพันธุกรรมทั้งหมดของสิ่งมีชีวิต ซึ่งมีเพียงหนึ่งเดียวในแต่ละสิ่งมีชีวิต ข้อมูลทางพันธุกรรมของมนุษย์ ซึ่งก็คือสายยาวของดีเอ็นเอเป็นคู่ที่พันขดกันอยู่บนโครโมโซมทั้ง 23 คู่ โดยข้อมูลทางพันธุกรรมที่เก็บอยู่บนสายยาวของดีเอ็นเอนั้น จะถูกจัดเก็บตามขนาดและลำดับของดีเอ็นเออย่างเฉพาะเจาะจงตามหน้าที่และจัดเก็บเป็นหมวดหมู่เพื่อใช้ในการสร้างโปรตีน ซึ่งเป็นวัตถุดิบที่ใช้สำหรับการทำงานต่าง ๆ ของเซลล์ของสิ่งมีชีวิต

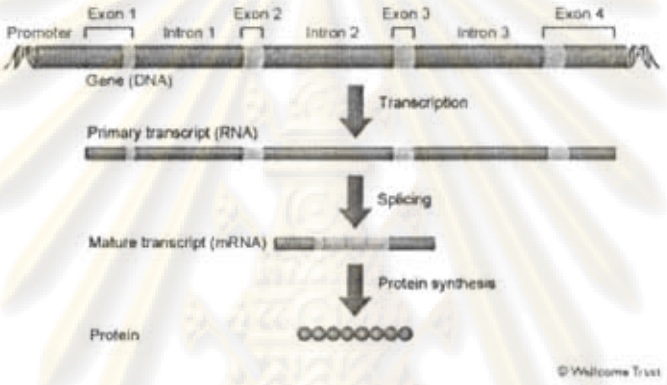
ยีนนั้นเป็นคำที่เริ่มมีการใช้โดย โยฮันสัน ในปีค.ศ. 1910 [6] ซึ่งใช้เรียกสิ่งที่เกิดจากการค้นพบของ เกรเกอร์ เมนเดล ซึ่งใช้เรียกลักษณะของแฟคเตอร์ที่เป็นตัวถ่ายทอดลักษณะต่างๆ จากรุ่นพ่อแม่ไปยังลูกหลานนั่นเอง ต่อมาได้มีการยืนยันความคิดของเกรเกอร์ เมนเดลโดย มอร์แกน (T.H. Morgan) จากการศึกษาการถ่ายทอดลักษณะต่าง ๆ ของแมลงหวี่ พบว่าลักษณะเฉพาะในการถ่ายทอดลักษณะทางพันธุกรรมซึ่งขณะนั้นเรียกว่า แฟคเตอร์ กับคำว่ายีนนั้นคือสิ่งเดียวกันนั่นเอง

กระบวนการสังเคราะห์โปรตีน (Protein synthesis)

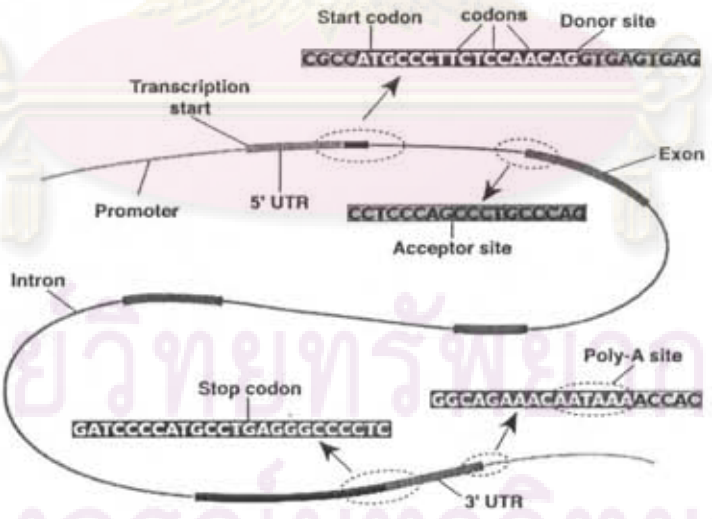
ในกระบวนการสร้างโปรตีน เริ่มจากยีนมีกระบวนการถอดรหัส (Transcription) จากลำดับนิวคลีโอไทด์ของยีนบนโครโมโซม (ขั้นตอนที่ 1 ในรูปที่ 2.4) เรียกสายนิวคลีโอไทด์ที่ได้จากการถอดรหัสว่า Pre-mRNAs (Primary transcript) ซึ่งเป็นการถอดรหัสของลำดับนิวคลีโอไทด์ต้นแบบออกมาให้เหมือนเดิม เปลี่ยนแปลงเพียงดีเอ็นเอไทมีนเป็นอาร์เอ็นเอ (RNA:Ribo nucleic acid) ชื่อยูราซิล (Uracil) หลังจากถอดรหัสแล้วจะมีการตัดเอาลำดับนิวคลีโอไทด์ที่ไม่ได้ใช้ในกระบวนการสังเคราะห์โปรตีนออก ซึ่งในกระบวนการนี้จะมีการนำลำดับนิวคลีโอไทด์ที่ได้มาตัดเอาลำดับนิวคลีโอไทด์ส่วนที่ไม่ใช้ออก โดยลำดับที่ไม่ได้ใช้ในการสังเคราะห์โปรตีนเรียกว่า **อินทรอน (Intron)** และลำดับที่ใช้ในการสังเคราะห์โปรตีนเรียกว่า **แอกซอน (Exon)** เรียกกระบวนการนี้ว่า การทำสไปลซิ่ง (Splicing) ดังรูปที่ 2.5 ซึ่งเกิดจากการทำหน้าที่ของ snRNA (small nuclear RNA) ทั้ง U1 U2AF U4 U5 และ U6 ที่ได้ทำการจดจำลักษณะและกระบวนการตัดนิวคลีโอไทด์ที่ไม่ได้ในการสังเคราะห์โปรตีนออก เราจะเรียกส่วนต่อกันระหว่างแอกซอนและอินทรอนว่าสไปลไซต์ (Splice sites) โดยสไปลไซต์ที่อ่านจากแอกซอนไปยังอินทรอนว่า**โดเนอร์ไซต์ (Donor site)** และเรียกสไปลไซต์ที่อ่านจากอินทรอนไปยังแอกซอนว่า **แอกเซพเตอร์ไซต์ (Acceptor site)** ดังรูปที่ 2.6



รูปที่ 2.4 แสดงการทำงานของยีนในการสร้างโปรตีน [5]



รูปที่ 2.5 กระบวนการสังเคราะห์โปรตีน



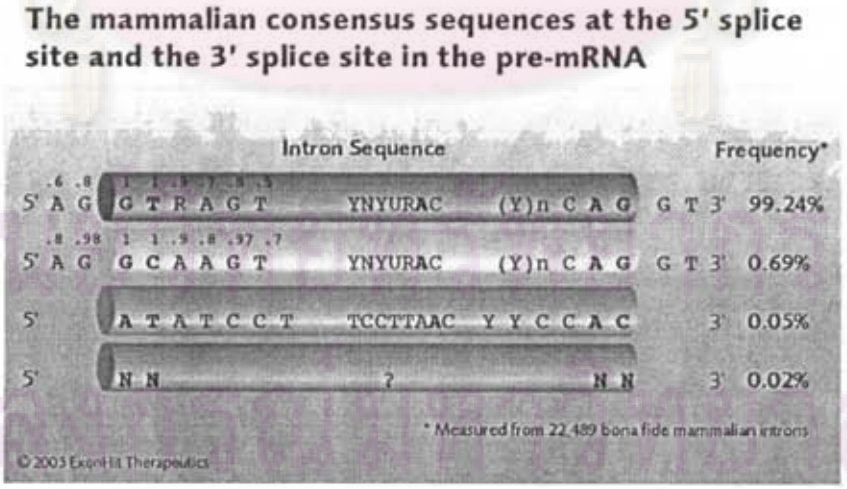
รูปที่ 2.6 แสดงส่วนประกอบต่าง ๆ ของยีน

(http://www.cs.cmu.edu/~epxing/research1.html)

ในกระบวนการสไปลซิ่งนั้น snRNA จะมีบทบาทอย่างมากในการจดจำตำแหน่งที่จะทำการตัดลำดับนิวคลีโอไทด์ ซึ่งจะมีจำนวนนิวคลีโอไทด์ในการจดจำอยู่แตกต่างกันไปตามชนิดของ snRNA เช่นใน U1 จะมีการจดจำลำดับนิวคลีโอไทด์ในส่วนของอินทรอนที่ 11 ตำแหน่ง และยังมี snRNA อีกหลายชนิดที่มีการจดจำแตกต่างกันไป ซึ่งหลังจากกระบวนการสไปลซิ่งแล้วจะได้ลำดับของอาร์เอ็นเอใหม่ คือ mRNAs (mature transcript) หลังจากนั้นจะนำ mRNA ที่ได้มาทำการแปลรหัส (Translation) เพื่อสร้างลำดับโปรตีนเพื่อใช้สำหรับกระบวนการต่าง ๆ ภายในเซลล์และในสิ่งมีชีวิตต่อไป

สไปลไซต์ (Splice sites)

ในการศึกษาความสัมพันธ์ต่าง ๆ ของสไปลไซต์ทั้งของโดเนอร์และแอคเซพเตอร์ว่ามีลักษณะสำคัญอย่างไร และมีปัญหาที่ต้องการการวิเคราะห์หาคำตอบอยู่มากมาย เช่น ร่างกายมนุษย์ทราบได้อย่างไรว่าส่วนใดในลำดับนิวคลีโอไทด์ที่เป็นแอกซอน และส่วนใดที่เป็นอินทรอน ทราบได้ว่าเมื่อใดจึงได้ตัดลำดับนิวคลีโอไทด์ที่ไม่ใช่แอกซอนออก และในกระบวนการสังเคราะห์โปรตีนนั้น มีรหัสใดเป็นตัวบ่งชี้ว่าโปรตีนชนิดนั้น ๆ ประกอบด้วยแอกซอนชิ้นใดและเรียงกันอย่างไร ซึ่งเป็นเรื่องทางชีวโมเลกุลที่พยายามหาคำตอบ ด้วยกระบวนวิธีต่างๆ ในทางชีวโมเลกุลได้อธิบายด้วย snRNA ซึ่งมีหน้าที่จดจำลักษณะของการคัดเลือกลำดับนิวคลีโอไทด์ที่ใช้ในการสังเคราะห์โปรตีน แต่ไม่ได้อธิบายว่า จะเลือกอย่างไรแล้วเลือกจดจำได้อย่างไร ส่วนในทางชีวสารสนเทศศาสตร์ ได้มีการทดลองนำลำดับนิวคลีโอไทด์ของสัตว์เลี้ยงลูกด้วยนม (Mammalian DNA sequences) บริเวณที่เป็นอินทรอนมาเพื่อสังเกตบริเวณสไปลไซต์ และทำการวัดร้อยละจำนวนของนิวคลีโอไทด์ต่าง ๆ [7] ได้ผลดังรูปที่ 2.7



รูปที่ 2.7 จำนวนร้อยละของนิวคลีโอไทด์ในบริเวณสไปลไซต์ของสัตว์เลี้ยงลูกด้วยนม [7]

ซึ่งแสดงให้เห็นว่าบริเวณสไปลไซต์ของโคเนอร์ไซต์ในส่วนที่เป็นอินทรอนจะพบลำดับนิวคลีโอไทด์เป็น "GT" และบริเวณสไปลไซต์ของแอคเซพเตอร์ไซต์ในส่วนที่เป็นอินทรอน จะพบลำดับนิวคลีโอไทด์เป็น "AG" สูงถึงร้อยละ 99.24% [7] และเป็นลำดับนิวคลีโอไทด์แบบอื่น ๆ เพียง 0.76% [7] จึงได้มีความพยายามที่จะค้นหาตำแหน่งที่แน่นอนของสไปลไซต์นี้ แต่เพียงการค้นพบลำดับนิวคลีโอไทด์ส่วนหน้าเป็น "GT" และส่วนท้ายเป็น "AG" ยังไม่เพียงพอที่จะชี้ตำแหน่งสไปลไซต์ได้ เนื่องจากในลำดับนิวคลีโอไทด์ยังมีรูปแบบนี้อีกมาก และไม่ได้เป็นสไปลไซต์อีกด้วย จึงได้มีการศึกษารูปแบบและพฤติกรรมอื่น ๆ ที่อาจใช้ในการตัดสินใจว่า บริเวณใดเป็นสไปลไซต์ได้อีกหรือไม่ อย่างไร ด้วยการพยายามค้นหาพฤติกรรมของ snRNA ซึ่งทำหน้าที่ในการจดจำลักษณะการตัดเอาอินทรอนออก จากการศึกษาลำดับนิวคลีโอไทด์ที่ snRNA ได้ทำการจดจำเพื่อตัดอินทรอนออก โดยเลือกที่จะพิจารณาลำดับนิวคลีโอไทด์ต่อจากโคเนอร์ไซต์ไปอีก 11 ตำแหน่ง และได้พิจารณาในส่วนของแอกเซพเตอร์ไซต์ที่ก่อนหน้าตำแหน่งอีก 24 ตำแหน่งตามการจดจำของ U2AF65 และได้พยายามศึกษาพฤติกรรมของแอกซอนของทั้งโคเนอร์ไซต์และแอกเซพเตอร์ที่ 5 ตำแหน่งด้วย รวมทั้งพยายามที่จะศึกษาพฤติกรรมของของลำดับนิวคลีโอไทด์ที่เป็นส่วนของแอกซอนอย่างเดียว และส่วนของอินทรอนอย่างเดียวประกอบการศึกษาไปด้วย ซึ่งน่าจะช่วยให้หาแนวทางการค้นหาตำแหน่งสไปลไซต์ได้ กล่าวโดยสรุปก็คือ **เราจะศึกษาลำดับนิวคลีโอไทด์รอบ ๆ โคเนอร์ไซต์ความยาว 16 ตำแหน่ง (ก่อนตำแหน่งสไปลไซต์ 5 ตำแหน่ง หลังตำแหน่งสไปลไซต์ 11 ตำแหน่ง) และในแอกเซพเตอร์ไซต์ความยาว 29 ตำแหน่ง (หน้าตำแหน่งสไปลไซต์ 24 ตำแหน่ง หลังตำแหน่งสไปลไซต์ 5 ตำแหน่ง)** ซึ่งเป็นบริเวณเดียวกับที่ snRNA ใช้ในการจดจำลักษณะของสไปลไซต์

ต้นไม้การตัดสินใจ (Decision tree)

ต้นไม้การตัดสินใจเป็นการเรียนรู้ของเครื่อง (Machine learning) ชนิดหนึ่ง ซึ่งประกอบขึ้นจากเซตของบัพ (nodes) และ เส้นเชื่อม (edges) มาจัดเรียงกันเป็นต้นไม้ ต้นไม้การตัดสินใจจึงเป็นเครื่องมือที่นิยมใช้สำหรับการแก้ปัญหาเกี่ยวกับการจำแนกและจัดกลุ่ม ซึ่งพบมากในปัญหาทางวิทยาศาสตร์และทางสาธารณสุข [8] โดยบัพจะแสดงเงื่อนไข คำถาม หรือการทดสอบที่ได้กำหนดไว้ และแสดงเส้นเชื่อมสำหรับคำตอบที่เกิดขึ้น ดังแสดงได้ในรูปที่ 2.8 โดยบัพจะมี 3 ลักษณะคือ

- บัพเริ่มต้นหรือราก (root node หรือ initial node) ซึ่งเป็นบัพเริ่มต้นของต้นไม้
- บัพภายใน (internal nodes หรือ test nodes) เป็นบัพที่มีเส้นเชื่อมจากบัพแม่ (parent node) โดยจะมีเงื่อนไข คำถาม หรือ แบบทดสอบ ในบัพนั้น และมีเส้นเชื่อมต่อลงไปสู่บัพลูก (child nodes) ตามคำตอบของเงื่อนไข คำถามหรือแบบทดสอบ
- ใบ (leaf nodes หรือ exterior nodes) คือบัพที่มีเพียงเส้นเชื่อมจากบัพแม่และไม่มีเงื่อนไข คำถาม หรือแบบทดสอบ เพื่อตัดสินใจไปที่บัพใดอีก หรือไม่มีบัพลูกนั่นเอง

ในรากหรือบัพเริ่มต้น และ บัพภายในนั้นอาจจะสร้างเส้นเชื่อมออกได้ตั้งแต่สองทางขึ้นไป แต่ที่นิยมใช้ต้นไม้การตัดสินใจกันมาก จะเป็นต้นไม้การตัดสินใจที่คัดแยกออกเป็นสองทางหรือเรียกว่า ต้นไม้การตัดสินใจแบบสองทาง (binary decision tree) ซึ่งสามารถแสดงตัวอย่างต้นไม้การตัดสินใจได้ดังรูปที่ 2.8

รูปที่ 2.8 ตัวอย่างของต้นไม้การตัดสินใจ

จากรูปที่ 2.8 รากและบัพภายในแสดงเงื่อนไข ซึ่งในตัวอย่างข้างต้นนี้แสดงถึงตำแหน่งของลำดับนิวคลีโอไทด์เช่น $x[5]="G"$ หมายความว่า ถ้าลำดับนิวคลีโอไทด์ที่ตำแหน่งที่ 5 มีนิวคลีโอไทด์เป็น "G" จริง (T) ให้ไปทางซ้ายของบัพ แต่ถ้าไม่ใช่หรือเป็นเท็จ (F) ให้ไปทางขวาของบัพ ส่วนใบของต้นไม้คือ Group ที่อยู่ส่วนปลายนั่นเอง

Maximal dependence decomposition (MDD)

Maximal dependence decomposition หรือ MDD คือกระบวนการสร้างต้นไม้การตัดสินใจแบบสองทาง (binary decision tree) ชนิดหนึ่ง ซึ่งเป็นการสร้างต้นไม้การตัดสินใจโดยการแบ่งกลุ่มของลำดับนิวคลีโอไทด์ออกตามความขึ้นต่อกันของตำแหน่ง ทั้งความขึ้นต่อกันอย่างมีนัยสำคัญของตำแหน่งที่อยู่ติดกันและตำแหน่งที่ไม่ได้อยู่ติดกัน โดยมีสมมติฐานเกี่ยวกับการมีอิทธิพลระหว่างกันของตำแหน่งที่อยู่บริเวณโดยรอบของสไปลไซต์ [9] วิธีการของการสร้าง MDD พร้อมตัวอย่างที่ 2.1 ประกอบการสร้าง ดังนี้

กำหนดกลุ่มของลำดับนิวคลีโอไทด์ D ซึ่งได้นำมาทำการปรับแนวของลำดับนิวคลีโอไทด์ความยาว λ ตำแหน่ง เริ่มจาก

(1) ทำการค้นหานิวคลีโอไทด์ที่มีจำนวนมากที่สุด (consensus nucleotide) ในแต่ละตำแหน่ง จากนั้นทำการทดสอบความเป็นอิสระกันของนิวคลีโอไทด์ที่มากที่สุดที่ตำแหน่ง i กับนิวคลีโอไทด์ที่ตำแหน่งที่ j โดยใช้ค่าสถิติ χ^2 ในการทดสอบสมมติฐานเป็นตารางการจร (contingency table) ขนาด 2×4 มีองศาความเป็นอิสระ (df : degree of freedom) ที่คำนวณได้จาก $(r-1)(c-1)$ เมื่อ r คือจำนวนแถวทั้งหมดของตารางการจรและ c คือจำนวนสดมภ์ทั้งหมดของตารางการจร ซึ่งมีค่า $df=(2-1)(4-1)=3$ และค่า P-value ซึ่งเป็นค่าน้อยที่สุดของระดับนัยสำคัญ (α) ที่จะทำให้ปฏิเสธสมมติฐาน H_0

จากนั้นทำการหาผลรวมของค่า χ^2 ของตำแหน่ง i (S_i) จากตำแหน่ง j โดยที่ $i \neq j$ โดยสามารถเขียนเป็นสมการได้เป็น

$$S_i = \sum_{i \neq j} \chi^2 \left(C_i, X_j \right) \tag{2.1}$$

$$\chi^2 = \sum_{m=1}^{r} \sum_{n=1}^{c} \frac{\left(O_{m,n} - E_{m,n} \right)^2}{E_{m,n}} \tag{2.2}$$

โดยที่ C_i มีค่าเป็น 1 ถ้านิวคลีโอไทด์ที่ตำแหน่ง i ตรงกับนิวคลีโอไทด์ที่มากที่สุด และมีค่าเป็น 0 เมื่อไม่ตรงกับนิวคลีโอไทด์ที่มากที่สุด

X_j คือนิวคลีโอไทด์ (a,c,g,t) ที่ตำแหน่ง j

$O_{m,n}$ จำนวนนิวคลีโอไทด์ X_j สดมภ์ที่ n ของ C_i ในแถวที่ m

$E_{m,n} = \dfrac{\left(\sum_{m=1}^{r} O_{m,n} \right) \left(\sum_{n=1}^{c} O_{m,n} \right)}{\sum_{m=1}^{r} \sum_{n=1}^{c} O_{m,n}}$ คือ ค่าคาดหมาย สดมภ์ที่ n แถวที่ m

m แถวของตารางการจร โดยที่ $m = \begin{cases} 1, C_i = 1 \\ 2, C_i = 0 \end{cases}$

n สดมภ์ของตารางการจร โดยที่ $n = \{1,2,3,4\}$ เมื่อ $X_j = a,c,g,t$ ตามลำดับ

(2) จากนั้นพิจารณาเลือกตำแหน่ง i ซึ่ง S_i มีค่ามากที่สุดเพื่อทำการแบ่งกลุ่มของลำดับนิวคลีโอไทด์ออกเป็น 2 กลุ่มโดยในกลุ่มที่ 1 (D_{i_1}) คือกลุ่มของลำดับนิคลีโอไทด์ที่ตำแหน่ง i ตรงกับนิวคลีโอไทด์ที่มากที่สุด ของตำแหน่ง i และกลุ่มที่ 2 คือ ลำดับนิวคลีโอไทด์ที่ไม่ตรงกับเงื่อนไขของ D_{i_1} ($D_{\bar{i}_1} = D - D_{i_1}$)

จากนั้นทำ (1) และ (2) ในแต่ละกลุ่ม (D_{i_1} และ $D_{\bar{i}_1}$) ก็จะได้ต้นไม้การตัดสินใจแบบสองทาง ซึ่งจะหยุดการสร้างต้นไม้การตัดสินใจ MDD เมื่อเกิดเหตุการณ์ใดเหตุการณ์หนึ่งต่อไปนี้

- ต้นไม้การตัดสินใจมีความลึก $\lambda - 1$ ชั้น
- ไม่มีนัยสำคัญทางสถิติระหว่างตำแหน่งของลำดับนิวคลีโอไทด์อีก โดยพิจารณาจากค่า P-value ที่ 0.001

- จำนวนของลำดับนิวคลีโอไทด์ในกลุ่มน้อยกว่าจำนวนที่เหมาะสมในการสร้างแบบจำลอง

ตัวอย่างที่ 2.1 แสดงการทำงานของ MDD ตามขั้นตอนข้างต้น

ตัวอย่างที่ 2.1 นำลำดับนิวคลีโอไทด์จากยีนของมนุษย์จำนวน 1,115 ยีนที่เป็นโคเนอร์ไซด์ความยาว 16 ตำแหน่งมาจำนวน 5,709 ลำดับ [3] มาจัดเรียงให้ตำแหน่งสไปลไซต์ตรงกัน โดยให้สไปลไซต์ที่ "G" อยู่ในตำแหน่งที่ 6 และ "T" อยู่ในตำแหน่งที่ 7 เมื่อทำการนับจำนวนนิวคลีโอไทด์ในแต่ละตำแหน่ง และหานิวคลีโอไทด์ที่มีจำนวนมากที่สุดจะได้ผลในตารางที่ 2.1

ตารางที่ 2.1 จำนวนนิวคลีโอไทด์ในแต่ละตำแหน่งเพื่อหานิวคลีโอไทด์ที่มากที่สุด

ตำแหน่ง (i)	$n(a_i)$	$n(c_i)$	$n(g_i)$	$n(t_i)$	consensus
1	1482	1573*	1294	1360	c
2	1539	1704*	1392	1074	c
3	1833	2141*	1063	672	c
4	3416*	750	749	794	a
5	493	181	4616*	419	g
6	0	0	5709*	0	g
7	0	0	0	5709*	t
8	2770*	163	2643	133	a
9	4007*	457	725	520	a
10	386	315	4719*	289	g
11	931	1021	1232	2525*	t
12	1484	1232	2028*	965	g
13	1119	1665*	1558	1367	c
14	1050	1676*	1610	1373	c
15	1043	1553	1797*	1316	g
16	1169	1570	1732*	1238	g

หลังจากหาลำดับนิวคลีโอไทด์ที่มากที่สุดในแต่ละตำแหน่งได้แล้ว จะเริ่มทำการคำนวณค่า χ^2 โดยจะแสดงตัวอย่างการคำนวณค่า $\chi^2(C_i, X_j)$ ในตำแหน่งที่ 1 เทียบกับตำแหน่งที่ 2 ($i=1, j=2$) สามารถเขียนตารางความถี่ในตารางที่ 2.2 และตารางค่าคาดหมาย (expected value table) ในตารางที่ 2.3

ตารางที่ 2.2 ตารางความถี่ระหว่าง C_1 เทียบกับ X_2

$O_{a,b}$	$n(A_2)$	$n(C_2)$	$n(G_2)$	$n(T_2)$	รวม
$C_1=1$	519	520	166	368	1573
$C_1=0$	1020	1184	1226	706	3956
รวม	1539	1704	1392	1074	5709

ตารางที่ 2.3 ตารางค่าคาดหมายระหว่าง C_1 เทียบกับ X_2 ของตารางที่ 2.2

$E_{a,b}$	$n(A_2)$	$n(C_2)$	$n(G_2)$	$n(T_2)$
$C_1=1$	$=\frac{1573\times1539}{5709}$ $=424.04$	$=\frac{1573\times1704}{5709}$ $=469.50$	$=\frac{1573\times1329}{5709}$ $=383.54$	$=\frac{1573\times1074}{5709}$ $=295.92$
$C_1=0$	$=\frac{3956\times1539}{5709}$ $=1114.96$	$=\frac{3956\times1704}{5709}$ $=1234.50$	$=\frac{3956\times1392}{5709}$ $=1008.46$	$=\frac{3956\times1074}{5709}$ $=778.08$

จะได้

$$\chi^2(C_1,X_2)=\frac{(519-424.04)^2}{424.04}+\frac{(520-469.50)^2}{469.50}+\frac{(166-383.54)^2}{383.54}+\frac{(368-295.92)^2}{295.92}$$
$$+\frac{(1020-1114.96)^2}{1114.96}+\frac{(1184-1234.50)^2}{1234.50}+\frac{(1226-1008.46)^2}{1008.46}+\frac{(706-778.08)^2}{778.08}$$
$$=231.395$$

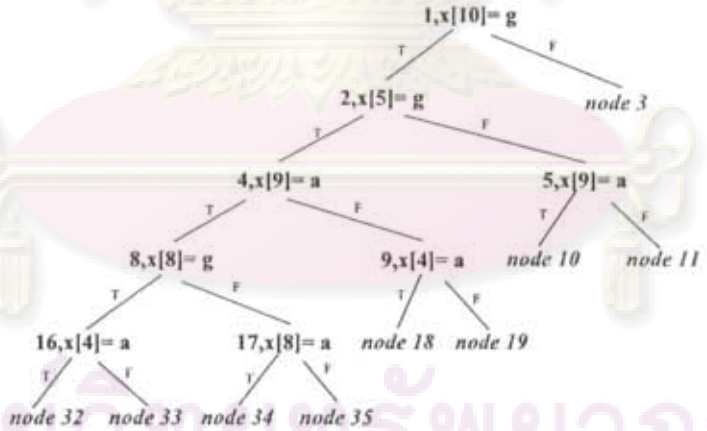
สำหรับ χ^2 ที่ความเชื่อมั่น 99.99% หรือค่า P-value ที่ 0.001 และองศาความเป็นอิสระเท่ากับ 3 จากตารางมีค่าเท่ากับ 16.3 ซึ่ง χ^2 ที่ได้จากการคำนวณมากกว่าในตาราง นั่นคือตำแหน่งที่ 1 และตำแหน่งที่ 2 ของกลุ่มของลำดับนิวคลีโอไทด์มีความขึ้นต่อกันอย่างมีนัยสำคัญทางสถิติด้วยความเชื่อมั่น 99.99%

และเมื่อทำการคำนวณหาค่า S_1 จากผลรวมของค่า χ^2 ที่ตำแหน่ง $j=2,...,16$ จะมีค่าเท่ากับ 341.61 สำหรับในตำแหน่งอื่น ๆ จะได้ค่า S_j ในตารางที่ 2.4

ตารางที่ 2.4 ค่า S_i ที่ได้จากการคำนวณ

ตำแหน่ง (i)	S_i	ตำแหน่ง (i)	S_i
1	341.61	9	1206.50
2	348.58	10	*1920.51
3	506.40	11	1482.07
4	1836.85	12	536.65
5	1693.36	13	697.18
6	0.00	14	702.87
7	0.00	15	817.67
8	1296.30	16	692.71

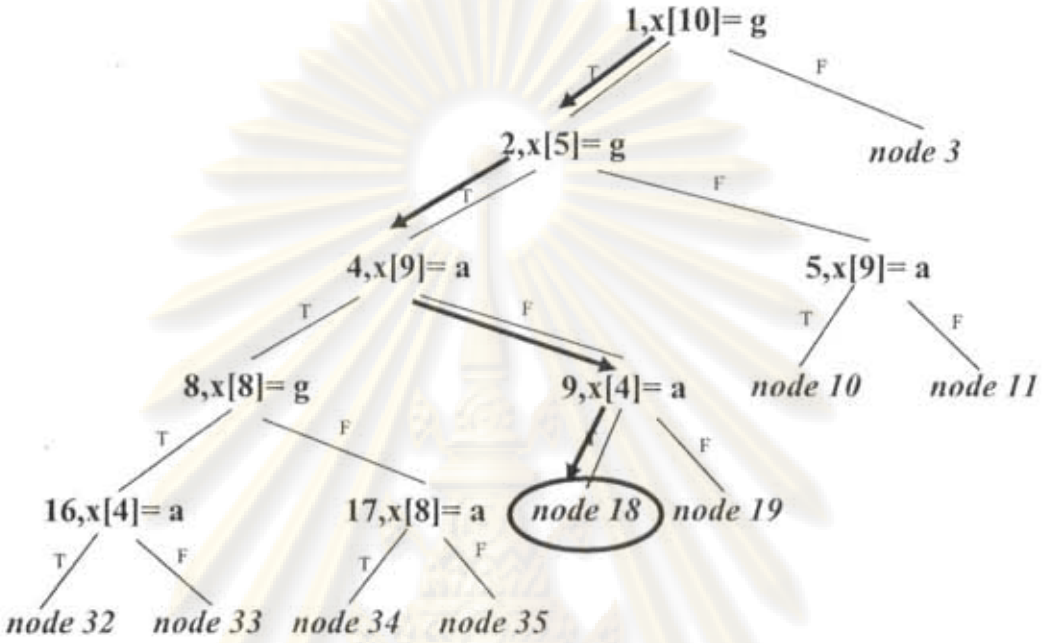
จากตารางที่ 2.4 จะพบว่าที่ตำแหน่งที่ 10 มีค่า S_i มากที่สุด คือ 1,920.51 จากการคำนวณค่า χ^2 ระหว่างตำแหน่งที่ 10 กับตำแหน่งอื่น ๆ ที่ไม่ใช่ตำแหน่งที่ 6, 7 และ 10 นั้นค่าองศาความเป็นอิสระของ χ^2 มีค่าเท่ากับ 12 โดยที่ค่า χ^2 ที่ได้จากการเปิดตารางเท่ากับ 32.9 เนื่องจากค่าที่ได้จากการคำนวณมีค่ามากกว่าค่าที่ได้จากการเปิดตาราง แสดงว่ายังมีนัยสำคัญทางสถิติระหว่างตำแหน่งอยู่ ดังนั้นจึงทำการแยกลำดับนิวคลีโอไทด์ที่มีนิวคลีโอไทด์ที่ตำแหน่งที่ 10 เป็น "G" ไว้ในกลุ่มที่ 1 และลำดับนิวคลีโอไทด์ที่ไม่ได้อยู่ในกลุ่มที่ 1 ให้แยกไว้ที่กลุ่มที่ 2 และกระทำไปเรื่อย ๆ และตรวจสอบการหยุดสร้างจนได้ต้นไม้การตัดสินใจ MDD ดังรูปที่ 2.9



รูปที่ 2.9 ต้นไม้การตัดสินใจ MDD ของโดเนอร์ไซต์

หลังจากที่ได้ทำการสร้างต้นไม้การตัดสินใจ MDD แล้ว การใช้งานต้นไม้การตัดสินใจ MDD นั้น เมื่อมีการรับเอาลำดับนิวคลีโอไทด์เข้ามา ยกตัวอย่างเช่น มีลำดับนิวคลีโอไทด์ "ataaggtgcgtgcatc" เข้ามาจะจัดไว้ในใบไหนของต้นไม้การตัดสินใจ MDD ในรูปที่ 2.9นั้น จะเริ่มจากการพิจารณาในบัพที่ 1 จะพบว่าในตำแหน่งที่ 10 มีนิวคลีโอไทด์เป็น 'g' ดังนั้นจึงไปพิจารณา

ต่อในบัพที่ 2 ซึ่งก็พบว่าตำแหน่งที่ 5 เป็น 'g' จึงได้ไปพิจารณาที่บัพที่ 4 แต่ในตำแหน่งที่ 9 เป็น 'c' ทำให้จัดลงไปพิจารณาต่อในบัพที่ 9 พบว่าในตำแหน่งที่ 4 เป็น 'a' จริง ดังนั้นจึงไปสู่ใบที่ 18 หรือกล่าวโดยสรุปก็คือ ลำดับนิวคลีโอไทด์ "ataaggtgcgtgcatc" เมื่อนำมาพิจารณาจัดกลุ่มด้วยต้นไม้การตัดสินใจ MDD ในรูปที่ 2.9 นั้น จะถูกจัดเข้าไปอยู่ในใบ กลุ่มที่ 18 แสดงได้ในรูปที่ 2.10 #



รูปที่ 2.10 ตัวอย่างของการจัดกลุ่มในต้นไม้การตัดสินใจ MDD ของโดเนอร์ไซต์

และได้กระทำเช่นเดียวกันด้วยข้อมูลแอกเซพเตอร์ไซต์ในยีนของมนุษย์

แบบจำลองมาร์กอฟ (Markov models)

แบบจำลองมาร์คอฟ เป็นแบบจำลองทางสถิติที่สร้างขึ้นจาก ลูกโซ่มาร์คอฟ (Markov chains) หรือเรียกอีกชื่อหนึ่งว่า แบบจำลองลูกโซ่มาร์คอฟ โดยแบบจำลองนี้ เป็นวิธีการวิเคราะห์พฤติกรรมของตัวแปรที่สนใจในเหตุการณ์ปัจจุบัน (present event) ที่สถานะ (state) ของเหตุการณ์ เพื่อพยากรณ์พฤติกรรมของตัวแปรนั้นในเหตุการณ์ของอนาคต (future event) โดยอาศัยข้อมูลพื้นฐานในปัจจุบันในการวิเคราะห์เหตุการณ์ที่จะเกิดขึ้นซึ่งมีความต่อเนื่องกัน แบบจำลองลูกโซ่มาร์คอฟนี้ถูกพัฒนาขึ้นโดยนักคณิตศาสตร์ชาวรัสเซียชื่อ Andrei A. Markov ในค.ศ. 1907 จากบทนิยาม

บทนิยาม ให้ $X=(X_n;n\geq 0)$ เป็นลำดับของตัวแปรสุ่มจากเซตจำกัด S เรียกว่า ปริภูมิสถานะ (state space) ถ้าทุก $n\geq 0$ และสำหรับทุกจำนวนที่เป็นไปได้ของ $i,k,k_0,\ldots,k_{n-1}$ จะได้

$$\begin{aligned} P(X_{n+1}=k|X_0=k_0,\ldots,X_n=i) &= P(X_{n+1}=k|X_n=i) \\ &= P(X_1=k|X_0=i) \end{aligned} \tag{2.3}$$

แล้วจะกล่าวได้ว่า X เป็นลูกโซ่มาร์คอฟหรือ X มีสมบัติมาร์คอฟ (Markov property) และสามารถเขียน $p_{ik}=P(X_1=k|X_0=i)$ เมื่อ $(p_{ik};i\in S,k\in S)$ คือความน่าจะเป็นในการเปลี่ยนสถานะของลูกโซ่มาร์คอฟ (transition probabilities of the chain) [10]

ตัวอย่างที่ 2.2 สมมติลำดับนิวคลีโอไทด์ความยาว 10 ตำแหน่ง จำนวน 1000 ลำดับ และจำนวนนิวคลีโอไทด์ในแต่ละตำแหน่งแสดงได้ในตารางที่ 2.5

ตารางที่ 2.5 จำนวนนิวคลีโอไทด์ที่ตำแหน่ง i

จำนวน	ตำแหน่ง (i)									
	1	2	3	4	5	6	7	8	9	10
$f(a_i)$	232	227	255	233	227	229	261	243	247	243
$f(c_i)$	263	259	233	271	231	272	270	251	234	277
$f(g_i)$	247	256	255	255	277	254	227	257	259	232
$f(t_i)$	258	258	257	241	265	245	242	249	260	248

ถ้าสมมติฐานของลำดับนิวคลีโอไทด์นี้คือ ตำแหน่งนิวคลีโอไทด์ในแต่ละตำแหน่งเป็นอิสระต่อกัน หรืออาจกล่าวได้ว่า ตำแหน่งที่ i ไม่ได้ขึ้นกับตำแหน่งที่ $i-1$ ทุก i ตั้งแต่ 1 ถึง 10 ดังนั้นถ้ากำหนด ลำดับนิวคลีโอไทด์ atcggcgcaa เราจะได้ว่า ความน่าจะเป็นที่ลำดับนิวคลีโอไทด์ที่กำหนดให้นี้จะอยู่ในกลุ่มตัวอย่าง คือ

$$\begin{aligned} P(atcggcgcaa) &= P(a_1)\cdot P(t_2)\cdot P(c_3)\cdot P(g_4)\cdot P(g_5)\cdot P(c_6)\cdot P(g_7)\cdot P(c_8)\cdot P(a_9)\cdot P(a_{10}) \\ &= \left(\frac{232}{1000}\right)\left(\frac{258}{1000}\right)\left(\frac{233}{1000}\right)\left(\frac{255}{1000}\right)\left(\frac{277}{1000}\right)\left(\frac{272}{1000}\right)\left(\frac{227}{1000}\right)\left(\frac{251}{1000}\right)\left(\frac{247}{1000}\right)\left(\frac{243}{1000}\right) \\ &= 0.000000961337 \end{aligned}$$

จากสมมติฐานข้างต้น เป็นการแสดงแบบจำลองมาร์คอฟอันดับ 0 (หมายถึงโอกาสที่เหตุการณ์ในอนาคตจะเกิดขึ้นจากเหตุการณ์ในอดีตจนถึงปัจจุบัน 0 เหตุการณ์) และการพยากรณ์จากแบบจำลองโดยใช้ข้อมูลจากตารางที่ 2.5 และจะเรียกตารางที่สร้างเป็นความน่าจะเป็นจากตารางที่ 2.5 นี้ว่า เมทริกซ์เปลี่ยนสถานะ (transition matrix)

จากตัวอย่างที่ 2.2 ถ้าตั้งสมมติฐานว่าโอกาสที่เหตุการณ์จะเกิดขึ้นในอนาคตจากเหตุการณ์ในอดีตถึงปัจจุบัน n เหตุการณ์ ก็จะเรียกเป็น แบบจำลองมาร์คอฟอันดับ n นั่นเอง

ดังนั้นถ้าต้องการหาแบบจำลองมาร์คอฟอันดับ 1 จากตัวอย่าง นั่นคือการสนใจพยากรณ์เหตุการณ์ในอนาคตจากเหตุการณ์ปัจจุบัน 1 เหตุการณ์ เมทริกซ์เปลี่ยนสถานะของแบบจำลองมาร์คอฟอันดับ 1 ซึ่งโดยทฤษฎีของเบย์ (Bayes' theorem) และสมบัติมาร์คอฟทำให้เราได้ว่า $p_{x_{i-1}x_i} = P(x_i | x_{i-1}) = \frac{P(x_{i-1}x_i)}{P(x_{x-1})}$ และเมื่อนับจำนวนนิวคลีโอไทด์ที่ต้องการจะได้ผลในตารางที่ 2.6

ตารางที่ 2.6 จำนวนนิวคลีโอไทด์ที่ตำแหน่งติดกัน ณ ตำแหน่ง i

$f(x_i x_{i-1})$	ตำแหน่ง (i)								
	2	3	4	5	6	7	8	9	10
$a_{i-1}a_i$	50	68	46	51	52	52	64	63	49
$a_{i-1}c_i$	65	50	74	48	71	72	74	50	72
$a_{i-1}g_i$	56	62	69	71	52	51	61	61	64
$a_{i-1}t_i$	61	47	66	63	52	54	62	69	62
$c_{i-1}a_i$	70	63	59	60	51	78	66	59	57
$c_{i-1}c_i$	68	67	61	63	72	69	66	61	73
$c_{i-1}g_i$	63	58	60	74	53	55	64	66	52
$c_{i-1}t_i$	62	71	53	74	55	70	74	65	52
$g_{i-1}a_i$	64	66	55	59	66	69	50	63	59
$g_{i-1}c_i$	57	56	81	62	67	58	61	68	73
$g_{i-1}g_i$	62	66	64	62	83	65	66	65	59
$g_{i-1}t_i$	64	68	55	72	61	62	50	61	68
$t_{i-1}a_i$	43	58	73	57	60	62	63	62	78
$t_{i-1}c_i$	69	60	55	58	62	71	50	55	59
$t_{i-1}g_i$	75	69	62	70	66	56	66	67	57
$t_{i-1}t_i$	71	71	67	56	77	56	63	65	66

สมมติตัวอย่างลำดับนิวคลีโอไทด์ *aaaaaaaaaa* ความน่าจะเป็นที่ตำแหน่งนี้จะอยู่ในกลุ่มตัวอย่างที่ 2.2 จะเป็น

$$\begin{aligned}P(aaaaaaaaaa) &= P(a_{10}|a_9a_8a_7a_6a_5a_4a_3a_2a_1)\cdot P(a_9|a_8a_7a_6a_5a_4a_3a_2a_1)\cdot P(a_8|a_7a_6a_5a_4a_3a_2a_1)\\ &\quad\cdot P(a_7|a_6a_5a_4a_3a_2a_1)\cdot P(a_6|a_5a_4a_3a_2a_1)\cdot P(a_5|a_4a_3a_2a_1)\cdot P(a_4|a_3a_2a_1)\\ &\quad\cdot P(a_3|a_2a_1)\cdot P(a_2|a_1)\cdot P(a_1)\end{aligned}$$

จากสมบัติมาร์คอฟ และเป็นแบบจำลองอันดับ 1 ซึ่งเราจะสนใจเหตุการณ์ในอดีต 1 เหตุการณ์ ดังนั้น

$$\begin{aligned}P(aaaaaaaaaa) &= P(a_{10}|a_9)\cdot P(a_9|a_8)\cdot P(a_8|a_7)\cdot P(a_7|a_6)\cdot P(a_6|a_5)\cdot P(a_5|a_4)\cdot P(a_4|a_3)\cdot P(a_3|a_2)\\ &\quad\cdot P(a_2|a_1)\cdot P(a_1)\\ &= \frac{n(a_{10}a_9)}{n(a_9)}\cdot\frac{n(a_9a_8)}{n(a_8)}\cdot\frac{n(a_8a_7)}{n(a_7)}\cdot\frac{n(a_7a_6)}{n(a_6)}\cdot\frac{n(a_6a_5)}{n(a_5)}\cdot\frac{n(a_5a_4)}{n(a_4)}\cdot\frac{n(a_4a_3)}{n(a_3)}\cdot\frac{n(a_3a_2)}{n(a_2)}\cdot\frac{n(a_2a_1)}{n(a_1)}\cdot\frac{n(a_1)}{n(x_1)}\end{aligned}$$

$$\begin{aligned}P(aaaaaaaaaa) &= \left(\frac{49}{247}\right)\cdot\left(\frac{63}{243}\right)\cdot\left(\frac{64}{261}\right)\cdot\left(\frac{52}{229}\right)\cdot\left(\frac{52}{227}\right)\cdot\left(\frac{51}{233}\right)\cdot\left(\frac{46}{255}\right)\cdot\left(\frac{68}{227}\right)\cdot\left(\frac{50}{232}\right)\cdot\left(\frac{232}{1000}\right)\\ &= \frac{3570624512}{9203245324668225} = 0.000000387975\end{aligned}$$

#

จากการคำนวณข้างต้นสามารถสังเกตได้ว่าในตำแหน่งที่ 1 ของลำดับนิวคลีโอไทด์ *aaaaaaaaaa* ไม่มีเหตุการณ์ในอดีตจึงถือว่าเป็นอิสระไม่ขึ้นกับเหตุการณ์ในอดีต หรือนั่นคือเป็นการใช้แบบจำลองมาร์คอฟอันดับ 0 ในการคำนวณแทน

บทที่ 3
เอกสารและงานวิจัยที่เกี่ยวข้อง

จากการศึกษาบทความทางชีวสารสนเทศที่เกี่ยวกับการค้นหาตำแหน่งสไปลไซต์ภายในยีนของมนุษย์นั้น พบว่ามีอยู่หลายบทความ ซึ่งแต่ละบทความใช้อัลกอริทึมและวิธีการเปรียบเทียบประสิทธิภาพของผลลัพธ์ที่แตกต่างกันอีกด้วย

บทความและงานวิจัยที่เกี่ยวข้องกับการค้นหาตำแหน่งสไปลไซต์ของยีน

โปรแกรม NetGene2 [11] ซึ่งได้ใช้เทคนิคของโครงข่ายประสาทเทียม (Neural network) แบบ feedforward ในการทำนายตำแหน่งของสไปลไซต์จากการเลื่อนหน้าต่างในการทำนายไปตลอดทั้งสายของยีน และตรวจสอบความแม่นยำโดยใช้ correlation coefficient ($C(X)$) ซึ่งถ้าค่า $C(X)$ เท่ากับ 1 แสดงว่าสามารถทำนายได้ถูกต้อง 100% และเท่ากับ -1 หมายถึงทำนายผิดพลาด 100% ซึ่งผลลัพธ์ของโปรแกรม NetGene2 มีค่า $C(X)$ มากที่สุดที่ 0.9244 สำหรับโดเนอร์ไซต์ และ 0.83 สำหรับแอกเซพเตอร์ไซต์

สำหรับโปรแกรมที่จะกล่าวถึงต่อไปก็คือ HSPL [12] ซึ่งได้พัฒนาการทำนายลำดับแอกซอนภายใน ในยีนของมนุษย์ ซึ่งใช้ฟังก์ชันดิสคริมิแนนต์เชิงเส้น (linear discriminant function) ที่รวบรวมข้อมูลเกี่ยวกับการมีฟังก์ชันคาบ 3 ของบริเวณต่าง ๆ รอบ ๆ สไปลไซต์ และลำดับนิวคลีโอไทด์สายสั้น ๆ ของแอกซอนและอินทรอน และตรวจสอบความแม่นยำด้วย $C(X)$ ซึ่งในโดเนอร์ไซต์ มีค่าเท่ากับ 0.63 และแอกเซพเตอร์ไซต์ที่ 0.47

อีกโปรแกรมที่น่าสนใจคือ โปรแกรม NNSplice [13] ซึ่งได้ทำการปรับปรุงโปรแกรม Genie ซึ่งจากเดิมใช้เพียงเทคนิคของ Generalize Hidden Markov Models (GHMMs) ในการทดลองและประมาณความน่าจะเป็นในองค์ประกอบของยีนโดยใช้กำหนดการพลวัต (dynamic programming) ในการรวบรวมข้อมูลจาก 2 ส่วนคือ เนื้อหาจากหลาย ๆ ส่วน และจากตัวรับรู้สัญญาณ (signal sensors) ซึ่ง NNSplice ได้ทำการปรับปรุงในส่วนของตัวรับรู้สัญญาณด้วยโครงข่ายประสาทเทียม 2 ชุดที่พัฒนาขึ้นมาใหม่ โดยมีพื้นฐานจากความถี่ของคู่นิวคลีโอไทด์ และตรวจสอบความแม่นยำของโปรแกรมด้วย 7-fold cross validation และหาค่าของ approximate coefficient (AC) ซึ่งเป็นทางเลือกในการทดสอบที่ดีกว่า $C(X)$ โดยค่า AC ที่มีค่าเท่ากับ 1 แสดงว่าทำนายได้ถูกต้อง 100% และเท่ากับ 0 คือทำนายผิดพลาด 100% ซึ่งจากเดิมโปรแกรม Genie ได้ค่าเฉลี่ย AC เท่ากับ 0.74 และเมื่อทำการปรับปรุงแล้วจะมีค่า AC เป็น 0.79

ต่อมาได้มีการพัฒนาโปรแกรมทำนายสไปลไซต์ของยีนขึ้นมาในช่วงปี ค.ศ. 2001 ซึ่งได้รวบรวมเทคนิคที่มีประสิทธิภาพต่าง ๆ ในการทำนายมาใช้ในโปรแกรมนี้ ซึ่งมีชื่อว่า GeneSplicer ซึ่งเป็นโปรแกรมค้นหายีนที่มีประสิทธิภาพ [3] โดยที่โปรแกรมนี้ใช้ต้นไม้การตัดสินใจ (Decision tree) ด้วยวิธี Maximal dependence decomposition (MDD) เพื่อพิจารณาความขึ้นต่อกันของตำแหน่งอย่างมีนัยสำคัญ ณ บริเวณโดยรอบของสไปลไซต์ของโดเนอร์ไซต์หรือแอกเซพเตอร์ไซต์ [9] และใช้แบบจำลองมาร์คอฟเพื่อเพิ่มความสามารถในการทำนายสไปลไซต์ โดยใช้ยีนของมนุษย์จำนวน 1,115 ยีนเป็นกลุ่มตัวอย่างในการศึกษา ซึ่งบทความได้แสดงให้เห็นประสิทธิภาพว่าสามารถทำนายตำแหน่งของสไปลไซต์ได้มีประสิทธิภาพดีกว่าโปรแกรมที่กล่าวมาข้างต้น จึงเป็นโปรแกรมที่น่าจะนำมาค้นคว้าและพัฒนาประสิทธิภาพของโปรแกรม GeneSplicer ให้ดียิ่งขึ้น ซึ่งในส่วนของโปรแกรม GeneSplicer นั้นได้พัฒนาความสามารถในการทำนายตำแหน่งสไปลไซต์ของสิ่งมีชีวิตชนิดอื่น ๆ นอกเหนือจากมนุษย์อีกด้วย

ระเบียบวิธีของ GeneSplicer

เริ่มจากการนำลำดับนิวคลีโอไทด์ ในกลุ่มตัวอย่างมาจัดเรียงกัน โดยให้ตำแหน่งที่เป็นสไปลไซต์อยู่ตรงกัน แยกตามชนิดของสไปลไซต์ (โดเนอร์ไซต์ และแอกเซพเตอร์ไซต์) แล้วสกัดเอาเฉพาะลำดับบริเวณรอบๆสไปลไซต์ โดยสำหรับโดเนอร์ไซต์ ตัดให้ได้ความยาว 16 ตำแหน่ง และสำหรับแอกเซพเตอร์ไซต์ ตัดให้ได้ความยาว 29 ตำแหน่ง ก็จะได้กลุ่มตัวอย่างที่จะนำไปทดสอบต่อไป จะเรียกกลุ่มตัวอย่างนี้ว่ากลุ่มสไปลไซต์จริง

จากนั้นสร้างต้นไม้การตัดสินใจ MDD ซึ่งในแต่ละใบ จะประกอบด้วยกลุ่มของลำดับนิวคลีโอไทด์สไปลไซต์ที่ใช้เป็นกลุ่มตัวอย่าง จากนั้นสร้างแบบจำลองมาร์คอฟอันดับ 1 (ในตำแหน่งที่ 1 ของลำดับนิวคลีโอไทด์สไปลไซต์นั้นจะสร้างแบบจำลองมาร์คอฟอันดับ 0 แทน) และสำหรับแบบจำลองมาร์คอฟของกลุ่มสไปลไซต์เท็จ (false splice sites) นั้น สร้างจากลำดับนิวคลีโอไทด์ที่มี "GT" ในกลุ่มเท็จของโดเนอร์ และ "AG" ในกลุ่มเท็จของแอกเซพเตอร์ ซึ่งลำดับนิวคลีโอไทด์เหล่านี้ไม่ได้เป็นสไปลไซต์ แต่นำมาสร้างแบบจำลองเช่นเดียวกับสไปลไซต์จริงเพื่อการเปรียบเทียบ

การปรับปรุงและพัฒนาความสามารถของโปรแกรมในการระบุตำแหน่งสไปลไซต์นั้นได้มีการเพิ่มเทคนิคเข้าไปในระบบ ซึ่งจากการสังเกตพบว่าสไปลไซต์จะถูกล้อมรอบด้วยแอกซอนกับอินทรอนเสมอ จึงได้สร้างแบบจำลองมาร์คอฟอันดับ 1 จำนวน 2 แบบจำลอง แบบจำลองแรกเป็นแบบจำลองแอกซอน และอีกแบบจำลองคือแบบจำลองอินทรอน จากบริเวณรอบ ๆ สไปลไซต์ความยาว 80 ตำแหน่งในแต่ละด้านของสไปลไซต์ (สำหรับแอกซอนและ

อินทรอนที่มีขนาดความยาวน้อยกว่า 80 ตำแหน่ง ถือเป็นส่วนน้อยที่พบ) คะแนนของสไปลไซต์ที่ตำแหน่ง k ในลำดับนิวคลีโอไทด์ จะคำนวณจากสูตรต่อไปนี้

สำหรับโดเนอร์ไซต์

$$S(k) = S_{comb}(k,16) + \left[S_{cod}(k-80) - S_{noncod}(k-80)\right] + \left[S_{noncod}(k+1) - S_{cod}(k+1)\right]$$

เมื่อ k เป็นตำแหน่งของสไปลไซต์โดเนอร์ไซต์ และ สำหรับแอกเซพเตอร์ไซต์

$$S(k) = S_{comb}(k,29) + \left[S_{noncod}(k-80) - S_{cod}(k-80)\right] + \left[S_{cod}(k+1) - S_{noncod}(k+1)\right]$$

เมื่อ k เป็นตำแหน่งของสไปลไซต์แอกเซพเตอร์ไซต์

โดยที่ $S_{comb}(k,i)$ เป็นคะแนนที่ได้จากการคำนวณของแบบจำลองมาร์คอฟในใบของต้นไม้การตัดสินใจ MDD

$S_{cod}(j)$ เป็นคะแนนที่ได้จากการคำนวณของแบบจำลองมาร์คอฟในบริเวณแอกซอนความยาว 80 ตำแหน่งเริ่มที่ตำแหน่ง j

$S_{noncod}(j)$ เป็นคะแนนที่ได้จากการคำนวณของแบบจำลองมาร์คอฟในบริเวณอินทรอนความยาว 80 ตำแหน่งเริ่มที่ตำแหน่ง j

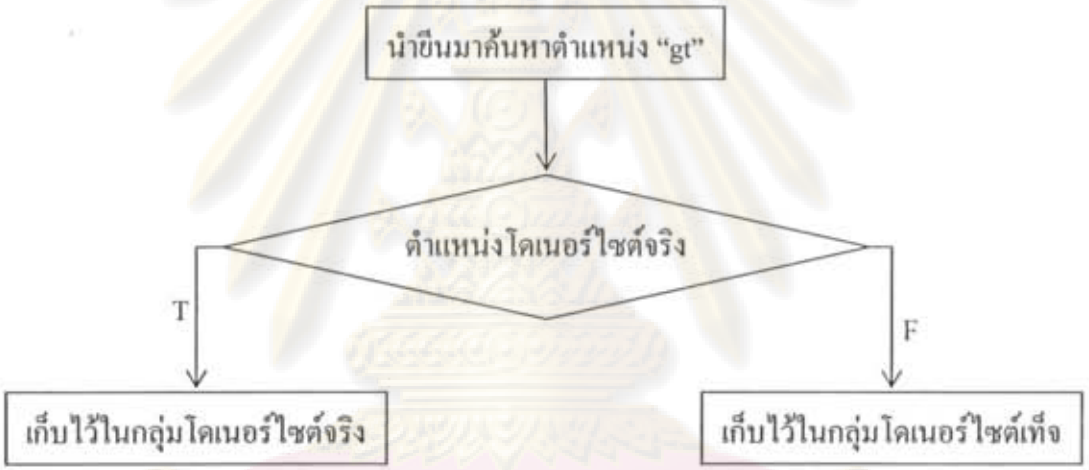
จากสมการการคิดคะแนนข้างต้น การคำนวณคะแนนของโปรแกรมจะเกิดจากการนำยีนที่ต้องการนำมาทดสอบมาหาตำแหน่งที่เป็นลักษณะของสไปลไซต์ แล้วนำลำดับนิวคลีโอไทด์ที่สกัดจากตำแหน่งที่พบมาทำการจัดกลุ่มในต้นไม้การตัดสินใจ MDD แล้วนำมาคำนวณคะแนน ซึ่งจะได้คะแนนออกมา โดยที่คะแนนนั้น จะใช้ในการคัดเลือกว่า ตำแหน่งสไปลไซต์ใดน่าจะเป็นสไปลไซต์จริงบ้างจากคะแนนของลำดับนิวคลีโอไทด์ ถ้ามีคะแนนสูง เกิดจากมีคะแนนในกลุ่มจริงสูง หรือมีคะแนนในกลุ่มเท็จต่ำ คะแนนจึงออกมาสูง และสไปลไซต์ใดที่มีคะแนนต่ำก็น่าจะมาจากมีคะแนนในกลุ่มจริง หรือ คะแนนในกลุ่มเท็จสูง ซึ่งจะเห็นว่ามีเพียงแบบจำลองอยู่ 2 แบบในการช่วยในการพิจารณา ซึ่งเราอาจจะเพิ่มเติมการทำงานของโปรแกรม GeneSplicer โดยการเพิ่มแบบจำลองที่น่าสนใจสักหนึ่งอย่างเพื่อช่วยให้มีการทำนายที่ดีขึ้น

บทที่ 4
วิธีที่ใช้ในการเรียนรู้

เพื่อให้เกิดความเข้าใจในวิธีการที่ใช้ในการเรียนรู้มากขึ้น จึงแสดงลักษณะการทำงานของโปรแกรม GeneSplicer ในแผนภาพที่ 4.1 และ 4.2 เพื่อแสดงขั้นตอนการเรียนรู้ของโปรแกรม

การเก็บข้อมูลที่ใช้ในการเรียนรู้และทดสอบ

ข้อมูลที่ใช้เป็นยีนของมนุษย์จำนวน 1,115 ยีน จาก GENBANK [3] และอ่านลำดับนิวคลีโอไทด์จาก 5' ไป 3' ในการค้นหาโคเนอร์ไซต์และแอกเซพเตอร์ไซต์ ในยีน



แผนภาพที่ 4.1 การนำยีนมาเก็บข้อมูลในการเรียนรู้และแบ่งกลุ่ม

แผนภาพที่ 4.1 แสดงการเก็บตำแหน่งที่พบ "gt" ถ้าเป็นตำแหน่งโคเนอร์ไซต์จริงจะเก็บไว้ในกลุ่มโคเนอร์ไซต์จริง ถ้าไม่ใช่จะเก็บไว้ในกลุ่มโคเนอร์ไซต์เท็จ

ตัวอย่างที่ 4.1 ลำดับนิวคลีโอไทด์ยีนของมนุษย์ (GENBANK accession number: U90224 REGION: 555 - 1441)

```
   1  ATGACTCCCC TCTGCCCTCG CCCCGCGCTC TGCTACCATT TCCTTACGTC TCTGCTTCGC TCAGCGATGC
  71  AAAACGCGCG AGGCACGGCA GAGGGCCGAA GCCGCGGTAC TCTCCGGGCC AGGCCCGCCC CTCGGCCGCC
 141  GGCGGCGCAG CACGGGATTC CCCGGCCGCT GTCCAGCGCT GGCCGCCTGA GCCAAGGCTG CCGCGGAGCC
 211  AGTACAGTCG GGGCCGCTGG CTGGAAGGGC GAGCTTCCTA AGGCGGGGGG AAGCCCGGCG CCGGGGCCGG
 281  gtaggaaagg cgggggaggg gctccggccg tctggaagga atccacgcgg cttgaggctg tggggaagta
 351  gggtggcgag cggtggttct gcgcgcgggg ggcggggggg tggggtggtc cattaggggc ccctggcgag
 421  ggggcggctt tctagtgtgt gaggcgacgc cctagaagct ccccttcaaa gttggcccca cgcgctgaat
 491  gtggaaagtt gactgggacc cagtagtttc ccatcccaaa cctgctttcc gagaagggct tcaaacccaa
 561  aatgtgaatc ccgcctcccc tctcaccaga actgtggact cgtcccgggg aggggcggtg ggtggggcgg
 631  ggctggcggg aaatttcggt tttggcgcgc tccctgcggc gacgctcatc gtgcgctctc ctcttccccc
 701  ggtggtctcc tcgctcgcct tctggctctg ccatgccctg ctctgaagAG ACACCCGCCA TTTCACCCAG
 771  TAAGCGGGCC CGGCCTGCGG AGGTGGGCGG CATGCAGCTC CGCTTTGCCC GGCTCTCCGA GCACGCCACG
 841  GCCCCCACCC GGGGCTCCGC GCGCGCCGCG GGCTACGACC TGTACAG
```

ในที่นี้เราจะใช้ตัวอักษรพิมพ์ใหญ่คือลำดับนิวคลีโอไทด์ที่เป็นแอกซอน และตัวอักษรพิมพ์เล็ก แทนลำดับนิวคลีโอไทด์ที่เป็นอินทรอน

เริ่มจากการหาข้อมูลของโดเนอร์ไซต์ ในตัวอย่างที่ 4.1 มีตำแหน่งที่นิวคลีโอไทด์เป็น 'g' และตำแหน่งถัดไปเป็น 't' เช่นในตำแหน่งที่ 48, 107, 171, ... ซึ่งค้นหาพบทั้งหมด 35 ตำแหน่ง เป็นตำแหน่งโดเนอร์ไซต์จริง 1 ตำแหน่งคือตำแหน่งที่ 281 หรือสังเกตได้จากในตัวอย่างคือเปลี่ยนจากตัวพิมพ์ใหญ่เป็นตัวพิมพ์เล็ก ซึ่งจะนำมาเก็บไว้ในกลุ่มโดเนอร์ไซต์จริง ส่วนตำแหน่งที่เหลืออีก 34 ตำแหน่ง เก็บไว้ในกลุ่มโดเนอร์ไซต์เท็จ

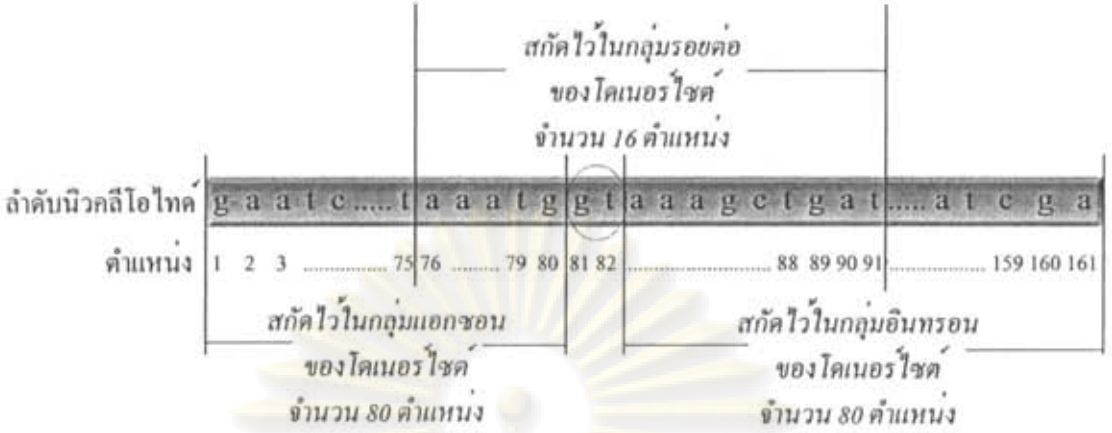
การค้นหาตำแหน่งแอกเซพเตอร์ไซต์นั้น จะค้นหาตำแหน่งที่พบนิวคลีโอไทด์ 'a' และตำแหน่งถัดไปเป็น 'g' เช่นในตำแหน่งที่ 65, 83, 92, ... ค้นหามาได้จำนวน 47 ตำแหน่ง เป็นแอกเซพเตอร์ไซต์จริง 1 ตำแหน่งคือตำแหน่งที่ 749 สังเกตจากตัวพิมพ์เล็กเป็นตัวพิมพ์ใหญ่ ซึ่งจะเก็บไว้ในกลุ่มแอกเซพเตอร์ไซต์จริง ตำแหน่งที่เหลือเก็บไว้ในกลุ่มแอกเซพเตอร์เท็จ #

ดังนั้นเมื่อนำข้อมูลทั้งหมดมาทำการค้นหาตำแหน่งสไปลไซต์ จะได้ผลของการค้นหาในตารางที่ 4.1

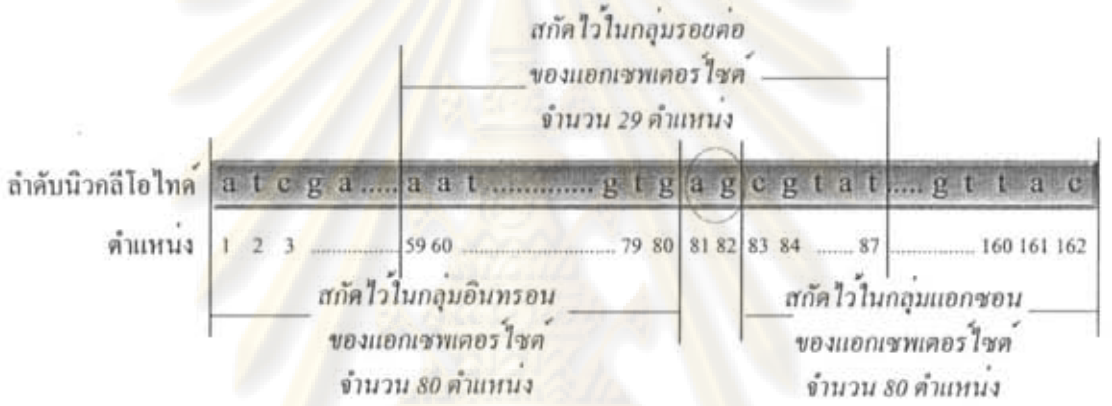
ตารางที่ 4.1 จำนวนของสไปลไซต์ ทั้งกลุ่มจริงและกลุ่มเท็จ จากยีนของมนุษย์ 1,115 ยีน

สไปลไซต์	สไปลไซต์จริง (true sites)	สไปลไซต์เท็จ (false sites)
โดเนอร์ไซต์	5,733	478,981
แอกเซพเตอร์ไซต์	5,733	650,089

จากการค้นหาตำแหน่งข้างต้น เราได้ทำการการสกัดลำดับนิวคลีโอไทด์จากตำแหน่งที่ค้นพบตามความยาวที่ต้องการต่าง ๆ กัน ซึ่งจะมีตำแหน่งที่ค้นพบแต่มีความยาวของลำดับนิวคลีโอไทด์ไม่ตรงตามที่กำหนด โดยตำแหน่งเหล่านี้จะทำการคัดออก แสดงตัวอย่างการสกัดได้ในรูปที่ 4.1 และ 4.2



รูปที่ 4.1 รูปแสดงการสกัดลำดับนิวคลีโอไทด์ไว้ในกลุ่มต่าง ๆ ของ โดเนอร์ไซต์



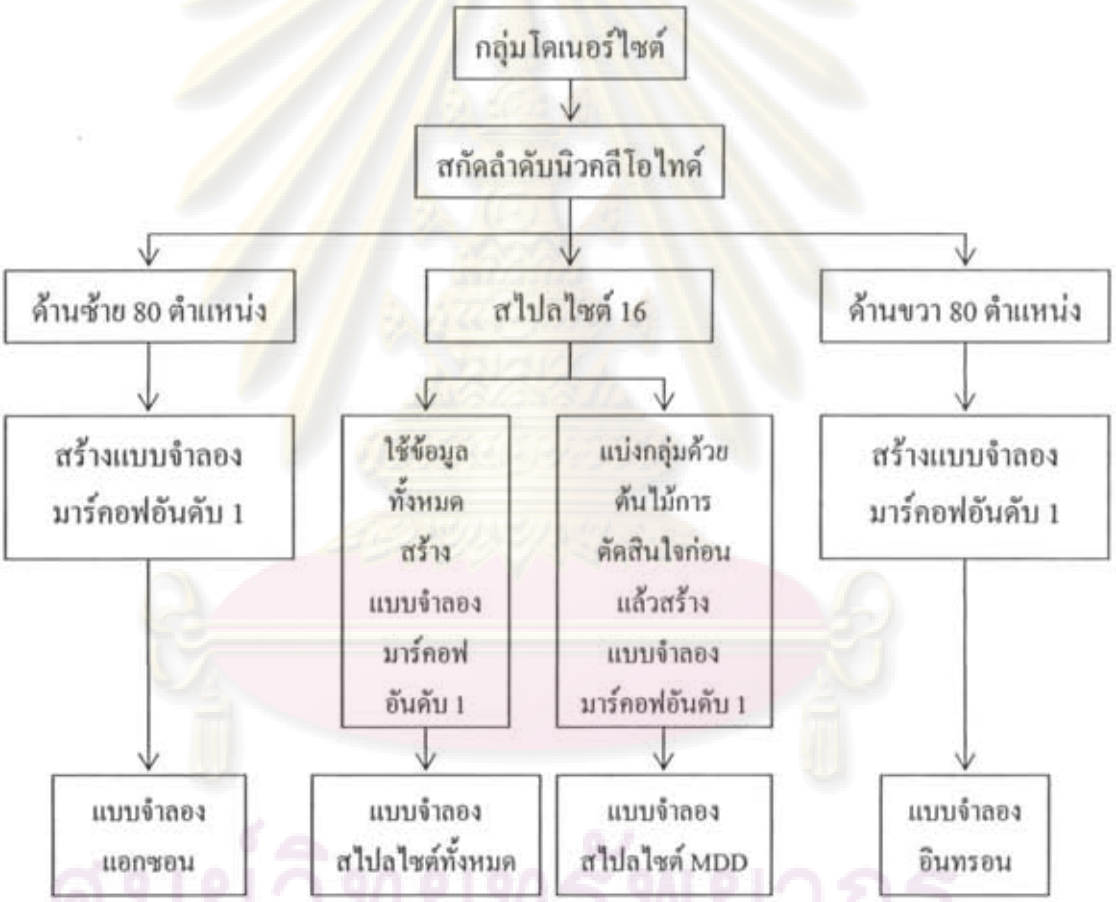
รูปที่ 4.2 รูปแสดงการสกัดลำดับนิวคลีโอไทด์ไว้ในกลุ่มต่าง ๆ ของแอกเซพเตอร์ไซต์

สำหรับโดเนอร์ไซต์และแอกเซพเตอร์ไซต์ ซึ่งจะทำการสกัดลำดับนิวคลีโอไทด์ ความยาว 16 ตำแหน่งสำหรับโดเนอร์ไซต์ และความยาว 29 ตำแหน่งสำหรับแอกเซพเตอร์ไซต์ ซึ่งลำดับนิวคลีโอไทด์ที่ได้จากการสกัดและนำมาจัดกลุ่มนั้นจะนำมาสร้างแบบจำลองมาร์คอฟบริเวณสไปลไซต์ รวมทั้งจะทำการสกัดลำดับนิวคลีโอไทด์ความยาว 80 ตำแหน่ง เก็บไว้ในกลุ่มแอกซอนและอินทรอน [3] ตามรูปที่ 4.1 และ 4.2 สำหรับสไปลไซต์แต่ละประเภท โดยที่ลำดับนิวคลีโอไทด์กลุ่มแอกซอนจะนำมาสร้างแบบจำลองมาร์คอฟบริเวณแอกซอน และลำดับนิวคลีโอไทด์กลุ่มอินทรอน จะนำมาสร้างแบบจำลองอินทรอน โดยที่ลำดับนิวคลีโอไทด์จากตำแหน่งสไปลไซต์จริงของแต่ละประเภทกลุ่ม ก็จัดไว้ในกลุ่มจริงของแบบจำลองนั้น ๆ และจากสไปลไซต์เท็จจะจัดไว้ในกลุ่มเท็จของแบบจำลองนั้น ๆ เช่นกัน ซึ่งผลลัพธ์จากการสกัดแสดงในตารางที่ 4.2

ตารางที่ 4.2 จำนวนลำดับนิวคลีโอไทด์ของโคเนอร์ไซต์และแอกเซพเตอร์ไซต์ ในกลุ่มต่าง ๆ

จำนวน	โคเนอร์ไซต์		แอกเซพเตอร์ไซต์	
	จริง	เท็จ	จริง	เท็จ
กลุ่มสไปลไซต์	5,709	478,219	5,732	648,113
กลุ่มแอกซอน	5,223	474,971	5,731	474,842
กลุ่มอินทรอน	5,425	643,421	5,731	643,883

ขั้นตอนการสร้างแบบจำลอง



แผนภาพที่ 4.2 การสร้างแบบจำลองของโปรแกรม GeneSplicer ในโคเนอร์ไซต์

โปรแกรม GeneSplicer มีการสร้างแบบจำลองที่ใช้ในการคำนวณคะแนนทั้งหมด 8 แบบจำลองสำหรับสไปลไซต์แต่ละประเภท โดยแบ่งออกเป็นกลุ่มจริงและกลุ่มเท็จ กลุ่มละ 4 แบบจำลอง ดังแสดงได้ในแผนภาพที่ 4.2 โดย 4 แบบจำลองนี้ประกอบด้วย แบบจำลองจากกลุ่มแอกซอน แบบจำลองจากกลุ่มอินทรอน และอีก 2 แบบจำลองที่เหลือจากกลุ่มสไปลไซต์ ซึ่งใช้วิธีสร้างแตกต่างกันคือ ในแบบจำลองแรกนั้นจะใช้ข้อมูลกลุ่มสไปลไซต์ทั้งหมดในการสร้าง กับอีกแบบจำลองจะใช้ต้นไม้การตัดสินใจ MDD เข้ามาแบ่งกลุ่มสไปลไซต์ในใบแต่ละบัพก่อน แล้วสร้างแบบจำลองในแต่ละใบนั้น

ในแบบจำลองทั้ง 8 นั้นเป็นแบบจำลองมาร์คอฟอันดับ 1 ทั้งหมด แต่ในตำแหน่งแรกจะเป็นแบบจำลองมาร์คอฟอันดับ 0 ซึ่งเขียนสมการของการคิดคะแนนได้เป็น

$$S(i,j) = \ln\left(f\left(x_i\right)\right) + \sum_{k=i+1}^{j} \ln\left(\frac{f\left(x_{k-1}x_k\right)}{f\left(x_{k-1}\right)}\right) \qquad (4.1)$$

$S(i,j)$ คือ คะแนนของลำดับนิวคลีโอไทด์จากตำแหน่ง i ถึงตำแหน่ง j

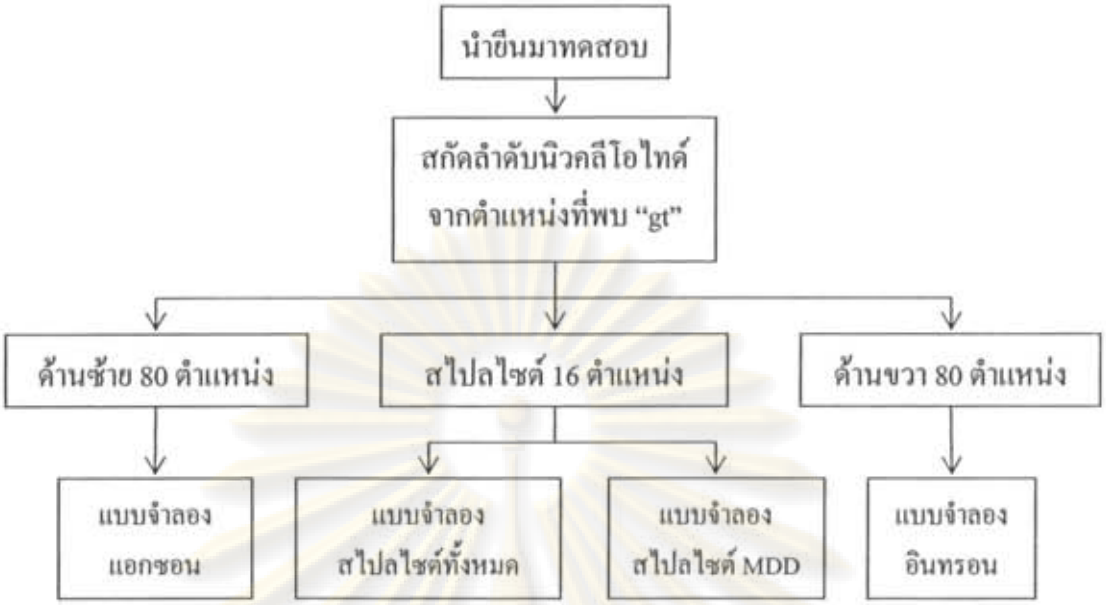
x_i คือ นิวคลีโอไทด์ x (a, c, g, t) ที่ตำแหน่ง i

$f\left(x_i\right)$ คือ ความน่าจะเป็นที่นิวคลีโอไทด์ x_i (a, c, g, t) ที่ตำแหน่ง i

$f\left(x_{k-1}x_k\right)$ คือ ความน่าจะเป็นที่นิวคลีโอไทด์ x_{k-1} ที่ตำแหน่ง $k-1$ และพบนิวคลีโอไทด์ x_k ที่ตำแหน่ง k

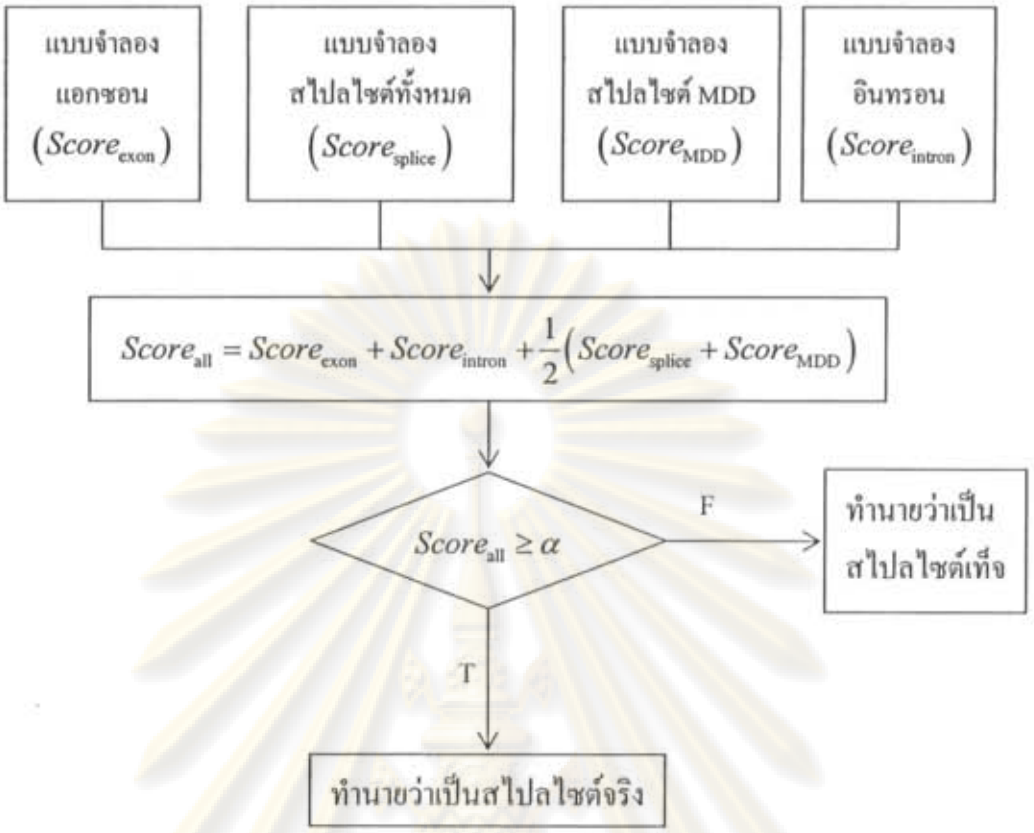
ขั้นตอนการทำนายตำแหน่งสไปลไซต์

การทำนายตำแหน่งสไปลไซต์ด้วยการคิดคะแนนจากลำดับนิวคลีโอไทด์ที่ความยาวต่าง ๆ ได้แสดงในแผนภาพที่ 4.3 – 4.5



แผนภาพที่ 4.3 การสกัดลำดับนิวคลีโอไทด์มาใช้คำนวณคะแนน

แผนภาพที่ 4.4 การคำนวณคะแนนในแต่ละแบบจำลอง



แผนภาพที่ 4.5 การคำนวณคะแนนของโปรแกรม GeneSplicer เพื่อหาคะแนนรวมมาทำนาย

เมื่อนำลำดับนิวคลีโอไทด์มาคิดคะแนนให้กับลำดับนิวคลีโอไทด์ที่ต้องการนำมาทดสอบ ซึ่งคะแนนจะมาจากทั้ง 8 แบบจำลองเป็นดังสมการ

$$Score_{all}(i) = Score_{exon}(i) + Score_{intron}(i) + \frac{1}{2}\left[Score_{splice}(i) + Score_{MDD}(i)\right] \quad (4.2)$$

$$Score_{model}(i) = S_{true}(i-k, i+l) - S_{false}(i-k, i+l) \quad (4.3)$$

โดยที่ i คือ ตำแหน่งที่ i ของสไปลไซต์ที่นำมาคิดคะแนน

$Score_{model}(i)$ คือ คะแนนรวมของสไปลไซต์ที่ตำแหน่ง i จากแบบจำลอง model (exon, intron, splice, MDD)

$S_{true}(i,j)$ คือ คะแนนของลำดับนิวคลีโอไทด์ ที่ตำแหน่ง i ถึง j จากแบบจำลองกลุ่มจริง

$S_{false}(i,j)$ คือ คะแนนของลำดับนิวคลีโอไทด์ ที่ตำแหน่ง i ถึง j จากแบบจำลองกลุ่มเท็จ

$Score_{all}(i)$ คือ คะแนนรวมของสไปลไซต์ที่ตำแหน่ง i จากทั้ง 4 แบบจำลอง

k, l คือ ค่าคงตัว ใช้สำหรับกำหนดความยาวของแต่ละสไปลไซต์

โดยที่มีค่าคงตัว k,l สำหรับแต่ละกลุ่มเป็นดังนี้

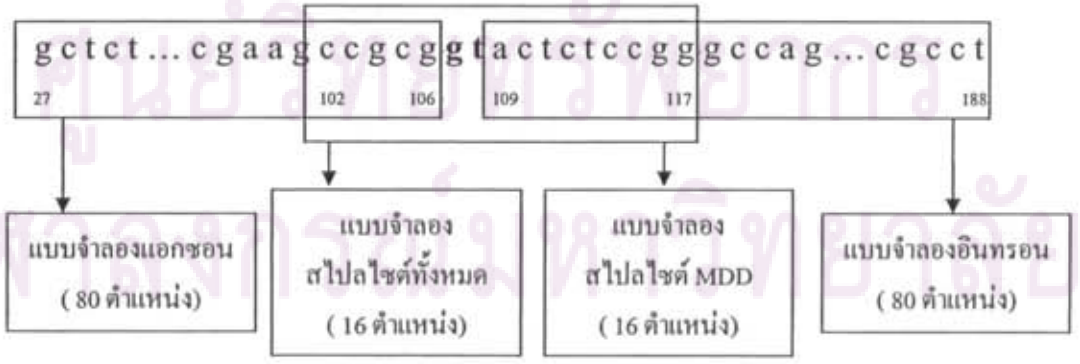
ตารางที่ 4.3 ค่าคงที่ k,l สำหรับแต่ละลำดับนิวคลีโอไทด์ของแต่ละสไปลไซต์

ประเภท / ค่าคงที่	โดเนอร์ไซต์			แอกเซพเตอร์ไซต์		
	แอกซอน	อินทรอน	สไปลไซต์	แอกซอน	อินทรอน	สไปลไซต์
k	81	-2	-5	-2	81	26
l	-1	81	10	81	-1	2

จะแสดงตัวอย่างในการคิดคะแนนเมื่อมีลำดับนิวคลีโอไทด์เข้ามาในตัวอย่างที่ 4.2

ตัวอย่างที่ 4.2 จากยีนในตัวอย่างที่ 4.1 พบว่าตำแหน่งนิวคลีโอไทด์ที่ 107 และ 108 เป็น "gt" จึงนำมาพิจารณาคำนวณคะแนนของโปรแกรม GeneSplicer

เริ่มจากสกัดลำดับนิวคลีโอไทด์ความยาว 162 ตำแหน่ง จากตำแหน่งที่ 27 จนถึง 188 จะได้ ลำดับออกมาเป็น "gctc tgctaccatt tccttacgtc tctgcttcgc tcagcgatgc aaaacgcgcg aggcacggca gagggccgaa gccgcg**gt**ac tctccgggcc aggcccgccc ctcggccgcc ggcggcgcag cacgggattc cccggccgct gtccagcgct ggccgcct" จากนั้นจะนำลำดับนี้ไปคำนวณคะแนนจากแบบจำลองที่ได้สร้างไปแล้วข้างต้น ทั้ง 4 แบบจำลอง โดยจะสกัดลำดับนิวคลีโอไทด์ตั้งแต่ตำแหน่งที่ 27 ถึง 106 จำนวน 80 ตำแหน่ง ("gctc tgctaccatt tccttacgtc tctgcttcgc tcagcgatgc aaaacgcgcg aggcacggca gagggccgaa gccgcg") นำไปคิดคะแนนในแบบจำลองแอกซอน และสกัดลำดับนิวคลีโอไทด์ตั้งแต่ตำแหน่งที่ 109 จนถึง 188 จำนวน 80 ตำแหน่ง ("ac tctccgggcc aggcccgccc ctcggccgcc ggcggcgcag cacgggattc cccggccgct gtccagcgct ggccgcct") เพื่อใช้คิดคะแนนในแบบจำลองอินทรอน สุดท้ายจะสกัดลำดับนิวคลีโอไทด์ตั้งแต่ตำแหน่งที่ 102 ถึง 117 จำนวน 16 ตำแหน่ง ("ccgcg**gt**ac tctccgg") ซึ่งจะสกัดมาคิดคะแนนในแบบจำลองสไปลไซต์ทั้งหมด กับแบบจำลองสไปลไซต์ MDD เขียนเป็นแผนภาพการสกัดได้ในแผนภาพที่ 4.6



แผนภาพที่ 4.6 การสกัดนิวคลีโอไทด์ ในตัวอย่างที่ 4.2

เริ่มจากการพิจารณาว่า ลำดับนิวคลีโอไทด์ที่สกัดมาคำนวณด้วยแบบจำลองสไปลไซต์ MDD นั้นจะนำมาพิจารณาไปอยู่ในใบของต้นไม้การตัดสินใจ MDD สำหรับโคเนอร์ไซต์ ในรูปที่ 2.9 พิจารณาในตำแหน่งที่ 10 ของลำดับนิวคลีโอไทด์ ("ccgcgg**g**tac tctccgg") ไม่ใช่ 'g' ดังนั้นจึงจัดอยู่ในใบที่ 3 และจะใช้เมทริกซ์เปลี่ยนสถานะของใบที่ 3 ในการคำนวณ

ในส่วนของการคิดคะแนนในแต่ละแบบจำลองโปรแกรม GeneSplicer ได้เก็บเมทริกซ์เปลี่ยนสถานะในรูปของ ลอการิทึมของความน่าจะเป็น ซึ่งในแต่ละแบบจำลองจะมีการ

1) แบบจำลองสไปลไซต์ทั้งหมด $\left(Score_{\text{splice}}\right)$ คิดคะแนนได้เป็น

$$S(1,16)=\ln\left(f\left(c_1\right)\right)+\ln\left(\frac{f\left(c_1c_2\right)}{f\left(c_1\right)}\right)+\ln\left(\frac{f\left(c_2g_3\right)}{f\left(c_2\right)}\right)+\ldots+\ln\left(\frac{f\left(g_{15}g_{16}\right)}{f\left(g_{15}\right)}\right)$$

เมื่อแทนค่าจากเมทริกซ์เปลี่ยนสถานะที่ได้เรียนรู้มาแล้วนั้น จะได้

$$S_{\text{true}}=\left(-1.28886\right)+\left(-1.10886\right)+\left(-2.40853\right)+\ldots+\left(-0.906296\right)=-20.8091$$

และ $S_{\text{false}}=\left(-1.39358\right)+\left(-1.10994\right)+\left(-2.26182\right)+\ldots+\left(-1.19355\right)=-22.0408$

ดังนั้น จะได้ $Score_{\text{splice}}=S_{\text{true}}-S_{\text{false}}=1.23165$

2) แบบจำลองสไปลไซต์ MDD $\left(Score_{\text{MDD}}\right)$ จะนำเมทริกซ์เปลี่ยนสถานะของใบที่ 3 ของต้นไม้การตัดสินใจ MDD ในโคเนอร์ไซต์ คำนวณ $S_{\text{true}}=-21.133$ และ $S_{\text{false}}=-21.796$ ทำให้ $Score_{\text{MDD}}=0.662971$

3) แบบจำลองแอกซอน $\left(Score_{\text{exon}}\right)$ คำนวณ $S_{\text{true}}=-112.2\dot{1}6$ และ $S_{\text{false}}=-117.067$ ทำให้ $Score_{\text{exon}}=4.85095$

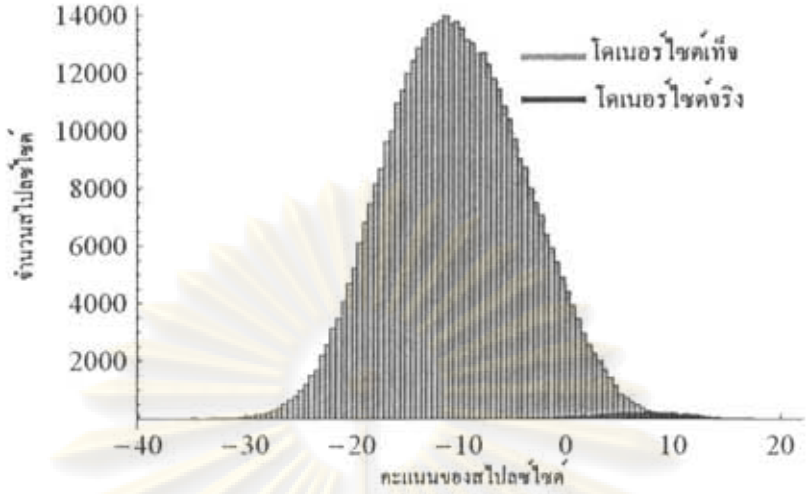
4) แบบจำลองอินทรอน $\left(Score_{\text{intron}}\right)$ คำนวณ $S_{\text{true}}=-114.217$ และ $S_{\text{false}}=-116.034$ ทำให้ $Score_{\text{intron}}=1.81696$

จากทั้ง 4 แบบจำลองจะคำนวณคะแนนสำหรับตำแหน่งนี้ได้เป็น

$$\begin{aligned}Score_{\text{all}}&=Score_{\text{exon}}+Score_{\text{intron}}+\frac{1}{2}\left(Score_{\text{splice}}+Score_{\text{MDD}}\right)\\&=4.85095+1.81696+\frac{1}{2}\left(1.23165+0.66971\right)\\&=7.61522\end{aligned}$$

ซึ่งจะนำคะแนนจากสไปลไซต์นี้ไปพิจารณาว่าผ่านเกณฑ์คะแนนหรือไม่ #

ดังนั้นเมื่อนำตำแหน่งของสไปลไซต์ทั้งกลุ่มจริงและกลุ่มเท็จมาคิดคะแนนเพื่อทำการทดสอบแยกกลุ่มจริงกับกลุ่มเท็จ แล้วนำมาเขียนฮิสโทแกรมในรูปที่ 4.3 เพื่อพิจารณาหาเกณฑ์คะแนนที่จะแบ่ง 2 กลุ่มนี้ให้ออกจากกัน (**ต่อไปจะขอเรียกเกณฑ์คะแนนนี้ว่า เกณฑ์** α)



รูปที่ 4.3 ฮิสโทแกรมคะแนนของโคเนอร์ไซต์กับจำนวนลำดับนิวคลีโอไทด์

คะแนนที่โปรแกรม GeneSplicer คำนวณได้นี้ ถ้าลำดับนิวคลีโอไทด์มีคะแนนมากกว่าเกณฑ์ α ก็จะทำนายว่าลำดับนิวคลีโอไทด์สไปลไซต์นี้น่าจะเป็นสไปลไซต์จริง ความผิดพลาดอันเกิดจากการทำนายจะพิจารณาที่โปรแกรมทำนายสไปลไซต์จริงเป็นสไปลไซต์เท็จ เรียกว่า false negative และการโปรแกรมทำนายสไปลไซต์เท็จเป็นสไปลไซต์จริง เรียกว่า false positive แสดงความสัมพันธ์ของกลุ่มต่าง ๆ ในตารางที่ 4.4

ตารางที่ 4.4 ความสัมพันธ์ของกลุ่มต่าง ๆ โดยใช้คะแนนและกลุ่มในการแบ่ง

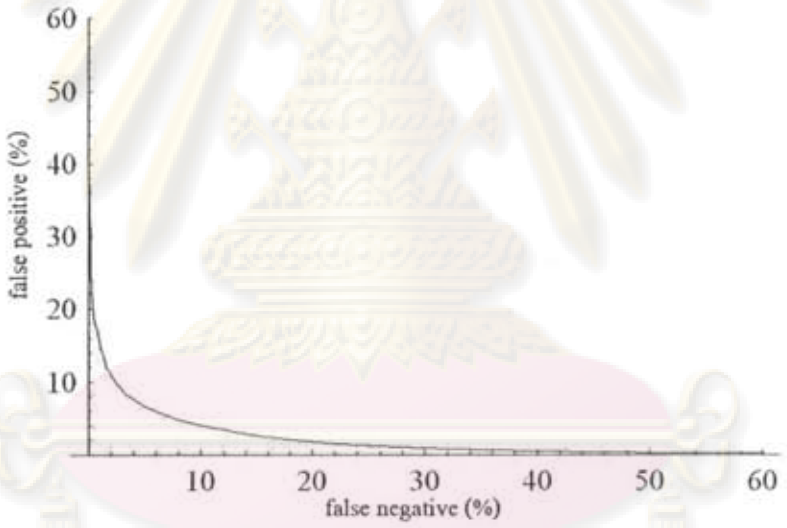
กลุ่มของลำดับนิวคลีโอไทด์	คะแนนของสไปลไซต์	
	ผ่านเกณฑ์	ไม่ผ่านเกณฑ์
สไปลไซต์จริง (true sites)	true positive	false negative
สไปลไซต์เท็จ (false sites)	false positive	true negative

โดยที่ความผิดพลาดจะพิจารณาร้อยละของ false negative (ร้อยละของจำนวน false negative เทียบกับจำนวนสไปลไซต์จริงทั้งหมด) **ซึ่งต่อไปจะขอเรียกว่า false negative** และร้อยละของ false positive (ร้อยละของจำนวน false positive เทียบกับจำนวนสไปลไซต์เท็จทั้งหมด) **ซึ่งต่อไปจะขอเรียกว่า false positive** จากเกณฑ์คะแนนที่กำหนดค่าหนึ่ง ๆ ซึ่งเมื่อใช้เกณฑ์คะแนนค่าต่าง ๆ แบ่งกลุ่มโคเนอร์ไซต์แล้วผลที่ได้แสดงในตารางที่ 4.5

ตารางที่ 4.5 false negative กับ false positive ที่เกณฑ์คะแนนต่าง ของกลุ่มโดเนอร์ไซต์

เกณฑ์คะแนนที่ใช้แบ่ง	false negative (%)	false positive (%)
-8.39	0.10	38.07
-5.97	0.19	25.64
-3.94	0.80	17.18
-2.00	2.01	10.87
0.43	7.00	5.37
3.36	20.00	1.81
4.86	29.99	0.90
6.05	39.99	0.50

สามารถเขียนกราฟ false negative กับ false positive จากตารางที่ 4.5 ในรูปที่ 4.4



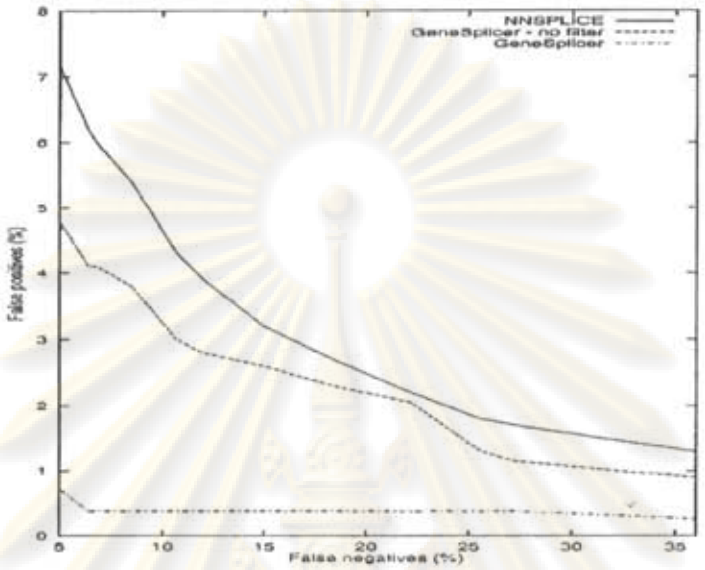
รูปที่ 4.4 กราฟแสดงความสัมพันธ์ระหว่าง false negative กับ false positive จากเกณฑ์ต่าง ๆ สำหรับโดเนอร์ไซต์

ผลของโปรแกรม GeneSplicer นั้น ตรวจสอบความแม่นยำโดยใช้วิธี 5-fold cross-validation มีวิธีการคือจะแบ่งข้อมูลที่มีออกเป็น 5 ส่วนอย่างสุ่มและใช้ 1 ส่วนเก็บไว้เพื่อเป็นกลุ่มทดสอบ และใช้ 4 ส่วนที่เหลือเป็นกลุ่มในการเรียนรู้ โดยความแม่นยำของโปรแกรมจะเกิดจากการเฉลี่ยผลของ false positive จากทั้ง 5 กลุ่มที่ทดสอบ ซึ่งสามารถแสดงได้ในตารางที่ 4.6

ตารางที่ 4.6 false negative กับ false positive จากการค้นหาตำแหน่งสไปลไซต์โดยวิธี 5-fold cross-validation ด้วยข้อมูลยีนของมนุษย์ [3]

	false negative (%)	false positives (%)					
		part. 1	part. 2	part. 3	part. 4	part. 5	average
แอกเซพเตอร์ไซต์ (กลุ่มจริง 5733 ตำแหน่ง)	3	7.18	10.57	13.22	7.17	8.34	9.3
	5	5.31	6.90	5.64	4.94	5.97	5.8
	7	4.31	5.21	4.99	3.91	5.16	4.7
	8	3.76	4.78	4.43	3.64	4.71	4.3
	10	3.16	4.10	3.80	3.22	4.16	3.7
	15	2.34	3.12	2.55	2.13	2.89	2.6
	20	1.60	2.48	2.07	1.46	2.17	2.0
	40	0.50	1.40	0.73	0.56	0.92	0.8
โดเนอร์ไซต์ (กลุ่มจริง 5733 ตำแหน่ง)	3	16.63	12.00	21.39	9.10	14.16	14.7
	5	5.98	6.66	5.91	5.21	8.02	6.4
	7	4.46	5.45	3.78	4.04	6.48	4.8
	8	3.96	4.34	3.45	3.39	5.58	4.1
	10	3.34	3.73	2.99	2.93	4.36	3.5
	15	2.41	2.65	2.02	1.99	3.39	2.5
	20	1.85	1.87	1.44	1.49	2.41	1.8
	40	0.75	0.52	0.51	0.66	0.99	0.7

และเมื่อทำการเปรียบเทียบความแม่นยำของโปรแกรมกับโปรแกรมค้นหาสไปลไซต์ของยีนอื่น ๆ นั้น ได้ใช้ข้อมูลทั้งหมดที่มีเป็นกลุ่มในการเรียนรู้ และใช้ข้อมูลของโปรแกรมอื่น ๆ ในการทดสอบ แสดงผลได้ในรูปที่ 4.5



รูปที่ 4.5 กราฟแสดงความสัมพันธ์ระหว่าง false negative กับ false positive สำหรับโดเนอร์ไซต์ของโปรแกรม GeneSplicer กับโปรแกรม NNSPLICE [3]

จากตารางที่ 4.5 เกณฑ์ α หนึ่งเกณฑ์จะได้ค่า false negative กับ false positive หนึ่งชุด เราคาดหวังให้เกณฑ์ α ทำให้ค่า false negative และ false positive ลดลง เมื่อเทียบค่ากับ false negative หรือ false positive ใด ๆ เดียวกัน ซึ่งนั่นหมายถึงความผิดพลาดอันเกิดจากการทำนายที่ลดน้อยลง

คะแนนของ GeneSplicer ที่คำนวณได้ เมื่อได้ทำการสังเกตและพิจารณาลักษณะของคะแนนในกลุ่มของ false positive ซึ่งยังคงมีคะแนนจากแบบจำลองกลุ่มจริงมาก หรือแบบจำลองกลุ่มเท็จน้อย ดังนั้นถ้ามีการปรับปรุงสร้างแบบจำลองมาร์คอฟอันดับ 1 ที่สร้างจากลำดับนิวคลีโอไทด์ในกลุ่ม false positive มาทดแทนแบบจำลองในกลุ่มเท็จเดิม โดยลำดับนิวคลีโอไทด์ใดที่มีลักษณะ false positive ย่อมมีคะแนนจากแบบจำลองกลุ่มนี้สูง และลำดับนิวคลีโอไทด์ที่ควรนำมาคิดคะแนนจากแบบจำลองนี้ควรมีคะแนนผ่านเกณฑ์ α มาแล้ว ซึ่งจากการปรับปรุงแบบจำลองในกลุ่มเท็จแทนด้วยแบบจำลองในกลุ่ม false positive นั้นจะขอเรียกแบบจำลองการคิดคะแนนใหม่นี้ว่า Enhanced GeneSplicer

ซึ่งจากแนวคิดดังกล่าว สามารถเขียนเป็นสมการในการคิดคะแนนใหม่นี้ได้เป็น

$$Score_{\text{Enchanced}}(i) = Score_{\text{exon}}(i) + Score_{\text{intron}}(i) + \frac{1}{2}\left[Score_{\text{splice}}(i) + Score_{\text{MDD}}(i)\right] \quad (4.4)$$

$$Score_{\text{model}}(i) = S_{true}(i-k, i+l) - S_{false\ positive}(i-k, i+l) \quad (4.5)$$

โดยที่

$S_{false\ positive}(i,j)$ คือคะแนนรวมจากแบบจำลองใหม่ที่สร้างจากกลุ่ม false positive ที่ผ่านเกณฑ์คะแนนจาก GeneSplicer

$Score_{\text{Enhanced}}(i)$ เป็นคะแนนที่รวมจากคะแนนใหม่ของแบบจำลองที่ปรับปรุงแล้ว

เมื่อพิจารณากลุ่มของ false positive ที่จะนำมาใช้สร้างแบบจำลอง Enhanced GeneSplicer นั้นจะต้องเกิดจากการคิดคะแนนเดิมก่อนและมีเกณฑ์ α ที่คัดกรองตำแหน่งของสไปลไซต์ซึ่งแต่ละเกณฑ์ α จะคัดเอา false positive ออกไม่เท่ากัน เราจึงทำการทดลองเลือกใช้เกณฑ์ α ที่ -5.97 ในโดเนอร์ไซต์มาทดลองเพื่อพิสูจน์วิธีการ Enhanced GeneSplicer นี้ว่าสามารถลดค่า false positive หรือ false negative ได้จริงหรือไม่

โดยการทดลองนี้เราได้นำกลุ่ม false positive ซึ่งมีคะแนนผ่านเกณฑ์ α ที่ -5.97 และนำลำดับนิวคลีโอไทด์ของกลุ่มนี้มาสร้างแบบจำลอง Enhanced GeneSplicer และพิจารณานำโดเนอร์ไซต์เฉพาะที่ผ่านเกณฑ์ $\alpha = -5.97$ มาคำนวณคะแนนจากแบบจำลอง Enhanced GeneSplicer ที่สร้าง และนำมาพิจารณาเกณฑ์จากคะแนนของแบบจำลอง Enhanced GeneSplicer **(ซึ่งต่อไปจะขอเรียกเกณฑ์ใหม่นี้ว่า เกณฑ์ β)** และหาค่า false negative กับ false positive รวมทั้งทดลองนำมาเปรียบเทียบกับค่า false negative กับ false positive ใน GeneSplicer เดิม โดยเลือกเปรียบเทียบผลจากค่า false negative เดียวกัน

ซึ่งสามารถเขียนแผนภาพการปรับปรุงแบบจำลอง Enhanced GeneSplicer ในแผนภาพที่ 4.3

แผนภาพที่ 4.6 แนวคิดการปรับปรุงโปรแกรม GeneSplicer ของโปรแกรม Enhanced GeneSplicer

ผลจากการใช้เกณฑ์ $\alpha = -5.97$ คัดเอาลำดับนิวคลีโอไทด์มาสร้างแบบจำลอง Enhanced GeneSplicer และคิดคะแนนจากแบบจำลองใหม่และใช้เกณฑ์ β คัดลำดับนิวคลีโอไทด์อีกครั้ง แสดงในตารางที่ 4.7

ตารางที่ 4.7 false negative และ false positive เปรียบเทียบ GeneSplicer กับ Enhanced GeneSplicer ที่สร้างจากเกณฑ์ $\alpha = -5.97$ ของโคเนอร์ไซต์

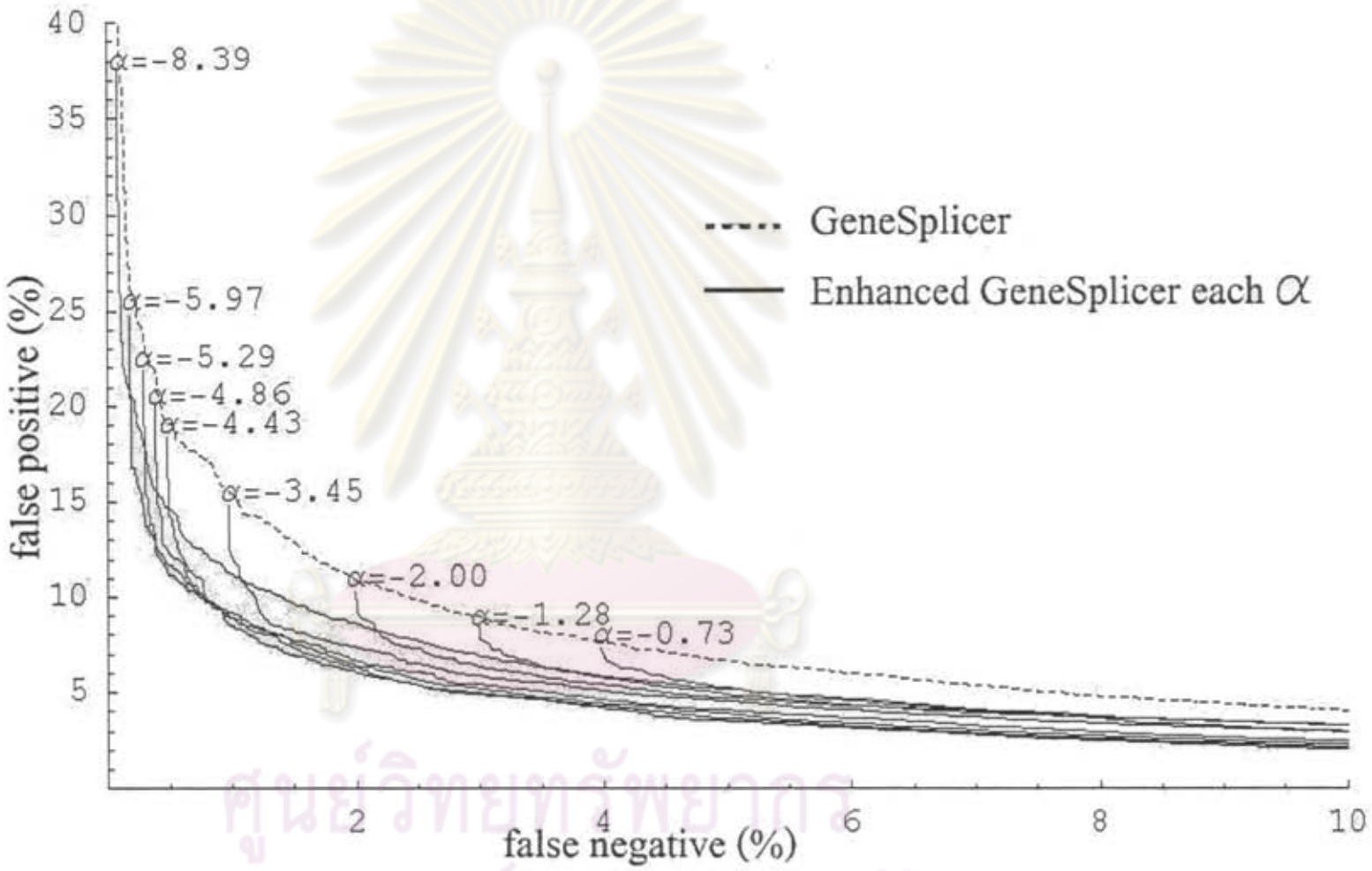
เกณฑ์ β	false negative (%)	false positive (%) GeneSplicer	false positive (%) Enhanced GeneSplicer
-5.06	0.19	25.64	16.78
-3.76	0.80	17.18	9.74
-1.81	2.01	10.87	6.04
-0.36	7.00	5.37	2.83
1.22	20.00	1.81	1.00
1.99	29.99	0.90	0.55
2.61	39.99	0.50	0.32

ผลในตารางที่ 4.7 ได้ยืนยันแนวคิดในการปรับปรุงโปรแกรม GeneSplicer ด้วยวิธีการของโปรแกรม Enhanced GeneSplicer ทำให้ค่า false positive มีค่าลดลงกว่าโปรแกรม GeneSplicer ที่มีการคิดคะแนนเพียงครั้งเดียว จากนั้นจึงได้พยายามทดลองนำเกณฑ์ α ต่าง ๆ มาทดลองปรับปรุงแบบจำลอง Enhanced GeneSplicer อีกหลายเกณฑ์ เพื่อหาเกณฑ์ α และ เกณฑ์ β ที่ให้ค่า false positive ลดลงได้มากที่สุด สำหรับการเปรียบเทียบที่ค่า false negative เดียวกัน จึงได้แสดงผลของการทดลอง ซึ่งคัดเลือกเกณฑ์ α โดยพิจารณาจากการทำให้ค่า false negative แตกต่างกัน โดยเลือก α ที่ -8.39, -5.97, -5.29, -4.87, -4.43, -3.45, -2.00, -1.28, -0.73 ซึ่งให้ค่า false negative จากแต่ละเกณฑ์เป็น 0.1%, 0.2%, 0.3%, 0.4%, 0.5%, 1%, 2%, 3% และ 4% ตามลำดับ แสดงผลการคำนวณในตารางที่ 4.8

ตารางที่ 4.8 false negative เทียบกับ false positive จากคะแนนของ Enhanced GeneSplicer ที่สร้างขึ้นตามเกณฑ์ α ต่าง ๆ กัน ในโดเนอร์ไซต์

false negative (%)	false positive (%) ของ Enhanced GeneSplicer จากเกณฑ์ α								
	-8.39 (0.1%)	-5.97 (0.2%)	-5.29 (0.3%)	-4.86 (0.4%)	-4.43 (0.5%)	-3.45 (1.0%)	-2.00 (2.0%)	-1.28 (3.0%)	-0.73 (4.0%)
0.2	20.53	**16.78**	-	-	-	-	-	-	-
0.5	14.68	**11.17**	11.71	12.37	14.30	-	-	-	-
2.0	8.57	**6.04**	6.35	6.40	6.63	7.35	9.28	-	-
3.3	6.56	4.77	**4.76**	4.93	5.19	5.61	6.06	6.72	-
5.0	5.20	**3.53**	3.66	3.92	4.08	4.56	4.73	4.92	5.24
10.0	3.01	**2.05**	2.16	2.30	2.51	2.99	3.33	3.33	3.32
20.0	1.54	**1.00**	1.07	1.19	1.36	1.84	2.30	2.35	2.36
30.0	0.82	**0.55**	0.58	0.67	0.81	1.23	1.70	1.81	1.82

ซึ่งจากตารางที่ 4.8 ค่า false positive ที่เป็นตัวหนานั้นแสดงว่าเลือกเกณฑ์ α และเกณฑ์ β ที่ทำให้ค่า false positive มีค่าที่น้อยที่สุดจากค่า false negative เดียวกัน ซึ่งจากตัวอย่างที่แสดง โปรแกรม Enhanced GeneSplicer ที่ปรับปรุงมีค่าร้อยละของ false positive น้อยกว่าของ GeneSplicer ได้จากการเลือกเกณฑ์ α และเกณฑ์ β ที่เหมาะสมต่าง ๆ กันตามค่าร้อยละของ false negative ที่ต้องการ จึงได้ทำการเขียนกราฟ false negative กับค่า false positive ของเกณฑ์ α แต่ละเกณฑ์เปรียบเทียบหาเกณฑ์ β ที่เหมาะสมของเกณฑ์ α ใด ๆ ที่ให้ค่า false positive น้อยที่สุด เมื่อเปรียบเทียบด้วยค่า false negative เดียวกัน แสดงในรูปที่ 4.6



รูปที่ 4.6 กราฟแสดง false negative กับ false positive จากเกณฑ์ α ต่าง ๆ กัน

ซึ่งสามารถเขียนแผนภาพการสร้างแบบจำลองในแผนภาพที่ 4.7 และ 4.8

แผนภาพที่ 4.7 การเรียนรู้ของโปรแกรม Enhanced GeneSplicer

แผนภาพที่ 4.8 การนำเกณฑ์ α มาสร้างแบบจำลองจากกลุ่ม false positive

บทที่ 5

ผลการทดลอง

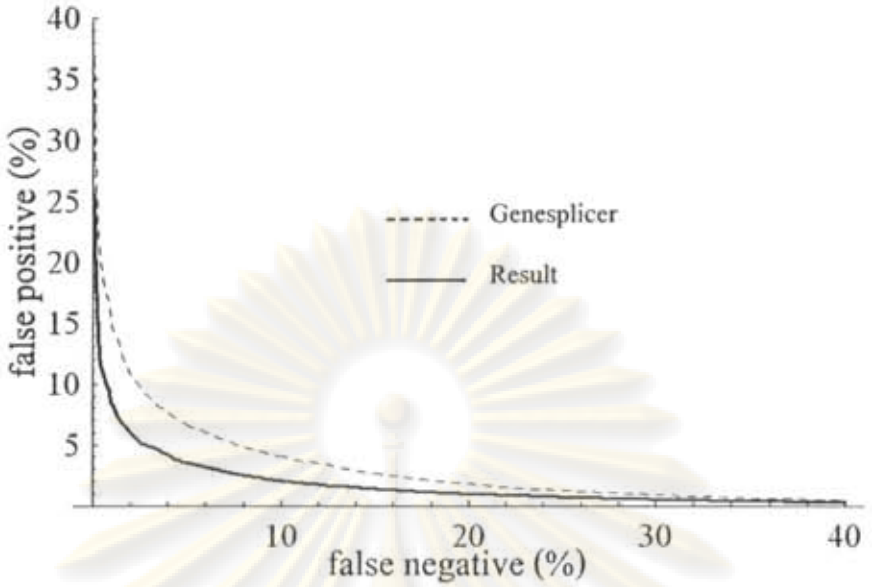
ผลของโดเนอร์ไซต์

จากวิธีการปรับปรุงของโปรแกรม Enhanced GeneSplicer ที่ได้กล่าวไปในบทที่ 4 นั้น เมื่อนำตำแหน่งของสไปลไซต์ของข้อมูลที่ใช้ในการเรียนรู้ มาคิดคะแนนเดิมและกำหนดเกณฑ์ α ที่เลือกไว้อย่างเหมาะสมแล้ว ถ้าตำแหน่งของสไปลไซต์ใดที่ผ่านเกณฑ์ α แล้ว จึงนำมาคิดคะแนนใหม่จากแบบจำลอง Enhanced GeneSplicer และเลือกเกณฑ์ β ที่ทำให้ค่า false negative ตามที่ต้องการไว้ แสดงผลจากการทดลองด้วยค่า false negative เทียบกับค่า false positive ในตารางที่ 5.1

ตารางที่ 5.1 ร้อยละของ false negative กับ false positive ของโดเนอร์ไซต์เปรียบเทียบผลลัพธ์ของโปรแกรม GeneSplicer กับ Enhanced GeneSplicer เมื่อใช้ข้อมูลทั้งหมดในการเรียนรู้และทดสอบ

false negative (%)	false positive (%) GeneSplicer	เกณฑ์ α	เกณฑ์ β	false positive (%) Enhanced GeneSplicer
0.2	25.50	-5.97	-5.06	18.48
0.5	19.06	-5.97	-3.34	11.16
1.0	15.40	-5.97	-2.06	8.49
2.0	10.88	-5.97	-1.81	6.07
5.0	6.67	-5.97	-0.75	3.53
10.0	4.05	-5.97	0.16	2.05
15.0	2.62	-5.97	0.72	1.42
20.0	1.81	-5.97	1.22	0.99
40.0	0.49	-5.97	2.61	0.32

กราฟในรูปที่ 5.1 ซึ่งแสดงความสัมพันธ์ระหว่างค่าร้อยละของ false negative กับค่าร้อยละของ false positive เปรียบเทียบผลลัพธ์ของ Enhanced GeneSplicer กับของ GeneSplicer ซึ่งแสดงให้เห็นว่า ผลลัพธ์ของ Enhanced GeneSplicer ที่ได้พัฒนาขึ้นนั้นมีค่าร้อยละของ false positive น้อยกว่าวิธีของโปรแกรม GeneSplicer



รูปที่ 5.1 กราฟแสดงความสัมพันธ์ระหว่างร้อยละ false negative กับ false positive ของโดเนอร์ไซต์เปรียบเทียบผลของโปรแกรม GeneSplicer กับ Enhanced GeneSplicer

และเมื่อทำการตรวจสอบความแม่นยำของ Enhanced GeneSplicer ด้วยการทำ 5-fold cross validation เช่นเดียวกับโปรแกรม GeneSplicer และเปรียบเทียบผลที่ได้เพื่อพิสูจน์ประสิทธิภาพของโปรแกรม Enhanced GeneSplicer แสดงในตารางที่ 5.2 สำหรับโดเนอร์ไซต์

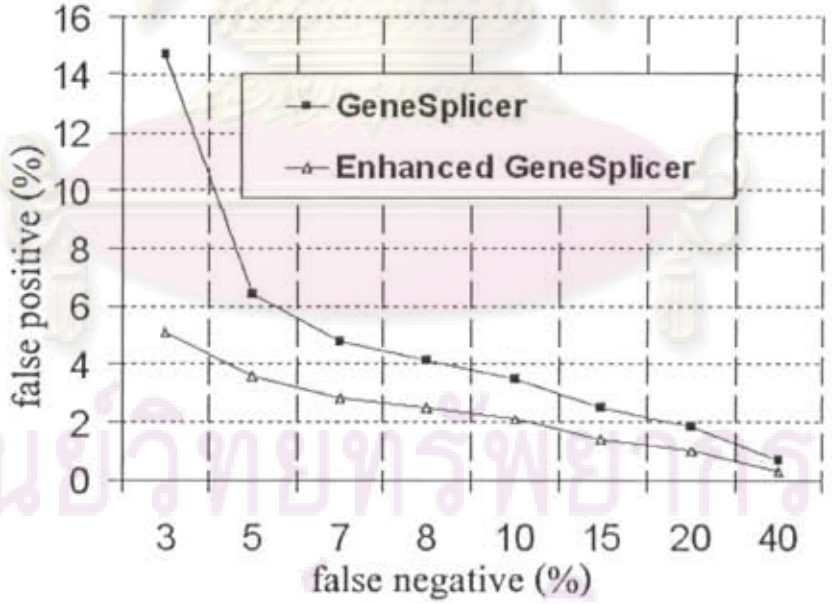
ตารางที่ 5.2 5-fold cross validation ของ Enhanced GeneSplicer ในโดเนอร์ไซต์

false negative (%)	false positive (%)					
	กลุ่มที่ 1	กลุ่มที่ 2	กลุ่มที่ 3	กลุ่มที่ 4	กลุ่มที่ 5	Average
0.2	27.08	23.15	22.95	19.58	21.00	22.75
0.5	12.94	13.97	12.85	12.66	12.92	13.07
1.0	9.68	9.08	9.32	9.35	9.85	9.46
2.0	6.53	6.53	6.48	6.37	6.51	6.48
5.0	3.44	3.75	3.68	3.67	3.67	3.64
10.0	2.10	2.17	2.12	2.05	2.07	2.10
20.0	1.04	1.04	1.08	1.00	1.03	1.04
40.0	0.31	0.35	0.34	0.33	0.33	0.33

ตารางที่ 5.3 เปรียบเทียบผลของ Enhanced GeneSplicer กับ GeneSplicer เมื่อใช้ 5-fold cross validation เพื่อทดสอบความแม่นยำของโปรแกรม

false negative (%)	false positive (%) (GeneSplicer)	false positive (%) (Enhanced GeneSplicer)
3	14.7	5.1
5	6.4	3.6
7	4.8	2.8
8	4.1	2.5
10	3.5	2.1
15	2.5	1.4
20	1.8	1.0
40	0.7	0.3

และเมื่อเปรียบเทียบผลกับโปรแกรม GeneSplicer ในจากตารางที่ 5.3 แสดงให้เห็นว่าโปรแกรม Enhanced GeneSplicer ที่พัฒนาขึ้นนั้นมีความแม่นยำกว่าโปรแกรม GeneSplicer ในทุก ๆ ค่าร้อยละของ false negative จากรูปที่ 5.2 ที่แสดงให้เห็นการลดลงของจำนวนของตำแหน่งสไปลไซต์ที่ทำนายผิดพลาดได้ลดลงกว่าของ GeneSplicer



รูปที่ 5.2 กราฟแสดงความสัมพันธ์ระหว่างร้อยละ false negative กับ false positive ของโดเนอร์ไซต์เปรียบเทียบผลของโปรแกรม GeneSplicer กับ Enhanced GeneSplicer จาก การทดสอบความแม่นยำด้วย 5-fold cross validation

เราได้ทดสอบความแม่นยำเพิ่มเติมด้วยการใช้ค่า precision และ recall โดยสามารถคำนวณได้ดังนี้

$$\text{recall} = \frac{\text{true positive}}{\text{true positive + false negetive}}$$

$$\text{precision} = \frac{\text{true positive}}{\text{true positive + false positive}}$$

ซึ่งจากสูตรดังกล่าวจะเห็นได้ว่าค่า recall นั้นคือค่าร้อยละของการทำนายที่ถูกต้องของสไปลไซต์จริงนี่เอง เมื่อได้เปรียบเทียบแล้วพบว่าค่า precision ยิ่งมีค่ามากถือว่ามีคุณภาพในการทำนายที่ดี ในตารางที่ 5.4 ได้แสดงการเปรียบเทียบของโปรแกรม Enhanced GeneSplicer กับ GeneSplicer

ตารางที่ 5.4 ค่า precision และ recall ของ Enhanced GeneSplicer กับ GeneSplicer ทดสอบความแม่นยำของโปรแกรม ในโดเนอร์ไซต์

false negative (%)	recall	precision of GeneSplicer	precision of Enhanced GeneSplicer
0.2	0.99809	0.0414	0.0620
0.5	0.99503	0.0548	0.0900
1	0.99006	0.0667	0.1146
2	0.97993	0.0909	0.1527
5	0.94991	0.1368	0.2299
10	0.90002	0.1978	0.3277
15	0.84993	0.2648	0.4001
20	0.80004	0.3297	0.4712
40	0.60008	0.5738	0.6762

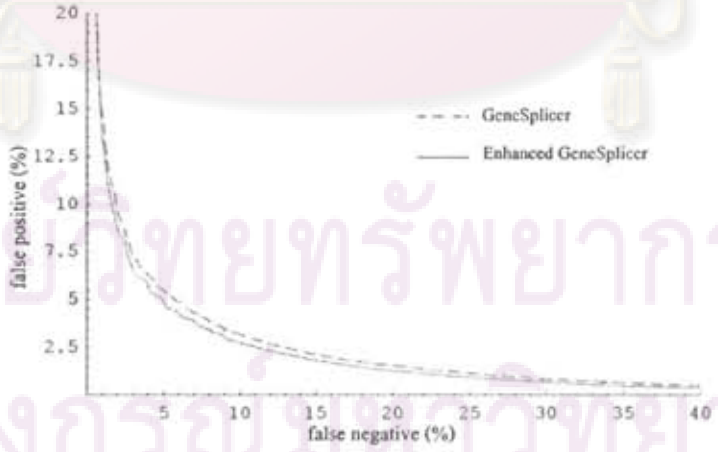
ผลของแอกเซพเตอร์ไซต์

ด้วยการใช้วิธีการ Enhanced GeneSplicer ในการเรียนรู้และทดสอบกับแอกเซพเตอร์ไซต์เช่นเดียวกับในโดเนอร์ไซต์ ผลลัพธ์ที่ได้จากการใช้ข้อมูลชุดเดียวกันในการเรียนรู้และทดสอบ เพื่อหาเกณฑ์ α และ β ที่มีค่า false positive ที่น้อยที่สุดในค่า false negative เดียวกัน แสดงผลได้ในตารางที่ 5.5

ตารางที่ 5.5 ร้อยละของ false negative กับ false positive ระหว่างโปรแกรม GeneSplicer กับ Enhanced GeneSplicer ในแอกเซพเตอร์ไซต์ เมื่อใช้ข้อมูลชุดเดียวกันในการเรียนรู้และทดสอบ

false negative (%)	false positive (%) GeneSplicer	เกณฑ์ α	เกณฑ์ β	false positive (%) Enhanced GeneSplicer
0.2	38.14	-11.53	-14.86	34.28
0.5	23.00	-6.70	-15.78	22.58
1.0	14.58	-3.62	-12.01	13.42
2.0	9.50	-1.49	-9.9	8.60
5.0	5.41	1.08	-11.52	4.81
10.0	3.17	2.63	-5.77	2.76
15.0	2.13	3.89	-5.20	1.83
20.0	1.55	4.63	-3.73	1.28
40.0	0.49	6.38	-0.18	0.37

และเมื่อแสดงการเปรียบเทียบผลลัพธ์อย่างชัดเจนมากยิ่งขึ้น จึงได้ใช้กราฟแสดงในรูปที่ 5.3



รูปที่ 5.3 กราฟแสดงความสัมพันธ์ระหว่างร้อยละ false negative กับ false positive ของแอกเซพเตอร์ไซต์เปรียบเทียบผลของโปรแกรม GeneSplicer กับ Enhanced GeneSplicer

ซึ่งผลลัพธ์ที่ได้จากโปรแกรม Enhanced GeneSplicer ก็ยังแสดงให้เห็นว่ามีประสิทธิภาพในการทำนายที่ดีกว่า GeneSplicer สำหรับแอกเซพเตอร์ไซต์เช่นเดียวกัน และเมื่อทำการทดสอบความแม่นยำของโปรแกรมด้วยเทคนิค 5-fold cross validation ผลลัพธ์ที่ได้แสดงในตารางที่ 5.6 และ 5.7

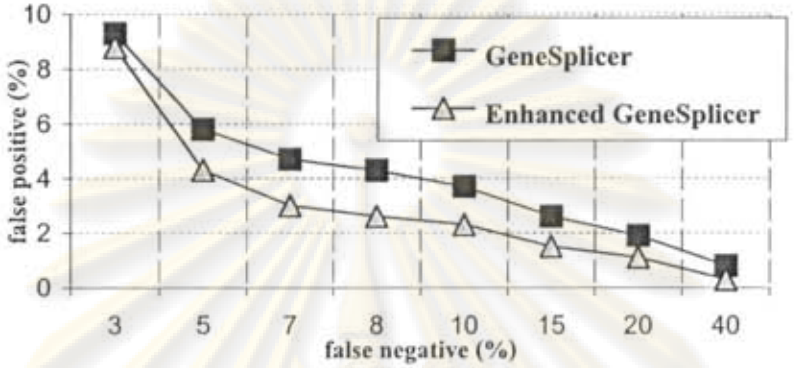
ตารางที่ 5.6 แสดงผลลัพธ์ 5-fold cross validation ของ Enhanced GeneSplicer

false negative (%)	false positive (%)					
	กลุ่มที่ 1	กลุ่มที่ 2	กลุ่มที่ 3	กลุ่มที่ 4	กลุ่มที่ 5	Average
0.2	52.63	31.12	30.43	30.04	34.28	35.70
0.5	46.94	18.42	17.55	19.40	19.29	24.32
1.0	25.35	12.76	13.10	12.45	12.74	15.28
2.0	17.00	8.14	8.25	7.96	8.30	9.93
5.0	4.44	4.22	4.32	4.33	4.34	4.33
10.0	2.32	2.33	2.33	2.34	2.38	2.34
20.0	1.20	1.06	1.05	1.10	1.09	1.10
40.0	0.31	0.34	0.32	0.31	0.32	0.32

ตารางที่ 5.7 เปรียบเทียบผลของ Enhanced GeneSplicer กับ GeneSplicer เมื่อใช้ 5-fold cross validation เพื่อทดสอบความแม่นยำของโปรแกรม

false negative (%)	false positive (%) (GeneSplicer)	false positive (%) (Enhanced GeneSplicer)
3	9.3	8.8
5	5.8	4.3
7	4.7	3.0
8	4.3	2.6
10	3.7	2.3
15	2.6	1.5
20	1.9	1.1
40	0.8	0.3

และเมื่อเปรียบเทียบผลกับโปรแกรม GeneSplicer ในจากตารางที่ 5.6 แสดงให้เห็นว่าโปรแกรม Enhanced GeneSplicer ที่พัฒนาขึ้นนั้นมีความแม่นยำยำกว่าโปรแกรม GeneSplicer ในทุก ๆ ค่าร้อยละของ false negative จากรูปที่ 5.2 ที่แสดงให้เห็นการลดลงของจำนวนของตำแหน่งสไปลไซต์ที่ทำนายผิดพลาดได้ลดลงกว่าของ GeneSplicer



รูปที่ 5.4 กราฟแสดงความสัมพันธ์ระหว่างร้อยละ false negative กับ false positive ของแอกเซพเตอร์ไซต์เปรียบเทียบผลของโปรแกรม GeneSplicer กับ Enhanced GeneSplicer จาก การทดสอบความแม่นยำด้วย 5-fold cross validation

ตารางที่ 5.8 ค่า precision และ recall ของ Enhanced GeneSplicer กับ GeneSplicer ทดสอบความแม่นยำของโปรแกรม ในแอกเซพเตอร์ไซต์

false negative (%)	recall	precision of GeneSplicer	precision of Enhanced GeneSplicer
0.2	0.99809	0.02179	0.02418
0.5	0.99503	0.03552	0.03615
1	0.99006	0.05464	0.05908
2	0.97993	0.08072	0.08842
5	0.94991	0.13004	0.14392
10	0.90002	0.19463	0.21726
15	0.84993	0.25356	0.28335
20	0.80004	0.30524	0.34726
40	0.60008	0.51036	0.57989

บทที่ 6
สรุปผลและข้อเสนอแนะ

สรุปผล

โปรแกรม GeneSplicer เป็นโปรแกรมในการทำนายตำแหน่งสไปลไซต์ที่ดีโปรแกรมหนึ่ง แต่จากผลในบทที่ 5 ได้แสดงให้เห็นอย่างชัดเจนว่าโปรแกรม Enhanced GeneSplicer ที่ได้ปรับปรุงการทำนายตำแหน่งสไปลไซต์จากโปรแกรม GeneSplicer เดิม ได้ลดความผิดพลาดในการทำนายตำแหน่งสไปลไซต์ให้น้อยลงกว่าเดิมได้ โดยในโดเนอร์ไซต์สามารถลดการทำนายโดเนอร์ไซต์เท็จได้จากเดิมที่ผิดพลาด 25.5% ลดลงเหลือ 18.48% จากการทำนายโดเนอร์ไซต์จริงผิดพลาดที่ 0.2% และแอกเซพเตอร์ไซต์สามารถลดการทำนายแอกเซพเตอร์ไซต์เท็จจากเดิม 38.14% ลดเหลือ 34.18% จากการทำนายแอกเซพเตอร์ไซต์จริงผิดพลาดที่ 0.2% โดยผลนี้ทดลองกับสไปลไซต์ของยีนมนุษย์บางส่วนเท่านั้น

การลด false negative

เนื่องจากนักชีววิทยาสามารถหาสไปลไซต์จริงได้ถูกต้องและแม่นยำ เพียงแต่ใช้เวลาในการทดสอบที่นานและตำแหน่งของสไปลไซต์จริงจะต้องมีอยู่ครบถ้วนในยีนนั้นถึงจะสามารถหาได้อย่างถูกต้อง ซึ่งโปรแกรม Enhanced GeneSplicer จะทำหน้าที่ช่วยลดภาระในการทำงานของการค้นหาสไปลไซต์ของยีนได้จากการลดตำแหน่งที่ไม่น่าจะเป็นสไปลไซต์ลงได้ สำหรับการใช้งานโปรแกรม หากผู้ใช้ต้องการให้โปรแกรมสามารถทำนายหาตำแหน่งสไปลไซต์จริงได้ครบทุกตำแหน่งโดยที่อนุญาตให้มีสไปลไซต์เท็จหลุดมาได้นั้น โปรแกรม Enhanced GeneSplicer ก็สามารถทำได้ตามที่ผู้ใช้ต้องการได้ด้วยเช่นกัน โดยจากผลการทดลองหากเรากำหนดค่า $\alpha = -18.37$ และ $\beta = -22.15$ ที่ทำให้ค่า false negative ที่ 0% ทำให้ลด false positive ลงได้จาก 98.6% เหลือ 90.17% สำหรับในโดเนอร์ไซต์ และในแอกเซพเตอร์ไซต์กำหนดค่า $\alpha = -26.35$ และ $\beta = -36.33$ ลดลงจาก 78.68% เหลือ 78.53% อันจะทำให้นักชีววิทยาสามารถทำงานด้วยความรวดเร็วมากยิ่งขึ้นในการตรวจสอบตำแหน่งสไปลไซต์จริงโดยที่ตำแหน่งสไปลไซต์จริงของยีนนั้นก็ยังคงไว้ให้นักชีววิทยาได้ทำการทดสอบหาตำแหน่งที่ถูกต้องอย่างแท้จริงได้ต่อไป

ข้อเสนอแนะ

แต่เนื่องจากการคำนวณของโปรแกรม Enhanced GeneSplicer นั้นอาจจะมีการคำนวณคะแนนจากตำแหน่งสไปลไซต์หนึ่ง ๆ ได้มากถึง 2 ครั้ง ซึ่งทำให้ต้องใช้เวลาในการคำนวณมากกว่าโปรแกรม GeneSplicer อยู่แล้ว แต่เนื่องจากเรามุ่งปรับปรุงโปรแกรมเพื่อลดความผิดพลาดอันเกิดจากการทำนายตำแหน่งสไปลไซต์ที่ผิดพลาด ทำให้เราต้องใช้เวลาในการตรวจสอบแยกหาข้อผิดพลาดเหล่านั้นอย่างช่วยไม่ได้

การปรับปรุงโปรแกรม Enhanced GeneSplicer ที่น่าสนใจประการหนึ่งคือ ในเรื่องของการคำนวณคะแนน ซึ่งถ้าเราต้องการลดความผิดพลาดในการทำนายตำแหน่งสไปลไซต์ให้ลดลงกว่าเดิมอีกนั้น สามารถทำได้โดยการปรับแบบจำลองกลุ่มเท็จนี้ด้วยวิธีการเดิมอีก และเพิ่มการคิดคะแนนได้อีกตามต้องการ แต่นั้นย่อมหมายความว่า เวลาในการคำนวณของโปรแกรมที่จะปรับปรุงย่อมใช้เวลามากขึ้นตามไปด้วย ต่อมาคือในส่วนของผลกับแอกเซพเตอร์ไซต์ซึ่งยังต้องใช้แบบจำลองจากเกณฑ์ α ที่มีจำนวนเยอะมาก การปรับปรุงโปรแกรมในส่วนของแบบจำลองที่นำมาใช้คัดกรองอาจจะต้องมีการพัฒนากันต่อไป ในส่วนสุดท้ายนี้ โปรแกรม GeneSplicer มีความสามารถในการทำนายสิ่งมีชีวิตอื่น ๆ นอกเหนือจากมนุษย์ได้อีก อาทิเช่น *Arabidopsis Thaliana* เป็นต้น ด้วยเทคนิคของโปรแกรม Enhanced GeneSplicer น่าจะได้มีการทดลองต่อไปว่าสามารถลดความผิดพลาดในการทำนายของโปรแกรมลงได้เช่นเดียวกัน

รายการอ้างอิง

[1] วสันต์ จันทราทิตย์, วีระพงศ์ อุลิดานนท์. 2544. ชีวสารสนเทศศาสตร์. สถาบันบัณฑิตวิทยาศาสตร์และเทคโนโลยี. กรุงเทพมหานคร : สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ.

[2] http://en.wikipedia.org/wiki/Bioinformatics.

[3] Pertea, M., Lin, X. and Salzberg, S.L.. 2001. GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Research Vol. 29, No. 5 : 1185-1190.

[4] Wiwatwanich A.. 2003. Digital Signal Processing Analysis of DNA sequences. Master's Thesis Computational science, Department of Mathematics Faculty of Science Chulalongkorn University.

[5] Karp, G.. 2005. Cell and molecular biology. 4th ed. Von Hoffmann Press : John Wiley & Son.

[6] ปรีชา สุวรรณพินิจ, นงลักษณ์ สุวรรณพินิจ. 2537. ชีววิทยา 2. 1st ed. กรุงเทพมหานคร : สำนักพิมพ์จุฬาลงกรณ์มหาวิทยาลัย.

[7] http://www.exonhit.com//index.php?page=63.

[8] Salzberg, S.L., Searls, D.B., Kasif, S.. 1999. Computational methods in molecular biology. 2nd ed. New comprehensive biochemistry Volume 32. Netherlands : Elsevier.

[9] Burge, C. and Karlin, S.. 1997. Prediction of Complete Gene Structures in Human Genomic DNA. Journal of Molecular Biology 268 : 78-94.

[10] Stirzaker D.. 1997. Elementary Probability: Cambridge University Press. United Kingdom. University Press.

[11] Brunak,S., Engelbrecht,J. and Knudsen,S. 1991. Prediction of human mRNA donor and acceptor sites from the DNA sequence. Journal of Molecular Biology Vol. 39, Issue 5 : 1257-1255.

[12] Solovyev,V.V., Salamov,A.A. and Lawrence,C.B. 1994. Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. Nucleic Acids Research Vol. 22, No. 24 : 5156-5163.

[13] Reese,M.G., Eeckman,F.H., Kulp,D. and Haulsser,D. 1997. Improved splice site detection in Genie. Jounal of Computational Biology Vol. 4, No. 3 : 311-324.

[14] Saxonov, S.,Daizaddeh, I., Fedorov, A. and Gibert W.. 2000. EID: Exon-Intron Database-an exhaustive database of protein-coding intron-containing genes. Nucleic Acids Research Vol. 28, No. 1 : 185-190.

ประวัติผู้เขียนวิทยานิพนธ์

นายสืบกุล กาญจนสุกร์ เกิดวันพฤหัสบดีที่ 21 พฤษภาคม พุทธศักราช 2524 ที่จังหวัดพิษณุโลก จบการศึกษาครุศาสตร์บัณฑิต สาขาวิชาการสอนคณิตศาสตร์ ภาควิชามัธยมศึกษา คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2545 ต่อมาเป็นผู้รับทุนพัฒนาอาจารย์สาขาขาดแคลนของประเทศ วิชาคณิตศาสตร์ จากมหาวิทยาลัยนเรศวร วิทยาเขตสารสนเทศพะเยา ในปีการศึกษา 2547 เพื่อศึกษาต่อในระดับปริญญามหาบัณฑิต สาขาวิชา วิทยาการคณนา ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย