

การวิจัยครั้งนี้เป็นการวิจัยเพื่อพิจารณาการแยกส่วนดีเอ็นเอบริเวณที่ไม่สามารถแปลรหัสของกรดอะมิโนได้ โดยวิธีการทดสอบแบบอัตราส่วนภาวะน่าจะเป็น โดยใช้ข้อมูลลำดับเบสของสิ่งมีชีวิตประเภทข้าว (*Oryza sativa*) จาก The Institute of Genomic Research (TIGR) พร้อมทั้งเปรียบเทียบความแม่นยำของโปรแกรมในการวิเคราะห์หาตำแหน่งการแยกส่วนดีเอ็นเอ 3 โปรแกรม ได้แก่ FGENESH RiceHMM และ GeneMark.hmm ซึ่งปรากฏผลดังนี้

โปรแกรม FGENESH มีความถูกต้องและแม่นยำมากที่สุด ซึ่งการตรวจสอบความแม่นยำในระดับนิวคลีโอไทด์ มีค่า Sensitivity, Specificity, Correlation Coefficient และ Simple Matching Coefficient เท่ากับ 0.859, 0.972, 0.805 และ 0.900 ตามลำดับ สำหรับการวิเคราะห์หาตำแหน่งการแยกส่วนดีเอ็นเอ โดยวิธีการทดสอบแบบอัตราส่วนภาวะน่าจะเป็น มีความถูกต้องในการทดสอบคิดเป็นร้อยละ 84.87 และวิธีการดังกล่าวสามารถแยกส่วนดีเอ็นเอได้ดีเมื่อ ข้อมูลลำดับเบสส่วนของ Intron ที่นำมาวิเคราะห์มีความแปรปรวนของการเกิดลำดับเบสแต่ละตำแหน่งไม่มากนัก ซึ่งจะส่งผลให้ค่าอัตราส่วนภาวะน่าจะเป็นมีความเฉพาะเจาะจงในการเกิดเบสแต่ละตำแหน่ง ในทางตรงกันข้ามกับส่วนของ Exon ที่มีความแปรปรวนของการเกิดลำดับเบสแต่ละตำแหน่งสูง ทำให้ค่าอัตราส่วนภาวะน่าจะเป็นไม่แตกต่างกัน เป็นเหตุให้การกระจายค่าคะแนนของข้อมูลจริงแตกต่างจากข้อมูลสุ่มและข้อมูลจีโนม ส่วนลำดับเบสคอนเซนซัสพบว่า ข้อมูลข้าวที่ใช้ในงานวิจัยนี้มีรูปแบบของตำแหน่ง Splice Site สอดคล้องกับรูปแบบของตำแหน่ง Splice Site ทั่วไปที่ทราบจากความรู้ที่ได้จากวิธีการทางห้องปฏิบัติการ

(วิทยานิพนธ์มีจำนวนทั้งสิ้น 73 หน้า)

Abstract

172373

This thesis presents the use of a log likelihood ratio test in non-coding genes. The tested data used in this research is DNA sequence of rice (*Oryza sativa*), which is a part of rice genome sequence data. This research focused on efficiency of, bases on accuracy and precision, splice site detection of three particular gene finding programs namely FGENESH, RiceHMM and GeneMark.hmm.

It has been found that FGENESH program gave the highest accuracy of splice site detection on the data set used. The sensitivity, specificity, correlation coefficient and simple matching coefficient were 0.859, 0.972, 0.805 and 0.900 respectively. Log likelihood ratio test was applied to distinguish the genuine splice sites from the pseudo sites. According to methods and data set used, FGENESH was 84.87 % correctness. The test is the most powerful while each individual base has a different probability distribution depending on its position in the sequence and each base in exons are independent and equally likely. Score distribution of real splice site data differs from those of random and genome data. Consensus pattern of splice site of rice sequence data used in this study agrees with general consensus pattern of splice site known experimentally.

(Total 73 pages)