Thesis Title RNA Secondary Structure Prediction and RNA Family Classification Using Conditional Random Field Models

Author Mr. Sitthichoke Subpaiboonkit

Degree Master of Science (Bioinformatics)

Thesis Advisory Committee Assoc. Prof. Dr. Jeerayut Chaijaruwanich Advisor

Dr. Chinae Thammarongtham Co-advisor

ABSTRACT

RNAs including non-coding RNAs (ncRNAs) have important biological functions in living cells based on the structures of non-coding RNAs. RNA family classification is a required task for annotating sequenced genomes. RNA families are defined based on RNA transcripts which perform similar functions in different species. Such functions have a strong relationship with RNA secondary structures but not their primary sequences. Thus RNA sequences alone are not sufficient to classify RNA families. Here, I focused on computational RNA secondary structure prediction by exploring primary sequences and complementary base-pair interactions, together with a suitable innovative feature extraction based on natural RNA's loop and stem

characteristics using the conditional random fields method (CRFs), In addition, I also focused on computational RNA family classification by exploring primary sequences of RNAs and their secondary structures as the selected feature to classify the RNA family also using the CRF method. Both problems are treated as a sequence labeling. The CRFs models can predict the RNA secondary structures and can classify the RNA families of the test RNAs with optimal F-score prediction between 56.61% - 98.20% and between 98.77% - 99.32%, respectively.

ชื่อเรื่องวิทยานิพนธ์ การทำนายโครงสร้างระดับทุติยภูมิของอาร์เอ็นเอ และการจำแนกชนิดของอาร์เอ็นเอด้วยวิธีคอนดิชันนอลแรนดอมฟิลด์

ผู้เขียน นาย สิทธิโชค ทรัพย์ไพบูลย์กิจ

ปริญญา วิทยาศาสตรมหาบัณฑิต (ชีวสารสนเทศศาสตร์)

คณะกรรมการที่ปรึกษาวิทยานิพนธ์ รศ.ดร. จีรยุทธ ไชยจารุวณิช อาจารย์ที่ปรึกษาหลัก
ดร. ชิเน ธำมรงค์ธรรม อาจารย์ที่ปรึกษาร่วม

บทคัดย่อ

อาร์เอ็นเอชนิดต่างๆ รวมทั้งอาร์เอ็นเอที่ไม่แปลรหัสเป็นโปรตีน มีหน้าที่ที่สำคัญทางชีววิทยาในเซลล์สิ่งมีชีวิต ปัจจัยสำคัญประการหนึ่งที่เกี่ยวข้องกับหน้าที่ของอาร์เอ็นเอที่ไม่แปลรหัสเป็นโปรตีนคือโครงสร้างของอาร์เอ็นเอชนิดนั้นๆ ดังนั้นการทำนายโครงสร้างระดับทุติยภูมิของอาร์เอ็นเอจึงเป็นงานวิจัยที่ได้รับความสนใจ นอกจากนี้ การจัดจำแนกชนิดของอาร์เอ็นเอนั้นเป็นเรื่องที่สำคัญเช่นกัน ทั้งนี้เพราะเป็นสิ่งจำเป็นในการศึกษาจีโนมของสิ่งมีชีวิตที่มีการหาลำดับเบสแล้ว โดยการจัดจำแนกอาร์เอ็นเอแต่ละชนิด อาศัยคุณลักษณะด้านหน้าที่การทำงานของอาร์เอ็นเอแต่ละตัวในการจัดจำแนก โดยอาร์เอ็นเอที่ทำหน้าที่เดียวกันหรือคล้ายคลึงกัน จะถูกจัดไว้ในกลุ่มเดียวกัน แม้ว่าจะเป็นอาร์เอ็นเอที่มาจากสิ่งมีชีวิตต่างชนิดกัน โดยอาร์เอ็นเอที่มีหน้าที่การทำงานเหมือนกันนั้น มักจะมีโครงสร้างระดับทุติยภูมิที่เหมือนกันหรือคล้ายกัน แม้ว่าอาร์เอ็นเอนั้นๆ จะมีลำดับเบสที่ต่างกัน ดังนั้นลำดับเบสของสายอาร์เอ็นเอเพียงอย่างเดียว จึงไม่สามารถใช้ในการจำแนกชนิดของอาร์เอ็นเอได้อย่างมีประสิทธิภาพมากนัก งานวิทยานิพนธ์นี้มุ่งเน้นที่การทำนายโครงสร้างอาร์เอ็นเอระดับทุติยภูมิด้วยวิธีการคำนวณ โดยใช้ประโยชน์จากข้อมูลลำดับเบสของสายอาร์เอ็นเอ และความสัมพันธ์ของการเข้าคู่กันของเบสคู่สม (ทำให้เกิดโครงสร้างระดับทุติยภูมิ) ในการสร้างวิธีสกัดคุณลักษณะของข้อมูลขึ้นมา โดยหลักการสำคัญคือการประยุกต์ใช้ความรู้ทางธรรมชาติของลูป และสเต็มของอาร์เอ็นเอที่เป็นส่วนประกอบของโครงสร้างระดับทุติยภูมิ แล้ว

นำมาประยุกต์ใช้กับแบบจำลองคอนดิชันนอล แรนดอมฟิลด์ หรือ ซีอาร์เอฟ ในการฝึกฝน และการทำนายข้อมูลด้วยเครื่อง นอกจากนี้ ในงานวิทยานิพนธ์นี้ ยังได้เพิ่มการทดลองการจำแนกชนิดของอาร์เอ็นเอด้วยวิธีการคำนวณ ซึ่งข้อมูลดิบคือลำดับเบสของสายอาร์เอ็นเอ และใช้โครงสร้างระดับทุติยภูมิเป็นคุณลักษณะข้อมูลในการประยุกต์ใช้กับแบบจำลองซีอาร์เอฟ ในกระบวนการเรียนรู้ด้วยเครื่อง ปัญหาทั้งสองนี้ถูกจัดอยู่ในกลุ่มปัญหาแบบการติดฉลากข้อมูล ความถูกต้องที่วัดด้วยค่าคะแนนเอฟจากการทำนายโครงสร้างระดับทุติยภูมิของอาร์เอ็นเอด้วยแบบจำลองซีอาร์เอฟ อยู่ระหว่าง 56.61-98.20% และค่าคะแนนเอฟที่ได้จากการจำแนกชนิดของอาร์เอ็นเอด้วยวิธีซีอาร์เอฟอยู่ระหว่าง 98.77% - 99.32%