

Thesis Title Predictive Models for Protein Classification

Author Mr. Watshsara Shoombuatong

Degree Doctor of Philosophy (Computer Science)

Thesis Advisory Committee	Assoc. Prof. Dr. Jeerayut Chaijaruwanich	Advisor
	Asst. Prof. Dr. Thanasak Mouktonglang	Co-advisor
	Assoc. Prof. Dr. Supapon Cheevadhanarak	Co-advisor
	Assoc. Prof. Dr. Kitsana Waiyamai	Co-advisor

ABSTRACT

Methods of protein sequence based computational prediction have a broad range of real-world applications. In this study, three types of gene and protein sequences were used for analysis, i.e. Human Leukocyte Antigen (HLA) Gene, Human Immunodeficiency Virus type 1 (HIV-1), and protein crystallization. We used a different computational method to classify each sequence dataset. Since, intuitively different proteins do not have the same sequence characteristics; no single computational method can predict all types of protein sequences.

In the first work, the k -nearest neighbor (k -NN) classifier based on spectrum kernel was used to classify HLA genes. The spectrum kernel was first used to transform HLA genes with suitable length of spectrum. The k -NN classifier was then applied to classify HLA genes. Both main classes and subclasses were classified using

this combination approach. Secondly, the HIV-1 coreceptor usage, especially in CRF_AF subtype, was classified by using a Support Vector Machine (SVM) based logistic model tree (LMT). The LMT method was used to select informative positions on the V3 loop sequence before classifying them with the SVM classifier. Finally, we were interested in predicting protein crystallization using the scoring card method (SCM). The SCM method exploits the effectiveness of dipeptide composition to construct a classifier. The SCM method is a simple and interpretable approach, and also comparable to the SVM method.

ชื่อเรื่องวิทยานิพนธ์

ตัวแบบการทำนายเพื่อการจำแนกกลุ่มโปรตีน

ผู้เขียน

นายวัชร ชุ่มบัวทอง

ปริญญา

วิทยาศาสตร์สุขภาพบัณฑิต (วิทยาการคอมพิวเตอร์)

คณะกรรมการที่ปรึกษาวิทยานิพนธ์

รศ.ดร. จิรยุทธ ไชยจารุณิช อาจารย์ที่ปรึกษาหลัก

ผศ.ดร. ธนะศักดิ์ หมวกทองหลาง อาจารย์ที่ปรึกษาร่วม

รศ.ดร. สุภาภรณ์ ชีวะธนรักษ์ อาจารย์ที่ปรึกษาร่วม

รศ.ดร. กฤษณะ ไวยมัย อาจารย์ที่ปรึกษาร่วม

บทคัดย่อ

การจำแนกกลุ่มของโปรตีนโดยใช้ตัวแบบทางคอมพิวเตอร์นั้นเป็นที่รู้จักในวงกว้างในการนำไปใช้ ในงานวิจัยชิ้นนี้ได้แบ่งข้อมูลออกเป็นสามประเภท, นั่นคือ ฮิวแมนลิวโคไซแอนติเจน (HLA), เชื้อไวรัสเอชไอวีชนิดที่ 1 (HIV-1), และ โปรตีนที่สามารถสร้างผลึกได้ เราได้ใช้ตัวแบบทางคอมพิวเตอร์ที่แตกต่างกันในการจัดกลุ่มข้อมูล ด้วยเหตุว่าแต่ละโปรตีนนั้นย่อมมีลักษณะเด่นที่แตกต่างกันออกไป ดังนั้นจึงไม่มีตัวแบบทางคอมพิวเตอร์ใดที่สามารถจำแนกกลุ่มของโปรตีนได้อย่างดีเสมอ โดยงานชิ้นแรกเค-เนียร์เรสเนเบอร์ร่วมกับสเปกตรัมเคอเนลใช้ในการจำแนกฮิวแมนลิวโคไซแอนติเจน ซึ่งในขั้นตอนแรกสเปกตรัมเคอเนลใช้ในการแปลงข้อมูลโดยใช้ค่าเอสของสเปกตรัมเคอเนลที่เหมาะสม จากนั้นใช้ตัวจำแนกเค-เนียร์เรสเนเบอร์ในการจำแนกกลุ่มของฮิวแมนลิวโคไซแอนติเจน โดยทั้งกลุ่มหลักและกลุ่มย่อยของฮิวแมนลิวโคไซแอนติเจนใช้วิธีดังกล่าวในการจำแนกกลุ่ม งานที่สองจำแนกเชื้อไวรัสเอชไอวีชนิดที่ 1 (HIV-1) โดยใช้ซัพพอร์ตเวกเตอร์แมชชีนร่วมกับโลจิสติกโมเดลทรี ซึ่งโลจิสติกโมเดลทรีใช้ในการเลือกตำแหน่งสำคัญที่ให้ข้อมูลที่ดิบบนสายลำดับของนิวคลีโอไทด์ ก่อนที่ซัพพอร์ตเวกเตอร์แมชชีนจะจำแนกเชื้อไวรัสเอชไอวีชนิดที่ 1 งานสุดท้ายวิธีของบัตรคะแนนนำมาใช้ในการจำแนกโปรตีนที่สามารถสร้างผลึกได้หรือไม่ โดยวิธีนี้จะใช้ข้อมูลองค์ประกอบของไดเปปไทด์มาใช้ในการสร้างตัวแบ่งแยก ซึ่งวิธีของบัตรคะแนนเป็นวิธีที่ง่ายและสามารถตีความได้และสามารถนำไปเปรียบเทียบกับได้กับซัพพอร์ตเวกเตอร์แมชชีน