วิธีการขั้นสูงที่เครื่องเรียนรู้เพื่อการทำนายโครงสร้างทุติยภูมิของโปรตีน

นายเกษมสันต์ ดูภานุมาต

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาคณิตศาสตร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2548

ISBN 974-53-2806-5

ADVANCED MACHINE LEARNING METHOD FOR PREDICTION OF PROTEIN SECONDARY STRUCTURE

Mr Kasemsant Kuphanumat

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Phylosophy Program in Computer Science
Department of Mathematics
Faculty of Science
Chulalongkorn University
Academic year 2005

ISBN 974-53-2806-5

Thesis Title	ADVANCED MACHINE LEARNING METHOD FOR PREDICTION OF PROTEIN SECONDARY STRUCTURE
By	Kasemsant Kuphanumat
Field of Study	Mathematics
Thesis Advisor	Professor Chidchanok Lursinsap, Ph.D.

Accepted by the Faculty of Science, Chulalongkorn University in Partial Fulfillment of the Requirements for the Doctor's Degree

..... Dean of the Faculty of Science
(Professor Piamsak Menasveta, Ph.D)

THESIS COMMITTEE

..... Chairman
(Associate Professor Jack Asavanant, Ph.D.)

..... Thesis Advisor
(Professor Chidchanok Lursinsap, Ph.D.)

..... Member
(Associate Professor Prasit Palittapongarnpim, M.D.)

..... Member
(Lerson Tanasugarn, Ph.D.)

..... Member
(Siripun Sanguansintukul, Ph.D.)

เกษมสันต์ คูภานุมาต : วิธีการขั้นสูงที่เครื่องเรียนรู้เพื่อการทำนายโครงสร้างทุติยภูมิของโปรตีน. (ADVANCED MACHINE LEARNING METHOD FOR PREDICTION OF PROTEIN SECONDARY STRUCTURE) อ. ที่ปรึกษา : ศ. ดร. ชิดชนก เหลือสินทรัพย์, 67 หน้า. ISBN 974-53-2806-5.

วิทยานิพนธ์ฉบับนี้ได้นำเสนอวิธีการใหม่ในการเข้ารหัสลำดับอะมิโนแอซิดของโปรตีนโดยใช้วิธีการของมาคอล์ฟโปรเซส ด้วยเทคนิคนี้ ลักษณะเด่นที่สำคัญของลำดับสายโปรตีนจะถูกสกัดออกและนำไปใช้สร้างชุดของเวคเตอร์สำหรับเป็นข้อมูลในการสอนเพื่อการจำแนกประเภทด้วยซัพพอร์ตเวคเตอร์แมชชีน(SVM) ได้อย่างมีประสิทธิภาพ วิธีการที่ใช้ในงานวิจัยนี้ให้ผลลัพธ์ที่โดดเด่นกว่าวิธีการอื่นๆ ที่มีในปัจจุบันเป็นอย่างมาก ด้วยวิธีการจำแนกประเภทโดยใช้ SVM ร่วมกับวิธีการเข้ารหัสของข้อมูลโดยใช้มาคอฟทรานสิชั่นเมตริกสามารถวัดค่าความถูกต้องในการจำแนกแบบสามกลุ่มได้ดังนี้คือ Q_3 = 82.49%, SOV = 77.18% โดยการประเมินจากกลุ่มข้อมูลทดสอบมาตรฐานของโปรตีนจำนวน 513 สาย (CB513) ซึ่งผลที่ได้นับว่าเป็นการพัฒนาเข้าใกล้ขีดจำกัดทางทฤษฎีได้อีกขั้นหนึ่ง

ภาควิชา คณิตศาสตร์ ลายมือชื่อนิสิต.....

สาขาวิชา วิทยาการคอมพิวเตอร์ ลายมือชื่ออาจารย์ที่ปรึกษา.....

ปีการศึกษา 2548

4473804823 : MAJOR COMPUTER SCIENCE

KEY WORD: PROTEIN SECONDARY STRUCTURE PREDICTION / NEW ENCODING SCHEME / MARKOV MODEL / MARKOV TRANSITION MATRIX / SUPPORT VECTOR MACHINE

KASEMSANT KUPHANUMAT: ADVANCED MACHINE LEARNING METHOD FOR PREDICTION OF PROTEIN SECONDARY STRUCTURE, THESIS ADVISOR: PROF. CHIDCHANOK LURSINSAP, Ph.D., 67 pp. ISBN 974-53-2806-5.

A new method based on Markov process to encode the protein sequences has been introduced. With this simple method, input vectors that contain the essential features of protein sequence can be extracted and efficiently used to train SVM classifiers. Our method achieved the remarkable result that out-performs other advanced methods at present. Using a seven-folded cross validation on the data set of 513 non-homologous protein chains (CB513), the SVM together with Markov transition matrix encoding scheme produces a three-state overall per-residue accuracy(Q_3) of 82.49 percent and a segment overlap accuracy(SOV) of 77.18 percent. That is the next improving step to reach the theoretical limitation.

Department	Mathematics	Student's signature.....
Field of study	Computer Science	Advisor's signature.....
Academic year	2005	

Acknowledgements

First, I specially thank my advisor, Prof. Dr. Chidchanok Lursinsap, for his continuous support in the Ph.D. program, Dr. Prasit Palittapongarnpim, deputy director at my workplace, BIOTEC, for his encouragement, thesis committee, Dr. Jack Asavanant, Dr. Lerson Tanasugarn, Dr. Prasit Palittapongarnpim, and Dr. Siripun Sanguansintukul, who asked me good questions, gave insightful comments and reviewed my work on a very short notice.

I am grateful to the financial support from Thailand Research Fund (TRF) and National Center for Genetic Engineering and Biotechnology (BIOTEC), National Science and Technology Development Agency (NSTDA).

In addition, I would like to thank all of my teachers, my colleagues and my friends at computing research laboratory, AVIC, and Shrimp Genome laboratory at department of biochemistry. Without warmly help and support from them, I could not have finished this dissertation. The software, SVMlight, and data set kindly provided by Prof. Thorsten Joachims, James A. Cuff, and Prof. Geoffrey J. Barton are highly appreciated.

Last, but not least, I thank my family: my parents for giving me life in the first place, for educating me with unconditional support and encouragement.

Table of Contents

Chapter

List of Tables

List of Figures