

วิธีการขั้นสูงที่เครื่องเรียนรู้เพื่อการทำนายโครงสร้างทุติยภูมิของโปรตีน



นายเกษมสันต์ คุณานูมาต

ศูนย์วิทยทรัพยากร  
จุฬาลงกรณ์มหาวิทยาลัย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต

สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาคณิตศาสตร์


คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2548

ISBN 974-53-2806-5

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

ADVANCED MACHINE LEARNING METHOD FOR PREDICTION OF  
PROTEIN SECONDARY STRUCTURE



Mr Kasemsant Kuphanumat

ศูนย์วิทยทรัพยากร  
จุฬาลงกรณ์มหาวิทยาลัย

A Dissertation Submitted in Partial Fulfillment of the Requirements  
for the Degree of Doctor of Philosophy 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.)

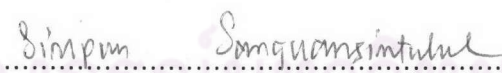
THESIS COMMITTEE

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

วิทยานิพนธ์ฉบับนี้ได้นำเสนอวิธีการใหม่ในการเข้ารหัสลำดับอะมิโนแอซิดของโปรตีนโดยใช้วิธีการของแมคคอล์ฟโปรเซส ด้วยเทคนิคนี้ ลักษณะเด่นที่สำคัญของลำดับสายโปรตีนจะถูกสกัดออกและนำไปใช้สร้างชุดของเวกเตอร์สำหรับเป็นข้อมูลในการสอนเพื่อการจำแนกประเภทด้วยซัพพอร์ตเวกเตอร์แมชชีน(SVM) ได้อย่างมีประสิทธิภาพ วิธีการที่ใช้ในงานวิจัยนี้ให้ผลลัพธ์ที่โดดเด่นกว่าวิธีการอื่นๆ ที่มีในปัจจุบันเป็นอย่างมาก ด้วยวิธีการจำแนกประเภทโดยใช้ SVM ร่วมกับวิธีการเข้ารหัสของข้อมูลโดยใช้มาคอฟทรานส์ชันเมตริกสามารถวัดค่าความถูกต้องในการจำแนกแบบสามกลุ่มได้ดังนี้คือ  $Q_3 = 82.49\%$ ,  $SOV = 77.18\%$  โดยการประเมินจากกลุ่มข้อมูลทดสอบมาตรฐานของโปรตีนจำนวน 513 สาย (CB513) ซึ่งผลที่ได้นับว่าเป็นการพัฒนาเข้าใกล้ขีดจำกัดทางทฤษฎีได้อีกขั้นหนึ่ง

ศูนย์วิทยทรัพยากร  
จุฬาลงกรณ์มหาวิทยาลัย

ภาควิชา คณิตศาสตร์

สาขาวิชา วิทยาการคอมพิวเตอร์

ปีการศึกษา 2548

ลายมือชื่อนิสิต..... 

ลายมือชื่ออาจารย์ที่ปรึกษา..... 

## 4473804823 : MAJOR COMPUTER SCIENCE

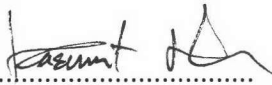
KEY WORD: PROTEIN SECONDARY STRUCTURE PREDICTION / NEW ENCODING SCHEME / MARKOV MODEL / MARKOV TRANSITION MATRIX / SUPPORT VECTOR MACHINE

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.

ศูนย์วิทยทรัพยากร  
จุฬาลงกรณ์มหาวิทยาลัย

Department Mathematics  
Field of study Computer Science  
Academic year 2005

Student's signature.....

Advisor's signature.....

## 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

Thai Abstract.....	iv
English Abstract.....	v
Acknowledgments.....	vi
List of Tables.....	ix
List of Figures.....	x
Chapter	
I Introduction.....	1
1.1 Secondary Structure Prediction.....	3
1.2 Problem Formulation and Proposed Solutions.....	4
1.3 The Contributions of Dissertation.....	5
1.4 Scope and Organization.....	7
II Prediction of Protein Secondary Structure .....	8
2.1 First Generation Methods.....	8
2.2 Second Generation.....	9
2.3 Third Generation.....	10
2.4 The Goals of Secondary Structure Prediction .....	11
2.5 Training and testing data sets .....	12
2.6 Protein Secondary Structure Definition .....	14
2.7 Performance measures .....	17
2.8 Cross Validation Method .....	20
III SVM and Markov Model Approach for Protein Secondary Structure	
Prediction .....	21

Chapter	
3.1 Support Vector Machine .....	21
3.2 Markov Model .....	25
3.3 Multi-Step Markov model .....	27
3.4 Vectors Representation of Protein Sequences .....	28
3.5 Normalization and Interpolation of Transition Matrix .....	38
3.6 Incorporation of the Evolutionary Information .....	39
3.7 Network Structure of Learning Model .....	40
3.8 Reliability Index and Filtering of the Predictions .....	44
3.9 Parameters optimization on SVM .....	46
IV Results and Discussion .....	47
4.1 Experimental Results .....	47
4.2 Conclusion .....	59
4.3 Discussion .....	59
4.4 Future Work .....	60
REFERENCES.....	61
APPENDIX.....	64
A Publications.....	65
Biography.....	67



## List of Tables

1.1	Three formulated problems and the proposed solutions. ....	6
2.1	Three and Eight Classes of Secondary Structure. ....	16
2.2	Three different predictions that all have the same per residue accuracy..	19
4.1	Performance of binary classifiers (%) respected to step and order of Markov model.....	48
4.2	Performance of binary classifiers (%) on 1st-layer with respect to window size.....	49
4.3	Performance of binary classifiers (%) on 2nd-layer with respect to window size.....	49
4.4	Ratio of number of SVs to all training samples.....	50
4.5	Results from tertiary classification, comparing of single and double layer network model.....	51
4.6	Results from ternary classification, comparing of optimum and combine step of Markov Model.....	52
4.7	Results from ternary classification, comparing of filtered and non-filtered result.....	52
4.8	Comparison with the results of other approach for three class classification.....	53

## List of Figures

3.1	Data flow of input patterns preparation process.....	32
3.2	Markov transition Matrix.....	33
3.3	An amino acid pattern constructed by sliding window. ....	34
3.4	The input vector created from 4th order Markov Chain.....	34
3.5	The constructing process of Input Vectors.....	35
3.6	The input vector that created from high-order Markov chain.....	36
3.7	The input vector constructed from multi-step Markov Chain.....	37
3.8	The single layer network model .....	42
3.9	double layer network model.....	43
4.1	Comparison of three-state overall per-residue accuracy (Q3) .....	55
4.2	Comparison of segment overlap accuracy (SOV) .....	56
4.3	Distribution of proteins predicted with different accuracies.....	57
4.4	Distribution of accuracy (Q3) and number of residues with different reliability index (Ri) .....	58

ศูนย์วิทยทรัพยากร  
จุฬาลงกรณ์มหาวิทยาลัย