

การรู้จำโปรโมเตอร์โดยใช้เทคนิคการเลือกด้วยวิธีทางสถิติจากการแทนด้วยเคออสเกมส์



นางสาวอรรรณ ตินนังวัฒนะ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต

สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาคณิตศาสตร์

คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2548

ISBN 974-53-2756-5

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

STATISTICAL FEATURE SELECTION FROM CHAOS GAME REPRESENTATION
FOR PROMOTER RECOGNITION

Miss Orawan Tinnungwattana

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 974-53-2756-5

481772

อรวรรณ ดินนังวัฒนะ : การรู้จำโปรโมเตอร์โดยใช้เทคนิคการเลือกด้วยวิธีทางสถิติจากการแทนด้วยเคออสเกมส์ (STATISTICAL FEATURE SELECTION FROM CHAOS GAME REPRESENTATION FOR PROMOTER RECOGNITION) อ. ที่ปรึกษา : ศ. ดร. ชิดชนก เหลือสินทรัพย์, 58 หน้า. ISBN 974-53-2756-5.

ปัญหาการรู้จำโปรโมเตอร์เป็นที่สนใจของนักวิจัยในปัจจุบัน แต่อัลกอริทึมที่มีอยู่ก็ยังไม่ให้การทดลองที่ไม่ดีพอ ดังนั้นเป้าหมายของวิทยานิพนธ์นี้คือการพัฒนาอัลกอริทึมที่สามารถแยกความแตกต่างระหว่างลำดับดีเอ็นเอที่เป็นโปรโมเตอร์และไม่ใช่โปรโมเตอร์ออกให้ได้ โดยไม่ได้เอารูปแบบที่มีอยู่ก่อนหน้า เช่น TATAAT-box และ TTGACA-box มาพิจารณา ความถูกต้องของการทำนายโปรโมเตอร์ขึ้นอยู่กับปัจจัย 2 ประการ คือ การแทนลำดับดีเอ็นเอและการเลือกคุณลักษณะที่สำคัญ หลักการคือการใช้เทคนิค Chaos Game Representation มาช่วยในการแปลงลำดับดีเอ็นเอซึ่งประกอบด้วยโปรโมเตอร์และไม่ใช่โปรโมเตอร์ให้เป็นภาพเพื่อที่จะเห็นรูปแบบได้ชัดเจนมากขึ้น จากนั้น จะทำการเลือกคุณสมบัติที่สำคัญที่ทำให้มีความแตกต่างกันมากที่สุดออกมาโดยใช้วิธีการเลือกทางสถิติ จุดประสงค์เพื่อลดขนาดของข้อมูลให้เล็กลงเพื่อให้โครงข่ายประสาทเทียมทำการรู้จำ วิธีการในงานวิจัยนี้สามารถใช้ได้ทั้งสิ่งมีชีวิตเซลล์เดียว และหลายเซลล์

ภาควิชา.....คณิตศาสตร์.....ลายมือชื่อนิสิต..... *Omra Siriratan*
 สาขาวิชา.....วิทยาการคอมพิวเตอร์.....ลายมือชื่ออาจารย์ที่ปรึกษา..... *C. Lu*
 ปีการศึกษา.....2548.....

4473847823 : MAJOR COMPUTER SCIENCE

KEY WORD: PROMOTER RECOGNITION/ CHAOS GAME REPRESENTATION / STATISTICAL
FEATURE SELECTION / ARTIFICIAL NEURAL NETWORK

ORAWAN TINNUNGWATTANA: STATISTICAL FEATURE SELECTION FROM
CHAOS GAME REPRESENTATION FOR PROMOTER RECOGNITION. THESIS
ADVISOR: PROF CHIDCHANOK LURSINSAP, Ph.D., 58 pp. ISBN 974-53-2756-5.

Recently, the recognition of promoters has attracted many researchers' attention. Unfortunately, most previous prediction algorithms did not provide high enough sensitivity and specificity. The aim of this dissertation is to provide a distinct classification between promoter and non-promoter sequences. We do not consider some well-known patterns around TSS, such as TATAAT-box and TTGACA-box, which were previously used by many researchers. The accuracy of promoter prediction is based on two factors, i.e., the representation of the given DNA sequence and the essential features of the sequence. A Chaos Game Representation (CGR) is adopted for transforming a DNA sequence having promoters and non-promoters into an image. The essential features of the CGR are selected by applying the concept of statistical feature selection. It is aimed at finding the smallest set of features that can distinguish the classes over the full set and reduce the dimension of the classifier. Recognition can then be performed by a supervised neural network. The method in this dissertation can be applied to both prokaryotic and eukaryotic organisms.

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

Acknowledgements

I am greatly indebted to my supervisor, Professor Dr. Chidchanok Lursinsap, for his suggestions, guidance and encouragement, help me overcome the necessary difficulties of the process of research and make this thesis possible.

I also wish to express my special thanks to the dissertation committee with their advice and guidance, help focus my research activities. And, I would like to thank Associate Professor Suchada Siripant, who looked after me when I stayed in AVIC research center to do my research.

This work is partially supported by Burapha University Chanthaburi IT Campus. I would also like to thank The Advanced Virtual and Intelligence Computing Research Center (AVIC) for their material support in enabling me to accomplish this research.

I am grateful to all my colleague and friends at the Burapha University Chanthaburi IT Campus and AVIC Research Center, especially Daranee Chotigunta, Sirisuda Bua-tongkue, Thararat Puangsuwan, Benchaporn Jantarakongkul, Krisana Chinasarn, Supaporn Bunrit, Kodchakorn Na Nakornphanom, for their care, having encouraged and supported me during my study.

Finally, I would like to express my sincere gratitude to my parents and family members for their love, hearty encouragement, unselfish, and, especially to my husband, Manit Chittakarn, for his love, wormest care, and being patient during my confusing and frustrating stage.

Table of Contents

Thai Abstract	iv
English Abstract	v
Acknowledgements	vi
Table of Contents	vii
List of Tables	x
List of Figures	xi
CHAPTER	
I INTRODUCTION	1
1.1 Problems Identification	2
1.2 Objective and Scope of the Research	3
1.3 Organization of the Dissertation	3
II BACKGROUND KNOWLEDGE	5
2.1 Molecular biology and Genetics	5
2.1.1 DNA and RNA	5
2.1.2 Gene and Chromosome	6
2.1.3 Synthesis of macromolecules and the central dogma	7
2.1.4 The Genetic Code	9
2.2 Basic introduction to promoter	10
2.3 The features of Promoter Sequences	13
2.3.1 TATA-Box and TTG-Box	13
2.3.2 CpG Islands	13
2.4 Artificial Neural Networks (ANNs)	14
2.4.1 What is a Neural Network?	14
2.4.2 How does Human Brain Learn?	14

CHAPTER	Page
2.4.3 From Human Neurones to Artificial Neurones	15
2.4.4 Multi-layer feedforward and Backpropagation Neural Networks . .	16
III LITERATURE REVIEWS	20
3.1 Promoter Recognition in Prokaryotes	20
3.2 Promoter Recognition in Eukaryotes	21
IV PROPOSED METHODS FOR PROMOTER RECOGNITION	25
4.1 Problem Formulation	25
4.2 Feature extraction and selection from DNA sequences	25
4.2.1 Chaos Game Representation (CGR)	26
4.3 Statistical Feature Selection of DNA Sequences	29
4.3.1 Statistical Feature Method	30
4.3.2 CpG island features	30
4.4 Architecture of the prediction system	31
V EXPERIMENTS AND RESULTS	33
5.1 Sequence Data Sets	33
5.1.1 Prokaryote Data Set	33
5.1.2 Eukaryote Training Sets	33
5.1.3 Evaluation sets in large genome sequences	35
5.2 Performance Evaluation of Promoter Recognition	35
5.3 The Results on <i>E.coli</i> Data Sets	36
5.4 The results on Eukaryote Training Set	36
5.4.1 The results on Training Set 1	36
5.4.2 The results on Training Set 2	38
5.4.3 The results on Training Set 3	39
5.5 The results on Evaluation sets in large genome sequences	39

CHAPTER	Page
VI CONCLUSION	42
References	44
Biography	47

List of Tables

TABLE		Page
5.1	Amount of available Eukaryote Training sets (No.sequences)	34
5.2	Amount of available test set (bp) for evaluating performance of pre- dictions.	34
5.3	Number of errors on several methods	37
5.4	Performances evaluation using precision, specificity and sensitivity . . .	37
5.5	The results on Training Set 1.	38
5.6	Comparing result on Data Set 1.	38
5.7	Comparing results on Test set A.	39
5.8	The results on Test set B.	40
5.9	Comparing results on Test set B.	40
5.10	Comparing result on Test set C.	41
5.11	Comparing result on Test set D.	41

List of Figures

FIGURE	Page
2.1 DNA structure and composition, formed by two complementary antiparallel chains of nucleotides.	6
2.2 Macromolecule synthesis and other eukaryote cell mechanisms. Protein synthesis and DNA replication.	8
2.3 The genetic code, written by convention in the form in which the codons appear in mRNAs.	10
2.4 A possible promoter gene relation structure of mRNA eukaryotic gene.	11
2.5 A possible structure of a Pol II promoter.	11
2.6 Component of Human Neurons	15
2.7 Components of Artificial Neural Network	15
2.8 Structure of multi-layer networks	17
4.1 Chaos Game Representation (CGR) suffix property. Sequences ending in a specific sub-string are in the square labeled with that suffix.	27
4.2 (a) Chaos Game Representation (CGR) of the first 10 nucleotides of <i>E.coli</i> gene <i>thrA</i> : ATGCCGAGTGT. (b) CGR of the full <i>thrA</i> sequence, totaling 2463 pairs of bases.	28
4.3 The CGR coordinates for the 2463 base pairs are plotted with the relative frequencies for each 8×8 quadrant represented as a grayscale(left). The distribution of counts in listed in the table(right).	29
4.4 The structure of neural network used in this problem.	32
5.1 Performance Evaluation of Promoter Recognition.	36