การรู้จำโปรโมเตอร์โดยใช้เทคนิคการเลือกด้วยวิธีทางสถิติจากการแทนด้วยเคออสเกมส์

นางสาวอรวรรณ ตินนังวัฒนะ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาคณิตศาสตร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2548

ISBN 974-53-2756-5

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

Thesis Title	STATISTICAL FEATURE SELECTION FROM CHAOS GAME REPRESENTATION FOR PROMOTER RECOGNITION
By	Miss Orawan Tinnungwattana
Field of Study	Computer Science
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 Peraphon Sophatsathit, Ph.D.)

..... Thesis Advisor

(Professor Chidchanok Lursinsap, Ph.D.)

..... Member

(Associate Professor Boonserm Kijsirikul, Ph.D.)

..... Member

(Rath Pichyangkura, Ph.D.)

..... Member

(Siripun Sanguansintukul, Ph.D.)

..... Member

(Kanuengnij Kubola, Ph.D.)

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 แต่อัลกอริทึมที่มีอยู่ก็ยังให้ผลการทดลองที่ไม่ดีพอ 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.....

สาขาวิชา..... **วิทยาการคอมพิวเตอร์**..... ลายมือชื่ออาจารย์ที่ปรึกษา.....

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.....

Field of study....**Computer Science**.....Advisor's signature.....

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 neccessary 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 Buatongkue, 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

CHAPTER Page

List of Tables

List of Figures