PREDICTION OF NON-CODING RNAS AND THEIR TARGETS IN *SPIRULINA PLATENSIS* GENOME

MR. TANAWUT SRISUK

THE THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE
(BIOINFORMATICS AND SYSTEMS BIOLOGY)
SCHOOL OF BIORESOURCES AND TECHNOLOGY AND
SCHOOL OF INFORMATION TECHNOLOGY
KING MONGKUT'S UNIVERSITY OF TECHNOLOGY THONBURI
2010

Prediction of non-coding RNAs and their targets in *Spirulina platensis* genome

Mr. Tanawut Srisuk B.Sc. (Biological Science)

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science (Bioinformatics and Systems Biology)
School of Bioresources and Technology
and
School of Information Technology
King Mongkut's University of Technology Thonburi
2010

Thesis Committee

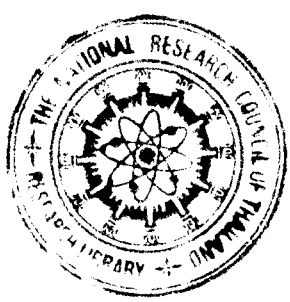
Kalyanee Paithoonrangsarid ..... Chairman of Thesis Committee
(Researcher, Kalyanee Paithoonrangsarid, Ph.D.)

Chinae Thammarongtham ..... Member and Thesis Advisor
(Researcher, Chinae Thammarongtham, D.Sc.)

..... Member and Thesis Co-advisor
(Assoc. Prof. Supapon Cheevadhanarak, Ph.D.)

Jittisak Senachak ..... Member
(Researcher, Jittisak Saenachak, Ph.D.)

..... Member
(Researcher, Duangdao Wichadakul, Ph.D.)



PREFACE

This thesis in the title of "Prediction of non-coding RNAs and their targets in *Spirulina platensis* genome" is my master study at King Mongkut's University of Technology Thonburi. The thesis is composed of four chapters which are Introduction, Literature Reviews, Materials and Methodology, and Results and Discussions parts. In addition, References and Appendices are also included.

Thesis Title	Prediction of non-coding RNAs and their targets in *Spirulina platensis* genome
Thesis Credit	12
Candidate	Mr. Tanawut Srisuk
Thesis Advisors	Dr. Chinae Thammarongtham Assoc. Prof. Dr. Supapon Cheevadhanarak
Program	Master of Science
Field of Study	Bioinformatics and Systems Biology
Faculty	School of Bioresources and Technology and School of Information Technology
B.E.	2553

Abstract

Non-coding RNAs (ncRNAs), transcripts that have function without being translated to protein, have a number of roles in the cell including important regulatory roles. Efforts to identify the whole set of ncRNAs and then to elucidate their functions would gain better biological understanding. Although ncRNA is another type of genome constituent, most of the genes for ncRNA are overlooked by standard genome annotation of genome sequencing projects. This also happens in *Spirulina platensis* genome sequencing project. It is because gene finding tools generally are able to identify only protein-coding genes but not non-protein-coding ones. In this study, *S. platensis* ncRNAs were detected by comparative genomics approach using computational tools, together with RNA secondary structure prediction. The results show that there are 334 ncRNA candidates. A set of 247 loci were classified into 13 known families. A majority of known ncRNAs which include 199 loci were classified into Group II intron components. The prediction pipe line was also applied to *Arthrospira maxima* and *Lyngbya* sp. PCC 8106 and ncRNA candidates for the species were identified. The predicted targets for some putative ncRNAs in *S. platensis* are also proposed.

Keywords: Non-Coding RNA Prediction/ Non-Coding RNA Target Prediction

หัวข้อวิทยานิพนท์	การทำนายหาอาร์เอนเอไม่แปลรหัสและเป้าหมายของอาร์เอนเอเหล่านั้นในจีโนมของ *Spirulina platensis*
หน่วยกิต	12
ผู้เขียน	นายธนาวุธ ศรีสุข
อาจารย์ที่ปรึกษา	ดร.ชิเน ธำมรงค์ธรรม รศ.ดร.สุภาภรณ์ ชีวะธนรักษ์
หลักสูตร	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	ชีวสารสนเทศและชีววิทยาระบบ
คณะ	ทรัพยากรชีวภาพและเทคโนโลยี และคณะเทคโนโลยีสารสนเทศ
พ.ศ.	2553

บทคัดย่อ

อาร์เอนเอไม่แปลรหัส(ncRNA)คือ RNA ที่ทำงานโดยไม่ต้องถูกแปลรหัสเป็นโปรตีน โดย ncRNA เหล่านี้ทำหน้าที่สำคัญภายในเซลล์ รวมไปถึงการควบคุมระบบต่างๆภายในเซลล์อีกด้วย การระบุหน้าที่ของ ncRNA เหล่านี้ย่อมนำไปสู่ความเข้าใจเกี่ยวกับชีววิทยาของสิ่งมีชีวิตมากขึ้น ถึงแม้ว่า ncRNA จะเป็นองค์ประกอบของจีโนมแต่ก็ถูกมองข้ามไปในขั้นตอนการวิเคราะห์จีโนมตามระเบียบวิธีของโครงการถอดรหัสจีโนมต่างๆและเกิดขึ้นกับโครงการถอดรหัสจีโนมของ *Spirulina platensis* เช่นกัน เนื่องจากเครื่องมือที่ใช้ระบุตำแหน่งยีนนั้นไม่สามารถระบุตำแหน่งของยีนที่สร้าง ncRNA ได้ โดยงานวิจัยนี้ได้ทำนายหาตำแหน่งของ ncRNA ด้วยวิธีเปรียบเทียบลำดับเบสในจีโนมร่วมกับการเปรียบเทียบโครงสร้างทุติยภูมิของ RNA ทำให้ได้ตำแหน่งที่คาดว่าจะเป็น ncRNA ทั้งหมด 334 ตำแหน่ง เป็นส่วนที่ทราบชนิดแล้ว 247 ตำแหน่ง แบ่งเป็น 13 ชนิด ในจำนวนนี้เป็นส่วนประกอบของ Group II intron ถึง 199 ตำแหน่ง วิธีการนี้ยังถูกใช้ทำนายหาตำแหน่งของ ncRNA ใน *Arthrospira maxima* และ *Lyngbya* sp. PCC 8106 และได้บริเวณที่คาดว่าจะเป็น ncRNA ในแบคทีเรียทั้งสองชนิดมาจำนวนหนึ่ง นอกจากนั้นแล้วเป้าหมายของ ncRNA แต่ละตัวที่ทำนายได้ก็ได้แสดงไว้ในงานวิจัยนี้ด้วย

คำสำคัญ: การทำนายหาอาร์เอนเอไม่แปลรหัส/ การทำนายหาเป้าหมายของอาร์เอนเอไม่แปลรหัส

ACKNOWLEDGEMENTS

First of all I would like to thank the National Center for Genetic Engineering and Biotechnology, Thailand (BIOTEC), and King Mongkut's University of Technology Thonburi for the scholarship in the Bioinformatics program. This thanks is extended to all Thai citizens who pay the tax which funds this scholarship.

Thank you for all the help and teaching from all teachers and staff in the Bioinformatics program at King Mongkut's University of Technology Thonburi, especially for Assoc. Prof. Dr. Supapon Cheevadhanarak my teacher and co-advisor. Her lessons filled my brain with many new ideas. I also thank all members in the *Spirulina* genome sequencing project, who took the time to provide the genome sequence data which was then used in this work.

Last, but not least, I would like to give a special thanks to Dr. Chinae Thammarongtham and Mr. Natapol Pornputtapong for their kind assistance. They taught me many things, from English grammar to biochemistry. Without their help, the work would not have been possible.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF TECHNICAL VOCABULARY AND ABBREVIATIONS

CM	= Covariance Model
DNA	= deoxyribonucleic acid
IG	= Intergenic Region
ORF	= open reading frame
ncRNA	= non-coding RNA
nt	= nucleotide
NA	= not available