



E46932



**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 MONCKUT'S UNIVERSITY OF TECHNOLOGY THONBURI**

2016

b00246555



E46932

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

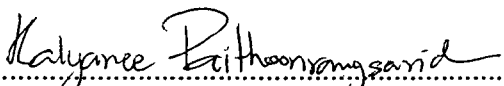
Mr. Tanawut Srisuk B.Sc. (Biological Science)

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

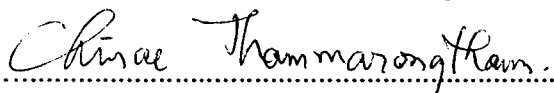
and

School of Information Technology
King Mongkut's University of Technology Thonburi
2010

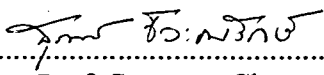
Thesis Committee


.....
(Researcher, Kalyanee Paithoonrangsarid, Ph.D.)

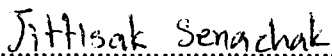
Chairman of Thesis Committee


.....
(Researcher, Chinae Thammarongtham, D.Sc.)

Member and Thesis Advisor


.....
(Assoc. Prof. Supapon Cheevadhanarak, Ph.D.)

Member and Thesis Co-advisor

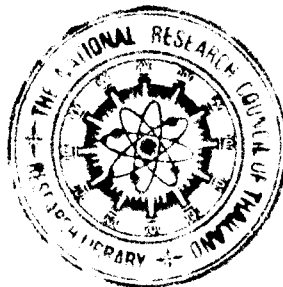

.....
(Researcher, Jittisak Saenachak, Ph.D.)

Member


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

Member

Copyright reserved



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 <i>Spirulina platensis</i> 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

E46932

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

หัวข้อวิทยานิพนธ์	การทำนายหาอาร์เอ็นเอไม่แปลรหัสและเป้าหมายของอาร์เอ็นเอเหล่านั้นใน จีโนมของ <i>Spirulina platensis</i>
หน่วยกิต	12
ผู้เขียน	นายธนาวุธ ศรีสุข
อาจารย์ที่ปรึกษา	ดร.จิเน ชำรงศรีธรรม รศ.ดร.ศุภาภรณ์ ชีวะธนรักษ์
หลักสูตร	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	ชีวสารสนเทศและชีววิทยาระบบ
คณะ	ทรัพยากรชีวภาพและเทคโนโลยี และคณะเทคโนโลยีสารสนเทศ
พ.ศ.	2553

บทคัดย่อ

E 46932

อาร์เอ็นเอไม่แปลรหัส(ncRNA)คือ RNA ที่ทำงานโดยไม่ต้องถูกแปลรหัสเป็นโปรตีน โดย ncRNA เหล่านี้ทำหน้าที่สำคัญภายในเซลล์ รวมไปถึงการควบคุมระบบต่างๆภายในเซลล์อีกด้วย การระบุหน้าที่ของ ncRNA เหล่านี้ย่อมนำไปสู่ความเข้าใจเกี่ยวกับชีววิทยาของสิ่งมีชีวิตมากขึ้น ถึงแม้ว่า ncRNA จะเป็นองค์ประกอบของจีโนมแต่ก็ถูกมองข้ามไปในขั้นตอนการวิเคราะห์จีโนมตามระเบียบวิธีของโครงการถอดรหัสจีโนมต่างๆและเกิดขึ้นกับโครงการถอดรหัสจีโนมของ *Spirulina platensis* เช่นกัน เนื่องจากเครื่องมือที่ใช้ระบุตำแหน่งยีนนั้นไม่สามารถระบุตำแหน่งของยีนที่สร้าง ncRNA ได้ โดยงานวิจัยนี้ได้ทำนายหาตำแหน่งของ ncRNA ด้วยวิธีเปรียบเทียบลำดับเบสในจีโนมร่วมกับการเปรียบเทียบโครงสร้างทุติยภูมิของ RNA ทำให้ได้ตำแหน่งที่คาดว่าจะ เป็น 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 และได้บริเวณที่คาดว่าจะ เป็น ncRNA ในแบคทีเรียทั้งสองชนิดมาจำนวนหนึ่ง นอกจากนั้นแล้วเป้าหมายของ ncRNA แต่ละตัวที่ทำนายได้ก็ได้แสดงไว้ในงานวิจัยนี้ด้วย

คำสำคัญ: การทำนายหาอาร์เอ็นเอไม่แปลรหัส/ การทำนายหาเป้าหมายของอาร์เอ็นเอไม่แปลรหัส

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 Thammamongtham 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

PREFACE	i
ENGLISH ABSTRACT	ii
THAI ABSTRACT	iii
ACKNOWLEDGEMENTS	iv
CONTENTS	v
LIST OF TABLES	vii
LIST OF FIGURES	viii
LIST OF TECHNICAL VOCABULARY AND ABBREVIATIONS	x
 CHAPTER 1 INTRODUCTION	 1
1.1 Background and Rationale	1
1.2 Objectives	2
1.3 Scope of works	2
1.4 Expected output	3
1.5 Conceptual work flow	3
 CHAPTER 2 LITERATURE REVIEWS	 4
2.1 <i>Spirulina platensis</i>	4
2.2 Non-coding RNA	4
2.3 Non-coding RNA identification	5
2.4 Computer based non-coding RNA gene identification	6
2.4.1 Homology based strategies	6
2.4.2 Ab-initio strategies	7
2.5 Non-coding RNA targets identification	8
2.5.1 Experimental identification	9
2.5.2 Computational prediction	9
2.6 Tools	10
2.6.1 BLAST	11
2.6.2 MUSCLE	11
2.6.3 RNAz	12
2.6.4 Infernal	13
2.6.5 IntaRNA	13
 CHAPTER 3 MATERIALS AND METHODOLOGY	 15
3.1 Software and hardware	15
3.1.1 BLAST+	15
3.1.2 MUSCLE 3.7	15
3.1.3 RNAz package	15
3.1.4 Infernal 1.0	15
3.1.5 IntaRNA	15
3.1.6 Vienna RNA package	15
3.1.7 CGView	15
3.1.8 Python scripts	16
3.1.9 Hardwares	16
3.2 Data sets	16
3.2.1 Bacterial genome sequences	16

3.2.2 Known ncRNA sequences and covariance models	18
3.3 Methodology	18
3.3.1 RNAz input preparation	18
3.3.2 RNAz scanning	19
3.3.3 RNAz result clustering	19
3.3.4 Covariance models construction	21
3.3.5 Classification by covariance model	21
3.3.6 Non-coding RNA target prediction	21
CHAPTER 4 RESULTS AND DISCUSSIONS	23
4.1 Non-coding RNA prediction	23
4.1.1 Pipeline Evaluation	23
4.1.2 Non-coding RNA in <i>S. platensis</i>	23
4.2 Non-coding RNA Classification	34
4.2.1 Yfr1	38
4.2.2 Yfr2b	39
4.2.3 Group II intron	40
4.2.4 Cobalamin riboswitch	45
4.2.5 Signal recognition particle RNA (SRP)	47
4.2.6 mir-598 homolog	48
4.2.7 Bacterial RNase P type A	49
4.2.8 CRISPR	50
4.2.9 Putative novel ncRNAs	52
4.3 Non-coding RNA target prediction in <i>S. platensis</i>	54
4.3.1 Yfr1 predicted targets	54
4.3.2 Yfr2b predicted targets	55
4.3.3 mir-598 homolog predicted targets	57
4.3.4 RNase P RNA predicted targets	59
4.3.5 Putative novel ncRNA predicted targets	59
4.4 Discussion	61
4.4.1 Non-coding RNA prediction	61
4.4.2 Non-coding RNA classification	61
4.4.3 Target prediction	61
REFERENCES	63
APPENDIX A. Non-coding RNA prediction in <i>Arthrospira maxima</i>	74
APPENDIX B. Non-coding RNA prediction in <i>Lyngbya</i> sp. PCC 8106	84
APPENDIX C. Predicted ncRNA classification in <i>Arthrospira maxima</i>	89
APPENDIX D. Predicted ncRNA classification in <i>Lyngbya</i> sp. PCC 8106	93
APPENDIX E. Mapping of RNase P targets ORF in <i>Spirulina platensis</i> genome	98
APPENDIX F. Comparing between position of ncRNA candidates and known ncRNA in <i>E. coli</i> K-12 substr. MG1655	106
CURRICULUM VITAE	112

LIST OF TABLES

TABLES	PAGES
3.1 Bacteria species and sources.	16
4.1 List of ncRNA candidates in <i>S. platensis</i> genome with adjacent ORF IDs and distance from the ORFs.	26
4.2 List of ncRNA which mathed to CM with cmsearch E-Value lower than 1e-5 and matched position on ncRNA.	35
4.3 Possible group II intron components in <i>S. platensis</i> .	41
4.4 Cobalamin riboswitch and downstream ORF(s)	46
4.5 Member in the largest group of possible novel ncRNA in <i>S. platensis</i> .	53
4.6 Top rank of predicted targets for Yfr1 homolog.	54
4.7 Top rank of predicted targets for Yfr2b in <i>S. platensis</i> .	56
4.8 Top rank of predicted targets for mir-598 homolog.	57
4.9 Top rank of predicted targets for RNase P type A in <i>S. platensis</i> .	59
4.10 Predicted target for sp42.	60
A.1 List of ncRNA candidates in <i>A. maxima</i> and adjacent ORFs with distance from the ORFs.	75
B.1 List of ncRNA candidates in <i>Lyngbya</i> sp. PCC 8106 and adjacent ORFs with distance from the ORFs.	85
C.1 Known ncRNA in <i>Arthrospira maxima</i> .	90
D.1 Known ncRNA in <i>Lyngbya</i> sp. PCC 8106.	94
F.1 Comparing between position of ncRNA candidates and known ncRNA in <i>E. coli</i> K-12 substr. MG1655	107

LIST OF FIGURES

FIGURES	PAGES
1.1 Conceptual work flow.	3
3.1 Non-coding RNA prediction processes.	20
3.2 Non-coding RNA classification processes.	22
3.3 Target prediction processes for non-coding candidates in <i>S. platensis</i> .	22
4.1 Whole genome mapping of predicted ORF and predicted ncRNA loci derived from using CGView tool.	25
4.2 Result from cmsearch represents matching between Yfr1 CM and locus sp289. Primary sequence and secondary structure aligning were described in Eddy (2009).	39
4.3 Result from cmsearch represents matching between Yfr2b CM and sp185 locus.	39
4.4 Predicted structure of Yfr2b in <i>S. platensis</i> . Thick line represents the regions which matched to Yfr2b CMs.	40
4.5 Matching between group II intron DV-DVI CM and sp26 locus from cmsearch result.	44
4.6 Full length ORF-less group II intron consensus structure in <i>S. platensis</i> . The structure was predicted with RNAalifold.	45
4.7 Result from cmsearch represents matching between CM of Cobalamin riboswitch and <i>S. platensis</i> ncRNA candidates.	47
4.8 Result from cmsearch represents matching between bacterial SRP CM and sp179 locus.	48
4.9 ORFs around mir-598 homolog in <i>S. platensis</i> .	48
4.10 Result from cmsearch represents matching between mir-598 CM and sp22 locus.	49
4.11 Result from cmsearch represents Bacterial RNase P type A CM matching region on <i>S. platensis</i> genome.	50
4.12 Result from cmsearch matching between CRISPR-DR57 CM and locus sp116.	51
4.13 Predicted structure for locus sp116.	52
4.14 Consensus structure of the largest group of putative novel ncRNAs in <i>S. platensis</i> .	53
4.15 Yfr1 predicted structure using RNAfold. Thick line represents interaction region and dot line represents A-U rich regions which are Hfq binding motif.	55
4.16 Predicted structure of Yfr2b in <i>S. platensis</i> . Thick line represents the regions which matched to Yfr2b CMs, dot-line represents predicted interaction regions.	57
4.17 Predicted structure of mir-598 homolog sp22 locus. Thick-line and dot-line represent predicted interaction sites for the targets in table 4.8.	58
E.1 Mapping of WD-40 repeat-containing protein ORF and adjacent ORFs on <i>S. platensis</i> genome.	99
E.2 Mapping of putative ATP-dependent helicase ORF and adjacent ORFs on <i>S. platensis</i> genome.	99
E.3 Mapping of twitching motility protein ORF and adjacent ORFs on <i>S. platensis</i> genome.	100
E.4 Mapping of cysteine desulfurase, SufS subfamily ORF and adjacent ORFs on <i>S. platensis</i> genome.	100

E.5	Mapping of sulfotransferase ORF and adjacent ORFs on <i>S. platensis</i> genome.	101
E.6	Mapping of aspartate carbamoyltransferase ORF and adjacent ORFs on <i>S. platensis</i> genome.	101
E.7	Mapping of rubredoxin-type Fe(Cys) ₄ protein ORF and adjacent ORFs on <i>S. platensis</i> genome.	102
E.8	Mapping of ribosomal protein L6 ORF and adjacent ORFs on <i>S. platensis</i> genome.	102
E.9	Mapping of von Willebrand factor type A ORF and adjacent ORFs on <i>S. platensis</i> genome.	103
E.10	Mapping of chitin synthase ORF and adjacent ORFs on <i>S. platensis</i> genome.	103
E.11	Mapping of photosystem II reaction center protein Z ORF and adjacent ORFs on <i>S. platensis</i> genome.	104
E.12	Mapping of MazG family protein ORF and adjacent ORFs on <i>S. platensis</i> genome.	104
E.13	Mapping of Pentapeptide repeat protein ORF and adjacent ORFs on <i>S. platensis</i> genome.	105

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