ห้องสมุดงานวิจัย สำนักงานคณะกรรมการวิจัยแห่งชาติ
E46298

EVOLUTIONARY SCENARIOS OF CYANOBCTERIAL LINEAGE AS DETERMINED BY COMPARTAIVE GENOMIC APPROACH (I)/ SINGLE NUCLEOTIDE POLYMORPHISMS (SNPs) AND GENOME WIDE ASSOCIATION STUDY (GWAS) CONTENT PACK (II)

MR. PALANG CHOTSIRI

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE (BIOINFORMATICS)
SCHOOL OF BIORESOURCES AND TECHNOLOGY) AND
SCHOOL OF INFORMATON TECHNOLOGY
KING MONGKUT'S UNIVERSITY OF TECHNOLOGY THONBURI
2011

b00256188

ห้องสมุดงานวิจัย สำนักงานคณะกรรมการวิจัยแห่งชาติ
E46298

Evolutionary scenarios of cyanobacterial lineages as determined by comparative genomic approach (I)

Mr. Palang Chotsiri B.Sc. (Physics)

A Thesis Submitted in Partial Fulfillment of the Requirements for
The Degree of Master of Science (Bioinformatics)
School of Bioresources and Technology and School of Information Technology
King Mongkut's University of Technology Thonburi
2011

Thesis Committee

K. Paithoonrangsarid ..... Chairman of Thesis Comittee (I)

(Researcher, Kalyanee Paithoonrangsarid, Ph.D.)

..... Member and Thesis Advisor (I)

(Researcher, Vethachai Plengvidhya, Ph.D.)

..... Member and Thesis Co-Advisor (I)

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

..... Member and Thesis Co-Advisor (I)

(Researcher, Jittisak Senachak, Ph.D.)

..... Member (I)

(Researcher, Kobkul Laoteng, D.Sc.)

..... Member (I)

(Researcher, Peerada Prommeenate, Ph.D.)

Copyright reserved

THE NATIONAL RESEARCH COUNCIL OF THAILAND
RESEARCH LIBRARY

PREFACE

This thesis was written in order to accomplish my master degree graduation in field of Bioinformatics, King Mongkut's University of Technology Thonburi (KMUTT). It consists of two separated parts. The first part is "Evolutionary scenarios of cyanobacterial lineage as determined by comparative genomic approach" or Thai title name is "การศึกษาเหตุการณ์ในสายวิวัฒนาการของสาหร่ายสีเขียวแกมน้ำเงินด้วยวิธีการศึกษาจีโนมแบบเปรียบเทียบ". This part was done at King Mongkut's University of Technology Thonburi (KMUTT), Bangkok, Thailand. The second part is "Single Nucleotide Polymorphisms (SNPs) and Genome Wide Association Study (GWAS) content pack" or Thai title name is "กล่องข้อมูล ภาวะหลากหลายรูปแบบบนนิวคลีโอไทด์เดี่ยว และ การศึกษาความเชื่อมโยงของทั่วทั้งจีโนม". This part was completed with Torrey Path Inc., Thailand.

Each part of thesis consist of five major chapters, which are an introduction, literature review, materials and methods, results and discussions, and conclusions and recommendations. Moreover, the programming code and the supported information are included in the appendixes parts.

Thesis Title (I)	Evolutionary Scenarios of Cyanobacterial Lineages as Determined by Comparative Genomic Approach
Thesis Credits (I)	6
Candidate	Mr. Palang Chotsiri
Thesis Advisor	Dr. Vethachai Plengvidhya
Co-Advisor	Associate Professor Dr. Supapon Cheevadhanarak Dr. Jittisak Senachak
Program	Master of Science
Field of Study	Bioinformatics
Faculty	School of Bioresources and Technology and School of Information Technology
B.E.	2554

E46298

Abstract

Cyanobacteria are central to carbon and nitrogen cycles, and have significantly contributed to global primary production. They are found in almost every possible habitat, including oceans, fresh water, bare rock, and soil, reflecting the broad range of biosynthetic capabilities of this bacterial group. However, their biological properties, habitats, and environmental niches that govern their genomic content and evolution are poorly understood. In this study, a comparative genomic analysis of 36 completed and 13 nearly completed cyanobacterial genomes from public databases was used to reveal the evolutionary dynamics of cyanobacterial genomes that were driven by the external factors. The analysis of the cyanobacterial orthologous group of proteins (cyanoCOGs) revealed that 570 out of 15,741 protein families are commonly encoded in every cyanobacterial species (strictly core cyanoCOGs). A parsimonious evolutionary scenario algorithm—which required the reconstructed phylogenetic tree from the concatenated core ribosomal protein genes and phyletic patterns of all cyanoCOGs as an input—was implemented in order to uncover the evolutionary scenarios of cyanobacterial genomes. It was found that the modern cyanobacterial genomes with genomic content ranging from 1,717 to 8,383 genes evolved from the last cyanobacterial common ancestor (LCCA) that possessed, conservatively, only 2,468 genes. Considering all core photosynthetic apparatus and some accessory genes, the LCCA was inferred as a photoautotroph, which conveyed their photosynthetic capability to the modern cyanobacteria. However, over evolutionary time, the descendent picocyanobacteria have lost the gene responsible for accessory pigments responsible for adaptation to their niches. It is interesting to note that the extensive gene gain via horizontal gene transfer was found in *Nostocales* and *Oscillatotiales*, particularly genes in the nitrogen fixation process, and extensive gene loss was found in the marine picocyanobacteria.

Keywords: Cyanobacteria / Evolution Scenarios / Comparative Analysis / Phylogenetic Analysis / Gene Loss / Gene Gain.

หัวข้อโครงการวิจัย (I)	การศึกษาเหตุการณ์ในสายวิวัฒนาการของสาหร่ายสีเขียวแกมน้ำเงินด้วยวิธีการศึกษาจีโนมแบบเปรียบเทียบ
หน่วยกิต (I)	6
ผู้เขียน	นายพลัง โชติศิริ
อาจารย์ที่ปรึกษา	ดร. เวทชัย เปล่งวิทยา
อาจารย์ที่ปรึกษาร่วม	รศ. ดร.สุภาภรณ์ ชีวะธนรักษ์ ดร. จิตติศักดิ์ เสนาจักร์
หลักสูตร	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	ชีวสารสนเทศ
คณะ	ทรัพยากรชีวภาพและเทคโนโลยี และ เทคโนโลยีสารสนเทศ
พ.ศ.	2554

บทคัดย่อ

E46298

สาหร่ายสีเขียวแกมน้ำเงิน (cyanobacteria) มีบทบาทสำคัญในวัฏจักรคาร์บอน และ ไนโตรเจน รวมทั้งเป็นแหล่งผลิตชั้นปฐมภูมิที่สำคัญของโลกอีกด้วย สาหร่ายสีเขียวแกมน้ำเงินสามารถพบได้ในสิ่งแวดล้อมทั่วไป อาทิ มหาสมุทร น้ำจืด ดิน แม้กระทั่งบนหิน ซึ่งสะท้อนถึงขีดความสามารถทางชีวภาพที่หลากหลาย ของสิ่งมีชีวิตกลุ่มนี้ แต่อย่างไรก็ตาม ความเข้าใจเกี่ยวกับ ความสัมพันธ์ระหว่างลักษณะทางชีวภาพ การดำรงชีวิต และสภาพแวดล้อม ที่ถูกกำหนดโดยจีโนม และวิวัฒนาการของจีโนมของสิ่งมีชีวิตเหล่านี้ยังคงมีน้อยอยู่ จากการศึกษาจีโนมแบบเปรียบเทียบระหว่าง 36 ชนิดของสาหร่ายสีเขียวแกมน้ำเงินที่ถอดลำดับสมบูรณ์แล้วและที่ยังไม่สมบูรณ์อีก 13 ชนิด พบว่า ความหลายหลายทางวิวัฒนาการของสิ่งมีชีวิตกลุ่มนี้ที่ถูกขับเคลื่อนด้วยปัจจัยทางสิ่งแวดล้อม ในการสร้างกลุ่มของโปรตีนที่มีหน้าที่เดียวกัน (cyanoCOGs) จำนวนทั้งหมด 15,741 กลุ่ม พบว่ามีโปรตีน 570 กลุ่ม จากทั้ง 49 สปีชีส์ (cyanoCOGs) สายวิวัฒนาการที่ถูกสร้างขึ้นมาจาก ไรโบโซมอล โปรตีนที่นำมาต่อกัน รวมกับ รูปแบบการปรากฏของยีนจากแต่ละจีโนม ถูกใช้เพื่อหาเหตุการณ์บนสายวิวัฒนาการของจีโนมของสาหร่ายสีเขียวแกมน้ำเงิน ซึ่งพบว่า สาหร่ายสีเขียวแกมน้ำเงินที่พบในปัจจุบันนี้ ที่มียีนตั้งแต่ 1,717 ยีน ไปจนถึง 8,383 ยีน จากการศึกษาในสายวิวัฒนาการพบว่าบรรพบุรุษร่วมสุดท้าย (LCCA) ที่มียีนอยู่ประมาณ 2,468 ยีน ผลการวิจัยยังบ่งชี้อีกว่า บรรพบุรุษร่วมตัวสุดท้ายของสาหร่ายสีเขียวแกมน้ำเงิน เป็นสิ่งมีชีวิตที่สามารถสร้างอาหารจากการสังเคราะห์แสงได้

E46298

โดยพบว่ามียีนหลักทั้งหมดที่เกี่ยวข้องกับกระบวนการสังเคราะห์แสง รวมทั้งยีนซึ่งประกอบด้วยยีนที่ใช้สังเคราะห์แสงที่มีร่วมกับทั่วทั้งกลุ่มของสาหร่ายสีเขียวแกมน้ำเงิน และยังมียีนเพิ่มเติมอีก โดยที่สาหร่ายสีเขียวแกมน้ำเงินขนาดเล็กในยุคปัจจุบันนี้ ได้ทำยีนเหล่านี้หายไปในช่วงวิวัฒนาการ เพื่อปรับตัวให้เข้ากับสภาพแวดล้อม แต่อย่างไรก็ตาม พบว่ามีการรับยีนใหม่ในสาหร่ายสีเขียวแกมน้ำเงินบางชนิด โดยผ่านกระบวนการส่งผ่านยีนในแนวขนาน (horizontal gene transfer) ถูกพบในกลุ่ม *Nostocales* และ *Oscillatotiales* โดยเฉพาะอย่างยิ่ง ยีนที่เกี่ยวข้องกับกระบวนการตรึงไนโตรเจน แต่อย่างไรก็ดี กลุ่มของสาหร่ายสีเขียวแกมน้ำเงินในทะเลขนาดเล็ก ถูกพบว่ามีการสูญเสียของยีนในสายวิวัฒนาการอย่างมาก

คำสำคัญ: สาหร่ายสีเขียวแกมน้ำเงิน / เหตุการณ์ในสายวิวัฒนาการ / การศึกษาจีโนมเชิงเปรียบเทียบ / การวิเคราะห์เชิงวิวัฒนาการ / การรับลักษณะทางพันธุกรรม / การหายไปของลักษณะทางพันธุกรรม

ACKNOWLEDGEMENT (I)

I would like to thank my advisor, Dr. Vethachai Plengvidhya, who always supports me with professional and valuable guidance. His suggestion and ceaseless contribution were most rewarding and informative to my thesis. I would like to thank Assoc. Prof. Dr. Supapon Cheevadhanarak and Dr. Jittisak Senachak, my co-advisors, for their kindness and all the helpful supervision.

I would like to express my appreciation to all thesis committee members, Dr. Kalyanee Paithoonrangsarid, Dr. Kobkul Laoteng and Dr. Peerada Prommeenate for their kindness and valuable comments throughout this work.

I am appreciative to all my lecturers in the Bioinformatics program at King Mongkut's University of Technology Thonburi (KMUTT) for all given knowledge. I would like to express my gratitude to the program for allowing me to carry out my master study. Moreover, I would like to convey my deepest gratitude to both National Center for Genetic Engineering and Biotechnology (BIOTEC), Thailand and KMUTT for the full scholarship that allows me to finish my master degree.

CONTENTS (I)

	PAGE
ENGLISH ABSTRACT (I)	i
THAI ABSTRACT (I)	ii
ACKNOWLEDGEMENTS (I)	iv
CONTENTS (I)	v
LIST OF TABLES (I)	vii
LIST OF FIGURES (I)	viii
LIST OF ABBREVIATION (I)	ix
CHAPTER	
1. INTRODUCTION (I)	**1**
1.1 Background and Rationale	1
1.2 Objectives	2
1.3 Scope of work	2
1.4 Expected outputs	2
2. LITERATURE REVIEWS (I)	**3**
2.1 Comparative genomic study	3
2.1.1 Homology detection stratigies	4
2.1.2 Evolutionary analysis	9
2.3 Cyanobacteria	10
2.2.1 Photosynthesis in cyanobacteria	13
2.2.2 Evolutionary study in cyanobacteria	15
3. MATERIALS AND METHODS (I)	**16**
3.1 Materials	16
3.1.1 Cyanobacteria genomic data	16
3.1.2 Computer resources	16
3.1.3 Programming language	16
3.1.4 Bioinformatics software and tools	16
3.2 Methodologies	19
3.2.1 Overviews	19
3.2.2 Construction of cyanobacterial clusters of orthologous groups of proteins (cyanoCOGs)	19
3.2.3 Cyanobacterial lineage reconstruction	23
3.2.4 Evolutionary scenarios reconstruction	25
4. RESULTS AND DISCUSSIONS (I)	**29**
4.1 Coverage of cyanobacterial genomes with cyanoCOGs	29
4.2 Phyletic pattern	37
4.3 Phylogenomic of cyanobacteria	41
4.4 Evolutionary scenarios in cyanobacteria lineage	43
4.5 Evolutionary scenarios of photosynthetic apparatus in cyanobacterial lineage	48

5. CONCLUSIONS AND RECOMMENDATIONS (I) **56**
5.1 Conclusion 56
5.2 Recommendations 56

REFERENCES (I) **58**

APPENDIX (I) **63**
A. Full photosynthesis-related genes in ancestral cyanobacterial genomes **63**
B. Implemented Python code for evolutionary scenario algorithm **71**

CURRICULUM VITAE (I) 77

LIST OF TABLES (I)

TABLE		PAGE
2.1	Comparison between various orthology and homology detection methods	6
2.2	Comparison of ortholog databases	8
2.3	Characteristics of the Cyanobacteria Subsections by using the morphological approaches	12
3.1	The cyanobacterial genomes have been included in this study and its general features	17
4.1	The NCBI COGs functional categories	36
4.2	The 25 most frequent phyletic patterns in the cyanoCOGs	39
4.3	The description of cyanobacterial group according to Figure 4.11 relate to their biological properties and environmental niches	47
4.4	The description of cyanobacterial group relate to their biological properties and environmental niches	49
A.1	All photosynthesis-related genes in ancestral cyanobacterial genomes	64

LIST OF FIGURES (I)

FIGURE		PAGE
2.1	Classification of the orthology strategies.	5
2.2	The reference photosynthesis pathway and photosynthetic proteins from KEGG databases	14
2.3	The reference photosynthesis antenna proteins and light harvesting complex from KEGG databases	14
3.1	Overall methodologies for finding the evolutionary scenario of cyanobacterial genes	20
3.2	Flow of the OrthoMCL algorithm for find orthologous group of proteins	21
3.3	Illustration of sequence relationships and similarity matrix construction	22
3.4	The concatenated ribosomal proteins are used for reconstructing the inferred evolution of cyanobacterial lineage	24
3.5	Patterns of events in a parent-children triple according to a parsimonious scenario	26
4.1	Coverage of cyanobacterial genomes with cyanoCOGs in green and NCBI COGs in red	30
4.2	The percentage coverage of cyanobacterial genomes with cyanoCOGs and NCBI COGs	31
4.3	Distribution of the number of species in cyanoCOGs	32
4.4	Distribution of the number of species in cyanoCOGs	33
4.5	Functional breakdown of the entire set of cyanoCOGs of each species the function of cyanoCOGs	34
4.6	The percentage functional breakdown of the entire set of cyanoCOGs of each species	35
4.7	Distribution of phyletic patterns by the number of cyanoCOGs	38
4.8	The phylogenetic networks	40
4.9	The evolutionary tree of cyanobacteria genome reconstructed by using concatenated ribosomal proteins	42
4.10	The cyanobacterial ancestral form represent by "LCCA", "A", "B", "C", "D", "E", "F", and "G"	44
4.11	The summary of gene gain and loss along the cyanobacterial lineages	45
4.12	The functional breakdown of the entire set of cyanoCOGs and the each level of evolutionary timeline the cyanoCOGs	46
4.13	The gain and loss events of the photosynthesis apparatus gene along the cyanobacterial lineages	50
4.14	The photosynthetic tree reconstructed from (A) *PsaI*, (B) *PsaM* proteins depicts that the protein from difference cyanoCOGs has clustered into the difference clades	52
4.15	The evolutionary tree of *PsaI* proteins from cyanobacteria with the *PsaI* protein from plastids of other organismal groups	53
4.16	The phylogenetic tree of *PsaA* (A) and *PsaD* (B) proteins of cyanobacteria with the same proteins from cyanophages	55

LIST OF ABBREVIATION (I)

arCOG	=	Archaea cluster of orthologous groups of proteins
COG	=	Cluster of orthologous groups of proteins
cyanoCOGs	=	Cyanobacterial cluster of orthologous groups of proteins
DNA	=	Deoxyribonucleic acid
EC	=	Enzyme category
HGT	=	Horizontal gene transfer
HMM	=	Hidden Markov model
KEGG	=	Kyoto encyclopedia of genes and genomes
KOG	=	Eukaryotic orthologous groups of proteins
LAB	=	Lactic acid bacteria
LACA	=	Last archaea common ancestor
LCCA	=	Last cyanobacterial common ancestor
LaCOGs	=	*Lactobaillales*-specific cluster of orthologous genes
MEGA	=	Molecular evolutionary genetic analysis
MCL	=	Markov clustering algorithms
MGI	=	Mouse genome informatics
NCBI	=	National center for biotechnology information
RBH	=	Reciprocal best hit
RIO	=	Resampled inference of orthology
RNA	=	Ribonucleic acid
RSD	=	Reciprocal smallest distance