CHARACTERIZATION OF CYANOBACTERIA GENE CONTENT VARIATIONS USING COMPARATIVE GENOMICS

MR. SIVAMOKE DISSOOK

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
2013

Characterization of Cyanobacteria Gene Content Variations Using Comparative Genomic

Mr. Sivamoke Dissook B.Sc. (Chemistry)

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
2013

Thesis Committee

…… Chairman of Thesis Committee
(Researcher, Jittisak Senachak, Ph.D.)

…… Member and Thesis Advisor
(Researcher, Weerayuth Kittichotirat, Ph.D.)

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

…… Member
(Researcher, Vethachai Plengvidhya, Ph.D.)

…… Member
(Researcher, Peerada Prommeenate, Ph.D.)

PREFACE

This thesis was for accomplishment of my master degree of Bioinformatics and Systems Biology. The topic of the study is “Characterization of cyanobacteria gene content variation using comparative genomic” or in Thai is “การศึกษาความหลากหลายในระดับยีนของไซยาโนแบคทีเรียโดยใช้การวิเคราะห์จีโนมเชิงเปรียบเทียบ” This work was done at King Mongkut’s University of Technology Thonburi (KMUTT), Thailand. This thesis consists of five main chapters, which are the introduction, theory and literature reviews, material and methods, results and discussion, and conclusion.

Thesis Title	Characterization of Cyanobacteria Gene Content Variations Using Comparative Genomic
Thesis Credits	12
Candidate	Mr. Sivamoke Dissook
Thesis Advisor	Dr. Weerayuth Kittichotirat
Thesis Co-advisor	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
Academic Year	2013

ABSTRACT

Through billions of years that cyanobacteria make their evolutionary journey, they become a very successful group of organisms on the face of the Earth. They can survive in very diverse environments, even in outer space. Their influences on biosphere could be as large as the mass extinction event. A large number of studies reveal unparalleled potentials of cyanobacteria, which could be applied to many useful applications such as bioactive compound production, novel material production, environmental remediation, food source, alternative energy and much more. Owing to their attractive properties, a large number of cyanobacterial genomes were characterized and made publicly available. This presents a great opportunity to apply comparative genomics to analyze and elucidate cyanobacteria's gene content variation. In this study, we compared the gene content of 100 cyanobacteria genomes of diverse species and identify clade specific genes, core genes, pan genes and genome specific genes according to the phylogenetic information. The integration of phylogenetic information with comparative genomics may yield basic information to answer what make each group of cyanobacteria so special. Results show a total of 28,931 species specific genes and 19,573 clade specific genes, the total specific genes in the set of 100 cyanobacteria accounted for 50% of all genes family, the core genome of cyanobacteria accounted for less than 1% of all genes family. Mathematical extrapolation of the data suggests that the cyanobacteria pan-genome is vast and that unique genes will continue to be identified even after adding hundreds of genomes. Interestingly this study found that there were no common genes encoded for component in the Krebs cycle of cyanobacteria.

Keywords : Comparative genomic / Core genes / Cyanobacteria / Gene family / Pan genes / Phylogenetic / Specific genes

หัวข้อวิทยานิพนธ์	การศึกษาความหลากหลายในระดับยีนของไซยาโนแบคทีเรียโดยใช้การวิเคราะห์จีโนมเชิงเปรียบเทียบ
หน่วยกิต	12
ผู้เขียน	นาย ศิวโมกข์ ดิษสุข
อาจารย์ที่ปรึกษา	ดร. วีรยุทธ กิตติโชติรัตน์
อาจารย์ที่ปรึกษาร่วม	รศ.ดร. สุภาภรณ์ ชีวะธนรักษ์
หลักสูตร	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	ชีวสารสนเทศและชีววิทยาระบบ
คณะ	ทรัพยากรชีวภาพและเทคโนโลยี และเทคโนโลยีสารสนเทศ
ปีการศึกษา	2556

บทคัดย่อ

ไซยาโนแบคที่เรียเป็นสิ่งมีชีวิตที่มีวิวัฒนาการยาวนานหลายพันล้านปีและสามารถปรับตัวให้อยู่รอดภายใต้สิ่งแวดล้อมที่หลากหลาย การศึกษาทำให้เห็นศักยภาพต่างๆ ของไซยาโนแบคทีเรียที่สามารถนำไปประยุกต์ใช้งานได้หลากหลาย เช่น ผลิตสารออกฤทธิ์ทางชีวภาพ ผลิตพลาสติก ฟื้นฟูสภาพแวดล้อม นำมาเป็นแหล่งอาหารของมนุษย์และสัตว์ นำมาสร้างพลังงานทางเลือก เป็นต้น เนื่องจากไซยาโนแบคทีเรียมีลักษณะที่น่าสนใจ จึงทำให้มีการศึกษา ทำข้อมูลจีโนม และเผยแพร่สู่สาธารณะเป็นจำนวนมาก ทำให้เป็นที่น่าสนใจที่จะศึกษาความหลากหลายในระดับยีนของไซยาโนแบคทีเรียโดยใช้การวิเคราะห์จีโนมเชิงเปรียบเทียบ ในการศึกษานี้ได้เปรียบเทียบข้อมูลจีโนมของไซยาโนแบคทีเรีย 100 สายพันธุ์ เพื่อที่จะวิเคราะห์หายีนที่จำเพาะต่อกลุ่มยีนที่มีร่วมกันหรือคอร์ยีน (core genes) ยีนที่ไม่ซ้ำกันทั้งหมดหรือแพนยีน (pan gene) และยีนที่จำเพาะต่อแต่ละสายพันธุ์ตามข้อมูลสายวิวัฒนาการ ซึ่งทำให้ได้ข้อมูลพื้นฐานที่สามารถอธิบายลักษณะพิเศษในแต่ละกลุ่มของไซยาโนแบคทีเรีย จากผลการทดลองพบว่ามียีนที่จำเพาะกับสายพันธุ์ 28,391 ยีน ยีนที่จำเพาะกับกลุ่ม 19,573 ยีน โดยยีนที่จำเพาะกับสายพันธุ์คิดเป็นร้อยละ 50 และคอร์ยีนคิดเป็นน้อยกว่าร้อยละ 1 ของยีนทั้งหมดของไซยาโนแบคทีเรีย การคาดการทางคณิตศาสตร์จากข้อมูลที่มีชี้ให้เห็นว่าจำนวนแพนยีนมีจำนวนมากและเมื่อเพิ่มจำนวนจีโนมที่นำมาวิเคราะห์ ยังจะพบยีนใหม่แม้หลังจากการเพิ่มข้อมูลอีกหลายร้อยจีโนม นอกจากนี้การศึกษายังพบสิ่งที่น่าสนใจ คือไม่พบยีนที่เกี่ยวข้องกับวัฏจักรเครปในคอร์ยีนของไซยาโนแบคทีเรียทั้งหมด

คำสำคัญ : การวิเคราะห์จีโนมเชิงเปรียบเทียบ / คอยีน / ยีนจำเพาะ / สายวิวัฒนาการแพนยีน / ไซยาโนแบคทีเรีย

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my thesis advisor, Dr. Weerayuth Kittichotirat, for keeping an eye on every major point of my work. His valuable suggestions, technical support and discussion were the most rewarding and informative to my thesis. I would like to express my gratitude to my thesis co-advisor, Assoc.Prof.Dr. Supapon Cheevadhanarak, for her kindness, invaluable guidance and dedication to help me with my work. In addition, I would like to thank all members in Systems Biology and Bioinformatics research group (SBI), who help me to complete this work.

I would like to thank all thesis committee members, Dr. Jittisak Senachak and Dr. Peerada Prommeenate and Dr. Vethachai Plengvidhya for the kindness to support my work, valuable comments and encouragement throughout my thesis work.

I would also like to thank the program of Bioinformatics and Systems Biology, King Mongkut's University of Technology Thonburi, for allowing me to carry out my master courses and workshops. I am very grateful to all my lecturers from both the School of Information Technology and the School of Bioresources and Technology for giving valuable knowledge during my study. In addition, I would like to thank National Center for Genetics Engineering and Biotechnology (BIOTEC, Thailand) and KMUTT for a full scholarship.

Finally, I am very grateful to my family for their love, support and encouragement throughout my life.

CONTENTS

		PAGE
ENGLISH ABSTRACT		ii
THAI ABSTRACT		iii
ACKNOWLEDGEMENTS		v
CONTENTS		vi
LIST OF TABLES		viii
LIST OF FIGURES		ix
LIST OF TECHNICAL VOCABULARY AND ABBREVIATIONS		xi
CHAPTER 1 INTRODUCTION		**1**
1.1	Background and rationale	1
1.2	Objectives	2
1.3	Scope of work	2
1.4	Expected outputs	2
1.5	Expected outcome	2
CHAPTER 2 THEORY AND LITERATURES REVIEW		**3**
2.1	Cyanobacteria	3
2.2	Comparative genomic studies	4
2.2.1	Homology	5
2.2.2	Orthologs identification strategies	8
2.2.3	Cluster of orthologous groups (COGs)	9
2.2.4	COGs construction tools	10
2.3	Organismal phylogeny	11
2.3.1	Raw material for molecular phylogeny reconstruction	11
2.3.2	Ribosomal RNA genes	11
2.3.3	Gene family	12
2.3.4	Gene content	12
2.3.5	Logic behind the alignments	12
2.3.6	Tree inference methods	13
2.3.7	Phylogenetic reconstruction tools	13
CHAPTER 3 MATERIALS AND METHODS		**16**
3.1	Genomic data	16
3.2	Data process and COGs construction	16
3.3	Additional annotation	17
3.4	Blast matrix	17
3.5	Phylogenetic tree reconstruction	18

PAGE

3.5.1 16s rRNA tree 18
3.5.2 Core genes tree 18
3.5.3 Phyletic pattern tree 20
3.6 In cooperating COG and tree data 20

CHAPTER 4 RESULTS AND DISCUSSION 21
4.1 Data collection 21
4.2 COGs construction 25
4.3 Additional annotation 31
4.3.1 NCBI functional categories 31
4.3.2 KEGG pathway annotation 32
4.3.3 Transporter annotation 33
4.3.4 Blast matrix 34
4.4 Phylogenetic tree reconstruction 36
4.4.1 16s rRNA 36
4.4.2 Core genes 36
4.4.3 Phyletic pattern 37
4.5 Cluster of orthologous groups and phylogenetic tree integration 41

CHAPTER 5 CONCLUSIONS 48
5.1 Conclusions 48
5.2 Recommendation 48

REFFERENCES 49

APPENDICES 55
Appendix 1 Code translation of NCBI COG categories 56
Appendix 2 Core genes of cyanobacteria 57
Appendix 3 Conserved hypothetical protein among 100 Cyanobacteria 60
Appendix 4 All KEGG pathway mapping results 61
Appendix 5 Full comparison result of all phylogenetic tree using TOPD/FMTS 67
Appendix 6 KEGG mapping result of Cyanobacteria core genes 69
Appendix 7 Supplementary data 73
Appendix 8 Programming script 74

CURRICULUM VITAE 94

LIST OF TABLES

TABLE		PAGE
2.1	Homology: terms and definitions	7
2.2	Comparison of COGs building methods	10
2.3	Comparison of accuracy among tree building tools	14
2.4	Comparison of computing time of maximum likelihood tree building tools	15
3.1	Selected genes for phylogenetic tree construction	18
4.1	Summary information of genomes used in this study	21
4.2	Percent dissimilarity matrix of all against all tree topology comparison among tree building methods	37

LIST OF FIGURES

FIGURE		PAGE
2.1	Used of comparative genomic for solving biological questions	4
2.2	Phylogenetic tree illustrating orthologous and paralogous relationships	5
2.3	Phylogenetic tree illustrating the effect of horizon gene transfer (HGT) on orthologous and paralogous relationships	6
2.4	Classification of orthology finding methods	8
3.1	Steps for choosing reasonably good genomes data in this study, the number in rectangular box shows the number of genomes that passed the criteria	16
3.2	Work flow of additional annotation process	17
3.3	Work flow of the in-house cluster of orthologous group construction method	19
3.4	Work flow of COG and phylogenetic tree integration process	20
4.1	Summary of the number of species in each genus	24
4.2	Scatter plot between genome sizes against number of gene in each genome	24
4.3	Distribution of COG based on number of present genomes	26
4.4	Distribution of specific COG based on each genomes	27
4.5	Core genome size (number of genes) vs number of genome	28
4.6	Pan genome size (number of genes) vs number of genome	28
4.7	Distribution of COG vs genes vs specific genes based on each genomes	29
4.8	Distribution of COGs/genes ratio of vs genome size	30
4.9	Percent of gene family annotated with NCBI functional category	32
4.10	Top 10 most KEGG pathway hit for Cyanobacteria gene families	32
4.11	Percent of type of transporter found in gene families of cyanobacteria	33
4.12	Blast matrix from core gene tree of clade 98	35
4.13	Phylogenetic tree reconstructed form 16s rRNA data. Color label indicates that the organisms came from the same genus	38
4.14	Phylogenetic tree reconstructed form core protein family data	39
4.15	Phylogenetic tree reconstructed form phyletic pattern data	40
4.16	Overall output from integration of COG and phylogenetic tree	41
4.17	General output format of the output file in this study	42

FIGURE **PAGE**

4.18 Number of core gene family annotated with NCBI functional category 42
4.19 Top ten most hit core gene family in Cyanobacteria core genome 43
4.20 Core metabolic pathway of all Cyanobacteria used in this study 44
4.21 Core metabolic pathway of all Cyanobacteria in clade 86 from core genes tree red lines indicated the mapped pathway 44
4.22 TCA component found in cyanobacteria pan genome 45

LIST OF TECHNICAL VOCABULARY AND ABBREVIATIONS

BLAST	=	Basic Local Alignment Search Tools
COG	=	Cluster of orthologous groups
HGT	=	Horizontal Genes Transfer
KAAS	=	Kegg Automatics Annotation Server
KEGG	=	Kyoto Encyclopedia of Genes and Genomes
LGT	=	Lateral Genes Transfer
NASA	=	National Aeronautics and Space Administration
NCBI	=	National Center for Biotechnology Information
PHA	=	Polyhydroxyalkanoates
rRNA	=	Ribosomal RNA
SRA	=	Short Reads Archive
TCA cycle	=	Tricarboxylic acid cycle

CHAPTER 1 INTRODUCTION

1.1 Background and Rationale

Cyanobacteria are photosynthetic prokaryotes found in almost every conceivable habitat on earth [1]. Evidently, cyanobacteria record extends back to around 3,500 million years ago [2]. A number of studies show that cyanobacteria could be used in a wide variety of areas such as cyanobacteria for bioactive compounds [3] cyanobacteria for bio plastics [4], cyanobacteria for bioremediation, cyanobacteria for alternative energy source, cyanobacteria as bio fertilizers, cyanobacteria as a healthy food source [5], etc.

As cyanobacteria are photosynthesis based organisms their life depended on only water, carbon dioxide, inorganic substances and light. However, some cyanobacteria are also able to use heterotrophic nutrition. Most of the time cyanobacteria are the first group of organisms to colonies bare areas of rock and soil. Adaptation capability to the environment of cyanobacteria such as ultraviolet absorbing sheath pigments is of key importance for the success of cyanobacteria. [6]. Moreover a number of cyanobacteria species are able to survive in the soil and other terrestrial habitats. Cyanobacteria are the origin of photosynthesis and plants on Earth. The chloroplast of the plants is actually cyanobacteria that live within the plant cells [7]. Additionally cyanobacteria have symbiotic relationship with other organisms namely bryophytes (Mosses, Hornworts and Liverworts), gymnosperms (Cycads), angiosperms (Gunnera), pteridophytes (Azolla), fungi, lichens, geosiphon pyriformis, and diatoms [8, 9].

Cyanobacteria have gained a lot of attention in many areas thanks to their wide range of exceptional properties. A number of strains of cyanobacteria's genomes have been sequenced and publicly available. Currently they are 250 cyanobacteria sequencing projects where 74 are completed sequence, 108 are scaffold or contig, 9 of them are SRA or traces and 59 of them still don't have data deposited yet (as of April 2014). This presents a great opportunity to apply comparative genomics to analyze and characterize cyanobacteria's gene content variation. In this study, we compared gene content of 100 cyanobacteria genomes of diverse species and identify group specific genes, core genes and genome specific genes according to the phylogenetic information. The in cooperation of phylogenetic information with comparative genomics may yield basic information to answer what make each interested group of cyanobacteria so special.

1.2 Objectives

To identify specific genes of diverse cyanobacteria species and develop comparative genomic pipeline for characterizing diverse species of organisms

1.3 Scopes of Work

1. Collect all available cyanobacteria genome data from a public database such as NCBI.
2. Process genome sequence data and create a cluster of orthologous group data to summarize gene content across all Cyanobacterial genomes.
3. Perform phylogenetic analyses to characterize evolutionary relationships of various Cyanobacteria species.
4. Identify specific genes of all cyanobacteria clade in the phylogenetic tree.

1.4 Expected outputs

1. An up to date collection of cyanobacterial genome sequence and cluster of orthologous group data.
2. A reliable phylogenetic tree summarizing evolutionary relationship of various cyanobacteria.
3. Sets of identified specific genes for every cyanobacteria clade

1.5 Expected outcome

The availability of such data will greatly facilitate researchers who are interested in studying the gene content variation in cyanobacteria.

CHAPTER 2 LITERATURE REVIEWS

As the cyanobacteria is an exceptionally interesting organism. In quest of answering what make each niche and each species of cyanobacteria so special, this studies aim to identify genes that are unique for each cyanobacteria group and species which may be responsible for cyanobacteria survival in different environment. This chapter summarizes the basic knowledge needed for better understanding of the project including the cyanobacteria, comparative genomic, homology, clusters of orthologous group (COGs) and their applications, methods for finding homology, COGs and organismal phylogeny reconstruction.

2.1 Cyanobacteria

"Cyanobacteria" or "blue-green algae" are a group of microorganisms that have gained a lot of attention in biotechnology, because they could be used in a wide variety of areas such as cyanobacteria for bioactive compounds, cyanobacteria for bio plastics (Polyhydroxy alkanoates, PHA), cyanobacteria for bioremediation, cyanobacteria for alternative energy source, cyanobacteria as bio fertilizers, cyanobacteria as a healthy food source, etc. Members of cyanobacteria comprise of unicellular to multicellular prokaryotes, which could perform photosynthesis.

The cyanobacteria or blue-green algae are an ancient group of prokaryotic organisms. Their record extends back to approximately 3,500 million years ago. Their long evolutionary history is considered a reason for the success of cyanobacteria in many habitats and their wide ecological tolerance as most of the time cyanobacteria are the first group of organisms to colonies bare areas of rock and soil. Cyanobacteria are found almost every conceivable habitat on Earth in environments as diverse as Antarctic soils to volcanic hot springs and often where no other vegetation can exist [9]. Even NASA has shown an interest in these organisms. Cyanobacteria have been taken into space to see if they can survive [10]. Cyanobacteria have the capability to carry out oxygen producing photosynthesis, using H_2O as an electron donor for CO_2 reduction, their life depended on only water, carbon dioxide, inorganic substances and light, this ability distinguishing them from all other prokaryotes. A study shows that some cyanobacteria able to use heterotrophic nutrition [11].

Currently they are 250 cyanobacteria sequencing projects, the most frequently sequenced genus were Prochlorococcus, Synechococcus, Microcystis, Cyanothece, and Arthospira. Prochlorococcus was known for their relatively small genome size of around 1.6-2.4 mb. with about 2000 genes comparing to algae which have over 10,000 genes but Prochlorococcus are one of the largest atmospherics oxygen producer [12]. Synechococcus was known for their relatively abundant comparing to others cyanobacteria in various

environment even in nutrient deficit environment [13]. Microcystis was known for their ability to form a harmful algal blooms (HABs) [14]. Cyanothece was known for their rhythm of day and night metabolism, Cyanothece will perform oxygenic photosynthesis during the day time and nitrogen fixation during the night time [15]. Arthospira was well known food supplement as they contained high protein and nutrient [16].

2.2 Comparative genomics studies

Comparative genomics is a prevailing discipline that compares two or more genomes to find out the similarity and differences between the genomes which could be used to answer many basic biology question (Figure 2.1). Owing to the rapid advancement of sequencing technology, large amount of genomic data have been released everyday thus it is very encouraging to use comparative genomic to study these data as a whole and maximizing the biological knowledge discovery power. Such knowledge could give an explanation to many aspect of biological events. For instant, if certain genomic sequence were long conserved over evolutionary lineage, that particular sequence might serve a vital function for the organism's survival. Moreover comparative study can help us identify both coding and non-coding genes and regulatory elements, characterizing the differences between organisms, and tells us what is common and what is unique between different species at the genome level. Furthermore, the function of complex organism such as human genes and other regions may be revealed by studying their counterparts in simpler model organisms [17].

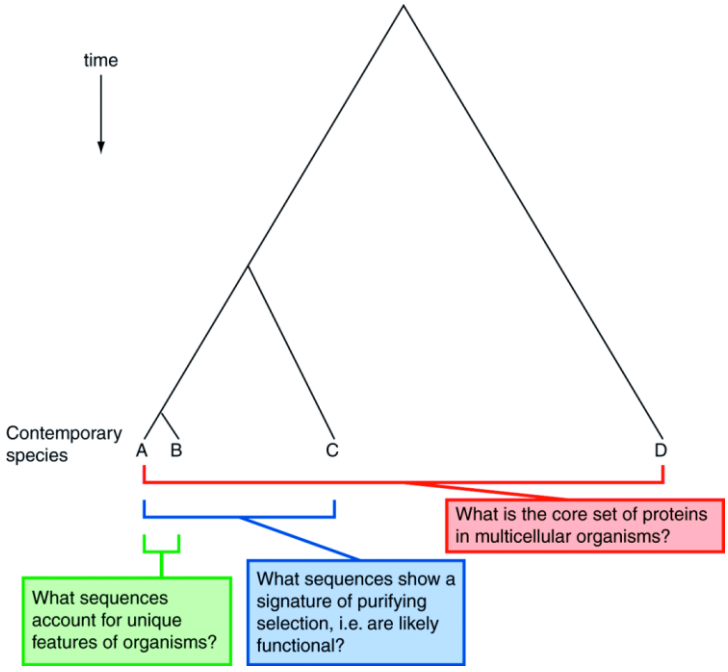


Figure 2.1 A lot of key questions in biology are basically comparative, by comparing genomes at different phylogenetic distance we are able to address different biological questions [18].

2.2.1 Homology

Before the first two complete genome sequences of cellular life forms, the bacteria *Haemophilus influenzae* and *Mycoplasma genitalium*, were released in 1995 the world of biology were known as pre genomic era. During that time, the field of homology were just emerging. The hundred years after that the term orthologs and paralogs were introduced by Walter Fitch [19] where orthologs are term homology was defined by Richard Owen in 1843 as "the same organ in different animals under every variety of form and function" cited by [20]. In 1860, after the publication of "Origin of Species" by Charles Darwin, Thomas Henry Huxley raised the term homology as the evidence of evolution cited by [20]. A genes that originating from a single ancestral gene in the last common ancestor of the compared genomes while paralogs are genes related by duplication. This is illustrated in Figure 2.2

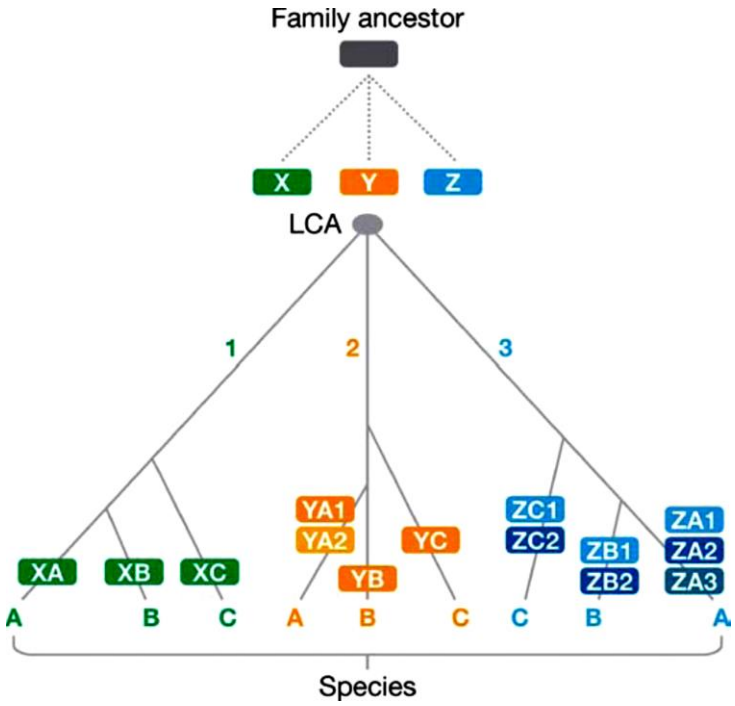


Figure 2.2 Phylogenetic tree illustrating orthologous and paralogous relationships [20].

The example in Figure 2.2 starts with the common ancestor of the entire family, which existed before the last common ancestor of all three compared species which already got three paralogous genes from the family ancestor and became the predecessors of the three branches of the tree. Therefore, each gene in branch 1 is a paralogs of each gene in branches 2 and 3, and vice versa. Branch 1 corresponds to a direct case whereby evolution from the last common ancestor involved only straight down inheritance. Thus the genes in different

species are orthologous to each other. Branch 2 illustrates, a lineage-specific (line of descent) duplication in species (A). Owing to the duplication happened in a single lineage, the paralogs in species (A) are orthologs with all other genes in this branch and since genes YA1 and YA2 have are the same because of they are a duplicates of one another they are called co-orthologs [21] to the genes YB and YC. In branch 3, the condition is more complicated by lineage-specific duplications in each of the species but from the definition genes ZA1-3 are co-orthologs to genes ZB1-2 and ZC1-2. What's more paralogs were classified into two sub-categories; the in-paralogs and out-paralogs where in-paralogs define genes which evolved more recently after the speciation event e.g. genes YA1 and YA2. On the other hand out-paralogs is the genes which evolved via duplication after the speciation event in an ancient time e.g. genes X, Y, and Z.

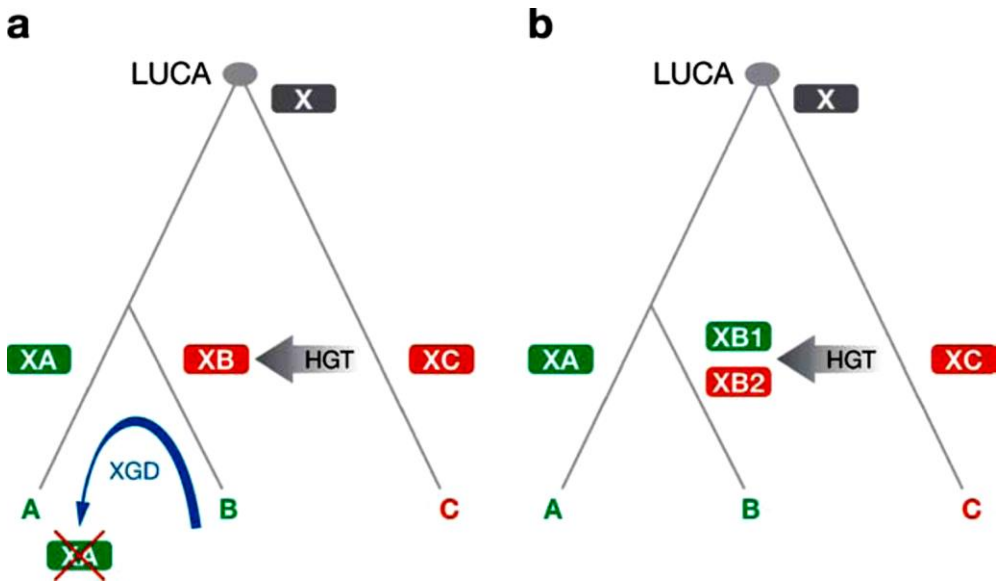


Figure 2.3 Phylogenetic tree illustrating the effect of horizon gene transfer (HGT) on orthologous and paralogous relationships [20].

Moreover, the incident known as horizon gene transfer (HGT) cause even more disturbance of the orthologous and paralogous relationships, HGT is an evolutionary progression that involves transfer of hereditary material among species but does not follow the line of descent from a parent to its offspring. In figure 2.3 a, two species (A and B) have homologous genes that one of them comes from ancestral organism but the other has been acquired through HGT from an outside source and displace the ancestral gene, this phenomenal is known as xenologous gene displacement (XGD). Functionally gene XA and XB would imitate orthologs but they are not, because they do not come from a single ancestral gene in the last common ancestor of the compared species therefore XA and XB are pseudoorthologs or xenologs. In addition, figure 2.3b when a species B acquires gene XB2 via HGT it is

homologous to a gene that already exist. The result of such an occurrence could be described as pseudoparalogy. The summary of homology terminology were described in table 2.1

Table 2.1 Homology: terms and definitions [20]

Homologs	**Genes sharing a common origin**
Orthologs	Genes originating from a single ancestral gene in the last common ancestor of the compared genomes.
Pseudoorthologs	Genes that actually are paralogs but appear to be orthologous due to differential, lineage-specific gene loss.
Xenologs	Homologous genes acquired via XGD by one or both of the compared species but appearing to be orthologous in pairwise genome comparisons.
Paralogs	**Genes related by duplication**
Inparalogs (symparalogs)	Paralogous genes resulting from a lineage-specific duplication(s) subsequent to a given speciation event (defined only relative to a speciation event, no absolute meaning).
Outparalogs (alloparalogs)	Paralogous genes resulting from a duplication(s) preceding a given speciation event (defined only relative to a speciation event, no absolute meaning).
Pseudoparalogs	Homologous genes that come out as paralogs in a single-genome analysis but actually ended up in the given genome as a result of a combination of vertical inheritance and HGT.

2.2.2 Orthologs identification strategies

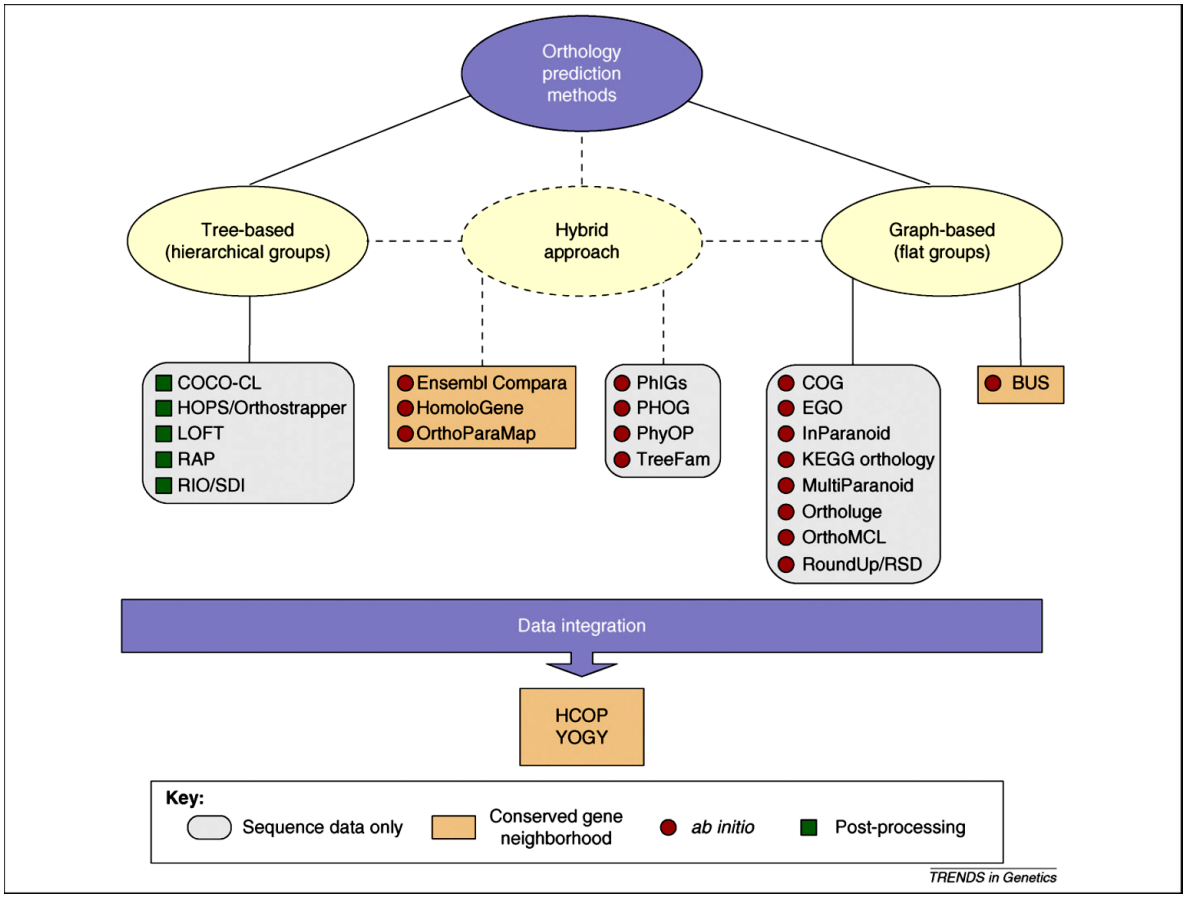


Figure 2.4 Classification of orthology finding methods [22]

There are three approaches for finding orthologous genes (figure 2.4). The first approach is the tree-based methods, which infer orthologous and paralogous relationships with the help of phylogenetic trees. General steps in this method including collecting homologous sequence, constructing multiple sequence alignment, building phylogenetic tree and then the orthologs and paralogs genes can be infer through the study of phylogenetic relations between species. The short comings of this method are; this method can be used if the species tree is reliable and the phylogenetic tree construction algorithms hardly produce totally reliable trees, this method rely heavily on biologically correct multiple sequence alignment thus incorrect alignment will make wrong evolutionary conclusion, lastly this method is very computational intensive therefore not feasible for very large data set.

The second approach is the graph based approach, this method depends on pairwise sequence similarities in all against all manners and assumption of sequence similarity on orthology for example sequences of orthologous genes are more alike to each other than they are to any other genes from the compared genomes, for example they form reciprocal best hits. The common similarity search algorithms are BLAST or Smith-Waterman. Result of orthologs

prediction is heavily rely on the scoring system, some graph based methods in cooperate the clustering technique like Markov clustering algorithm to construct multi species orthologous group. This method is suitable for large dataset but orthologs and paralogs are often inseparable.

The third approach is the hybrid approach, this method combine the principle of the first two methods; for example by using a species tree to guide the clustering process [22].

2.2.3 Cluster of orthologous groups (COGs)

COGs is the way of grouping homologous genes together. It allow us to get maximum information from completely sequenced genomes, each COG made of orthologous genes or orthologous group of paralogs from three or more organisms so it mean that if two proteins from different organisms were group in the same COG, they are orthologs. Orthologs often retain their function so it is possible to infer the unknown function genes from the known function genes exist in the same COG. Conversely after the construction of COGs the remaining conserved genes which their function are not known is a very good candidate for wet lab functional studies. So COGs help to screen the potentially important genes, which functions are unknown. Another benefit of COGs is that after the comparison of all genes, it is possible to identify all highly conserved genes which exist in one group of organisms but not the others. This information might be useful e.g. it can be used to identify the virulence factor of particular viral species and further studies can be made like mapping all the virulence factor genes on to pathway to search for vulnerability or drugs target and more. Furthermore COGs provide a reasonably good raw material for evolution study by identifying all the genes that are conserved across the interested group of organisms.

There are several studies that utilized the COGs information. 2005, Tettelin et al. [23] created *Streptococcus agalactiae* COG consists of 8 Streptococcus genomes to study disease causing genes and have concluded that a universal vaccine is possible only by including dispensable genes. 2006, Armen et al. [24] created cyanobacteria COG consists of 15 cyanobacteria genomes to study the evolution of photosynthetic systems. They suggest that photosynthesis originated in the cyanobacterial lineage under the selective pressures of UV light and depletion of electron donors. 2006, Makarovaa et al. [25] build Lactic acids bacteria COGs consists of 12 Lactic acids bacteria to study their evolutionary relationship. They found combination of extensive gene loss and key gene acquisitions via horizontal gene transfer during the coevolution of lactic acid bacteria with their habitats. 2012, Yuri et al. [26] built archea COGs consists of 120 archaea genomes to study horizontal genes transfer in archaea. They found that gene exchange between major groups of Archaea appears to be largely random, with no major 'highways' of horizontal gene transfer.

2.2.4 COGs construction tools

This part compares a few` well-known orthologous detection and clustering tools as well as the clustering tools that was used in this study (in-house method). Table 2.2 compare well known orthology calling methods. RIO and Orthostrapper apply the phylogeny analysis strategy, which is not feasible for a large data set. RSD, RBH, SBH, and BLASTP apply the strategy based on sequence similarity thus feasible for large data set, but they're lack of clustering capability so they're not feasible for multi-species comparison. Inparanoid, OrthoMCL, KOG, TribeMCL have grouping capability. However Inparanoid can group only 2 species at a time. Thus Inparanoid was not feasible for comparison of a large number of species. KOG, on the other hand is a manual curated database thereby not applicable for this type of study. OrthoMCL and TribeMCL are well known orthologs clustering software which employed Markov clustering algorithm. Nevertheless they do not support incomplete genome data. Therefore the most appropriate method for this study, which make comparison of a large number of genomes and includes incomplete data is the in-house method [27, 28].

Table 2.2 Comparison of COGs building methods [27, 28]

Methods	Strategy	Apply to Proteins	Grouping Capability	Parameter analyzed	Support Incomplete Genomes
RIO	Phylogeny	Pfam domains	NO	Orthology bootstrap cutoff	NO
Orthostrapper	Phylogeny	Pfam domains	NO	Orthology bootstrap cutoff	NO
RSD	Distance	YES	NO	BLASTP E-value cutoff, Divergence cutoff	NO
RBH	BLASTP	YES	NO	BLASTP E-value cutoff	NO
Inparanoid	BLASTP	YES	YES(2 species)	BLASTP E-value cutoff	NO
OrthoMCL	BLASTP	YES	YES	BLASTP E-value cutoff, MCL inflation index	NO
KOG	BLASTP	YES	YES	N/A	NO
SBH	Homology	YES	NO	BLASTP E-value cutoff	NO
BLASTP	Homology	YES	NO	BLASTP E-value cutoff	NO
TribeMCL	Homology	YES	YES	BLASTP E-value cutoff, MCL inflation index	NO
In-house Method	BlLASTN, BLASTP, TBLASTN	YES	YES	BLASTN,BlLASTP, TBLASTN E-value cutoff, sequence identity, alignment coverage	YES

2.3 Organismal phylogeny

When talking about evolutionary study, organismal phylogeny is the way to classify the organism's taxonomy which conventionally can be done using taxonomy key based on observable characteristics or phenotypic of the organism such as morphology, physiology, whole cell properties and so on [29]. However the advancement in DNA sequencing technology have open a rich source of evolutionary information (genotypic data), paving a very promising way to reconstruct the organismal phylogeny [30].

There are huge different between using phenotypic and genotypic data in phylogenetic reconstruction. To start with, the number of definable information of the genotypic data is far more than that of the phenotypic data. For every measurable phenotype, there are responsible genetic sequences that contain thousands of nucleotide that are keeping the accurate record of the organism's evolutionary journey. Moreover, unlike phenotypic innovation, the genotypic data is constantly changing regardless of the changing of the phenotype, which some evolutionists refer as "evolutionary clock" [31]. What's more, the genotypic data is a precisely defined data e.g. 4 letters of nucleotide or 20 character of amino acid. On the other hand, the phenotypic data are defined purely on observation and comparison, which may be affected by the observers. In other words, it is comparable to digital signal versus analogue signal.

2.3.1 Raw material for molecular phylogeny reconstruction

As illustrated in chapter 2.3 that the genotypic data is far more superior to the phenotypic data, the next question is which portion of genome or genes should to be used in phylogeny reconstruction. The problem is the whole genome of an organism consists of some portion that doesn't share the same history with the organism itself. For instance, the genome of endosymbiotically organelle in eukaryote and horizontal gene transfer (HGT) or lateral gene transfer (LGT) in prokaryote which evidently occurring extensively in prokaryotes [32]. Therefore it is essential to identify the portion of the genome, which is not subjected to HGT and share the same history with the organism. Many studies supporting that the genes which involved in the vital activity of the cell like rRNAs genes are suitable for phylogenetic reconstruction [33, 34]. Others argue that phylogeny should be inferred from multiple genes and information from single gene is not enough [35]. Others argue otherwise [36].

2.3.2 Ribosomal RNA genes

Ribosomal ribonucleic acid (rRNA) is the RNA component of the ribosome and is essential for protein synthesis in all living organisms. Ribosomes contain two major type of rRNAs namely large sub unit (LSU) and small sub unit (SSU) rRNAs. For prokaryote the large subunit contains 23s and 5s rRNA, the small subunit contains 16s rRNA. For eukaryote the large subunit contains 28s, 5.8s and 5s rRNA, the small subunits contain 18s rRNA. Focusing

on prokaryotes, the 23s rRNA genes contain about 2900 base pairs, the 5s rRNA contain about 120 base pairs and the 16s rRNA contain about 1500 base pairs. Considering the size of genes, the longer the gene is the more information it contains, which should results in better taxonomy classification. This is true when comparing the phylogenetic tree result of 16s rRNA with the 5s rRNA [37-39]. But in the case of 23s rRNA, is not entirely true because the average rate of sequence change of 23s rRNA is significantly higher than 16s rRNA. Therefore, the longer length of 23s rRNA doesn't give better informative signal than 16s rRNA. Generally 16s rRNA and 23s rRNA tree likely to be very similar. Another reason why rRNAs are good molecule for phylogenetic study especially the small subunit rRNA is that its mutation tend to be nucleotide replacements and the insertion deletion events are occur in a certain area of the molecule. This is important because the rapidly changing position in the sequence added a random noise and increasing the chance of error in the analysis.

2.3.3 Gene family

Gene family or Cluster of orthologous groups contain orthologous genes, which may come from many organisms in the compared set. Thus the core cluster (orthologous gene that found across all genomes) could be used as material for phylogeny reconstruction. However not all core cluster are good for building tree. For example, paralogous genes may cause bias problem if not choose carefully, or cluster which happened to be transferred horizontally. Lerat et.al. [35] suggested that the universally single copy cluster of orthologous group are robust to horizontal gene transfer and should be used to build phylogenetic tree.

2.3.4 Gene content

Another way of inferring that one organism is closer to another is by comparing their gene content. The more genes they have in common, the closer to each other they are. In order to use gene content to build phylogenetic tree, one have to identify the homologous relationship of all genes in all target organisms and then extract the present and absent pattern of all genes known as phyletic pattern. Finally the phylogenetic tree could be inferred by similarity of the phyletic pattern. This method only looks at the present moment and does not concern about horizontal gene transfer. Thus the result of this method should be concern as the similarity of the genome rather than an evolutionary relationship. Moreover the incompleteness of the genome sequence data would have a direct effect on tree topology [40].

2.3.5 Logic behind the alignment

The soul justification to believe in phylogenetic tree is phylogenetic tree based on the assumption that the compared sequences come from a common ancestor. Therefore the first step in planting a molecular phylogenetic tree starts with multiple sequence alignment. The reasons behind this are to organize the sequence so that the homologous part of the sequence

is in the same column. After that we assume that all residues placed in each position is the same residue that come from the common ancestor. Then the differences can be compared.

2.3.6 Tree inference methods

There are three main methods for inferring phylogenetic tree using genotypic information; distance [41], maximum parsimony [42], and maximum likelihood [43].

Distance method is based on the assumption that the organisms which share common ancestor should be more similar to each other than that which common ancestor was more ancient. This method doesn't take the information from the sequences directly but calculates all against all pairwise distance of the sequences and transforms them into dissimilarity matrix and finally grouping the closest sequences together. The advantage of this method is that it requires minimal amount of computational power. On the other hand it makes the least use of the sequence information because it doesn't utilize information of individual sequence position.

Maximum parsimony method is the method that analyzes the sequence information directly. This method is based on the principle known as Occam's razor [44] "other things being equal, the simplest explanation --- the most parsimonious one--- should be chosen". Believing in parsimony, a good phylogenetic model is the model that makes minimal steps of evolution or change. This method is good when the amount of change is small. However, when the amount of differences is large in many organisms, the systematic bias will increase [45].

Maximum likelihood method is the most popular method to date. It measures the probability that the generated tree model would result in the observed data set. The down side of this method is that all possible trees are considered thus it requires considerably a lot of computing power [46].

Benchmarking of these three methods show that the maximum likelihood produce more correct tree than any other methods [47].

2.3.7 Phylogenetic reconstruction tools

As described in section 2.3.4, the maximum likelihood is the best phylogenetic reconstruction method. There are a number of phylogenetic reconstruction tools available. The major factors that were considered in choosing phylogenetic tree building tool were accuracy and computational time. Often time that the most accurate tree building method takes considerable amount of time, which is not feasible for this study.

Table 2.3 Comparison of accuracy among tree building tools [48]

#Sequences	250	1,250	5,000	78,132
Type	a.a.	a.a	a.a.	nt
RAxML 7 (JTTCAT + SPRs)	90.50%	88.40%	88.40%	--
PhyML 3.0 (Γ_4 + SPRs)	89.90%	--	--	--
FastTree 2.0.0 (JTT+CAT or JC+CAT)	86.90%	83.70%	84.30%	92.10%
PhyML 3.0 (Γ_4, no SPR)	86.00%	--	--	--
PhyML 3.0 (no gamma, no SPR)	81.70%	80.10%	--	--
FastME 1.1 (log-corrected distances)	79.60%	77.70%	75.30%	--
BIONJ (max-lik. distances)	77.70%	73.70%	73.10%	--
Parsimony (RAxML)	76.80%	76.50%	69.40%	--
BIONJ (log-corrected distances)	76.60%	73.00%	72.30%	--
Neighbor-joining (log-corrected distances)	76.00%	72.60%	71.60%	66.10%
Clearcut 1.0.8 (log-corrected distances)	75.50%	72.30%	71.50%	58.10%

Table 2.3 illustrates the Comparison of accuracy among tree building tools based on non-maximum likelihood methods. The number of percentage shows topological accuracy for simulated alignments with varying numbers of sequences. RAxML 7 is the best tool for building phylogenetic tree in term of the accuracy of tree topology, followed by PhyML 3.0 with SPR and FastTree 2.0.

Table 2.4 Comparison of computing time of maximum likelihood tree building tools [48]

Alignment	# Distinct Sequences	#Positions	FastTree2.0	RAxML7	PhyML 3
16s rRNA, subsets	500	1287 nt	0.02	2.2	2.9
COGs, subsets	500	65 - 1099 a.a.	0.02	5.2	7.2
COGs, subsets	2,500	197 - 384 a.a	0.11	61	-
Efflux permeases	8,236	394 a.a.	0.25	197	>1200
16S ribosomal RNAs, families	15,011	1287 nt.	0.66	64	>2000
ABC transporter	39,029	214 a.a	1.02	-	-
16S rRNAs, all	237,882	1287 nt.	21.8	-	-

Table 2.4 illustrates comparison of computing time of maximum likelihood tree building tools in hours. All runs used a single thread of execution. All runs accounted for variable rates across sites, using CAT for RAxML 7 and FastTree 2 or C4 for PhyML 3. All FastTree runs include local SH-like supports and all RAxML runs include branch lengths under C4. RAxML and PhyML were run without support values (no bootstrap). For random subsets of 500 16S rRNAs or for COGs, the study showed average running times. For alignments with over 1,000 sequences, the study used RAxML 7's fast convergence option [48]. Although RAxML7 shows supreme topology accuracy, it is clear that FastTree 2.0 is about 100 and 150 times faster than RAxML 7 and PhyML 3 respectively with nucleotide data and much faster with amino acid data. Considering that this study involved using the concatenated nucleotide from core orthologous genes, which is around 10,000 nucleotide long with 100 distinct sequences, the FastTree 2.0 was logical choice for this study. Besides, FastTree 2.0 yield 92.10 % for topological accuracy with nucleotide data, which is better than its amino acid mode.

CHAPTER 3 MATERIALS AND METHODS

This chapter illustrates all the methods and materials in this study. Most of the tools used in this study were written or put together as pipe lines using Perl language. Perl is a high-level, general-purpose, programming language. Perl is one of the programming languages that is popularly used in bioinformatics applications. In fact, Perl is the language that used to process data in the human genomes project. Perl is available for a huge variety of operating systems that people uses today; Windows, UNIX, Linux, Mac, etc. Moreover there are available tools that can facilitate biological computation which written in Perl known as bio Perl package [49].

3.1 Genomic data

All genomic data were gathered from NCBI database in genbank format. This study selected only a reasonably good genomes data by using a series of filtering criteria (Figure 3.1). Firstly, genome file must contain annotation data. Secondly, the annotated genomes must contain 16s rRNA gene. Thirdly, genome data must contain less than 500 scaffolds and finally any genomes data causing an error during the COGs building process were eliminated.

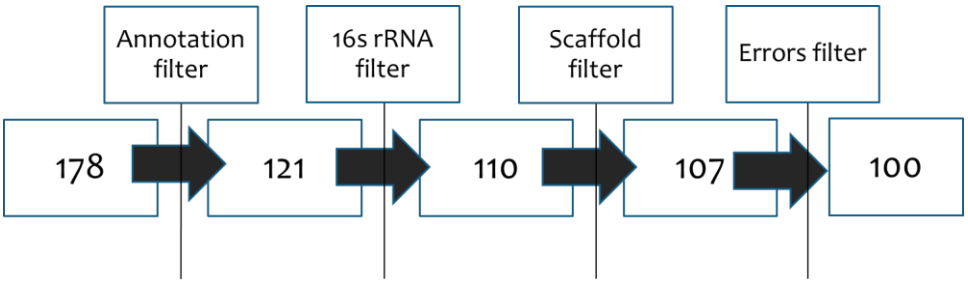


Figure 3.1 Steps for choosing reasonably good genomes data in this study, the number in rectangular box shows the number of genomes that passed the criteria

3.2 Data process and COGs construction

The selected genbank files were processed using BioPerl package and Perl script to extract biological data and prepare file for COGs construction**.**

Having included incomplete genome data into the study, this study adopted the method of creating the cluster of orthologous group from Kittichotirat, Weerayuth, et al. [28] (in-house method). This takes advantages of multiple BLAST programs to capture all of possible genes in uncompleted genomic data. In this study the criteria for determining the homologous genes were 30% sequence identity 50% sequence coverage and BLAST e-value of less than 1e-6. The overall COG building workflow is described in figure 3.3.

3.3 Additional annotation

Once the COGs is done, this study perform cellular function assignment by using representative sequence of each COG, which is the gene with the longest length, and blast against the NCBI cog database [50]. The best BLAST hit result with e-value less than 10e-6 is then used for the cellular function annotation. The code for functional categories were describe in appendix 1. Moreover this study also performed the pathway mapping annotation for each COG by sending the representative sequence as described earlier to KEGG pathway Automatic Annotation Server (KAAS) [51]. Using the best hit (BEH) method with the Cyanobacteria' pathway data set and server's default criteria to obtain the annotations. Additionally, the in-house database for transporter proteins were in cooperate into the COG data using representative sequence of each COG blast against the transporter database and used the best hit result with e-value less than 10e-6 for annotation.

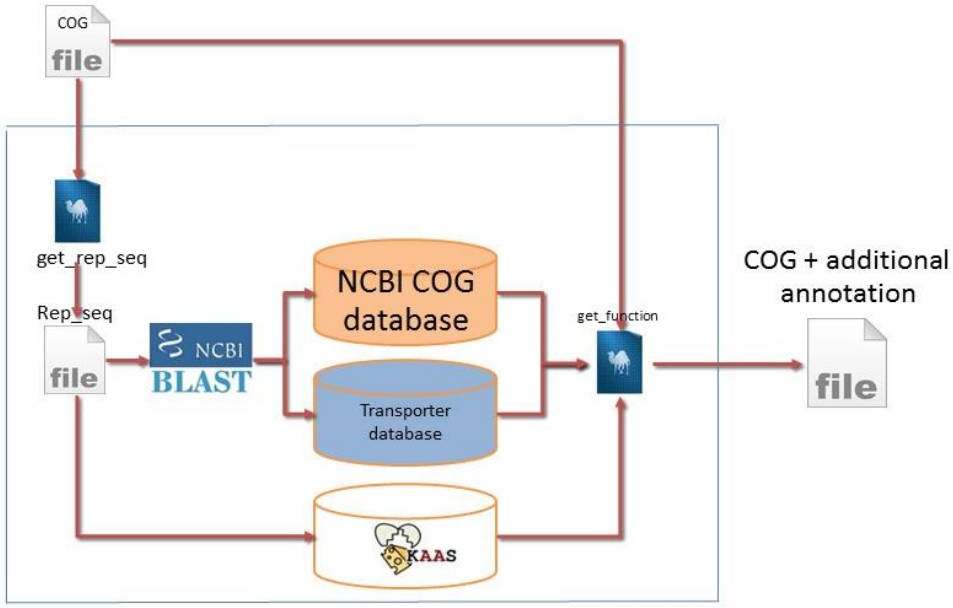


Figure 3.2 Work flow of additional annotation process

3.4 Blast matrix

The BLAST matrix is a visual representation of a pairwise proteome comparison using BLAST (Basic Local Alignment Tool) implemented in CMG biotools [52]. In this study, we created a Perl script for controlling CMG biotools to construct the BLAST matrix for every phylogenetic clade to provide additional information regarding pairwise strains comparison within clade using default parameters.

3.5 Phylogenetic tree reconstruction

Owing to the hot debates in the field of evolutionary study, this study identified the clade specific genes based on three approaches for making phylogenetic tree.

3.5.1 16s rRNA tree

The 16s rRNA genes of all cyanobacterial genomes were called using RNAmmer 1.2 [53]. The nucleotide sequences 16s rRNA genes were aligned using ClustalW version 2.1 with default parameters [54]. All columns with gaps were removed from alignment results, producing an alignment of 1,192 base pairs in FASTA format. FasTreeMP program version 2.1 was used to construct the tree using default parameter.

3.5.2 Core gene tree

Twelve core genes (Table 3.1) that are found to be universally single copy across 100 cyanobacteria strains were selected for analysis. The nucleotide sequences of individual genes were aligned using ClustalW version 2.1 with default parameters. All columns with gaps were removed from alignment results, which in turn were concatenated to produce a single alignment (a total length of 13,090 base pairs) in FASTA format. FastTreeMP program version 2.1 was used to construct the phylogenetic tree.

Table 3.1 selected genes for phylogenetic tree construction

Cluster ID	Product description	Length(bp.)
p-cluster063051	glutamate/cysteine ligase	1218
p-cluster099073	50S ribosomal protein L6	576
p-cluster129454	ATP synthase, F1 delta subunit	564
p-cluster170423	phosphoglycerate kinase	1242
p-cluster281234	glycine/serine hydroxymethyltransferase	1371
p-cluster295329	protein translocase subunit secY/sec61 alpha	1398
p-cluster341618	SSU ribosomal protein S18P	228
p-cluster354858	translation elongation factor P	600
p-cluster385376	4-hydroxy-3-methylbut-2-enyl diphosphate reductase	1281
p-cluster396713	preprotein translocase subunit SecA	2937
p-cluster416475	transfer RNA-Gln reductase	1392
p-cluster440341	glycyl-tRNA synthetase beta chain	2238

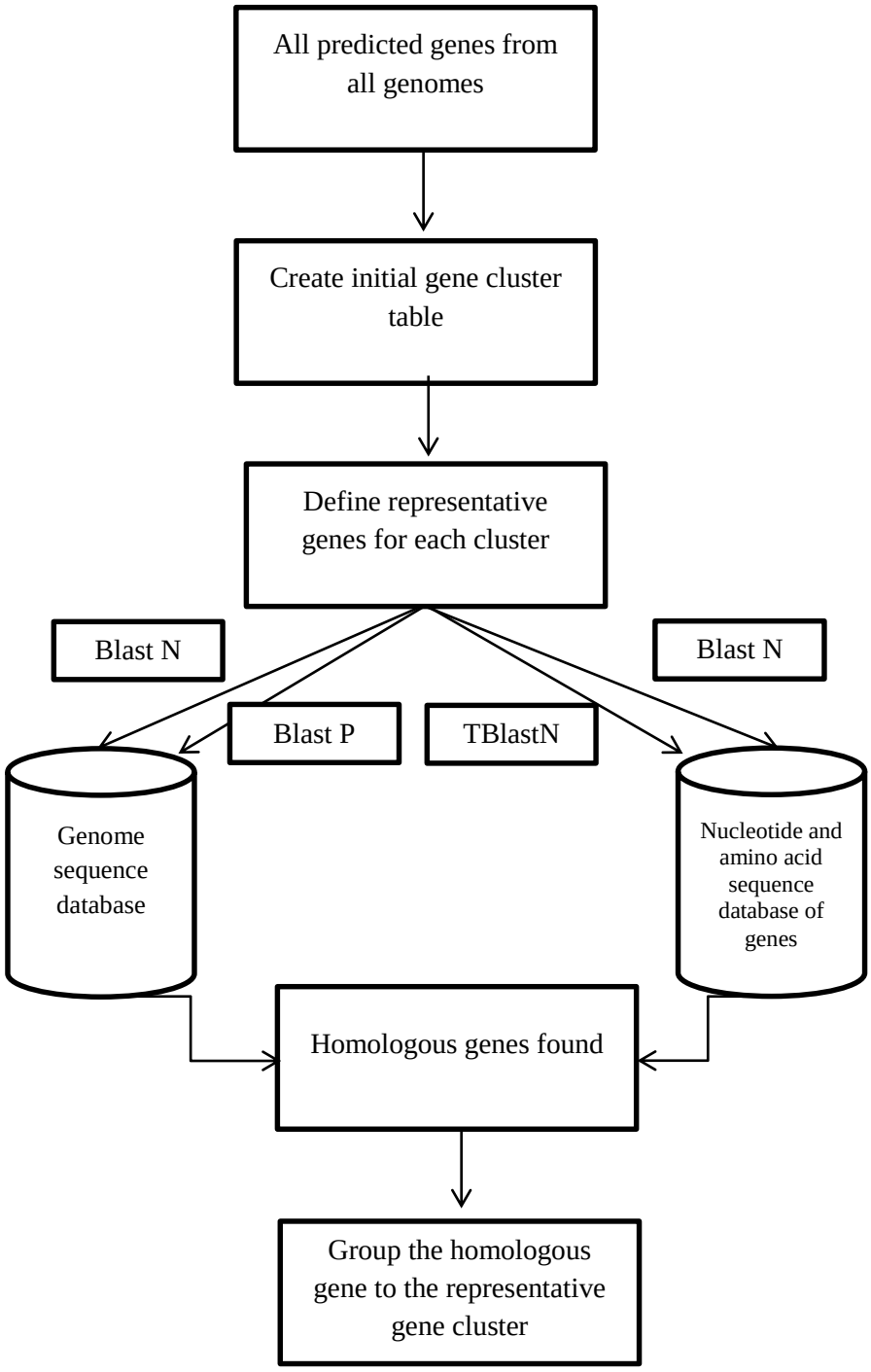


Figure 3.3 Work flow of the in-house cluster of orthologous group construction method

3.5.3 Phyletic pattern tree

The construction of dendrogram gene content tree from phyletic patterns was performed using a phyletic pattern indicating the presence/absence of the protein coding genes of all species present in the Clusters of Orthologous Group (COGs) data. The pattern then was cluster using MeV program [55]. All phylogenetic trees were compared in the aspect of agreement with TOPD/FMTS program [56].

3.6 In cooperating COG and tree data

Phylogenetic information as Newick format were used to group phylogenetically related cyanobacteria species from smallest group expanding to the largest group which cover the entire study organism set (100 genomes). The grouping information then further use to group COG which are specific to each clade or cyanobacterial strains. Moreover the grouping information was further used to determine the core and pan genome of each cyanobacteria clade.

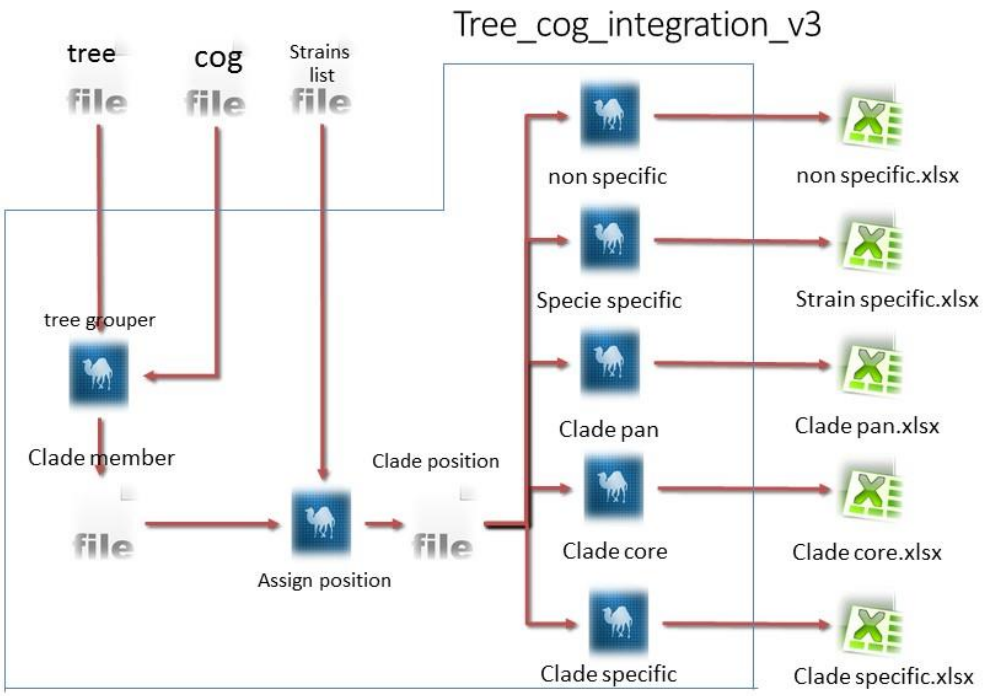


Figure 3.4 Work flow of COG and phylogenetic tree integration process

CHAPTER 4 RESULTS AND DISCUSSION

4.1 Data collection

Initially 178 genomes were downloaded from NCBI database, 57 of them contain no annotations so were eliminated, 11 of them contain no 16s rRNA annotation and were eliminated, 3 of them have more than 500 scaffold and were eliminated and 7 of them caused errors during the COGs construction therefore were eliminated. Therefore a total of 100 genomes were chosen for the study (Table 4.1).

Table 4.1 Summary of genome sequence data used in this study

Organism_name	Number of genes	Genome size (Mbp)	Number of Coding sequences	Number of Contigs/ Scaffold
Acaryochloris marina MBIC11017	6329	6503.724	6254	1
Anabaena cylindrica PCC 7122	5393	6395.836	5320	1
Anabaena sp. 90	4570	5305.675	4511	5
Anabaena variabilis ATCC 29413	5102	6365.727	5043	1
Arthrospira maxima CS-328	5728	6003.314	5690	129
Arthrospira platensis C1	6153	6089.21	6108	1
Arthrospira platensis NIES-39	6676	6788.435	6630	1
Calothrix sp. PCC 6303	5446	6767.834	5393	1
Calothrix sp. PCC 7507	6031	7023.215	5950	1
Chamaesiphon minutus PCC 6605	5561	6284.095	5498	1
Chroococcidiopsis thermalis PCC 7203	5463	6315.792	5408	1
Crinalium epipsammum PCC 9333	4529	5315.554	4474	1
Crocosphaera watsonii WH 8501	5996	6238.156	5958	323
Cyanobacterium aponinum PCC 10605	3410	4114.099	3358	1
Cyanobacterium stanieri PCC 7202	2889	3163.381	2837	1
cyanobacterium UCYN-A	1241	1443.806	1199	1
Cyanobium gracile PCC 6307	3330	3342.364	3280	1
Cyanothece sp. ATCC 51142	4813	4934.271	4762	1
Cyanothece sp. ATCC 51472	5126	5428.187	5083	48
Cyanothece sp. PCC 7424	5280	5942.652	5227	1
Cyanothece sp. PCC 7425	5019	5374.574	4966	1
Cyanothece sp. PCC 7822	5478	6091.62	5422	1
Cyanothece sp. PCC 8801	4309	4679.413	4260	1
Cyanothece sp. PCC 8802	4368	4669.813	4320	1
Cylindrospermopsis raciborskii CS-505	3500	3879.03	3449	93
Cylindrospermum stagnale PCC 7417	5892	7003.56	5811	1
Dactylococcopsis salina PCC 8305	3388	3781.008	3337	1
Fischerella sp. JSC-11	4671	5380	4627	34
Geitlerinema sp. PCC 7407	3870	4681.111	3815	1
Gloeobacter violaceus PCC 7421	4478	4659.019	4430	1

Organism_name	Number of genes	Genome size (Mbp)	Number of Coding sequences	Number of Contigs/ Scaffold
Halothece sp. PCC 7418	3763	4179.17	3708	1
Leptolyngbya sp. PCC 6406	5236	5606.57	5190	377
Leptolyngbya sp. PCC 7375	7900	9422.968	7828	5
Leptolyngbya sp. PCC 7376	4278	5125.95	4228	1
Lyngbya aestuarii BL J	6568	6873.508	6515	439
Lyngbya sp. PCC 8106	6185	7037.511	6142	110
Microcoleus sp. PCC 7113	6106	7470.429	6027	1
Microcoleus vaginatus FGP-2	5209	6698.929	5134	40
Microcystis aeruginosa NIES-843	6360	5842.795	6312	1
Microcystis aeruginosa PCC 7941	4563	4801.899	4520	77
Microcystis aeruginosa PCC 9432	4803	4994.942	4760	132
Microcystis aeruginosa PCC 9443	4824	5185.041	4780	221
Microcystis aeruginosa PCC 9701	4527	4755.998	4483	323
Microcystis aeruginosa PCC 9717	4880	5300.338	4836	264
Microcystis aeruginosa PCC 9806	4038	4262.564	3994	93
Microcystis aeruginosa PCC 9807	4828	5155.707	4784	267
Microcystis aeruginosa PCC 9808	4888	5051.045	4845	141
Microcystis aeruginosa PCC 9809	4724	5011.016	4680	303
Microcystis aeruginosa SPC 777	5289	5454.996	5241	278
Microcystis aeruginosa TAIHU98	5404	4849.611	5356	4
Microcystis sp. T1-4	4478	4693.747	4434	449
Nodularia spumigena CCY9414	4904	5316.258	4860	204
Nostoc punctiforme PCC 73102	6186	8234.322	6086	1
Nostoc sp. PCC 7107	5327	6329.823	5237	1
Nostoc sp. PCC 7120	5425	6413.771	5365	1
Nostoc sp. PCC 7524	5415	6635.03	5355	1
Oscillatoria acuminata PCC 6304	5779	7689.443	5704	1
Oscillatoria nigro-viridis PCC 7112	5859	7479.014	5781	1
Oscillatoria sp. PCC 6506	5891	6676.705	5822	377
Oscillatoriales cyanobacterium JSC-12	4831	5530.491	4780	1
Pleurocapsa sp. PCC 7327	4319	4986.817	4268	1
Prochlorococcus marinus str. AS9601	1961	1669.886	1920	1
Prochlorococcus marinus str. MIT 9211	1897	1688.963	1854	1
Prochlorococcus marinus str. MIT 9215	2021	1738.79	1982	1
Prochlorococcus marinus str. MIT 9301	1946	1641.879	1906	1
Prochlorococcus marinus str. MIT 9303	3046	2682.675	2997	1
Prochlorococcus marinus str. MIT 9312	1852	1709.204	1810	1
Prochlorococcus marinus str. MIT 9313	2321	2410.873	2269	1
Prochlorococcus marinus str. MIT 9515	1945	1704.176	1905	1

Organism_name	Number of genes	Genome size (Mbp)	Number of Coding sequences	Number of Contigs/ Scaffold
Prochlorococcus marinus str. NATL1A	2234	1864.731	2193	1
Prochlorococcus marinus str. NATL2A	2203	1842.899	2162	1
Pseudanabaena biceps PCC 7429	4803	5476.421	4757	464
Pseudanabaena sp. PCC 7367	3613	4557.046	3561	1
Raphidiopsis brookii D9	3057	3186.511	3007	47
Rivularia sp. PCC 7116	6670	8698.463	6609	1
Rubidibacter lacunae KORDI 51-2	3501	4153.658	3457	99
Stanieria cyanosphaera PCC 7437	4377	5041.209	4328	1
Synechococcus elongatus PCC 6301	2574	2696.255	2523	1
Synechococcus elongatus PCC 7942	2662	2695.903	2612	1
Synechococcus sp. BL107	2553	2285.034	2507	1
Synechococcus sp. CC9311	2942	2606.748	2892	1
Synechococcus sp. CC9605	2696	2510.659	2645	1
Synechococcus sp. CC9902	2357	2234.828	2306	1
Synechococcus sp. JA-3-3Ab	2813	2932.766	2760	1
Synechococcus sp. PCC 6312	3557	3697.276	3513	1
Synechococcus sp. PCC 7002	2872	3008.047	2824	1
Synechococcus sp. PCC 7502	3297	3510.253	3248	1
Synechococcus sp. RCC307	2579	2224.914	2534	1
Synechococcus sp. RS9916	3009	2664.873	2961	1
Synechococcus sp. RS9917	2820	2584.918	2770	1
Synechococcus sp. WH 5701	3401	3280.236	3346	116
Synechococcus sp. WH 7803	2583	2366.98	2533	1
Synechococcus sp. WH 8016	3037	2694.843	2992	16
Synechococcus sp. WH 8102	2569	2434.428	2519	1
Synechocystis sp. PCC 6803	3207	3569.561	3159	1
Synechocystis sp. PCC 6803 substr. PCC-N	3216	3570.103	3168	1
Synechocystis sp. PCC 6803 substr. PCC-P	3216	3570.114	3168	1
Thermosynechococcus elongatus BP-1	2521	2593.857	2476	1
Trichodesmium erythraeum IMS101	4494	7750.108	4451	1
Xenococcus sp. PCC 7305	5414	5929.641	5373	234

Of all 100 cyanobacteria genomes, 20 were in the genus of Synechococcus, 13 were Microcystis, 10 were Prochlorococcus, 7 were Cyanothece, 4 were Nostoc, and others were described in figure 4.1

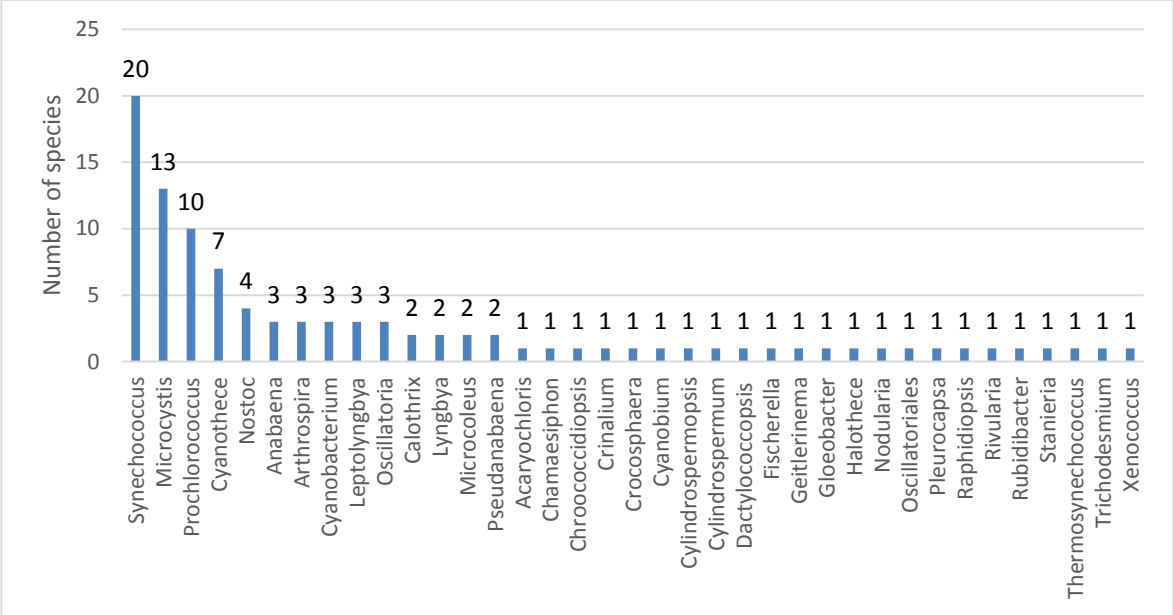


Figure 4.1 Summary of the number of species in each genus

The scatter plot between genome sizes against number of gene in each genome (figure 4.2) shows positive linear correlation, which was expected of a reasonably good genes calling and annotated genomes.

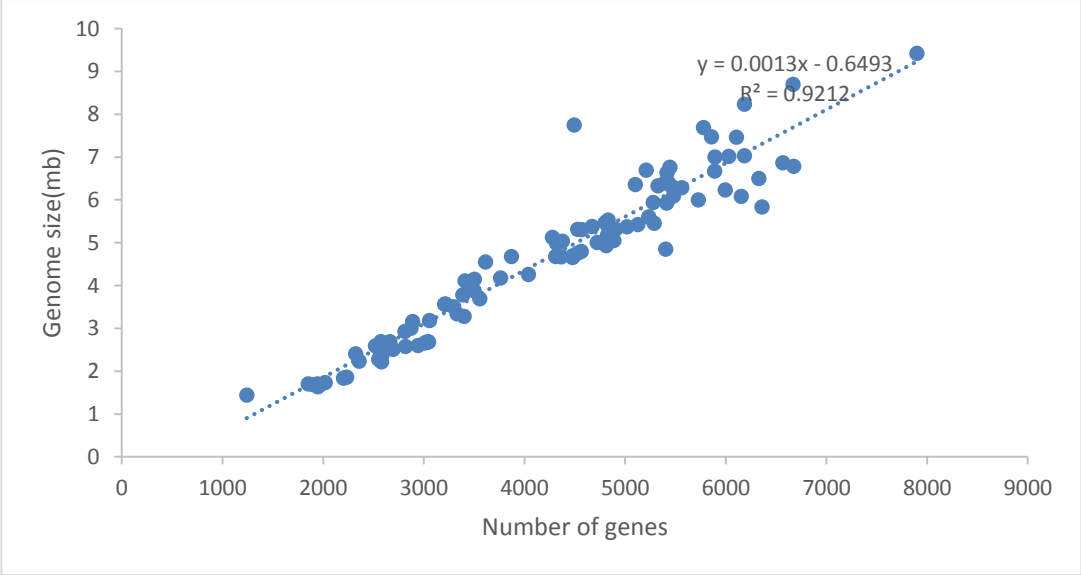


Figure 4.2 Scatter plot between genome sizes against number of gene in each genome

4.2 COGs construction

A total of 424,662 protein coding genes from 100 cyanobacteria genomes were classified into 58,920 cluster of orthologous group using the algorithm describe in the material and methods. Notably 42,484 of cluster (72.1%) were annotated as hypothetical protein suggesting that there are much to discover about the biology of cyanobacteria. Figure 4.3 illustrates the distribution of the number of genes per COG. Interestingly, although the criteria for grouping homologous group is not stringent comparing to the study of *Aggregatibacter actinomycetemcomitans* [28], the results suggested that about 50% of COG (27762/58920) are unique genes which can be found in single cyanobacteria strain only. Figure 4.4 describe the distribution of the number of unique genes found in each genome. The cyanobacteria that have a largest number of genome specific gene is *Leptolyngbya sp. PCC 7375* [57] and the second largest is *Acaryochloris marina MBIC11017* [58].

Moreover, some COG have a large number of genes. The largest COG contains 2,706 genes, which inherited the annotation of ABC transporter which is known to be the largest gene group [59, 60]. The second largest COG contains 1340 genes, corresponds to response regulator receiver modulated diguanylate cyclase [61].

A total of 401 genes family were found to be core genes (share in all strains) (appendix 2) and 56,520 are considered variable genes (found in some but not all strains). Most of the core genes have some functional annotation associated with the house keeping functions. This suggests that the highly conserved genes among variety of cyanobacteria play important roles that may be necessary for organism's survival. Nevertheless this study found 23 hypothetical protein in the core gene set which might be interesting to make further investigations (appendix 3). Interestingly the core genes of cyanobacteria accounts for less than 1% of any single genome. Comparing to other bacterial species. Welch et al. [62] studied three strains of *Escherichia coli* and revealed that less than 40% of the genomes were core genes. Seemingly the size of the core genome will get smaller with additional *E. coli* strains included for comparison. In a study of 12 Procholorococcus isolates, the size of the core genome approached 1,250 genes, or from 40% to 67% of the genomes of individual isolates. In study of 14 *Aggregatibacter actinomycetemcomitans,* core genes found to be accounted for 70.6% to 83.3% of any single genome [28]. This is reasonable owing to the organism that are categorized as cyanobacteria can be very diverse and therefore their genomes can be very different and share less core genes. Figure 4.5 illustrates that the number of conserved genes of cyanobacteria decreased as the new genomes were introduced. This result is consistent with [28] and [62]. Figure 4.6 illustrate that the number of pan genes of cyanobacteria increased as new genomes were introduced. The equation describing the trend line in figure 4.6 estimated that for every one genome added around 5000 new genes were added to the pan genome. The result suggested that each Cyanobacteria contain a lot of unique genes which is consistent with result from Figure 4.3 and 4.4.

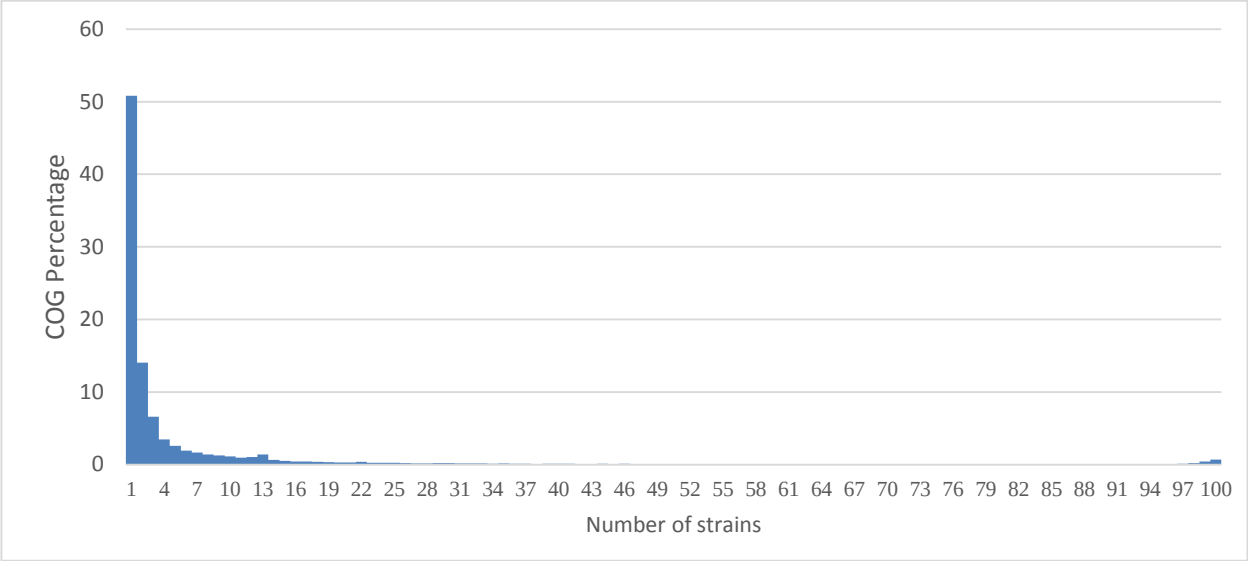


Figure 4.3 Distribution of COG based on number of present genomes

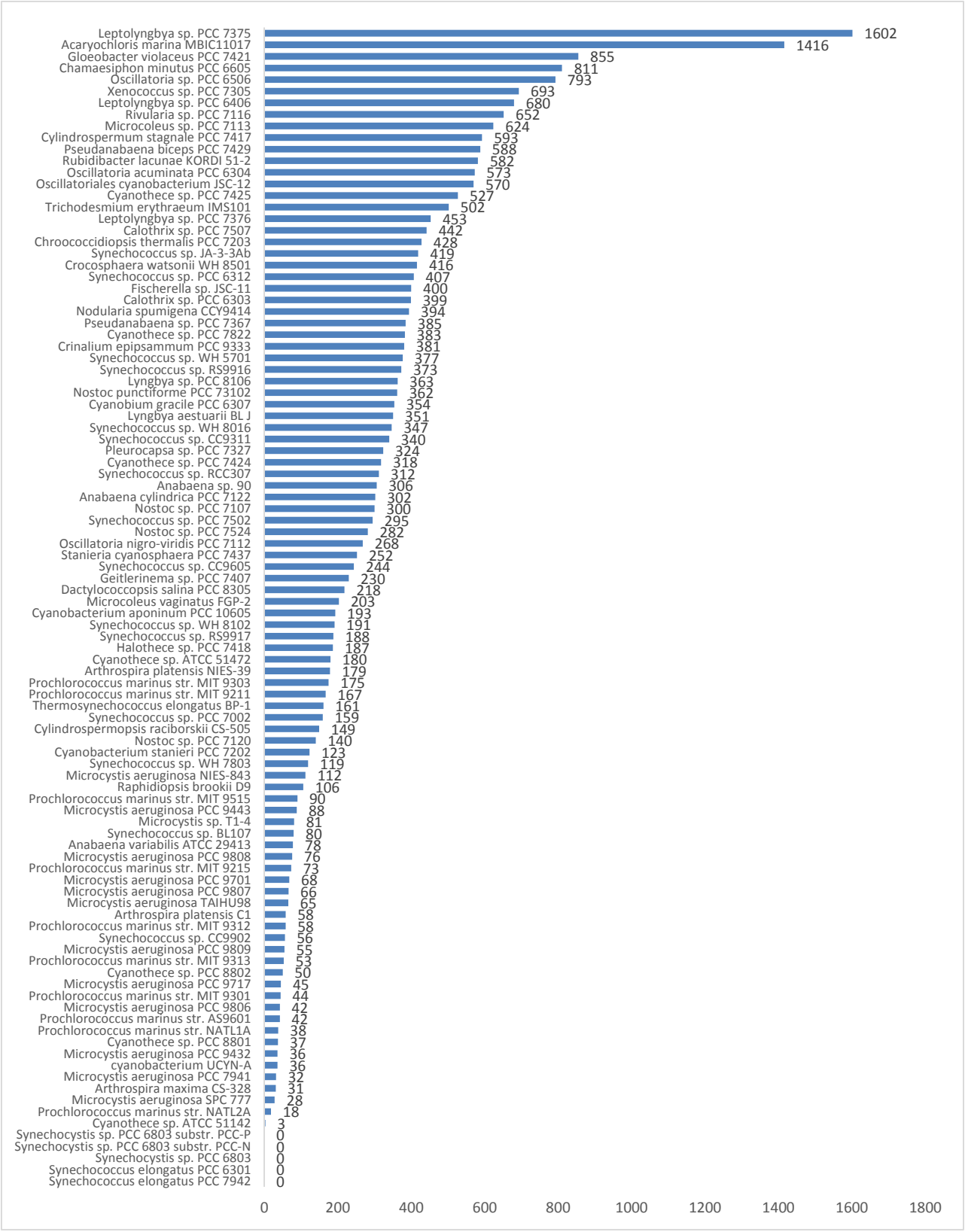


Figure 4.4 Distribution of specific COG based on each genomes

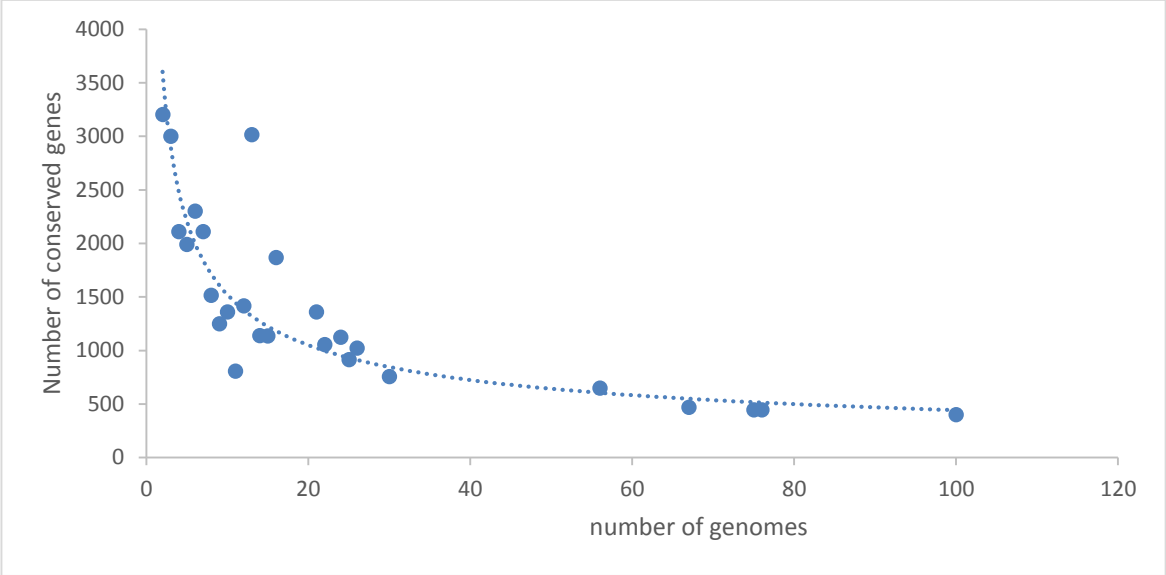


Figure 4.5 Core genome size (number of genes) vs number of genome

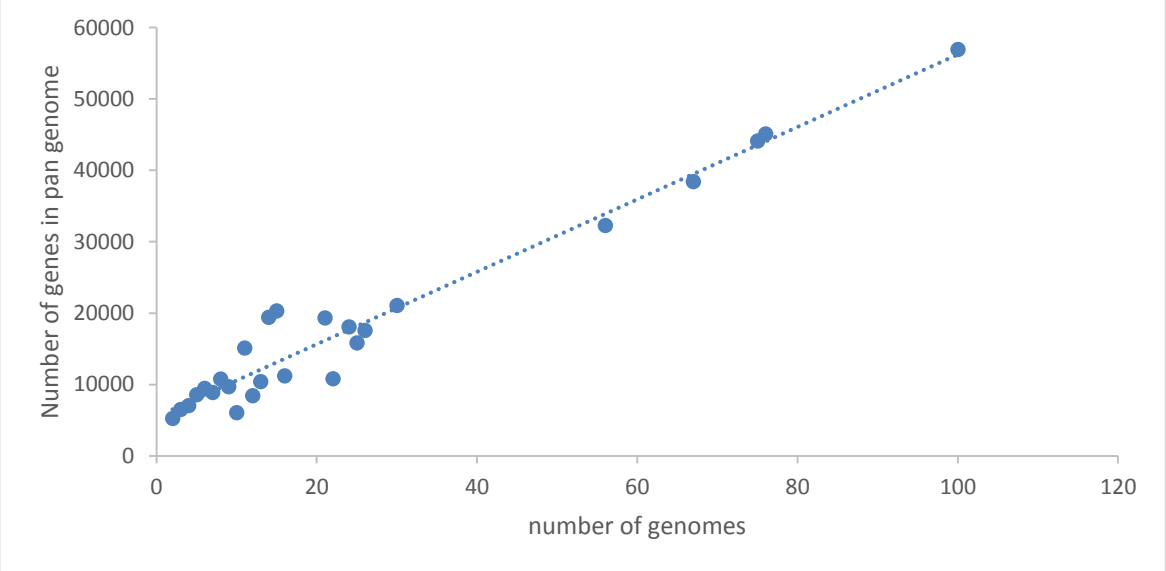


Figure 4.6 Pan genome size (number of genes) vs number of genome

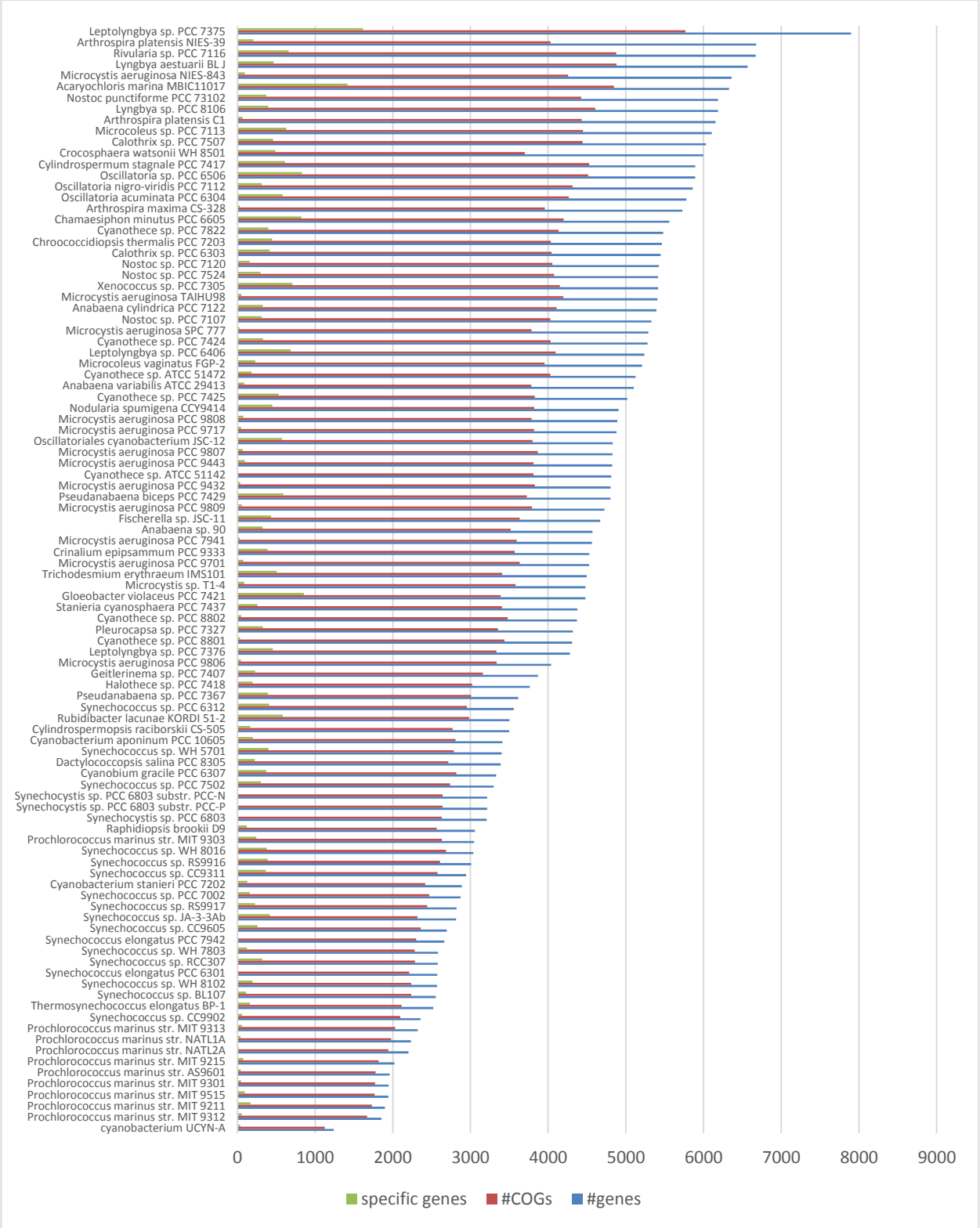


Figure 4.7 Distribution of COG vs genes vs specific genes based on each genomes

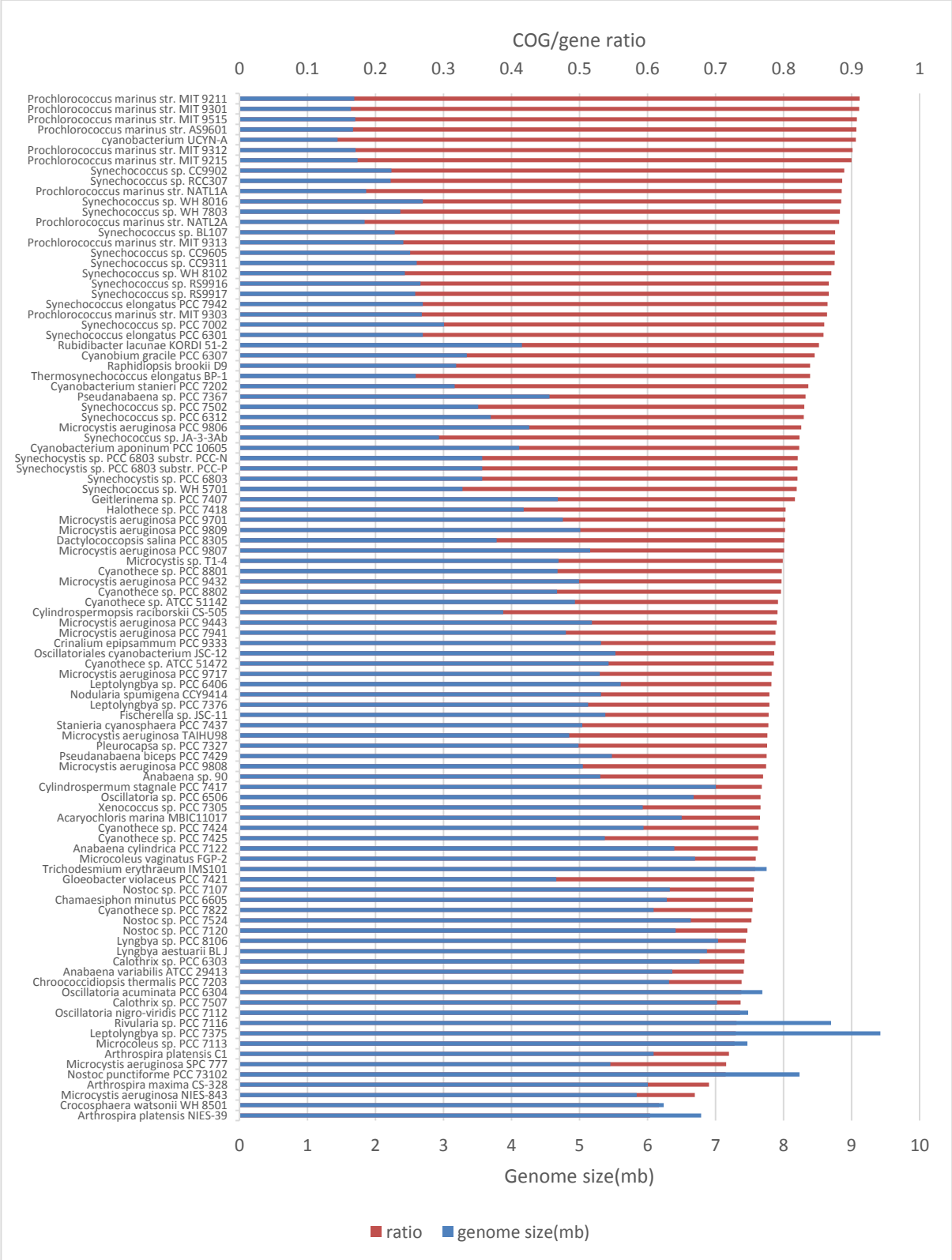


Figure 4.8 Distribution of COGs/genes ratio of each genome vs genome size

The comparison plot between the distribution of COGs, all genes and specific genes in each genome (figure 4.7) shows that the number of genes duplication (difference between gene and cog number) have a positive correlation with the genome size while specific genes doesn't have any correlation with COGs number or genes number. It could be hypothesized that the specific genes for each strains depend directly on their environment. The ratio which defined by number of COGs divided by the number of genes compared with the genome size shows negative correlation (figure 4.8) suggesting that the smaller the genome size are, the less gene duplication are. The smallest number of gene duplication was found in *Prochlorococcus marinus str. MIT 9211*, the largest number of gene duplication was found in *Arthrospira platensis NIES-39*. *Prochlorococcus marinus str. MIT 9211* was isolated from the equatorial Pacific at a depth of 83 meters and *Arthrospira platensis NIES-39* was isolated from salt lake in Republic of Chad, Central Africa. It could be hypothesized that *Prochlorococcus marinus str. MIT 9211* live in Deep Ocean where the ecological variables are stable thus adaptation is not necessary in this stable environment, results in small number of gene duplication. On the other hand, *Arthrospira platensis NIES-39,* which live in salt lake, might have to respond to a certain environment and results in high number of gene duplication. Nonetheless, these are only speculations which need to be further investigated.

4.3 Additional annotation

4.3.1 NCBI functional categories

Interestingly, only 2,558 COG (9.5%) were found to have significant similarity with protein from NCBI COG database with the criteria described in chapter 3.3. In addition, within 2,558 function assignment, 497 of them (19.4%) were "S" category which translated to "function unknown", 379 of them (14.8%) were "R" category which translated to "General function prediction only" figure 4.9 illustrate number of gene family annotated with single NCBI functional category. This result suggested that very little are known about the biology of cyanobacteria.

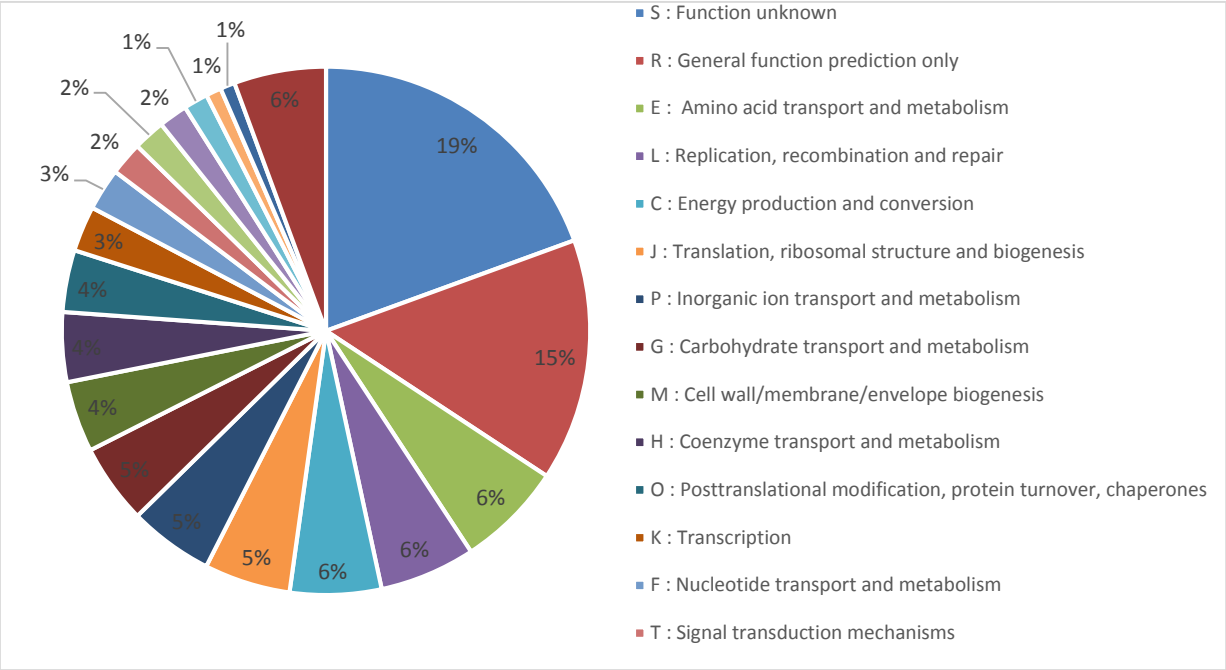


Figure 4.9 Percent of gene family annotated with NCBI functional category

4.3.2 KEGG pathway annotation

A total of 5,210 gene families (9.15%) were found to have significant similarity with KEGG pathway database based on the KAAS server criteria. Figure 4.10 shows top 10 most hit of gene family found in KEGG pathway database. The id in front of the pathway name is the accession number for pathway is KEGG system. The full list of the pathway that were found to have significant similarity with KEGG pathway database can be found in appendix 4.

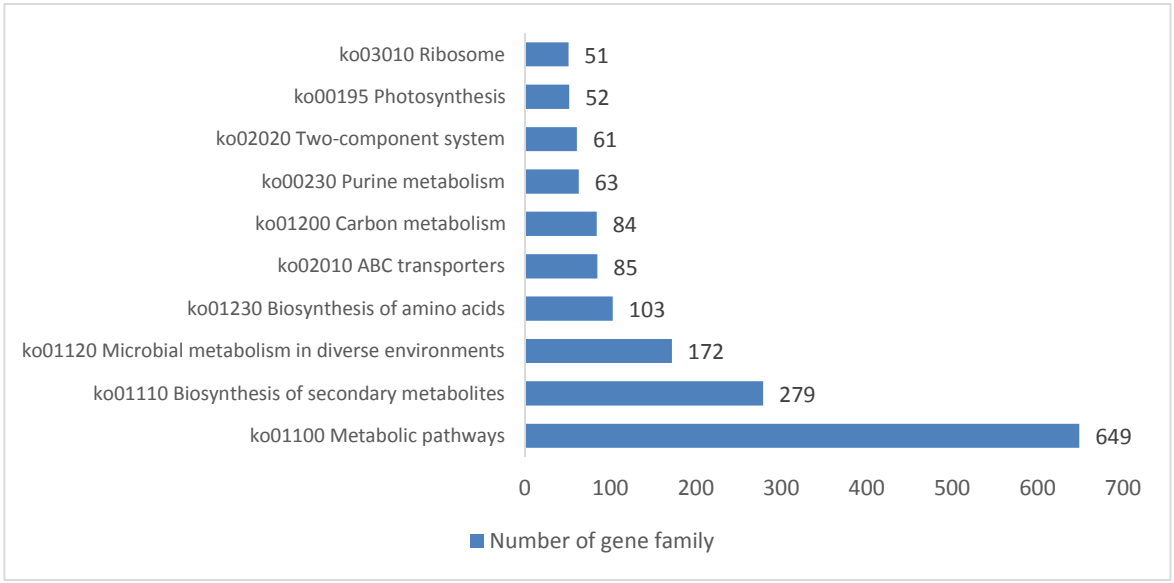


Figure 4.10 Top 10 most KEGG pathway hit for Cyanobacteria gene families.

4.3.3 Transporter annotation

In order to survive, cell of any single cellular or multicellular organism must acquire enough energy and nutrient, stabilize the appropriate cell condition within the cell such as pH, osmotic pressure, ion concentration etc. Therefore transporter is an important system for organisms to be able to survive in various environments. This study also in cooperated the transporter data from an in-house [63] transporter protein database comprise of 36,285 transporter proteins. A total of 1,833 gene (3.2%) family were found to have significant similarity with transporter database with the criteria described in chapter 3.3. Three main type of transporter were found; ATP-Dependent (51%) secondary transporter (30%) and ion channels (figure 4.11). The result suggests that cyanobacteria rely on energy dependent transporter in order to survive in their environment. This might be due to the fact that cyanobacteria have the ability to convert and use solar energy.

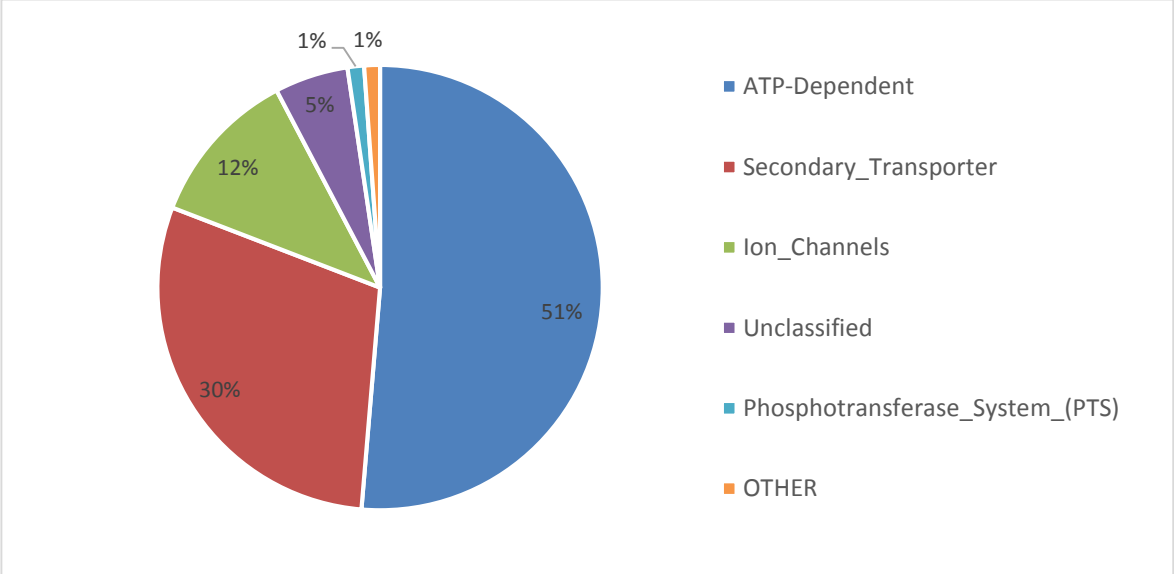


Figure 4.11 Percent of type of transporter found in gene families of cyanobacteria

4.3.4 Blast matrix

The BLAST matrix was calculated for every cyanobacterial clade. The high-resolution picture could be found in the supplementary data link. The result shows that the conservation between genomes is generally higher within closely related strains of cyanobacteria. For example, *Synechococcus sp*. and *Microcystis aeruginosa*. Figure 4.12 illustrates an example of the blast matrix result for clade 98, which consists of *Synechococcus* and *Prochlorococcus* genera. It could be observed that in the *Prochlorococcus* genus, some of the strains proteome are more similar to the *Synechococcus* than other *Prochlorococcus*. The author hypothesized that *Prochlorococcus* and *Synechococcus* adapt to their environment differently. The dissimilarity within *Prochlorococcus* genus may be the result of *Prochlorococcus* strains from different places and the high similarity observed between *Prochlorococcus* and *Synechococcus* was the indication that these strains came from the similar environment. This would imply that *Synechococcus* doesn't change much in different environment. The author hypothesized that the gene content of *Synechococcus* is sufficient to support survival in diverse environment with minimal adaptation.

Similarity of cyanobacteria's proteome across all cyanobacteria in this study ranges from 4.8-99.9%. The internal homology (red squares) ranges from 1.2–17.9%. In comparison, a study performed on genomes from the Vibrionaceae family showed that different strains of *Vibrio cholerae* share between 70–80% proteins while the similarity to organisms outside the species ranged from 30–45% [64]. From that same study, the internal homology (red squares) ranges from 1.3–5.3%. Another study analyzed the similarity between Enterobacteriaceae genomes, and found a 76–98.8% similarity between 7 genomes of *Escherichia coli* [65]. The same study showed an internal homology of approximately 0.3–3% for the 7 *Escherichia coli*. The result suggested that cyanobacteria in this study are very diverse.

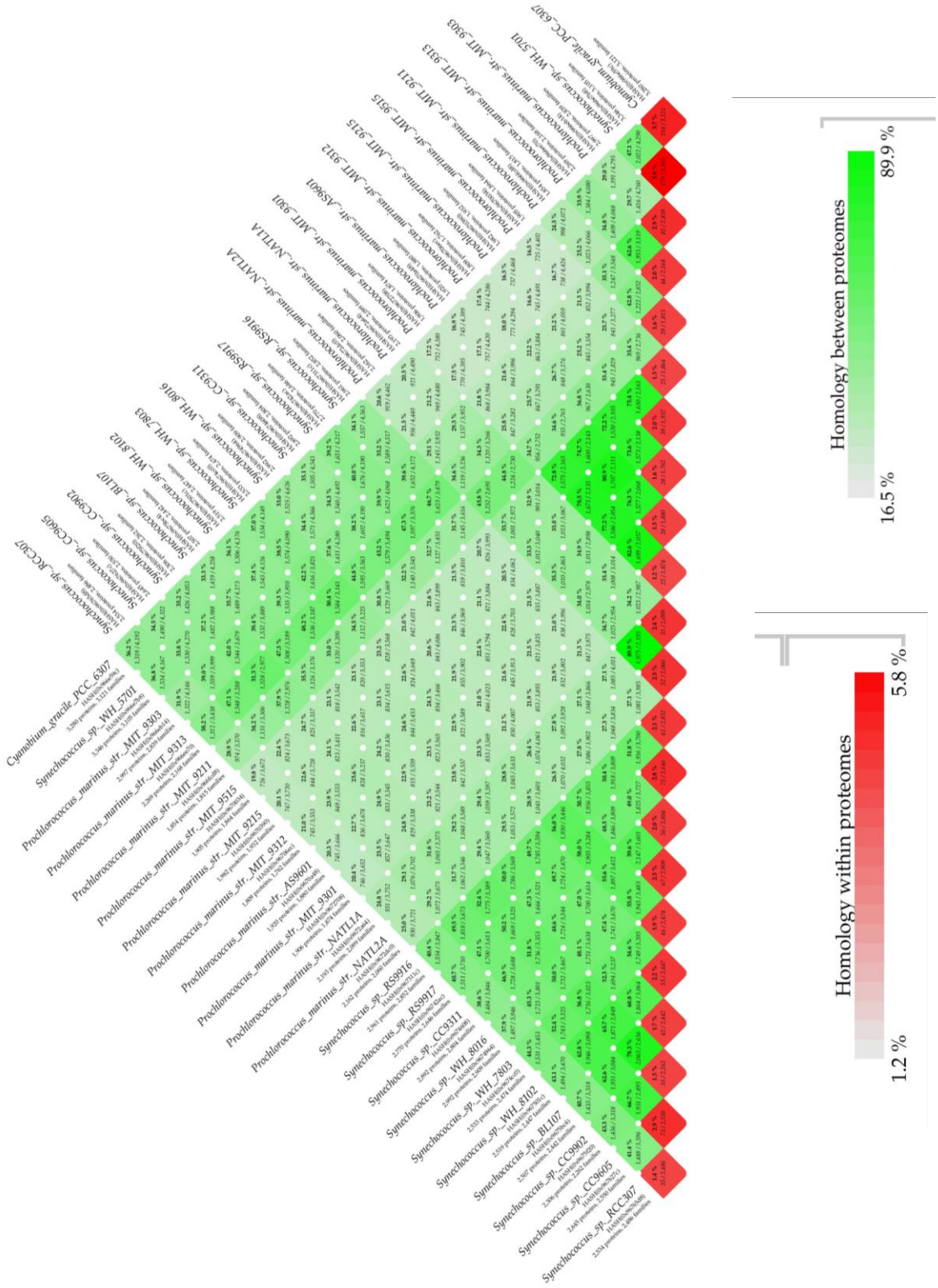


Figure 4.12 Blast matrix from core gene tree of clade 98

4.4 Phylogenetic tree reconstruction

4.4.1 16s rRNA

The strategy of using 16s rRNA genes as the information for reconstructing bacterial phylogenetic tree has remained common for very long time. This is because it is found in nearly all bacteria. It also held a vital function for their survival and thus it is not changed a lot over time, which results in a good molecule for time measurement [33]. In addition, it held relatively enough information (about 1,500 base pairs) for phylogeny analysis, which could differentiate most of the genus and species level. However for the closely related strains, which were recently diverged, their differences might lie elsewhere and their 16s rRNA gene are still identical. Therefore the use of 16s rRNA gene alone is insufficient to completely resolved the tree topology [66, 67]. In figure 4.13 it can be clearly observed that the tree can't be resolved between closely related strains e.g. Synechococcus and Prochlorococcus. Another observation that can be made in figure 4.12 is that the tree separates some of the organisms that are from the same genus e.g. Cyanothece and Leptolyngbya. This raised many questions about cyanobacteria classification. If the assumption that the 16s rRNA phylogenetic tree was correct, then this raised the question whether the separated group of cyanobacteria should be classify as the same genus or not. On the other hand if the assumption that the 16s rRNA sequences of these particular group lead to a questionable tree topology, then this raised the question whether the strategy to use only 16s rRNA sequence to reconstruct the phylogenetic tree is appropriate for the evolutionary study of cyanobacteria or not.

4.4.2 Core genes

To address the lack of power to differentiate closely related strains of the 16s rRNA genes, this study uses cluster of orthologous gene family that is shared among all cyanobacterial strains. However, this alone might not give a good information to reflect the organismal phylogeny owing to the horizontal gene transfer (HGT) or known in other term as lateral gene transfer (LGT) event, which cause complicated gene history and disagreement with the rRNA gene tree [68]. To cope with this complication, this study used the universally single copy core genes in each strains as information for the organismal phylogeny. These genes are suggested to be robust to HGT event even in the organisms with high HGT rates [35, 69, 70]. The described criteria for selecting genes resulted in 12 core genes as illustrated in table 3.1. Figure 4.6 clearly shown that when in cooperating information from 12 genes, the topology of the closely related organism was totally resolved or distinguishable. In addition, genus Cyanothece and Leptolyngbya are also better grouped in the core gene tree compared to 16s rRNA tree.

4.4.3 Phyletic pattern

Inferring phylogenetic tree using phyletic pattern doesn't reflex much about the history as this method looks at the present and absent of the gene content and compare their pattern similarity rather than using the information contain within the evolving genomic sequence. However it is an interesting method for grouping organisms that shared similar life style together and this information might be useful for the extraction of the genes that are responsible for a unique feature of closely related organisms.

Comparing the tree from all approaches, core genes tree and phyletic tree show less similarity percentage to each other than to 16s rRNA tree (table 4.2). However considering each clade, most of the closely related cyanobacteria strains were grouped together in the same clade except for the 16s rRNA tree which separated Cyanothece and *Leptolyngbya sp. PCC 6406* from *Leptolyngbya sp. PCC 7375* while trees from others approach agree. The full comparison data could be found in appendix 5. The disagreement between evolutionary trees (core gene tree) and gene content tree (phyletic tree) might be owing to the high frequency of HGT event, however this is only a speculation, which need further investigation.

Table 4.2 Percent dissimilarity matrix of all against all tree topology comparison among tree building methods

	16s	core	Phyletic
16s	0	43	42
core	43	0	32
phyletic	42	32	0

As for the recommendation on phylogenetic tree construction, this study recommended that multiple core genes family that are universally single copy should be considered as primary source of information. Because it relies on multiple genes, even if some of them are horizontally transferred, which is unlikely, the remaining genes should give the right information. On the other hand, 16s rRNA genes approach, which rely on only single gene, will result in catastrophic result if something wrong with that gene. After all, even 16s rRNA gene itself can be subjected to horizontal gene transfer. Though, the phyletic pattern utilized the whole genome information, wrong gene calling, and incompleteness of genome would cause a direct and adverse impact on tree topology therefore it is not recommended for incomplete genome.

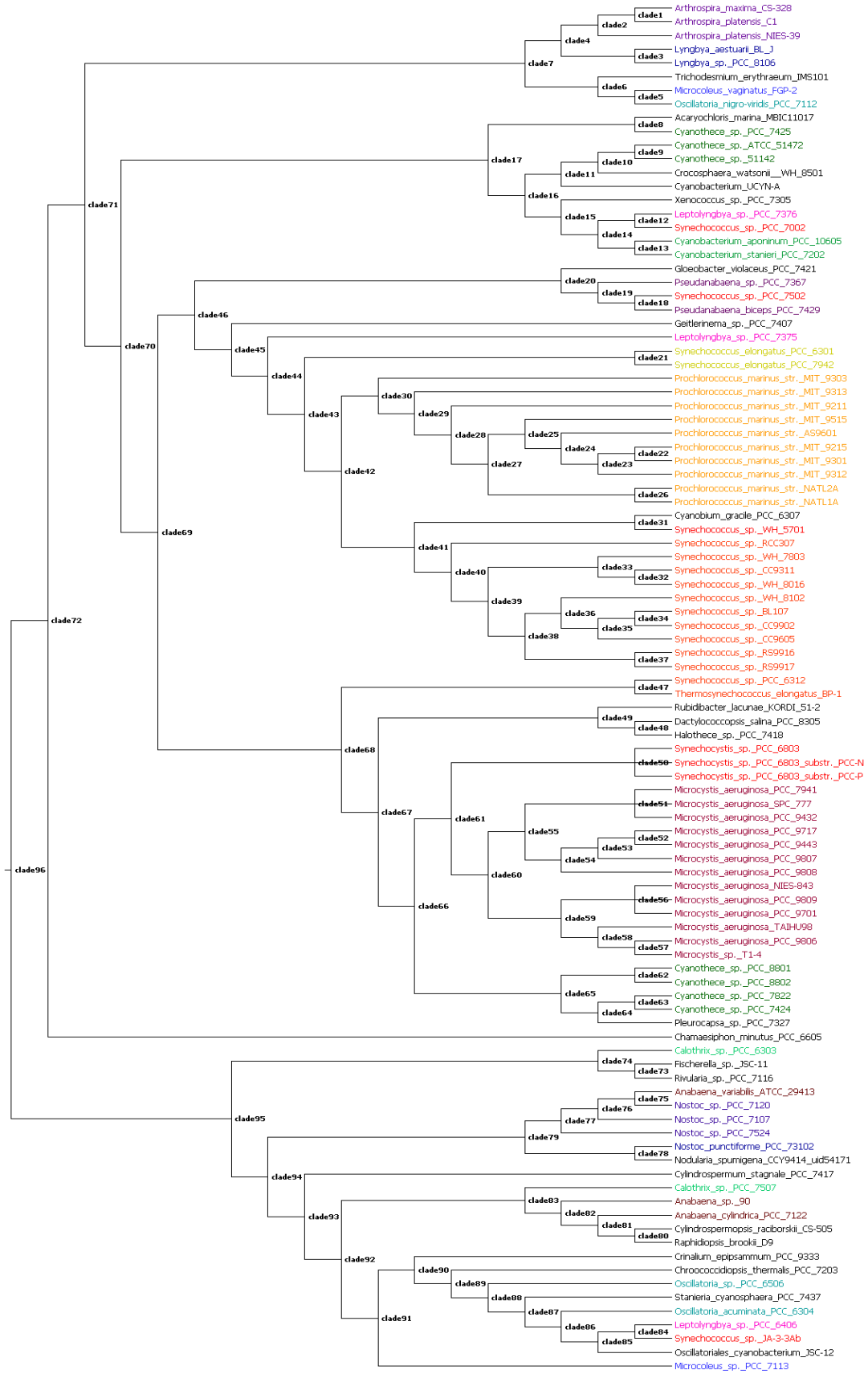


Figure 4.13 Phylogenetic tree reconstructed form 16s rRNA data. Color label indicates that the organisms came from the same genus.

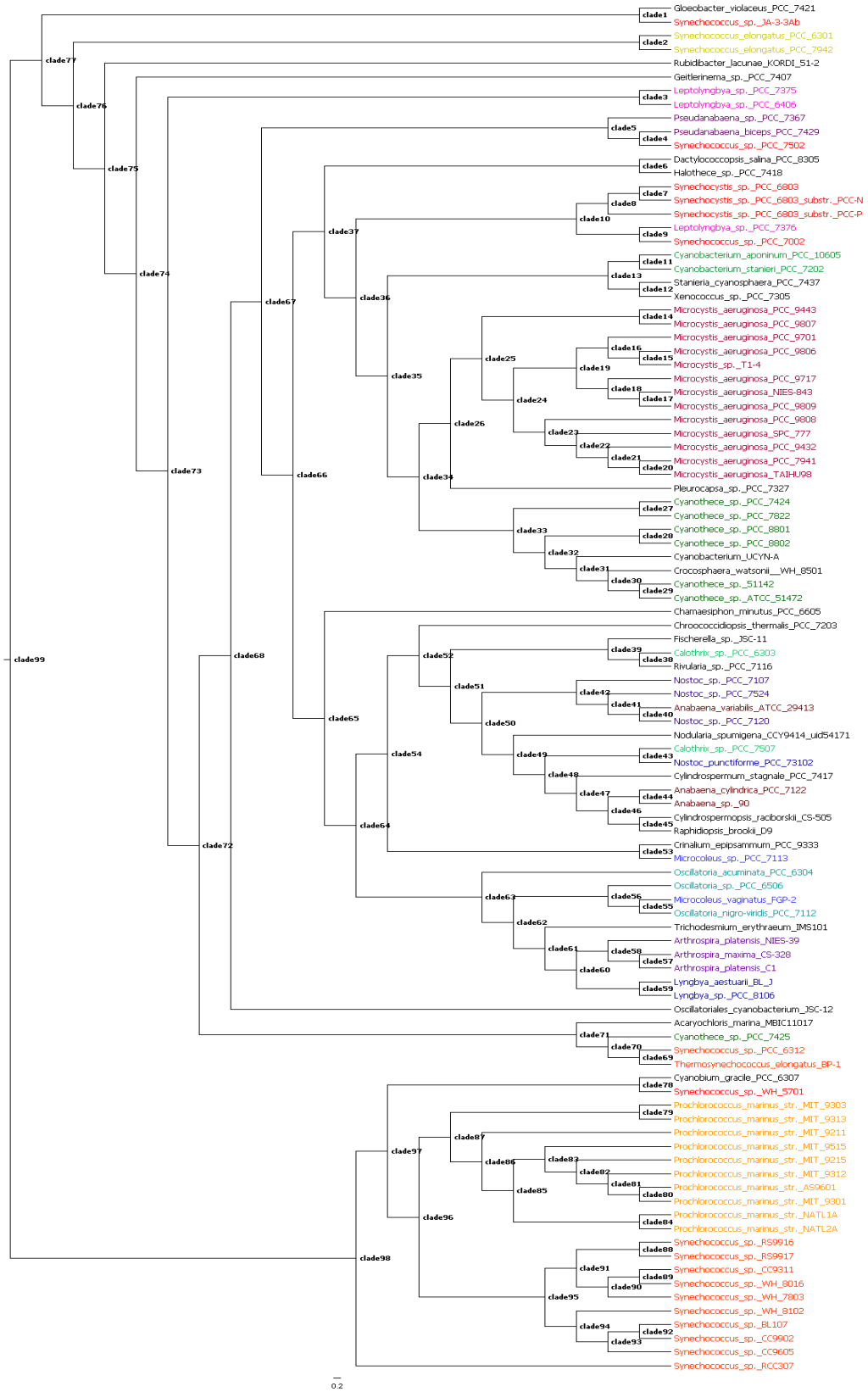


Figure 4.14 Phylogenetic tree reconstructed form core protein family data. Color label indicates that the organisms came from the same genus.

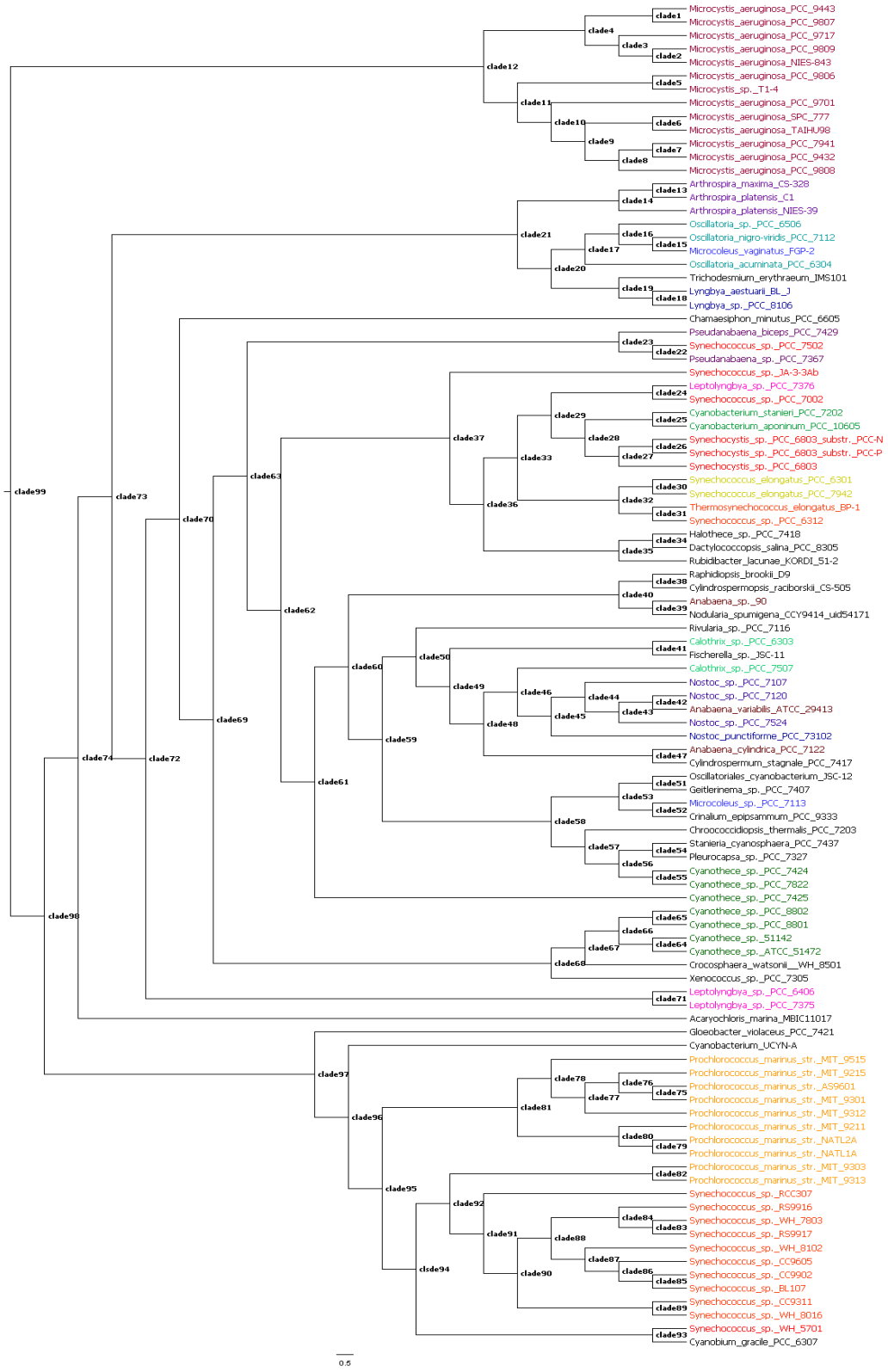


Figure 4.15 Phylogenetic tree reconstructed form phyletic pattern data. Color label indicates that the organisms came from the same genus.

4.5 Cluster of orthologous group and phylogenetic tree integration

The aim of this study is to search for the clade and species-specific genes that may contribute to the special characteristics of that particular cyanobacterial clade or species. In addition to the clade/species specific genes, which defined by the genes that only present in that particular clade or species, this study also identified the core and pan genome of each clade. The core genes were defined by the genes that are shared among that particular clade regardless of the other clade's information, the pan genes were defined as the genes that found in at least one species in that particular clade. Moreover, this study also annotated additional information that might be useful for further study namely KEGG pathway, NCBI COG categories, and transporter (Figure 4.16, 4.17). Additionally, the blast matrix for each clade is available.

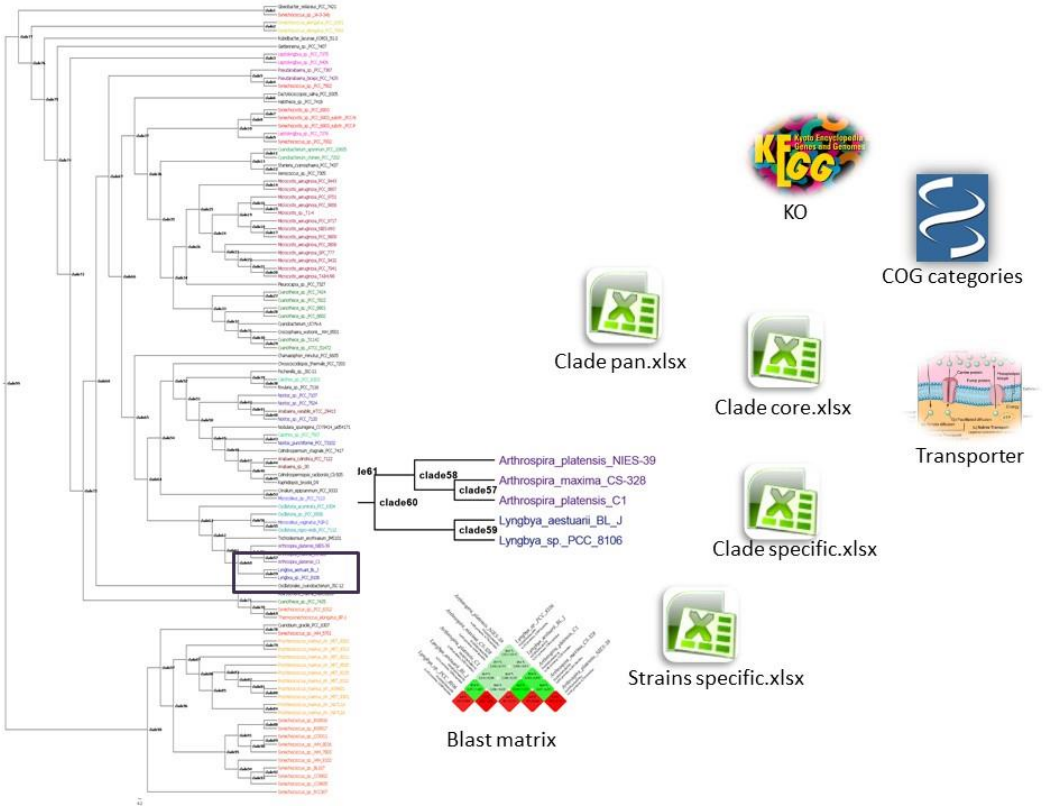
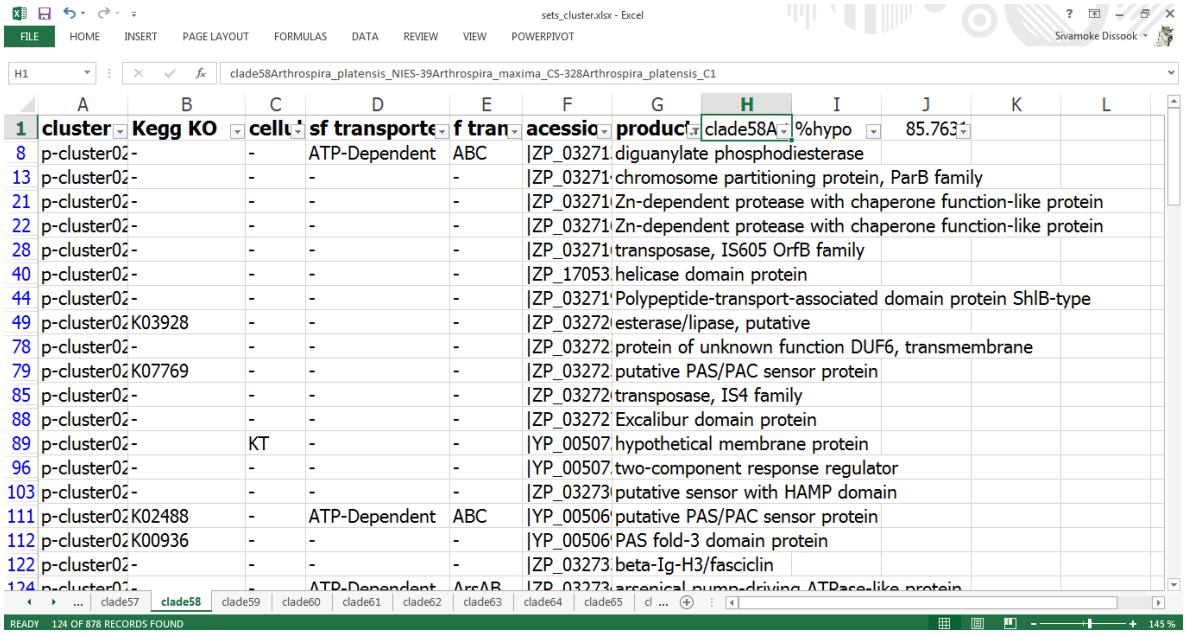


Figure 4.16 Overall output from integration of COG and phylogenetic tree

Result of clade classification using the 16s rRNA, core genes, and phyletic pattern are shown in figure 4.13, .4.14, 4.15, respectively. The clade with the largest number of genes is the core clade, which would report the same set of genes in the specific genes and core genome for every phylogenetic construction methods. The core genome of this study set consists of

401 genes where 219 gene families (55%) were annotated with the NCBI COG categories. The result shows that majority of the core genes are the house keeping genes (figure 4.18).



	A	B	C	D	E	F	G	H	I	J	K	L
1	cluster	Kegg KO	cellu	sf transporte	f tran	acessio	product	clade58A	%hypo	85.763		
8	p-cluster02	-	-	ATP-Dependent	ABC	|ZP_03271	diguanylate phosphodiesterase					
13	p-cluster02	-	-	-	-	|ZP_03271	chromosome partitioning protein, ParB family					
21	p-cluster02	-	-	-	-	|ZP_03271	Zn-dependent protease with chaperone function-like protein					
22	p-cluster02	-	-	-	-	|ZP_03271	Zn-dependent protease with chaperone function-like protein					
28	p-cluster02	-	-	-	-	|ZP_03271	transposase, IS605 OrfB family					
40	p-cluster02	-	-	-	-	|ZP_17053	helicase domain protein					
44	p-cluster02	-	-	-	-	|ZP_03271	Polypeptide-transport-associated domain protein ShlB-type					
49	p-cluster02	K03928	-	-	-	|ZP_03272	esterase/lipase, putative					
78	p-cluster02	-	-	-	-	|ZP_03272	protein of unknown function DUF6, transmembrane					
79	p-cluster02	K07769	-	-	-	|ZP_03272	putative PAS/PAC sensor protein					
85	p-cluster02	-	-	-	-	|ZP_03272	transposase, IS4 family					
88	p-cluster02	-	-	-	-	|ZP_03272	Excalibur domain protein					
89	p-cluster02	-	KT	-	-	|YP_00507	hypothetical membrane protein					
96	p-cluster02	-	-	-	-	|YP_00507	two-component response regulator					
103	p-cluster02	-	-	-	-	|ZP_03273	putative sensor with HAMP domain					
111	p-cluster02	K02488	-	ATP-Dependent	ABC	|YP_00506	putative PAS/PAC sensor protein					
112	p-cluster02	K00936	-	-	-	|YP_00506	PAS fold-3 domain protein					
122	p-cluster02	-	-	-	-	|ZP_03273	beta-Ig-H3/fasciclin					

Figure 4.17 General output format of the output file in this study

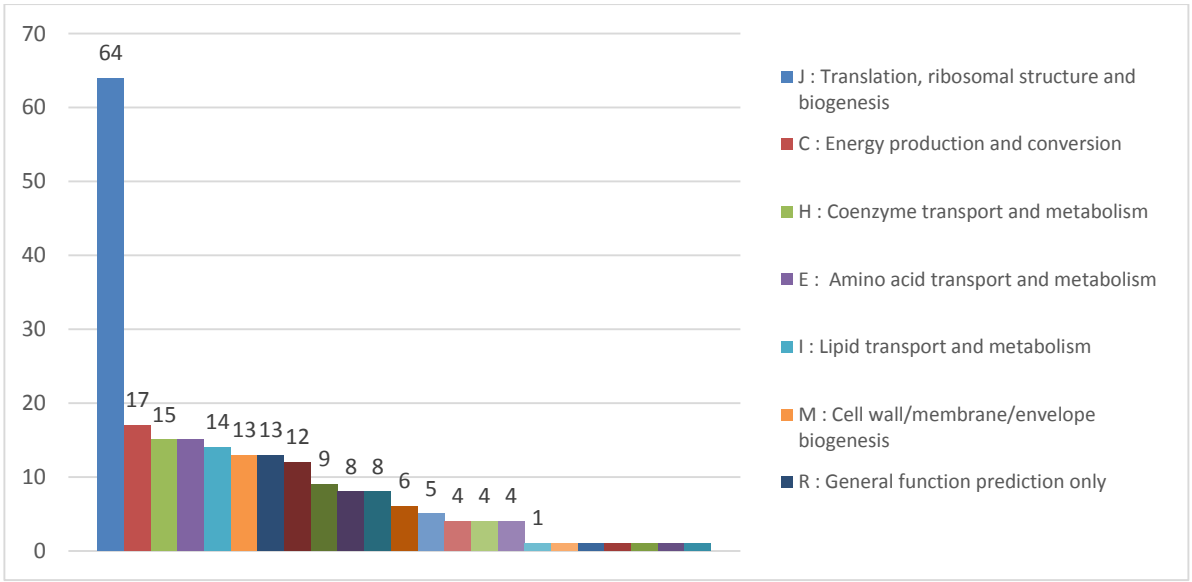


Figure 4.18 Number of core gene family annotated with NCBI functional category

About 83.2 % of core genome were annotated with KEGG pathway mostly involved with housekeeping operation where the top 10 pathways from in core genome of cyanobacteria was illustrated in figure 4.19. The full list of core pathways of Cyanobacteria could be found in appendix 6.

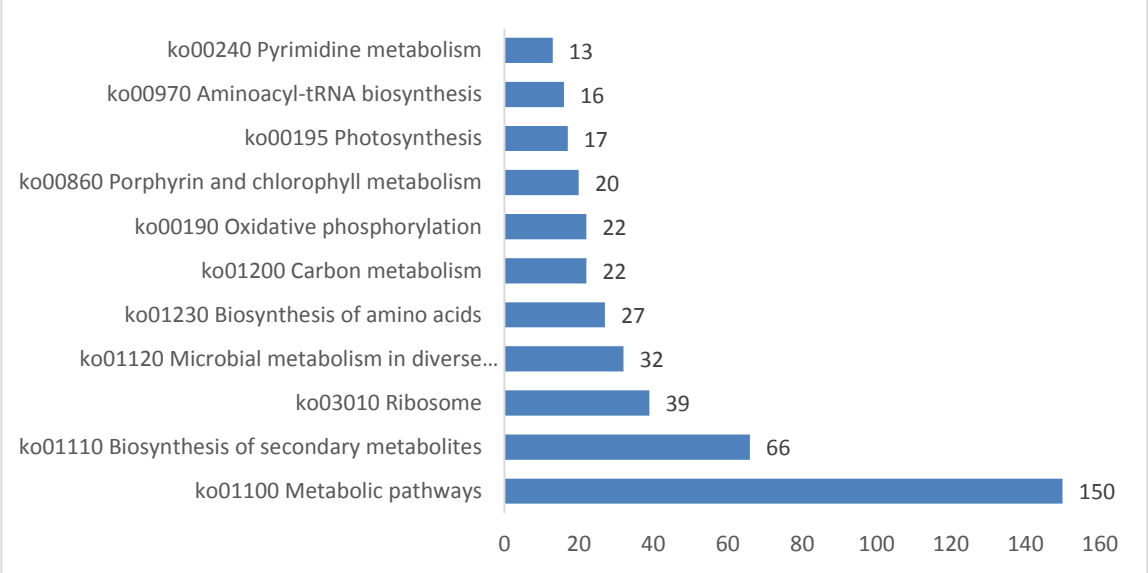


Figure 4.19 Top ten most hit core gene family in Cyanobacteria core genome

Interestingly, by looking at the shared metabolic pathway one would expected to see that the important metabolic pathway used by all aerobic organisms to generate energy such as the TCA cycle should be shared among cyanobacteria. However the study found that TCA cycle were not shared by all cyanobacteria in this study (Figure 4.20). Reviews show that even a crucial metabolic pathway like TCA cycle do evolved over time [71]. It could be hypothesized that cyanobacteria used in this study came from a diverse origin that their TCA cycle evolved so far apart that the criteria used to identify orthologs in this study couldn't detect their orthologous relationship. When further investigation was conducted by looking at the core genes from a group of closely related cyanobacteria from clade 86 in the core genes tree which consist of *Prochlorococcus marinus*, this study found that this group of cyanobacteria's TCA cycle were shared (figure 4.21) thereby supporting the hypothesis. Nonetheless it is a very interesting subject to make a further research on the evolution of TCA cycle in cyanobacteria.

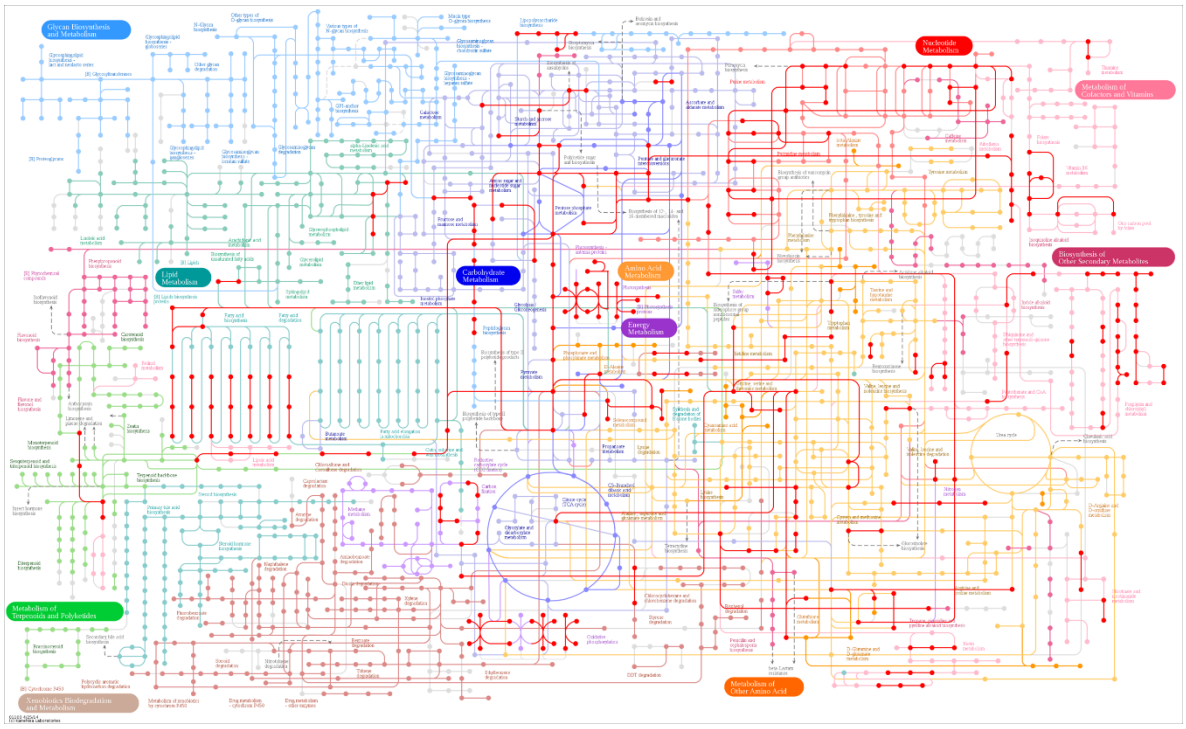


Figure 4.20 Core metabolic pathway of all cyanobacteria used in this study red lines indicated the mapped pathway

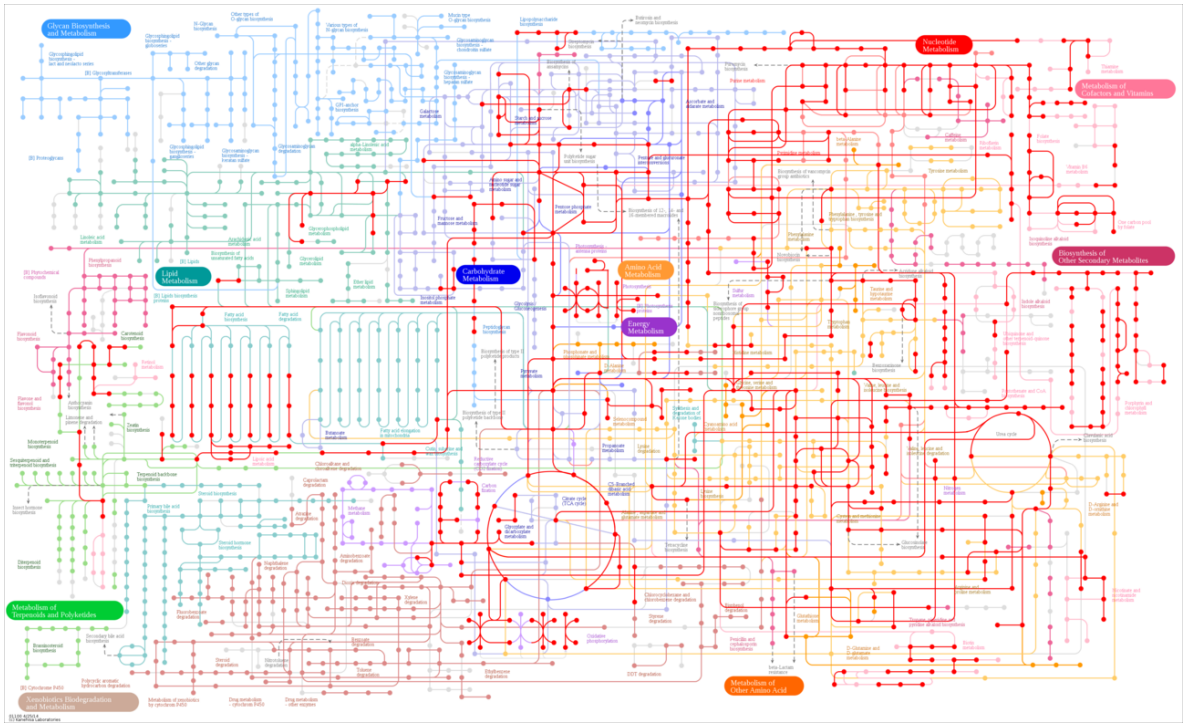


Figure 4.21 Core metabolic pathway of all Cyanobacteria in clade 86 from core genes tree red lines indicated the mapped pathway

As mentioned earlier, the investigation of TCA cycle of cyanobacteria is another interesting topic. This research makes a further study on cyanobacteria's TCA cycle using KEGG pathway database. Preliminary results show that out of 54 enzymes found in KEGG TCA cycle, only 29 enzyme were found in the pan genome of cyanobacteria in this study (Figure 4.22). Moreover, all of the TCA components were identified for every clade in the core gene tree. Comparisons of these components show that cyanobacteria have variety of TCA cycle types. The author hypothesized that the differences in TCA cycle of cyanobacteria resulted from the adaptation to their environment. This is a very interesting topic to make a further investigation as this could result in better understanding in making a better bioengineered organism. All of the TCA component identification result could be found in appendix 7.

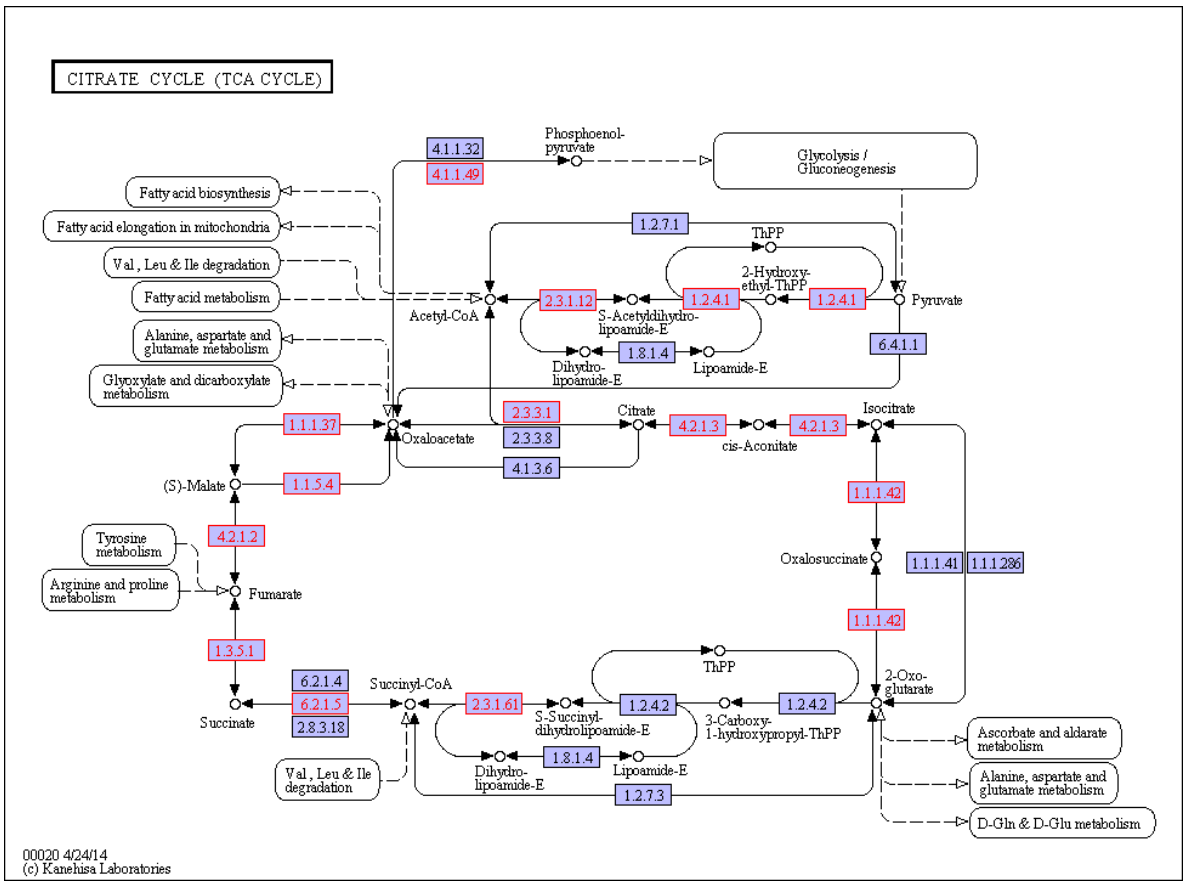


Figure 4.22 TCA component found in cyanobacteria pan genome (marked in red)

When combining the classified clade of all trees with the habitat information clade 42 of the 16s rRNA tree, clade 98 of the core gene tree and clade 95 of the phyletic tree do agree on grouping all the *Prochlorococcus marinus*, majority of *Synechococcus sp.* and *Cyanobium gracile PCC 6307* together. All of the species in these clade were from marine environment except for *Cyanobium gracile PCC 6307*, which came from fresh water environment. Further analysis found that the core protein family of these clades consist of 1,026 genes, the pan genome consist of 10,225 genes and the clade specific genes consist of 46 genes. They are 3 specific genes that could be mapped to KEGG pathway, one of which is in the pentose phosphate pathway annotated with E.C. 2.7.1.15 (ribokinase). The ribokinase catalyst a reversible reaction between D-Ribose-5 phosphate to D-Ribose (2-Deoxy-D-ribose 5-phosphate + ADP <=> Deoxyribose + ATP). Another one found in photosynthesis and pathway annotated with photosystem II PsbJ protein. On the transporter front one transporter was found to be specific to this clade, which annotated with ABC.X1.S: putative ABC transport system substrate-binding protein. Further conclusion is difficult to drawn out as the 73.9% of the clade specific genes were annotated with hypothetical protein. Another group of organism that all tree agrees on were the clade 25, 60 and 12 of the 16s rRNA tree core gene tree and phyletic tree respectively. These clade comprise of the organisms from Microcystis genus, which come from fresh water environment. Further analysis found that the core protein family of these clade consist of 2,466 genes, the pan genome consist of 8,689 genes and 37 genes clade specific genes. dTDP-glucose 4,6-dehydratase (E.C. 4.2.1.46) was found to be specific for this clade, the KEGG pathway mapping found that E.C. 4.2.1.46 were used in streptomycin biosynthesis pathway, polyketide sugar unit biosynthesis pathway and biosynthesis of vancomycin group antibiotics pathway. Another gene that is found to be specific to this clade is the E.C. 1.1.3.6: cholesterol oxidase in the steroid degradation pathway. Further conclusion is hard to come by as the 72.2% of the clade specific genes were annotated with hypothetical protein. Looking at smaller clade, for example clade 58 of the core gene tree, which consists of only the organism from *Arthrospira* genus. *Arthrospira* naturally live in alkaline habitat and this study hypothesized that the unique genes that are found among this clade should somehow contribute to their survival in their habitat. This study found that among the specific genes in this clade, there are three gene families that corresponding to cation efflux system protein NrsA. This result is the evidence that the integration of COG and phylogenetic data could yield basic information to make a better understanding of the organism's characteristics.

In other clades, the habitat or ecological niche information usually mixed together. This result raised a very important question whether the organisms from particular ecological niche can only live in that particular ecological environment or not. If the answer is yes, the niche results of the organisms grouping based on their evolution should have a higher degree of correlation with the ecological niche data. However this is only a speculation and they are many variables to consider such as the correctness of the phylogenetic tree, the bio sample,

which used to sequence these organisms, the diversity of the ecological environment. The ecological niche is a very broad term and most of the time the parameter in which a certain strains was isolated are limited. It would be very useful in the niche classification if the data such as the temperature, photon density, height, depth, time of isolation, and other organisms in the area were available. Additionally, metagenome data of the isolate site would be interesting because it may provide a variety of potential genes that might be in cooperated into the cyanobacteria genome during their course of evolution. The habitat information were describe in supplementary data “organisms data” in appendix 7. All of the data could be found in appendix 7.

CHAPTER 5 CONCLUSIONS

5.1 Conclusions

In this research, a comparative genomic method was used to characterize gene content variation across 100 cyanobacteria genomes in order to search for specific genes that might contribute to special characteristics of each group or strains of cyanobacteria. 63 complete cyanobacterial genomes and 37 in-progress cyanobacterial genomes were obtained from National Center for Biotechnology Information genomics database. Cluster of Orthologous Groups (COGs) of cyanobacteria were created using the in-house method. The phylogenetic trees were constructed and used to identify the biologically meaningful clades of cyanobacteria and a set of species. Clade specific genes were also identified along with their pathway and transporter annotation information.

The results from COGs construction and annotation suggested that cyanobacteria are very diverse group of organism that contained within their genome a lot of unique/species-specific genes. There are much to discover about the biology of cyanobacteria as about 70% of their genes are still annotated as hypothetical. While to date the debates on how to build the organismal phylogeny is still ongoing, the different approaches of organismal phylogeny construction suggested that for closely related organisms, more information should be considered as a raw data for building a phylogenetic tree, at lease for the purpose of building a totally resolved tree that are able to separate all genomes.

This study identified a total of 19,573 clade specific genes across all clades specified by phylogenetic tree and a total of 28,931 species specific genes across 100 cyanobacterial genomes along with associated KEGG pathway, NCBI COG annotation and transporter annotation which might be useful information for researchers who are interested in cyanobacteria.

5.2 Recommendation

According to high amount of hypothetical protein, it is difficult to make a decisive conclusion regarding their specific genes. Thus an advanced annotation pipeline could make this study more informative. Another suggestion is that a visualization tools for these huge amount of data would help to make a quicker and better understanding of the data.

REFERENCES

1. Chorus, I. and Bartram, J., 1999, **Toxic Cyanobacteria in Water: A Guide to Their Public Health Consequences, Monitoring and Management**, Spon Press, Pages.

2. Schopf, J. W., and Kudryavtsev, A. B., 2012, "Biogenicity of Earth's earliest fossils: a resolution of the controversy", **Gondwana Research**, Vol. 223, pp. 761-771.

3. Gupta, V., Ratha, S. K., Sood, A., Chaudhary, V., and Prasanna, R., 2013, "New insights into the biodiversity and applications of cyanobacteria blue-green algae—Prospects and challenges", **Algal Research**, Vol. 22, pp. 79-97.

4. Asada, Y., Miyake, M., and Miyake, J., 1998, "Production of bioplastics and hydrogen gas by photosynthetic microorganisms", **Chinese Journal of Oceanology and Limnology**, Vol. 161, pp. 91-104.

5. Abed, R. M. M., Dobretsov, S., and Sudesh, K., 2009, "Applications of cyanobacteria in biotechnology", **Journal of applied microbiology**, Vol. 1061, pp. 1-12.

6. Whitton, B.A. and Potts, M., 2012, "Introduction to the Cyanobacteria", In **Ecology of Cyanobacteria II**, Springer, pp 1-13.

7. Douglas, S. E., 1998, "Plastid evolution: origins, diversity, trends", **Current opinion in genetics and development**, Vol. 86, pp. 655-661.

8. Krings, M., Hass, H., Kerp, H., Taylor, T. N., Agerer, R., and Dotzler, N., 2009, "Endophytic cyanobacteria in a 400-million-yr-old land plant: A scenario for the origin of a symbiosis?", **Review of Palaeobotany and Palynology**, Vol. 1531, pp. 62-69.

9. Rai, A.N., Bergman, B., and Rasmussen, U., 2002, **Cyanobacteria in Symbiosis**, Springer, Pages.

10. de La Torre, R., Sancho, L.G., Horneck, G., Ríos, A.d.l., Wierzchos, J., Olsson-Francis, K., Cockell, C.S., Rettberg, P., Berger, T., and de Vera, J.-P.P., 2010, "Survival of Lichens and Bacteria Exposed to Outer Space Conditions–Results of the Lithopanspermia Experiments", **Icarus**, Vol. 208, No. 2, pp. 735-748.

11. Fay, P., 1965, "Heterotrophy and Nitrogen Fixation in Chlorogloea Fritschii", **Journal of general microbiology**, Vol. 39, No. 1, pp. 11-20.

12. Munn, C., 2003, **Marine Microbiology: Ecology & Applications**, Garland Science

13. Partensky, F., Blanchot, J., and Vaulot, D., 1999, "Differential Distribution and Ecology of Prochlorococcus and Synechococcus in Oceanic Waters: A Review", **BULLETIN-INSTITUT OCEANOGRAPHIQUE MONACO-NUMERO SPECIAL-**, Vol., No., pp. 457-476.

14. Oberholster, P., Botha, A., and Grobbelaar, J., 2004, "Microcystis Aeruginosa: Source of Toxic Microcystins in Drinking Water", **African Journal of Biotechnology**, Vol. 3, No. 3.

15. Reddy, K., Haskell, J.B., Sherman, D., and Sherman, L., 1993, "Unicellular, Aerobic Nitrogen-Fixing Cyanobacteria of the Genus Cyanothece", **Journal of bacteriology**, Vol. 175, No. 5, pp. 1284-1292.

16. Feng, D.-l. and Wu, Z.-c., 2006, "Culture of Spirulina Platensis in Human Urine for Biomass Production and O2 Evolution", **Journal of Zhejiang University SCIENCE B**, Vol. 7, No. 1, pp. 34-37.

17. O'Brien, S.J., Menotti-Raymond, M., Murphy, W.J., Nash, W.G., Wienberg, J., Stanyon, R., Copeland, N.G., Jenkins, N.A., Womack, J.E., and Graves, J.A.M., 1999, "The Promise of Comparative Genomics in Mammals", **Science**, Vol. 286, No. 5439, pp. 458-481.

18. Hardison, R.C., 2003, "Comparative Genomics", **PLoS biology**, Vol. 1, No. 2, p. e58.

19. Fitch, W.M., 1970, "Distinguishing Homologous from Analogous Proteins", **Systematic Biology**, Vol. 19, No. 2, pp. 99-113.

20. Koonin, E.V., 2005, "Orthologs, Paralogs, and Evolutionary Genomics 1", **Annu. Rev. Genet.**, Vol. 39, No., pp. 309-338.

21. Sonnhammer, E.L. and Koonin, E.V., 2002, "Orthology, Paralogy and Proposed Classification for Paralog Subtypes", **TRENDS in Genetics**, Vol. 18, No. 12, pp. 619-620.

22. Kuzniar, A., van Ham, R.C., Pongor, S., and Leunissen, J.A., 2008, "The Quest for Orthologs: Finding the Corresponding Gene across Genomes", **TRENDS in Genetics**, Vol. 24, No. 11, pp. 539-551.

23. Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., and Durkin, A.S., 2005, "Genome Analysis of Multiple Pathogenic Isolates of Streptococcus Agalactiae: Implications for the Microbial "Pan-Genome"", **Proceedings of the National Academy of Sciences of the United States of America**, Vol. 102, No. 39, pp. 13950-13955.

24. Mulkidjanian, A.Y., Koonin, E.V., Makarova, K.S., Mekhedov, S.L., Sorokin, A., Wolf, Y.I., Dufresne, A., Partensky, F., Burd, H., and Kaznadzey, D., 2006, "The Cyanobacterial Genome Core and the Origin of Photosynthesis", **Proceedings of the National Academy of Sciences**, Vol. 103, No. 35, pp. 13126-13131.

25. Makarova, K., Slesarev, A., Wolf, Y., Sorokin, A., Mirkin, B., Koonin, E., Pavlov, A., Pavlova, N., Karamychev, V., and Polouchine, N., 2006, "Comparative Genomics of the Lactic Acid Bacteria", **Proceedings of the National Academy of Sciences**, Vol. 103, No. 42, pp. 15611-15616.

26. Wolf, Y.I., Makarova, K.S., Yutin, N., and Koonin, E.V., 2012, "Updated Clusters of Orthologous Genes for Archaea: A Complex Ancestor of the Archaea and the Byways of Horizontal Gene Transfer", **Biol Direct**, Vol. 7, No., p. 46.

27. Chen, F., Mackey, A.J., Vermunt, J.K., and Roos, D.S., 2007, "Assessing Performance of Orthology Detection Strategies Applied to Eukaryotic Genomes", **PLoS One**, Vol. 2, No. 4, p. e383.

28. Kittichotirat, W., Bumgarner, R.E., Asikainen, S., and Chen, C., 2011, "Identification of the Pangenome and Its Components in 14 Distinct Aggregatibacter Actinomycetemcomitans Strains by Comparative Genomic Analysis", **PLoS One**, Vol. 6, No. 7, p. e22420.

29. Sneath, P.H., 1957, "Some Thoughts on Bacterial Classification", **Journal of general microbiology**, Vol. 17, No. 1, pp. 184-200.

30. Zuckerkandl, E. and Pauling, L., 1965, "Molecules as Documents of Evolutionary History", **Journal of theoretical biology**, Vol. 8, No. 2, pp. 357-366.

31. Kimura, M., 1987, "Molecular Evolutionary Clock and the Neutral Theory", **Journal of molecular evolution**, Vol. 26, No. 1-2, pp. 24-33.

32. Koonin, E.V., Makarova, K.S., and Aravind, L., 2001, "Horizontal Gene Transfer in Prokaryotes: Quantification and Classification 1", **Annual Reviews in Microbiology**, Vol. 55, No. 1, pp. 709-742.

33. Woese, C.R., 1987, "Bacterial Evolution", **Microbiological reviews**, Vol. 51, No. 2, p. 221.

34. Doolittle, W.F., 1999, "Phylogenetic Classification and the Universal Tree", **Science**, Vol. 284, No. 5423, pp. 2124-2128.

35. Lerat, E., Daubin, V., and Moran, N.A., 2003, "From Gene Trees to Organismal Phylogeny in Prokaryotes: The Case of the Γ-Proteobacteria", **PLoS biology**, Vol. 1, No. 1, p. e19.

36. Bapteste, E., Boucher, Y., Leigh, J., and Doolittle, W.F., 2004, "Phylogenetic Reconstruction and Lateral Gene Transfer", **Trends in microbiology**, Vol. 12, No. 9, pp. 406-411.

37. Hori, H. and Osawa, S., 1987, "Origin and Evolution of Organisms as Deduced from 5s Ribosomal Rna Sequences", **Molecular biology and evolution**, Vol. 4, No. 5, pp. 445-472.

38. Distel, D., Lane, D., Olsen, G., Giovannoni, S., Pace, B., Pace, N., Stahl, D., and Felbeck, H., 1988, "Sulfur-Oxidizing Bacterial Endosymbionts: Analysis of Phylogeny and Specificity by 16s Rrna Sequences", **Journal of bacteriology**, Vol. 170, No. 6, pp. 2506-2510.

39. Lane, D., Harrison, A., Stahl, D., Pace, B., Giovannoni, S., Olsen, G., and Pace, N., 1992, "Evolutionary Relationships among Sulfur-and Iron-Oxidizing Eubacteria", **Journal of bacteriology**, Vol. 174, No. 1, pp. 269-278.

40. Wolf, Y.I., Rogozin, I.B., Grishin, N.V., and Koonin, E.V., 2002, "Genome Trees and the Tree of Life", **TRENDS in Genetics**, Vol. 18, No. 9, pp. 472-479.

41. Fitch, W.M. and Margoliash, E., 1967, "Construction of Phylogenetic Trees", **Science**, Vol. 155, No. 760, pp. 279-284.

42. Swofford, D.L. and Documentation, B., 1989, "Phylogenetic Analysis Using Parsimony", **Illinois Natural History Survey, Champaign**, Vol., No.

43. Felsenstein, J., 1981, "Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach", **Journal of molecular evolution**, Vol. 17, No. 6, pp. 368-376.

44. Sabes, P.N. and Jordan, M.I.,1995. "Advances in Neural Information Processing Systems". **In G. Tesauro & D. Touretzky & T. Leed (Eds.), Advances in Neural Information Processing Systems**,

45. Felsenstein, J., 1978, "Cases in Which Parsimony or Compatibility Methods Will Be Positively Misleading", **Systematic Biology**, Vol. 27, No. 4, pp. 401-410.

46. Olsen, G.J. and Woese, C.R., 1993, "Ribosomal Rna: A Key to Phylogeny", **The FASEB journal**, Vol. 7, No. 1, pp. 113-123.

47. Saitou, N. and Imanishi, T., 1989, "Relative Efficiencies of the Fitch-Margoliash, Maximum-Parsimony, Maximum-Likelihood, Minimum-Evolution, and Neighbor-Joining Methods of Phylogenetic Tree Construction in Obtaining the Correct Tree", **Mol. Biol. Evol**, Vol. 6, No. 5, pp. 514-525.

48. Price, M.N., Dehal, P.S., and Arkin, A.P., 2010, "Fasttree 2–Approximately Maximum-Likelihood Trees for Large Alignments", **PLoS One**, Vol. 5, No. 3, p. e9490.

49. Cozens, S. and Wainwright, P., 2000, **Beginning Perl**, Wrox Press, Pages.

50. Tatusov, R.L., Koonin, E.V., and Lipman, D.J., 1997, "A Genomic Perspective on Protein Families", **Science**, Vol. 278, No. 5338, pp. 631-637.

51. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., and Kanehisa, M., 2007, "Kaas: An Automatic Genome Annotation and Pathway Reconstruction Server", **Nucleic acids research**, Vol. 35, No. suppl 2, pp. W182-W185.

52. Vesth, T., Lagesen, K., Acar, Ö., and Ussery, D., 2013, "Cmg-Biotools, a Free Workbench for Basic Comparative Microbial Genomics", **PLoS One**, Vol. 8, No. 4, p. e60120.

53. Lagesen, K., Hallin, P., Rødland, E.A., Stærfeldt, H.-H., Rognes, T., and Ussery, D.W., 2007, "Rnammer: Consistent and Rapid Annotation of Ribosomal Rna Genes", **Nucleic acids research**, Vol. 35, No. 9, pp. 3100-3108.

54. Larkin, M.A., Blackshields, G., Brown, N., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., and Lopez, R., 2007, "Clustal W and Clustal X Version 2.0", **Bioinformatics**, Vol. 23, No. 21, pp. 2947-2948.

55. Chu, V.T., Gottardo, R., Raftery, A.E., Bumgarner, R.E., and Yeung, K.Y., 2008, "Mev+ R: Using Mev as a Graphical User Interface for Bioconductor Applications in Microarray Analysis", **Genome Biol**, Vol. 9, No. 7, p. R118.

56. Puigbò, P., Garcia-Vallvé, S., and McInerney, J.O., 2007, "Topd/Fmts: A New Software to Compare Phylogenetic Trees", **Bioinformatics**, Vol. 23, No. 12, pp. 1556-1558.

57. Shih, P.M., Wu, D., Latifi, A., Axen, S.D., Fewer, D.P., Talla, E., Calteau, A., Cai, F., de Marsac, N.T., and Rippka, R., 2013, "Improving the Coverage of the Cyanobacterial Phylum Using Diversity-Driven Genome Sequencing", **Proceedings of the National Academy of Sciences**, Vol. 110, No. 3, pp. 1053-1058.

58. Swingley, W.D., Chen, M., Cheung, P.C., Conrad, A.L., Dejesa, L.C., Hao, J., Honchak, B.M., Karbach, L.E., Kurdoglu, A., and Lahiri, S., 2008, "Niche Adaptation and Genome Expansion in the Chlorophyll D-Producing Cyanobacterium Acaryochloris Marina", **Proceedings of the National Academy of Sciences**, Vol. 105, No. 6, pp. 2005-2010.

59. Tatusov, R.L., Mushegian, A.R., Bork, P., Brown, N.P., Hayes, W.S., Borodovsky, M., Rudd, K.E., and Koonin, E.V., 1996, "Metabolism and Evolution of< I> Haemophilus Influenzae</I> Deduced from a Whole-Genome Comparison with Escherichia Coli", **Current biology**, Vol. 6, No. 3, pp. 279-291.

60. Tomii, K. and Kanehisa, M., 1998, "A Comparative Analysis of Abc Transporters in Complete Microbial Genomes", **Genome research**, Vol. 8, No. 10, pp. 1048-1059.

61. De, N., Navarro, M.V., Raghavan, R.V., and Sondermann, H., 2009, "Determinants for the Activation and Autoinhibition of the Diguanylate Cyclase Response Regulator Wspr", **Journal of molecular biology**, Vol. 393, No. 3, pp. 619-633.

62. Welch, R., Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D., Buckles, E., Liou, S.-R., Boutin, A., and Hackett, J., 2002, "Extensive Mosaic Structure Revealed by the Complete Genome Sequence of Uropathogenic Escherichia Coli", **Proceedings of the National Academy of Sciences**, Vol. 99, No. 26, pp. 17020-17024.

63. Wanchai, V., 2010, **A study of niche adaptation in Cyanobacteria via comparative genomic**, Master's thesis in Bioinformatics and Systems Biology, School of Bioresources and Technology, King Mongkut's University of Technology Thonburi.

64. Vesth, T., Wassenaar, T.M., Hallin, P.F., Snipen, L., Lagesen, K., and Ussery, D.W., 2010, "On the Origins of a Vibrio Species", **Microbial ecology**, Vol. 59, No. 1, pp. 1-13.

65. Willenbrock, H., Petersen, A., Sekse, C., Kiil, K., Wasteson, Y., and Ussery, D.W., 2006, "Design of a Seven-Genome Escherichia Coli Microarray for Comparative Genomic Profiling", **Journal of bacteriology**, Vol. 188, No. 22, pp. 7713-7721.

66. Janda, J.M. and Abbott, S.L., 2007, "16s Rrna Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls", **Journal of clinical microbiology**, Vol. 45, No. 9, pp. 2761-2764.

67. Patel, J.B., 2001, "16s Rrna Gene Sequencing for Bacterial Pathogen Identification in the Clinical Laboratory", **Molecular diagnosis**, Vol. 6, No. 4, pp. 313-321.

68. Philippe, H. and Douady, C.J., 2003, "Horizontal Gene Transfer and Phylogenetics", **Current opinion in microbiology**, Vol. 6, No. 5, pp. 498-505.

69. Daubin, V., Lerat, E., and Perrière, G., 2003, "The Source of Laterally Transferred Genes in Bacterial Genomes", **Genome Biol**, Vol. 4, No. 9, p. R57.

70. Pedulla, M.L., Ford, M.E., Houtz, J.M., Karthikeyan, T., Wadsworth, C., Lewis, J.A., Jacobs-Sera, D., Falbo, J., Gross, J., and Pannunzio, N.R., 2003, "Origins of Highly Mosaic Mycobacteriophage Genomes", **Cell**, Vol. 113, No. 2, pp. 171-182.

71. Huynen, M.A., Dandekar, T., and Bork, P., 1999, "Variation and Evolution of the Citric-Acid Cycle: A Genomic Perspective", **Trends in microbiology**, Vol. 7, No. 7, pp. 281-291.

APPENDICES

Appendix 1 Code translation of NCBI COG categories

INFORMATION STORAGE AND PROCESSING
[J] Translation, ribosomal structure and biogenesis
[A] RNA processing and modification
[K] Transcription
[L] Replication, recombination and repair
[B] Chromatin structure and dynamics

CELLULAR PROCESSES AND SIGNALING
[D] Cell cycle control, cell division, chromosome partitioning
[Y] Nuclear structure
[V] Defense mechanisms
[T] Signal transduction mechanisms
[M] Cell wall/membrane/envelope biogenesis
[N] Cell motility
[Z] Cytoskeleton
[W] Extracellular structures
[U] Intracellular trafficking, secretion, and vesicular transport
[O] Posttranslational modification, protein turnover, chaperones

METABOLISM
[C] Energy production and conversion
[G] Carbohydrate transport and metabolism
[E] Amino acid transport and metabolism
[F] Nucleotide transport and metabolism
[H] Coenzyme transport and metabolism
[I] Lipid transport and metabolism
[P] Inorganic ion transport and metabolism
[Q] Secondary metabolites biosynthesis, transport and catabolism

POORLY CHARACTERIZED
[R] General function prediction only
[S] Function unknown

Appendix 2 Cyanobacteria core genes

YP_001516146.1, YP_001519896.1, YP_007156849.1, YP_007098439.1, YP_007158580.1, YP_321216.1, YP_322408.1, YP_322684.1, YP_323239.1, YP_323719.1, YP_324391.1, YP_324650.1, ZP_03271376.1, YP_005070439.1, ZP_03272834.1, ZP_03275085.1, YP_007079860.1, ZP_03276107.1, YP_005068594.1, ZP_17050794.1, ZP_17055478.1, ZP_17055624.1, ZP_17055678.1, YP_001549895.1, YP_005069051.1, YP_005072837.1, YP_007135870.1, YP_007137168.1, YP_007139336.1, YP_007064422.1, YP_007069161.1, YP_007095542.1, YP_007098227.1, YP_007098285.1, YP_007098933.1, YP_007099200.1, YP_007100056.1, NP_898323.1, YP_007089476.1, YP_007089592.1, YP_007089596.1, YP_474616.1, YP_007090104.1, YP_007090580.1, YP_007093504.1, YP_007094541.1, YP_007094834.1, YP_007141585.1, YP_007141764.1, YP_007142298.1, YP_007142620.1, YP_007143028.1, YP_007143589.1, YP_007144359.1, ZP_00519113.1, ZP_00516385.1, ZP_00515727.1, ZP_00513937.1, ZP_07109404.1, YP_474614.1, YP_007160709.1, YP_007163122.1, YP_007142502.1, YP_007044998.1, YP_007045037.1, YP_007096092.1, YP_007045634.1, YP_007045803.1, YP_007046102.1, YP_007046285.1, YP_001228480.1, YP_007046593.1, YP_007046743.1, YP_007047310.1, YP_007047524.1, YP_001801498.1, YP_001801877.1, YP_001801976.1, YP_001802761.1, YP_001802911.1, YP_001803623.1, YP_001550916.1, YP_001803982.1, YP_002376093.1, YP_002377137.1, YP_002481817.1, YP_002481935.1, NP_926858.1, EKQ68864.1, YP_002482498.1, YP_002483274.1, YP_002484060.1, YP_002484799.1, YP_002485374.1, YP_003885758.1, YP_003888882.1, YP_003890529.1, YP_002371171.1, YP_003137580.1, ZP_06306890.1, ZP_06306896.1, ZP_06307934.1, NP_925730.1, YP_007147339.1, YP_007147750.1, YP_007148312.1, YP_007149074.1, YP_007150275.1, AFZ49705.1, AFZ51167.1, AFZ52046.1, AFZ52067.1, AFZ52069.1, AFZ52118.1, ZP_08984203.1, ZP_08984971.1, YP_007450905.1, ZP_08986984.1, ZP_08987679.1, YP_007109375.1, YP_007110636.1, NP_924450.1, NP_925288.1, NP_924675.1, NP_925257.1, NP_926339.1, NP_926385.1, NP_926863.1, AFZ42754.1, AFZ43168.1, AFZ45478.1, ZP_21048532.1, ZP_21048328.1, ZP_21046197.1, ZP_01083827.1, ZP_21046066.1, ZP_21044742.1, ZP_21044166.1, ZP_21044029.1, ZP_21043710.1, ZP_18906390.1, ZP_18907334.1, ZP_18909059.1, ZP_18909089.1, ZP_18909712.1, ZP_18910622.1, ZP_18911293.1, YP_007070343.1, YP_007071299.1, ERT06269.1, EAW39229.1, EAW38339.1, YP_007120518.1, YP_007121320.1, YP_007122363.1, YP_007123438.1, YP_007123669.1, ZP_21045909.1, YP_007124219.1, YP_007124592.1, YP_007125318.1, YP_007125338.1, YP_007125368.1, ZP_08495272.1, ZP_08492163.1, ZP_08491600.1, ZP_08493151.1, ZP_08495055.1, YP_001655186.1, ZP_18835036.1, ZP_18850722.1, ZP_18842464.1, ZP_18825652.1, ZP_18850989.1, ZP_16389726.1, ZP_18838561.1, ZP_18847023.1, ZP_10227589.1, ZP_10228788.1,

ZP_01629276.1, ZP_01630061.1, ZP_01628595.1, YP_007053738.1, YP_001868151.1, YP_001867331.1, YP_001865446.1, YP_001869651.1, YP_007052223.1, NP_486142.1, NP_486852.1, YP_007073718.1, YP_007074918.1, YP_007075668.1, YP_007075810.1, YP_007078569.1, YP_007078696.1, YP_007084994.1, YP_007086556.1, YP_007086644.1, YP_007086799.1, YP_007087602.1, YP_007088765.1, EKQ70183.1, EKQ70899.1, EKQ69902.1, EKQ68469.1, EKQ68652.1, EKQ68665.1, EKQ68869.1, EKQ67497.1, EKQ67759.1, NP_484953.1, YP_007113626.1, YP_007114316.1, YP_007114720.1, YP_007115493.1, YP_007117780.1, YP_007117902.1, YP_007118220.1, ZP_07110645.1, ZP_07111772.1, ZP_07112635.1, ZP_07113400.1, ZP_07113736.1, YP_007079987.1, YP_007080238.1, YP_007080456.1, YP_007080676.1, YP_007080908.1, YP_007081809.1, YP_007081997.1, YP_001550202.1, YP_001550227.1, YP_001550288.1, YP_001550901.1, ZP_18830096.1, YP_001551283.1, YP_001551405.1, YP_001551556.1, YP_001551558.1, YP_001483733.1, YP_001016161.1, YP_001016493.1, YP_001016519.1, YP_001016520.1, YP_001016791.1, YP_001017521.1, YP_001017595.1, YP_001017682.1, YP_001802300.1, YP_001017746.1, YP_001018261.1, YP_001018320.1, YP_001018775.1, YP_001018832.1, NP_894101.1, NP_894540.1, NP_894868.1, NP_895204.1, NP_895825.1, YP_001010479.1, YP_001012084.1, YP_001014318.1, YP_001014597.1, YP_001015489.1, YP_001015715.1, ZP_21065141.1, YP_007103186.1, ZP_21066325.1, ZP_21066444.1, ZP_21066827.1, ZP_21067064.1, ZP_21067793.1, ZP_21068201.1, ZP_21068217.1, ZP_21068333.1, YP_474566.1, ZP_21068529.1, YP_007100901.1, YP_007100930.1, YP_007101252.1, YP_007101456.1, YP_007101514.1, YP_007101587.1, YP_007101770.1, YP_007101943.1, YP_007101995.1, YP_007102037.1, YP_007102128.1, YP_007102464.1, YP_007102564.1, YP_007102640.1, YP_007102833.1, YP_007102930.1, YP_007102949.1, YP_007103084.1, YP_007103367.1, YP_007103380.1, YP_007103445.1, YP_007103851.1, YP_007104020.1, YP_007104075.1, YP_007104102.1, YP_007104286.1, ZP_06304519.1, YP_007055724.1, YP_007054189.1, YP_007055430.1, YP_007056008.1, YP_007059347.1, ERN42045.1, YP_170723.1, YP_171372.1, YP_171473.1, YP_171649.1, ZP_01471616.1, ZP_01469586.1, ZP_01468327.1, ZP_18820586.1, ZP_01085502.1, YP_001801549.1, ZP_21047302.1, YP_730214.1, YP_730770.1, YP_731402.1, YP_376306.1, YP_377521.1, YP_473833.1, YP_473858.1, YP_474338.1, YP_474428.1, YP_474599.1, YP_474618.1, YP_474628.1, YP_474684.1, YP_474893.1, YP_475120.1, YP_475609.1, YP_475982.1, YP_476072.1, YP_382787.1, YP_007062201.1, YP_007062276.1, YP_001226342.1, ZP_01086598.1, YP_001226642.1, YP_001227095.1, YP_001227486.1, YP_001227821.1, YP_001228059.1, YP_001228223.1, YP_001228633.1, ZP_01472547.1, ZP_01471603.1, ZP_18828344.1, ZP_01081045.1, ZP_01084352.1, ZP_01085507.1, ZP_01085618.1, YP_001224775.1, YP_001225454.1, ZP_08954985.1, ZP_08955388.1, ZP_08956077.1, YP_007146880.1, ZP_08957332.1, NP_896415.1, NP_898131.1, NP_898465.1,

YP_007450902.1, YP_007450940.1, YP_007451290.1, YP_007449943.1, YP_007452284.1, YP_007452423.1, YP_007452598.1, YP_007450306.1, YP_007450540.1, YP_007450607.1, NP_681664.1, NP_682200.1, NP_682216.1, NP_682336.1, NP_683174.1, YP_720352.2, YP_720495.1, YP_722556.1, YP_722574.1, YP_723165.1, YP_723520.1, ZP_21057918.1, ZP_21055260.1, ZP_21053729.1, ZP_21053425.1,

Appendix 3 Conserved hypothetical protein among 100 Cyanobacteria strains

cluster	Kegg KO	NCBI COG categories	acession	product
p-cluster017974	-	-	YP_322408.1	hypothetical protein
p-cluster057551	-	-	YP_007090580.1	hypothetical protein
p-cluster125606	-	-	YP_007150275.1	hypothetical protein
p-cluster180122	-	-	EAW39229.1	hypothetical protein
p-cluster189623	-	-	YP_007123669.1	hypothetical protein
p-cluster259726	-	-	ZP_01630061.1	hypothetical protein
p-cluster309890	-	-	YP_007080456.1	hypothetical protein
p-cluster322654	-	-	YP_001018261.1	hypothetical protein
p-cluster326258	-	-	NP_894868.1	hypothetical protein
p-cluster327217	-	-	NP_895825.1	hypothetical protein
p-cluster329258	-	-	YP_001012084.1	hypothetical protein
p-cluster338193	-	-	ZP_21065141.1	hypothetical protein
p-cluster342258	-	-	YP_007100901.1	hypothetical protein
p-cluster363830	-	-	YP_171473.1	hypothetical protein
p-cluster369442	-	-	ZP_01469586.1	hypothetical protein
p-cluster372354	-	-	YP_001801549.1	hypothetical protein
p-cluster378602	-	-	YP_731402.1	hypothetical protein
p-cluster385117	-	-	YP_474338.1	hypothetical protein
p-cluster385206	-	-	YP_474428.1	hypothetical protein
p-cluster397857	-	-	YP_001227486.1	hypothetical protein
p-cluster403565	-	-	ZP_01081045.1	hypothetical protein
p-cluster411098	-	-	ZP_08954985.1	hypothetical protein
p-cluster431093	-	-	NP_682200.1	hypothetical protein

Appendix 4 all genes KEGG pathway mapping result

KO number / pathway	Number of gene hit in pathway
ko01100 Metabolic pathways	649
ko01110 Biosynthesis of secondary metabolites	279
ko01120 Microbial metabolism in diverse environments	172
ko01230 Biosynthesis of amino acids	103
ko02010 ABC transporters	85
ko01200 Carbon metabolism	84
ko00230 Purine metabolism	63
ko02020 Two-component system	61
ko00195 Photosynthesis	52
ko03010 Ribosome	51
ko00190 Oxidative phosphorylation	48
ko00860 Porphyrin and chlorophyll metabolism	46
ko00240 Pyrimidine metabolism	46
ko00520 Amino sugar and nucleotide sugar metabolism	42
ko00330 Arginine and proline metabolism	38
ko00620 Pyruvate metabolism	33
ko00270 Cysteine and methionine metabolism	32
ko00680 Methane metabolism	32
ko00010 Glycolysis / Gluconeogenesis	31
ko00260 Glycine, serine and threonine metabolism	31
ko00500 Starch and sucrose metabolism	31
ko00630 Glyoxylate and dicarboxylate metabolism	27
ko00970 Aminoacyl-tRNA biosynthesis	26
ko00030 Pentose phosphate pathway	26
ko00720 Carbon fixation pathways in prokaryotes	25
ko00400 Phenylalanine, tyrosine and tryptophan biosynthesis	25
ko00250 Alanine, aspartate and glutamate metabolism	24
ko01210 2-Oxocarboxylic acid metabolism	23
ko03440 Homologous recombination	21
ko00710 Carbon fixation in photosynthetic organisms	20
ko00910 Nitrogen metabolism	20
ko00051 Fructose and mannose metabolism	19
ko00790 Folate biosynthesis	18
ko00130 Ubiquinone and other terpenoid-quinone biosynthesis	18
ko03430 Mismatch repair	17
ko00920 Sulfur metabolism	17
ko00760 Nicotinate and nicotinamide metabolism	16
ko00020 Citrate cycle(TCA cycle	16
ko00640 Propanoate metabolism	16
ko00340 Histidine metabolism	16

KO number / pathway	Number of gene hit in pathway
ko00550 Peptidoglycan biosynthesis	15
ko00196 Photosynthesis - antenna proteins	15
ko00650 Butanoate metabolism	15
ko00561 Glycerolipid metabolism	15
ko00770 Pantothenate and CoA biosynthesis	15
ko01212 Fatty acid metabolism	15
ko00670 One carbon pool by folate	14
ko00380 Tryptophan metabolism	14
ko00350 Tyrosine metabolism	14
ko00480 Glutathione metabolism	14
ko00360 Phenylalanine metabolism	14
ko03410 Base excision repair	14
ko03030 DNA replication	13
ko00900 Terpenoid backbone biosynthesis	13
ko03018 RNA degradation	13
ko03060 Protein export	13
ko00290 Valine, leucine and isoleucine biosynthesis	12
ko00300 Lysine biosynthesis	12
ko00730 Thiamine metabolism	11
ko01220 Degradation of aromatic compounds	11
ko04112 Cell cycle - Caulobacter	11
ko00627 Aminobenzoate degradation	11
ko03070 Bacterial secretion system	11
ko00740 Riboflavin metabolism	11
ko00564 Glycerophospholipid metabolism	11
ko00780 Biotin metabolism	10
ko00906 Carotenoid biosynthesis	10
ko04122 Sulfur relay system	10
ko00540 Lipopolysaccharide biosynthesis	10
ko00052 Galactose metabolism	10
ko00061 Fatty acid biosynthesis	9
ko00071 Fatty acid degradation	9
ko00450 Selenocompound metabolism	9
ko00625 Chloroalkane and chloroalkene degradation	9
ko00362 Benzoate degradation	9
ko02030 Bacterial chemotaxis	9
ko00521 Streptomycin biosynthesis	8
ko00310 Lysine degradation	8
ko03420 Nucleotide excision repair	8
ko00983 Drug metabolism - other enzymes	7
ko00312 beta-Lactam resistance	7
ko00660 C5-Branched dibasic acid metabolism	7
ko01054 Nonribosomal peptide structures	7
ko00440 Phosphonate and phosphinate metabolism	7

KO number / pathway	Number of gene hit in pathway
ko00562 Inositol phosphate metabolism	6
ko00600 Sphingolipid metabolism	6
ko00950 Isoquinoline alkaloid biosynthesis	6
ko00280 Valine, leucine and isoleucine degradation	6
ko00430 Taurine and hypotaurine metabolism	6
ko00410 beta-Alanine metabolism	6
ko04146 Peroxisome	6
ko00361 Chlorocyclohexane and chlorobenzene degradation	5
ko00791 Atrazine degradation	5
ko05152 Tuberculosis	5
ko00040 Pentose and glucuronate interconversions	5
ko00750 Vitamin B6 metabolism	5
ko00960 Tropane, piperidine and pyridine alkaloid biosynthesis	5
ko00253 Tetracycline biosynthesis	5
ko00643 Styrene degradation	5
ko00053 Ascorbate and aldarate metabolism	5
ko00982 Drug metabolism - cytochrome P450	4
ko00633 Nitrotoluene degradation	4
ko03020 RNA polymerase	4
ko00909 Sesquiterpenoid and triterpenoid biosynthesis	4
ko00622 Xylene degradation	4
ko04066 HIF-1 signaling pathway	4
ko05204 Chemical carcinogenesis	4
ko00621 Dioxin degradation	4
ko00460 Cyanoamino acid metabolism	4
ko00401 Novobiocin biosynthesis	4
ko00471 D-Glutamine and D-glutamate metabolism	4
ko00511 Other glycan degradation	4
ko04910 Insulin signaling pathway	4
ko00523 Polyketide sugar unit biosynthesis	4
ko00624 Polycyclic aromatic hydrocarbon degradation	3
ko01053 Biosynthesis of siderophore group nonribosomal peptides	3
ko00930 Caprolactam degradation	3
ko00785 Lipoic acid metabolism	3
ko00311 Penicillin and cephalosporin biosynthesis	3
ko04141 Protein processing in endoplasmic reticulum	3
ko01040 Biosynthesis of unsaturated fatty acids	3
ko05206 MicroRNAs in cancer	3
ko05200 Pathways in cancer	3
ko04626 Plant-pathogen interaction	3
ko00940 Phenylpropanoid biosynthesis	3

KO number / pathway	Number of gene hit in pathway
ko03013 RNA transport	3
ko04726 Serotonergic synapse	3
ko03320 PPAR signaling pathway	3
ko05134 Legionellosis	3
ko04727 GABAergic synapse	3
ko00590 Arachidonic acid metabolism	3
ko04070 Phosphatidylinositol signaling system	3
ko00626 Naphthalene degradation	3
ko00364 Fluorobenzoate degradation	3
ko00140 Steroid hormone biosynthesis	2
ko05016 Huntington's disease	2
ko04940 Type I diabetes mellitus	2
ko04020 Calcium signaling pathway	2
ko00980 Metabolism of xenobiotics by cytochrome P450	2
ko00945 Stilbenoid, diarylheptanoid and gingerol biosynthesis	2
ko00591 Linoleic acid metabolism	2
ko00473 D-Alanine metabolism	2
ko00984 Steroid degradation	2
ko00965 Betalain biosynthesis	2
ko04728 Dopaminergic synapse	2
ko05034 Alcoholism	2
ko04978 Mineral absorption	2
ko04724 Glutamatergic synapse	2
ko00623 Toluene degradation	2
ko04918 Thyroid hormone synthesis	2
ko05014 Amyotrophic lateral sclerosis(ALS	2
ko05120 Epithelial cell signaling in Helicobacter pylori infection	2
ko01052 Type I polyketide structures	2
ko05111 Vibrio cholerae pathogenic cycle	2
ko04370 VEGF signaling pathway	2
ko00903 Limonene and pinene degradation	2
ko00908 Zeatin biosynthesis	2
ko05030 Cocaine addiction	2
ko04068 FoxO signaling pathway	2
ko05340 Primary immunodeficiency	2
ko05031 Amphetamine addiction	2
ko05203 Viral carcinogenesis	1
ko00100 Steroid biosynthesis	1
ko04915 Estrogen signaling pathway	1
ko04011 MAPK signaling pathway - yeast	1
ko00901 Indole alkaloid biosynthesis	1
ko04930 Type II diabetes mellitus	1

KO number / pathway	Number of gene hit in pathway
ko05211 Renal cell carcinoma	1
ko05133 Pertussis	1
ko01051 Biosynthesis of ansamycins	1
ko00830 Retinol metabolism	1
ko00510 N-Glycan biosynthesis	1
ko04666 Fc gamma R-mediated phagocytosis	1
ko04614 Renin-angiotensin system	1
ko01055 Biosynthesis of vancomycin group antibiotics	1
ko03450 Non-homologous end-joining	1
ko05322 Systemic lupus erythematosus	1
ko05410 Hypertrophic cardiomyopathy(HCM	1
ko04973 Carbohydrate digestion and absorption	1
ko04612 Antigen processing and presentation	1
ko00281 Geraniol degradation	1
ko03050 Proteasome	1
ko04723 Retrograde endocannabinoid signaling	1
ko05146 Amoebiasis	1
ko04916 Melanogenesis	1
ko04113 Meiosis - yeast	1
ko05222 Small cell lung cancer	1
ko04917 Prolactin signaling pathway	1
ko00524 Butirosin and neomycin biosynthesis	1
ko05215 Prostate cancer	1
ko05140 Leishmaniasis	1
ko04920 Adipocytokine signaling pathway	1
ko04064 NF-kappa B signaling pathway	1
ko05020 Prion diseases	1
ko04921 Oxytocin signaling pathway	1
ko00363 Bisphenol degradation	1
ko04668 TNF signaling pathway	1
ko02060 Phosphotransferase system(PTS	1
ko02040 Flagellar assembly	1
ko04210 Apoptosis	1
ko04913 Ovarian steroidogenesis	1
ko04142 Lysosome	1
ko05010 Alzheimer's disease	1
ko04974 Protein digestion and absorption	1
ko05142 Chagas disease(American trypanosomiasis	1
ko04964 Proximal tubule bicarbonate reclamation	1
ko04914 Progesterone-mediated oocyte maturation	1
ko00121 Secondary bile acid biosynthesis	1
ko05205 Proteoglycans in cancer	1
ko00642 Ethylbenzene degradation	1
ko01057 Biosynthesis of type II polyketide products	1

KO number / pathway	Number of gene hit in pathway
ko05100 Bacterial invasion of epithelial cells	1
ko00941 Flavonoid biosynthesis	1
ko00072 Synthesis and degradation of ketone bodies	1
ko00120 Primary bile acid biosynthesis	1
ko04151 PI3K-Akt signaling pathway	1
ko03008 Ribosome biogenesis in eukaryotes	1
ko04621 NOD-like receptor signaling pathway	1

Appendix 5 full comparison result of all phylogenetic tree using TOPD/FMTS program

Tree comparison data

###################################### topd 16s - core ##############################

* Percentage of taxa in common: 100.0%

* Split Distance [differents/possibles]: 0.56701030927835 [110 / 194]

* Disagreement [taxa disagree / all taxa]: [43 / 100], New Split Distance: 0.203703703703704, Taxa disagree: (MicrocystisaeruginosaPCC9701 MicrocystisaeruginosaPCC9717 MicrocystisaeruginosaTAIHU98 OscillatoriaPCC6506 AcaryochlorismarinaMBIC11017 Anabaena90 AnabaenacylindricaPCC7122 CalothrixPCC6303 CalothrixPCC7507 ChamaesiphonminutusPCC6605 ChroococcidiopsisthermalisPCC7203 CrinaliumepipsammumPCC9333 CyanothecePCC7425 CylindrormumstagnalePCC7417 FischerellaJSC11 GeitlerinemaPCC7407 GloeobacterviolaceusPCC7421 LeptolyngbyaPCC6406 LeptolyngbyaPCC7375 MicrocoleusPCC7113 MicrocystisaeruginosaPCC7941 MicrocystisaeruginosaPCC9432 MicrocystisaeruginosaPCC9808 MicrocystisaeruginosaSPC777 NodulariamigenaCCY9414uid54171 NostocPCC7107 NostocPCC7524 NostocpunctiformePCC73102 OscillatoriaacuminataPCC6304 OscillatorialescyanobacteriumJSC12 ProchlorococcusmarinusAS9601 ProchlorococcusmarinusMIT9215 ProchlorococcusmarinusMIT9301 ProchlorococcusmarinusMIT9303 ProchlorococcusmarinusMIT9312 ProchlorococcusmarinusMIT9313 RivulariaPCC7116 RubidibacterlacunaeKORDI512 StanieriacyanoaeraPCC7437 SynechococcusJA33Ab SynechococcusRCC307 TrichodesmiumerythraeumIMS101 XenococcusPCC7305)

###################################### topd 16s - phyletic ##########################

* Percentage of taxa in common: 100.0%

* Split Distance [differents/possibles]: 0.670103092783505 [130 / 194]

* Disagreement [taxa disagree / all taxa]: [42 / 100], New Split Distance: 0.272727272727273, Taxa disagree: (SynechococcusWH7803 MicrocystisaeruginosaPCC9701 MicrocystisaeruginosaPCC9717 MicrocystisaeruginosaSPC777 ProchlorococcusmarinusAS9601 RivulariaPCC7116 StanieriacyanoaeraPCC7437 TrichodesmiumerythraeumIMS101 AcaryochlorismarinaMBIC11017 Anabaena90 AnabaenacylindricaPCC7122 CalothrixPCC7507 ChamaesiphonminutusPCC6605 ChroococcidiopsisthermalisPCC7203 CrinaliumepipsammumPCC9333 CrocoaerawatsoniiWH8501 CyanobacteriumUCYNA CyanothecePCC7425 CylindrormumstagnalePCC7417 GeitlerinemaPCC7407 GloeobacterviolaceusPCC7421 LeptolyngbyaPCC6406 LeptolyngbyaPCC7375 MicrocoleusPCC7113 MicrocystisaeruginosaPCC9808 MicrocystisaeruginosaTAIHU98 NodulariamigenaCCY9414uid54171 NostocPCC7107 NostocPCC7524

NostocpunctiformePCC73102 OscillatoriaPCC6506 OscillatoriaacuminataPCC6304 OscillatorialescyanobacteriumJSC12 ProchlorococcusmarinusMIT9211 ProchlorococcusmarinusMIT9303 ProchlorococcusmarinusMIT9313 PseudanabaenaPCC7367 PseudanabaenabicepsPCC7429 SynechococcusJA33Ab SynechococcusPCC7502 SynechococcusRCC307 XenococcusPCC7305)

###################################### topd core - phyletic ##########################

* Percentage of taxa in common: 100.0%

* Split Distance [differents/possibles]: 0.54639175257732 [106 / 194]

* Disagreement [taxa disagree / all taxa]: [32 / 100], New Split Distance: 0.246153846153846, Taxa disagree: (SynechococcusRCC307 CyanobacteriumUCYNA MicrocystisaeruginosaTAIHU98 RivulariaPCC7116 StanieriacyanoaeraPCC7437 SynechococcusWH7803 SynechococcusJA33Ab AcaryochlorismarinaMBIC11017 Anabaena90 AnabaenacylindricaPCC7122 CalothrixPCC7507 ChamaesiphonminutusPCC6605 ChroococcidiopsisthermalisPCC7203 CrocoaerawatsoniiWH8501 CyanothecePCC7425 CylindrormumstagnalePCC7417 GeitlerinemaPCC7407 MicrocystisaeruginosaPCC9808 MicrocystisaeruginosaSPC777 NodulariamigenaCCY9414uid54171 NostocpunctiformePCC73102 OscillatoriaacuminataPCC6304 OscillatorialescyanobacteriumJSC12 ProchlorococcusmarinusMIT9211 ProchlorococcusmarinusMIT9215 ProchlorococcusmarinusMIT9312 PseudanabaenaPCC7367 PseudanabaenabicepsPCC7429 RubidibacterlacunaeKORDI512 SynechococcusPCC7502 TrichodesmiumerythraeumIMS101 XenococcusPCC7305)

Appendix 6 core genes KEGG pathway mapping result

KO number / pathway	Number of gene hit in pathway
ko01100 Metabolic pathways	149
ko01110 Biosynthesis of secondary metabolites	65
ko03010 Ribosome	39
ko01120 Microbial metabolism in diverse environments	33
ko01230 Biosynthesis of amino acids	27
ko01200 Carbon metabolism	23
ko00190 Oxidative phosphorylation	21
ko00860 Porphyrin and chlorophyll metabolism	19
ko00195 Photosynthesis	17
ko00970 Aminoacyl-tRNA biosynthesis	16
ko00240 Pyrimidine metabolism	13
ko00230 Purine metabolism	12
ko00550 Peptidoglycan biosynthesis	9
ko03440 Homologous recombination	9
ko00620 Pyruvate metabolism	9
ko01212 Fatty acid metabolism	9
ko00010 Glycolysis / Gluconeogenesis	8
ko00061 Fatty acid biosynthesis	8
ko00300 Lysine biosynthesis	8
ko00900 Terpenoid backbone biosynthesis	8
ko00260 Glycine, serine and threonine metabolism	7
ko03018 RNA degradation	7
ko00520 Amino sugar and nucleotide sugar metabolism	7
ko02010 ABC transporters	6
ko00030 Pentose phosphate pathway	6
ko00710 Carbon fixation in photosynthetic organisms	5
ko00680 Methane metabolism	5
ko04112 Cell cycle - Caulobacter	5
ko03070 Bacterial secretion system	5
ko03060 Protein export	5
ko00250 Alanine, aspartate and glutamate metabolism	5
ko00564 Glycerophospholipid metabolism	5
ko00730 Thiamine metabolism	4
ko00720 Carbon fixation pathways in prokaryotes	4
ko00253 Tetracycline biosynthesis	4
ko00740 Riboflavin metabolism	4
ko00780 Biotin metabolism	4
ko00400 Phenylalanine, tyrosine and tryptophan biosynthesis	4
ko00640 Propanoate metabolism	4
ko00630 Glyoxylate and dicarboxylate metabolism	4
ko00920 Sulfur metabolism	4
ko01210 2-Oxocarboxylic acid metabolism	4

KO number / pathway	Number of gene hit in pathway
ko00270 Cysteine and methionine metabolism	4
ko00330 Arginine and proline metabolism	3
ko02020 Two-component system	3
ko03020 RNA polymerase	3
ko00561 Glycerolipid metabolism	3
ko00500 Starch and sucrose metabolism	3
ko00540 Lipopolysaccharide biosynthesis	3
ko04066 HIF-1 signaling pathway	3
ko03420 Nucleotide excision repair	3
ko00910 Nitrogen metabolism	3
ko00471 D-Glutamine and D-glutamate metabolism	3
ko03030 DNA replication	3
ko00480 Glutathione metabolism	3
ko00670 One carbon pool by folate	2
ko00790 Folate biosynthesis	2
ko00040 Pentose and glucuronate interconversions	2
ko04122 Sulfur relay system	2
ko00020 Citrate cycle TCA cycle	2
ko00310 Lysine degradation	2
ko03410 Base excision repair	2
ko00473 D-Alanine metabolism	2
ko01040 Biosynthesis of unsaturated fatty acids	2
ko05152 Tuberculosis	2
ko00760 Nicotinate and nicotinamide metabolism	2
ko05134 Legionellosis	1
ko05203 Viral carcinogenesis	1
ko00521 Streptomycin biosynthesis	1
ko00960 Tropane, piperidine and pyridine alkaloid biosynthesis	1
ko04930 Type II diabetes mellitus	1
ko04940 Type I diabetes mellitus	1
ko00983 Drug metabolism - other enzymes	1
ko00290 Valine, leucine and isoleucine biosynthesis	1
ko01051 Biosynthesis of ansamycins	1
ko05010 Alzheimer's disease	1
ko04070 Phosphatidylinositol signaling system	1
ko03430 Mismatch repair	1
ko00130 Ubiquinone and other terpenoid-quinone biosynthesis	1
ko00312 beta-Lactam resistance	1
ko05111 Vibrio cholerae pathogenic cycle	1
ko00052 Galactose metabolism	1
ko00906 Carotenoid biosynthesis	1
ko04626 Plant-pathogen interaction	1
ko00051 Fructose and mannose metabolism	1

KO number / pathway	Number of gene hit in pathway
ko00350 Tyrosine metabolism	1
ko03013 RNA transport	1
ko04146 Peroxisome	1
ko00053 Ascorbate and aldarate metabolism	1
ko00523 Polyketide sugar unit biosynthesis	1
ko00460 Cyanoamino acid metabolism	1
ko00450 Selenocompound metabolism	1
ko00625 Chloroalkane and chloroalkene degradation	1
ko00650 Butanoate metabolism	1
ko03320 PPAR signaling pathway	1

Appendix 7 supplementary data

All of the supplementary file could be found in the compact disk of this thesis, the supplementary file includes:

1. 16s tree clade specific genes
2. 16s tree clade core genes
3. 16s tree clade pan genes
4. 16s tree nonspecific genes
5. core genes tree clade specific genes
6. core genes tree clade core genes
7. core genes tree clade pan genes
8. core genes tree nonspecific genes
9. phyletic tree clade specific genes
10. phyletic tree clade core genes
11. phyletic tree clade pan genes
12. phyletic tree nonspecific genes
13. species specific genes
14. clade TCA component
15. phylogenetic tree data
16. COG data
17. Organisms data
18. Blast matrix data

Appendix 8 Programming script

8.1 Script for integrating COGs data in Dr. Weerayuth Kittichotirat's format with phylogenetic tree data in newick format

```
#!/usr/bin/perl -w
#program for grouping genes accroding to phylogenetic tree applied to Dr. Weerayuth K. cog data structure
#written by Sivamoke Dissook Q1 2014; sivamoke.bif@mail.kmutt.ac.th
#v2 add non resolved tree grouping capability
#v3 add pan core finding function

use strict;
use Excel::Writer::XLSX;
use List::Util qw(sum);
use Data::Dumper;
use Cwd 'realpath';

#global variable
sub main;                  #jointed operation center
sub cog_count;             #count cog line
sub tree_grouper;          #take newick tree and group the organisms from the tree
sub assign_position;       #search and assign organism position in cogs table
sub species_specific;      #search for species specific genes
sub set_specific;          #grouping set specific genes together
sub non_specific;          #search for non specific genes
sub find_tree;             #search for tree file in data folder
sub check_arguments;       #check the inputs
sub find_core;             #extract core for each clade
sub find_pan;              #extract pan for each clade

#path to data dir which contains:
#rooted tree file in newick format ended with .tree
#list.strains.01.development
#list.strains.02.development <<<----fixed name incase of organism's name > 31 characters
#gene_table.protein.tab.development (COGs table)


main(@ARGV); #USAGE: cog_tree_integration_v1 dataDir outDir
```

```
sub main
{
################################### ~DATA VARIABLES~
############################################
   my $timestamp = localtime(time);
   print "[$timestamp] cog tree protocol initiated... \n";
   my $data_dir = realpath shift(@_);
   my $out_dir = realpath shift(@_);
   if (check_arguments($data_dir, $out_dir) eq '')
   {
      my $cog_file = "$data_dir/gene_table.protein.tab";
      my $sets_file = tree_grouper(find_tree($data_dir), $out_dir);
      my $list_file = "$data_dir/list.strains.01.development";
      my $list_file_l31 = "$data_dir/list.strains.02.development";
      my $position = assign_position($cog_file, $sets_file, $list_file, $out_dir);
      my $sets_specific = "$out_dir/group_cluster.txt";
      my $species_specific = "$out_dir/species_cluster.txt";
      my $cog_number = cog_count($cog_file);

#########################################################################
########################
      #searching clade specific genes
      $timestamp = localtime(time);
      print "[$timestamp] working on clade specific genes\n";
      my $set_genes = set_specific ($cog_file, $position, $out_dir);

      #searching species specfic genes
      $timestamp = localtime(time);
      print "[$timestamp] working on species specific genes\n";
      my $species_genes = species_specific ($cog_file, $list_file_l31, $out_dir);

      #searching for non specific genes
      $timestamp = localtime(time);
      print "[$timestamp] working on non specific genes\n";
      my $ngenes_count = non_specific($cog_file, $sets_specific, $species_specific,
$out_dir);
      #searching for pan genes in each clade
      $timestamp = localtime(time);
      print "[$timestamp] working on pan genes\n";
```

```
        find_pan($cog_file, $position, $out_dir);
        #searching for core genes in each clade
        $timestamp = localtime(time);
        print "[$timestamp] working on core genes\n";
        find_core($cog_file, $position, $out_dir);

        $timestamp = localtime(time);
        print "[$timestamp] process completed \n";
        print "total genes in cog = $cog_number\n
            total set specific genes = $set_genes\n
            total species specific genes = $species_genes\n
            total non specific genes = $ngenes_count\n";

    }
}

sub find_pan
{
    my $cog = shift @_;
    my $position = shift @_;
    my $out_dir = shift @_;

    my $workbook = Excel::Writer::XLSX->new("$out_dir/pan_genome.xlsx");

    # Create a format for the column headings
    my $header = $workbook->add_format();
    $header->set_bold();
    $header->set_size(12);

    unless ( open(FILE, $position) ) {print "Cannot open file \"$position\"\n\n";}
    my @column = <FILE>;
    close(FILE);
    open (OUTFILE, ">$out_dir/pan_cluster.txt");
    open (OUTFILE2, ">$out_dir/pan_report.txt");

    my @hypo_values;
    my @genes_values;
    for(my $j=0;scalar(@column)>$j;$j++)
    {
```

```
chomp $column[$j];
my @a01 = split(/\|/, $column[$j]);
my @a02 = split(/\t/, $a01[1]);
my @a03 = split(/\t/, $a01[0]);

my $timestamp = localtime(time);
print "[$timestamp] Building pan genome for $a03[0] \n";
print OUTFILE "$a03[0]\t";
print OUTFILE2 "$a03[0]\t";

my $group = $workbook->add_worksheet("$a03[0]");
$group->write(0, 0, 'cluster', $header);
$group->write(0, 1, 'Kegg KO', $header);
$group->write(0, 2, 'cellular f(n)', $header);
$group->write(0, 3, 'acession', $header);
$group->write(0, 4, 'product', $header);
$group->write(0, 5, $a01[0]);

unless ( open(FILE, $cog) ) {print "Cannot open file \"$cog\"\n\n";}
my $row = 1;
my $hypo = 0;
my $group_genes = 0;
while (<FILE>)
{
        chomp $_;
        my @cog_line = split (/\t/, $_);
              for(my $l=0;scalar(@a02)>$l;$l++)
              {
                 if($cog_line[$a02[$l]] =~ /1,.+\|/)
                 {
                        if($cog_line[4] =~ m/hypothetical protein/)
                    {
                       $hypo++;
                    }
                        print OUTFILE "$cog_line[0]\t";
                        $group->write($row, 0,$cog_line[0]);
                        $group->write($row, 1,$cog_line[1]);
   $group->write($row, 2,$cog_line[2]);
           $group->write($row, 3,$cog_line[3]);
```

```
                        $group->write($row, 4,$cog_line[4]);
                        $row++;
                        $group_genes++;
                        last;
                    }
                }
            }
            print OUTFILE "\n";

        if($group_genes > 0)
        {
            my $hypo_percent = ($hypo/$group_genes)*100;
            push (@hypo_values, $hypo_percent);
            push (@genes_values, $group_genes);
            print OUTFILE2 "$group_genes\n";
            $group->write(0, 6,'%hypo');
            $group->write(0, 7,$hypo_percent);
        }
        else
        {
            push (@hypo_values, 0);
            print OUTFILE2 "\n";
    }
  }
  close(FILE);
  close(OUTFILE2);
  close(OUTFILE);
}

sub find_core
{
  my $cog = shift @_;
  my $position = shift @_;
  my $out_dir = shift @_;

  my $workbook = Excel::Writer::XLSX->new("$out_dir/core_genome.xlsx");

  # Create a format for the column headings
  my $header = $workbook->add_format();
```

```
$header->set_bold();
$header->set_size(12);


unless ( open(FILE, $position) ) {print "Cannot open file \"$position\"\n\n";}
my @column = <FILE>;
close(FILE);
open (OUTFILE, ">$out_dir/core_cluster.txt");
open (OUTFILE2, ">$out_dir/core_report.txt");

my @hypo_values;
my @genes_values;
for(my $j=0;scalar(@column)>$j;$j++)
{
	chomp $column[$j];
	my @a01 = split(/\|/, $column[$j]);
	my @a02 = split(/\t/, $a01[1]);
	my @a03 = split(/\t/, $a01[0]);

	my $timestamp = localtime(time);
	print "[$timestamp] Building core genome for $a03[0] \n";
	print OUTFILE "$a03[0]\t";
	print OUTFILE2 "$a03[0]\t";

	my $group = $workbook->add_worksheet("$a03[0]");
	$group->write(0, 0, 'cluster', $header);
	$group->write(0, 1, 'Kegg KO', $header);
	$group->write(0, 2, 'cellular f(n)', $header);
	$group->write(0, 3, 'acession', $header);
	$group->write(0, 4, 'product', $header);
	$group->write(0, 5, $a01[0]);

	unless ( open(FILE, $cog) ) {print "Cannot open file \"$cog\"\n\n";}
	my $row = 1;
	my $hypo = 0;
	my $group_genes = 0;
	while (<FILE>)
	{
		chomp $_;
```

```
        my @cog_line = split (/\t/, $_);
            my $count = 0;
            my $pcount =0;
            for(my $l=0;scalar(@a02)>$l;$l++)
            {
               if($cog_line[$a02[$l]] =~ /1,.+\|/)
               {
                        $count ++;
               }
               if($cog_line[$a02[$l]] =~ /P,.+/)
               {
                  $pcount++;
               }
        }
            if ($count > 0 && $count+$pcount == scalar(@a02))
            {
               if($cog_line[4] =~ m/hypothetical protein/)
               {
                 $hypo++;
               }
                     print OUTFILE "$cog_line[0]\t";
                     $group->write($row, 0,$cog_line[0]);
                     $group->write($row, 1,$cog_line[1]);
$group->write($row, 2,$cog_line[2]);
        $group->write($row, 3,$cog_line[3]);
                  $group->write($row, 4,$cog_line[4]);
                  $row++;
                  $group_genes++;
        }
      }
if($group_genes > 0)
{
   my $hypo_percent = ($hypo/$group_genes)*100;
   push (@hypo_values, $hypo_percent);
   push (@genes_values, $group_genes);
   $group->write(0, 6,'%hypo');
   $group->write(0, 7,$hypo_percent);
   print OUTFILE2 "$group_genes\n";
   print OUTFILE "\n";
```

```
            }
            else
            {
               push (@hypo_values, 0);
               print OUTFILE "\n";
               print OUTFILE2 "\n";
      }
   }
   close(FILE);
   close(OUTFILE);
   close(OUTFILE2);
}

sub non_specific
{
   my $cog = shift(@_);
   my $group_cluster = shift(@_);
   my $species_cluster = shift(@_);
   my $out_dir = shift(@_);
   open (OUTFILE,">$out_dir/non_specific.txt") || die;
   my $workbook = Excel::Writer::XLSX->new("$out_dir/non_specific_genes.xlsx");
   my $header = $workbook->add_format();
   $header->set_bold();
   $header->set_size(12);
   my $ngenes = $workbook->add_worksheet('non_specific');
   $ngenes->write(0, 0, 'cluster', $header);
   $ngenes->write(0, 1, 'Kegg KO', $header);
   $ngenes->write(0, 2, 'cellular f(n)', $header);
   $ngenes->write(0, 3, 'acession', $header);
   $ngenes->write(0, 4, 'product', $header);

   #store group cluster and species cluster in hash
   my %group;
   my %species;

   unless ( open(FILE, $group_cluster) ) {print "Cannot open file \"$group_cluster\"\n\n";}
   while(<FILE>)
   {
      my @a01 = split(/\t/, $_);
```

```
        for(my $i=1;scalar(@a01)>$i;$i++)
        {
           $group{$a01[$i]}=$a01[0];
        }
     }
     close(FILE);
     unless ( open(FILE, $species_cluster) ) {print "Cannot open file
\"$species_cluster\"\n\n";}
     while(<FILE>)
     {
        my @a01 = split(/\t/, $_);
        for(my $i=1;scalar(@a01)>$i;$i++)
        {
           $species{$a01[$i]}=$a01[0];
        }
     }
     close(FILE);
     #print Dumper(\%group);
     #search the cogs
     unless ( open(FILE, $cog) ) {print "Cannot open file \"$cog\"\n\n";}
     my $row = 1;
     my $ngenes_count = 0;
     while(<FILE>)
     {
        chomp $_;
        my @a01 = split(/\t/, $_);
        if(!defined $species{$a01[0]})
        {
           if(!defined $group{$a01[0]})
           {
              print OUTFILE "$a01[0]\t$a01[3]\t$a01[4]\t";
              $ngenes->write($row, 0,$a01[0]);
                        $ngenes->write($row, 1,$a01[1]);
                        $ngenes->write($row, 2,$a01[2]);
                              $ngenes->write($row, 3,$a01[3]);
                              $ngenes->write($row, 4,$a01[4]);

                              my $j = 5;
              for (my $i=6;scalar(@a01)>$i;$i++)
```

```
                {

                    if ($a01[$i] =~ /1,.+\|/||$a01[$i] =~ /P,.+/)
                    {

                            my @a02 = split(',', $a01[$i]);
                            my @a03 = split(/\|/, $a02[1]);
                            print OUTFILE "$a03[0]\t";
                            $ngenes->write($row, $j,$a03[0]);
                            $j++;

                        }
            }
            $row++;
                            $ngenes_count++;
            print OUTFILE "\n";
          }
       }
    }
    close(FILE);
    close(OUTFILE);
    return $ngenes_count;
}

sub species_specific
{
    my $cog = shift(@_);
    my $position = shift(@_);
    my $out_dir = shift(@_);
    my $workbook = Excel::Writer::XLSX->new("$out_dir/species_cluster.xlsx");
    open (OUTFILE, ">$out_dir/species_cluster.txt");
    open (OUTFILE2, ">$out_dir/species_specfic_report");
    print OUTFILE2 "Organism\t#spec_cog\n";

    # Create a format for the column headings
    my $header = $workbook->add_format();
    $header->set_bold();
    $header->set_size(12);
```

```
unless ( open(FILE, $position) ) {print "Cannot open file \"$position\"\n\n";}
my @column = <FILE>;
close(FILE);

unless ( open(FILE, $cog) ) {print "Cannot open file \"$cog\"\n\n";}
my @total_genes_count;
for(my $j=0;scalar(@column)>$j;$j++)
{
        chomp $column[$j];
        my @a01 = split(/\t/, $column[$j]);

   my $position = $j + 6;
        my $timestamp = localtime(time);
        print "[$timestamp] processing  $a01[2] \n";
        print OUTFILE "$a01[0]\t";
        print OUTFILE2 "$a01[0]\t";

        my $strain = $workbook->add_worksheet("$a01[0]");
        $strain->write(0, 0, 'cluster', $header);
        $strain->write(0, 1, 'Kegg KO', $header);
        $strain->write(0, 2, 'Cellular f(n)', $header);
        $strain->write(0, 3, 'acession', $header);
        $strain->write(0, 4, 'product', $header);

        unless ( open(FILE, $cog) ) {print "Cannot open file \"$cog\"\n\n";}
        my $row = 1;
        my $genes_count=0;
        while (<FILE>)
        {
                chomp $_;
                my @cog_line = split (/\t/, $_);
                my $count = 0;
                if ($cog_line[$position] =~ /1,.+\|/||$cog_line[$position] =~ /P,.+/)
                        {
                                $count ++;
                        }

                        if ($count == 1)
                        {
```

```
                                my $positive = 0;
                                for(my $j=6;scalar(@cog_line)>$j;$j++)
                                {
                                        if ($cog_line[$j] =~ /1,.+\|/||$cog_line[$j] =~ /P,.+/)
                                        {
                                                $positive++;

                                        }
                                }

                                if ($positive == 1)
                                {
                                        print OUTFILE "$cog_line[0]\t";
                                        $strain->write($row, 0,$cog_line[0]);
                                        $strain->write($row, 1,$cog_line[1]);
                                        $strain->write($row, 2,$cog_line[2]);
                                        $strain->write($row, 3,$cog_line[3]);
                                        $strain->write($row, 4,$cog_line[4]);
                                        $row++;
                                        $genes_count++;
                                }
                        }
                }
                print OUTFILE2 "$genes_count\n";
                if($genes_count > 0)
                {
                   push(@total_genes_count, $genes_count);
                }

                print OUTFILE "\n";
        }
        if(scalar(@total_genes_count)>0)
        {
           my $ave_unique_genes = sum(@total_genes_count)/scalar(@total_genes_count);
        }
        else
        {
           my $ave_unique_genes = "no unique gene";
        }
```

```
    close(OUTFILE);
    close(OUTFILE2);
    close(FILE);
    return sum(@total_genes_count);
}

sub set_specific
{
    my $cog = shift @_;
    my $position = shift @_;
    my $out_dir = shift @_;

    my $workbook = Excel::Writer::XLSX->new("$out_dir/sets_cluster.xlsx");

    # Create a format for the column headings
    my $header = $workbook->add_format();
    $header->set_bold();
    $header->set_size(12);

    unless ( open(FILE, $position) ) {print "Cannot open file \"$position\"\n\n";}
    my @column = <FILE>;
    close(FILE);

    open (OUTFILE, ">$out_dir/group_cluster.txt");

    my @hypo_values;
    my @genes_values;
    for(my $j=0;scalar(@column)>$j;$j++)
    {
            chomp $column[$j];
            my @a01 = split(/\|/, $column[$j]);
            my @a02 = split(/\t/, $a01[1]);
            my @a03 = split(/\t/, $a01[0]);

            my $timestamp = localtime(time);
            print "[$timestamp] processing  $a03[0] \n";
            print OUTFILE "$a03[0]\t";

            my $group = $workbook->add_worksheet("$a03[0]");
```

```
$group->write(0, 0, 'cluster', $header);
$group->write(0, 1, 'Kegg KO', $header);
$group->write(0, 2, 'cellular f(n)', $header);
$group->write(0, 3, 'acession', $header);
$group->write(0, 4, 'product', $header);
$group->write(0, 5, $a01[0]);

unless ( open(FILE, $cog) ) {print "Cannot open file \"$cog\"\n\n";}
my $row = 1;
my $hypo = 0;
my $group_genes = 0;
while (<FILE>)
{
	chomp $_;
	my @cog_line = split (/\t/, $_);
		my $count = 0;
		for(my $l=0;scalar(@a02)>$l;$l++)
		{
			if ($cog_line[$a02[$l]] =~ /1,.+\|/||$cog_line[$a02[$l]] =~
/P,.+/)
			{
				$count ++;
			}
		}
		if ($count == scalar(@a02))
		{
			my $positive = 0;
			for(my $j=6;scalar(@cog_line)>$j;$j++)
			{
				if ($cog_line[$j] =~ /1,.+\|/||$cog_line[$j] =~
/P,.+/)
				{
					$positive++;
				}
			}

			if ($positive == scalar(@a02))
			{
			   if($cog_line[4] =~ m/hypothetical protein/)
```

```
                                    {
                                       $hypo++;
                                    }
                                        print OUTFILE "$cog_line[0]\t";
                                        $group->write($row, 0,$cog_line[0]);
                                        $group->write($row, 1,$cog_line[1]);
                                        $group->write($row, 2,$cog_line[2]);
                                        $group->write($row, 3,$cog_line[3]);
                                        $group->write($row, 4,$cog_line[4]);
                                        $row++;
                                        $group_genes++;
                                }
                        }
                }

        if($group_genes > 0)
        {
           my $hypo_percent = ($hypo/$group_genes)*100;
           push (@hypo_values, $hypo_percent);
           push (@genes_values, $group_genes);
           $group->write(0, 6,'%hypo');
           $group->write(0, 7,$hypo_percent);
        }
        else
        {
           push (@hypo_values, 0);
        }

        print OUTFILE "\n";
}
close(FILE);

my $mean_hypo = mean(@hypo_values);
close(OUTFILE);

sub mean
{
   return sum(@_)/@_;
}
```

```
    return sum(@genes_values);

}

sub assign_position
{
    my $cogs = shift(@_);
    my $organism_group = shift(@_);

    my $strains_list = shift(@_);
    my $out_dir = shift(@_);

    open (OUTFILE, ">$out_dir/organisms.positions");
    my $result_path = "$out_dir/organisms.positions";

    #put group in an array
    unless ( open(FILE, $organism_group) ) {print "Cannot open
file\"$organism_group\"\n\n";}
    my @group = <FILE>;
    close(FILE);

    #put column of each strain in to hash
    my %strains;
    unless ( open(FILE, $strains_list) ) {print "Cannot open file \"$organism_group\"\n\n";}
    my $position = 6;
    while(<FILE>)
    {
            my @a01 = split(/\t/, $_);
            my $organism = $a01[0];
            my $column = $position;
            $strains{$organism} = $column;
            $position++;
    }
    close(FILE);

    #map group to column and store in one variable
    for(my $k=0;scalar(@group)>$k;$k++)
    {
            chomp $group[$k];
```

```
            my @a01 = split(/\t/, $group[$k]);
            print OUTFILE "$group[$k]|";
            for(my $l=1;scalar(@a01)>$l;$l++)
            {
               if(!defined $strains{$a01[$l]})
               {
                  print "$a01[$l] is not match, check organism's name\n";
                  die;
               }
               else
               {
                        print OUTFILE "$strains{$a01[$l]}\t";
                    }
            }
            print OUTFILE "\n";
   }
   close(OUTFILE);
   return $result_path;
}

sub tree_grouper
{
   my $file = shift(@_);#path to tree file
   my $out_dir = shift(@_);
   unless ( open(FILE, $file) ) {print "Cannot open file \"$file\"\n\n";}
   open (OUTFILE, ">$out_dir/organisms.sets");
   my $result_path = "$out_dir/organisms.sets";
   my $tree = <FILE>;
   $tree =~ s/\d\.\d\d\d//g; #remove supporting value
   $tree =~ s/:\d\d//g;  #removing branch length
   $tree =~ s/\/\d\-\d\d\d\d//g; #removing number from gap remove program
   $tree =~ s/\*//g; #remove * for complete genome indicator
   #print $tree;
   my $string = $tree;
   my ($whole_string, @parts) = flatten(parse(\$string));

   for my $i (0 .. $#parts)
   {
      print OUTFILE "clade", $i + 1, "\t", "$parts[$i]\n";
```

```
}


sub parse
{
   my ($string_ref) = @_;
   if($$string_ref =~ s/\A([(])//) #expecting '(x,y,z,...)'
   {
      my @secondList = ();
      my $first = parse($string_ref); #find x
      push(@secondList, $first);

      while($$string_ref =~ s/\A[,]//) #loop for 2nd, 3rd, ... elements
      {
         my $second = parse($string_ref);
         push(@secondList, $second);
      }

      if(scalar(@secondList) == 1)
      {
         die "Expected atleast one comma [,]";
      }

      $$string_ref =~ s/\A[)]// or die "Expected a closing paren [)]";
      return \@secondList;
   }
   elsif ($$string_ref =~ s/\A([^,()]+)//) #expecting any alphabets
   {
      return $1;
   }
   else #ill-formed
   {
      die "Expected [(] or [^,()]";
   }
}


sub flatten
{
```

```
        my ($data) = @_;
        if (ref $data eq 'ARRAY')
        {
           my ($first, @secondList) = @$data;
           my @second_str = ();
           my @second_others = ();
           my ($first_str,  @first_others ) = flatten($first );

           foreach my $elm (@secondList)
           {
              my ($temp, @temp_others) = flatten($elm);
              push(@second_str, $temp);
              push(@second_others, @temp_others);
           }

           my $str = join("\t",$first_str,@second_str);
           my @others = (@first_others, @second_others, $str);
           return $str, @others;
        }
        elsif (ref $data eq '')
        {
           return $data;
        }
        else
        {
           die "Unknown data type ", ref $data;
        }
     }
     close(OUTFILE);
     return $result_path;
}

sub cog_count
{
     my $cog = shift(@_);
     unless ( open(FILE, $cog) ) {print "Cannot open file \"$cog\"\n\n";}
     my $count = 0;
     while(<FILE>)
     {
```

```
        $count++;
    }
    return $count;
}

sub find_tree
{
    my $data = shift(@_);
    opendir (DIR, $data);
    my @dir = readdir(DIR);
    closedir(DIR);
    my $tree;
    foreach(@dir)
    {
        next if($_ =~ /^\./);
        if ($_ =~ m/\.tree/)
        {
            $tree = $_;
        }
    }
    return "$data/$tree";
}
sub check_arguments
{
    if (!defined $_[0]||!defined $_[1])
    {
        print "USAGE: cog_tree_integration_v1 dataDir outDir\n
        dataDir is a folder which contains\n
        #rooted tree file in newick format ended with .tree\n
        #list.strains.01.development\n
        #list.strains.02.development <<<----fixed name incase of organism's name > 31
characters\n
        #gene_table.protein.tab.development (COGs table)\n
        outDir is the destination folder for storing results";
        die;
    }
    return ";
```

CURRICULUM VITAE

NAME	Mr. Sivamoke Dissook
DATE OF BIRTH	14 September 1989
EDUCATIONAL RECORD	
HIGH SCHOOL	Wichai Wittaya Bilingual School, 2007
BACHELOR'S DEGREE	Bachelor of Science (Chemistry) Chiangmai University, 2011
MASTER'S DEGREE	Master of Science (Bioinformatics and Systems Biology) King Mongkut's University of Technology Thonburi, 2013
SCHOLARSHIP	Full scholarship, Master's Degree in Bioinformatics and Systems Biology King Mongkut's University of Technology Thonburi and National Center of Genetic Engineering and Biotechnology (BIOTEC), Thailand, 2012 – 2013
PUBLICATION	Dissook, S., Kittichotirat, W., Cheevadhanarak, S., Senachak, J. and Prommeenate, P., 2014, "Characterization of cyanobacteria gene content variation using comparative genomic", **Next Generation Sequencing Conference (NGS2014)**, 29-30 July 2014, Bangkok, Thailand.