GMR
Genetics and Molecular Research
ISSN 16765680

Gene annotation and functional analysis of a newly sequenced *Synechococcus* strain

Y. Li, N.N. Rao, Y. Yang, Y. Zhang and Y.N. Gu

Center for Information in BioMedicine, School of Life Science and Technology,
University of Electronic Science and Technology of China, Chengdu, China

Corresponding author: N.N. Rao
E-mail: raonn@uestc.edu.cn

Genet. Mol. Res. 14 (4): 12416-12426 (2015)
Received May 8, 2015
Accepted July 29, 2015
Published October 16, 2015
DOI http://dx.doi.org/10.4238/2015.October.16.8

ABSTRACT. *Synechococcus* sp PCC 7336 represents a newly sequenced strain, and its genome is obviously different from that of other *Synechococcus* strains. In this analysis, local alignment and annotation databases were constructed and combined with various bioinformatic tools to carry out gene annotation and functional analysis of this strain. From this analysis, we identified 5096 protein-coding genes and 47 RNA genes. Of these, 116 genes that were classified into 9 categories were associated with photosynthesis, and type V polymerase proteins that were identified are unique for this strain. An additional 107 genes were closely related to signal transduction pathways, which primarily comprised parts of two-component regulatory systems. Gene ontogeny analysis showed that 2377 genes were annotated with a total number of 9791 functional categories, and specifically that 41 genes distributed in 4 protein complexes were involved in oxidative phosphorylation. Clusters of orthologous groups classification showed that there were 1463 homologous proteins associated with 17 specific metabolic pathways, and that most of the proteins participated in primary metabolic processes such as binding and catalysis. The phylogenetic tree based on 16S rRNA sequences indicated that *Synechococcus* PCC 7336 is highly likely to represent a new branch.

Key words: *Synechococcus*; Gene; Function; Metabolism; Phylogeny

INTRODUCTION

The study of microbial genomics has developed rapidly over the past 20 years from gene cloning to genetic engineering. For instance, 16s rRNA plays an important role as a molecular marker in the classification and identification of microorganisms, which has promoted the study of phylogenetics to a new level. In the early stages of phylogenetic research, scientists primarily focused on housekeeping genes because of the inadequacy of information on functional genes and methodologic limitations. In comparison, whole genome sequencing technology brings us to a deep understanding of microbial genomes; this field has been termed species genomics. The whole genome sequencing and publication of *Haemophilus influenzae* in 1995 was a milestone, which drove traditional biology into large-scale, high flux research (Fleischmann et al., 1995). To date, the number of genomes recorded in the National Center for Biotechnology Information (NCBI) microbial genome database has reached 5696, and the nucleotide sequences contained therein number more than 2.4 million. Gene, RNA, and protein research requires a large quantity of sequencing data, and microbial genomics can provide support for these areas.

Synechococcus is a unicellular cyanobacterium that is very widespread in marine environments. It was first described in 1979 (Waterbury et al., 1979) and was originally defined to include “small unicellular cyanobacteria with ovoid to cylindrical cells that reproduce by binary traverse fission in a single plane and lack sheaths. *Synechococcus*is one of the most important components of the prokaryotic autotrophic picoplankton in the temperate to tropical oceans, and scientists have studied it over many years. Recently, two orthogonal constitutive promoter libraries were built and tested for *Synechococcus* sp PCC 7002 (Markley et al., 2014). These promoter libraries were then combined to create and optimize a series of IPTG inducible cassettes. The results demonstrated that a synthetic biology tool box could enable the accelerated engineering of *S.* PCC 7002. In another study, the genome-wide metabolic network of *S. elongatus* PCC 7942 was reconstructed by metabolic flux simulation models, and the applicability of the model was demonstrated by simulating the autotrophic growth conditions of this strain (Triana et al., 2014). Comparatively, parallel microarray-based analysis of gene expression and gene knockout experiments were conducted for *Synechococcus* sp WH 8102, and the results showed that the sodT::sodB exconjugants were unable to grow at low Ni concentrations while the sodN::sodB exconjugants displayed higher growth rates at low Ni concentrations than did the wild type(Dupont et al., 2012).The genome sizes of the common *Synechococcus* strains are less than 3Mb and the numbers of genes are less than 3000. However, the genome of the newly sequenced strain *S.*PCC7336 is over 5Mb, and is therefore likely to carry unique structures and functional genes. In order to explore the molecular characteristics of *S.* PCC 7336, gene prediction and annotation were initially performed, then gene function, classification, and regulatory pathways were examined. Finally, a phylogenetic tree was constructed to illustrate the process of evolution of this strain. We expect that the results from these analyses will provide insight into this species.

MATERIAL AND METHODS

Materials

The shotgun sequencing data of *S.* PCC 7336 were downloaded from NCBI (accession No. NZ_ALWC01000000); all of the 16S rRNA sequences utilized in this study were also obtained from the NCBI database.

Gene prediction and annotation

Open reading frames were predicted by GeneMarkS (Besemer et al., 2001). Then, the non-redundant database packages local_b2g_db.zip, go_2013011-assocdb-data.gz, gene_info.gz, gene 2 accession.gz, b2g4pipe_v2.5.zip, and idmapping.tb.gz were downloaded from NCBI and a local blast2go database was constructed for open reading frame (ORF) annotation. All the ORFs were concomitantly submitted to SWISS-PROT, TREMBL, CDD, PFAM, and Cog databases to generate multiple alignments, and the annotation results were merged for similarity (>30% and e <1 e^{-5}).

Gene function classification

GO classification was carried out for all the predicted genes, and their statistical distributions among cellular components, molecular functions, and biological pathways were gathered. COG classification was subsequently performed to determine the numbers and functions of homologous genes.

Regulatory pathway analysis

The Kegg orthology (KO) numbers of the genes were predicted by KAAS (Moriya et al., 2007), and the KO numbers were mapped to the corresponding pathways of the KEGG database. Multiple alignments and the nonribosomal peptide synthetases and polyketide synthases (NRPS-PKS) database were combined to predict the secondary metabolic gene clusters.

Phylogenetic analysis

We chose 47 representative 16S rRNA sequences of cyanobacteria to generate a phylogenetic tree using MEGA6.0 (Tamura et al., 2013), and the evolutionary characteristics of *S.* PCC 7336 were analyzed.

RESULTS

Gene prediction and annotation

An ORF in prokaryotes can be viewed as a protein coding gene; 5096 coding and 47 RNA genes were obtained from *S.* PCC 7336. The total length of coding gene sequence was determined to be 4,251,462 bp, and the average coding length was 834 bp; the coding rate was 82.73%. The rRNA genes were found to consist of two 4925 bp clusters that were comprised of 5S, 16S, and 23S RNA, along with Ala- and Ile-tRNA, but these clusters were ascertained to be distinct copies because of their differing orders of gene arrangements. This structure differs from that of other *Synechococcus* that have two completely identical copies of the rRNA cluster. 43 tRNA have 40 different functions, most of which exist in dispersive arrangements and thus do not form a gene cluster. The GC content of all the RNA genes is not high and is consistent with that of the entire genome. The results of gene annotation are shown in Table 1. We obtained 4470 known genes using a local non-redundant database while the identification of annotated genes using the other protein databases was relatively less, which might be attributed to different sequence formats in various databases as well as to the inclusion of a certain number of non-coding RNA genes.

Table 1. Gene annotation in different databases.

	Gene		Database					
	RNA	Protein-coding	NR	SWISS-PROT	CDD	PFAM	TREMBL	Cog
Number	47	5096	4470	2174	2665	2285	3977	1953
Ratio (%)			87.72	42.66	52.3	44.84	78.04	38.32

Protein-coding genes

Following our analysis, 116 genes were found to be related to photosynthesis; these can be divided to 9 families including photosystems I and II, ATP synthase, CO_2 fixation, NADH dehydrogenase, phycobilisome, cytochrome oxidase, cytochrome b6lf complex, and electron carriers. The PsaABCDEFL and PsbABCDEFHOPUV clusters primarily participate in photosynthesis. Genes in the Pet family have two different roles; those in the PetABCDG cluster encode the cytochrome b6lf complex whereas the PetEFHJ cluster is associated with the electron transport chain. ATPase polymerase complex genes were classified into types F and V. Type F is comprised of 8 subunits: alpha, beta, gamma, delta, a, b, c, and epsilon; while type V constitutes 7 subunits: A-E, I, and K. The type V polymerase proteins in *S.* PCC 7336 are unique compared with those in other cyanobacteria. Type V proteins are highly conservative polymerase proteins that are rare in cyanobacteria while are widely distributed in eukaryotic organisms such as plants. The formation of plant chloroplasts is closely related to cyanobacteria and plants obtain their photosynthetic ability through a symbiotic relationship with cyanobacteria. Antenna proteins are primarily involved in light absorption and 3 clusters of antenna proteins were found in *S.* PCC 7336 including ApcABDEF for phycocyanobilin, CpcABCDEFG for phycocyanin, and CpeABCDRSTUYZ for phycoerythrin. All of the antenna proteins participate in photosystem II reactions. However, although *S.* PCC 7336 has the same type V polymerase proteins as do plants, it lacks the light concentration complex LHCI, and its photosystem I protein complex presents as a polymer instead of in single molecule formation, which demonstrates that the photosystem of *S.* PCC 7336 differs from those of other cyanobacteria as well as from those of plants.

The two-component regulatory system is the primary means of providing a basic stimulus-response coupling mechanism to allow cyanobacteria to sense and respond to changes in many different environmental conditions. In *S.* PCC 7336, a total of 107 related genes were found including 16 normal histidine kinases, 7 diguanylate cyclases/phosphodiesterases, 6 cyclic adenosine monophosphate kinases, 1 circadian rhythm kinase, 1 polyphosphate kinase, and 1 CO_2 induced protein. In addition, 55 regulators were identified that contained 39 transcriptional regulators, 4 two-component regulators, 4 DNA-binding regulators, 5 kinase regulators, and specific regulators such as AbrB, Crp/Fnr, and HxlR. These kinases and regulators can stimulate gene expression in response to environmental stresses, indicating that *S.* PCC 7336 has flexible regulatory mechanisms.

Metabolic pathways analysis

As indicated by the results of the GO program, 2377 genes were annotated with a total of 9791 functional categories. The distribution of genes in each category can be seen in Figure 1. On the cellular components level, gene products were mainly distributed in the cytoplasm or the cell membrane. The cell membrane is important to cyanobacteria as material exchange between the inside and outside of the cell is entirely reliant upon the cell membrane channels. On the molecular function level, the gene products were found to be primarily involved in binding and catalytic processes. Of these, 1589 genes were seen to participate in catalytic process, accounting for almost half of all functional genes. The remaining 1388 genes participated in the binding process. In addition to those involved in standard nucleic acid and protein binding, the remainder of the genes appeared to play a role in as many as 130 binding processes such as RNA polymerase binding, specific motifbinding, and ATP binding. On the biological process level, most of the gene products mainly participated in primary metabolic processes, and the remaining genes were found to be involved in secondary metabolic pathways. Cyanobacteria is strongly adaptable to diverse environments including oceans, fresh water, soil, desert, and polar regions; therefore flexible and changeable regulatory pathways area guarantee of viability, and provide specific response mechanisms that can be activated when the organisms are faced with various challenges.

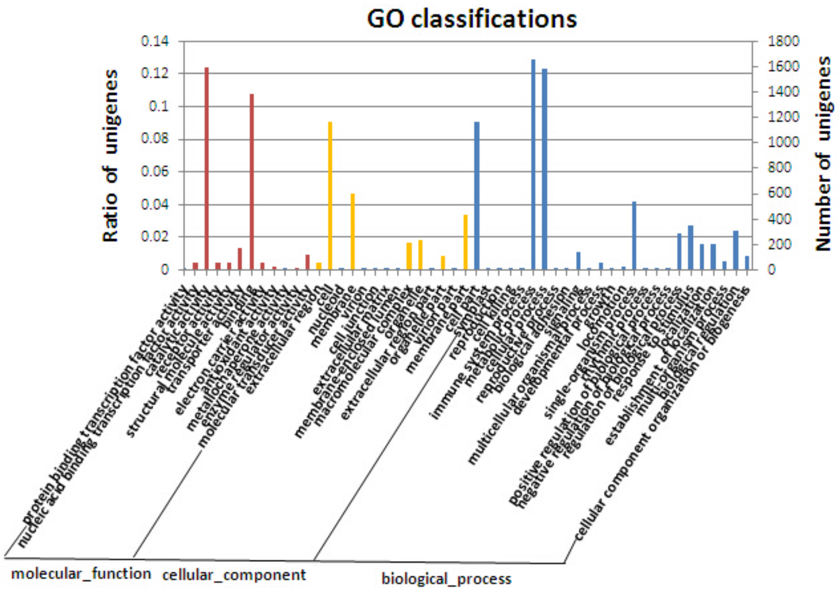

Figure 1. GO classification of genes in *Synechococcus* PCC7336. 2377 genes were divided into 12 groups in molecular function, 16 groups in cellular component and 23 groups in biological process respectively.

Synthesis of secondary metabolites

Cyanobacteria contribute a significant source of bioactive substance, the macromolecular nitrogen compounds which are synthesized primarily by the multi-modular PKS or NRPS systems. PKS has three types of structure: typeI-modular, typeII-iterative or aromatic and typeIII-chalcone synthase (Funa et al., 1999), and NRPS also has three types: type A- linear, type B-iterative and type C-nonlinear (Mootz et al., 2002). Comparison and feature analysis between *S.* PCC 7336 genes and those from strains containing NRPS/PKS modules in the database reveals a total of 11 secondary metabolite clusters, the coordination and structure of which can be seen in Figure 2. The secondary metabolites are primarily synthesized through the PKS pathway, with only one being synthesized through the NRPS pathway and another through the NRPS-TLPKS hybrid pathway. The longest NRPS-TLPKS hybrid gene cluster has a length of 99,552 bp and contains 46 genes. The NRPS/PKS gene cluster consists of four parts: synthetic, transport, regulatory, and other genes. The synthetic genes are the core of the whole synthesis system since the type of proteins encoded by them usually determines the structure and function of the secondary metabolites. On the other hand, the transport or regulatory genes are not necessary for these clusters. The synthesis of secondary metabolites is also regulated by global regulators. The No.11 NRPS-TLPKS hybrid cluster encodes catalytic enzymes with its first and last synthetic genes, and the other 11 genes encode specific amino acids. In this study, the NRPSPredictor2 SVM program (Röttig et al., 2011) was used to predict the products of the NRPS-TLPKS pathway, and the arrangements were listed as (pro) + (mal) + (asp) + (asp-gly) + (gly) + (mal) + (leu-lys) + (mal) + (pk-mal). Their physiological and biochemical characteristics can be determined through HPLC or ELISA. To date, over 3000 kinds of secondary metabolites in cyanobacteria have been identified, most of which have been identified as microcystins due to their toxicity. These are also harmful to other organisms during outbursts of water bloom. Our results illustrated the diversity of secondary metabolite clusters in *S.* PCC 7336. It is meaningful for environmental management to explore the synthesis mechanisms and structure of such metabolites.

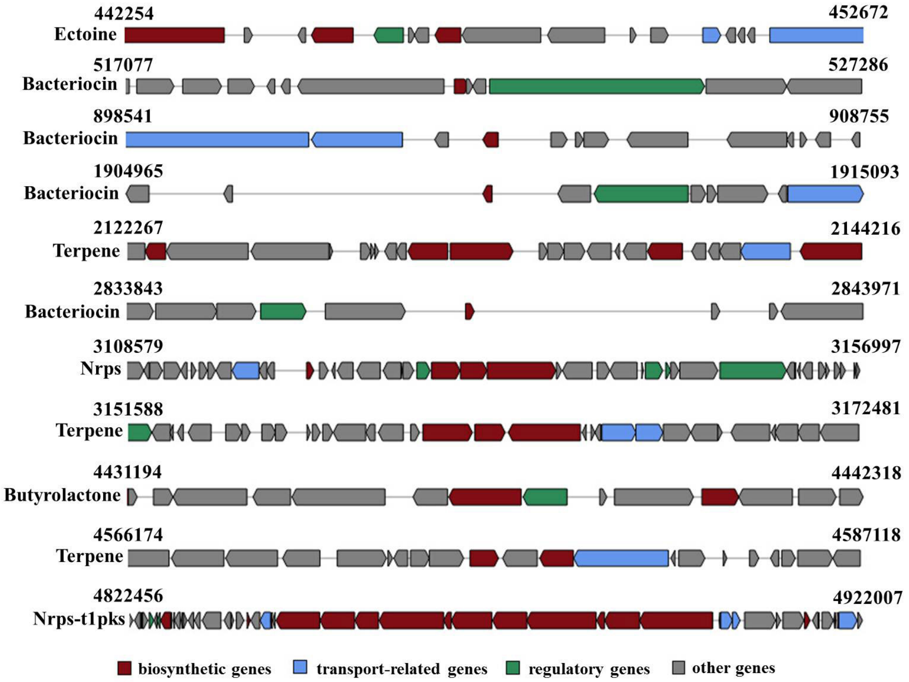

Figure 2. Locations and structures of secondary metabolite clusters in *Synechococcus* PCC7336.

Oxidative phosphorylation

Oxidative phosphorylation is a vital part in the respiration process of cyanobacteria.

Although the structure and function of the electron transfer chain in cyanobacteria and those in eukaryotes are similar, more diverse enzymes and substrates are involved in the electron transport chain in cyanobacteria (Nealson, 1999). Cyanobacteria utilize a large variety of substances to provide and accept electrons, which enables their survivalin various environments. Many reductants and oxidants can activate oxidative phosphorylation. The neutral-point of chemicals can be used to measure the energy released by oxidation or reduction reactions. As for eukaryotes, cyanobacteria utilize the energy released by substrate oxidation to pump ions out of the cell membrane to create an electrochemical gradient. There are 5 protein compounds involved in the oxidative phosphorylation process in cyanobacteria: the NADH dehydrogenase, fumarate reductase, cytochrome reductase, cytochrome oxidase, and ATP synthase complexes. We mapped the annotated genes of *S.* PCC7336 onto these 5 complexes and discovered that the genes were present in all complexes except for the cytochrome reductase complex (Figure 3). Of these, 11 genes, including 8 oxidoreductase subunits and 3 dehydrogenase subunits, were found to take part in the NADH dehydrogenase process (EC: 1.6.5.3). An additional 2subunit-coding genes of succinate dehydrogenase, SdhA and SdhB, were found to be involved in the process of transferring succinic acid into fumaric acid (EC: 1.3.99.1). A group of 7 genes in the cytochrome oxidase complex (EC: 1.9.3.1) were shown to promote substrate decomposition into oxygen, hydrogen ions, and water. In addition, these genes also play a role in the two-component regulatory system. Both types F and V subunits are involved in the synthesis of ATP, which is correspondent with the analysis of photosynthesis mentioned above. Furthermore, the previous studies indicated that oxidative phosphorylation in cyanobacteria possesses unique characteristics. For example, certain strains of cyanobacteria contain two complete but totally different electron transfer chains located on the surface of the cell membrane or the thylakoid.

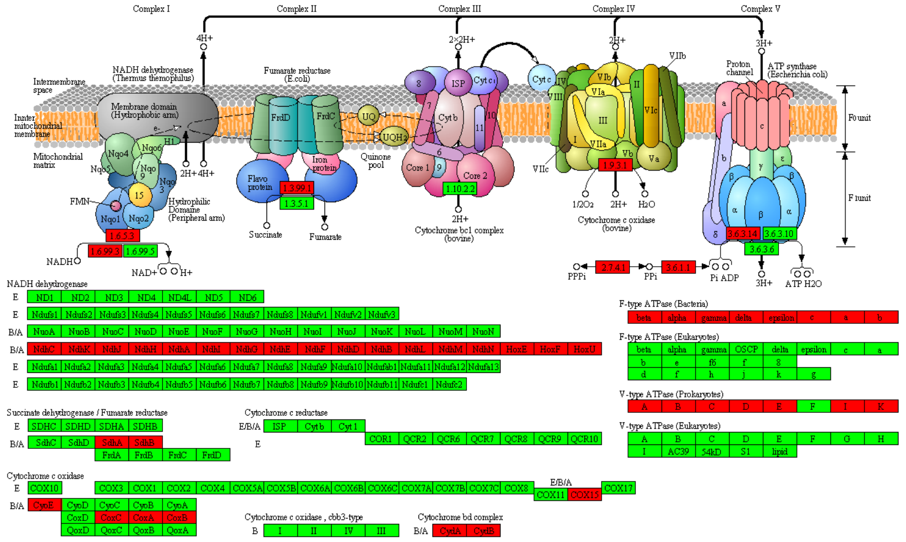
Complex I
Complex II
Complex III
Complex IV
Complex V
4H+
2×2H+
2H+
3H+
Intermembrane space
Innter mitochondrial membrane
Mitochondrial matrix
NADH dehydrogenase (Thermus themophilus)
Membrane domain (Hydrophobic arm)
e-
H1
Nqo4
Nqo6
Nqo5
Nqo 9
Nqo 3
15
FMN
Nqo1
Nqo2
Hydrophilic Domaine (Peripheral arm)
2H+ 4H+
1.6.5.3
1.6.99.3
1.6.99.5
NADH
NAD+ H+
Fumarate reductase (E.coli)
FrdD
FrdC
UQ
UQH2
Quinone pool
Flavo protein
Iron protein
1.3.99.1
1.3.5.1
Succinate
Fumarate
8
ISP
Cyt c1
7
Cyt b
10
11
6
Core 1
9
Core 2
1.10.2.2
2H+
Cytochrome bc1 complex (bovine)
Cyt c
IV
VIII
VIb
VIa
II
III
VIc
I
VIIa
Vb
Va
VIIb
VIIc
1.9.3.1
1/2O2
2H+
H2O
Cytochrome c oxidase (bovine)
Proton channel
ATP synthase (Escherichia coli)
a
c
b
γ
ε
β
α
δ
3.6.3.14
3.6.3.10
3.6.3.6
F0 unit
F1 unit
PPPi
2.7.4.1
PPi
3.6.1.1
Pi ADP
3H+
ATP H2O
NADH dehydrogenase
E ND1 ND2 ND3 ND4 ND4L ND5 ND6
E Ndufs1 Ndufs2 Ndufs3 Ndufs4 Ndufs5 Ndufs6 Ndufs7 Ndufs8 Ndufv1 Ndufv2 Ndufv3
B/A NuoA NuoB NuoC NuoD NuoE NuoF NuoG NuoH NuoI NuoJ NuoK NuoL NuoM NuoN
B/A NdhC NdhK NdhJ NdhH NdhA NdhI NdhG NdhE NdhF NdhD NdhB NdhL NdhM NdhN HoxE HoxF HoxU
E Ndufa1 Ndufa2 Ndufa3 Ndufa4 Ndufa5 Ndufa6 Ndufa7 Ndufa8 Ndufa9 Ndufa10 Ndufab1 Ndufa11 Ndufa12 Ndufa13
E Ndufb1 Ndufb2 Ndufb3 Ndufb4 Ndufb5 Ndufb6 Ndufb7 Ndufb8 Ndufb9 Ndufb10 Ndufb11 Ndufc1 Ndufc2
Succinate dehydrogenase / Fumarate reductase
E SDHC SDHD SDHA SDHB
B/A SdhC SdhD SdhA SdhB
FrdA FrdB FrdC FrdD
Cytochrome c reductase
E/B/A ISP Cytb Cyt1
E COR1 QCR2 QCR6 QCR7 QCR8 QCR9 QCR10
Cytochrome c oxidase
E COX10 COX3 COX1 COX2 COX4 COX5A COX5B COX6A COX6B COX6C COX7A COX7B COX7C COX8
E/B/A COX11 COX15 COX17
B/A CyoE CyoD CyoC CyoB CyoA
CoxD CoxC CoxA CoxB
QoxD QoxC QoxB QoxA
Cytochrome c oxidase , cbb3-type
B I II IV III
Cytochrome bd complex
B/A CydA CydB
F-type ATPase (Bacteria)
beta alpha gamma delta epsilon c a b
F-type ATPase (Eukaryotes)
beta alpha gamma OSCP delta epsilon c a
b e f6 f 8
d f h j k g
V-type ATPase (Prokaryotes)
A B C D E F I K
V-type ATPase (Eukaryotes)
A B C D E F G H
I AC39 54kD S1 lipid

Figure 3. Functions of genes involved in oxidative phosphorylation. The whole process of oxidative phosphorylation was shown. Complex I: NADH dehydrogenase; Complex II: fumaratereductase; Complex III: cytochrome bc1 complex; Complex IV: Cytochrome c oxidase; Complex V: ATP synthase. The lower part lists all of the genes involved in the oxidative phosphorylation. Those found only in *Synechococcus* PCC7336 were marked in red and their functions correspond to the red marked loci on the upper part. The detail reactions could be found in KEGG database through the number.

Function and classification of homologous proteins

All proteins within COGs are presumed to derive from a single ancestor and can be divided into orthologs and paralogs. The orthologous proteins evolving from vertical lines typically maintain the same function as the irancestor, whereas paralogous proteins are derived from gene duplication and are likely to have gained new functions. The results of COG classification showed that a total of 1463 proteins were involved in 17 specific metabolic processes in *S.* PCC 7336 (except for those with unknown functions); only one gene was shown to take part in cell movement (Figure 4). This finding is reasonable as *Synechococcus*, a unicellular strain, lacks flagellum to ensure its mobility. The separation between *S.* PCC 7336 and other strains demonstrated that the genetic material of *S.* PCC 7336 underwent little exchange with that of other organisms (Stucken et al., 2013), which suggested a relatively independent metabolic regulation and specific living environments compared to those of other cyanobacterial strains. Proteins gathered in type J, E, L, and O corresponded to binding and catalytic processes, which suggested that primary regulatory processes such as DNA duplication and translation are critical for *S.*PCC7336. However, the homologous to annotated protein ratio in *S.* PCC7336 is low compared with that of the other *Synechococcus* strains, and the remaining unknown proteins need to be examined by other methods.

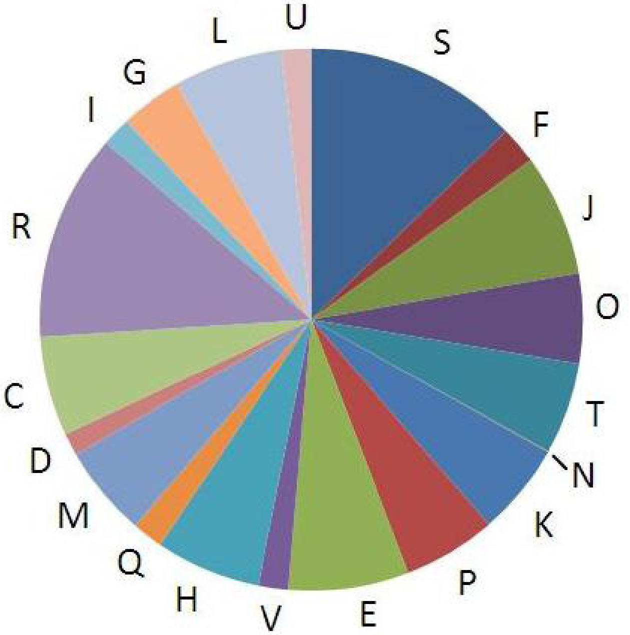

Figure 4. COG classification of genes. The regulatory pathways represented by each letter are as listed below. E: Amino acid transport and metabolism; C: energy manufacture and conversion; J: translation and ribosome biogenesis; R: predicted general function; L: replication, recombination, and repair; H: coenzyme transport and metabolism; G: transport and metabolism of carbohydrates; P: inorganic ion transport and metabolism; O: post transcriptional modification, protein turnover, and molecular chaperone; T: signal transduction; M: cell wall and cell membrane formation; I: lipid transport and metabolism; S: unknown function; K: transcription; F: transport and metabolism of nucleotide; N: cell movement; V: defense mechanism; U: intracellular transport, material secretion, and vesicular transport; Q: synthesis, transport, decomposition, and metabolism of secondary product; D: cell cycle control, cell division, and chromosome distribution.

Phylogenetic analysis

16S ribosomal RNA is a component of the 30S small subunit of prokaryotic ribosomes. It is of moderate size and consists of both highly conserved and highly variable domains, the latter of which is commonly used in phylogeny research. We selected 47 16S rRNAs for use in constructing a phylogenetic treere presenting all the *Synechococcus* strains as well as other representative strains; the results are shown in Figure 5. We can see that the phylogenetic discrepancy of *Synechococcus* is greater than that of *Prochlorococcus* though both belong to marine cyanobacteria. All the *Prochlorococcus* strains were classified as one cluster, which did not occur for *Synechococcus*. *S.* PCC7336 formed a single branch which is consistent with its differences in genomic size, gene number, and protein classification as mentioned above. The evolutionary processes of *Synechococcus* sp JA-2-3B'a (2-13), *Synechococcus* sp JA-3-3Ab, and *Gloeobacter violaceus* PCC 7421 are most closely related to *S.* PCC 7336. So far the studies on *S.* JA-2-3B'a (2-13) and *S.* JA-3-3Ab have been almost blank, and a study of *G. violaceus* PCC7421 showed that it has some similar characteristics with *S.* PCC7336. The genome of *G. violaceus* PCC 7421 is 4,659,019 bp, which covers 4430 coding genes and 45 tRNA genes. There are only 610 homologous proteins in *G. violaceus* PCC 7421, even less than in *S.* PCC7336. Furthermore, *G. violaceus* PCC 7421 is the only strain lacking a thylakoid membrane, and its phycobilisomes are located in the cytoplasm (Nakamura et al., 2003). Both of these strains belong to unicellular gram negative bacteria without regulatory genes in their photosystems. We presumed from the tree that the 16S rRNA sequence of*S.*PCC7336 might have changed due to mutation or other factors that caused it to evolve into a unique branch, or that it actually represents a new subspecies only discovered recently. The evolutionary process of *S.* PCC 7336 remains to be further studied through additional biological evidence.

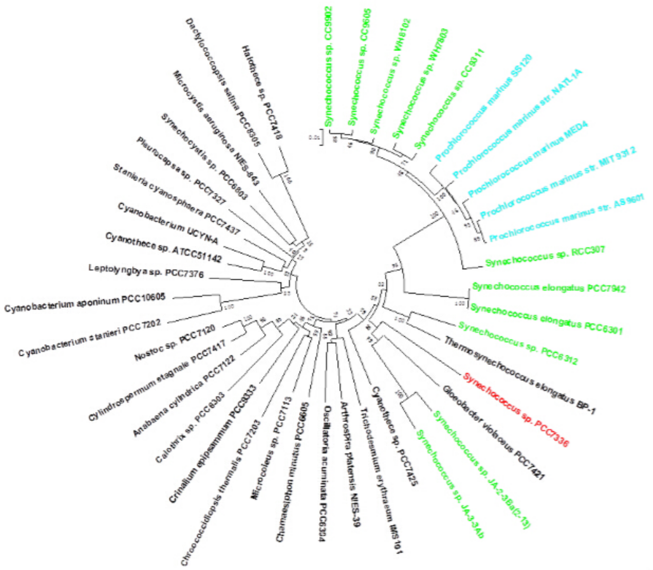
Synechococcus sp. CC9902
Synechococcus sp. CC9605
Synechococcus sp. WH8102
Synechococcus sp. WH7803
Synechococcus sp. CC9311
Prochlorococcus marinus SS120
Prochlorococcus marinus str. NATL1A
Prochlorococcus marinus MED4
Prochlorococcus marinus str. MIT 9312
Prochlorococcus marinus str. AS9601
Synechococcus sp. RCC307
Synechococcus elongatus PCC7942
Synechococcus elongatus PCC6301
Synechococcus sp. PCC6312
Thermosynechococcus elongatus BP-1
Synechococcus sp. PCC7336
Gloeobacter violaceus PCC7421
Synechococcus sp. JA-2-3Ba(2-13)
Synechococcus sp. JA-3-3Ab
Cyanothece sp. PCC7425
Trichodesmium erythraeum IMS101
Arthrospira platensis NIES-39
Oscillatoria acuminata PCC6304
Chamaesiphon minutus PCC6605
Microcoleus sp. PCC7113
Chroococcidiopsis thermalis PCC7203
Crinalium epipsammum PCC9333
Calothrix sp. PCC6303
Anabaena cylindrica PCC7122
Cylindrospermum stagnale PCC7417
Nostoc sp. PCC7120
Cyanobacterium stanieri PCC7202
Cyanobacterium aponinum PCC10605
Leptolyngbya sp. PCC7376
Cyanothece sp. ATCC51142
Cyanobacterium UCYN-A
Stanieria cyanosphaera PCC7437
Pleurocapsa sp. PCC7327
Synechocystis sp. PCC6803
Microcystis aeruginosa NIES-843
Dactylococcopsis salina PCC8305
Halothece sp. PCC7418

Figure 5. The phylogenetic tree based on 16S rRNA. *Prochlorcoccus* strains are marked with blue; *Synechococcus* strains are marked with green; *Synechococcus* PCC7336 is marked with red. Neighbor-joining method was used and the numbers on each node are bootstrap values of 1000 replicates, which indicate homologous degree between different strains.

DISCUSSION

In this study, we utilized methods derived from bioinformatics to conduct gene prediction and functional analysis of *S.* PCC7336, and identified 5096 coding and 47 RNA genes. Of the coding genes, 2541 (49.8%) are protein coding. We focused on the genes involved in photosynthesis, signal transduction, and secondary metabolism because they are vital to cell activity. According to our results, *S.* PCC7336 has unique gene functions and types, which provide a reference for further experiments. For example, *S.* PCC 7336 has a variety of specific regulators, and the interactions between these regulators and targets can be a potential research direction. Through these regulatory relationships, we might discover the differences in cell activities between *S.* PCC7336 and other strains. The synthesis of secondary metabolites is always a hotspot of research. It is important to figure out such synthesis mechanisms in order to improve our environment, as we are now faced with the challenge of water blooms of cyanobacteriain large areas. The phylogenetic tree based on 16S rRNA sequence reveals the evolutionary process of *S.* PCC7336 which differs from that of other *Synechococcus* strains but shares similarity with *G. violaceus* PCC 7421. It is therefore plausible that *S.* PCC7336represents a new branch on the phylogenetic tree when all of the evidence is combined.

In conclusion, our results demonstrated that the genomic size and gene number of *S.* PCC 7336 is almost twice that of other *Synechococcus* strains, and that it contains specific regulators across various family groups. GO and COG classification showed that most of its coding genes participate in primary regulatory pathways, and the construction of a phylogenetic tree indicated that *S.* PCC 7336 is likely to be an independent branch.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by the NSFC (National Natural Science Foundation of China, Grant #81171411) and I would like to thank all members of the Neuro-Information of the Ministry of Education Genomic and Medical Informatics Lab for useful comments and discussions.

REFERENCES

Besemer J, Lomsadze A and Borodovsky M (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res*. 29: 2607-2618.

Dupont CL, Johnson DA, Phillippy K, Paulsen IT, et al. (2012). Genetic identification of a high-affinity Ni transporter and the transcriptional response to Ni deprivation in *Synechococcus* sp. strain WH8102. *Appl. Environ. Microbiol*. 78: 7822-7832.

Fleischmann RD, Adams MD, White O, Clayton RA, et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269: 496-512.

Funa N, Ohnishi Y, Fujii I, Shibuya M, et al. (1999). A new pathway for polyketide synthesis in microorganisms. *Nature* 400: 897-899.

Markley AL, Begemann MB, Clarke RE, Gordon GC, et al. (2014). A synthetic biology toolbox for controlling gene expression in the cyanobacterium *Synechococcus* sp. PCC 7002. *ACS Synth. Biol*. [Epub Ahead of Print].

Mootz HD, Schwarzer D and Marahiel MA (2002). Ways of assembling complex natural products on modular nonribosomal peptide synthetases. *Chembiochem* 3: 490-504.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, et al. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res*. 35: W182-W185.

Nakamura Y, Kaneko T, Sato S, Mimuro M, et al. (2003). Complete genome structure of *Gloeobacter violaceus* PCC 7421, a cyanobacterium that lacks thylakoids. *DNA Res*. 10: 137-145.

Nealson KH (1999). Post-Viking microbiology: new approaches, new data, new insights. *Origins Life Evol. Biosph*. 29: 73-93.

Röttig M, Medema MH, Blin K, Weber T, et al. (2011). NRPS predictor2-a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res*. 39: W362-367.

Stucken K, Koch R and Dagan T (2013). Cyanobacterial defense mechanisms against foreign DNA transfer and their impact on genetic engineering. *Biol. Res*. 46: 373-382.

Tamura K, Stecher G, Peterson D, Filipski A, et al. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol*. 30: 2725-2729.

Triana J, Montagud A, Siurana M, Fuente D, et al. (2014). Generation and evaluation of a genome-scale metabolic network model of *Synechococcus elongatus* PCC7942. *Metabolites* 4: 680-698.

Waterbury JB, Watson SW, Guillard RR and Brand LE (1979). Widespread occurrence of a unicellular, marine, planktonic, cyanobacterium. *Nature* 277: 293-294.