Proteomics Assisted Genome Sequence of *Cyanothece* 51142, a Unicellular N_2-fixing Cyanobacterium

Eric A. Welsh[1], Michelle Liberton[1], Jana Stöckel[1], Hanayo Sato[1], Thomas Loh[1], Chunyan Wang[2], Aye Wollam[2], Robert S. Fulton[2], Sandra W. Clifton[2], Jon M. Jacobs[3], Rajeev Aurora[4], **Louis A. Sherman[5]**, Richard D. Smith[3], Richard K. Wilson[2] and Himadri B. Pakrasi[1]

[1]Department of Biology, Washington University, St. Louis, MO 63130, USA; [2]Genome Sequencing Center, Washington University School of Medicine, St. Louis, MO 63110, USA; [3]Pacific Northwest National Laboratory, Richland, WA 99352, USA; [4]Department of Molecular Microbiology & Immunology, Saint Louis University School Of Medicine, St. Louis, MO 63103, USA; [5]Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA.

Cyanothece sp. ATCC 51142 is a unicellular diazotrophic marine cyanobacterium capable of separating the incompatible processes of oxygenic photosynthesis and nitrogen fixation temporally within the same cell. It performs photosynthesis during the day and nitrogen fixation at night. The genome of *Cyanothece* 51142 was sequenced and found to contain a unique arrangement of one large circular chromosome, four small plasmids, and one linear chromosome, the first report of such a linear element in a photosynthetic bacterium.

Phylogenetic analysis suggests that nitrogen fixation is an ancient process that arose early in evolution and has subsequently been lost in many cyanobacterial strains. During a diurnal period, *Cyanothece* cells actively accumulate and degrade different storage inclusion bodies for the products of photosynthesis and N_2-fixation.

This ability to utilize metabolic compartmentalization and energy storage makes *Cyanothece* an ideal system for bioenergy research, as well as studies of how a unicellular organism balances multiple, often incompatible, processes in the same cell.

The Genome of *Cyanothece* sp. PCC 51142

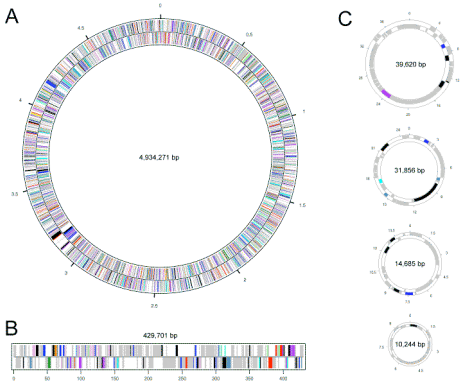


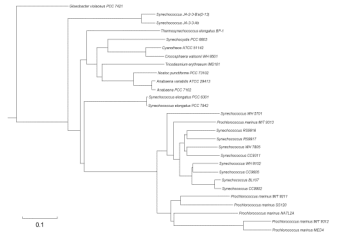
Figure 1. Distribution of gene functions within the *Cyanothece* genome. Labels are given in Mb pairs for the circular chromosome (A), and Kb pairs for the linear chromosome (B) and plasmids (C). Genes are colored by functional category as follows: ncRNA (yellow); energy, fatty acid, phospholipid metabolism (red); cell envelope (orange); cellular processes (steel blue); central intermediary metabolism (light green); photosynthesis and respiration (dark green); regulation (cyan), DNA, transcription, translation (dark blue), small molecule biosynthesis (magenta), transport and binding (purple), unknown/hypothetical (grey), other (black). The two rRNA operons on the circular chromosome are indicated in yellow at 3.95 Mbp and 4.10 Mbp (A). The glycolysis/fermentation-related cluster on the linear chromosome is shown in light green and red between 375 kbp and 400 kbp (B).

The genome of *Cyanothece* 51142 was sequenced at the Washington University Genome Sequencing Center (St. Louis, MO) and the final assembly independently confirmed using an optical restriction map generated by OpGen, Inc. (Madison, Wisconsin). The 5.5 Mb *Cyanothece* genome consists of a 4.93 Mb circular chromosome, four plasmids ranging in size from 10 kb to 40 kb, and notably, a 430 kb linear chromosome (Fig. 1, Table 1). The annotation of genes of unknown function was greatly aided by data from a high-throughput proteomics analysis, covering approximately 51% of the predicted proteome, with 24% of all hypothetical genes re-annotated as unknowns due to their observation in the proteomic library. Additionally, peptides from 53 low confidence open reading frames were identified that would not have otherwise been included in the final genome annotation. The combined analysis of proteome and genome data is an important new approach that resulted in the inclusion or reclassification of 510 genes, and lent an additional level of validation to the genome annotation.

Table 1. General features of the *Cyanothece* 51142 genome. Both the rRNA operons are located on the circular chromosome, as well as all 41 tRNAs. Of the two cobalamin riboswitches present in the genome, one is located on the linear chromosome, the only non-coding RNA outside of the circular chromosome.

	Chrom 1	(%)	Chrom 2	(%)	pCT42a	(%)	pCT42b	(%)	pCT42c	(%)	pCT42d	(%)
Size (bp)	4934270		429625		39620		31856		14685		10244	
G+C content (%)	37.9		38.6		36.8		41.5		38.1		37.0	
Protein-coding genes	4738	100.0	438	100.0	37	100.0	23	100.0	20	100.0	13	100.0
Assigned/Probable function	1719	36.3	72	16.4	3	8.1	5	21.7	3	15.0	1	7.7
Possible function	1403	29.6	112	25.6	10	27.0	7	30.4	2	10.0	3	23.1
Conserved unknown	229	4.8	1	0.2	--	--	--	--	--	--	--	--
Conserved hypothetical	239	5.0	10	2.3	1	2.7	1	4.3	1	5.0	--	--
Unknown	200	4.2	16	3.7	3	8.1	4	17.4	4	20.0	--	--
Hypothetical	948	20.0	227	51.8	20	54.1	6	26.1	10	50.0	9	69.2
Maximum ORF size (bp)	7689	86.3	5919	94.0	5523	90.2	3792	62.1	2481	73.2	2442	82.6
Average ORF size (bp)	899		822		966		860		538		651	
Regulation	188	4.0	12	2.7	--	--	1	4.3	--	--	--	--
Transport and Binding	234	4.9	14	3.2	1	2.7	--	--	--	--	--	--
Transposon-related	80	1.7	9	2.0	2	5.4	1	4.3	2	10.0	1	7.7
RNA												
rRNA operons	2		-		-		-		-		-	
tRNA	41		-		-		-		-		-	
SRP	1		-		-		-		-		-	
tmRNA	1		-		-		-		-		-	
6S / SsrS RNA	1		-		-		-		-		-	
TPP Riboswitch	1		-		-		-		-		-	
Cobalamin riboswitch	1		1		-		-		-		-	
ydaO/yuaA leader	1		-		-		-		-		-	
Ribosomal protein L10 leader	1		-		-		-		-		-	
Group II intron	1		-		-		-		-		-	

Figure 2. Phylogenetic tree of cyanobacteria. The tree was generated from the analysis of 435 sets of proteins, co-orthologous in *all* 28 of the analyzed strains. Each of the sets of orthologous proteins was aligned using MAFFT[3], concatenated into a single combined protein alignment of all 435 sets, then pruned of columns containing gaps. The PHYLIP[4] package was used to bootstrap, calculate distances, and generate the final tree using the Fitch-Margoliash method, designating *Gloeobacter* as the outgroup. All branches have bootstrap values of 100.



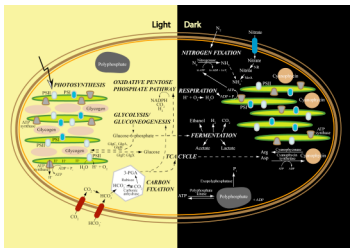


Figure 3. Processes involved in daily metabolic cycling. Photosynthesis fixes carbon during the day, which is stored in glycogen granules. Glycogen is rapidly consumed during a burst of respiration in the early dark period, which coincides with the peak of nitrogenase activity, fermentation, and a minimum of photosynthetic capacity[7]. Fixed nitrogen is stored in cyanophycin granules, which are depleted during the following day.

SUMMARY

- **A combination of sequencing and optical mapping has solved the genomic structure of *Cyanothece* sp. ATCC 51142 into a large circular chromosome, a small linear chromosome and 4 plasmids.**
- **Proteomics analysis has resulted in an improved gene annotation.**
- **Transcriptomics, proteomics and metabolomics are being performed as a function of time to detail the daily metabolic and circadian rhythms.**

OpGen optical mapping reveals circular and linear Chromosomes

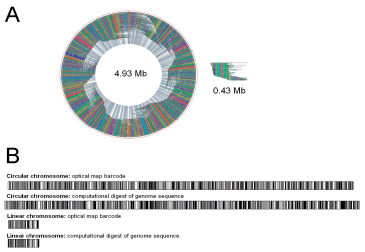


Figure 4. Optical restriction maps. (A) Optical restriction digest maps of the circular and linear chromosomes. (B) Comparison of optical restriction digest "barcode" patterns to those of the final genome sequence, showing agreements between the optical and final genome assemblies. The optical maps are slightly compressed relative to the computational maps, due to the contraction of fragment ends upon digestion.

Transcription during the diurnal cycle demonstrates the regulation of the metabolic rhythms. A complete model of cell division, proteomics, metabolomics and transcription will be developed now that the genome is completed.

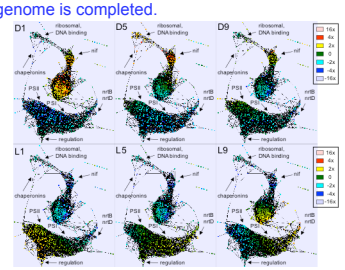


Figure 5. Transcription network as a function of the diurnal cycle. The colors indicate differential transcription, with red indicating up-regulation and blue down-regulation.

Initial funding for this project was from the Danforth Foundation at Washington University. This work is part of a Membrane Biology EMSL Scientific Grand Challenge project at the W. R. Wiley Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the U.S. Department of Energy's Office of Biological and Environmental Research (BER) program located at Pacific Northwest National Laboratory. PNNL is operated for the Department of Energy by Battelle.