# **Database tool**

# CyanoClust: comparative genome resources of cyanobacteria and plastids

#### Naobumi V. Sasaki\* and Naoki Sato

Department of Life Sciences, Graduate School of Arts and Sciences, University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo, 153-8902, Japan

\*Corresponding author: Tel: +81 3 5454 6631; Fax: +81 3 5454 6998; Email: naobumi@bio.c.u-tokyo.ac.jp

Submitted 8 October 2009; Revised 9 December 2009; Accepted 11 December 2009

Cyanobacteria, which perform oxygen-evolving photosynthesis as do chloroplasts of plants and algae, are one of the best-studied prokaryotic phyla and one from which many representative genomes have been sequenced. Lack of a suitable comparative genomic database has been a problem in cyanobacterial genomics because many proteins involved in physiological functions such as photosynthesis and nitrogen fixation are not catalogued in commonly used databases, such as Clusters of Orthologous Proteins (COG). CyanoClust is a database of homolog groups in cyanobacteria and plastids that are produced by the program Gclust. We have developed a web-server system for the protein homology database featuring cyanobacteria and plastids. Database URL: http://cyanoclust.c.u-tokyo.ac.jp/.

# Introduction

Chloroplasts are the sites of photosynthesis within the cells of land plants and algae. In non-photosynthetic plant tissues, these are called plastids (1). Various lines of evidence, such as the similarity of the photosynthetic machinery and photosynthetic metabolic pathways, suggest that cyanobacteria are related to the ancestor of chloroplasts (2,3). Cyanobacteria existed on the earth at least 2.7 billion years ago (4). Contemporary cyanobacteria exhibit wide morphological diversity (unicellular, colonial, filamentous, etc.) and are adapted ecologically to a wide spectrum of niches (freshwater, marine, terrestrial, thermophilic, etc.) (5). In a classical review of the molecular evolution of cyanobacteria, Doolittle (6) raised three questions: (i) what is the proper phylogenetic position of the cyanobacteria within the prokaryotes; (ii) what phylogenetic relationships exist within the cyanobacteria; and (iii) what evolutionary relationships do cyanobacteria bear to eukaryotic photosynthesis?

To answer the first and second questions, the clade structure and lineage of cyanobacteria have been revealed by careful molecular phylogenetic analyses. Based on interest in their ecological importance, genome sequences of cyanobacteria have been analyzed by the Kazusa DNA Research Institute in Japan (http://www.kazusa.or.jp/) and DOE Joint Genome Institute in the USA (http://www .jgi.doe.gov/). In addition, genomic data repositories for cyanobacteria such as CyanoBase (7), CYORF (http://cyano .genome.jp/) and Cyanosite (http://www-cyanosite.bio .purdue.edu/) assist the experimental analysis of cyanobacteria.

Referring to the third question, some reports have suggested that the gene transfer from the endosymbiont to the host genome was the key event (8). Nevertheless, despite the intuitive idea that the plastids and host plant cells co-evolved, the actual evolutionary relationship between the host cells and plastids remains complex, as seen in secondary endosymbiosis (9–11). It has also been pointed out that the evolution of plastids themselves was discontinuous (12). To study the evolution of plastids, some useful platforms are available for comparative genomic analyses, such as ChloroplastDB (13), but resources for the comparative genomics of plastids and cyanobacteria are still limited.

We developed the CyanoClust database web server to provide reliable orthologs over many cyanobacteria and plastids, or just various cyanobacteria. It is also used as a utility for comparative genome analysis. To compare orthologs among loosely related organisms, the Gclust software,

<sup>©</sup> The Author(s) 2010. Published by Oxford University Press.

This is Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http:// creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

which was developed in our laboratory (14), was used to construct the database. In this article, we present the database system and web interface of the CyanoClust database. This interface makes the CyanoClust web system a useful platform for comparative genomic analysis.

# Features of the CyanoClust database

#### Database content and organization

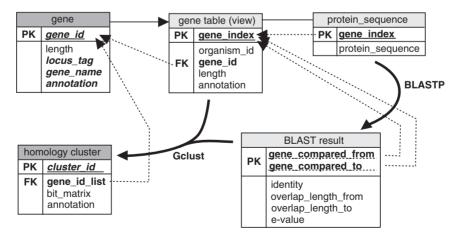
CyanoClust (version 2009-03) contains protein homology information for 38 cyanobacteria, 59 plastids, five anoxygenic photosynthetic bacteria and one chromatophore of *Paulinella*. In addition to these photosynthetic organisms, five non-photosynthetic bacteria were included and are expected to serve as outgroups.

The original genome data of these organisms were obtained from the NCBI/RefSeq (http://www.ncbi.nlm.nih.gov/genomes/) and processed into three tables describing protein data (Figure 1). This was performed with Perl scripts available from the Gclust server website (http:// gclust.c.u-tokyo.ac.jp/). The database-building process is summarized in Figure 1. We used the Gclust software, which is based on the Entropy Optimized Organism Count (EOOC) method (15), for automatic clustering to obtain a homolog cluster table. The EOOC method automatically simulates human-driven manual identification of protein clusters with some parameter settings. In total, 179 056 proteins were classified into 40526 clusters, including 26373 singleton clusters for which no orthologs were detected in any of the other genomes. Each cluster is accessed by a serial ID, and each cluster entry contains the list of gene IDs. In addition, the group number of the large, loosely

conserved synteny that we call virtual linkage groups (VLG) (16) is indicated with a probable-consensus annotation when it is assigned. Referencing to each gene entry is processed internally by CGI scripts on the web system. Each cluster is displayed as a bit matrix. The 'gene' tables and 'homology cluster' table are used mainly in the CyanoClust web system. The fields emphasized in italic bold font in these tables (Figure 1) were used as search keys in the web interface.

#### Organism grouping and evolution of photosynthesis

The main characteristic of the CyanoClust database is the selection of the genome set, a process especially designed to analyze the evolution of photosynthetic organisms such as cyanobacteria and plastids. In this respect, the five genomes of non-photosynthetic organisms serve as references or outgroups. The five genomes of anoxygenic photosynthetic bacteria are also good references for comparison with the oxygenic photosynthetic organisms. The plastid genomes in the genome set are classified into several groups: standard plastids, secondary plastids, Cyanidiales plastids and the Paulinella chromatophore genome. Paulinella chromatophora is a strange protist that has a plastid-like organelle and is thought to have originated from a cyanobacterium related to Prochlorococcus or Synechococcus (17). Therefore, the origin of the Paulinella chromatophore is believed to differ from the origin of plastids in plants and algae, which are believed to have branched from an ancestral cyanobacterium near the root cyanobacteria (9,16). In the bacterial group, of Rhodopseudomonas palsustris and Rhodospirillum rubrum are purple non-sulfur bacteria that perform anoxygenic photosynthesis, unlike cyanobacteria. Chlorobium tepidum

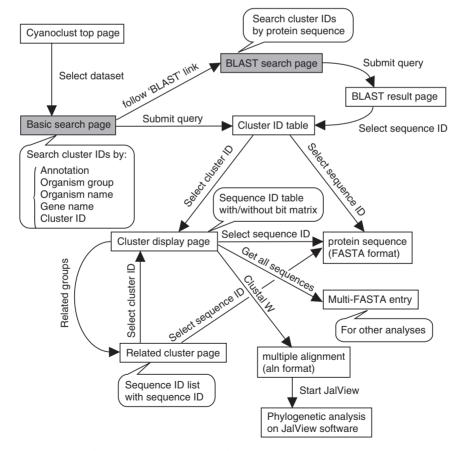


**Figure 1.** The process of database construction. Each box represents a conceptual object representing the structure of data. The topmost line in each box shows the name of the relational table. PK, primary key; FK, foreign key. Each dotted arrow indicates the relation of an entity to the reference. Bold arched arrows indicate an external operation to produce the relations (the main program names are given near the arrows). Light gray background of the title, transaction (temporary and internal) relation; dark gray background, resulting relation.

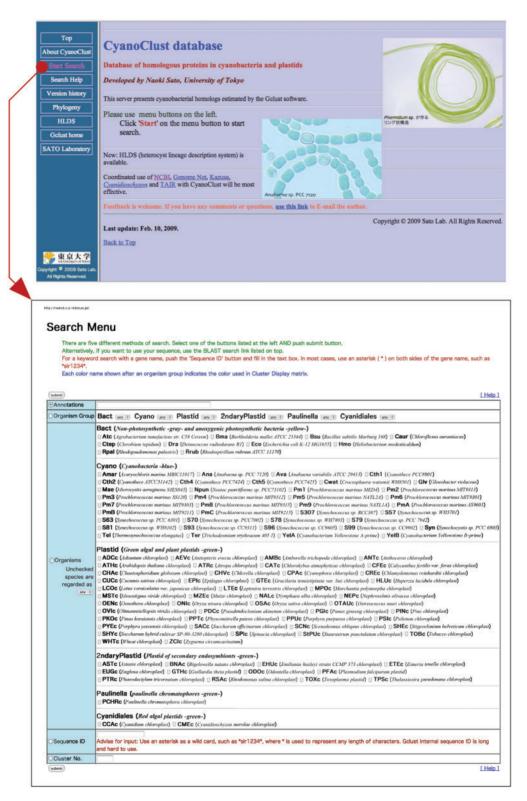
is another anoxygenic photosynthetic bacteria classified into green sulfur bacteria. *Heliobacterium* is a photosynthetic firmicute. The photosynthetic machineries of these bacteria differ somewhat from that of either cyanobacteria or plastids. Namely, anoxygenic bacteria have only one kind of photosystem containing bacteriochlorophylls, whereas cyanobacteria and plastids use two photosystems in series that contain chlorophylls to oxidize water to oxygen. Although the evolutionary relationship between anoxygenic photosynthesis and oxygenic photosynthesis is still unclear, the CyanoClust resources will assist further analysis of their evolution.

## The CyanoClust interface

The web system of the CyanoClust consists of search schemes, as shown in Figure 2, which are driven by CGI scripts. Most search tasks can be performed in the 'Basic Search' and 'Blast Search' pages. The 'Basic Search' page offers two kinds of searches: normal searches and phylogenetic profiling. The normal search is used to find clusters by annotations, sequence ID, and cluster number. To search clusters by annotation, any descriptive word(s) are allowed, such as 'DNA-binding protein' or 'Photosystem I', and the query is searched by string matching (wildcards are also available). These annotation keywords were originally taken from the gene annotation. A search using the GenBank gene ID is also possible in 'Sequence ID Search.' For expert use, a direct search using the cluster ID is also available. The phylogenetic profile search is run using two methods: group selection and organism selection. Each organism group has an attribute: 'yes,' 'no' or 'any.' The attribute 'yes' is used to select clusters that are shared by the designated group, and 'no' is used to select clusters that are not shared by the organism group. The attribute 'any' is used to exclude the organism group from the current selection. In the organism selection method, each organism can be selected one at a time. On submitting a guery, the 'Search Result' page is displayed, presenting the sequence ID and representative annotation of each cluster. Figure 3 shows an example of a guery that contains the keyword 'phycobilisome' in gene



**Figure 2.** Page transition diagram of the CyanoClust web interface. Each box indicates a web page, and each arrow indicates the page transition caused by an action. Each balloon indicates a simplified explanation of the indicated page. The gray boxes indicate the two types of initial search page, which are also linked with each other.



**Figure 3.** Screen shots showing the top page and Basic Search page. The Basic Search page opens upon following the link in the top page, as indicated by an arrow. The 'Basic Search' page accepts a search query consisting of an individual gene name or annotations, gene conservation in specified organism groups or species, or just the cluster ID. By submitting a query, a cluster ID table is displayed under the name 'Search Result'.

No.	Cluster Number	Sec	quence ID	Length	seqs	Annotations
1		6 Ava Ava 2925		172	75	Phycobilisome protein
2		7 Ter_Tery_4799		132		Phycobilisome protein
3	-	4 S78 SynWH7803 047	8	197		Conserved hypothetical protein in phycobilisome ro
4		4 S93_sync_2251=cpeD		301		Phycobilisome linker polypeptide
15		4 Ava Ava 2926		162		Phycobilisome protein
6		7 Mae MAE 42240		138		CpcD phycobilisome linker-like
7		7 Glv_gvip267=apcF		161		phycobilisome core component
, 8		Z CMEc_apcE		840		phycobilisome linker polypeptide; Cyanidioschyzon
9		2 Pm2 PMT1682=cpeA		155		Phycobilisome protein (phycoerythrin, alpha-subuni
10		2 S63_syc0494_c=cpcE		273		phycobilisome maturation protein CpcE
11		1 PCHRc PCC 0249=ap		176		phycobilisome core component-allophycocyanin beta-
12				80		CpcD phycobilisome linker-like
		2 <u>Ava_Ava_2933</u>				
13		4 Ana_asr0023=apcC		68		phycobilisome core linker protein Lc7.8
14		0 Ana_pCC7120deltap0	I=nblA	61		phycobilisome degradation protein
15		1 Ana_alr0535=cpcG2		247		phycobilisome rod-core linker protein
16		2 YelA_CYA_0523		159		putative phycobilisome protein
17		5 S81_SYNW1999=cpe0		294		phycobilisome linker polypeptide
18	<u>231</u>	6 S81_SYNW1997=cpc0	32	252	13	possible phycobilisome rod-core linker polypeptide
19	<u>260</u>	5 <u>S78_SynWH7803_048</u>	3	201	11	Conserved hypothetical protein in phycobilisome ro
20	<u>274</u>	5 S63_syc2065_d=cpcG		250	10	phycobilisome rod-core linker polypeptide
21	<u>282</u>	6 S78_SynWH7803_050	0	56	10	Conserved hypothetical protein in phycobilisome ro
22	<u>336</u>	0 S78_SynWH7803_049	0	144	8	Conserved hypothetical protein in phycobilisome ro
23	<u>384</u>	0 S78 SynWH7803 050	14	224	6	Conserved hypothetical protein in phycobilisome ro
24	385	2 <u>S81 SYNW1989</u>		300	6	Possible phycobilisome linker polypeptide
25	385	3 Ava Ava 2932		286		CpcD phycobilisome linker-like
26		4 S81_SYNW2001=cpeB		244		possible phycobilisome linker polypeptide
			-	= · · ·	-	
	ister Dis		nees : 42 Final threshold	192	Rec Million	
Clu		play	ices : 42 Final threshold		6 VLG :	Data Set : cyanoclust1 [ 2009-0
Clu	ister Dis ter No. : 507	play Number of Sequer		: 1.00e-06.	VLG :	Data Set : cyanoclust1 [ 2009-0         1       Annotation : hypothetical hypothetical protein         Image: State St
Clu	ister Dis ter No. : 507	play Number of Sequer		: 1.00e-06.	VLG :	Data Set : cyanoclust1 [ 2009-0         1       Annotation : hypothetical hypothetical protein       [His         1       8       8       1       1
Clu	ter No. : 507	play Number of Sequer		: <b>1.00e-06.</b>	VLG :	Data Set : cyanoclustl [ 2009-0         1       Annotation : hypothetical hypothetical protein       [His
Clus	ter No. : 507	play Number of Sequer	http://nsato6.c.u-tokyo.ac	: <b>1.00e-06.</b>	VLG : itiple Aligin_exec	Data Set : cyanoclust1 [ 2009-0         1       Annotation : hypothetical hypothetical protein         Image: Set : cyanoclust1 [ 2009-0         gnment Cluster-No:507         cgi?DATA_SET=cyanoclust18VERSION_NO=currer         Qr         Google
Clus Clus	Ister Dis ter No. : 507 CLUS Syn Ana Cwat_S	play Number of Sequer	http://nsato6.c.u-tokyo.ac http://nsato6.c.u-tokyo.ac litiple sequence align -HVSNNDKSNRVFLYBVMINPK	: 1.00e-06.	VLG : itiple Aligin_exec	Data Set : cyanoclustl [ 2009-4         1       Annotation : hypothetical hypothetical protein         Image: Set : cyanoclustl ( 2009-4         Image: Set : cy
Clu Clus No.	Ister Dis ter No. : 507	play Number of Sequer	http://nsato6.c.u-tokyo.ac http://nsato6.c.u-tokyo.ac httple sequence align -HVSNNDKSNRVFLYEVMINPK HLEANNQGENQIFLYEVVFSG-	: 1.00e-06.	VLG : itiple Alig ign_exec art Jalvie ORRGKPY ORFSRTV	Data Set : cyanoclustl [ 2009-0         1       Annotation : hypothetical hypothetical protein         Image: State of the state of th
Clus Clus	Ister Dis ter No. : 507 CLUS Syn Ana Cwat ( Mae_M Aya NgunD	play Number of Sequer C @ R + STAL W (1.83) mu 16341 E 42240 pNPDR040 iz2087	http://nsato6.c.u-tokyo.ac http://nsato6.c.u-tokyo.ac litiple sequence align -HVSNNDKSNRVFLYBVMINPK	: 1.00e-06.	VLG : itiple Aligin_exection cart Jalvie ORRGKPY QRFSRTV jp/cgi-b	Data Set : cyanoclustl [ 2009-0         1       Annotation : hypothetical hypothetical protein         1       Annotation : hypothetical hypothetical protein         1       Manotation : hypothetical hy
Clus Clus No. 1 2 3 4	Ister Dis ter No. : 507 CLUS Sym Ama Ava NepunD Mpu Syn_s: Mae M	play Number of Sequer C & F2 + STAL W (1.83) mu 16341 E 42240 pNPDR040 1r2087 E 51310	http://nsato6.c.u-tokyo.ac      http://nsato6.c.u-tokyo.ac      http://nsato6.c.u-tokyo.ac      http://nsato6     http:/nsato6     http://nsato6     ht	:1.00e-06.	VLG : itiple Aliji iign_exec art Jalvie orror Jalvie orror Jalvie iign_exec	Data Set : cyanoclust1 [ 2009-0         1       Annotation : hypothetical hypothetical protein       [He         mment Cluster-No:507       Coogle         cgi?DATA_SET=cyanoclust1&VERSION_NO=currer       Q= Google         w       Google       Google         1       0.14       YHE CY6 2030         0.39       1.2       YHE CY6 2030         0.39       50 Sympt7492,2118         0.39       570 SYNC7F02,A386         0.30       570 SYNC7F02,A386
Clus No. 1 2 3 4 5	Ister Dis ter No. : 507 CLUS Syn Nou Nou Syn si Syn Syn Si Syn Si	play Number of Sequer C @ P + STAL W (1.83) mu 16341 1E_42240 pNPDR040 1r2087 1E_51310 NPCC7002_A1850	http://nsato6.c.u-tokyo.ac      iltiple sequence align     -//VSNNDKSNRVPLYEVMINPK     MLEANNOCENOITLYEVVPSc-	: 1.00e-06. 3 jp/cgi-bin/ma ment sr MVKPI0INGLF SSGSLP SSGSLP SSGSLP SSGSLP SSGSLP SSGSLP SSGSLP SSGSLP SSGSLP SSGSLP SSGSLP 	VLG: itiple Alivie ign_exec cart Jalvie ORRGKPY ORFSRTV Jp/cgi-b 100 IYY FGP IYY FGP	Data Set : cyanoclustl [ 2009-0         1       Annotation : hypothetical hypothetical protein       Heg         nment Cluster-No:507       Git and the state of the stat
Clus Clus No. 1 2 3 4 5 6	Ister Dis ter No. : 507 CLUS Syn Ava Mae M Syn_se Mou Mae M Syn_se Syn_se Pmc veila_c	play Number of Sequer C @ P + STAL W (1.83) mu 16341 12: 42240 pNPDR040 17:2087 12: 51310 10: 10: 10: 10: 10: 10: 10: 10: 10: 10:	http://nsato6.c.u-tokyo.ac      iltiple sequence align     -//v9NNDK8NRVPLYEVMINPK     MLEANNOCENOITLYEVVPSc-          Orex_05341	<pre>: 1.00e-06.</pre>	VLG: triple Alia ign_exec art Jalvie ORRGKPY ORFSRTV jp/cgi-b 100 IVT FGP IVT FGP IVT FGP	Data Set : cyanoclustl [ 2009-0]         1       Annotation : hypothetical hypothetical protein
Clus Clus No. 1 2 3 4 5 6 7	Ister Dis ter No. : 507 CLUS Syn Ana Cwat_ Hae M Mae M Syn st Syn syn st Pm Sr0 SN Pm Yeih C PmA Ana ai Cwat Awa Awa	play Number of Sequer C @ @ + STAL W (1.83) mu 16341 EL 42240 pNPDR040 12007 12097 12097 12090 1	http://nsato6.c.u-tokyo.ac      iltiple sequence align     -KV9NNDK8NRVPLYEVMINPK     MLEANNQCENQIFLYEVVF8c-     Cwar_g6543     Cwar_g6543     FHMM     Mac_MAE_25131     FMMM     Mac_MAE_51315     FMMM     Mac_MAE_51315     FMMM     Mac_MAE_51315     FMMM	<pre>: 1.00e-06.</pre>	VLG: tiple Alivie ign_exec art Jalvie ORRGKPY ORPSRTV ip/cgi-b 100 IVF FGP IVF FGP IVF FGP IVF FGP	Data Set : cyanoclustl ( 2009-0         1       Annotation : hypothetical hypothetical protein
Clus Clus No. 1 2 3 4 5 6 7 8	Ister Dis ter No. : 507 CLUS Syn CLUS Syn CLUS Syn CLUS Syn CLUS Syn CLUS Syn Syn Syn Syn Syn Syn Syn Syn Syn Sy	play Number of Sequer C D F + STAL W (1.83) mu 16341 E_42240 pNPDR040 172007 18551310 17000702_A1850 174_0290 174_0290 175_020 175_	http://nsato6.c.u-tokyo.ac      iltiple sequence align     -///www.subscription.com/         www.subscription.com/         wwww.subscription.com/         www.subscription.com/         www.subscription.com/	: 1.00e-06.	VLG : tiple Alivie ign_exec art Jalvie ORRGKPY ORFSRTV ip/cgi-b 100 LYT FGP TYT FGP TYT FGP TYT FGP TYT FGP TYT FGP	Data Set : cyanoclustl [ 2009-0         1       Annotation : hypothetical hypothetical protein       Heg         nment Cluster-No:507       Colspan="2">Colspan="2"Co
Clus Clus No. 1 2 3 4 5 6 7 8 9	Ister Dis ter No. : 507 CLUS Syn. CLUS Syn. Syn. ac Wae M Ava. Npun Mae M Hae M Yela_C PmM Ana_ai Sofs Syn.ai Syn.	play Number of Sequer C D F + STAL W (1.83) mu 16341 E 42240 pNPDR040 172007 18551310 17000702_A1850 174_0290 174_0290 175_0200 175_0200000000000000000000000000000000000	http://nsato6.c.u-tokyo.ac      iltiple sequence align      -KVSNNDKSNNVLLYEVMINPK      MLEANNOGCROIFLYEVVFSG-      O http://nsato6      Cont.g534     Fillion      Neuro_NVDR040     Fillion      You J.VIS007     R KKW      You J.VIS007     Fillion      You J.VIS007     Fillion      You J.VIS007      Fillion      You J.VIS007      Fillion      You J.VIS007      Fillion      F	I 1.00e-06.	VLG: itiple Align_exec ign_exec cart Jalvie ORRGKPY ORFSRTV IVF FGP IVF FGP	Data Set : cyanoclustl [ 2009-0         1       Annotation : hypothetical hypothetical protein
Clus Clus No. 1 2 3 4 5 6 7 8 9 10	Ister Dis ter No. : 507 CLUS Syn. CLUS Syn. Syn. as Yela. Char Yela. Char Yela. Char Yela. Char Yela. Char Yela. Char Yela. Char Yela. Syn. as Son Yela. Syn. as Son Yela. Syn. as Son Syn. Syn. Syn. Syn. Syn. Syn. Syn. Syn. Syn.	play Number of Sequer C G R + STAL W (1.83) mu (6341 E_42240 pNPDR0400 IE_51310 NPCC7002_A036 NPCC7002_A0386	http://nsato6.c.u-tokyo.ac      iltiple sequence align     KVSNNDKSNNVELYEVMINPK      MLEANNQGENVIFLYEVMINPK      MLEANNQGENVIFLYEVVIPSG-      Omt gf531      There MAR 25249      Omt gf531      There MAR 25249      Chart gf531      There MAR 2524      Chart gf531      Chart gf53      Chart gf5      Chart gf53      Chart gf53      Chart gf53      Chart gf5      Chart gf53      Chart gf5      Ch	: 1.00e-06.	VLG: itiple Alivie iart Jalvie cart Jalvie ORRGKPY ORFSRTV IP/cgi-b 100 IV FGP VY FGP VY FGP VY GP VY GP VY GP VY GP VY GP	Data Set : cyanoclustl { 2009-0           1         Annotation : hypothetical hypothetical protein         [He           gnment Cluster-No:507         cgi?DATA_SET=cyanoclustl&VERSION_NO=currer         Q= Google           w         Image: Comparison of the compa
Clus Clus No. 1 2 3 4 5 6 7 7 8 9 10 11	Ister Dis ter No. : 507 CLUS Syn Ana Cwat_ Mae M Ava Nyn as Pm 70 Si Pm 70 Si Si Si Si Si Si Si Si Si Si Si Si Si S	play Number of Sequer Number of Sequer C @ @ + STAL W (1.83) mu 16341 E_42240 pNPDR040 12087	http://nsato6.c.u-tokyo.ac      iltiple sequence align     KVSNNDKSNRVELYEVMINPK      MLEANNOGENGIFLYEVVFSG-      O http://nsato6      MuEANNOGENGIFLYEVVFSG-      O http://nsato6      O http	: 1.00e-06.	VLG: tiple Ali iign_exec art Jalvie oRRGKPY oRFSRT 100 107 107 107 107 107 107 107	Data Set : cyanoclustl [ 2009-0           1         Annotation : hypothetical hypothetical protein         Heg           nment Cluster-No:507         Immediate in the intervention of the interventinte intervention of the intervention of the interventio
Clus Clus No. 1 2 3 4 5 6 7 7 8 9 10 11 11 12	Ister Dis ter No. : 507 CLUS Syn. CLUS Syn. CLUS Syn. Syn. Syn. Syn. Syn. Syn. Syn. Syn	play Number of Sequer C @ R <sup>3</sup> + STAL W (1.83) mu 16341 152087 1	http://nsato6.c.u-tokyo.ac     iltiple sequence align     -KV9NNDK8NRVPLYEVHINPK     MLEANNQCEN0IFLYEVYF8c-     Cwa2_0534     Ms_MAE_2247     Ms_MAE_2247     Ms_MAE_2247     Ms_MAE_2247     Ms_MAE_2247     Ms_MAE_2247     Ms_MAE_2247     Ms_MAE_2331     THWM     Ms_MAE_2331     THWM     Ms_MAE_2331     THWM     Ms_MAE_2337     THWM     Ms_MAE_2357     Ms_MAE_2357     MAHW     MAE_MAE_2357     MAHW     MAEMMAE_MAE_2357     MAHW     MAEMMAE_MAE_2357     MAHW     MAEMMAE_MAE_2357     MAHW     MAEMMAE_MAEMMAEMMAEMMAEMMAEMMAEMMAEMMAEM	: 1.00e-06.	VLG: VLG:	Data Set : cyanoclustl { 2009-0           1         Annotation : hypothetical hypothetical protein         LHc           gnment Cluster-No:507         cgi?DATA_SET=cyanoclustl&VERSION_NO=currer         Q= Google           w         Image: Comparison of the compa
Clus Clus No. 1 2 3 4 5 6 7 7 8 9 10 11 11 12	Ister Dis ter No. : 507 CLUS Syn Ava No. Mae M Sro 87 Sro 87 S	play Number of Sequer Number of Sequer C @ @ + STAL W (1.83) mu (6341 E_42240 pNPDR040 ir2007 iE_51310 NPCC7002_A1850 ir4007 ir4007 ir208	http://nsato6.c.u-tokyo.ac     iltiple sequence align     -//v9NNDK8NRVPL/YEV/HINPK     //stankogENQIFLYEV/FSG     //stankog	: 1.00e-06.	VLG: VLG:	Data Set : cyanoclustl ( 2009-0           1         Annotation : hypothetical hypothetical protein         Etc.           gnment Cluster-No:507         cgi?DATA_SET=cyanoclust1&VERSION_NO=currer         Q= Google           w         Image: Comparison of the comp
Clus Clus No. 1 2 3 4 5 6 7 7 8 9 10 11 12 13	Ister Dis ter No. : 507 CLUS Syn Syn CLUS Syn CLUS Syn CLUS Syn CLUS Syn Syn Syn Syn Syn Syn Syn Sy	play Number of Sequer C	http://nsato6.c.u-tokyo.ac      http://nsato6.c.u-tokyo.ac      http://nsato6.c.u-tokyo.ac      http://nsato6.c.u-tokyo.ac      http://nsato6     http:/nsato6     http://nsato6     http://nsato6     http://nsato6     http://nsato6     http://nsato6     http://nsato6     http://nsato6     http:/nsato6     http	: 1.00e-06.	VLG: VLG: VLG: VLG: VLG: VTCO	Data Set : cyanoclustl [ 2009-0         1       Annotation : hypothetical hypothetical protein       He         nment Cluster-No:507       Image: Cluster-No:507       Image: Cluster-No:507         cgi?DATA_SET-cyanoclust1&VERSION_NO-currer       Image: Cluster-No:507       Image: Cluster-No:507         cgi?DATA_SET-cyanoclust1&VERSION_NO-currer       Image: Cluster-No:507       Image: Cluster-No:507         cgi?DATA_SET-cyanoclust1&VERSION_NO-currer       Image: Cluster-No:507       Image: Cluster-No:507         Image: Cluster-No:507       Image: Cluster-No:507
Clus No. 1 2 3 4 5 6 7 8 9 10 11 12 13 14	Ister Dis ter No. : 507 CLUS Syn CLUS Syn CLUS Syn CLUS Syn CLUS Syn CLUS Syn Syn Syn Syn Syn Syn Syn Sy	play Number of Sequer Number of Sequer C @ R + STAL W (1.83) mu (6341 E 42240 pxPDx040 iz2087 KE 51310 NPCC7002_A1850 ix160 *a_4414 NpR1189 iz0957 E 28390 NPCC7002_A0386 *c1974_d	http://nsato6.c.u-tokyo.ac      iltiple sequence align     KVSNNDKSNRVELYEVMINPK      MLEANROGENDIFLYEVVISSG-      Omt g6531     Omt g653     Omt g65     Omt g653     Omt g65	: 1.00e-06.	VLG: tiple Ali lign_exec art Jalvie art Jalvie typop typ	Data Set : cyanoclustl [ 2009-0           1         Annotation : hypothetical hypothetical protein         He           gnment Cluster-No:507         git         <
Clus No. 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15	Ister Dis ter No. : 507 CLUS Syn Ana CVat_ Mae, MA Ava Nou Mae, MA Ava Nou Nou Mae, MA Ava Nou Nou Mae, MA Ava Nou Nou Mae, MA Ava Nou Nou Mae, MA Ava Nou Nou Mae, MA Ava Nou Nou Mae, MA Ava Nou Nou Mae, MA Ava Ava Ava Ava Ava Ava Ava Av	play Number of Sequer Number of Sequer C @ C + STAL W (1.83) mu (6341 E_42240 pNPDR040 ir2007	http://nsato6.c.u-tokyo.ac      iltiple sequence align     KVSNNDKSNRVELYEVMINPK      MLEANROGENDIFLYEVVISSG-      Omt g6531     Omt g653     Omt g65     Omt g653     Omt g65	: 1.00e-06.	VLG: tiple Ali lign_exec art Jalvie art Jalvie typop typ	Data Set : cyanoclustl [ 2009-0           1         Annotation : hypothetical hypothetical protein         LHc           ymment Cluster-No:507         cgi?DATA_SET=cyanoclust1&VERSION_NO=currer         Coogle           w         Image: Cluster-No:507         cgi?DATA_SET=cyanoclust1&VERSION_NO=currer         Coogle           ymment Cluster-No:507         cgi?DATA_SET=cyanoclust1&VERSION_NO=currer         Coogle         Coogle           w         Image: Cluster-No:507         cgi?DATA_SET=cyanoclust1&VERSION_NO=2009-01&CCLU         Cluster-No:502           intervention         intervention         Cluster-No:502         Cluster-No:502         Cl
Clus No. 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16	Ister Dis ter No. : 507 CLUS Syn. CLUS Syn. CLUS Syn. CLUS Syn. Cut. Syn.	play Number of Sequer C	http://nsato6.c.u-tokyo.ac      iltiple sequence align     KVSNNDKSNRVELYEVMINPK      MLEANROGENDIFLYEVVISSG-      Omt g6531     Omt g653     Omt g65     Omt g653     Omt g65	: 1.00e-06.	VLG: tiple Ali lign_exec art Jalvie art Jalvie typop typ	Data Set : cyanoclustl ( 2009-0           1         Annotation : hypothetical hypothetical protein         LHc           gnment Cluster-No:507         ggrad
Clus No. 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17	Ister Dis ter No. : 507 CLUS Syn CLUS Syn CLUS Syn CLUS Syn CLUS Syn CLUS Syn Syn Syn Syn Syn Syn Syn Sy	play Number of Sequer Number of Sequer C @ @ + STAL W (1.83) mu (6341 E 42240 pNPDR040 iz2087 E 51310 NPCC7002_A1850 YA_0290 Y	http://nsato6.c.u-tokyo.ac     iltiple sequence align	: 1.00e-06.	VLG: tiple Ali lign_exec art Jalvie art Jalvie typop typ	Data Set : cyanoclustl [ 2009-0         1       Annotation : hypothetical hypothetical protein       LHc         ymment Cluster-No:507       cgi?DATA_SET=cyanoclust1&VERSION_NO=currer       Coogle         cgi?DATA_SET=cyanoclust1&VERSION_NO=currer       Coogle       cgi?DATA_SET=cyanoclust1&VERSION_NO=2009-01&CLLU         0       get_dnd.cgi?DATA_SET=cyanoclust1&VERSION_NO=2009-01&CLLU       cgi?DATA_SET=cyanoclust1&VERSION_NO=2009-01&CLLU         0       get_dnd.cgi?DATA_SET=cyanoclust1&VERSION_NO=2009-01&CLLU       cgi?Data_2030         0       get_dnd.cgi?DATA_SET=cyanoclust1&VERSION_NO=2009-01&CLLU         0       get_dnd
Clus No. 1 2 3 4 5 6 7 7 8 9 9 10 11 12 13 14 15 16 17 18	Ister Dis ter No. : 507 CLUS Syn CLUS Syn CLUS Syn CLUS Syn CLUS Syn CLUS Syn CLUS Syn Syn Syn Syn Syn Syn Syn Sy	play Number of Sequer C	http://nsato6.c.u-tokyo.ac      http://nsato6.c.u-tokyo.ac      http://nsato6.c.u-tokyo.ac      http://nsato6.c.u-tokyo.ac      http://nsato6     http:/nsato6     http://nsato6     http://nsato6     http:/nsato6     http://nsato6     http:/nsato6     http:/ns	: 1.00e-06.	VLG: tiple Ali lign_exec art Jalvie art Jalvie typop typ	Data Set : cyanoclustl ( 2009-0           1         Annotation : hypothetical hypothetical protein         LHc           ymment Cluster-No:507         cgi?DATA_SET=cyanoclust1&VERSION_NO=currer         Coogle           cgi?DATA_SET=cyanoclust1&VERSION_NO=currer         Coogle         Coogle           ww         Coogle         Coogle         Coogle           1         0.14         Yelk C10, 2030         Coogle           1         0.20         Coogle         Soft Symptify 2010           1         0.14         Yelk C10, 2030         Soft Symptify 2010           1         0.20         0.21         Soft Symptify 2010           1         0.20         0.22         Soft Symptify 2010
Clus No. 1 2 3 4 5 6 7 8 9 9 10 11 12 13 14 15 16 17 18 19	Ister Dis ter No. : 507 CLUS Syn CLUS Syn CLUS Syn CLUS Syn CUS Syn Syn Syn Syn Syn Syn Syn Sy	play Number of Sequer C C (1) (1) (1) (1) (1) (1) (1) (1) (1) (1)	http://nsato6.c.u-tokyo.ac     iltiple sequence align	: 1.00e-06.	VLG: tiple Ali lign_exec art Jalvie art Jalvie typop typ	Data Set : cyanoclustl [ 2009-0           1         Annotation : hypothetical hypothetical protein         Heg           nment Cluster-No:507         gramment Cluster-No:507         gramment Cluster-No:507           cgi?DATA_SET=cyanoclust1&VERSION_NO=currer         Q- Google         gramment Cluster-No:507           ww         Image: Cluster Clu
Clus No. 1 2 3 4 5 6 7 8 9 9 10 11 12 13 14 15 16 17 18 19	Ister Dis ter No. : 507 CLUS Syn CLUS Syn CLUS Syn CLUS Syn CLUS Syn Syn Syn Syn Syn Syn Syn Sy	play Number of Sequer Number of Sequer C @ @ + STAL W (1.83) mu (6341 E_42240 pNPDR040 ir2007 iE_51310 NPCC7002_A1850 ir3007 ir4007 ir200	http://nsato6.c.u-tokyo.ac      http://nsato6.c.u-tokyo.ac      http://nsato6.c.u-tokyo.ac      http://nsato6.c.u-tokyo.ac      http://nsato6     http:/nsato6     http://nsato6     http://nsato6     http:/nsato6     http://nsato6     http:/nsato6     http:/ns	: 1.00e-06.	VLG: tiple Ali lign_exec art Jalvie art Jalvie typop typ	Data Set : cyanoclustl [ 2009-0           1         Annotation : hypothetical hypothetical protein         LHc           gnment Cluster-No:507         cgi?DATA_SET=cyanoclustl&VERSION_NO=currer < Q= Google

#### Search results 44 hit(s) Key:Annotations Val:Phycobilisome

Figure 4. Example analysis of a cluster. The upper figure shows a cluster ID table in a 'Search Results' page. Upon clicking the cluster number in the cluster ID table (red circle), a 'Cluster Display' page is shown. Below is a stacked view of a 'Cluster Display' page, a multiple alignment, and a dendrogram driven by the program Jalview. In the 'Cluster Display' page, a bit matrix is used to show the similarity of homologs for small clusters (hidden behind other windows in this example), but only a list is shown if the cluster size is large.

annotation. If a keyword search was used to show the page, the matched keyword is shown in red. This page is used to access clusters or protein sequences. Users can fetch any sequence in FASTA format for further analysis. The 'Cluster Display' page shows a matrix of homologs, if the matrix is not too large. Below the matrix are the buttons 'ClustalW' and 'Get All Sequences.' An additional button, 'Related Groups,' is also shown if homologs exist that were not used to construct the cluster.

In addition, an amino acid sequence can be searched from the 'Blast Search' page, which is accessible by following the 'Blast Search' link on top of the 'Basic Search' page. The CyanoClust web system also implements a convenient phylogenetic analysis tool using Jalview (18) software (Figure 4) after displaying an alignment, which requires JDK runtime version 1.4.2 or higher on the client. With these interfaces, the CyanoClust web system is used as a platform for comparative genomic analysis.

## **Prospects**

#### Quality control of cluster annotation

The current policy of the CyanoClust database is that we do not modify the annotation given for each protein in the original database. In the 'Search Result' page, the annotation field inherits the gene annotation of the top entry of the gene list. Currently, we do not curate the annotations because any annotations given in the original databases may be subject to changes with research advances. Different annotations are often given for the entries of a single cluster, but comparison of these different descriptions will give an idea of the function of the orthologs. A future enhancement of the database will include well-curated annotations of the clusters. However, this will be possible only after extensive efforts at manual curation. Although extensive, the complete curated annotation will be possible using the CyanoClust database, as described below.

#### CyanoClust as an annotation tool

The increase in genome sequence data in the public repositories will facilitate comparison of various different genomes, but the current database and software development focus mainly on human-related genomes. The development of comparative genomics and annotation of photosynthetic organisms is still in progress. In this respect, the CyanoClust database may be useful as a starting point for evidence-based detailed annotation. Experimental data have accumulated for *Synechocystis* sp. PCC 6803 and *Anabaena* sp. PCC 7120. For many other cyanobacteria, strain-specific physiology information is being accumulated. By combining all available experimental data through the orthologous relationship defined by CyanoClust, we will be able to give a more reliable annotation to each conserved cluster. Another promising application of CyanoClust is its use as a reference to annotate new genomes that are being sequenced with new generation sequencers. Formerly, annotation of new genome involved extensive blast searches, but it is usually difficult to judge a correct threshold. Re-clustering of all new possible proteins with all of the sequences in CyanoClust will give reliable clusters, which will have plausible annotations. Only orphan proteins will have to be studied individually.

In summary, CyanoClust will serve as a potent tool for the annotation of both sequenced genomes and genomes to be sequenced.

# Funding

This work was supported in part by Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan (17018010, 18017005, 20017006) and the Global COE program 'From Earth to Earths' to N.S., and a Grant-in-Aid from the Japan Society for the Promotion of Science (JSPS) (2011425) to N.V.S. Funding to pay the Open Access publication charges for this article was provided by Ministry of Education, Culture, Sports, Science and Technology, Japan.

Conflict of interest. None declared.

### References

- 1. Kirk, J.T.O. and Tilney-Bassett, R.A.E. (1967) *The Plastids; their Chemistry, Structure, Growth, and Inheritance* (W. H. Freeman, London, San Francisco).
- 2. Sandelius, A.S. (2008) *The chloroplast*, Springer, New York. Springer, New York.
- 3. Castenholz, R. (1992) Species usage, concept, and evolution in the cyanobacteria (Blue-Green-Algae). J. Phycol., 28, 737–745.
- Knoll,A.H. (2008) Cyanobacteria and earth history. In: Herrero,A. and Flores,E. (eds), *The Cyanobacteria: Molecular Biology, Genomics, and Evolution*, Caister Academic Press, Norfolk, UK, pp. 1–20.
- 5. Whitton,B.A. and Potts,M. (2000) Introduction to the cyanobacteria. In: Whitton,B.A. and Potts,M. (eds), *The Ecology of Cyanobacteria: their Diversity in Time and Space*, Kluwer Academic, Boston, pp. 1–12.
- Doolittle,W.F. (1982) Molecular evolution. In: Carr,N.G. and Whitton,B.A. (eds), *The Biology of Cyanobacteria*, Botanical monographs, University of California Press, Berkeley, Vol. 19, pp. 307–332.
- Nakamura, Y., Kaneko, T. and Tabata, S. (2000) CyanoBase, the genome database for *Synechocystis* sp strain PCC6803: status for the year 2000. *Nucleic Acids Res.*, 28, 72.
- Martin, W., Rujan, T., Richly, E. et al. (2002) Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc. Natl Acad. Sci. USA, 99, 12246–12251.

- Sato, N. (2006) Origin and evolution of plastids: genomic view on the unification and diversity of plastids. In: Wise, R.R. and Hoober, J.K. (eds), *The Structure and Function of Plastids*, Advances in photosynthesis and respiration v. 23, Springer, Dordrecht, Vol. 23, pp. 75–102.
- 10. Kim,E. and Archibald,J.M. (2008) Diversity and evolution of plastids and their genomes. In: Sandelius,A.S. and Aronsson,H. (eds), *The Chloroplast*, Springer, New York, pp. 1–39.
- 11. Cavalier-Smith, T. (2002) Nucleomorphs: enslaved algal nuclei. *Curr. Opin. Microbiol.*, **5**, 612–619.
- Sato,N. (2001) Was the evolution of plastid genetic machinery discontinuous? Trends Plant Sci., 6, 151–155.
- Cui,L.Y., Veeraraghavan,N., Richter,A. et al. (2006) ChloroplastDB: the chloroplast genome database. Nucleic Acids Res., 34, D692–D696.

- 14. Sato, N., Ishikawa, M., Fujiwara, M. *et al.* (2005) Mass identification of chloroplast proteins of endosymbiont origin by phylogenetic profiling based on organism-optimized homologous protein groups. *Genome Inform.*, **16**, 56–68.
- Sato, N. (2009) Gclust: trans-kingdom classification of proteins using automatic individual threshold setting. *Bioinformatics*, 25, 599–605.
- Sasaki,V. and Sato,N. (2009) Elucidating genome structure evolution by analysis of isoapostatic gene clusters using statistics of variance of gene distances. *Genome Biol. Evol.*, doi:10.1093/gbe/evp051.
- Nowack, E.C.M., Melkonian, M. and Glockner, G. (2008) Chromatophore genome sequence of Paulinella sheds light on acquisition of photosynthesis by eukaryotes. *Curr. Biol.*, 18, 410–418.
- Clamp, M., Cuff, J., Searle, S.M. et al. (2004) The Jalview Java alignment editor. Bioinformatics, 20, 426–427.