Published online 30 October 2009 *Nucleic Acids Research, 2010, Vol. 38, Database issue* ***D379–D381***
doi:10.1093/nar/gkp915

CyanoBase: the cyanobacteria genome database update 2010

Mitsuteru Nakao[1,2,*], Shinobu Okamoto[1,2], Mitsuyo Kohara[1], Tsunakazu Fujishiro[1], Takatomo Fujisawa[1], Shusei Sato[1], Satoshi Tabata[1], Takakazu Kaneko[3] and Yasukazu Nakamura[1,4]

[1]Kazusa DNA Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu, Chiba 292-0812, [2]Database Center for Lifescience, Research Organization of Information and Systems, Faculty of Engineering Bldg.12, The University of Tokyo 2-11-16 Yayoi, Bunkyo-ku, Tokyo 113-0032, [3]Kyoto Sangyo University, Motoyama, Kamigamo, Kita-Ku, Kyoto-City 603-8555 and [4]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization of Information and Systems, Yata, Mishima 411-8540, Japan

Received September 15, 2009; Accepted October 7, 2009

ABSTRACT

CyanoBase (http://genome.kazusa.or.jp/cyanobase) is the genome database for cyanobacteria, which are model organisms for photosynthesis. The database houses cyanobacteria species information, complete genome sequences, genome-scale experiment data, gene information, gene annotations and mutant information. In this version, we updated these datasets and improved the navigation and the visual display of the data views. In addition, a web service API now enables users to retrieve the data in various formats with other tools, seamlessly.

INTRODUCTION

Cyanobacteria are prokaryotic organisms that have served as important model organisms for studying oxygenic photosynthesis and have played a significant role in the Earth's history as primary producers of atmospheric oxygen. *Synechocystis* sp. PCC 6803 was the first cyanobacteria to have its genome sequenced in 1996 (1). CyanoBase is a comprehensive and freely accessible web database of Cyanobacteria genome information; the data and the web site are licensed under the Creative Commons CC0 public domain license. The database contains the entire 3.9 Mb genome sequence of *Synechocystis* sp. PCC 6803 in six circular genomic molecules (chromosome and plasmids), with a total of 3725 genes. CyanoBase was developed as the genome database not only for *Synechocystis* sp. PCC but also for the other cyanobacteria species (2,3). CyanoGenes/mutants, released in 1998, were designed to facilitate the sharing of information on mutants and manual gene annotations submitted by the research community (4). As a result of several genome sequencing projects involving cyanobacteria species, CyanoBase now includes 35 completely sequenced genomes. In addition, various genome scale experimental datasets have been produced, including gene expression profiles and protein–protein interaction data.

In this update of the database, we have redesigned and improved the user interface in order to support both biologists, who work around the experiments, and bioinformaticians, who tend to deal with emerging genome-scale data. We have expanded the accessibility of the data navigation based on its hierarchical nature. We have also developed a new keyword search system to allow the user to access deep information about the genome. For improving reusability of the data, a web service was released to enable processing of the data using other tools. Useful external links were also updated to serve an information hub for the cyanobacteria genes.

IMPROVED DATA ACCESSIBILITY

CyanoBase provides access paths to the genome and gene information for cyanobacteria. To improve the accessibility of the data, the organization of the data was re-arranged. CyanoBase consists of pages for viewing each genome resource, including genome projects (DataSetView), an individual genome project (SpeciesView), individual chromosomes (MapView), multiple genes (GenesView), gene function classification (GeneCategoryView), word clouds (WordCloudView), search results (SearchView) and individual genes (GeneView). Pages are linked according to the hierarchy and connectivity of the data. We also designed the navigation to conform to the structure for genome information as used by biologists.

*To whom correspondence should be addressed. Tel: +81 3 5841 6754; Fax: +81 3 5841 8090; Email: mn@kazusa.or.jp, mn@dbcls.jp

The GeneView page can be reached via multiple paths from the species page, including the (i) chromosome circle map (MapView), (ii) gene list (GenesView), (iii) function classification (GeneCategoryView) and (iv) search results (SearchView). The hierarchical relationship is displayed on every page as a breadcrumbs list in the header. We formulated URLs to correspond to the hierarchical navigation. For instance, the URL of the slr1311 GeneView page for *Synechocystis* sp. PCC 6803 is http://genome.kauzsa.or.jp/cyanobase/Synechocystis/genes/slr1311, in which each part of the path refers to a step in the hierarchy: the data source name (Synechocystis), the scope name (genes) and the gene ID (slr1311).

The word cloud is generated automatically from text descriptions of a gene set to facilitate a visual inspection of an outline of the gene descriptions. This view helps users to summarize the gene set and explore related gene sets. The word frequencies in the text of a gene description are summed to construct a word cloud view that captures the character of the selected gene set (Figure 1).

We also improved the keyword search feature. In a full text search, users can now select a target data scope. The search targets include gene symbols, definitions, function classifications, descriptions in automatic annotations and information on mutants.

IMPROVED DATA REPRESENTATION

In this update, we introduced several new data representations to improve viewing of and navigation through data.

Genome context

A graphical image of the genomic context, generated by Gbrowse (5), indicates the length, direction and function of the gene and the surrounding genes. It also provides the navigation links among genes in GeneView.

TableView

TableView provides an enhanced user interface that is sortable by column and contains related links. The tabular representation is suitable for the species list, BLAST hits (homologs) and InterProScan (6) matches. It is useful to analyze intra-/inter-genomic data using these simple statistics and rankings.

Protein domains

An image and a table of predicted InterPro functional domains enable the user to glance at the arrangement of the protein functional domains and peruse these descriptions on the GeneView page. The InterPro IDs in the table have links to lists of the genes to be matched within or between species in CyanoBase.

Word clouds

A word cloud image of the gene function, created using Wordle (http://www.wordle.net), provides a summarized view of the gene function of a species on the SpeciesView page. The graphics work as an icon of the gene function of the species.

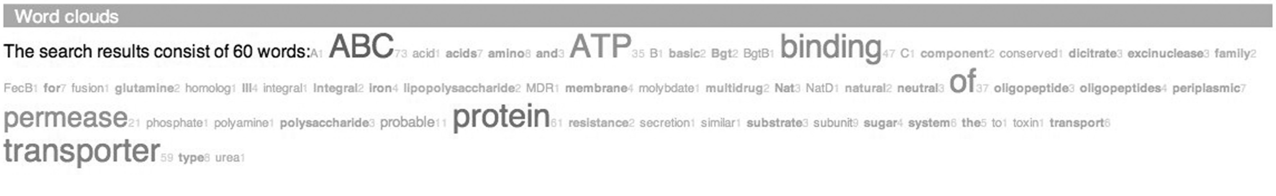
NEW DATA AND RESOURCES

The resources present in CyanoBase as of September 2009 are shown in Table 1. The annotations, along with the additional genome and gene information, are accumulated continually. Genome projects on cyanobacteria species have produced 35 complete sequenced genomes. CyanoBase imported the genome information from both GenBank and the original sites of the genome projects.

We also updated the curated resources. First, we updated 301 open reading frames (ORFs) based on comments from the research community: 200 ORFs were improved in the translational initiation site, 99 new ORFs were added and 2 ORFs were deleted. Second, we updated the functional annotations of 338 genes based on information registered in CyanoGenes and CYORF. Third, mutant information and curated gene descriptions were collected directly from biologists for 1700 cases and 688 genes via CyanoMuntats. We integrated the mutant information into the GeneView page and released the new summary page for the mutants. Fourth, we added a

Table 1. Data and annotations in CyanoBase (September 2009)

Annotation	Data
Species (genome projects)	35
Nucleotides	266 584 858
Genes	117 435
Protein-coding genes	114 783
RNA genes	2652
PubMed references for genes	2260
Mutant information for 688 genes	1700



Word clouds

The search results consist of 60 words: A$_{1}$ ABC$_{73}$ acid$_{1}$ acids$_{7}$ amino$_{8}$ and$_{3}$ ATP$_{35}$ B$_{1}$ basic$_{2}$ Bgt$_{2}$ BgtB$_{1}$ binding$_{47}$ C$_{1}$ component$_{2}$ conserved$_{1}$ dicitrate$_{3}$ excinuclease$_{3}$ family$_{2}$ FecB$_{1}$ for$_{7}$ fusion$_{1}$ glutamine$_{2}$ homolog$_{1}$ III$_{4}$ integral$_{1}$ Integral$_{2}$ iron$_{4}$ lipopolysaccharide$_{2}$ MDR$_{1}$ membrane$_{4}$ molybdate$_{1}$ multidrug$_{2}$ Nat$_{3}$ NatD$_{1}$ natural$_{2}$ neutral$_{3}$ of$_{37}$ oligopeptide$_{3}$ oligopeptides$_{4}$ periplasmic$_{7}$ permease$_{21}$ phosphate$_{1}$ polyamine$_{1}$ polysaccharide$_{3}$ probable$_{11}$ protein$_{61}$ resistance$_{2}$ secretion$_{1}$ similar$_{1}$ substrate$_{3}$ subunit$_{9}$ sugar$_{4}$ system$_{6}$ the$_{5}$ to$_{1}$ toxin$_{1}$ transport$_{6}$ transporter$_{59}$ type$_{8}$ urea$_{1}$

Figure 1. Word cloud links of gene annotations. The display represents the search results from the SearchView page by the keyword 'ABC' for *Synechocystis* sp. PCC 6803 genes at SearchView page. The size of each word corresponds to the number of times it appears in the results.

publication list for each gene. Publications that described the cyanobacteria genes were curated manually and listed on the GeneView page. The curating effort has continued to operate on a portion of the Gene Indexing project using KazusaAnnotation, a social genome annotation system (http://a.kazusa.or.jp). Finally, the GeneView page now includes genome-scale experimental data, including protein–protein interaction data for *Synechocystis* sp. PCC 6803 collected by the yeast two-hybrid method (7).

We added automated annotations for each gene on the GeneView page. These annotations include putative orthologs based on finding a reciprocal best hit using the BLAST program, protein functional domains based on the InterProScan system and a Gene Ontology gene association based on the ipr2go mapping. Users can browse and analyze these data using the TableView.

Useful links were added to the GeneView external links section, for example, to the web sites for Gclust (8) and MBGD (9) (automated ortholog gene cluster databases) and Fluorome (10) (a database of a large-scale analysis of chlorophyll fluorescence kinetics). Moreover, it is possible to link many more external databases via KEGG/GENES (11) and LinkDB (12).

WEB SERVICES

CyanoBase provides a URL-based REST web-service interface for reusing data with other tools and computer programs. Data are available in several file formats: tab-delimitered text, CSV, FastA and gff3. Tools such as Galaxy (13), BioMart (14) and spreadsheet programs are able to import the data seamlessly. The URLs are indicated by the orange-colored icons and are specified in the KazusaAPI section on the relevant pages.

The SearchView page has a special export function for a set of genes in the search results. Users can easily obtain the gene set in plaintext format. This gene set export function also allows users to obtain the set of genes belonging to a gene category on the GeneCategoryView page.

CyanoBase also provides alternative ways to export sequence and gene annotation data. KazusaMart (http://mart.kazusa.or.jp), a BioMart system, is able to filter and export the data. Also, a Gbrowse service can be used to select and export the genome sequence and the features, with web and DAS interfaces (5).

ACKNOWLEDGEMENTS

The authors thank Mr Manabu Yamada for generating the data of CyanoBase. The authors thank Dr Paul Horton for reviewing the draft manuscript.

FUNDING

Kazusa DNA Research Institute Foundation. Funding for open access charge: Kazusa DNA Research Institute Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Kaneko,T., Sato,S., Kotani,H., Tanaka,A., Asamizu,E., Nakamura,Y., Miyajima,N., Hirosawa,M., Sugiura,M., Sasamoto,S. *et al.* (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.*, **3**, 109–136.
2. Nakamura,Y., Kaneko,T., Hirosawa,M., Miyajima,N. and Tabata,S. (1998) CyanoBase, a www database containing the complete nucleotide sequence of the genome of *Synechocystis* sp. *strain PCC6803*. *Nucleic Acids Res.*, **26**, 63–67.
3. 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.
4. Nakamura,Y., Kaneko,T., Miyajima,N. and Tabata,S. (1999) Extension of CyanoBase. CyanoMutants: repository of mutant information on Synechocystis sp. strain PCC6803. *Nucleic Acids Res.*, **27**, 66–68.
5. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
6. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
7. Sato,S., Shimoda,Y., Muraki,A., Kohara,M., Nakamura,Y. and Tabata,S. (2007) A large-scale protein protein interaction analysis in *Synechocystis* sp. *PCC6803*. *DNA Res.*, **14**, 207–216.
8. Sato,N. (2009) Gclust: trans-kingdom classification of proteins using automatic individual threshold setting. *Bioinformatics*, **25**, 599–605.
9. Uchiyama,I. (2007) MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups. *Nucleic Acids Res.*, **35**, D343–D346.
10. Ozaki,H., Ikeuchi,M., Ogawa,T., Fukuzawa,H. and Sonoike,K. (2007) Large-scale analysis of chlorophyll fluorescence kinetics in *Synechocystis* sp. PCC 6803: identification of the factors involved in the modulation of photosystem stoichiometry. *Plant Cell Physiol.*, **48**, 451–458.
11. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
12. Fujibuchi,W., Goto,S., Migimatsu,H., Uchiyama,I., Ogiwara,A., Akiyama,Y. and Kanehisa,M. (1998) DBGET/LinkDB: an integrated database retrieval system. *Pac. Symp. Biocomput.*, 683–694.
13. Giardine,B., Riemer,C., Hardison,R.C., Burhans,R., Elnitski,L., Shah,P., Zhang,Y., Blankenberg,D., Albert,I., Taylor,J. *et al.* (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, **15**, 1451–1455.
14. Haider,S., Ballester,B., Smedley,D., Zhang,J., Rice,P. and Kasprzyk,A. (2009) BioMart Central Portal–unified access to biological data. *Nucleic Acids Res.*, **37**, W23–W27.