CYORF: Community Annotation of Cyanobacteria Genes

Miho Furumichi[1] miho@scl.kyoto-u.ac.jp
Yoko Sato[2] ysatoh@fqs.fujitsu.com
Tatsuo Omata[3] omata@nuagr1.agr.nagoya-u.ac.jp
Masahiko Ikeuchi[4] mikeuchi@bio.c.u-tokyo.ac.jp
Minoru Kanehisa[1] kanehisa@kuicr.kyoto-u.ac.jp

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan
[2] Fujitsu Kyushu System Engineering Limited, 2-2-1 Momochihama, Sawara-ku, Fukuoka 814-8589, Japan
[3] Graduate School of Bioagricultural Sciences, Nagoya University, Chigusa-ku, Nagoya 464-8601, Japan
[4] Graduate School of Arts and Sciences, University of Tokyo, Meguro-ku, Tokyo 153-8902, Japan

Keywords: database, *Synechocystis* genome, cyanobacteria, community annotation

1 Introduction

The complete genome sequence of the cyanobacterium *Synechocystis* sp. PCC 6803 was determined in 1996 by the Kazusa DNA Research Institute. In comparison to other bacterial species, such as *Escherichia coli* and *Bacillus subtilis*, the *Synechocystis* genome contained a larger proportion of unknown genes, because cyanobacteria and their genes had not been well studied despite their importance in the evolution of life and the maintenance of biosphere. Thus, the availability of the genome sequence and a complete set of genes was a great boon to cyanobacteriologists, accelerating their research and resulting in a number of publications. However, accelerated research poses a database problem. When new gene functions are identified, the published result is stored in PubMed, but not in any sequence database. The lack of an up-to-date, well-annotated database is not limited to *Synechocystis*; it is a problem for the majority of the prokaryotic genomes thus far sequenced. The primary repositories of GenBank, EMBL, and DDBJ are not well maintained because the genome entries can only be updated by the sequencing teams who submitted the original data, but usually they are too busy doing next genomes. The providers of annotated databases such as SWISS-PROT and KEGG are simply unable to keep up with the rapidly increasing amount of data.

2 Community Annotation Database

Perhaps, the only solution to the current database problem is to get the research community actively involved in the annotation process. We have developed a database system named CYORF using the PostgreSQL relational database system with a Web interface for data annotation, so that anyone in the research community can make modifications to the database. The annotation involves editing of gene name, definition, and comment fields based on published results, which are linked to PubMed. Because CYORF incorporates various capabilities of KEGG and DBGET at the GenomeNet, it is possible to examine, for example, orthologs in other species, paralogs within the species, Pfam and Prosite motifs present, neighbor genes on the chromosome, and positions in the KEGG pathways, as well as to predict localization sites and protein 3D structures, or to search similar sequences in cyanobacterial draft genomes. The public (read-only) version of CYORF with all these capabilities is available at `http://cyano.genome.ad.jp/` (Figure 1).

The CYORF database was first released in May 2002 as a product of the three-way collaboration among genome scientists at Kazusa who had provided an updated list of ORFs for the *Synechocystis* genome, biologists in the Cyanobacteria DNA Chip Consortium who would be willing to do annotations, and bioinformaticians in Kyoto who had expertise in developing and maintaining the GenomeNet database resource.

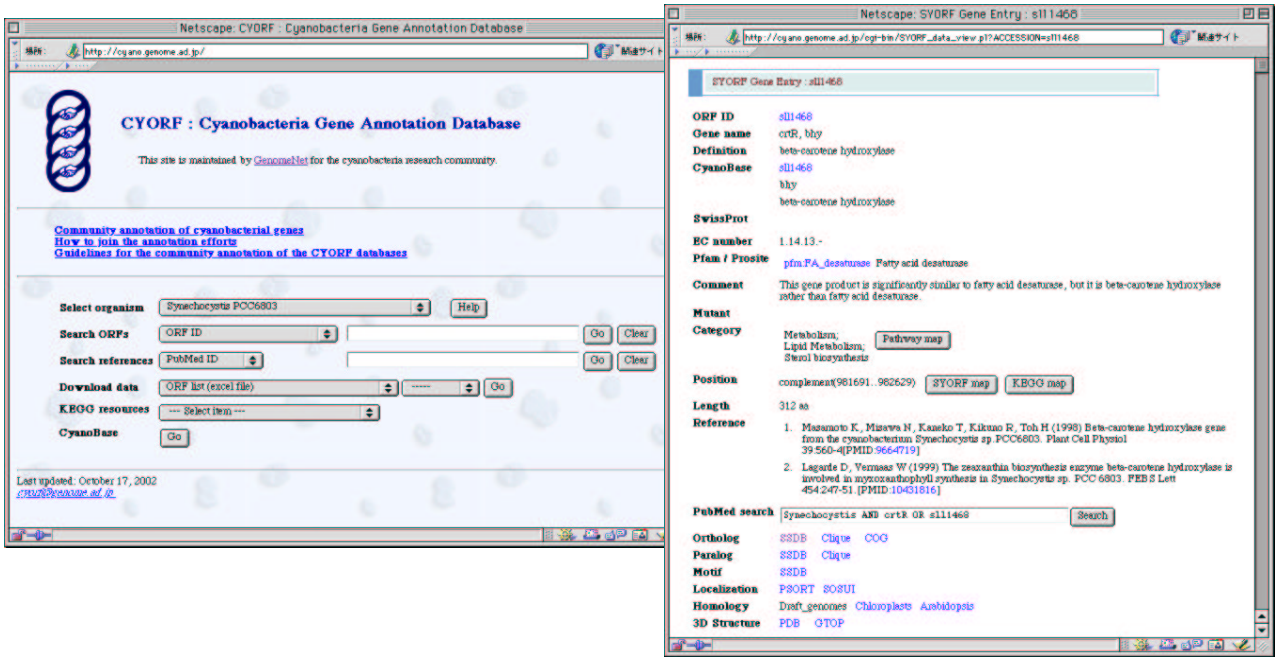

Figure 1: The public version of CYORF.

Although the first release was made available by the Japanese community reannotating about one thousand genes, the database is now open to the international community. Once the annotators of CYORF access the annotators' version of CYORF with the username and password, they can add new annotations to any ORF on-line. The results will be immediately incorporated in the public version of CYORF with their name shown as the last annotator. Weekly updates of CYORF are sent to Kazusa for use in updating their CyanoBase. As of October 2002 there are 30 researchers, 25 Japanese and 5 from abroad, who are actively involved in the community annotation of *Synechocystis* and also *Anabaena* (Table 1). We plan to expand the resource to include all completely sequenced cyanobacterial genomes and to improve system capabilities for cross-species comparisons and annotations.

Table 1: Update Status of the CYORF database (as of October 30, 2002).

Species	Number of all ORFs	Number of reannotated ORFs	Reference
Synechocystis PCC6803	3267	1106	[2]
Anabaena PCC7120	6132	119	[1]
Thermosynechococcus elongatus BP-1	2475	3	[3]

References

[1] Kaneko, T. *et al.*, Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120, *DNA Res.*, 8:205–213, 2001.

[2] Kaneko, T. *et al.*, 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, 1996.

[3] Nakamura, Y. *et al.*, Complete genome structure of the thermophilic cyanobacterium *Thermosynechococcus elongatus* BP-1, *DNA Res.*, 9:123–130, 2002.