CyanoBase, a www database containing the complete nucleotide sequence of the genome of *Synechocystis* sp. strain PCC6803

Yasukazu Nakamura*, Takakazu Kaneko, Makoto Hirosawa, Nobuyuki Miyajima and Satoshi Tabata

Kazusa DNA Research Institute, 1532-3 Yana, Kisarazu, Chiba 292, Japan

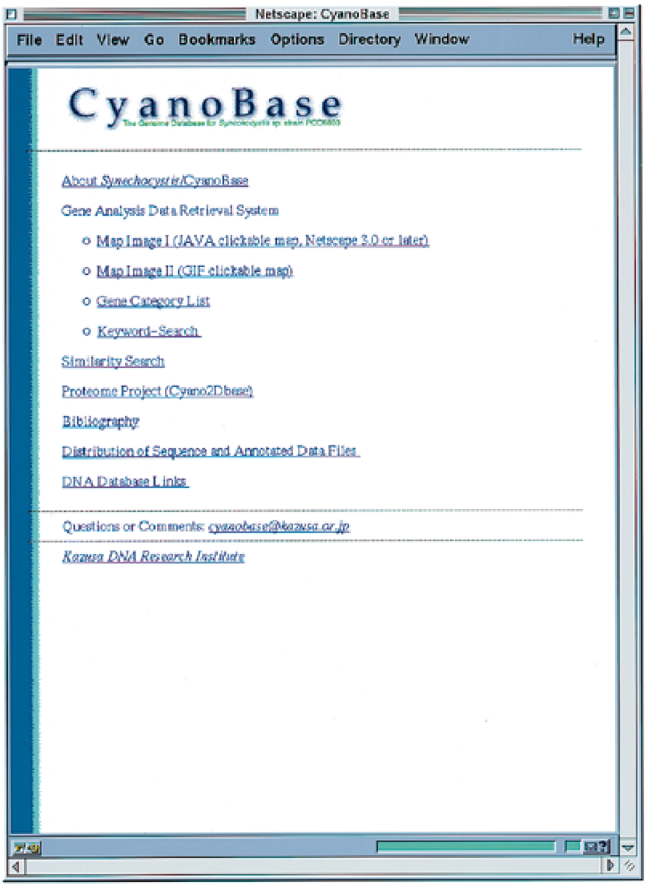
Received October 6, 1997; Accepted October 8, 1997

ABSTRACT

CyanoBase (http://www.kazusa.or.jp/cyano/) is a database containing genomic information on the cyanobacterium *Synechocystis* sp. strain PCC6803. It furnishes an annotation to each of the 3168 protein genes deduced from the entire nucleotide sequence of this genome. Information on the genome can be directly accessed through three different menus: a clickable physical map of the genome, a gene classification list, and a keyword search menu, all of which are accessible from the main page of the database. The entry page for a gene annotation contains the following information: the location of the gene on the genome, the nucleotide and deduced amino acid sequence of the gene, the result of a similarity search, and the classification of the deduced gene product according to its function. This page has reverse-links to the local physical map and gene classification list so that relevant genes can be searched in terms of their location on the genome and their function. In addition, the main page of CyanoBase provides engines for similarity searches between a query sequence and the entire genome sequence and for keyword searches, in addition to numerous links to pages containing related information.

INTRODUCTION

Cyanobacteria are prokaryotic microorganisms which carry a complete set of genes for oxygenic photosynthesis. Since their apparatus for photosynthesis resembles that of plants, they have been used as model organisms by many scientists to investigate the structure and function of plant-type photosynthesis. Cyanobacteria are also interesting organisms from an evolutionary viewpoint because they are believed to be of ancient origin and to have survived a number of major changes in the earth's environment. Moreover, it is widely held that plant chloroplasts evolved from cyanobacterial ancestors that had developed an endosymbiotic relationship with eukaryotic host cells. *Synechocystis* sp. strain PCC6803 is a unicellular cyanobacterium. It has been widely used for the study of the mechanism of photosynthesis because this strain has the ability to grow both photoautotrophically and photoheterotrophically (reviewed in 1), allowing disruption of the photosynthetic protein genes without lethality.



CyanoBase

The Genome Database for Synechocystis sp. strain PCC6803

About *Synechocystis*/CyanoBase

Gene Analysis Data Retrieval System

- Map Image I (JAVA clickable map, Netscape 3.0 or later)
- Map Image II (GIF clickable map)
- Gene Category List
- Keyword-Search

Similarity Search

Proteome Project (Cyano2Dbase)

Bibliography

Distribution of Sequence and Annotated Data Files

DNA Database Links

Questions or Comments: *cyanobase@kazusa.or.jp*

Kazusa DNA Research Institute

Figure 1. The main page of CyanoBase.

In order to determine all the genomic information of this organism, we initiated a genomic sequencing project and a web

*To whom correspondence should be addressed. Tel: +81 438 52 3935; Fax: +81 438 52 3934; Email: ynakamu@kazusa.or.jp

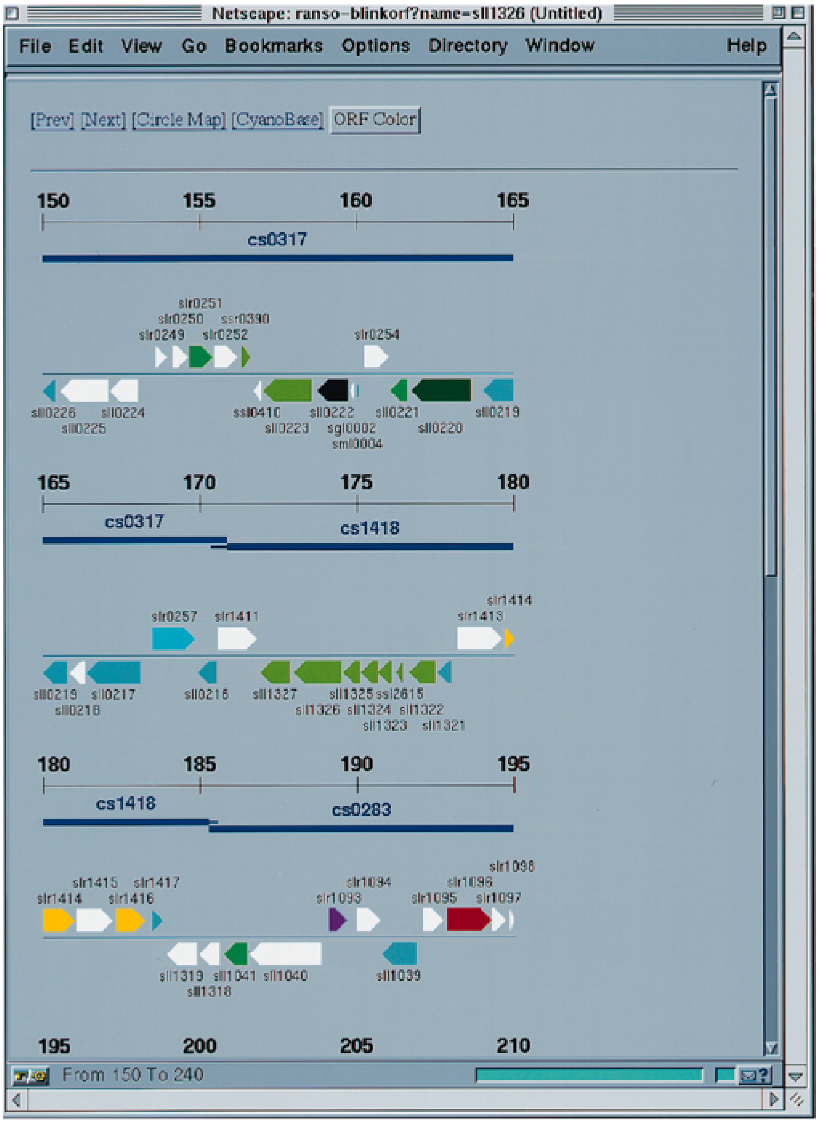


Figure 2. Java based map image of a local area. Each blue bar (a clone) or color-coded box (assigned gene) is a hyperlink to information on that area.

site was set up in 1995 (2) as a repository for the 818 potential protein genes deduced from a continuous 1.0 megabase sequence that represented all of the known sequence up until then (3). With completion of sequencing of the entire 3 573 470 bp genome in March 1996 (4), we developed the current version of CyanoBase, which contains annotations for 3168 potential protein genes. The aim of this database is to provide detailed information on potential protein genes using a user-friendly interface, including Java clickable genome maps and a hypertext classification list. We will continue to update the database by posting new information as it becomes available.

CONTENTS

Entrance from the main page

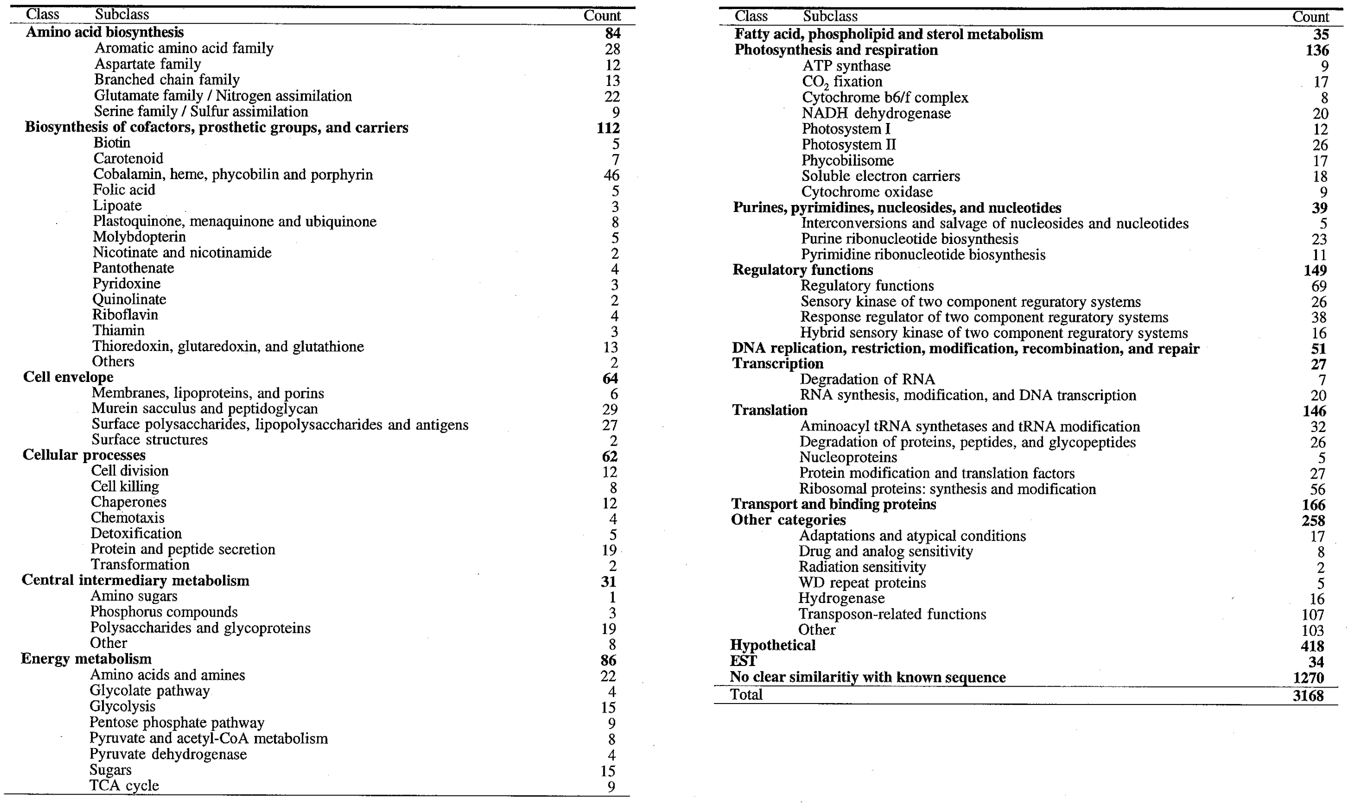
The URL of CyanoBase is http://www.kazusa.or.jp/cyano/ (Fig. 1).

Each potential gene assigned to the *Synechocystis* genome has a standard name for identification, as described previously (4). The standard name consists of a three-letter code where the first letter represents the species name (s: *Synechocystis*), the second, the length and/or the method of identification of the open reading frame (ORF) [l: longer than 300 bp, s: 150–297 bp, m: shorter than 150 bp, g: a gene predicted only by the computer program GenMark (5)], and the third, the reading direction (l or r: leftward or rightward). The three letter code is followed by a four-digit number. The standard name has been added to each CDS in flat files of the DDBJ/EMBL/GenBank databases according to the format '/note=standard_name:' tag, and all the genes are represented by their standard names in CyanoBase.

The annotation for each gene can be accessed through three menus on the main page of CyanoBase.

Physical maps of the genome. Map Image I and II show restriction maps of the circular genome (6) in Java and GIF formats,

Table 1. Summary of the classification of the genes on the genome of *Synechocystis* sp. strain PCC6803. The function of each gene was predicted through sequence comparison with genes of known function



Class	Subclass	Count
Amino acid biosynthesis		**84**
	Aromatic amino acid family	28
	Aspartate family	12
	Branched chain family	13
	Glutamate family / Nitrogen assimilation	22
	Serine family / Sulfur assimilation	9
Biosynthesis of cofactors, prosthetic groups, and carriers		**112**
	Biotin	5
	Carotenoid	7
	Cobalamin, heme, phycobilin and porphyrin	46
	Folic acid	5
	Lipoate	3
	Plastoquinone, menaquinone and ubiquinone	8
	Molybdopterin	5
	Nicotinate and nicotinamide	2
	Pantothenate	4
	Pyridoxine	3
	Quinolinate	2
	Riboflavin	4
	Thiamin	3
	Thioredoxin, glutaredoxin, and glutathione	13
	Others	2
Cell envelope		**64**
	Membranes, lipoproteins, and porins	6
	Murein sacculus and peptidoglycan	29
	Surface polysaccharides, lipopolysaccharides and antigens	27
	Surface structures	2
Cellular processes		**62**
	Cell division	12
	Cell killing	8
	Chaperones	12
	Chemotaxis	4
	Detoxification	5
	Protein and peptide secretion	19
	Transformation	2
Central intermediary metabolism		**31**
	Amino sugars	1
	Phosphorus compounds	3
	Polysaccharides and glycoproteins	19
	Other	8
Energy metabolism		**86**
	Amino acids and amines	22
	Glycolate pathway	4
	Glycolysis	15
	Pentose phosphate pathway	9
	Pyruvate and acetyl-CoA metabolism	8
	Pyruvate dehydrogenase	4
	Sugars	15
	TCA cycle	9
Fatty acid, phospholipid and sterol metabolism		**35**
Photosynthesis and respiration		**136**
	ATP synthase	9
	CO_2 fixation	17
	Cytochrome b6/f complex	8
	NADH dehydrogenase	20
	Photosystem I	12
	Photosystem II	26
	Phycobilisome	17
	Soluble electron carriers	18
	Cytochrome oxidase	9
Purines, pyrimidines, nucleosides, and nucleotides		**39**
	Interconversions and salvage of nucleosides and nucleotides	5
	Purine ribonucleotide biosynthesis	23
	Pyrimidine ribonucleotide biosynthesis	11
Regulatory functions		**149**
	Regulatory functions	69
	Sensory kinase of two component reguratory systems	26
	Response regulator of two component reguratory systems	38
	Hybrid sensory kinase of two component reguratory systems	16
DNA replication, restriction, modification, recombination, and repair		**51**
Transcription		**27**
	Degradation of RNA	7
	RNA synthesis, modification, and DNA transcription	20
Translation		**146**
	Aminoacyl tRNA synthetases and tRNA modification	32
	Degradation of proteins, peptides, and glycopeptides	26
	Nucleoproteins	5
	Protein modification and translation factors	27
	Ribosomal proteins: synthesis and modification	56
Transport and binding proteins		**166**
Other categories		**258**
	Adaptations and atypical conditions	17
	Drug and analog sensitivity	8
	Radiation sensitivity	2
	WD repeat proteins	5
	Hydrogenase	16
	Transposon-related functions	107
	Other	103
Hypothetical		**418**
EST		**34**
No clear similaritiy with known sequence		**1270**
Total		**3168**

respectively. When a given position on the map is clicked, a local map covering the corresponding 90 kb area appears (Fig. 2). In each local map, the positions of the cosmid and lambda phage clones used for sequence determination, and the long-PCR products used to close gaps between clones are shown by blue bars, and the assigned protein genes are designated by color-coded boxes under the bars. Each bar and box provides a link to detailed information on the corresponding clone or gene, as described in the next section.

Gene category list. 1864 genes out of the 3168 potential protein genes assigned to the *Synechocystis* genome showed similarity to sequences already registered in the public DNA or protein databases. These genes have been classified according to their predicted biological function (3,4), as summarized in Table 1. On the web page, a table of gene classification is presented in a hierarchical manner, with a link to the annotation to each gene (Fig. 3).

Keyword-Search. A keyword search box is provided to allow a search for information about the gene. This is achieved by either submitting the gene name (a three-letter genetic name), the name of the gene product, or a standard name.

Annotation—information on predicted protein genes

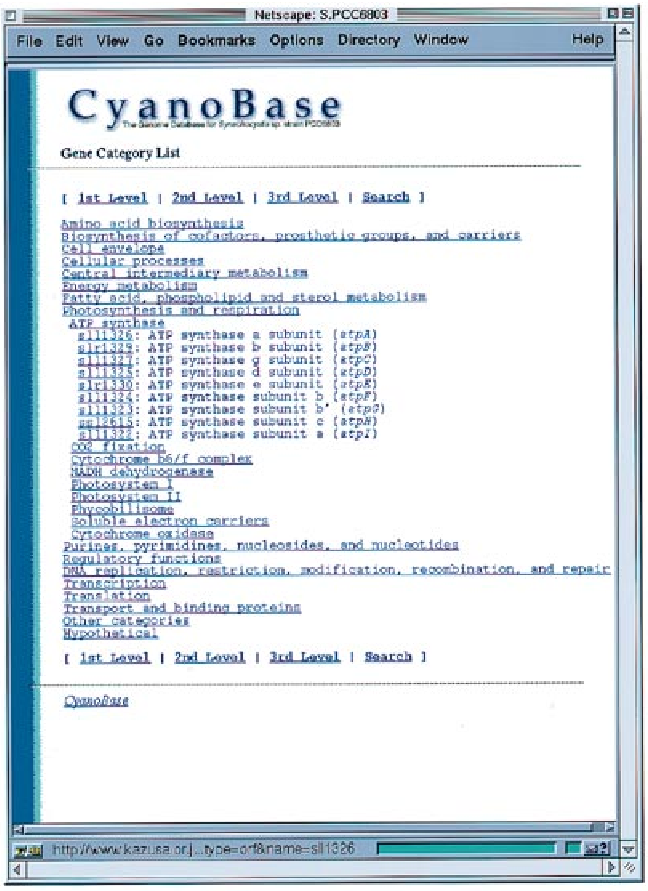
A sample image of an annotation page to a potential protein gene is shown in Figure 4.

Location. Nucleotide positions of the initiation and termination points of a coding region and the coding direction on the physical map are indicated. A reverse-link to a linear physical map is provided which enables users to obtain information about genes in adjacent regions.

Sequence retrieval links. Both the nucleotide and amino acid sequences of a gene can be obtained through the links. For nucleotide sequence information, two input boxes are provided to allow specification of the initiation and termination positions of the sequence to be retrieved. The positions of the first and the last nucleotides of a coding region are given as default numbers.

Results of similarity searches. A link to the results of a Smith–Waterman search for each sequence in the protein sequence database (MPSRCH result) and summary information on the most similar sequence are provided.

Classification. The gene category to which a gene of interest belongs is presented in a hierarchical manner. A link to a gene



CyanoBase
The Genome Database for Synechocystis sp. strain PCC6803

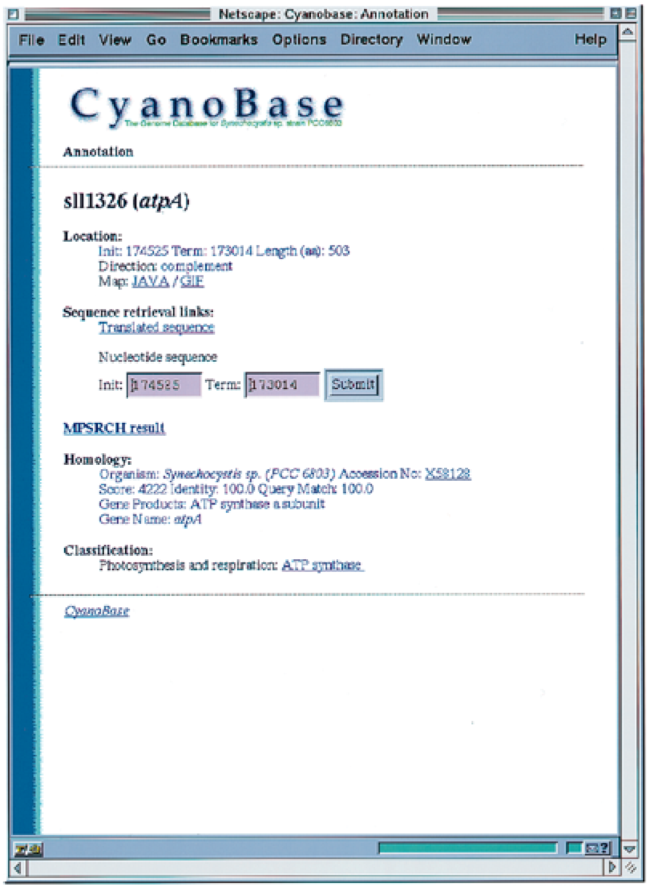
Gene Category List

[1st Level | 2nd Level | 3rd Level | Search]

Amino acid biosynthesis
Biosynthesis of cofactors, prosthetic groups, and carriers
Cell envelope
Cellular processes
Central intermediary metabolism
Energy metabolism
Fatty acid, phospholipid and sterol metabolism
Photosynthesis and respiration
 ATP synthase
  sll1326: ATP synthase a subunit (*atpA*)
  slr1329: ATP synthase b subunit (*atpB*)
  sll1327: ATP synthase g subunit (*atpC*)
  sll1325: ATP synthase d subunit (*atpD*)
  slr1330: ATP synthase e subunit (*atpE*)
  sll1324: ATP synthase subunit b (*atpF*)
  sll1323: ATP synthase subunit b' (*atpG*)
  ssl2615: ATP synthase subunit c (*atpH*)
  sll1322: ATP synthase subunit a (*atpI*)
 CO2 fixation
 Cytochrome b6/f complex
 NADH dehydrogenase
 Photosystem I
 Photosystem II
 Phycobilisome
 Soluble electron carriers
 Cytochrome oxidase
Purines, pyrimidines, nucleosides, and nucleotides
Regulatory functions
DNA replication, restriction, modification, recombination, and repair
Transcription
Translation
Transport and binding proteins
Other categories
Hypothetical

[1st Level | 2nd Level | 3rd Level | Search]

CyanoBase



CyanoBase
The Genome Database for Synechocystis sp. strain PCC6803

Annotation

sll1326 (*atpA*)

Location:
Init: 174525 Term: 173014 Length (aa): 503
Direction: complement
Map: JAVA / GIF

Sequence retrieval links:
Translated sequence

Nucleotide sequence

Init: 174525 Term: 173014

MPSRCH result

Homology:
Organism: *Synechocystis sp. (PCC 6803)* Accession No: X58128
Score: 4222 Identity: 100.0 Query Match: 100.0
Gene Products: ATP synthase a subunit
Gene Name: *atpA*

Classification:
Photosynthesis and respiration: ATP synthase

CyanoBase

Figure 3. The hypertext-based gene classification list in CyanoBase. A class (Photosynthesis and respiration) and a subclass (ATP synthase) are depicted as examples.

Figure 4. An annotation page for a deduced protein gene (sll1326).

category list provides information on other genes with the same or similar biological function.

Other services on the main page

Similarity search. An input form for a similarity search is provided. Users can choose one out of two combinations of program and reference sequence: a BLASTP program with the library for 3168 deduced protein sequences and a BLASTN program with the library for the complete nucleotide sequence of the entire genome of *Synechocystis*.

Links. (i) Proteome project. A link to the main page of the proteome project of *Synechocystis*. (7) is provided. (ii) Distribution of Sequence and Annotated Data Files (ftp link). The following files are provided for distribution; two concatenated files containing the nucleotide and deduced amino acid sequences of the assigned 3168 protein genes, a file in either flat file or Macintosh Excel format containing a table summarizing the annotations to 3168 potential protein genes, a file containing the complete nucleotide sequence of the genome, and the file ‘cyanoace’ which contains both the genomic sequence and annotated information in ACeDB format (8). (iii) DNA Database Links. This is a link to the nucleotide sequences that have been submitted to the international DNA databases. Since 3 573 470 bp of consecutive sequence exceeds the size limit set for a single LOCUS in the international DNA databases, the sequence was divided into 27 files and each file was separately registered. Information can be obtained for each entry through the DBGET integrated database retrieval system on the GenomeNet WWW Server. (iv) Bibliography. This provides information on related publications through links to the PubMed publication search service at NCBI/NLM.

IMPLEMENTATION

Map interfaces and the presentation of the annotation to each gene were implemented in the JMGD system with the Sybase SQL server system 11 as a DBMS. JMGD was developed by N.M. and is available from http://www.kazusa.or.jp/jmgd/ . The package includes the Java source codes and a set of perl scripts. The ACeDB database file which is the core portion of CyanoBase was prepared separately. The data for ACeDB can also be obtained through our anonymous ftp server.

ACKNOWLEDGEMENTS

We thank Drs Y.Akiyama and S.Goto (Institute for Chemical Research, Kyoto University) for providing us with a browser for the gene category list. The browser was originally developed for KEGG (Kyoto Encyclopedia of Genes and Genomes, http://www.genome.ad.jp/kegg/kegg.html). This work was supported by the Kazusa DNA Research Institute Foundation.

REFERENCES

1 Packer,L. and Glazer,A.N. (eds) (1988) *Methods Enzymol.* **167**.
2 Hirosawa,M., Kaneko,T. and Tabata,S. (1995) in Hagiya,M., Miyano,K., Nakai,K., Suyama,A., Yokomori,T. and Takagi,T. (eds), *Proceedings of Genome Informatics Workshop 1995.* Universal Academy Press, Tokyo. pp. 102–103.
3 Kaneko,T., Tanaka,A., Sato,S., Kotani,H., Sazuka,T., Miyajima,N., Sugiura,M. and Tabata,S. (1995) *DNA Res.* **2**, 153–166.
4 Kaneko,T., Sato,S., Kotani,H., Tanaka,A., Asamizu,E., Nakamura,Y., Miyajima,N., Hirosawa,M., Sugiura,M., Sasamoto,S., *et al.* (1996) *DNA Res.* **3**, 109–136.
5 Borodovsky,M. and McIninch,J. (1993) *Comput. Chem.*, **17**, 123–133.
6 Kotani,H., Kaneko,T., Matsubayashi,T., Sato,S., Sugiura,M. and Tabata,S. (1994) *DNA Res.* **1**, 303–307.
7 Sazuka,T. and Ohara,O. (1997) *Electrophoresis* **18**, 1252–1258.
8 Durbin,R. and Mieg,J.T. (1991) ftp://ncbi.nlm.nih.gov/repository/acedb/