Development of Community Annotation Databases for Linking Genomes to Cellular Functions

Miho Furumichi[1] miho@scl.kyoto-u.ac.jp
Yoko Sato[2] ysatoh@fqs.fujitsu.com
Toshiaki Katayama[3] katayama@kuicr.kyoto-u.ac.jp
Shuichi Kawashima[1] shuichi@kuicr.kyoto-u.ac.jp
Minoru Kanehisa[1] kanehisa@kuicr.kyoto-u.ac.jp

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan
[2] Fujitsu Kyushu System Engineering, 2-2-1 Momochihama, Sawara-ku, Fukuoka 814-8589, Japan
[3] Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan

Keywords: community database, KEGG, cyanobacteria, *B. subtilis*, *E. coli*

1 Introduction

The roles of biological databases are changing in two important aspects. First, the data content about molecular structures and molecular functions is certainly not sufficient for the emerging field of systems biology or the Genomes to Life initiatives. It is necessary to somehow capture higher-level structures and functions, such as molecular interaction networks and cellular processes, and create new databases that integrate both molecular and higher-level information. Secondly, although the current form of data submissions by individual authors works well for the sequence and 3D structure databases, it is not appropriate for collecting higher-level biological knowledge from the research community. For the first aspect we have been developing KEGG [2], which integrates cellular functions with genomes and chemistry. For the second aspects we are promoting 'community database' where individual researchers in the community not only obtain information from the database, but also enter their knowledge into the database, so that the community as a whole can share most up-to-date information and knowledge. From the viewpoint of a database provider, it is most effective to get specialists in the research community actively involved in the annotation process in order to collect higher-level biological knowledge. We have developed community database systems named BSORF (`http://bacillus.genome.ad.jp`) and CYORF (`http://cyano.genome.ad.jp`) for *Bacillus subtilis* and cyanobacteria research communities, respectively. Furthermore, we are ready for the ECORF database (`http://ecoli.genome.ad.jp`) for *Escherichia coli* research community. We plan to develop other community databases using the same framework and share information with different communities.

2 Results and Discussion

Because BSORF and CYORF incorporate various capabilities of the KEGG and DBGET [1] systems at the GenomeNet (`http://www.genome.ad.jp`), it is possible to examine, for example, biological pathways, ortholog clusters, and conserved operon structures. BSORF/CYORF have been community database for gene annotations, but now they are being expanded to integrate functional genomics and proteomics data and to infer higher-order functions based on the technologies developed for KEGG.

In *B. subtilis*, various types of data are accumulated by systematic experiments, including DNA array data, mutant information, and northern blotting data, which are all released as part of BSORF, The BSTF database for transcriptional factors is also linked to/from BSORF. Reference information is being added to the database and Two-hybrid experiment data will also be added in the future. On the other hand, CYORF is focused on comparative genomics, because genome sequencing projects are completed or in progress for a number of different cyanobacterial species. Currently there are 9 species in the database and we will expand the resource to include all completely sequenced cyanobacterial genomes and to improve system capabilities for cross-species comparisons and annotations. In the case of the new ECORF database, we plan to incorporate rich biological knowledge about *E.coli* which will be linked to other organisms through the manually annotated KO (KEGG Orthlogy) database and the computationally derived SSDB (Sequence Similarity DataBase) / OC (Ortholog Cluster) database.

Table 1: Current status of our community databases.

	CYORF	BSORF	ECORF
Organisms	*Synechocystis* *Anabaena* *T.elongatus* *G.violaceus* *P.marinus* (3) *Synechococcus* (2)	*B.subtilis*	*E.coli* MG1655 *E.coli* W3110
Community annotation	International community	Japanese community	Will be released
DNA array data	In preparation	Released	In preparation
Two -hybrid data	In preparation	In the planning stage	
Transcription factors and signals	In preparation	BSTF, DBTBS	
Genome comparison tool	Released	-	
Hierarchical classifications of gene functions	KO	KO	

3 Acknowledgments

This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan, The Japan Society for the Promotion of Science, and the Japan Science and Technology Corporation. The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

References

[1] Fujibuchi, W., Goto, S., Migimatsu, H., Uchiyama, I., Ogiwara, A., Akiyama, Y., and Kanehisa, M., DBGET/LinkDB: an integrated database retrieval system, *Proc. Pacific Symp. Biocomputing '98*, World Scientific, 683–694, 1998.

[2] Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A., The KEGG databases at GenomeNet, *Nucleic Acids Res.*, 30:42–46, 2002.