Construction of the *Bacillus subtilis* ORF Database (**BSORF DB**)

A. Ogiwara [1]
M. Watanabe [3]
ogi@nibb.ac.jp
mari@ims.u-tokyo.ac.jp

N. Ogasawara [2]
T. Takagi [3]
nogasawa@bs.aist-nara.ac.jp
takagi@ims.u-tokyo.ac.jp

[1] National Institute for Basic Biology,
38 Nishigounaka, Myoudaiji-cho, Okazaki 444 Japan

[2] Graduate School of Biological Science,
Nara Institute of Science and Technology,
8916-5 Takayama-cho, Ikoma, Nara 630-01 Japan

[3] Human Genome Center,
Institute of Medical Science, The University of Tokyo,
4-6-1 Shirokanedai, Minato-ku, Tokyo 108 Japan

1 Introduction

The BSORF DB, *Bacillus subtilis* Open Reading Frames database, was established in order to release the results of the international sequencing project of Bacillus subtilis genome. The international project is being carried out under collaboration of Japanese and European research groups. The Japanese groups are responsible to 7 regions, which cover about 30% of the whole genome, in total.

The BSORF DB was firstly released with the sequencing result of one Japanese responsible region of length 215kb. It is provided by WWW (`http://bacillus.genome.ad.jp/ BSORF-DB.html`) [1].

As the progress of the sequencing project, the Japanese responsible regions have almost been finished. Now, the BSORF DB has been improved to accept multiple contigs that correspond to the corresponding regions. We have also made an overall revision in the procedures that accept initial data and convert into html files for the WWW service.

2 Database Construction

Database constructor programs consist of several modules written in Perl version 5, and many of the modules are made in conformity with Perl 5 object oriented style. For example, BSORFaccess class was defined to access BSORF internal data files and to set/refer values of each field of the BSORF entry.

The construction of the BSORF DB begins with the data conversion of a flat file entry, and ends with the generation of HTML documents or GIF images. Since the determined sequences are also to be deposited to the public database like EMBL or DDBJ, it is convenient to take in the data in flat file format. In the current version, both GenBank and EMBL formats are supported as a data source. An entry including one contig of the genomic sequence and having CDS items that correspond to each

ORF can be acceptable to the BSORF DB. Thus, the BSORF system is easily applicable to other genome data.

Of course, converting a public database entry into BSORF format is not sufficient to generate the BSORF DB. There are BSORF specific data items such as ‘products’, ‘functions’, ‘category’, ‘homology’, and so on. Such kind of information is annotated by each investigator of the sequencing project. The BSORF constructor includes the annotation workbench, which is also accessed using WWW. The access to the editors’ workbench is restricted to the project member.

Other kind of annotation data, like physicochemical properties and sequence motifs, are automatically calculated by the constructor system. The results of homology search, as well as the above data, are integrated into the BSORF entries.

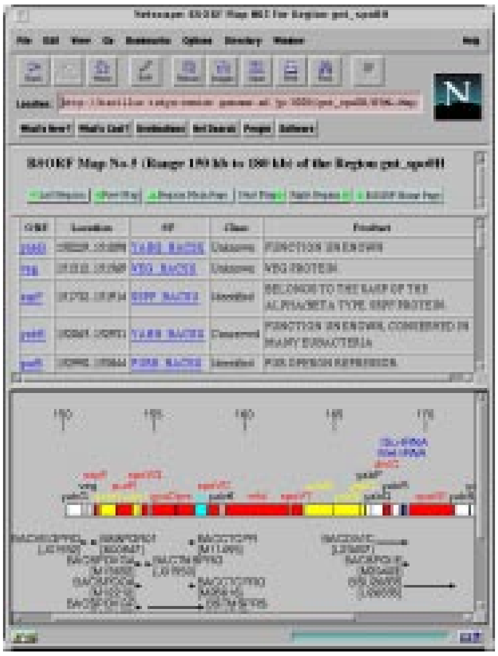
Finally, the entries are converted into HTML for the presentation. In the current WWW mechanism, it is more efficient to keep all the data in HTML beforehand, rather than generating the HTML on the fly, as far as the number of queries is limited and the answers are static. Some data, e.g. maps of ORF position, are represented graphically. These are represented in GIF format and are also automatically generated in the construction procedure.

3 Classification and Analysis

Classification of a gene into certain category is a good way to clearly describe the function. However, it is difficult to classify the whole biological functions into the complete and disjoint categories, for the criterion may not be unique.

We started the classification with simple cases. Since the metabolic system is one of the most clearly classified system, we firstly picked up ORFs that are describe to have an enzyme activity. The KEGG database (`http://www.genome.ad.jp/kegg/kegg.html`) constructed in Kyoto University provides good references of the metabolic system.

Integration of various sequence analysis method such as prediction structural topology, prediction of localization, and comparison with other genomes are being carried out. These results will be expected to give some useful suggestions for the functional analysis of the *Bacillus* genome, which will become the main theme in the next stage after the sequencing project will have been completed.



BSORF Map No.5 (Range 150 kb to 180 kb) of the Region gnt_spo0H

ORF	Location	SP	Class	Product
[illegible]	[illegible]	[illegible]	Unknown	FUNCTION UNKNOWN
veg	[illegible]	VEG_BACSU	Unknown	VEG PROTEIN
sspF	[illegible]	SSPF_BACSU	Identified	BELONGS TO THE SASP OF THE ALPHA/BETA TYPE SSPF PROTEIN
[illegible]	[illegible]	[illegible]	Conserved	FUNCTION UNKNOWN, CONSERVED IN MANY EUBACTERIA
purR	[illegible]	PURR_BACSU	Identified	PUR OPERON REPRESSOR

References

[1] A. Ogiwara, T. Takagi & N. Ogasawara, “A WWW Database of *Bacillus Subtilis* ORFs determined by the International Project of Sequencing *B. Subtilis* Genome”, *Proc. Genome Informatics Workshop 1995*, pp. 162-163, 1995.