# Comprehensive Sequence Analysis of *B. subtilis* Genome Using the BSORF Database

**Atsushi Ogiwara** [1]
ogi@nibb.ac.jp

**Naotake Ogasawara** [2]
nogasawa@bs.aist-nara.ac.jp

**Mari Watanabe** [3]
mari@ims.u-tokyo.ac.jp

**Toshihisa Takagi** [3]
takagi@ims.u-tokyo.ac.jp

[1] National Institute for Basic Biology
38 Nishigounaka, Myoudaiji, 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

**Abstract**

The sequencing of the whole *Bacillus subtilis* genome has been completed. The genome contains about 4,000 ORFs, and information of these ORFs are stored in **BSORF** database, which was formerly reported as a database of Japanese groups' achievements [1, 2]. Using the complete genome database, we performed comprehensive sequence analyses based on the sequence similarity. First one is clustering of paralogous genes, and another one is searching orthologous genes in *Escherichia coli* and *Synechocystis sp.* genomes. A part of these results were used to classify *B. subtilis* genes into several functional categories. The KEGG pathway database was applied to classify candidates of enzymes for intermediate metabolisms [3]. These results are also stored in the BSORF database in order to assist the functional analysis of *B. subtilis* genome.

## 1 Clustering of Paralogues

Though bacterial genome contains several thousand of genes, variety is not so much. For example, there are many transport system of small molecules on bacterial cell membrane. Each of them corresponds to the target molecule, however, there are only limited number of types of the transporter system. These genes are thought to constitute paralogues. Searching and classification of paralogous genes is expected to give us useful information about the constitution of the organism.

Amino acids sequences of 4,020 ORFs determined in *B. subtilis* genome were compared each other by BLASTP and FASTA program, and pairs of ORFs that were mutually more similar than certain threshold were selected. Then, they were clustered using single-linkage algorithm. Various sets of parameters were tested during the selection of similar pairs and the clustering, and found that P-value (or similarity score) was the most dominant. We finally got 471 clusters with $10^{-10}$ P-value cutoff and some additional parameters of identity and alignment length. About a half of ORFs belonged to certain cluster, i.e., about a half seemed to be orphan or unknown genes.

Though strong relation was observed in many clusters to biological functions, it did not always correspond one to one. For example, almost all regulators of two component system seemed to consist a single cluster, however, sensor kinases broke up to some clusters. Other prominent clusters were transporters, transcription regulators, and some enzymes like dehydrogenases.

## 2 Searching for Orthologous genes

Various bacterial genomes have been determined since 1995, and comparison with other organisms also give us useful suggestions for genes yet to be known.

Searching for orthologous genes was achieved basically the same method as the paralog clustering. We searched similar sequence pairs between *B. subtilis* ORFs and *Escherichia coli* ORFs or *Synechocystis sp.* ORFs, using BLASTP. In this procedure, selection by the threshold P-value was also the most effective. We could obtain at last 1,726 candidate orthologues to *E. coli*, and 1,366 candidates for *Synechocystis*, in *B. subtilis* genome. These orthologous genes spread almost uniformly along the genome, and no global correlation could be observed in the arrangement of orthologues. But some operon structures were observed to be conserved in order of the component genes.

## 3 Application to the Functional Classification

Finally, we tried to find out genes that code enzymes. Information of orthologues were used to assign the EC number to candidate enzymes. We checked the results by marking the corresponding enzymes on the KEGG pathway database. Since some essential pathway like glycolysis and TCA cycle are known to exist, we checked the completeness or closure of the essential pathway.

These results, paralogous genes, orthologous genes, and correspondence to some pathway database, were implemented into the newly released BSORF database.

## Acknowledgments

We thank Prof. Kanehisa and the KEGG project team for the metabolic pathway database. Mr. Sato, the chief of the programming team of the KEGG project, kindly provided us useful interface for the pathway database. The computation time was provided both by the Human Genome Center, The University of Tokyo, and the Supercomputer Laboratory, Kyoto University.

## 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.

[2] A. Ogiwara, N. Ogasawara, M. Watanabe & T. Takagi, "Construction of the *Bacillus subtilis* ORF database (**BSORF DB**)", *Proc. 7th Workshop on Genome Informatics,* pp. 228–229, 1996.

[3] H. Bono, H. Ogata, S. Goto & M. Kanehisa, "Genome scale prediction of enzyme genes utilizing the knowledge of metabolic interactions.", *Proc. 7th Workshop on Genome Informatics,* pp. 252–253, 1996.