Systematic Classification of *B. subtilis* Genes and Construction of a Knowledge Base to Represent Functional Information

Atsushi Ogiwara
ogi@nibb.ac.jp
National Institute for Basic Biology
38 Nishigounaka, Myoudaiji, Okazaki 444-8585, Japan

1 Introduction

Bacills subtilis is a typical Gram-positive bacterium whose whole genomic sequence was determined. Before the genome sequencing was completed, we had constructed a *B. subtilis* genome database called BSORF [2]. This was not only a database of sequences, but also to store as many functional annotations as possible. Since the complete genome sequence of *Bacillus subtilis* had been determined [1], we also tried to classify whole ORF products in the viewpoint of functional categories. We started with the clustering of paralogous genes [3], and using the result of paralog clustering, we annotated transporter genes using other features like motifs and operon structures. To rearrange the resulting annotation, I constructed a knowledge base system using Java language. Since the main focus of *B. subtilis* and other microbial genome projects was shifted to the functional analysis, genome-informatics should also treat the functional information. Here I'll present a framework for functional information representation that is suitable for WWW.

2 Classification of transporter genes

As reported previously [1], there seemed to exist 4,100 ORFs in *Bacillus subtilis* strain 168 genome. Paralog clustering revealed that about a half of them consisted of paralogues. The largest paralogous group, which contained 79 ORFs, corresponded to the ATP binding subunit of the ATP-binding cassette (ABC) transporters. Thus, to categorize the whole ORFs systematically, we started with classifying the transporter genes.

2.1 ABC transporters

The ABC transporter is known to consist of some components. (1) ATP-binding subunit, (2) membrane spanning subunit, and in some cases, (3) substrate specific binding subunit. At least, two ATP-binding subunits and two membrane spanning subunits are required. Each component consists of either homo-dimer or hetero-dimer. In the case of *B. subtilis*, genes of these components tend to constitute operons. i.e., these seem to be co-regulated on transcription. There are some other features: ATP-binding subunit having a specific ATP binding motif, and membrane spanning subunit having multiple membrane spanning features. Thus, by combining these features, I classified 79 candidate ATP-binding subunits of ABC transporters into 5 major groups. These groups seem to correspond fairly well to the substrate specificity, judging from the annotation of orthologues. In addition, sub-clustering of these ATP-binding subunits suggested that there seemed to be strong relationship between sequence homology and substrate specificity. It may be interpreted to reflect the history of paralogous evolution.

2.2 Phosphotransferase systems

Phosphotransferase system (PTS) consists of 3 components: a type I enzyme that transfers a phosphate group from phosphoenol pyruvate to an HPr component, and a type II enzyme that accepts the phosphate group from HPr, imports certain sugar and phosphorylates it. In the case of *B. subtilis*, operon structures are diversed especially for type II enzymes. There are 4 component in type II enzymes, but at least, component B and C seem to be essential. Though some PTS specific motifs and transmembrane features are observed, it was difficult to utilize the information of operon structure to discriminate the PTS system.

2.3 Multidrug resistance families

Many drug resistance family homologues were observed in *B. subtilis*. Some of them were ABC type transporters, like eucaryotic *mdr*. Other kinds of multidrug resistance family, like major facilitator superfamily (MFS) and small multidrug resistance family (SMR) were found. However, no specific motifs nor operon structure were observed in these families. Thus, I used only homology information and membrane spanning features to discriminate these classes.

3 Knowledge base system by Java language

To rearrange these results, I made a knowledge base system not only to store each fact, but also to represent the feature of certain function. I adopted Java language for knowledge representation. Since Java is a kind of object-oriented language, it is reasonable to represent a functional category as a class, and each genes as instances. Since Java was originally designed to construct a Web content, it is quite easy to construct a system which is suitable for the Web presentation. In the strict sense of the word of "knowledge base", the system might be inadequate to call the knowledge base because there is no general mechanism to achieve rule-based reasoning. But I implemented some "reasoning" mechanism in a class-dependent manner. For example, we can discriminate whether an operon belongs to the ABC transporter family or not.

Acknowledgements

The author thanks Professor Naotake Ogasawara in Nara Institute of Science and Technology for valuable suggestions and discussions. Staffs of the Human Genome Center, especially Ms Mari Watanabe assisted in constructing the database. This work was supported by a Grant-in-Aid for Scientific Research on Priority Area "Genome Science" from the Ministry of Education, Science, Sports and Culture of Japan. The computational time was provided by the Human Genome Center, Institute of Medical Science, the University of Tokyo, and by the Supercomputer laboratory, Institute for Chemical Research, Kyoto University.

References

[1] Kunst, F. *et al.*, The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*, *Nature*, 390:249–256, 1997.

[2] Ogiwara, A., Ogasawara, N., Watanabe, M., and Takagi, T., Construction of the *Bacillus subtilis* ORF database (BSORF DB), *Genome Informatics 1996*, Universal Academy Press, 228–229, 1996.

[3] Ogiwara, A., Ogasawara, N., Watanabe, M., and Takagi, T., Comprehensive Sequence Analysis of *B. subtilis* Genome Using the BSORF Database, *Genome Informatics 1997*, Universal Academy Press, 320–321, 1997.