The systematic sequencing of the
*Bacillus subtilis* genome in Japan

Naotake Ogasawara and Hiroshi Yoshikawa

Author for correspondence: Naotake Ogasawara. Tel: +81 7437 2 5430. Fax: +81 7437 2 5439. e-mail: nogasawa@bs.aist-nara.ac.jp

Graduate School of Biological Sciences, Nara Institute of Science and Technology, Ikoma, Nara 630-01, Japan

**Keywords**: *Bacillus subtilis*, genome project, genome sequencing

A European/Japanese co-operative programme for the systematic sequencing of the *Bacillus subtilis* genome is now in progress. It involves a large number of laboratories with interests in the molecular biology and physiology of *B. subtilis*. The research interests of these laboratories are diverse and include, for example, studies on DNA replication, transcription, recombination, protein synthesis and secretion, sporulation and cell division. In 1985, the number of genes identified in this bacterium, 535, was second only to the 1000 genes identified in *Escherichia coli*. As a result of discussions during the bi-annual International *Bacillus* Conferences it was agreed to seek funding for systematic sequencing of the whole genome. The initial genome project started in Europe within the framework of the EC-funded Science Programme (September 1989–December 1991). In response to a request from the European network, a Japanese project team was organized by N. Ogasawara and H. Yoshikawa who, by 1985, had already reported the sequencing of some 10 kb in the region of the replication origin of the *B. subtilis* chromosome (Moriya *et al.*, 1985). At that time, this was the longest sequenced bacterial chromosome segment.

The systematic sequencing of the *B. subtilis* genome was officially initiated in Japan in 1991. It was funded by a research grant from Monbusho (The Ministry of Education, Science, Culture and Sport of Japan) as part of the Japanese Human Genome Program (April 1991–May 1996). The *B. subtilis* project was included in the project as a model system for the development of techniques for large-scale, multi-laboratory sequencing. International co-operation was a condition for the receipt of funds from the programme. Five contiguous segments of the genome, totalling 1 Mb, and one separate segment of 0.25 Mb, were originally assigned to six Japanese groups. In addition, a new segment of 0.15 Mb was assigned to a seventh Japanese group. In total, the Japanese groups were therefore responsible for the sequencing of 1.4 Mb of the genome. By March 1996, the groups had completed the sequencing of more than 95% of their assigned regions. Ten papers resulting from the programme were published between 1993 and 1996 (Akagawa *et al.*, 1995; Fujishima & Yamane, 1995; Ogasawara *et al.*, 1994; Oyama *et al.*, 1995; Takemura *et al.*, 1995; Yamamoto *et al.*, 1996a; Yoshida *et al.*, 1994, 1995a, b, c), and further progress is reported in this issue (Mizuno *et al.*, 1996; Yamamoto *et al.*, 1996d; Yamane *et al.*, 1996; Yasumoto *et al.*, 1996; Yoshida *et al.*, 1996) and elsewhere (Yamamoto *et al.*, 1996b, c). Our entire sequence data will be released, by the end of September 1996, to the DDBJ/EMBL/GenBank data banks and to the *B. subtilis* ORF database that is under development in the Human Genome Center of Tokyo University (http://www.genome.ad.jp). This database includes a graphical representation of the disposition of ORFs, together with detailed information of their properties, including a summary of similarity searches.

The Japanese groups have followed a single sequencing strategy. λ linking clones, covering each of the assigned regions, were isolated using as probes previously sequenced and mapped genes, as well as *NotI* and *SfiI* linking clones isolated by Itaya & Tanaka (1991). Many regions were found to be difficult to clone into the λ phage vectors. We have largely overcome this problem by use of the inverse PCR method. Finally, the development of Long Range PCR has increased the rate of cloning and sequencing of the remaining regions. Sequence was determined by shotgun cloning the λ phage inserts or Long Range PCR products into a M13 vector. The target DNA was fragmented by digestion with DNase I. Inserts of sub-clones were amplified directly from plaques by PCR and their sequences were determined by an automatic dye sequencer. Sequence data from random sub-clones corresponding to a mean redundancy of about 6 were usually sufficient to cover most of the target fragments. The remaining gaps were filled using region-specific primers. To avoid errors, we routinely determined the sequences of both strands even when the reading from one strand seemed to be unambiguous. In addition, we inspected raw data from every sequence chromatograph and used sequence outputs only from regions where the base peaks were well separated.
Analysis of 1037 kb sequence from the Japanese Project (January, 1996) revealed 1007 putative ORFs, including 262 ORFs already characterized genetically or biochemically. The likely functions of 233 additional ORFs were suggested by the significant similarities of their products to known protein sequences in the current databases. It should be noted that no functional information could be deduced from the remaining 512 ORFs (51%). The Japanese project has now been oriented to undertake a systematic analysis of the functions of the ORFs located in the segments assigned to, and sequenced by, the Japanese groups. The project is included in the second phase of the Japanese Human Genome Program, a 5 year project that started in April 1996, and which is supported by a grant-in-aid for Special Project Research from Monbusho. Information obtained on the functional means will also be included in the B. subtilis ORF database mentioned above.

REFERENCES


Received 26 June 1996; accepted 13 August 1996.