Sequence Analysis of the 36-kb Region Between gntZ and trnY Genes of Bacillus Subtilis Genome

Yasuhiro Kasahara, Sumiko Nakai, and Naotake Ogasawara

Graduate School of Biological Sciences, Nara Institute of Science and Technology, 891-6, Takayama, Ikoma, Nara 630-01, Japan, and Department of Genetics, Osaka University Medical School, 2-2, Yamadaoka, Suita, Osaka 565, Japan

(Received 13 March 1997)

Abstract

Within the framework of an international Bacillus subtilis genome sequencing project, we have determined a 36-kb sequence covering the region between the gntZ and trnY genes. In addition to five genes sequenced and characterized previously, 27 putative protein coding sequences (open reading frame; ORF) were identified. A homology search for the newly identified ORFs revealed that six of them had similarities to known proteins. It is notable that new ORFs belonging to response-regulator aspartate phosphatase (Rap) and its regulator (Phr) families, and response regulator and sensory kinase families of two-component signal transduction systems have been identified. Furthermore, we found that some 180-bp non-coding sequence, that might be an remnant of an ancient IS element, is preserved in at least five loci of the B. subtilis genome.

Key words: Bacillus subtilis, gntZ, trnY, DNA sequencing

A European and Japanese cooperative program for the systematic sequencing of the Bacillus subtilis genome is in progress. The entire genome sequence is expected to appear in the middle of this year, and it will be the first genome sequence of a free-living Gram-positive bacteria. Our understanding of the biochemistry, physiology and genetics of B. subtilis is second only to that of Escherichia coli. B. subtilis cells become naturally competent under certain growth conditions. This unique property makes insertional mutagenesis of target genes much easier in B. subtilis, than that in E. coli. A new project for the systematic functional analysis of newly identified genes in the B. subtilis genome sequence has already started to understand the whole picture of the bacterial cell. Our group is responsible for sequencing the region between 344° (gntZ) and 11° (sigH) of the genetic map of B. subtilis. In a previous communication, we reported the 180-kb sequence between trnY (355°) and sigH. This paper deals with the cloning and sequence analysis of the remaining 36-kb region.

1. Cloning and Sequencing of the Region Between gntZ and trnY

We used a B. subtilis genomic DNA library described previously to isolate lambda phages containing inserts of the region between gntZ and trnY (Fig. 1). Two clones, D29 and D3, were obtained by screening the library using sequences of gntZ and trnY as probes, respectively. Then, a clone, D10, containing an insert covering the gap between inserts of D29 and D3, was isolated by a second screening of the library using the terminal fragments of the D29 and D3 inserts as probes.

The inserts of the three clones were amplified by long polymerase chain reaction (PCR), and subjected to shotgun sequencing as described previously. All the sequences were determined using the Dye-Primer Cycle sequencing Kit and 373A sequencer from Applied Biosystems (CA, USA). We have determined the sequences of both DNA strands, except for the regions whose sequence are already registered in data banks. The 36,201-bp sequence thus determined was deposited in the GSDB/DDBJ/EMBL/NCBI data banks with accession number D78193.
2. Analysis of Protein Coding Sequences (ORF) in the 36-kb Sequence

From the sequence data, all six possible reading frames were surveyed for protein coding sequences which started with ATG, GTG or TTG codons preceded by a putative Shine-Dalgarno (SD) sequence. Based on these criteria, 30 putative ORFs were identified (Table 1). In addition, we assigned as ORFs two regions where proteins more than 150 aa in size were encoded, although no apparent SD sequence preceded them. The location and direction of translation of the identified ORFs are schematically shown in Fig. 1, together with the location of putative transcription termination signals.

Five of the ORFs corresponded to genes sequenced and characterized previously: rocDEFR, encoding arginine utilization proteins and a regulator of their expression,\(^6,7\) and bglA, encoding 6-phospho-beta-glucosidase.\(^8\) In addition, characterization of the ahpCF operon, encoding alkyl hydroperoxide reductase subunits, has recently been published, after the completion of our sequencing.\(^9,10\)

Comparison of the aa sequences of the remaining putative ORFs with the non-redundant protein sequence database is summarized in Table 2. The detailed results of the search are available in BSORFDB data base (http://bacillus.tokyo-center.genome.ad.jp/). YydK protein showed similarities with transcriptional regulators of Escherichia coli and B. subtilis. The adjacent bglA gene was reported to be induced by aromatic beta-glucosides,\(^8\) and the involvement of the YydK protein in the bglA induction is being investigated. The YydI protein contains an ATP-GTP binding motif and showed similarities to probable ATP transport proteins of Staphylococcus. In addition, YydH and YydJ proteins are potential membrane proteins containing five transmembrane segments (by PSORT analysis, http://psort.nibb.ac.jp/). Thus, the yydHIJ
products seem to constitute an ABC transport system. YycR showed strong similarities to formaldehyde dehydrogenase (66.3% identity in 404 aa overlap) and formaldehyde dimutasate (54.9% identity in 406 aa overlap) of Pseudomonas putida, suggesting its involvement in formaldehyde metabolism in B. subtilis. The RapG protein is the seventh member of response-regulator aspartate phosphatase (Rap) family of B. subtilis. 11 Four of the rap genes reported previously are associated with a small gene, phosphatase regulator (phr), that encodes a peptide which may be secreted from the cell and serve as a quorum sensor by inhibiting the phosphatase activity. 11 The amino-domain of each Phr protein is very hydrophobic and the carboxyl domain is hydrophilic. The rapG gene is also followed by a small ORF (phrG) which encodes a polypeptide having the characteristics of the Phr family. These families are unique to B. subtilis and their known biological role is in the fine regulation of sporulation. Experiments are in progress to see if RapG and PhrG have any role in the regulation of sporulation. YycK protein belongs to HtrA/HhoA/HhoB family of serine protease. Three genes each belonging to this family have been identified in E. coli, Haemophilus influenzae, and Synechocystis sp. PCC6803. However, yycK is the only member that has been identified in B. subtilis at present. From the nt sequence, it seemed that the yycK gene constitutes an operon with five additional ORFs, yycF to yycJ, and we have confirmed it by Northern analysis (data not shown). The products of the first two ORFs of the operon, yycF and yycG, belong to the families of response regulator and sensory kinase genes of two-component signal transduction system, respectively. The operon is expressed in the exponential phase of the B. subtilis growth, and the YycF regulator is essential for the growth (unpublished results). The function of the remaining 16 ORFs is unknown.

Table 1. ORFs identified in the 36-kb sequence between the gntZ and trnY genes.

| ORF   | Endpoints (nt number) | Size (aa) | SD sequence and initiation codon) | Product                        |
|-------|-----------------------|----------|-----------------------------------|-------------------------------|
| ahpC  | 496 1059              | 187      | AGGAGGAATACATTATAG                | Alkyl Hydroperoxide            |
| ahpF  | 1073 2602             | 509      | AAGGAGTCATTCAATAG                 | Reductase Subunit              |
| bgdA  | 4151 2712             | 479      | AAGGAGGAATATACATTATAG             | 6-Phospho-beta-Glucosidase     |
| yydK  | 4739 5449             | 236      | AAATGAGGATCCGTATAG                | DNA Binding Regulator          |
| yydJ  | 6488 5706             | 240      | GGGGCTTGAGATAG                    | Function Unknown               |
| yydI  | 7136 6505             | 209      | AGGAGTATGATATAG                   | ABC Transport Protein          |
| yydH  | 8046 7288             | 252      | AATGATGATGATAG                    | Function Unknown               |
| yydG  | 8986 8027             | 319      | GGGGATTTTCATAG                    | Function Unknown               |
| yydF  | 9193 9044             | 49       | AGGAGGGATATAG                     | Function Unknown               |
| yydE  | 9575 11590            | 671      | AAAAAATGATGATAG                   | Function Unknown               |
| yydD  | 13884 12124           | 586      | AACAAGGAGTTAG                     | Function Unknown               |
| yycG  | 14282 13884           | 132      | AAAAAATGATGATAG                   | Function Unknown               |
| yycB  | 15720 14275           | 481      | AAGAGGATGATGATAG                  | Function Unknown               |
| yycA  | 16461 15982           | 139      | GGGGATGATGATAG                    | Function Unknown               |
| yycC  | 16868 17311           | 137      | GGGGATGATGATAG                    | Formaldehyde Dehydrogenase     |
| yycR  | 18571 17345           | 408      | GGGGATGATGATAG                    | or Dismutase                   |
| yycQ  | 19157 18909           | 82       | AAGGAGGTAGTATAG                   | Function Unknown               |
| yycP  | 20536 19173           | 387      | AAGGAGGTAGTATAG                   | Function Unknown               |
| yycG  | 21084 20947           | 245      | AAGGAGGTAGTATAG                   | Function Unknown               |
| yycN  | 21690 20537           | 145      | AAGGAGGTAGTATAG                   | Function Unknown               |
| rapG  | 21806 22903           | 365      | AAGGAGGTTCATAG                    | Protein Asp Phosphatase        |
| phrG  | 22904 23902           | 38       | AAGGAGGTTCATAG                    | Phosphatase Regulator          |
| rocF  | 24146 23256           | 296      | AAGGAGGTGCAAGATG                  | Arginase                      |
| rocE  | 25623 24220           | 467      | AAGGAGGTGCAAGATG                  | Amino Acid Permease RocE       |
| rocD  | 27051 25846           | 401      | AAGGAGGATGATGATAG                 | Ornithine Aminotransferase     |
| rocC  | 27292 28677           | 461      | GGGGATGATGATAG                    | Arginine Utilization Regulator |
| yycK  | 30314 29112           | 400      | AAGGAGGATGATGATAG                 | Probable Protease             |
| yycJ  | 31189 30833           | 268      | AAGGAGGATGATGATAG                 | Function Unknown               |
| yycI  | 32053 31211           | 280      | AAGGAGGATGATGATAG                 | Function Unknown               |
| yycH  | 33407 32040           | 455      | AAGGAGGATGATGATAG                 | Function Unknown               |
| yycG  | 35232 33937           | 611      | AAGGAGGATGATGATAG                 | Probable Sensor Kinase         |
| yycF  | 35947 35240           | 235      | AAGGAGGATGATGATAG                 | Probable Response Regulator    |

a) The same nucleotides as those of a consensus SD sequence, AAGGAGGTGTA, and the initiation codon are indicated by bold letters.
Table 2. Summary of comparison of 27 ORFs with the protein databases.

| ORF       | Similar Product                          | Database Entry | Score  | Identity |
|-----------|------------------------------------------|----------------|--------|----------|
| YdyK      | Fatty acyl response regulator (240 aa)   | SP:FARR.ECOLI  | 291    | 27.4%    |
|           | Repressor of the trehalose operon (238 aa)| gp:BSTREAPR.3  | 287    | 26.2%    |
|           | Alkylphosphate uptake regulator (241 aa) | SP:PHNF.ECOLI  | 245    | 26.0%    |
| YydI      | Probable transport protein               | pir:SP42925    | 246    | 26.9%    |
| YydA      | Hypothetical 17.3 kd protein (155 aa)    | SP:YBEA.ECOLI  | 290    | 30.8%    |
| YycR      | Hypothetical protein H10033 (155 aa)     | SP:YBEA.HAEIN  | 281    | 30.8%    |
| YycK      | Glutathione-independent formaldehyde     | SP:FADH.PSEPU  | 1746   | 66.3%    |
|           | dehydrogenase (398 aa)                   |                |        |          |
|           | Formylmethionine synthetase - P. putida (398 aa) | pir:JC2516  | 1438   | 54.9%    |
| YycQ      | Hypothetical protein (82 aa)             | YbeF           | 123    | 30.0%    |
| YycN      | Hypothetical protein (157 aa)            | YdgE           | 328    | 37.9%    |
| RapG      | Protein Asp phosphatase RapB (377 aa)    | prf:2102249B   | 572    | 29.4%    |
|           | Protein Asp phosphatase RapA (372 aa)    | gp:BSU55043.6  | 528    | 27.8%    |
|           | Protein Asp phosphatase RapD (354 aa)    | gp:BS283337.4  | 426    | 24.6%    |
| YycK      | Proteas HhoB precursor (355 aa)          | SP:HHOB.ECOLI  | 510    | 33.5%    |
|           | Proteas HhoA precursor (435 aa)          | SP:HHOA.ECOLI  | 567    | 38.1%    |
|           | Proteas HhoA precursor (474 aa)          | SP:HTRA.ECOLI  | 549    | 36.6%    |
| YycE      | Sensor kinase ResE (589 aa)              | SP:RESE.BACSU  | 929    | 29.5%    |
|           | Phosphate regulon sensor kinase (579 aa)  | SP:PHOR.BACSU  | 722    | 27.9%    |
|           | Phosphate regulon sensor kinase (431 aa)  | SP:PHOR.ECOLI  | 525    | 31.1%    |
| YycF      | Phosphate regulon regulator (240 aa)      | SP:PHOF.BACSU  | 812    | 53.1%    |
|           | Transcriptional regulator ResD (240 aa)   | SP:RESD.BACSU  | 643    | 47.4%    |
|           | Phosphate regulon regulator (229 aa)      | SP:PHOF.ECOLI  | 576    | 41.7%    |

a) Abbreviations for the protein databases: SP; Swiss-Prot aa database, gp; translated proteins from NCBI-GenBank nt database, pir; Protein Research Foundation aa database. b) Optimized fasta score. c) % aa identity/overall length. d) H. Yoshikawa, personal communication. e) Y. Kasahara, Unpublished result.

No apparent similarities to reported sequences was found for them, except for three ORFs. Products similar to YydA protein were found in E. coli and H. influenzae. YycQ and YycN showed similarities to the YbeF and YdgE proteins of B. subtilis, respectively.

3. Conservation of a 180-bp Non-coding Sequence Between the rocF and phrG Genes in the B. subtilis genome

No insertion sequences (IS) and transposons indigenous to B. subtilis have yet been described, and the overall structure of the B. subtilis genome is considered to be very stable. However, comparison of our nt sequence with the non-redundant DNA sequence database revealed that a 180-bp non-coding sequence between rocF and phrG gene is conserved in the non-coding regions of at least 4 other loci of the B. subtilis genome (Fig. 2).

Figure 2. Alignment of a 180-bp non-coding sequence between the rocF and phrG genes with the other four non-coding sequences of the B. subtilis genome. Gaps indicated by dashes are introduced to optimize the alignment, and highly conserved regions are boxed. An additional 52-bp sequence is inserted in (B) at the position indicated by #. Duplicated sequences flanking each repetitive sequences are indicated by italic letter. Conserved bases in each duplication are indicated by upper case. (A) 180-bp sequence between rocF and phrG (at 23228 to 23445); (B) 175-bp sequence between ygbL and sfp (at 33050 to 33224 of BSSRFAP, 36° of the genetic map); (C) 235-bp sequence between ydbT and ydcA (Y. Kasahara, unpublished result, 40° of the genetic map); (D) 171-bp sequence between ydbN and ydbO (at 30905 to 31075 of BSGENR, 33° of the genetic map); (E) 179-bp sequence downstream of ydbL (at 3381 to 3201 of BSU11882, 28° of the genetic map). The accession numbers of BSSRFAP, BSGENR and BSU11882 are X70556, X73124 and U11882, respectively. Accession number of a sequence containing the ydbT and ydcA genes is AB001488.
To our knowledge, this is the first report of the existence of repetitive sequence of unknown function in the *B. subtilis* genome. Duplication of sequences flanking the repetitive regions seems to have occurred in parallel with the acquisition of the repeats, although base changes have been accumulated in the duplicated sequences and the receptive sequences themselves. These results suggested that these sequences might be remnants of an ancient IS element that was active during the *B. subtilis* evolution.

**Acknowledgments:** This work was supported by a Grant-in-Aid for Creative Basic Research (Human Genome Program) from the Ministry of Education, Science, Sports, and Culture of Japan.

**References**

1. Ogasawara, N. and Yoshikawa, H. 1996, The systematic sequencing of the *Bacillus subtilis* genome in Japan, *Microbiol.*, 142, 2993–2994.
2. Moszer, I., Kunst, F., and Danchin, A. 1996. The European *Bacillus subtilis* genome sequencing project: Current status and accessibility of the data from a new World Wide Web site, *Microbiol.*, 142, 2987–2991.
3. Harwood, C. R. and Wipat A. 1996. Sequence and functional analysis of the genome of *Bacillus subtilis* strain 168, *FEBS Lett.*, 389, 84–87.
4. Biaudet, V., Samson, F., Anagnostopoulos, C., Ehrlich, S. D., and Bessières, P. 1996. Computerized genetic map of *Bacillus subtilis*, *Microbiol.*, 142, 2669–2729.
5. Ogasawara, N., Nakai, S., and Yoshikawa, H. 1994, Systematic sequencing of the 180 kilo bases region of the *Bacillus subtilis* chromosome containing the replication origin, *DNA Res.*, 1, 1–14.
6. Calogero, S., Gardan, R., Glaser, P., Schweizer, J., Rapoport, G., and Debarbouille, M. 1994. RocR, a novel regulatory protein controlling arginine utilization in *Bacillus subtilis*, belongs to the NtrC/NifA family of transcriptional activators, *J. Bacteriol.*, 176, 1234–1241.
7. Gardan, R., Rapoport, G., and Debarbouille, M. 1995, Expression of the *rocDEF* operon involved in arginine catabolism in *Bacillus subtilis*, *J. Mol. Biol.*, 249, 843–856.
8. Zhang, J. and Aronson, A. 1994, A *Bacillus subtilis* bglA gene encoding phospho-beta-glucosidase is inducible and closely linked to a NADH dehydrogenase-encoding gene, *Gene*, 140, 85–90.
9. Antelamann, H., Engelmann, S., Schmid, R., and Hecker, M. 1996, General and oxidative response in *Bacillus subtilis*: Cloning, expression, and mutation of the alkyl hydroperoxide reductase operon, *J. Bacteriol.*, 178, 6571–6578.
10. Bsat, N., Chen, L., and Helmann, J. D. 1996, Mutation of the *Bacillus subtilis* alkyl hydroperoxide reductase (ahpCF) operon reveals compensatory interactions among hydrogen peroxide stress genes, *J. Bacteriol.*, 178, 6579–6586.
11. Perego, M., Glaser, P., and Hoch, J. A. 1996, Aspartyl-phosphate phosphatases deactivate the response regulator components of the sporulation signal transduction system in *Bacillus subtilis*, *Mol. Microbiol.*, 19, 1151–1157.
12. Itaya, M. 1993, Stability and asymmetric replication of the *Bacillus subtilis* 168 chromosome structure, *J. Bacteriol.*, 175, 741–749.
