Complete nucleotide sequence of a *skin* element excised by DNA rearrangement during sporulation in *Bacillus subtilis*

Ken-ichi Takemaru,[†] Motoki Mizuno, Tsutomu Sato, Michio Takeuchi and Yasuo Kobayashi

Author for correspondence: Yasuo Kobayashi. Tel: +81 423 67 5706. Fax: +81 423 67 5715.
e-mail: ykobayas@cc.tuat.ac.jp

Faculty of Agriculture,
Tokyo University of
Agriculture and Technology,
Fuchu, Tokyo 183, Japan

As part of the *Bacillus subtilis* genome sequencing project, we have determined the complete nucleotide sequence of a *skin* element which is located between *spoIVCB* and *spoIIIC*. The entire sequence of this element is 48 032 bp in length, and contains 57 ORFs with putative ribosome-binding sites. Two of them correspond to previously sequenced and characterized genes, *cwlA* and *spoIVCA*. Furthermore, seven ORF products identified in this element show interesting similarities with known proteins present in data banks, including the ϕ 105 immunity repressor, the ϕ 105 Cro-like protein and the SPP1 terminase. These results indicate the possibility that the *skin* element is a cryptic remnant of an ancestral temperate phage.

Keywords: *Bacillus subtilis*, genome sequencing, DNA rearrangement, *skin* element, ancestral temperate phage

INTRODUCTION

The final stages of sporulation in *Bacillus subtilis* involve the maturation of the spore and its release. The mother-cell compartment of the sporangium produces the coat proteins that surround the spore (Aronson & Fitz-James, 1976) and then it lyses, completing its terminal differentiation program. Expression of the coat proteins is controlled by a late sporulation-specific σ factor, σ^K (Kroos *et al.*, 1989), which is functional only in the mother cell. The coding sequence for σ^K in the chromosome of vegetative cells is separated into two truncated genes, named *spoIVCB* and *spoIIIC*. During sporulation, *spoIVCB*, which encodes the amino-terminal portion of σ^K , and *spoIIIC*, which encodes the carboxy-terminal portion of the factor, are brought together in-frame by site-specific recombination between repeated 5 bp sequences to create an intact coding sequence, *sigK* (Stragier *et al.*, 1989). The recombination event takes place specifically in the mother cell 3 h after the onset of sporulation and is mediated by a site-specific recombinase encoded by the *spoIVCA* gene (Sato *et al.*, 1990; Stragier

et al., 1989). In contrast, the intervening DNA between *spoIVCB* and *spoIIIC*, which includes *spoIVCA*, is excised as a circular DNA molecule by this event (Kunkel *et al.*, 1990). It was also demonstrated that the intervening DNA, called the *skin* element (*sigK* intervening element), is about 42 kb long and contains no genes essential for viability (Kunkel *et al.*, 1990).

As part of the international *B. subtilis* sequencing project, we report here cloning and sequencing of this *skin* element. These studies indicate the possibility that the *skin* element is a cryptic remnant of an ancestral temperate phage.

METHODS

Bacterial strains, phages and plasmids. *B. subtilis* strain JH642 (*trpC2 pheA1*), obtained from the *Bacillus* Genetic Stock Center (Ohio, USA), was used as a source of chromosomal DNA to construct λ libraries and as template for PCR amplification for direct sequencing. λ DASH II and EMBL3 (Stratagene) were used to construct libraries of *B. subtilis* chromosomal DNA after partial digestion with *EcoRI*, *HindIII* or *MboI*. *Escherichia coli* strain P2392 was used for propagation of the λ library. M13 phage vectors and *E. coli* strain JM105 were used to prepare randomly overlapping libraries for shotgun sequencing.

DNA sequencing. All sequences were determined using the *Taq* Dye Primer Cycle Sequencing Kit and the 373A sequencer from Applied Biosystems. We have determined the sequence of both

[†] **Present address:** Department of Genetics, The Graduate University for Advanced studies, 1111 Yata, Mishima, Shizuoka-ken 411, Japan.

The DDBJ accession number for the nucleotide sequence reported in this paper is D32216.

Fig. 1. Organization of the *skin* element. The top line represents ORFs identified in this element. Genes transcribed from left to right are shown above the line and those transcribed from right to left below the line. Shaded boxes, genes sequenced and characterized; hatched boxes, ORFs encoding putative proteins similar to known proteins present in databases; closed triangles, overlaps in ORFs; asterisks, putative transcription terminator sequences. The middle line shows GenBank entries with corresponding accession numbers. The bottom line indicates regions cloned in λ phages, together with those obtained directly from the chromosome by PCR. These sequence data appear in the EMBL/GenBank/DBJ Nucleotide Sequence Data Libraries under the accession number D32216.

DNA strands, except for regions whose sequences were already registered in data banks.

Randomly overlapping libraries for shotgun sequencing were prepared according to the method of Sambrook *et al.* (1989) after sonication of target fragments. Inserts of M13 phages were amplified by PCR from a suspension of phages picked up from plaques using primers having identical sequences to -21M13 and M13RP1 sequencing primers from Applied Biosystems. Amplified products were used as templates for sequencing reactions after polyethylene glycol precipitation. We determined the concentration of purified products by measuring UV absorbance, and used them for the sequencing reaction after dilution to $20 \text{ ng kb}^{-1} \mu\text{l}^{-1}$. Sequence information from random inserts corresponding to four to five times the length of the starting fragment was enough to cover most of the fragment. The sequences of the remaining gap regions were filled using synthesized oligonucleotide primers.

Data handling and computer analysis. DNA sequence assembly was performed with the ARSQ program from Software Development. To avoid errors in the number of bases, we inspected every chart and used sequence outputs only from regions, usually up to about 400 bp, where base peaks were well separated. The compiled sequence was further analysed for the locations of possible ORFs using the GENETYX-MAC 6.2.0 program from Software Development.

RESULTS AND DISCUSSION

Cloning and sequencing of the *skin* element

To completely clone the *skin* element, we used several cloning methods. We first screened genomic libraries constructed in the bacteriophage λ DASH II and EMBL3 using *spoIIIC* and *spoIVCA* genes as probes. Then, extensive chromosomal walking was performed using the end of each λ DNA insert as a probe (Fig. 1). It turned out, however, that a remaining gap was difficult to clone into the λ phage vector. Because the *skin* element was found to be about 42 kb in length (Kunkel *et al.*, 1990), we investigated the size of the remaining gap by Southern hybridization analysis of chromosomal DNA. The size of the gap was estimated to be 6 kb (data not shown), suggesting that the entire length of this element is about 48 kb. The remaining gap missing in the λ phage library was then amplified by PCR directly from the chromosome, followed by direct sequencing of the amplified products.

Sequencing was performed by a strategy based on a combination of PCR and random sequencing, i.e. preparation of a random library, amplification of inserts subcloned in M13 phages by PCR, sequencing of ran-

Table 1. ORF products of the *skin* element showing similarities to extant protein sequences

ORF	Size (aa)	Gene	GenBank accession no.	Description	Homology score	Percentage match (over aa)
1	346	<i>arsB</i>	M86824	Arsenic efflux pump protein – <i>Staphylococcus aureus</i> plasmid pI258	92	18.3 (197)
3	105	<i>arsR</i>	M86824	<i>ars</i> operon regulatory protein – <i>Staphylococcus aureus</i> plasmid pI258	156	38.4 (73)
5	375	<i>gsiAA</i>	X56679	Glucose starvation-inducible gene – <i>B. subtilis</i>	1047	46.9 (375)
7	116		M11920 X02799	Immunity repressor – <i>B. subtilis</i> temperate phage ϕ 105	94	20.2 (109)
8	76		X06090	Cro-like protein – <i>B. subtilis</i> temperate phage ϕ 105	80	34.5 (55)
50	431	<i>gene2</i>	X56064	Terminase – <i>B. subtilis</i> virulent phage SPP1	330	26.9 (398)
109	139	<i>arsC</i>	M86824	Arsenate reductase – <i>Staphylococcus aureus</i> plasmid pI258	480	64.2 (134)

domly selected inserts, and filling of the gaps by region-specific primers (Ogasawara *et al.*, 1994). The entire DNA sequence was determined on both strands to ensure high quality, except for regions whose sequences were already registered in data banks and were identical to our sequence data. However, when discrepancies were found, we determined both strands to confirm our result. The entire sequence of the *skin* element is 48032 bp in length, 5939 bp of which was previously registered in data banks (Fig. 1). The base composition of this element clearly shows a high AT content (61 %) compared with the mean (55 %) (McClelland *et al.*, 1987).

There are no exact criteria for evaluating mistakes during a genome sequencing project. To avoid errors in the number of bases, we inspected every chart of raw data from the sequencer and used sequence outputs only from regions where base peaks were well separated. We have found quite a few discrepancies between our sequences and those in data banks. We have reexamined our sequences and have not found any ambiguity suggesting errors in our sequence. Errors due to mutation during PCR amplification may cause errors in the sequence data. To avoid such errors, we have determined the sequence of at least two independent clones in the same region using random DNA libraries of PCR products.

Identification of putative coding sequences in the *skin* element

Putative ORFs were searched for in all six possible translation frames. We considered the longest possible ORF starting with ATG, GTG or TTG and preceded by a Shine–Dalgarno sequence complementary to the 3' end of the 16S rRNA of *B. subtilis* (3' UCUUCCUCCACUAG) (McLaughlin *et al.*, 1981) to be a valid ORF. Based

on these criteria, 57 putative ORFs were identified. The initiation codon appeared to be ATG in 44 cases, TTG in 10 cases and GTG in three cases. ORF products composed of 100–200 amino acids are most abundant in this element. Overlaps in ORFs were found in 11 regions (Fig. 1). Locations and directions of ORFs are illustrated in Fig. 1 together with their functional assignment as discussed in the following section. There are two large non-coding gaps between ORF71 and ORF55, and ORF94 and ORF5. The nucleotide sequences of these regions were carefully examined for errors, but none were detected. Following the terminology of Daniels *et al.* (1992), these areas can be called 'grey holes'.

Functional assignment of ORFs in the *skin* element

Among the 57 ORFs identified in this study, sequences of one truncated and five complete ORFs were reported previously (Fig. 1). Two of their products, *cwlA* (Kuroda & Sekiguchi, 1990; Foster, 1991) and *spoIVCA* (Sato *et al.*, 1990), were characterized genetically and biochemically. Comparison of the protein sequences of putative ORFs with known protein sequences present in data banks using the Lipman & Pearson programs (1985) revealed interesting similarities to known proteins. Table 1 presents the results from this analysis, as well as specific features of the sequences. Alignments of the amino acid sequences of the predicted products of ORFs 7, 8, 50 and 5 with their counterparts are shown in Fig. 2.

The *skin* element is a remnant of an ancestral temperate phage

Interestingly, the ORF7 product shows 20.2 % identity to the immunity repressor of *B. subtilis* temperate phage ϕ 105 (Dhaese *et al.*, 1985). Moreover, the ORF8 product has

Fig. 2. Alignments of the amino acid sequences of the predicted products of ORFs 7, 8, 50 and 5 with their counterparts. (a) ORF7 product vs ϕ 105 immunity repressor (ϕ 105 ORF1); (b) ORF8 product vs ϕ 105 Cro-like protein (ϕ 105 ORF4); (c) ORF50 product vs SPP1 terminase; (d) ORF5 product vs *B. subtilis* GsiAA protein. The α -helix-turn- α -helix motifs proposed for the ϕ 105 immunity repressor (Dhaese *et al.*, 1985) and Cro-like protein (Kaer *et al.*, 1987) are underlined. Hyphens indicate gaps introduced to improve the alignments. Asterisks and dots represent identical and conserved amino acid residues, respectively.

34.5% identity to the ϕ 105 Cro-like protein (Kaer *et al.*, 1987). These two ORF products appear to contain a α -helix-turn- α -helix configuration in their N terminus (Fig. 2). This three-dimensional structure is the trademark of a DNA-binding protein and has been found in many other repressors (Pabo & Sauer, 1984). Furthermore, SpoIVCA, which is a site-specific recombinase involved in the excision of the *skin* element from the chromosome, has 24.1% identity to the ϕ 105 ORF3 product (Y. Kobayashi and others, unpublished data). ORF3 also encodes a site-specific recombinase involved in the excision and insertion of a prophage from the chromosome (Y. Kobayashi and others, unpublished data). These

Fig. 3. Comparison of the predicted sequences of ORFs 7, 8 and *spoIVCA* with those of the immunity region of *B. subtilis* temperate phage ϕ 105. The percentage of identical amino acids between corresponding genes is shown. Sizes of ORFs are to scale. Numbers below the gene name refer to the number of amino acid residues. The DNA sequence data for the ϕ 105 immunity repressor region are from Kaer *et al.* (1987). P_M is a promoter of the ORF1 operon.

results are represented schematically in Fig. 3. The organization of the corresponding genes as well as the position of recombination sites are very similar. Interestingly, however, the orientation of the *spoIVCA* is opposite to that of the ORF3, suggesting an inversion in this region. ORF3 is under the control of the P_M promoter (Kaer *et al.*, 1987), but *spoIVCA* (ORF3 homologue) is freed from the control of P_M by the inversion and is under sporulation control (Sato *et al.*, 1994) in the *B. subtilis* chromosome. In addition, the protein encoded by ORF50 shows 26.9% identity to the terminase of the *B. subtilis* virulent phage SPP1 (Chai *et al.*, 1992). We also found that the *skin* element ORF5 product, which is a homologue of *B. subtilis* GsiAA, shows 33.2% identity with the ϕ 105 ORF5 product in the N-terminal amino acid sequence (data not shown).

From these observations, we propose that the *skin* element is a cryptic remnant of an ancestral temperate phage. This possibility is further supported by the following facts. Firstly, this element contains the *cwlA* gene, encoding the cell wall lytic enzyme called autolysin. The *B. subtilis* DNA fragments containing *cwlA* were cloned in *E. coli*, producing a lytic zone on an agar medium containing *B. subtilis* cell walls (Kuroda & Sekiguchi, 1990; Foster, 1991). The *cwlA* gene was mapped near *aroD* on the *B. subtilis* chromosome, which is consistent with our results. The insertionally inactivated mutant was able to grow and divide as the wild-type. In addition, Foster (1993) reported that *cwlA* was never expressed under the conditions tested, whereas this gene seemed to be expressed from its own promoter in *E. coli* and so is in itself functional. Importantly, PBSX, a *B. subtilis* defective prophage, encodes two lytic enzymes which cross-react with CwlA antisera (Foster, 1993). Presumably, CwlA is a phage-encoded lytic enzyme, suggesting that the *skin* element is

derived from an ancestral temperate phage. Secondly, Kunkel *et al.* (1990) reported that deletion of the entire *skin* element did not impair growth and sporulation. They also demonstrated that strains genetically engineered to contain the rearranged gene, *sigK*, sporulated normally, and they were rendered independent of the recombinase gene and the two truncated genes. Thirdly, some closely related bacilli have the *sigK* homologue without any DNA rearrangement (Adams *et al.*, 1991). We also found several *B. subtilis* stock strains obtained from the University of Tokyo have the *sigK*, but no *spoIVCA* gene (unpublished data). These results lead us to speculate that the *skin* element is the remnant of an ancestral temperate phage that was inserted into the *sigK* gene during the course of evolution, and the expression of its *spoIVCA* homologue, ORF3, is now under sporulation control.

ACKNOWLEDGEMENTS

We are grateful to Drs Naotake Ogasawara and Jun-ichi Sekiguchi for helpful comments and suggestions. This work was supported by a Grant-in-Aid for Creative Basic Research 'Human Genome Analysis' from the Ministry of Education, Science and Culture, Japan.

REFERENCES

- Adams, L. F., Brown, K. L. & Whiteley, H. R. (1991). Molecular cloning and characterization of two genes encoding sigma factors that direct transcription from a *Bacillus thuringiensis* crystal protein gene promoter. *J Bacteriol* **173**, 3846–3854.
- Aronson, A. I. & Fitz-James, P. (1976). Structure and morphogenesis of the bacterial spore coat. *Bacteriol Rev* **40**, 360–402.
- Chai, S., Bravo, A., Luder, G., Nedlin, A., Trautner, T. A. & Alonso, J. C. (1992). Molecular analysis of the *Bacillus subtilis* bacteriophage SPP1 region encompassing genes 1 to 6. *J Mol Biol* **224**, 87–102.
- Daniels, D. L., Plunkett, G., III, Burland, V. & Blattner, F. R. (1992). Analysis of the *Escherichia coli* genome: DNA sequence of the region from 84.5 to 86.5 minutes. *Science* **257**, 771–778.
- Dhaese, P., Seurinck, J., De Smet, B. & Van Montagu, M. (1985). Nucleotide sequence and mutational analysis of an immunity repressor gene from *Bacillus subtilis* temperate phage ϕ 105. *Nucleic Acids Res* **13**, 5441–5455.
- Foster, S. J. (1991). Cloning, expression, sequence analysis and biochemical characterization of an autolytic amidase of *Bacillus subtilis* 168 *trpC2*. *J Gen Microbiol* **137**, 1987–1998.
- Foster, S. J. (1993). Analysis of *Bacillus subtilis* 168 prophage-associated lytic enzymes; identification and characterization of CWLA-related prophage proteins. *J Gen Microbiol* **139**, 3177–3184.
- Kaer, L. V., Montagu, M. V. & Dhaese, P. (1987). Transcriptional control in the *EcoRI*-F immunity region of *Bacillus subtilis* phage ϕ 105: identification and unusual structure of the operator. *J Mol Biol* **197**, 55–67.
- Kroos, L., Kunkel, B. & Losick, R. (1989). Switch protein alters specificity of *B. subtilis* RNA polymerase containing a compartment-specific sigma factor. *Science* **243**, 526–529.
- Kunkel, B., Losick, R. & Stragier, P. (1990). The *Bacillus subtilis* gene for the developmental transcription factor σ^K is generated by excision of a dispensable DNA element containing a sporulation recombinase gene. *Genes & Dev* **4**, 525–535.
- Kuroda, A. & Sekiguchi, J. (1990). Cloning, sequencing and genetic mapping of a *Bacillus subtilis* cell wall hydrolase gene. *J Gen Microbiol* **136**, 2209–2216.
- Lipman, D. J. & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science* **227**, 1435–1441.
- McClelland, M., Jones, R., Patel, Y. & Nelson, M. (1987). Restriction endonucleases for pulsed-field mapping of bacterial genomes. *Nucleic Acids Res* **15**, 5985–6005.
- McLaughlin, J. R., Murray, C. L. & Rabinowitz, J. C. (1981). Unique features in the ribosome-binding site sequence of the Gram-positive *Staphylococcus aureus* β -lactamase gene. *J Biol Chem* **256**, 11283–11291.
- Ogasawara, N., Nakai, S. & Yoshikawa, H. (1994). Systematic sequencing of the 180 kilobase region of the *Bacillus subtilis* chromosome containing the replication origin. *DNA Res* **1**, 1–14.
- Pabo, C. O. & Sauer, R. T. (1984). Protein–DNA recognition. *Annu Rev Biochem* **53**, 293–321.
- Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: a Laboratory Manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.
- Sato, T., Samori, Y. & Kobayashi, Y. (1990). The *cisA* cistron of *Bacillus subtilis* sporulation gene *spoIVC* encodes a protein homologous to a site-specific recombinase. *J Bacteriol* **172**, 1092–1098.
- Sato, T., Harada, K., Ohta, Y. & Kobayashi, Y. (1994). Expression of the *Bacillus subtilis* *spoIVCA* gene, which encodes a site-specific recombinase, depends on the *spoIIGB* product. *J Bacteriol* **176**, 935–937.
- Stragier, P., Kunkel, B., Kroos, L. & Losick, R. (1989). Chromosomal rearrangement generating a composite gene for a developmental transcription factor. *Science* **243**, 507–512.

Received 13 July 1994; revised 5 September 1994; accepted 21 September 1994.