A 10.3 kbp segment from *nprB* to *argJ* at the 102° region of the *Bacillus subtilis* chromosome

A. Levine, F. Vannier, B. Roche, S. Autret, D. Mavel and S. J. Sror

Author for correspondence: S. J. Sror. Tel: +33 1 69 15 57 14. Fax: +33 1 69 15 78 08.
e-mail: seror@igmors.u-psud.fr

Institut de Gntique et
Microbiologie, URA CNRS
1354, Universit Paris XI,
Btiment 409, 91405 Orsay
Cedex, France

The approximately 10 kbp region encompassing *nprB* and *argJ* at 102 on the *Bacillus subtilis* chromosome was sequenced, revealing 12 ORFs, four known genes (*argJ*, *argC*, *ipi* and *nprB*) and two genes, *yitY* and *yitS*, whose products respectively display significant homology with L-gulonon--lactone oxidase of rat and dihydrofolate reductase of *Staphylococcus aureus*. The data also indicated that *nprB* mapped to a different position than previously published.

Keywords: *Bacillus subtilis*, genome sequencing, 102 region, *argJ*, *nprB*

Within the European programme for sequencing the *Bacillus subtilis* genome, a 60 kbp DNA fragment located between *nprB* (95) and *argJ* (102) was originally assigned to our group. Since, as described in this paper, *nprB* was, in fact, found to be located approximately 8 kb from *argJ*, the assignment was redefined and the region to be sequenced was extended from *argJ* (102) to *addAB* (89). In this first communication, we describe our cloning and sequencing strategy and we report the sequencing and gene features of the 10336 bp region from *nprB* to *argJ*.

For this study, we employed a chromosome walking strategy using the integrative plasmid pDIA5304 (Glaser *et al.*, 1993), into which we inserted a 1 kb fragment from *argJ* obtained from pUL710. This allowed us, after digestion of the chromosomal DNA with an appropriate restriction enzyme (*SacI*), to pick up a fragment of more than 10 kbp adjacent to *argJ*. The recombinant plasmid, pOMG3001, was propagated in *Escherichia coli* TP611, a mutant strain (*pcnB*) which facilitates cloning of *B. subtilis* DNA by reducing the copy number of ColE1 derivatives (Xu *et al.*, 1993). The integrity of the clone was verified by Southern hybridization. For shotgun analysis, pOMG3001 was subjected to a nebulization procedure and 1–1.5 kb fragments were subcloned into pUC18. After identification and selection by Southern hybridization, the fragments corresponding to the insert were sequenced. The sequencing reactions were performed using the Taq Dye Primer Cycle Sequencing kit and the Taq Dye Deoxy Terminator Cycle Sequencing kit, followed by analysis with a 373A DNA sequencer

from Applied Biosystems. The sequences were compiled using the Autoassembler program (Perkin Elmer) and analysed for possible ORFs (GENEMARK). The predicted amino acid sequence of the products of each ORF was used to search for similarity to sequences reported previously in databases using BLAST/FASTA.

From the sequence data, all six possible reading frames were screened with the DNA Strider program. Based on starting codons ATG, GTG or TTG preceded by a putative Shine–Dalgarno (SD) sequence, 12 putative ORFs were detected. The direction of transcription and translation of all except *orf2* (*yitS*) and *orf5* (*yitU*), was, like the majority of genes on the *B. subtilis* chromosome,

Fig. 1. Physical map and gene organization of the *nprB*–*argJ* region of the *B. subtilis* chromosome 102 area. The scale, in kb, starts at the T residue of the *MboII* site in the *nprB* gene and stops at nucleotide 10336, the G residue of the *BspWI* site in the *argJ* region. The numbers 1–12 indicate the ORFs, with the arrows indicating the direction of transcription and translation. The positions of previously identified genes, e.g. *argJ*, are also indicated.

Abbreviation: SD, Shine–Dalgarno.

The EMBL accession number for the nucleotide sequence reported in this paper is Z79580.

Table 1. Putative orfs in the *nprB*–*argJ* region

ORF	Endpoints (nt)	Size of product (aa)	SD consensus sequence (upper case) and initiation codon (bold)
<i>orf1</i>	39 > 1652	538	AGGAGGA tgtgtt tg
<i>orf2</i>	1702 < 2553	283	ggaaAGAAGGgaagca atg
<i>orf3</i>	2691 > 3533	280	gagcGGAGGAacatg ccatg
<i>orf4</i>	3648 > 4007	119	aAGGAGGtatcaggcgga atg
<i>orf5</i>	4492 < 5304	270	AAGGAGcttttc atg
<i>orf6</i>	5425 > 6192	255	aGGAGGggcacatt gtg
<i>orf7</i>	6256 > 6564	102	gaaaaAAGGAGtgagggaa ctg
<i>orf8</i>	6619 > 6831	70	AGctGAAaggag tg
<i>orf9</i>	6889 > 8289	466	GAAAAAGAAAttgcttcattcg ctt tg
<i>orf10</i>	8334 > 8828	164	cAcAAAGcAGc gtg
<i>orf11</i>	9035 > 10075	346	AAGAGGGtGAActt gtt tg
<i>orf12</i>	10095 > 10334	80 (partial)	GGAGAGAAtcgaac catg

Table 2. Similarity of predicted *orf* products to other proteins

Product	Gene name	Similar protein in database	Identity observed or BLAST probabilities
Orf1	<i>nprB</i>	Neutral protease of <i>B. subtilis</i>	100 % for the entire length
Orf2	<i>yitS</i>	Sadired_2; <i>S. aureus</i>	$3.5e^{-32}$
Orf3	<i>yitT</i>	Orf1 encoded by <i>ipi</i> of <i>B. subtilis</i>	100 % for the entire length
Orf4	<i>ipi</i>	BsuPI intracellular proteinase inhibitor encoded by <i>ipi</i> (Orf2) of <i>B. subtilis</i>	100 % for the entire length
Orf5	<i>yitU</i>	YidA; <i>E. coli</i>	$8.6e^{-31}$
Orf6	<i>yitV</i>	Yjfp; <i>E. coli</i>	$2.9e^{-27}$
Orf7	<i>yitW</i>	GPU00013_4; <i>M. leprae</i>	$1.7e^{-24}$
Orf8	<i>yitX</i>		
Orf9	<i>yitY</i>	L-Gulono- γ -lactone oxidase; rat	$9.1e^{-15}$
Orf10	<i>yitZ</i>		
Orf11	<i>argC</i>	ArgC of <i>B. subtilis</i>	100 % for the entire length
Orf12	<i>argJ</i>	ArgJ of <i>B. subtilis</i> (partial sequence)	100 % for the entire length

coincident with the direction of DNA replication (Fig. 1). The endpoints of the ORFs in the sequence, the size of the proteins encoded, putative SD sequences and

initiation codons are shown in Table 1. Five of the 12 genes were predicted to use ATG as the initiation codon, whilst four and three use GTG and TTG, respectively.

Fig. 2. Hydropathy analysis of potential membrane proteins YitT (a) and YitZ (b).

In the case of *orf10*, the predicted SD sequence only precedes the anticipated starting codon by a single base. This is rare in *B. subtilis* and confirmation of expression of this gene would be required *in vivo*.

Comparison of the amino acid sequences of the putative products of each ORF with the non-redundant protein sequence database (NCBI) is indicated in Table 2. The amino acids of Orf9 showed similarity (52% in the C-terminal region) with rat L-gulonolactone oxidase (Nishikimi *et al.*, 1992). The *orf7* product showed 50% amino acid sequence identity with protein GPU00013 from *Mycobacterium leprae* (K. Robinson, Gen Pept database). Orfs 5 and 6 showed high similarity (65% and 59%, respectively) to the C-terminal regions of YidA, a hypothetical 29.7 kDa protein of *E. coli* (Adachi *et al.*, 1984), and the C-terminal region of YifP, a hypothetical 27.6 kDa protein of *E. coli* (Burland *et al.*, 1995). Orf2 showed 51% amino acid sequence identity with the dihydrofolate reductase of *Staphylococcus aureus* (Dale *et al.*, 1993).

Five of the ORFs corresponded to genes previously sequenced: *argJ*, *argC* (Mountain *et al.*, 1984; O'Reilly & Devine, 1994), *ipi*, *yitT* (Shiga *et al.*, 1993) and *nprB* (Tran *et al.*, 1991). The reported sequence for these genes was found to be identical to our sequence, except for one position in *argC* (in our sequence, G at position 9737 instead of T), two positions downstream and upstream of *ipi* (in our sequence, an additional C at position 4332 and a G at position 3366 instead of C) and within the preceding gene, *yitT* (in our sequence GCC at position 2483 instead of CGG). In addition, our results revealed a 246 bp sequence (8674–8920) upstream of *argC* which was previously mapped downstream of the *dap* operon (148°) in strain GS11 (Chen *et al.*, 1993). In fact, Southern blotting of chromosomal DNA from strain 168 restricted by *EcoRI* and *Clal*, using the 246 bp fragment as a probe, confirmed a unique band compatible with our sequencing data. This, therefore, strongly suggests that a DNA rearrangement occurred downstream of the *dap* operon, giving rise to the difference between strain 168 and strain GS11. Finally, it is important to note that *nprB* was, in fact, previously mapped at 95° (Tran *et al.*, 1991), whilst the location of *ipi* on the chromosome was not reported earlier.

Amongst the predicted ORFs, we also searched for potential membrane or secreted proteins. For this purpose, the amino acid sequences were scanned with the DNA Strider hydrophathy program (Kyte–Doolittle algorithm; Kyte & Doolittle, 1982). These results indicated two small integral membrane proteins, as shown in Fig. 2: YitZ, with three to four predicted transmembrane domains (TMDs), and YitT, with a minimum of five predicted TMDs. In addition, NprB (538 amino acids) was predicted to be a secreted protein. In fact, this protein was previously identified as an extracellular neutral protease (Tran *et al.*, 1991).

ACKNOWLEDGEMENTS

We are grateful to Philippe Glaser for providing pDIA5304 and *E. coli* TP611, and to Margaret C. M. Smith for providing pUL710. We also thank Laurent Mallet for technical advice in this study. We are pleased to acknowledge support from CNRS, University Paris XI, the EU Biotechnology Programme (no. BIO2-CT94-2011) and GREG (no. 118/94).

REFERENCES

- Adachi, T., Mizuuchi, K., Menzel, R. & Gellert, M. (1984). DNA sequence and transcription of the region upstream of the *E. coli* *gyrB* gene. *Nucleic Acids Res* 12, 6389–6395.
- Burland, V., Plunkett, G., Sofia, H. J., Daniels, D. L. & Blattner, F. R. (1995). Analysis of the *Escherichia coli* genome VI: DNA sequence of the region from 92.8 through 100 minutes. *Nucleic Acids Res* 23, 2105–2119.
- Chen, N. Y., Jiang, S. Q., Klein, D. A. & Paulus, H. (1993). Organization and nucleotide sequence of the *Bacillus subtilis* diaminopimelate operon, a cluster of genes encoding the first three enzymes of diaminopimelate synthesis and dipicolinate synthase. *J Biol Chem* 268, 9448–9465.
- Dale, G. E., Then, R. L. & Stueber, D. (1993). Characterization of the gene for chromosomal trimethoprim-sensitive dihydrofolate reductase of *Staphylococcus aureus* ATCC 25923. *Antimicrob Agents Chemother* 37, 1400–1405.
- Glaser, P., Kunst, F., Arnaud, M. & 14 other authors. (1993). *Bacillus subtilis* genome project: cloning and sequencing of the 97 kb region from 325° to 333°. *Mol Microbiol* 10, 371–384.
- Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 157, 105–132.
- Mountain, A., Mann, N. H., Munton, R. N. & Baumberg, S. (1984). Cloning of a *Bacillus subtilis* restriction fragment complementing auxotrophic mutants of eight *Escherichia coli* genes of arginine biosynthesis. *Mol Gen Genet* 197, 82–89.
- Nishikimi, M., Kaiwai, T. & Yagi, K. (1992). Guinea pigs possess a highly mutated gene for L-gulonolactone oxidase, the key enzyme for L-ascorbic acid biosynthesis missing in this species. *J Biol Chem* 267, 21967–21972.
- O'Reilly, M. & Devine, K. M. (1994). Sequence and analysis of the citrulline biosynthetic operon *argC–F* from *Bacillus subtilis*. *Microbiology* 140, 1023–1025.
- Shiga, Y., Yamagata, H. & Udaka, S. (1993). Characterization of the gene encoding an intracellular proteinase inhibitor of *Bacillus subtilis* and its role for regulation of major intracellular proteinase. *J Bacteriol* 175, 7130–7137.
- Tran, L., Wu, X.-C. & Wong, S.-L. (1991). Cloning and expression of a novel protease gene encoding an extracellular neutral protease from *Bacillus subtilis*. *J Bacteriol* 173, 6364–6372.
- Xu, F., Lin-Chao, S. & Cohen, S. N. (1993). The *Escherichia coli* *pcnB* gene promotes adenylation of antisense RNAI of ColE1-type plasmids *in vivo* and degradation of RNAI decay intermediates. *Proc Natl Acad Sci USA* 90, 6756–6780.

Received 18 June 1996; revised 2 September 1996; accepted 20 September 1996.