Sequence and analysis of a 31 kb segment of the *Bacillus subtilis* chromosome in the area of the *rrnH* and *rrnG* operons

Hongtu Liu, Koki Haga, Kenji Yasumoto, Yukio Ohashi,
Hirofumi Yoshikawa and Hideo Takahashi

Author for correspondence: Hideo Takahashi. Tel: +81 3 3812 2111 ext. 7825. Fax: +81 3 3813 0539.

Institute of Molecular and
Cellular Biosciences, The
University of Tokyo,
Bunkyo-ku, Tokyo 113,
Japan

A 31 141 bp continuous nucleotide sequence in the region from *trnI* to pNEXT52 in the *Bacillus subtilis* 168 genome was determined. In the region, there were 22 ORFs, two complete rRNA operons, and five tRNA genes. It was deduced that the function of one of the ORFs was similar to that of a sigma factor belonging to the ECF (extra-cytoplasmic functions) subfamily. The gene cluster *feuA*, *B*, *C* reported previously for other strains of *B. subtilis* was also found in strain 168 and located in this region.

Keywords: *Bacillus subtilis*, genome sequencing, ECF sigma factor, ribosomal RNA operons

The *NotI* and *SfiI* restriction maps of the entire circular *Bacillus subtilis* 168 chromosome have been constructed and the whole genome size is considered to be 4188 kb (Itaya & Tanaka, 1991). In the framework of the international *B. subtilis* genome sequencing project, a defined region of the chromosome was assigned to each laboratory of the consortium. The DNA region located between *spo0H* (10°) and pNEXT9 (24°) was assigned to us. We previously reported the sequence of an approximately 50 kb region between *spo0H* and *rrnH* (15°) (Yasumoto *et al.*, 1996). In this study, we determined the sequence of 31 141 bp in the region between 15° and 18°, between the *rrnH* and the pNEXT52 *NotI* linking clone. Since there were few genes known in this region, we used *NotI* linking clones as initial sequencing material. Subsequently, the regions between known sequences were amplified by PCR.

All sequences were determined by shotgun sequencing. Cycle sequencing reactions were performed on both strands using the Catalyst 800 Molecular Biology LabStation and ABI Dye Primer Cycle Sequencing Ready Reaction kits (Applied Biosystems), and thereafter they were analysed with a 373A DNA sequencer (Applied Biosystems). The compiled sequences were analysed using a GenomeNet FASTA server in the Human Genome Center, Institute of Medical Science, University of Tokyo.

Abbreviation: ECF, extra-cytoplasmic function.

The GenBank accession numbers for the updated sequences reported in this paper are D84213 and D84214.

According to the physical map of *B. subtilis* 168 (Itaya & Tanaka, 1991), three *NotI* sites involved in linking clones pNEXT5, pNEXT3 and pNEXT52 were assumed to be located in our region. Therefore, we first determined the sequence of the inserts of these linking clones (Fig. 1). Secondly, we obtained cosmid clones which were constructed from *B. subtilis* 168 (S. Hiraga, personal communication). By hybridization analysis, these clones contained fragments located in the 22N *NotI* fragment of the physical map (Itaya & Tanaka, 1991; M. Itaya, personal communication). We, therefore, used these clones as the sequence source of our region. Partial sequencing of these clones disclosed that the known sequence of the *feuA*, *B*, *C* gene cluster (Quirk *et al.*, 1994; accession no. L19954) was located in our region upstream of pNEXT3. Hitherto, there had been three 'islands' in the region between *rrnG* and pNEXT52 (Fig. 1). We amplified additional segments by PCR (LA PCR kit; Takara Shuzo) to try to bridge each gap using synthesized primers based on known sequences. Since sequence L19954 from the database is derived from a different *B. subtilis* strain than 168, we finally confirmed the sequence of this region by determining that of strain 168.

A shotgun sequencing protocol cannot be directly applied to a fragment containing two rRNA operons because of their highly conserved sequences. Our initial attempt to amplify the *rrnH* and *rrnG* regions was carried out with primers based on *trnI* and pNEXT5 sequences and several kinds of fragments were produced, probably due to the high identity among rRNA gene sequences. The longest product of a 9 kb fragment

Fig. 1. Protein and RNA coding regions in the 31 kb sequence. Coding regions that are transcribed from left to right are shown above the scale, and those transcribed in the opposite direction are shown below it. Boxes, which indicate the protein coding regions (ORFs) or RNA coding regions, are shaded according to their function: A, functionally identified proteins or genes; B, strong identity with functionally identified proteins or genes; C, strong identity with functionally unknown proteins or genes; D, weak identity with functionally identified proteins or genes; E, no homology in the database; F, RNAs. Thick black bars indicate the initial four islands and the dashed line represents the region contained in the cosmid clone. Arrowheads indicate *NotI* sites.

appeared to contain two rRNA operons and each operon was amplified individually from this 9 kb fragment and sequenced. In the course of the study, in which we confirmed that two sets of rRNA operons were located in the 9 kb fragment, we performed PCR with a series of primers, sl-1 (5' CACGTGTTACTCACCCGTCGCCG 3'), sl-2 (5' GAATGATGTCACATCTGTTATCTAG 3'), sr-1 (5' CACGTGTTACTCACCCGTCGCCG 3'), and sr-2 (5' GCGCCGCGGGTC-CATCTGTAAGTGG 3'). Two sr primers were designed from the 5S rRNA gene sequence in a downstream direction and another two from the 16S rRNA gene sequence in an upstream direction (Fig. 2a). The results shown in Fig. 2(b) (lanes 5–8) clearly indicate that the 9 kb fragment contains two sets of rRNA operons and only one spacer. On the other hand, when chromosomal DNA was used as template, we observed at least four

bands instead of the three which corresponded to the spacer region between *rrnH* and *rrnG*, *rrnI* and *rrnH*, and *rrnJ* and *rrnW* [from the bottom to the second top bands, respectively in Fig. 2(b) lanes 1–4]. It was assumed that the band that migrated most slowly corresponded to another spacer region between contiguous rRNA operons.

In the genome of *B. subtilis*, ten rRNA operons were mapped (Vold, 1985; Okamoto, *et al.*, 1993). According to the results of Okamoto *et al.* (1993) using a YAC library as template, the existence of an 11th ribosomal operon, which had been proposed as *rrnF* by Bott *et al.* (1984) and Anagnostopoulos *et al.* (1993), is unlikely. Our results, however, suggest that there would be at least one unmapped rRNA operon or *rrn*-like region in the *B. subtilis* genome. Although the discrepancy may be

Fig. 2. Amplification of spacer region between 5S and 16S rRNA genes of *rrn* operons of *B. subtilis* by PCR. (a) Sets of oligonucleotide primers used (see text). (b) PCR performed with the following combinations of primers: lanes 1 and 5, sl-1 and sr-1; lanes 2 and 6; sl-1 and sr-2; lanes 3 and 7, sl-2 and sr-1; lanes 4 and 8, sl-2 and sr-2. Genomic DNA (lanes 1–4) or a 9 kb fragment (see text, lanes 5–8) was used as template.

Table 1. ORFs in the 31 kb sequence

ORF	Endpoints*	Length (nt/aa)	SD sequence and initiation codon
<i>ybaP</i>	10615 < 10752	138/46	GTTTTTCTTAGGACTTATTTatg
<i>ybaQ</i>	10902 > 11015	114/38	GAAACCAGCTCGGGTTGCCAAatg
<i>ybaR</i>	11196 > 12629	1434/478	ACTAAAAAGGAGTCGTTATATttg
<i>ybaS</i>	12847 > 13404	558/186	TCAGTTTTTAAATGGGTATGAatg
<i>ybaT</i>	13409 > 13696	288/96	CTTATTTGCAAAATCCATTGATttg
<i>ybbA</i>	13712 < 14278	567/189	CAGCAAAGGCTCCCTGAACTCgtg
<i>feuC</i>	14286 < 15467	1182/394	AACGAAAAGGAGGGGAGCAAAatg
<i>feuB</i>	15463 < 16464	1002/334	TAATTCAGAGTAGGTTTTTTGTatg
<i>feuA</i>	16486 < 17436	951/317	ATTGAACAGGAGGCTCTATAGatg
<i>ybbB</i>	17530 < 19116	1587/529	TTCTGAAAAGGTGGTCTACTCgtg
<i>ybbC</i>	19310 < 20551	1242/414	CGCTTTAATAAGGAGGGACACatg
<i>ybbD</i>	20568 < 22493	1926/642	AGACGATAGGAGGAAGTTTCTatg
<i>ybzA</i>	22524 < 22754	231/77	AGGCGGAACGGAAGTAAAAGCatg
<i>ybbE</i>	22762 < 23685	924/308	AAGCTAAGGAAGGTAGATCAGatg
<i>ybbF</i>	23935 < 25272	1338/446	TCTTGAAGGGGAGGGGTTTTAatg
<i>ybbH</i>	25301 < 26215	915/305	AACAAGCACAGACATTACTCGatg
<i>ybbI</i>	26215 < 27132	918/306	AGTTGAAATATCATAACAGAAatg
<i>ybbJ</i>	27255 < 27671	417/139	ATATTGTAAGGAGGCCACCCCatg
<i>ybbK</i>	27687 < 28139	453/151	AGCAAGAGAGGAGAAACCAATatg
<i>ybbL</i>	28955 > 29515	561/187	GAAGGCAGAGGTTAGATAAAATatg
<i>ybbM</i>	29532 > 29747	216/72	ATCTTTAAGTGGGGTGATGAAatg
<i>ybbP</i>	30322 > 31140	819/273	GAAATCTTGGAGGACGAGGAAatg

* Endpoints correspond to the relative location in our region. ' > ' indicates the coding region which is transcribed from *oriC* to *terC*, while ' < ' indicates the opposite direction.

Table 2. Similarities of ORF products to proteins in the databases

Product of ORF	Best score	Characteristics of protein	FASTA score	Identity (%)	Overlap (aa)
YbaR (478 aa)	gpu:D90906_30	Hypothetical protein in <i>Synechocystis</i> sp.	1505	56.4	477
	gp:MSU50335_2	Hypothetical protein (ORF2) in <i>Mycobacterium smegmatis</i>	1055	46.9	467
YbaS (186 aa)	sp:MRED_HAEIN	Rod-shape-determining protein	108	22.8	123
	prf:2121326J	NADH dehydrogenase subunit IV	101	18.8	186
	pir:S26019	Ubiquinol-cytochrome- <i>c</i> reductase	83	18.1	138
	sp:POLR_EPMV	RNA replicase polyprotein	77	19.7	142
FeuC (394 aa)	gp:BSFHUDBG_3	Ferrichrome transport system	761	41.4	302
	sp:FECDC_ECOLI	Iron(III) dicitrate transport system permease	502	31.7	315
	pir:S54822	CbrC protein in <i>Erwinia chrysanthemi</i>	481	33.3	312
YbbB (529 aa)	sp:YIJO_ECOLI	Hypothetical transcriptional regulator	188	27.9	122
	sp:RAFR_PEDPE	Raffinose operon transcriptional regulator	174	22.3	247
	sp:XYLR_ECOLI	Xylose operon regulatory protein	163	29.7	101
	sp:ADAA_BACSU	Methylphosphotriester-DNA alkyltransferase	158	28.7	108
YbbD (642 aa)	gp:ALTCHT60_1	Chitinase (<i>N</i> -acetyl- β -glucosaminidase) 60 precursor	856	35.0	597
	prf:2018215A	Acetylglucosaminidase	361	43.0	193
YbbE (308 aa)	sp:Y959_HAEIN	Homologue of family 3 of glycosyl hydrolases	266	30.9	282
	sp:DAC_STRSQ	D-Alanyl-D-alanine carboxypeptidase precursor	158	25.2	151
	sp:PBP4_NOCLA	Penicillin-binding protein subunit IV	155	25.9	162
	prf:1404319A	Acetyl-CoA carboxylase	153	21.0	162
YbbF (446 aa)	sp:FMDD_BACNO	Possible fimbrial assembly protein	103	23.6	144
	sp:PTSB_KLEPN	PTS, sucrose-specific IIBC component	636	36.3	397
	sp:PTSB_SALTY	PTS, sucrose-specific IIBC component	631	36.4	398
YbbH (305 aa)	prf:1717372A	Sucrose transporter EIIScr (glutamine synthetase)	599	35.9	398
	sp:Y143_HAEIN	Hypothetical protein which belongs to the YebK/YfhH/YjcY family	460	29.1	289
	sp:YNGB_CLOPE	Hypothetical protein which belongs to the YebK/YfhH/YjcY family	429	29.0	276
YbbI (306 aa)	sp:Y754_HAEIN	Hypothetical protein in <i>Haemophilus influenzae</i>	875	48.4	285
	gpu:D90907_87	Hypothetical protein in <i>Synechocystis</i> sp.	716	42.2	275
YbbL (SigW) (187 aa)	prf:2209365A	AlgU in <i>Azotobacter vinelandii</i>	387	36.8	185
	sp:RPSH_PSEAE	RpoE in <i>Pseudomonas aeruginosa</i>	377	37.1	186
	gp:PHRRPOE_1	RpoE in <i>Photobacterium</i> sp.	369	36.8	182
	sp:RPOE_ECOLI	RpoE in <i>Escherichia coli</i>	350	36.6	183
YbbP (273 aa)	gp:SYCSLLH_52	Hypothetical protein in <i>Synechocystis</i> sp.	422	38.2	254
	gpu:MPAE00057_10	Hypothetical protein in <i>Mycoplasma pneumoniae</i>	103	39.4	109

ascribed to the difference in the template for PCR, the whole genome sequence should be analysed before conclusions can be drawn.

We searched for ORFs in all six possible translation frames. We considered that ORFs longer than 30 aa starting with ATG, GTG, or TTG and preceded by a Shine-Dalgarno (SD) sequence were most likely to be coding sequences. On this basis, a total of 22 putative ORFs were identified (Table 1, Fig. 1). These genes to which functions could not be assigned were designated according to the guidelines of the SWISS-PROT protein sequence database as *ybaP* through *ybbP*. In addition, a cluster of tRNA genes was identified by homology in comparison with other identified tRNA genes (Fig. 1).

ORFs in the region which shared homology with

proteins in the database are shown in Table 2. Here, we describe certain ORFs which may not be sufficiently described in the Table.

1. FeuC was partially deduced by the nucleotide sequence as described by Quirk *et al.* (1994). A complete frame of 404 aa residues was identified in the region. Better identities were found between it and its homology with the third frame of BSFHUDBG (GenBank, 41.4% in 302 aa overlap), FEUCD_ECOLI (31.7% in 315 aa overlap) and S54822 (PIR, 33.3% in 312 aa overlap) which are iron(III)-uptake proteins. Comparison of the structure of the *B. subtilis* operons of *fhuD*, *B*, *G* and *C* (accession no. X93092) and *feuA*, *B* and *C* revealed that the products of FeuA, FeuB and FeuC corresponded to FhuD, FhuB and FhuG, respectively, based on their

identities, although the directions of *fhuD* and *feuA* were divergent. The genes *fhuD*, *B*, *G* and *C* were located at 286° on the map (Schneider & Hantke, 1993).

2. YbbL (SigW) was found to be significantly similar to the sigma factor, with identities of about 30% in the whole region of YbbL, in particular, to the ECF subfamily of sigma factors including the σ^{24} factor of *E. coli* and the RpoE of *Pseudomonas aeruginosa*. We therefore designated the ORF as SigW. To date, several ORFs including SigX and SigY similar to the ECF sigma factors have been detected in the genome of *B. subtilis* (Huang *et al.*, 1997; Yoshida *et al.*, 1996), though their detailed function has not yet been studied.

ACKNOWLEDGEMENTS

The authors thank Dr M. Itaya (Mitsubishi Kagaku Institute of Life Science) for providing the *NotI* linking clones and Professor S. Hiraga (Kumamoto University) for his generous gift of cosmid clones. This research was supported by a Grant-in-Aid for Creative Research (Human Genome Project) from the Ministry of Education, Science and Culture of Japan.

REFERENCES

- Anagnostopoulos, C., Piggot, P. J. & Hoch, J. A. (1993). The genetic map of *Bacillus subtilis*. In *Bacillus subtilis and Other Gram-positive Bacteria*, pp. 425–461. Edited by A. L. Sonenshein, J. A. Hoch & R. Losick. Washington, DC: American Society for Microbiology.
- Bott, K. F., Stewart, G. C. & Anderson, A. G. (1984). Genetic mapping of cloned ribosomal RNA genes. In *Genetics and Biotechnology of Bacilli*, pp. 19–34. Edited by A. T. Ganesan & J. A. Hoch. Orlando, FL: Academic Press.
- Huang, X., Decatur, A., Sorokin, A. & Helmann, J. D. (1997). The *Bacillus subtilis* σ^X protein is an extracytoplasmic function σ factor contributing to survival at high temperature. *J Bacteriol* 179, 2915–2921.
- Itaya, M. & Tanaka, T. (1991). Complete physical map of the *Bacillus subtilis* 168 chromosome constructed by a gene-directed mutagenesis method. *J Mol Biol* 220, 631–648.
- Okamoto, K., Serror, P., Azevedo, V. & Vold, B. (1993). Physical mapping of stable RNA genes in *Bacillus subtilis* using polymerase chain reaction amplification from yeast artificial chromosome library. *J Bacteriol* 175, 4290–4297.
- Quirk, P. G., Guffanti, A. A., Clejan, S., Cheng, J. & Krulwich, T. A. (1994). Isolation of Tn917 insertional mutants of *Bacillus subtilis* that are resistant to the protonophore carbonyl cyanide *m*-chlorophenylhydrazine. *Biochim Biophys Acta* 1186, 27–34.
- Schneider, R. & Hantke, K. (1993). Iron-hydroxamate uptake systems in *Bacillus subtilis*: identification of lipoproteins as part of a binding protein-dependent transport system. *Mol Microbiol* 8, 111–121.
- Vold, B. S. (1985). Structure and organization of genes for transfer ribonucleic acid in *Bacillus subtilis*. *Microbiol Rev* 49, 71–80.
- Yasumoto, K., Liu, H., Jeong, S. M., Ohashi, Y., Kakinuma, S., Tanaka, K., Kawamura, F., Yoshikawa, H. & Takahashi, H. (1996). Sequence analysis of a 50 kb region between *spo0H* and *rrnH* on the *Bacillus subtilis* chromosome. *Microbiology* 142, 3039–3046.
- Yoshida, K., Shindo, K., Sano, H., Seki, S., Fujimura, M., Yanai, N., Miwa, Y. & Fujita, Y. (1996). Sequencing of a 65 kb region of the *Bacillus subtilis* genome containing the *lic* and *cel* loci, and creation of a 177 kb contig covering the *gnt*–*sacXY* region. *Microbiology* 142, 3113–3123.

Received 17 March 1997; accepted 14 April 1997.