Organization of the \textit{Bacillus subtilis} 168 chromosome between \textit{kdg} and the attachment site of the \textit{SP\beta} prophage: use of Long Accurate PCR and yeast artificial chromosomes for sequencing

Véronique Capuano, Nathalie Galleron, Petar Pujic, Alexei Sorokin and S. Dusko Ehrlich

Author for correspondence: Alexei Sorokin. Tel: +33 1 34 65 25 33. Fax: +33 1 34 65 25 21. 
e-mail: sorokine@biotec.jouy.inra.fr

\textbf{INTRODUCTION}

Two types of approach are presently being followed in the sequencing of small genomes. One is based on random collections of cloned genome segments. The segments are sequenced to a high redundancy, typically greater than sixfold above the genome size, the resulting sequence is assembled into contigs and the gaps between the contigs filled by a variety of methods. This approach was used to sequence the complete genomes of \textit{Haemophilus influenzae} and \textit{Mycoplasma genitalium} (Fleischmann \textit{et al.}, 1995; Fraser \textit{et al.}, 1995). Another approach is based on the use of ordered collections of genome segments. The segments are sequenced independently to completion and the final sequence is then assembled. This approach was used to

\textbf{Keywords:} \textit{Bacillus subtilis}, genome sequencing, yeast artificial chromosome, Long Accurate PCR
Within the *Bacillus subtilis* genome sequencing project we are following an approach based on the ordered collection of segments in yeast artificial chromosomes (YACs; Azavedo et al., 1993). Sequencing of one YAC, 15-6B, carrying an insert of about 105 kb, involved: (i) cloning of YAC DNA in phage M13-based vectors and random sequencing of the clones; (ii) contig ordering, by combinatorial PCR and (iii) gap closure by direct sequencing approach was used to sequence this YAC. Its DNA was originally provided by C. Anagnostopoulos (Jouy-en-Josas, France) and covers a part of the chromosome region assigned to our chromosome to *B. subtilis* chromosome and about 50% had homology with known genes.

**PCR procedures.** Oligonucleotides were synthesized on DNA synthesizer ‘Oligo 1000’ (Beckman). For amplification of fragments shorter than 2 kb, standard PCR was carried out, using either M13 ssDNA clones or *B. subtilis* chromosomal DNA as template. LA PCR was carried out essentially as described by Cheng et al. (1994). Oligonucleotides of 20–23 bases, 12 of which were G or C residues, were used under the following reaction conditions: 20 mM Tricine, pH 8.7; 85 mM KOAc; 1 mM Mg(OAc)₂; 8% (v/v) glycerol; 2% (v/v) DMSO; 0.2 mM each dNTP (Perkin Elmer); 0.2 μM each primer; 2 U Tth polymerase (Perkin Elmer); 0.05 U Vent polymerase (Biobal) and 0.1 μg *B. subtilis* chromosomal DNA prepared as described by Sorokin et al. (1996b). The final reaction volume was 50 μl. Cycling conditions were: 94°C, 5 min; 12 cycles of 10 s at 94°C and 12 min at 68°C and 24 cycles increasing the extension time by 15 s each cycle.

**PCR fragment subcloning.** LA PCR reaction products, corresponding to the xpt, degR, degR-thyB or thyB-attSPβ interval were partially cut by HpaII, TagII or SaeIII. The cleaved DNA was electrophoresed on agarose gels and the fragments of 500–1500 bp were purified and inserted into the Acl or BamHI site of M13mp18 or pSGMU2 (Errington, 1986). For direct sequencing and sometimes for cloning, the LA PCR products, obtained by mixing 10 independent reactions, were purified using a Wizard PCR Preps Kit (Promega) or, after electrophoresis on agarose gels, using GeneClean (Bio 101).

**Sequencing.** ssDNA of phage clones was prepared by the perchlorate method (Zimmermann et al., 1989) as previously described (Sorokin et al., 1993). Plasmid DNA was prepared as described by Sorokin et al. (1995). PCR with biotinylated primers (Dynal) carried out as specified by the supplier, was used to prepare DNA for reverse sequencing of the M13 phage inserts. Direct and reverse sequencing was carried out with Applied Biosystems PRISM and Reverse Sequencing kits, respectively, on the thermal cycler (Perkin Elmer) or Catalyst station (Applied Biosystems).

**Computing.** R. Staden’s *xhAP* program (Dear & Staden, 1991) was used for gel assembly and generation of a consensus sequence. Sequence homologies were searched using *FASTA* (Pearson & Lipman, 1988) or *BLAST* (Altschul et al., 1990) algorithms, realized on the NCBI e-mail server.

**Nomenclature.** The nomenclature for coding sequences (CDSs) is in accordance with the *B. subtilis* genome project as described by Sorokin et al. (1996a).

**RESULTS AND DISCUSSION**

**Sequencing of the region between *kdg* and *attSPβ**

YAC 10-9 contains an insert which was estimated to be ~100 kb by PFGE (Azavedo et al., 1993) and thus represents approximately 2.5% of the total *B. subtilis* genome. The sequence several *Saccharomyces cerevisiae* chromosomes (Oliver et al., 1992; Dujon et al., 1994).

**METHODS**

**Strains and growth conditions.** *Bacillus subtilis* strain 168, originally provided by C. Anagnostopoulos (Jouy-en-Josas, France) was used as the chromosomal DNA donor. *Escherichia coli* strain JEC128F’ (araD 139 A(ara-leu)7G96galE 15galK16 A(lac)Y7#1 pg YAC DNA per 500 ml yeast cell culture to be obtained.

**Genomic banks and screening.** *B. subtilis* chromosomal DNA was partially digested by pancreatic DNase I or nebulization. Randomized DNA was electrophoresed on agarose gels and fragments from 600 to 1500 bp were purified using the GeneClean kit (Bio 101). Ends of fragments generated by pancreatic DNase I or by nebulization were filled using the Klenow fragment of *E. coli* DNA polymerase I. DNA fragments were cloned into M13mp18 DNA linearized at the Smal polynuker site and dephosphorylated. Yield of phage plaques was about 10⁷ (μg M13 DNA)⁻¹ for cloning of *Alul* or nebulized DNA fragments. Some 10% of the plaques were white on plates containing IPTG. The cloning of fragments generated by pancreatic DNase I yielded 10⁸ plaques (μg M13 DNA)⁻¹, 50% of which were white. Plaques were replicated on Biodyne A membranes (Pall) and hybridized with 10 ng ³²P-labelled YAC 10-9 DNA using the Random Primed DNA Labeling Kit (Boehringer-Mannheim). Phages corresponding to the hybridizing plaques were propagated on *E. coli* strain JIC128F’ for 15 h in 48-well plates. The following steps were carried out with the aid of a BIOMEC 1000 Laboratory workstation (Beckman). Phage suspensions were stored in 96-well plates and 15 μl of each were used to prepare the filters (Hybond N⁺, Amersham) for a second hybridization screening. Positive clones (410 were identified) were stored and used for sequencing. About 90% of the positive clones carried inserts from the target region, as deduced from the analysis of the sequence. The remaining clones carried inserts from other regions of the *B. subtilis* chromosome and about 50% had homology with known genes.
YAC DNA was purified and used as a probe for hybridization with a total chromosomal DNA bank, constructed in the M13mp18 phage vector. Of the 60000 phage clones tested, 410 were positive. This number is almost fourfold lower than expected from the ratio of the YAC to chromosome length, which may be due to several mutually non-exclusive factors: (i) the size of the YAC may have been overestimated; (ii) a cloning bias against the regions covered by the YAC may exist; and (iii) the hybridization procedure used may not have allowed detection of all clones homologous to the probe.

All hybridizing inserts were sequenced, resulting in a total sequence length of about 54 kb, distributed among 125 contigs. Homology searches with protein data banks allowed us to identify the contigs corresponding to ilvA, metB, degR, thyB and xpt genes and also the contigs carrying the SPβ immunity region (immSPβ). A contig which has homology with GenBank entries M81760 and M81762, corresponding to attSPβ, was also identified. Oligonucleotides corresponding to these contigs and to the kdgT gene were designed and used for LA PCR amplification. Overlapping PCR products which covered the region between kdg and attSPβ were thus obtained, allowing construction of the genetic and physical map (Fig. 1a). The region between attSPβ and immSPβ was not amplified, which indicates that the distance between the two sites is longer than 25 kb, the maximal amplifiable fragment.

The LA PCR products were used as hybridization probes to attribute the 410 phage clones, and thus the sequenced contigs, to the intervals of the genetic map. The number of contigs and the length of the determined sequence in each of the three amplified 12 kb intervals is shown in Fig. 1(a). The total corresponding contig length was 24.4 kb, that is 68% of the size of this region, which indicates that the random sequencing had generated some 70% of the sequence of the YAC 10-9 insert. Since we did not intend to sequence the phage SPβ region present in the YAC, the sequences of the clones containing phage fragments were not used.

To complete the sequencing of the area delimited by the kdg and attSPβ markers, the three contiguous LA PCR products, xpt–degR, degR–thyB and thyB–attSPβ, were
### Table 1. General features of the CDSs

<table>
<thead>
<tr>
<th>CDS*</th>
<th>Endpoints (nt)t</th>
<th>Size (aa)</th>
<th>Translation start$</th>
<th>PIR entry</th>
<th>Description</th>
<th>Match (%)§</th>
</tr>
</thead>
<tbody>
<tr>
<td>ypbP (opr)</td>
<td>1426-2010</td>
<td>194</td>
<td>AcGAGGTAgaCagATG</td>
<td>reea</td>
<td>Adenine phosphoribosyl transferase (E. coli)</td>
<td>27 (141)</td>
</tr>
<tr>
<td>ypgP (gmuX)</td>
<td>2007-3323</td>
<td>438</td>
<td>AGGGGATcattcATG</td>
<td>s4223</td>
<td>Uracil transport protein (E. coli)</td>
<td>26 (410)</td>
</tr>
<tr>
<td>yphP (harA)</td>
<td>3397-4494</td>
<td>365</td>
<td>AaGAGGTTGATcgcATG</td>
<td>p0079</td>
<td>Chalcone synthase (plants)</td>
<td>26 (364)</td>
</tr>
<tr>
<td>yphQ</td>
<td>4498-5004</td>
<td>168</td>
<td>AaGGGGeGtcggATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yprR</td>
<td>5340-8921</td>
<td>1193</td>
<td>AaGGGGrGtrcggATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yphR</td>
<td>8986-9243</td>
<td>85</td>
<td>AAGGGAAGAaGcggATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yppQ (ewol)</td>
<td>9564-10455</td>
<td>296</td>
<td>AaGGGgaGTtatttcATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ypsQ</td>
<td>10633-11031</td>
<td>132</td>
<td>AaGGGAATcTcattcATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ypaR</td>
<td>11031-11720</td>
<td>229</td>
<td>AaGGGAaGcGGATcggATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yppP</td>
<td>11803-12483</td>
<td>26</td>
<td>AaGGGAAGAaGcggATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yppQ</td>
<td>12476-12658</td>
<td>60</td>
<td>AaGGGAAGAaGcggATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ypaR</td>
<td>13111-13293</td>
<td>60</td>
<td>AaGGGAAGAaGcggATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ypgP (hosA)</td>
<td>16416-16898</td>
<td>160</td>
<td>AaGGGAAGAaGcggATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ypgQ</td>
<td>16914-17351</td>
<td>205</td>
<td>AaGGGAAGAaGcggATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yphR</td>
<td>17591-18724</td>
<td>377</td>
<td>AaGGGAAGAaGcggATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>inoD</td>
<td>19013-20655</td>
<td>558</td>
<td>AaGGGAAGAaGcggATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ypbP</td>
<td>20796-21230</td>
<td>144</td>
<td>actaGAGGggtttttGTG</td>
<td>s5641</td>
<td>Homoserine succinyl transferase (E. coli)</td>
<td>47 (221)</td>
</tr>
<tr>
<td>yplP</td>
<td>21527-22045</td>
<td>172</td>
<td>gggGAGGaagaACTG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yplP</td>
<td>22303-22916</td>
<td>203</td>
<td>AgaGGGAaatsagcaggATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yplQ</td>
<td>22934-23467</td>
<td>177</td>
<td>AAGGAGGgtctcccaggATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>thiB</td>
<td>23551-24345</td>
<td>264</td>
<td>AAGGAGGgtctcccaggATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dfrA</td>
<td>24342-24848</td>
<td>168</td>
<td>AAGGAGGgtctcccaggATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yplK</td>
<td>24849-25469</td>
<td>206</td>
<td>GGAGGgtctcccaggATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yplQ</td>
<td>25508-26149</td>
<td>213</td>
<td>AAGGAGGgtctcccaggATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yplP</td>
<td>26170-27165</td>
<td>331</td>
<td>AAGGAGGgtctcccaggATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>is A</td>
<td>27414-28682</td>
<td>422</td>
<td>AAGGAGGgtctcccaggATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ypmP</td>
<td>28768-29019</td>
<td>83</td>
<td>AAGGAGGgtctcccaggATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ypmQ</td>
<td>29167-29748</td>
<td>193</td>
<td>AAGGAGGgtctcccaggATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ypmR</td>
<td>29826-30593</td>
<td>255</td>
<td>AAGGAGGgtctcccaggATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ypmS</td>
<td>30603-31166</td>
<td>187</td>
<td>AAGGAGGgtctcccaggATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ypmT</td>
<td>31179-31373</td>
<td>64</td>
<td>AAGGAGGgtctcccaggATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yppR</td>
<td>31445-32782</td>
<td>445</td>
<td>AAGGAGGgtctcccaggATG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yppP</td>
<td>32832-33257</td>
<td>141</td>
<td>AAGGAGGgtctcccaggATG</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Nomenclature for the putative CDSs is described in the text.
† Corresponding to L77246 GenBank entry.
‡ The putative start codon is indicated in bold and nucleotides complementary to 16S RNA are indicated in upper case letters.
§ The number of amino acids over which the percentage match was determined is shown in parentheses.
§§ SwissProt entry. §§§ GenBank entry. ** Position of the attachment site of SPβ is indicated.
used to construct M13 or pSGMU2 banks. About 300 clones were chosen at random and used as templates for ~ 500 sequencing reactions. Some 200000 bases were read, and a consensus sequence of 25 kb established. A total of 21 point mutations was revealed in the cloned final reading redundancy of the 355 kb sequence between kdg and by a T-rich sequence that could form a p-independent discrepancy, the correct sequence was determined by thus read, and a consensus sequence of 25 kb established.

putative start codon of 28 CDSs, while TTG and GTG were filled by direct sequencing of the PCR products. The procedure was therefore estimated to be approximately 1 in 10000 bp (21 /200 000), which is far below the error rate of a standard sequencing run, and thus should not affect the quality of the final sequence. In all cases of sequence discrepancy, the correct sequence was determined by sequencing several independent clones and/or direct sequencing of the product mixture from independent LA PCR experiments. The gaps remaining in the sequence were filled by direct sequencing of the PCR products. The final reading redundancy of the 35-5 kb sequence between kdg and attSPb was about 6. It should be possible to significantly reduce this value, by replacing some or most of the shotgun sequencing of the LA PCR products by primer walking. All of the sequence was read at least once on each strand.

Identification of putative genes

CDSs larger than 60 codons, exhibiting ATG, TTG or GTG as start codon and which had an appropriately located upstream sequence complementary to the 3' end of 16S rRNA (McLaughlin et al., 1981) were analysed. By these criteria, the region contains 38 CDSs and one truncated coding sequence, interrupted by the insertion of the SPb prophage. As reported in Table 1, ATG is the putative start codon of 28 CDSs, while TTG and GTG are the initiation codons for 9 and 1 CDSs, respectively. The genetic organization of the region is shown in Fig. 1(b). A transcriptional organization is also proposed, as deduced from the presence of inverted repeats followed by a T-rich sequence that could form a p-independent transcription terminator (Fig. 1b). The genes localized between the adjacent transcription terminators and transcribed in the same direction were considered to be in the same operon. We found 19 putative transcriptional units, 13 of which were co-directional with chromosome replication, a proportion similar to that observed by Zeigler et al. (1990). Sequence analysis indicated that 19 of the 38 proteins encoded by the CDSs were found to have homologues in the data banks, a value higher than the 38 proteins encoded by the CDSs were found to have homologues, and the transcriptional features of the region, are discussed below.

Gene functions in the xpt–attSPb region

ypaP–ypaQ (xpt–pbruX). The xanthine phosphoribosyltransferase gene (xpt), mutations which confer resistance to purine analogues, maps in this region (Saxild & Nygaard, 1987) and probably corresponds to ypaP. This hypothesis was confirmed by sequencing of the corresponding mutations (EMBL/GenBank/DDBJ accession no. X83878). The other gene of this operon, ypaQ, probably encodes a xanthine transport protein.

ypbp (bcsA). Since ypbp is the first bacterial gene related to plant chalcone synthase (Table 1) we propose the name bcsA (bacterial chalcone synthase A). Chalcone synthase catalyses the first step in the branch of phenylpropanoid metabolism leading to the synthesis of the ubiquitous flavonoid pigments and UV protectants. This pathway also leads to the synthesis of the isoflavonoid antibiotic phytoalexin, in response to wounding or microbial attack, which inhibits microbial growth by disrupting the cell membrane (for review, see Phillips & Kapulnik, 1995). In plants, different external stimuli, such as fungal elicitors or UV illumination, can induce the expression of chalcone synthase activity (Ryder et al., 1987). It is conceivable that BcsA is involved in bacterial adaptive and protective responses to environmental stress.

ypcP (exol). The product encoded by ypcP has > 35% identity with the N-terminal domain of DNA polymerase I (PolI) from E. coli and other species. The small product of partial PolII proteolysis (residues 1–300) carries the 5'-3' exonuclease activity while the large fragment contains 3'-5' exonuclease and DNA polymerase activities. Genetically encoded separated functions of PolII-type polymerases are known for bacteriophage T5, in which the gene denoted D15 encodes a 5'-3' exonuclease (Leavitt & Ito, 1989). D15 exonuclease is thought to be part of the bacteriophage T5 transcription–replication enzyme complex and may be involved in nick translation or in accumulation of ssDNA (Sayers & Eckstein, 1991). PolII has been purified from B. subtilis and the exonuclease activity was reported to be absent (Okazaki & Kornberg, 1964). However, the polA gene that encodes PolII was sequenced in our laboratory and the intrinsic 5'-3' exonuclease domain has been found (unpublished results). The exonuclease function of the YpcP protein might therefore indicate the redundant nature of this activity in B. subtilis.

ypdPQ. The ypdPQ gene encodes a protein which is 40% identical to a cell wall enzyme (elsB gene product) from Enterococcus faecalis (Bensing & Dunay, 1993). Conjugation in Ent. faecalis is thought to be mediated by an aggregation substance which may interact with a surface protein (binding substance) of the recipient cell. It was proposed that the els locus is related to the synthesis or assembly of the binding substance and that ElsB may act as a cell wall hydrolyase. The 70 aa N-terminal region of YpdPQ also exhibits similarity to the N-terminal region of ribonuclease HI from E. coli. Conserved amino acids, presumed to be at the catalytic site of RNase HI (Itaya & Kondo, 1991), are also present in YpdQ.

degR. The degR gene, involved in the regulation of transcription of the extracellular degradative enzyme genes, known also as prtR, has been identified and sequenced earlier by Yang et al. (1987).

ypfQ (cspD). ypfQ encodes a protein which is 82% identical to the major cold-shock protein CspB from B. subtilis
Fig. 2. Alignment of amino acid sequences. (a) Comparison of the B. subtilis CspD (Bs CspD) and the B. subtilis cold-shock protein B (Bs CspB). (b) Comparison of the B. subtilis yppP gene product (Bs YppP) and the E. coli B12-binding protein (Ec BtuE) and Sacch. cerevisiae glutathione peroxidase homologue (Sc GSH). (c) Comparison of the B. subtilis yplf gene product (Bs YplP) and bacterial transcriptional activators Klebsiella pneumoniae NifA (Kp NifA) and NtrC (Kp NtrC), B. subtilis LevR (Bs LevR) and RocR (Bs RocR). (d) Comparison of the B. subtilis yppP and yppQ gene products (Bs YppP and Bs YppQ), the tomato DNA-binding E4 protein (E4) and the N. gonorrhoeae PilB protein (Ng PilB). Black boxes indicate identical amino acids.

(Willinsky et al., 1992). A gene encoding a protein homologous to CspB and named cspC has also been detected in B. subtilis (GenBank accession no. D25230). We therefore propose to designate yppQ as cspD. As shown in Fig. 2(a), the CspD protein has only 12 amino acid differences from CspB, all of which are in the non-structured regions of the protein (Schindelin et al., 1993; Schnuchel et al., 1993). It is thus very likely that CspD can substitute for CspB. The limits of the role of CspB have been investigated by using a B. subtilis cspB mutant (Willinsky et al., 1992). It was noted that the viability of the mutant is drastically affected at a freezing temperature, but cryotolerance increases when the cells are preincubated for 2 h at 10°C before freezing. It was therefore suggested that genes other than cspB are involved in this tolerance. Probable candidates are cspC and cspD. It should be noted that the sequence 5' ATTGG, named the Y-box and thought to be recognized by CspB (Schröder et
The origin of replication. Almost perfect
the scale bar. The corresponding references are compiled in
monocistronic unit in the direction from the terminus to
probable that this CDS, homologous to homoserine
metB.
38 bp upstream from the translational start codon of
metB. The occurrence of a putative terminator down-
stream from metB suggests that it is transcribed as a
monocistronic unit from the terminus to
the origin of replication. Almost perfect — 10 (TATAAT)
and —35 (TTGAAA) promoter sequences are present
38 bp upstream from the translational start codon of metB.
In E. coli and S. typhimurium, transcription of the metB
counterpart is negatively regulated by the MetR protein
(Mares et al., 1992). In the putative regulatory region of
the B. subtilis metB gene we did not detect any sequence
resembling the E. coli MetR binding site.

ypgP (bsaA). The translational start codon of ypgP is
separated by 341 bp from that of the metB gene and the
two genes are transcribed divergently. The first gene of
the ypg operon, ypgP, has very high similarity (more than
50%, Table 1) with plant salt-stress-associated proteins.
The corresponding homologue in the orange is encoded
by esaA (cisus salt-stress-associated). The very high
similarity of these proteins over their whole length
strongly suggests that ypgP is involved in a stress response
in B. subtilis. We therefore propose the designation bsaA
(bacillus stress-associated) for this gene. This is the first
example of a potential esaA counterpart found in bacteria,
although one is known in yeast (PIR accession no.
S46121). The similarity of these proteins to glutathione
peroxidases may suggest that free radical levels increase
during the salt stress response (Holland et al., 1993). The
third CDS of this operon (ypgR) shares homology with the
mpv gene product from cyanobacteria. The homology
is, however, restricted to only a part of the two
proteins, which suggests that they may have a conserved
domain rather than a similar function.

thyB and dfrA. Thymidylate synthase, specified in B.
subtilis by the thyA and thyB genes, catalyses the synthesis
of dTMP from dUMP, while dihydrofolate reductase
dfrA) is involved in the reduction of dihydrofolate to
tetrahydrofolate. The thyB and dfrA genes are closely
linked and were cloned and characterized by Iwakura et al.
(1988). Our translated sequence differs by one amino acid
from both the ThyB (aa 144) and DfrA (aa 165) sequences
of that study. Verification of the sequencing gels indicated
no error in our sequence. A strain difference could
possibly account for the discrepancy.

ilvD and ilvA. The genes ilvA and ilvD, which are involved
in the synthesis of isoleucine, leucine and valine, have been
generically mapped in this part of the B. subtilis
chromosome (Barat et al., 1965). The B. subtilis ilvA
gene has been cloned and sequenced (GenBank accession no.
M58606). This sequence is identical to ours over 1-6 kb,
but the 5'-terminal 130 bp are completely different. We
believe that our sequence is correct, since it was de-
cermined from two different templates — the chromosomal
DNA cloned in M13 phage and the LA PCR product,
which was sequenced directly. The ilvD gene was
identified by the homology of the encoded product to the
corresponding E. coli enzyme (Table 1). In the species
closest to B. subtilis from which ilv genes have been
sequenced, Lactococcus lactis, as well as in E. coli and other
tenterobacteriaceae, ilvA and ilvD belong to the same
operon (Godon et al., 1992; Renault et al., 1995; Driver &
Lawther, 1985). In B. subtilis, however, the ilvA and ilvD
genes are separated by 10 kb.

ypLP. The N-terminal part of the ypLP gene product is
homologous to the central domain of the NifA and NtrC
proteins, which control nitrogen assimilation in Gram-
negative bacteria (Drummond et al., 1986). These proteins

Fig. 3. Alignment of the B. subtilis 168 genetic and CDS maps
between spoVAF and attSPβ. The CDS map is compiled from
Sorokin et al. (1993, 1996a) and the results of this study.
Experimentally confirmed gene functions are shown above
the scale bar. The corresponding references are compiled in
Table 2.
Table 2. Experimentally defined gene functions in the lysA-attSPβ region of the *B. subtilis* chromosome

<table>
<thead>
<tr>
<th>Gene*</th>
<th>Gene product function</th>
<th>Biological function or gene category</th>
<th>Reference†</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>lysA</em></td>
<td>DAP decarboxylase</td>
<td>Lysine biosynthesis</td>
<td>Yamamoto et al. (1991)</td>
</tr>
<tr>
<td><em>ppIB</em></td>
<td>Peptidyl-prolyl-isomerase</td>
<td>Unknown</td>
<td>Herrler et al. (1994)</td>
</tr>
<tr>
<td><em>sipS</em></td>
<td>Signal peptidase</td>
<td>Protein secretion</td>
<td>van Dijl et al. (1992)</td>
</tr>
<tr>
<td><em>ribG</em></td>
<td>Deaminase</td>
<td>Riboflavin biosynthesis</td>
<td>Mironov et al. (1994)</td>
</tr>
<tr>
<td><em>ribH</em></td>
<td>Riboflavin synthase (α-subunit)</td>
<td>Riboflavin biosynthesis</td>
<td>Schott et al. (1990)</td>
</tr>
<tr>
<td><em>ribA</em></td>
<td>GTP cyclohydrolase II</td>
<td>Riboflavin biosynthesis</td>
<td>Mironov et al. (1994)</td>
</tr>
<tr>
<td><em>ribT</em></td>
<td>Reductase</td>
<td>Riboflavin biosynthesis</td>
<td>Ludwig et al. (1987)</td>
</tr>
<tr>
<td><em>dabB</em></td>
<td>δ-D-Carboxypeptidase</td>
<td>Cell wall biogenesis</td>
<td>Buchanan &amp; Ling (1992)</td>
</tr>
<tr>
<td><em>tpmA</em></td>
<td>Unknown</td>
<td>Spore maturation</td>
<td>Popham et al. (1995)</td>
</tr>
<tr>
<td><em>tpmB</em></td>
<td>Unknown</td>
<td>Spore maturation</td>
<td>Popham et al. (1995)</td>
</tr>
<tr>
<td><em>resD</em></td>
<td>Signal transduction regulator</td>
<td>Respiration regulation</td>
<td>Sun et al. (1996)</td>
</tr>
<tr>
<td><em>resE</em></td>
<td>Protein kinase</td>
<td>Respiration regulation</td>
<td>Sun et al. (1996)</td>
</tr>
<tr>
<td><em>aroC</em></td>
<td>3-Dehydroquinate</td>
<td>Aromatic amino acid biosynthesis</td>
<td>Sorokin et al. (1993)</td>
</tr>
<tr>
<td><em>serA</em></td>
<td>Phosphoglycerate dehydrogenase</td>
<td>Serine biosynthesis</td>
<td>Sorokin et al. (1993)</td>
</tr>
<tr>
<td><em>gpsA</em></td>
<td>Glycerol-3-phosphate dehydrogenase</td>
<td>Membrane biosynthesis</td>
<td>Moribidoni et al. (1995)</td>
</tr>
<tr>
<td><em>spoIVA</em></td>
<td>Spore coat assembly protein</td>
<td>Spore maturation</td>
<td>Driks et al. (1994)</td>
</tr>
<tr>
<td><em>mitA</em></td>
<td>GTP cyclohydrolase 1</td>
<td>Folic acid biosynthesis</td>
<td>Babitzke et al. (1992)</td>
</tr>
<tr>
<td><em>mitB</em></td>
<td>RNA-binding transcription regulator</td>
<td>Tryptophan biosynthesis regulation</td>
<td>Babitzke et al. (1992)</td>
</tr>
<tr>
<td><em>gerC1</em></td>
<td>Unknown</td>
<td>Spore germination</td>
<td>Yazdi &amp; Moir (1990)</td>
</tr>
<tr>
<td><em>gerC2</em></td>
<td>Unknown</td>
<td>Spore germination</td>
<td>Yazdi &amp; Moir (1990)</td>
</tr>
<tr>
<td><em>gerC3</em></td>
<td>Unknown</td>
<td>Spore germination</td>
<td>Yazdi &amp; Moir (1990)</td>
</tr>
<tr>
<td><em>cheR</em></td>
<td>Chemotactic methyltransferase</td>
<td>Chemotaxis</td>
<td>Ordal et al. (1985)</td>
</tr>
<tr>
<td><em>aroH</em></td>
<td>Chorismate mutase, isoyzmes 1 and 2</td>
<td>Aromatic amino acid biosynthesis</td>
<td>Gray et al. (1990)</td>
</tr>
<tr>
<td><em>trpE</em></td>
<td>Anthranilate synthase</td>
<td>Tryptophan biosynthesis</td>
<td>Henner et al. (1984)</td>
</tr>
<tr>
<td><em>trpD</em></td>
<td>Anthranilate phosphoribosyltransferase</td>
<td>Tryptophan biosynthesis</td>
<td>Henner et al. (1984)</td>
</tr>
<tr>
<td><em>trpC</em></td>
<td>Indol-3-glycerol phosphate synthase</td>
<td>Tryptophan biosynthesis</td>
<td>Henner et al. (1984)</td>
</tr>
<tr>
<td><em>trpF</em></td>
<td>Phosphoribosylanthranilate isomerase</td>
<td>Tryptophan biosynthesis</td>
<td>Henner et al. (1984)</td>
</tr>
<tr>
<td><em>trpB</em></td>
<td>Tryptophan synthase β</td>
<td>Tryptophan biosynthesis</td>
<td>Henner et al. (1984)</td>
</tr>
<tr>
<td><em>trpA</em></td>
<td>Tryptophan synthase α</td>
<td>Tryptophan biosynthesis</td>
<td>Henner et al. (1984)</td>
</tr>
<tr>
<td><em>hisH</em></td>
<td>Histidinol-phosphate aminotransferase</td>
<td>Histidine biosynthesis</td>
<td>Henner et al. (1986)</td>
</tr>
<tr>
<td><em>tyrA</em></td>
<td>Prephenate dehydrogenase</td>
<td>Tyrosine biosynthesis</td>
<td>Henner et al. (1986)</td>
</tr>
<tr>
<td><em>aroE</em></td>
<td>3-Enolpyruvylshikimate-phosphate synthase</td>
<td>Aromatic amino acid biosynthesis</td>
<td>Henner et al. (1986)</td>
</tr>
<tr>
<td><em>gerA</em></td>
<td>Rieske iron-sulphur protein</td>
<td>Aerobic respiration</td>
<td>Yu et al. (1995)</td>
</tr>
<tr>
<td><em>gerB</em></td>
<td>Cytochrome <em>bc</em> complex b domain</td>
<td>Aerobic respiration</td>
<td>Yu et al. (1995)</td>
</tr>
<tr>
<td><em>gerC</em></td>
<td>Cytochrome <em>bc</em> complex c domain</td>
<td>Aerobic respiration</td>
<td>Yu et al. (1995)</td>
</tr>
<tr>
<td><em>hisD</em></td>
<td>Transcription regulator</td>
<td>Biotin biosynthesis regulation</td>
<td>Bower et al. (1995)</td>
</tr>
<tr>
<td><em>aspB</em></td>
<td>Aspartate aminotransferase</td>
<td>Aspartate biosynthesis</td>
<td>Hoch &amp; Mathews (1972)</td>
</tr>
<tr>
<td><em>dnaD</em></td>
<td>Unknown</td>
<td>DNA replication</td>
<td>Bruand et al. (1995)</td>
</tr>
<tr>
<td><em>cotD</em></td>
<td>Spore coat protein</td>
<td>Spore maturation</td>
<td>Donovan et al. (1987)</td>
</tr>
<tr>
<td><em>rnpB</em></td>
<td>RNase P RNA component</td>
<td>tRNA processing</td>
<td>Reich et al. (1986)</td>
</tr>
<tr>
<td><em>xpt</em></td>
<td>Xanthine phosphoribosyltransferase</td>
<td>Purine nucleotide biosynthesis</td>
<td>Saxild &amp; Nygaard (1987)</td>
</tr>
<tr>
<td><em>degR</em></td>
<td>Unknown</td>
<td>Protein secretion regulation</td>
<td>Tanaka et al. (1987)</td>
</tr>
<tr>
<td><em>metB</em></td>
<td>Homoserine succinyl transferase</td>
<td>Methionine biosynthesis</td>
<td>Barat et al. (1965)</td>
</tr>
<tr>
<td><em>lkb</em></td>
<td>Dihydroxyacid dehydratase</td>
<td>Isoleucine biosynthesis</td>
<td>Barat et al. (1965)</td>
</tr>
<tr>
<td><em>thyB</em></td>
<td>Thymidylate synthetase B</td>
<td>Thymidine biosynthesis</td>
<td>Iwakura et al. (1988)</td>
</tr>
<tr>
<td><em>dfrA</em></td>
<td>Dihydrofolate reductase</td>
<td>Folic acid metabolism</td>
<td>Iwakura et al. (1988)</td>
</tr>
<tr>
<td><em>ilvA</em></td>
<td>Threonine dehydratase</td>
<td>Isoleucine biosynthesis</td>
<td>Barat et al. (1965)</td>
</tr>
</tbody>
</table>

*Given in the order of their position on the map in Fig. 3.
† The most direct evidence of the product or biological function.
belong to the so-called NtrC/NifA family of transcription activators, which includes gene products involved in many different physiological functions, such as pilin formation (PilR), catabolism of xylene (XylR) or C4 dicarboxylate transport (DctD). In *B. subtilis*, the two examples known until now are involved in fructose and sucrose hydrolysis (LevR, Martin-Verstaete *et al.*, 1994) and in arginine catabolism (RocR, Calogero *et al.*, 1994). All the genes from the family contain a highly conserved domain, revealed by multiple alignments (Fig. 2c). This domain is supposed to be essential for formation of open promotor complexes with RNA polymerase (Popham *et al.*, 1989). It is possible that *yplP* encodes a protein which belongs to this family of regulators.

**yppP and yppQ.** These two CDSs, which are transcribed divergently from *yppP*, have a strong homology with the same protein, the *pilB* gene product of *Neisseria gonorrhoeae* (Taha *et al.*, 1988). The *yppP* sequence aligns with *PilB* from amino acids 227 to 347 while *yppQ* aligns with *PilB* from amino acids 390 to 515 (Fig. 2d). *PilB* acts negatively in trans on the pilin promoter. It has been shown that the 280 N-terminal residues of *PilB* are sufficient for repressor activity. We propose that *PilB* of *N. gonorrhoeae* might be a bifunctional protein which interacts with DNA via its N-terminal domain. In both *N. gonorrhoeae* and *B. subtilis*, the *pilB* and *yppP* genes are located upstream of a transcription regulator gene and are transcribed divergently from it. It is tempting to suggest that *yppP*, *yppQ* and *yppQ* may have a function related to the same physiological process. Inactivation of expression of pilin genes in *N. gonorrhoeae* results in strains that exhibit marked competence deficiency. While in *B. subtilis* no structure resembling a pilus has been observed, the late competence *comG* protein is similar to a type IV *N. gonorrhoeae* pilus protein (Albano *et al.*, 1989). It might be interesting to test whether inactivation of *YppP*, *YppP* or *YppQ* affects *B. subtilis* competence.

**ypqP (pipL).** This CDS is interrupted by the *SPβ* prophage. Obviously, its integrity is not essential for *B. subtilis*, since the corresponding protein is produced only when the phage is excised. We have designated this CDS *pipL* (Phage-Interrupted Protein near the Left attachment site).

**Conclusions**

The complete nucleotide sequence of the 35·5 kb *kdg-attSPβ* region was determined in this study. A contiguous sequence of 153 kb, extending between the *fisA* gene and the *SPβ* prophage, and corresponding to the region assigned to our laboratory within the *B. subtilis* genome sequencing project, was thus completed. The genetic and CDS organization of this region, the latter based on current and previous work (Sorokin *et al.*, 1993, 1996a) are summarized in Fig. 3. The gene functions from this region which have been defined experimentally, and the corresponding references, are listed in Table 2. The sequencing strategy used here relied on LA PCR. The procedure has proved to be very powerful for ordering of the contigs formed by random sequencing, and for generating DNA both for subcloning and for direct sequencing. The sequence errors due to LA PCR are not significant. We conclude that LA PCR products can be used directly within genome sequencing projects.

**REFERENCES**


Received 18 April 1996; revised 11 June 1996; accepted 24 June 1996.