INTRODUCTION

As part of the Bacillus subtilis genome sequencing project (Moszer et al., 1996) we were assigned the region between adeC (130°) and polC (147°). In previous communications we have described the sequencing of several segments of DNA from this region (Daniel & Errington, 1993; Daniel et al., 1994, 1996; Yanouri et al., 1993). Here we describe the sequence of a 28 kbp segment of DNA, obtained by plasmid integration and excision, using as a starting point the previously characterized spoVM gene (Levin et al., 1993). The sequence of the region is analysed and shown to contain genes with a wide range of functions.

METHODS

Bacterial strains. B. subtilis strain 168 (from C. Anagnostopoulos, INRA, Jouy-en-Josas, France) was used in all plasmid integration and excision experiments. Subcloning was performed in Escherichia coli strain DH5α (Life Technologies; (F−)endA1 hsdR17(s− m−) supE44 recA1 gyrA96 thi−1 relA1 Δ(lacZYA−argF)U169 φ80 galM ΔM15). During chromosome walking experiments plasmids were excised and recovered in E. coli strains DH5α or TP611 (from P. Glaser, Institute Pasteur, France; thi−1 thr− leuB6 lacY1 tonA21 supE44 hsdR hsdM recBC lop11 cya−610 pcr80 sad::Tn10).

DNA manipulations. The technique of chromosome walking was used to clone DNA from B. subtilis by the recovery of integrated plasmids from the chromosome (Niaudet et al., 1982) as described by Errington (1990). DNA manipulations were carried out using standard procedures (e.g. Sambrook et al., 1989). Plasmids pSG1302 (Stevens et al., 1992) or pSG2 (Fort & Errington, 1985) were used as the cloning vectors. To initiate chromosome walking downstream from spoVM, plasmid pSG2012 was constructed by cloning the 0.4 kbp HindIII/XhoI fragment from pSC6002 and subcloning it into pSG1301. pSG2012 was then transformed into B. subtilis 168 and a transformant was used for chromosome walking. To initiate the chromosome walking upstream from spoVM, a second plasmid (pSG2013) was constructed by isolating the 1.3 kb EcoRI/HindIII insert from pSC6002 and subcloning it into pSG1301. The resulting plasmid was then used for chromosome walking as described above.

Transformation. B. subtilis 168 was transformed by the method of Anagnostopoulos & Spizizen (1961) as modified by Jenkinson (1983). Transformants of B. subtilis containing integration plasmids were selected on Oxoid nutrient agar containing chloramphenicol (5 μg ml−1). E. coli strains DH5α and TP11 were transformed by the method of Sambrook et al. (1989).

PCR. PCR products were generated from template DNA of B. subtilis 168 using the conditions described by Innis & Gelfand (1990) or long range PCR (GeneAmp XL PCR kit, Perkin Elmer).

Sequence determination. The clones obtained by plasmid integration and excision were sequenced by primer walking using custom synthesized oligonucleotide primers. Plasmid DNA was prepared for sequencing using Hybaid recovery plasmid mini preps or the Qiagen Plasmid Midi Kit. The PCR products were purified with a QIAEX II Gel Extraction Kit (Qiagen). The purified DNA was sequenced using the ABI PRISM Dye Terminator Cycle Sequencing Kit (Perkin Elmer) and the Applied Biosystems 373 automatic DNA sequencer (Perkin Elmer). All of the sequence was determined on both strands.

The sequence of a 28 kbp segment of DNA surrounding the spoVM gene of Bacillus subtilis 168 (lying at approximately 145° on the standard genetic map) has been determined. The region contains 27 ORFs, a number of which have predicted products significantly similar to proteins in sequence databases, particularly to proteins involved in macromolecular synthesis of nucleic acids, proteins and phospholipids. A pair of closely linked genes encode a likely serine protein phosphatase and a serine protein kinase, respectively. Such proteins play important regulatory roles in eukaryotic cells but are rare in prokaryotes.

Keywords: Bacillus subtilis, spoVM
Sequence analysis. The DNA sequence was analysed using the Nucleotide Interpretation Program (Staden package) and protein databases were searched using FASTA and BLASTX (GGG). Comparison of the predicted protein sequences to the Prosite library of sequence motifs was searched using Motifs (GGG). Tmpred was used to predict membrane-spanning regions and their orientation (Hofmann & Stoffel, 1993).

RESULTS AND DISCUSSION

Cloning and sequencing

The spoVM-encoded protein is required for a late step in spore development (Levin et al., 1993) and the gene lies at 145° on a recent genetic map (Anagnostopoulos et al., 1993). The sequence of the previously cloned and sequenced spoVM gene was used as a base from which to chromosome-walk by plasmid integration and excision (Fig. 1). To begin walking 0.4 kbp (pSG2012) and 1.3 kbp (pSG2013) segments of DNA from the spoVM region were subcloned into the integration plasmid pSG1301. Initial clones obtained extending downstream from spoVM showed a different restriction pattern from that obtained previously by Levin et al. (1993). In retrospect, it appears that plasmid pSC603 obtained by Levin et al. (1993) must have contained two non-contiguous HindIII fragments of DNA, rather than having arisen by partial digestion as the authors suggested. We confirmed that our physical map of the spoVM region was correct by PCR amplification of DNA from B. subtilis strain 168 using appropriately positioned synthetic oligonucleotide primers.

In a series of steps a 28 kbp segment of DNA extending both upstream and downstream of spoVM was obtained (Fig. 1). After each round of chromosome walking, we confirmed that the clones obtained were contiguous with previous clones, either by plasmid walking back through the region or by PCR amplification of DNA from B. subtilis strain 168 using appropriately positioned synthetic oligonucleotide primers. The plasmids obtained were sequenced by primer walking with synthetic oligonucleotides and an automatic DNA sequencer.

ORFs

The sequence obtained (27779 bp) was analysed for the presence of ORFs that would encode peptides of at least 67 aa (with the exception of spoVM, the product of which is only 27 aa long as described by Levin et al., 1993) and 27 were identified. All except yloA and rpmB would be transcribed from left to right as shown in Fig. 1, which corresponds to the same direction as used by DNA replication forks traversing the region. The translation initiation signals for the putative ORFs are listed in Table 1. The initiation codon for most (17) of the ORFs was ATG, the next most frequent was TTG (5) and 4 ORFs used GTG (not including the two putative start sites of yloL; Table 1). Each start codon was preceded by a likely Shine–Dalgarno sequence; most of these were reasonably complementary to the 3′ end of the 16S rRNA of B. subtilis (3′ UCUUUCUCU-CCACUAG; Mountain, 1989). The spacing between the 3′ end of the Shine–Dalgarno sequence (5′ AGGAGG) and the first base of the initiation codon ranged from 11 to 5 bp. The most common spacings were 8 (25%), 6 (21%), 10 (18%) and 7 bp (14%).

ORFs with no known function

The amino acid sequence predicted for each ORF was searched for similarity to previously reported protein sequences using FASTA, BLASTX and Motifs. The results are summarized in Table 2. Three predicted products,
YloA, YloS and YlpC, shared no significant similarity to any known proteins in the databases. Several other sequences, yloC, yloM, yloN, yloQ, yloU and yloV, appear to encode proteins that are highly conserved in other bacteria but for which no function has yet been assigned. To try and obtain more information from the sequence of these proteins, the program TMPred was used to predict membrane-spanning regions and their orientation. YloU and YloQ have one strong putative transmembrane helix (N terminus inside the cell). YloA, YloC, YloM, YloN, YloS and YloV have no predicted transmembrane helices and they are therefore unlikely to be secreted or directly anchored to the membrane.

**ORFs to which probable functions can be assigned**

A substantial proportion of the ORFs showed strong similarity to genes of known function, allowing tentative functions to be assigned to the newly sequenced genes. The functions uncovered were mainly associated with macromolecular synthesis. Thus, there were genes involved in synthesis of DNA (priA, recG), protein synthesis (def, fmt and rpbM) and phospholipid synthesis (plsX, fabD and fabG). Others were involved in intermediary metabolism (dfp, araD and cfeE) or nucleotide precursor synthesis (gmk). One is probably involved in the movement of cations across the membrane (pacL) and its product is predicted to have nine membrane-spanning helices, as has Pacl from *Synechococcus sp.*

**ORFs with unusual features**

Two groups of genes found in the spoVM region are worthy of more extensive discussion. yloO and yloP encode proteins with signature motifs and similarity to families of eukaryotic proteins that act as serine protein phosphatases and kinases, respectively. The yloO and

---

**Table 1. Putative translation start signals**

<table>
<thead>
<tr>
<th>ORF</th>
<th>Endpoints (bp)*</th>
<th>Putative ribosome-binding site and translation start†</th>
</tr>
</thead>
<tbody>
<tr>
<td>yloA</td>
<td>592 &lt; 625</td>
<td>aatgcaggagaAtgAGAGG ggtTa TGc cat aTG</td>
</tr>
<tr>
<td>yloB</td>
<td>690 &lt; 724</td>
<td>aagagcagggaGATgGAGACG aGcgg aTG aag Gta</td>
</tr>
<tr>
<td>yloC</td>
<td>3445 &lt; 3479</td>
<td>aacagggagaAtgGAGG ttaaa aTG ata cGa</td>
</tr>
<tr>
<td>yloD</td>
<td>4552 &lt; 4588</td>
<td>attagcctcatacAcGAGcG aagaacc cTG cag cGg</td>
</tr>
<tr>
<td>yloH</td>
<td>5287 &lt; 5326</td>
<td>atataagaaaAtgGGAGG TGaaataaca aTg tta gAT</td>
</tr>
<tr>
<td>yloI</td>
<td>5577 &lt; 5610</td>
<td>gtaagcagaggagGGAGa attca cTG ctt aac</td>
</tr>
<tr>
<td>yloJ</td>
<td>6793 &lt; 6827</td>
<td>ctctcaaaacAAcCggAGG gcgag tcG aat tTt</td>
</tr>
<tr>
<td>yloK</td>
<td>9235 &lt; 9271</td>
<td>aagatcatcagctctggGGAGG gtaatcg TGa cag cGg</td>
</tr>
<tr>
<td>yloL†</td>
<td>9722 &lt; 9758</td>
<td>gaaactagggAtAtgGaAGG atgatttg aTG acG aag</td>
</tr>
<tr>
<td>yloM</td>
<td>10665 &lt; 10698</td>
<td>gctctggagtagGtggAGG aaAca aTG aag aaA</td>
</tr>
<tr>
<td>yloN</td>
<td>12009 &lt; 12045</td>
<td>tgtcagcatgAGAAAGgAGG gaaagag aTG gag gaa</td>
</tr>
<tr>
<td>yloO</td>
<td>13106 &lt; 13143</td>
<td>agctcgaaggAGAgAGG TGATgagg aTG tta cca</td>
</tr>
<tr>
<td>yloP</td>
<td>13863 &lt; 13901</td>
<td>tgtgctcaagatgAGGAGG gttgat gcG atc cGc</td>
</tr>
<tr>
<td>yloQ</td>
<td>15828 &lt; 15862</td>
<td>aagagtagaatAGACGAGG гGaaa tGc cag</td>
</tr>
<tr>
<td>yloR</td>
<td>16725 &lt; 16763</td>
<td>tttaacgagcAgAGCAGG Tgttagcata aTG ata aag</td>
</tr>
<tr>
<td>yloS</td>
<td>17452 &lt; 17488</td>
<td>tttcgttagcatgAAgAGG attcagag aTG aag aca</td>
</tr>
<tr>
<td>spoVM</td>
<td>18169 &lt; 18205</td>
<td>gggcagactgAtAGGAGG gGcAAggA aTG aaA aag</td>
</tr>
<tr>
<td>rpmB</td>
<td>18565 &lt; 18530</td>
<td>gtttggcacaAGgAGGAGG gaAaCaa aTG gca cGt</td>
</tr>
<tr>
<td>yloU</td>
<td>19166 &lt; 19201</td>
<td>aacgcagagcGGGAGGAGG ggtgaagatgc aTG aat cTT</td>
</tr>
<tr>
<td>yloV</td>
<td>19400 &lt; 19437</td>
<td>ggttagcggagAGGAGG aGcagag aTG aag aCC</td>
</tr>
<tr>
<td>yloW</td>
<td>20967 &lt; 21001</td>
<td>agctgctggtcAtAGGAGG aatagc aTG aaa aAC</td>
</tr>
<tr>
<td>ylpA</td>
<td>21673 &lt; 21689</td>
<td>tagcgaagagGGGAGGAGG aGcagatg aTG tgt cTT</td>
</tr>
<tr>
<td>ylpB</td>
<td>22530 &lt; 22569</td>
<td>tggcaggagctGAGGAGG gGcAGattc aTG ctt cGg</td>
</tr>
<tr>
<td>ylpC</td>
<td>24388 &lt; 24726</td>
<td>tagttcatagGtccGGAGG gttTtagat aTG aag aga</td>
</tr>
<tr>
<td>ylpD</td>
<td>25271 &lt; 25306</td>
<td>ctaatatattcAtAAGGAGG aGatg aTG aag aca</td>
</tr>
<tr>
<td>ylpE</td>
<td>26287 &lt; 26326</td>
<td>acagatgaatGGtggAGGAGG TrTtacatc aTG agr aag</td>
</tr>
<tr>
<td>ylpF</td>
<td>27235 &lt; 27272</td>
<td>aatgcaggagaAtgGGAGG aGatgagc aTG ctt aat</td>
</tr>
</tbody>
</table>

16S rRNA 3'-OH-UCUUUCUCUCC ACUAG

* The direction of the ORF is indicated: > as DNA replication; < counter to DNA replication.
† Bases shown in bold upper case letters represent bases complementary to 16S rRNA. Upper case letters in plain type indicate the putative start codon.
‡ There are two potential start sites for yloL. The putative ribosome-binding sites and translation starts are shown for each ORF.
Table 2. Properties of ORFs and their predicted products

<table>
<thead>
<tr>
<th>ORF</th>
<th>Motifs</th>
<th>Size (aa)</th>
<th>Name of related gene</th>
<th>Description (putative product/function)</th>
<th>Database accession no.*</th>
<th>Percentage identity (over aa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>yloA</td>
<td>-</td>
<td>201†</td>
<td>-</td>
<td>No significant identities found</td>
<td>gp:D16486</td>
<td>-</td>
</tr>
<tr>
<td>yloB</td>
<td>E1-E2 ATPase phosphorylation site (eukaryotic transport ATPase)</td>
<td>890</td>
<td>pcwL</td>
<td>Cation-transporting ATPase - Synechococcus sp.</td>
<td>gp:M99223</td>
<td>41 (994)</td>
</tr>
<tr>
<td>yloC</td>
<td>-</td>
<td>291</td>
<td>-</td>
<td>Hypothetical protein - E. coli†</td>
<td>sp:P23839</td>
<td>30 (291)</td>
</tr>
<tr>
<td>yloD</td>
<td>(i) Guanylate kinase signature (ii) ATP/GTP-binding motif A (P-loop)</td>
<td>244</td>
<td>gskI</td>
<td>S' Guanylate kinase - Saccharomyces cerevisiae</td>
<td>sp:P15454</td>
<td>41 (183)</td>
</tr>
<tr>
<td>yloE</td>
<td>-</td>
<td>67</td>
<td>rpoZ</td>
<td>DNA-directed RNA polymerase II subunit - E. coli†</td>
<td>sp:P08374</td>
<td>31 (60)</td>
</tr>
<tr>
<td>yloF</td>
<td>-</td>
<td>406</td>
<td>-</td>
<td>Cation-transporting ATPase - Ratius norvegicus</td>
<td>sp:P24285</td>
<td>41 (408)</td>
</tr>
<tr>
<td>yloG</td>
<td>-</td>
<td>805</td>
<td>prtA</td>
<td>Primosomal protein N' (replication factor Y) - E. coli†</td>
<td>sp:P17888</td>
<td>38 (379)</td>
</tr>
<tr>
<td>yloH</td>
<td>-</td>
<td>160</td>
<td>def</td>
<td>N-Formylmethionylaminoacyl-tRNA defformylase - E. coli†</td>
<td>sp:P27251</td>
<td>47 (143)</td>
</tr>
<tr>
<td>yloI</td>
<td>-</td>
<td>317</td>
<td>fntA</td>
<td>Methionylaminoacyl-tRNA formyltransferase - E. coli†</td>
<td>sp:P38882</td>
<td>46 (299)</td>
</tr>
<tr>
<td>yloJ</td>
<td>-</td>
<td>447</td>
<td>fmu</td>
<td>Fmu protein - E. coli†</td>
<td>sp:P36929</td>
<td>33 (404)</td>
</tr>
<tr>
<td>yloK</td>
<td>-</td>
<td>363</td>
<td>-</td>
<td>Hypothetical 40-kDa protein - E. coli‡</td>
<td>sp:P36979</td>
<td>36 (357)</td>
</tr>
<tr>
<td>yloL</td>
<td>-</td>
<td>254</td>
<td>-</td>
<td>Serine/threonine protein phosphatase - Mycoplasma genitalium</td>
<td>sp:P47354</td>
<td>28 (249)</td>
</tr>
<tr>
<td>yloM</td>
<td>-</td>
<td>648</td>
<td>pkn2</td>
<td>Serine/threonine protein kinase - Mycoplasma genitalis</td>
<td>sp:Z80233</td>
<td>34 (564)</td>
</tr>
<tr>
<td>yloN</td>
<td>-</td>
<td>317</td>
<td>fntA</td>
<td>Serine/threonine protein kinase - Mycobacterium tuberculosis</td>
<td>sp:M94878</td>
<td>35 (727)</td>
</tr>
<tr>
<td>yloP</td>
<td>-</td>
<td>298</td>
<td>-</td>
<td>Hypothetical 37-kDa protein - E. coli</td>
<td>sp:P39286</td>
<td>37 (229)</td>
</tr>
<tr>
<td>yloQ</td>
<td>-</td>
<td>217</td>
<td>cseE</td>
<td>Hypothetical protein - M. genitalium</td>
<td>sp:P47356</td>
<td>39 (229)</td>
</tr>
<tr>
<td>yloR</td>
<td>Ribulose-phosphate 3-epimerase family signature</td>
<td>553</td>
<td>-</td>
<td>rpoZ</td>
<td>Pentose-5-phosphate 3-epimerase precursor - Salmonella typhimurium</td>
<td>sp:P43843</td>
</tr>
<tr>
<td>yloS</td>
<td>-</td>
<td>70</td>
<td>sfpVM</td>
<td>Stage V sporulation protein M - B. subtilis</td>
<td>sp:P37817</td>
<td>100 (66)</td>
</tr>
<tr>
<td>yloW</td>
<td>-</td>
<td>220</td>
<td>yhaQ</td>
<td>Hypothetical protein MG850 - M. genitalium</td>
<td>sp:P46109</td>
<td>31 (773)</td>
</tr>
<tr>
<td>yloY</td>
<td>-</td>
<td>533</td>
<td>-</td>
<td>Hypothetical protein MG860 - M. genitalium</td>
<td>sp:P47609</td>
<td>31 (773)</td>
</tr>
<tr>
<td>yloZ</td>
<td>-</td>
<td>244</td>
<td>yhaP</td>
<td>t-Serine dehydratase part 1 - E. coli</td>
<td>sp:P46263</td>
<td>47 (73)</td>
</tr>
<tr>
<td>yloA</td>
<td>-</td>
<td>220</td>
<td>yhaQ</td>
<td>t-Serine dehydratase 1 - E. coli (see text)</td>
<td>sp:P16095</td>
<td>23 (212)</td>
</tr>
<tr>
<td>yloB</td>
<td>-</td>
<td>120</td>
<td>aspB2</td>
<td>t-Serine dehydratase 2 - E. coli (see text)</td>
<td>sp:P30744</td>
<td>27 (223)</td>
</tr>
<tr>
<td>yloC</td>
<td>-</td>
<td>533</td>
<td>-</td>
<td>Hypothetical protein ORFS1 - Lactobacillus sake</td>
<td>sp:G98238</td>
<td>64 (209)</td>
</tr>
<tr>
<td>yloD</td>
<td>-</td>
<td>220</td>
<td>yhaQ</td>
<td>t-Serine dehydratase part 2 - E. coli</td>
<td>sp:P46263</td>
<td>37 (263)</td>
</tr>
<tr>
<td>yloE</td>
<td>-</td>
<td>188</td>
<td>-</td>
<td>No significant identities found</td>
<td>sp:M96793</td>
<td>37 (526)</td>
</tr>
<tr>
<td>yloF</td>
<td>-</td>
<td>333</td>
<td>pksX</td>
<td>Fatty acid/phospholipid synthesis protein - E. coli</td>
<td>sp:M94927</td>
<td>32 (328)</td>
</tr>
<tr>
<td>yloG</td>
<td>-</td>
<td>317</td>
<td>fabD</td>
<td>Malonyl CoA-acyl carrier protein transacylase - E. coli</td>
<td>sp:M97040</td>
<td>44 (309)</td>
</tr>
<tr>
<td>yloH</td>
<td>-</td>
<td>172†</td>
<td>fabG</td>
<td>3-Ketoacyl-acyl carrier protein reductase - E. coli</td>
<td>sp:P23716</td>
<td>50 (171)</td>
</tr>
</tbody>
</table>

* gp, Genpept; sp, SWISS-PROT; ge, translation of a GenBank/EMBL DNA sequence. The number following an underscore denotes the position of the ORF in the sequence (i.e. _1 denotes the first ORF etc.).
† Partial coding sequence.
‡ A similar protein is found in Haemophilus influenzae.

yloP genes are adjacent and slightly overlapping, so it seems likely that their products act on a common target or targets. Proteins of this class are relatively rare in prokaryotes but possible homologues of yloO and yloP were found in other bacteria (Table 2). In eukaryotes these proteins generally act as regulators controlling a range of different functions (Hanks et al., 1988; Hunter, 1995). In bacteria, there are two well characterized examples of serine phosphorylation. One system serves to regulate the activity of σ factors, such as σ27 (Errington, 1996; Stratiger & Losick, 1996). The SpoIIB kinase and SpoIE phosphatase have opposing activities that regulate the state of phosphorylation of a specific serine residue on the SpoIIB protein. Accumulation of the non-phosphorylated form of SpoIIA leads to release of σ factor activity. The other system involves regulation of the sugar transport system (PTS), which is effected in part by phosphorylation of the Hpr protein on a specific serine residue (Voskuil & Chambliss, 1996). Since yloO and yloP appear to be embedded in an operon that
contains a gene involved in sugar metabolism (yloR/cfxE), it is attractive to suppose that yloP encodes the missing serine protein kinase for Hpr.

The organization of the yloO–yloQ region is also interesting because there is a similar arrangement of ORFs in Mycoplasma genitalium. Thus, MG108 encodes a putative serine/threonine protein phosphatase, MG109 encodes a putative serine/threonine protein kinase and MG110 encodes a hypothetical protein similar to YloQ of B. subtilis. This region (135,337–140,022 bp) of the M. genitalium chromosome also contains an ORF similar to CfxE of Alcaligenes eutrophus and YloR of B. subtilis, which also lies in close proximity to the yloO–yloQ region (Fig. 1).

Another interesting question is posed by the two contiguous ORFs, yloW and ylpA. YloW is similar to YhaQ and to the N-terminal regions of two full-length kinase and MG10 encodes a hypothetical protein similar to YloQ of B. subtilis. This region contains an ORF similar to CfxE of Alcaligenes eutrophus and YloR of B. subtilis, which also lies in close proximity to the yloO–yloQ region (Fig. 1).

More interestingly, it is remarkable to note that YloW and YlpA overlaps in the central region of L-serine dehydratase of E. coli. Moreover the organization of yloW and ylpA in B. subtilis is similar to that of yhaQ and yhab in E. coli. Although the alignment of YloW and YlpA overlaps in the central region of l-serine dehydratase, this is a region of only weak similarity (YloW shows 9% identity and YlpA shows 16% identity to SdaA of E. coli in this region). Database searches using a six-frame translation of the DNA sequence from the yloW/ylpA region, using BLASTX, revealed no evidence of frame shifts in the DNA sequence. Thus, they almost certainly represent two separate genes. It seems possible that they encode two subunits of an enzyme with a function related to l-serine dehydratase.

ACKNOWLEDGEMENTS

This work was funded by the BIOTECH programme of the European Community. We thank Annette Presscott and Alice Taylor for help with the DNA sequencing and Simon Cutting for providing the plasmids pSC600 and pSC603.

REFERENCES


Received 6 August 1997; revised 31 October 1997; accepted 7 November 1997.