Sequence analysis of the *Bacillus subtilis* chromosome region between the serA and kdg loci cloned in a yeast artificial chromosome

Alexei Sorokin, Vasco Azevedo,† Emmanuelle Zumstein,‡ Nathalie Galleron, S. Dusko Ehrlich and Pascale Serror

Author for correspondence: Alexei Sorokin. Tel: +33 1 34652533. Fax: +33 1 34652521, e-mail: sorokine@biotec.jouy.inra.fr

The standard strategies of genome sequencing based on λ-vector or cosmid libraries are only partially applicable to AT-rich Gram-positive bacteria because of the problem of instability of their chromosomal DNA in heterologous hosts like *Escherichia coli*. One complete collection of ordered clones known for such bacteria is that of *Bacillus subtilis*, established by using yeast artificial chromosomes (YACs). This paper reports the results of the direct use of one of the YAC clones from the above collection for the sequencing of the region cloned in it. The strategy applied consisted of the following: (i) construction of M13 banks of the partially purified YAC DNA and sequencing of 800 M13 clones chosen at random; (ii) directed selection of M13 clones to sequence by using marginal contig fragments as hybridization probes; (iii) direct sequencing of joining PCR fragments obtained by combinations of primers corresponding to the ends of representative contigs. The complete 104 109 bp insert sequence of this YAC clone was thus established. The strategy used allowed us to avoid resequencing the two largest, previously sequenced, contigs (13695 and 20303 bp) of the YAC insert. We propose that the strategy used can be applied to the sequencing of the whole bacterial genome without intermediate cloning, as well as for larger inserts of eukaryotic origin cloned in YACs. Sequencing of the insert of the YAC clone 15-6B allowed us to establish the contiguous sequence of 127 kb from *spolI A* to *kdg*. The organization of the newly determined region is presented. Of the 138 ORFs identified in the *spolI A-kdg* region, 57 have no clear putative function from their homology to proteins in the databases.

**Keywords**: *Bacillus subtilis* genome, yeast artificial chromosomes, combinatorial PCR

**INTRODUCTION**

The genomes of the bacteria *Haemophilus influenzae* and *Mycoplasma genitalium* have been completely sequenced (Fleishmann et al., 1995; Fraser et al., 1995) and several others, including *Escherichia coli* (Daniels et al., 1992), *Bacillus subtilis* (Kunst et al., 1995; Ogasawara et al., 1995) and *Mycobacterium leprae* (Honóře et al., 1993), are currently being sequenced. The sequence of the genome of the lower eukaryote *Saccharomyces cerevisiae* is also expected to be available soon (Williams, 1995). In most cases, the sequencing strategies used have been based on the construction of ordered λ- or cosmid-based vector collections in *E. coli* (Daniels et al., 1992; Glaser et al., 1993; Honóře et al., 1993; Oliver et al., 1992; Dujon et al., 1994; Wilson et al., 1994). The sequencing of the *H. influenzae* genome by a random approach was made feasible by the construction and use of a λ-based library (Fleishmann et al., 1995). In the case of the *B. subtilis* genome, and probably other AT-rich genomes of Gram-positive bacteria, many segments cannot be stably maintained in *E. coli*.
A. SOROKIN and OTHERS

In order to overcome this difficulty we have developed a sequencing strategy based on yeast artificial chromosomes (YACs; Burke et al., 1987). YACs have greatly facilitated genome mapping (Stalling et al., 1994), but until now had not been used for direct sequencing.

An ordered collection of fragments of the B. subtilis genome in YACs has been constructed (Azevedo et al., 1993a). Here we report the sequencing of one of the YACs from the collection, designated 15-6B. The YAC carries a 104 kb insert from the region of the chromosome near 205° which was assigned to our laboratory within the framework of the European B. subtilis genome project (Kunst & Deivine, 1991). The strategy used can presumably be applied to the megabase-scale YACs and to whole small genomes. The analysis of the nucleotide sequence is presented, focusing on the new genes for which a putative function can be postulated according to BLAST (Altschul et al., 1990) algorithms available on the NCBI e-mail server were used for rapid homology searches.

METHODS

Strains and growth conditions. Yeast cells containing artificial chromosomes were grown as described by Azevedo et al. (1993a). E. coli TG1 (K-12lac-pro) supE thi hsdR F' [traD36 proAB lacIq Δ(lacZ)M15] was used to make lawns on M13 selection plates and for growing phages in liquid medium. E. coli JJ128F' (araD139 araE139 leu2-3,112 metC16 galE15 galK16 Δ(lacZ)M15 hsdR hsdM' Str rif F' [lacIq Δ(lacZ)M15 traD36]), which reproducibly gave efficiency of electroporation by M13 DNA approaching 5 × 10⁸ p.f.u. µg⁻¹, was used in cloning experiments. Electrocompetent cells of JJ1C28F' were prepared according to the protocol described by Dowet et al. (1988) and stored at −80 °C. Plasmid pCGS966 (Smith et al., 1990) was a gift from Dr D. Smith (Genome Therapeutics).

Isolation of YAC DNA. A derivative of the yeast strain 15-6B (Azevedo et al., 1993a), called rec9, was constructed by transforming 15-6B with the EcoRI-NdeI fragment of pCGS966. Recombinants between the pCGS966 fragment and the left YAC15-6B arm, carrying the pBR322 replication, were selected on a medium containing selective agents as described by Smith et al. (1990). A stable clone with an amplified (four copies) YAC, designated 15-6Brec9, was identified by direct screening of the yeast chromosomal DNA from resulting transformants. The identity of the insert in 15-6Brec9 with the corresponding B. subtilis chromosome region was verified by Southern hybridization. Plugs containing yeast chromosomes were made as described by Anand et al. (1989). Electrophoresis was performed in 1×TBE at 10 V cm⁻¹ at 18 °C (1×TBE is 90 mM Tris/borate, 2 mM EDTA). The switching interval was from 0-3 to 6 s in forward migration and 0-1 to 2 s in reverse migration, for 20 h. After separation, YAC DNA was electroeluted in a dialysis membrane (Amicon) for 4 h at 10 V cm⁻¹. The solution containing YAC DNA was then concentrated fivefold by butanol treatment and DNA was treated with phenol/chloroform before ethanol precipitation. This protocol allowed us to obtain about 5–10 µg YAC DNA from 500 ml yeast cell culture.

M13 bank construction. The problem of constructing a representative and clean M13 library from purified YAC DNA was analysed recently (Yaudin et al., 1995). The authors suggested performing double purification of YAC by PFGE in order to avoid contamination by yeast DNA. We used another protocol, described below. Purified YAC DNA (0.5–10 µg) was partially digested by one of the following restriction enzymes: AalI, TagI, HpaII or HphI. The digested DNA was separated in agarose gel and segments between 500 and 1500 bp were purified as previously described (Sorokin et al., 1993). Purified DNA was ligated with 50 ng Smal- or AccI-cut and dephosphorylated M13mp19 vector in 20 µl ligase buffer (Boehringer), precipitated by 2-propanol using tRNA or glycogen as a carrier, rinsed with 70% ethanol, dried, dissolved in deionized water and used for electroformation of JJ1C28F' cells. The yield of phage plaques varied from 10⁶ to 10⁷ per µg M13 DNA, with a ratio of white to blue clones from 1 to 10.

Selection of M13 clones for random and directed sequencing. White M13 plaques were propagated for 5 h in 48-well plates in 2 x YT medium on JJ1C28F' or TG1 cells as a host. After centrifugation on GPKR centrifuge (Beckman), 150 µl aliquots of the phage supernatants were distributed into 96-well plates by using a BIOMEC 1000 laboratory workstation (Beckman); 50 µl aliquots of the same supernatants were used to prepare filters corresponding to the stock plates. The resulting filters were used to identify the M13 clones which carried YAC DNA inserts. For this purpose we used purified YAC DNA, labelled with ³²P by using a random-prime labelling kit (Boehringer), as a probe. The hybridized clones were randomly chosen for sequencing. The M13 clones sequenced during the directed sequencing step were selected by hybridization with the inserts of the phages corresponding to the ends of contiguous fragments (contigs) obtained during random sequencing. These inserts were amplified by PCR and pooled before labelling and hybridization. When only small gaps (estimated to be less than 500 bp) between contigs remained to be sequenced, the phage inserts used as probes were pooled in four different mixtures and only the M13 clones hybridizing with two of these mixtures were used for sequencing.

Sequencing. ssDNA of M13 phages was prepared as previously described (Sorokin et al., 1993). For reverse sequencing we used either M13 dsDNA prepared as described (Sorokin et al., 1995) or DNA generated by PCR using biotinylated primers and magnetic beads according to the supplier’s protocol (Dynal). Direct and reverse PCR sequencing was performed by using an Applied Biosystems PRISM direct or reverse sequencing kit on a Perkin Elmer 9600 thermal cycler or a Catalyst station from Applied Biosystems. ssDNA sequencing using Sequenase was performed as previously described (Sorokin et al., 1993).

Oligonucleotide synthesis and PCR. Oligonucleotides were purchased from Eurogentec or synthetized on an Oligo 1000 DNA synthesizer (Beckman). PCR was performed by using M13 DNA or supernatants and B. subtilis or yeast chromosomal DNA in conditions previously described (Sorokin et al., 1993).

Computing. The xbase and stnix programs (Dear & Staden, 1991) were used for gel assembling and consensus sequence analysis. The regions encoding proteins longer than 70 aa were further analysed for the presence of a sequence complementary to the 3’ end of 16S rRNA of B. subtilis, which is 3’ UCUUUCCUCCACUAG which is stream of a potential start codon (AUG, UUG or GUG) considering that the largest ORF is the most probably functional. In cases of doubt, the statistical tests, namely codon usage or uneven positional base frequencies, offered by the stnix program were also applied. FASTA (Pearson & Lipman, 1988) or BLAST (Altschul et al., 1990) algorithms available on the NCBI e-mail server were used for rapid homology searches.
**ORF nomenclature.** In the framework of the *B. subtilis* Genome Project the names of the ORFs of the *lysA*-ile-A region should start with *yp*-. All the putative ORFs identified in this region were therefore named *ypXZ*, with *x* assigning them to a putative operon or divergon proposed on the basis of locations of *p*-independent transcription terminators, and *Z* distinguishing different ORFs in the same putative operon or divergon. The names of genes are used instead of, or along with, the *yp*-names when the function of an ORF has been determined experimentally or can be proposed as a result of a homology search. Three sequences (Bruand *et al.*, 1995; Sorokin *et al.*, 1995; Bower *et al.*, 1995; accession numbers U11289, U11687 and L38424) from the YAC15-6B region were released before the *yp*-nomenclature had been chosen. The ORF names in these releases started with *yp*- instead of *yp*-. They have been therefore correspondingly renamed here.

**Sequence accession numbers.** The sequences corresponding to the YAC15-6B insert are available from GenBank under accession numbers L09228 (positions 1–13695 in 15-6B), L47648 (12396–37282), M80926 (37015–39348), M80245 (39345–57318), L47709 (57126–80900) and L47838 (80824–104109). The entries L47648, L47709 and L47838 contain the newly determined sequence.

### RESULTS AND DISCUSSION

**Direct sequencing of a YAC**

**Sequencing strategy.** Genomic sequencing generally comprises two main steps, one random and the other directed. The first is very efficient in the early stages of the project, and allows rapid generation of the bulk of the sequence, but it becomes inefficient later, since most of the randomly determined sequence is not novel. A switch to directed sequencing is therefore required to avoid an inordinate level of sequence redundancy. Our approach also used random and directed steps, carried out as follows.

Random sequencing was carried out on YAC segments subcloned in M13-derived vectors as described below. A number of contigs was thus established. Directed sequencing involved two phases. In the first, the clones from the M13 bank that allowed extension of the sequence were identified by hybridization, using as probes segments from the ends of the contigs. In the second, the contigs were ordered by combinatorial PCR and the gaps between them were closed by sequencing the appropriate PCR products.

**Isolation of YAC DNA.** To facilitate preparation of YAC DNA, we attempted to modify the copy number of YAC15-6B, using the approach described by Smith *et al.* (1990). For this purpose, pCGS966 was cleaved with *EcoRI* and *NdeI* and introduced by transformation into a yeast strain carrying YAC15-6B, selecting for expression of the thymidine kinase gene. The chromosomes of the transformants, cultivated under conditions inducing YAC amplification, were analysed by PFGE. Unexpectedly, YAC amplification was not induced in any of the clones. However, screening of clones grown on standard selective medium (Anand *et al.*, 1989) revealed that one of them, designated *repP*, contained about four copies of the YAC. The reasons for this increase in copy number are not understood. Rearrangements had not taken place in the *B. subtilis* moiety of the amplified YAC, as judged by Southern hybridization (not shown), and DNA of this amplified YAC was therefore used for sequence analysis. The DNA was purified by PFGE and used for the construction of M13 libraries. Hybridization of these collections with purified YAC DNA indicated that 20–50% of the M13 clones contained a DNA insert from the YAC. Other M13 phages forming white plaques presumably contained yeast chromosomal DNA.

**Random and directed sequencing step.** About 800 M13 clones of the YAC15-6B library were randomly chosen for sequencing. This generated ~90 kb of the consensus sequence, organized in 80 contigs, with a redundancy of 2:8. About 86% of the 104 kb YAC insert was thus sequenced. At this point we decided to switch to the direct sequencing. Two contiguous regions sequenced earlier (positions 1–13695 and 37015–57318; Sorokin *et al.*, 1993; Roels *et al.*, 1992; Henner *et al.*, 1990) were excluded from further analysis. We first attempted to extend the sequence of 40 of the largest contigs. For this purpose inserts of 80 M13 clones corresponding to the ends of the contigs were amplified by PCR, pooled in series of 20 and used as probes for hybridization with M13 clones that had not been previously sequenced. Sequencing of 300 positive clones extended the total contig length to ~96 kb, organized in 47 contigs. Iteration of this procedure revealed new hybridizing M13 clones, but did not allow us to extend the sequence significantly. This indicated that the appropriate M13 clones were absent or extremely rare in the M13 collection.

A similar conclusion was reached in attempts to identify the M13 clones carrying inserts that bridge two contigs and therefore hybridized with two different probe pools. Only five clones among 3000 tested joined the contigs. Given that the average distance between two contigs was ~250 bp (~12 kb missing sequence distributed among 47 contigs), and the average M13 insert size was about 500 bp, many more bridging clones should have been identified. This supports the assumption that some chromosomal regions were severely under-represented in the M13 collection.

**Combinatorial PCR.** To complete the YAC sequence we decided to synthesize the missing regions by PCR, using *B. subtilis* chromosomal DNA as template. Forty oligonucleotides corresponding to the ends of the 20 longest contigs were used for this purpose. PCR products were obtained for all missing regions, varying in size from 0:1 to 6 kb. They were used as probes for screening 2000 M13 clones which had not been sequenced previously; 258 hybridizing clones were detected. Sequencing of the clones allowed us to make only eight additional contig joins, and several new contigs were formed. This result confirmed that the M13 bank was biased, and suggested that some *B. subtilis* regions cannot be cloned in *E. coli* even in short segments on M13 phages. This conclusion is supported by our failure to clone the PCR fragments in *E. coli* directly. The sequence of the YAC was therefore completed by primer walking over the PCR products.

The length of the novel sequence is 70111 bp.
Fig. 1. Organization of the *B. subtilis* chromosome region between the spoIIA and kdg loci. Thick arrows correspond to ORFs, detected by sequence analysis. Thin arrows show putative transcripts. Small ‘t’s show the positions of stem-loop structures which may function as transcription terminators. The insert in YAC15-6B starts from the middle of ypuL (position 14512) and ends in the middle of ypuA (position 118620). Nomenclature of ORFs is explained in the text. The regions from spoIIE to serA and from spoIVA to aroE have been discussed elsewhere (Sorokin et al., 1993; Henner et al., 1990). The ORFs with unknown functions from the first publication are renamed as follows: orfX9 to ypuA; orfX1 to ypuC; orfX4 to ypuD; orfX5 to ypuE; orfX6 to ypuF; orfX7 to ypuG; orfX8 to ypuH; orfX13 to ypuI; orfX21 to ypuN. The involvement of the previously found orfX11 and orfX12 in spore maturation, giving the names spmA and spmB, was found recently (Popham et al., 1995). resABCDE were renamed from orfX14–18 after establishing the respiratory function of this operon (Sun et al., 1995). orfX20 was renamed sigX because this gene encodes a σ-factor (Lonetto et al., 1994). Other genes indicated above the scale bar encode the following: spo, spolB regulation genes; lysA, diaminopimelate decarboxylase; ppIB, peptidyl-prolyl isomerase; rib, riboflavin biosynthesis operon; dacB, D-9-carboxypeptidase; aroc, 3-dehydroquinase dehydratase; serA, phosphoglycerate dehydrogenase; fer, ferredoxin; recQ, product homologous to the *E. coli* recQ gene product; gud, glutamate dehydrogenase; cmk, CMP kinase; rpsA, S1 ribosomal protein; era, product homologous to the *E. coli* era gene product; gpsA, glycerol phosphate dehydrogenase; dbpA, DNA-binding protein; mtr, S-methyltryptophan resistance; gerC, germination deficiency; ndk, nucleotide diphosphate kinase; cheR, chemotactic methyl transferase; ara, general pathway of aromatic amino acid biosynthesis; trp, tryptophan biosynthesis; hisH, histidine biosynthesis; tyrA, tyrosine biosynthesis; qcrABC, respiration bf-complex; dapB, dihydropicolinate reductase; pap5, poly(A) polymerase; birA, biotin biosynthesis regulation; pan, pantothentic acid biosynthesis; dinG, product homologous to the *E. coli* dinG gene product; aspB, aspartate aminotransferase; asnS, asparagine-tRNA synthetase; dnaD, a protein involved in the initiation of replication; nth, endonuclease III; ponA, penicillin-binding protein 1; cotD, spore-coat protein; rnpB, RNase P RNA component; kdu, kdg, enzymes of pectin and galacturonic catabolism. Numbering starts from the first base pair of the SalI site in the spoIVA gene, the same as for the entry with GenBank accession number L09228 (Sorokin et al., 1993).
the newly determined sequence to the known regions of 13695 and 20303 bp, mentioned above.

**Analysis of the sequenced region**

The sequenced region was analysed for the presence of ORFs encoding proteins larger than 70 amino acids and having an appropriate ribosome-binding site. The presence of structures characteristic of p-independent transcriptional terminators was also examined. The organization of the spoIVA–kgd region of the *B. subtilis* chromosome is presented in Fig. 1. The spoIVA–serA and spoIVC–araE regions were described earlier (Sorokin et al., 1993; Roels et al., 1992; Henner et al., 1990). In the *serA–kgd* area, which is 90 kb long, 98 ORFs were detected. The average gene size is therefore 0.9 kb, a figure very similar to that previously reported (Glaser et al., 1993; Ogasawara et al., 1994). For 89 ORFs, that is 91% of those detected, the direction of transcription is colinear with the replication fork movement. The ypJ–ypD–birA, aspB–nth and ponA regions were described earlier (Sorokin et al., 1993). More detailed results of the database homology analysis are presented in Table 1 and discussed below.

**Genetic organization of the sequenced region**

**serA and ypaA.** The *serA* and *ypaA* genes are organized in two divergent transcription units which share the regulatory region. *serA* was partially sequenced before and shown to be involved in serine biosynthesis (Sorokin et al., 1993).

**ypbA (fer) and ypbC.** The *ypbA (fer)* product is 79% identical to ferredoxin of *Bacillus stearothermophilus*, suggesting that *ypbA* encodes a ferredoxin. It shares the regulatory region with the divergent *JypbE*–*JpbF* operon. The protein encoded by *ypbC* shares high homology with RecQ of *E. coli*, which is believed to be involved in the SOS response (Irino et al., 1986). Similarity of a region of these proteins (RecQ and YbpC) to RNA helicases of prokaryotes, they form a novel family of regulators, their functional and structural characterization of the newly discovered genes (Sorokin et al., 1995; Bower et al., 1995; Bruand et al., 1995; Popham & Setlow, 1995). The functional and structural characterization of the cotD and rmpP genes was also previously reported (Donovan et al., 1987; Reich et al., 1986). More detailed results of the database homology analysis are presented in Table 1 and discussed below.

**ypfC (cmk) and ypfD (rpsA).** The products of the genes *ypfC* and *ypfD* are homologous to the *E. coli* *mscA* and *rpsA* gene products, respectively, the latter encoding the S1 ribosomal protein. The similar organization of this region in *B. subtilis* and *E. coli* led us to suggest that these genes are counterparts (Sorokin et al., 1995), although the function of the second gene in *B. subtilis* remains unclear. It has been shown recently that *mscA* of *E. coli* encodes cytidine monophosphate kinase and the gene was therefore renamed *cmk* (Fricke et al., 1995). It was reported that, in contrast to *E. coli*, *B. subtilis* contains only one pyrimidine ribonucleoside monophosphate kinase, synthesizing both UMP and CMP (Walch & Ingraham, 1976). This is in good agreement with our finding that *ypfC* is essential (Sorokin et al., 1995).

**ypgA.** This gene has the highest homology with the hypothetical protein ORF6 from the *Erwinia herbicola* carotenoid biosynthesis gene cluster. It also shares a homologous region with flavocorticobromes b and 2-hydroxy-acid oxidases, shown to be essential for the activity of these proteins (Diép Lé & Lederer, 1991). However, several amino acids essential for the binding of flavin are absent in *ypgA* and therefore no function for this gene can be suggested.

**ypbC (era).** The product of this gene shows homology with the *E. coli* Era protein, an essential GTPase with a regulatory function. These proteins are structurally similar to RAS proteins of eukaryotes (Ahn et al., 1986). In prokaryotes, they form a novel family of regulators, comprising at least five members (Ahn et al., 1986; Trach & Hoch, 1989; Yamashita et al., 1993; Ogasawara et al., 1994), but the function of none has yet been established. A distinctive feature of YpbC from the proteins of this family is that it has two GTP-binding motifs (not shown). Studies of Era-depleted temperature-sensitive mutants of *E. coli* indicated that the protein is required for adaptation to thermal stress (Lerner & Inoye, 1991) and cell division (Gollop & March, 1991). The closest protein for which activities and tertiary structure are known in detail is EF-Tu (Weijland & Parmeggiani, 1993).

**ypfD (gpsA).** The protein encoded by *ypfD* has high homology to the NADH-dependent glyceraldehyde-3-phosphate dehydrogenases. Mutants devoid of this activity, designated *glyc*, were isolated and studied in two laboratories (Oh et al., 1973; Freese & Oh, 1974; Lindgren & Rutberg, 1974). Three such mutations were mapped by Rutberg's group and high linkage was detected between *trpC*, *his* and *glyc* markers by transformation. Taking into account the known distance between *trpC* and *hisH* (5 kb), the approximate distance between these markers might be estimated as 10 kb, with the order *glyc-trpC-hisH*
Table 1. Results of comparison of ORFs with sequences in the PIR protein data base

<table>
<thead>
<tr>
<th>ORF</th>
<th>End-points (bp)*</th>
<th>Size (aa)</th>
<th>PIR entry</th>
<th>Description (size, aa)</th>
<th>Homology (FASTA score)</th>
<th>Match (%)†</th>
</tr>
</thead>
<tbody>
<tr>
<td>serA</td>
<td>14013–12436c</td>
<td>526</td>
<td>deecpg</td>
<td>Phosphoglycerate dehydrogenase E. coli (410)</td>
<td>324</td>
<td>30.5 (324)</td>
</tr>
<tr>
<td>ynaA</td>
<td>14510–15082</td>
<td>191</td>
<td>s02153</td>
<td>NADH dehydrogenase (ubiquinone)</td>
<td>130</td>
<td>28.6 (126)</td>
</tr>
<tr>
<td>yphA (for)</td>
<td>15370–15122c</td>
<td>83</td>
<td>FEBSSF</td>
<td>Ferredoxin B. stearothermophilus (81)</td>
<td>369</td>
<td>79.0 (81)</td>
</tr>
<tr>
<td>yphC (reqQ)</td>
<td>16687–18177</td>
<td>497</td>
<td>bveeqq</td>
<td>reQ protein of E. coli (610)</td>
<td>788</td>
<td>40.3 (380)</td>
</tr>
<tr>
<td>ypcA (gadL)</td>
<td>21749–23029</td>
<td>427</td>
<td>a38168</td>
<td>Glutamate dehydrogenase Clostridium difficile (421)</td>
<td>1055</td>
<td>51.6 (419)</td>
</tr>
<tr>
<td>ypdA</td>
<td>23138–24112</td>
<td>325</td>
<td>s18998</td>
<td>Hypothetical protein B. licheniformis (122)</td>
<td>454</td>
<td>77.0 (122)</td>
</tr>
<tr>
<td>yph/C (cmk)</td>
<td>28377–29051</td>
<td>225</td>
<td>s29160</td>
<td>Hypothetical protein Clostridium pasteurianum (308)</td>
<td>183</td>
<td>28.4 (296)</td>
</tr>
<tr>
<td>yphD (rpsA)</td>
<td>29284–30432</td>
<td>383</td>
<td>r3ec1</td>
<td>Ribosomal protein S1 E. coli (557)</td>
<td>662</td>
<td>39.0 (333)</td>
</tr>
<tr>
<td>yphD (gprA)</td>
<td>34909–35946</td>
<td>345</td>
<td>a26118</td>
<td>Polynucleotide phosphorylase E. coli (711)</td>
<td>214</td>
<td>24.2 (236)</td>
</tr>
<tr>
<td>yphC (gcrA)</td>
<td>60000–60503</td>
<td>168</td>
<td>s13643</td>
<td>PRP22 S. cerevisiae (1145)</td>
<td>172</td>
<td>21.2 (255)</td>
</tr>
<tr>
<td>yphD (gcrB)</td>
<td>60505–61077</td>
<td>191</td>
<td>ckb16p</td>
<td>Cytochrome b Chlorella protothecoides (215)</td>
<td>404</td>
<td>42.3 (182)</td>
</tr>
<tr>
<td>yphE (gcrC)</td>
<td>61215–61982</td>
<td>256</td>
<td>wmsp17</td>
<td>Cytochrome b Spinacia oleracea (382)</td>
<td>174</td>
<td>29.1 (117)</td>
</tr>
<tr>
<td>yphA</td>
<td>62129–62686</td>
<td>186</td>
<td>b31383</td>
<td>Hypothetical 83K protein (gpr locus)</td>
<td>107</td>
<td>24.7 (81)</td>
</tr>
<tr>
<td>yphE (dapB)</td>
<td>64950–65753</td>
<td>268</td>
<td>rdecpd</td>
<td>Dihydrodipicolinate reductase Bacillus subtilis (731)</td>
<td>480</td>
<td>44.9 (205)</td>
</tr>
<tr>
<td>yphH</td>
<td>66881–68014</td>
<td>372</td>
<td>a61112</td>
<td>Trans-acting transcription factor SPT14</td>
<td>288</td>
<td>21.6 (380)</td>
</tr>
<tr>
<td>yphI (papS)</td>
<td>68019–69212</td>
<td>398</td>
<td>s11887</td>
<td>mRNA nucleotide transferase E. coli (412)</td>
<td>224</td>
<td>30.4 (224)</td>
</tr>
<tr>
<td>birA (birA)</td>
<td>69213–70190</td>
<td>326</td>
<td>bvecbf</td>
<td>birA bifunctional protein E. coli (321)</td>
<td>342</td>
<td>26.7 (322)</td>
</tr>
<tr>
<td>ypkA (panB)</td>
<td>70420–71253</td>
<td>278</td>
<td>P31057S</td>
<td>Ketopantoate hydroxymethyltransferase E. coli (264)</td>
<td>630</td>
<td>47.1 (263)</td>
</tr>
<tr>
<td>ypkB (panC)</td>
<td>71255–72115</td>
<td>287</td>
<td>P31663S</td>
<td>Panthothenate synthetase E. coli (283)</td>
<td>566</td>
<td>40.8 (282)</td>
</tr>
<tr>
<td>ypkC (panD)</td>
<td>72117–72500</td>
<td>128</td>
<td>P31664S</td>
<td>Aspartate 1-decarboxylase E. coli (126)</td>
<td>254</td>
<td>48.3 (116)</td>
</tr>
<tr>
<td>yplA (dinG)</td>
<td>72626–75421</td>
<td>932</td>
<td>ag2717</td>
<td>dinG protein E. coli (716)</td>
<td>366</td>
<td>22.4 (606)</td>
</tr>
<tr>
<td>aspB (aspB)</td>
<td>76251–77432</td>
<td>394</td>
<td>a38621</td>
<td>Aspartate aminotransferase Bacillus sp. (392)</td>
<td>1428</td>
<td>71.3 (387)</td>
</tr>
<tr>
<td>yplN (aum)</td>
<td>77576–78888</td>
<td>431</td>
<td>syecn7</td>
<td>Asparaginyl-tRNA synthetase E. coli (466)</td>
<td>624</td>
<td>41.1 (265)</td>
</tr>
<tr>
<td>dnaD (dnaD)</td>
<td>78961–79659</td>
<td>233</td>
<td></td>
<td>No homologies</td>
<td>&lt; 70</td>
<td></td>
</tr>
<tr>
<td>ypbO (prtO)</td>
<td>79666–80355</td>
<td>220</td>
<td>a32412</td>
<td>Endonuclease III E. coli (211)</td>
<td>439</td>
<td>46.0 (208)</td>
</tr>
<tr>
<td>penA (penA)</td>
<td>83468–89094c</td>
<td>915</td>
<td>s28033</td>
<td>Penicillin-binding protein 1a Streptococcus pneumoniae (719)</td>
<td>1083</td>
<td>39.1 (586)</td>
</tr>
<tr>
<td>yppB</td>
<td>84290–83670</td>
<td>207</td>
<td>y23k_stror$</td>
<td>Hypothetical 23 kDa protein in pan.A S' region Streptococcus oralis (198)</td>
<td>477</td>
<td>50.5 (179)</td>
</tr>
</tbody>
</table>

*ORF end-points in base pairs (bp).
†Match (%) indicates the percentage of amino acids identical in the alignment.
‡PIR entry numbers are as follows: deecpg (P31663S), s02153 (P31664S), ag2717 (P31057S).
§y23k_stror$ represents a hypothetical protein.
ypjE (dapB). This gene encodes a protein highly homologous to the \( E.\ coli \) dihydropicolinate reductase, an enzyme involved in the early steps of lysin biosynthesis. An enzyme with such activity was isolated from sporing \( B.\ subtilis \) cells (Kimura, 1975). Since the protein isolated by Kimura (1975) is different from that encoded by \( yplE \), as judged by molecular mass (18.5 kDa and 29.5 kDa, respectively), it is possible that \( B.\ subtilis \) uses two enzymes for this pathway. This is in agreement with the finding of strikingly different dihydropicolinate reductases in \( Bacillus cereus \) and \( B. megaterium \) (Kimur & Goto, 1977), leading to the notion of the existence of two different types of this enzyme in bacilli (Kimura et al., 1978).

**ypjH.** The \( yplH \) gene product has homology with the yeast SPT14 and the human PIG-A proteins, which were recently shown to be involved in glycosylphosphatidylinositol (GPI) synthesis (Schönbächler et al., 1995). GPI is involved in anchoring of some proteins to the plasma membrane of eukaryotes. The presence of such a gene in \( B.\ subtilis \) suggests the existence of GPI-anchored proteins in this bacterium.

**ypjI.** About half of the protein encoded by \( yplI \) is highly homologous to the \( E.\ coli \) mRNA polyadenylase and tRNA nucleotidyltransferase. Multiple alignments of these proteins (not shown) indicate that it is more probable that they have a common motif, possibly involved in RNA-binding, rather than a common function.

### Table 1. (cont.)

<table>
<thead>
<tr>
<th>ORF</th>
<th>End-points (bp)*</th>
<th>Size (aa)</th>
<th>PIR entry</th>
<th>Description (size, aa)</th>
<th>Homology (FASTA score)</th>
<th>Match (%)†</th>
</tr>
</thead>
<tbody>
<tr>
<td>ypgB</td>
<td>87653–88159</td>
<td>169</td>
<td>wqeb3t</td>
<td>Phosphotransferase system enzyme II, glucose specific ( S.\ typhimurium ) (530)</td>
<td>357</td>
<td>40.1 (167)</td>
</tr>
<tr>
<td>yprA</td>
<td>88262–90511</td>
<td>750</td>
<td>s00986</td>
<td>Initiation factor ( eIF-4A) ( M.\ mucedo ) (407)</td>
<td>121</td>
<td>20.0 (140)</td>
</tr>
<tr>
<td>cotD</td>
<td>92081–92308</td>
<td>76</td>
<td>d27393</td>
<td>cotD protein ( B.\ subtilis ) (75)</td>
<td>418</td>
<td>100.0 (75)</td>
</tr>
<tr>
<td>yprA</td>
<td>95679–97604</td>
<td>642</td>
<td>ding._ecoli</td>
<td>ATP-dependent helicase ( DinG) ( E.\ coli ) (716)</td>
<td>145</td>
<td>33.3 (135)</td>
</tr>
<tr>
<td>yprA (kluD)</td>
<td>98409–97645c</td>
<td>255</td>
<td>s17711</td>
<td>Deoxygluconate oxidoreductase ( E.\ coli ) (253)</td>
<td>659</td>
<td>55.0 (251)</td>
</tr>
<tr>
<td>yprB (kda1)</td>
<td>99238–98411c</td>
<td>276</td>
<td>s17710</td>
<td>Deoxyxuronate isomerase ( E.\ coli ) (278)</td>
<td>753</td>
<td>46.0 (278)</td>
</tr>
<tr>
<td>yprC (kdgR)</td>
<td>99460–100479</td>
<td>340</td>
<td>s15318</td>
<td>( ap-4) protein ( B.\ subtilis ) (334)</td>
<td>425</td>
<td>27.8 (334)</td>
</tr>
<tr>
<td>yprD (kdgK)</td>
<td>100517–101491</td>
<td>325</td>
<td>jq0782</td>
<td>Fructokinase ( V.\ alginolyticus ) (307)</td>
<td>251</td>
<td>27.5 (287)</td>
</tr>
<tr>
<td>yprE (kdgA)</td>
<td>101493–102083</td>
<td>197</td>
<td>adecoq</td>
<td>2-Keto-4-hydroxyglutarate aldolase ( E.\ coli ) (213)</td>
<td>279</td>
<td>36.4 (165)</td>
</tr>
<tr>
<td>yprF (kdgT)</td>
<td>102128–103120</td>
<td>331</td>
<td>jq0113</td>
<td>2-Keto-4-deoxygluconate permease ( E.\ coli ) (398)</td>
<td>894</td>
<td>51.4 (315)</td>
</tr>
</tbody>
</table>

* Positions in the YAC15-6B insert.
† The number of amino acids over which the match percentage was determined is shown in parentheses.
‡ GenBank entry, homology detected by BLAST search.
§ SwissProt entry.
A. SOROKIN and OTHERS

very high similarity of its product to the structure of the enzyme from Bacillus sp. strongly suggested its function.

ypnA (asnS). The ypnA gene product is highly homologous to asparaginyl-tRNA-synthetase but its function needs to be experimentally tested. In the cases of tyrosyl- and threonine-tRNA-synthetase two active genes were found, having concerted regulation and complementing each other’s activity (Glaser et al., 1990; Putzer et al., 1992; Grundy & Henkin, 1993).

ypgB. The protein encoded by this gene is homologous to the histidinyl domain of phosphotransferase system enzyme II of Salmonella typhimurium. According to the nomenclature suggested by Saier & Reizer (1992), this domain is the first phosphorylation site of the IIA domain. This finding suggests similarities between the B. subtilis system and the glucose permease of E. coli or the sucrose permease of S. typhimurium.

ypA. Substantial homology of part of the YprA protein is homologous to asparaginyl-tRNA-synthetase but its function needs to be experimentally tested. In the cases of tyrosyl- and threonine-tRNA-synthetase two active genes were found, having concerted regulation and complementing each other’s activity (Glaser et al., 1990; Putzer et al., 1992; Grundy & Henkin, 1993).

ypB. The protein encoded by this gene is homologous to asparaginyl-tRNA-synthetase but its function needs to be experimentally tested. In the cases of tyrosyl- and threonine-tRNA-synthetase two active genes were found, having concerted regulation and complementing each other’s activity (Glaser et al., 1990; Putzer et al., 1992; Grundy & Henkin, 1993).

ypF (kdgF). Several genes involved in pectin catalysis were isolated and sequenced from Erwinia chrysanthemi (Condemine & Robert-Baudouy, 1991; Hugouvieux-Cotte-Pattar & Robert-Baudouy, 1994). The characterized genes from E. chrysanthemi also include the regulator of this system, kdgR (Reverchon et al., 1991). The relevant aldolase, the gene of which in Gram-negative bacteria is called kdgA or eda, is one of the two enzymes of the Entner–Doudoroff pathway of gluconate assimilation, which does not play an essential role in Bacillus species (Goldman & Blumenthal, 1963). Our homology search revealed six genes encoding a system for polygalacturonate catalysis, including aldolase and the putative regulator. Interestingly, the product of the regulatory gene has highest homology with the putative catalytic affer of B. subtilis CcpA.

Transcriptional features of the sequenced region

The transcription signals which can be reliably predicted in bacteria from sequencing data are the promoter-independent transcription terminators (Daniels et al., 1992; Glaser et al., 1993). Identification of transcription terminators was shown to be useful for prediction of the transcription map which matches well with the experimentally defined one (Azvedo et al., 1993b). Fig. 2 lists the 36 stem–loop structures which might work as promoter-independent transcription terminators that were detected by analysis of the sequence of the region between the serA gene and the kdg operon. This list also contains terminators from the spoIVA–aroE region, which was sequenced by others and was not discussed above, but is relevant to this study. Experimental studies of transcription of ypfC–ypfD (Sorokin et al., 1995), spoIVA (Roels et al., 1992; Stevens et al., 1992), the trpE regulatory region (Henner & Yanofski, 1993), trpC (Sun et al., 1992; Stevens et al., 1992; Yu et al., 1995), cotD (Zheng & Losick, 1990) and rnpB (Reich et al., 1986) have already been reported. We can therefore consider the corresponding terminators as experimentally proven to be active. Some features of the promoter regions deduced from the sequence and which are potentially relevant to the regulation are discussed below.

ORFs ypbC and yplA encode proteins that might be involved in the SOS response (see above). This implies that the corresponding promoters belong to the SOB regulon, which is characterized by the presence of the GAAC-N,-GTTC structure in the promoter regions (Cheo et al., 1991). Promoter structures and the SOB consensus sequence can easily be detected upstream of ypbC and yplA (Fig. 3a, b). Although not so evident, a similar sequence exists also in the yprA leader region (Fig. 3c).

Surprisingly, the ypfD–birA operon encodes several genes having apparently different functions (see above, and Bowser et al., 1995). This might be due to the fact that the biotin biosynthesis regulator, BirA, is a negative regulator of biotin biosynthesis and at the same time a positive regulator of the ypfD–birA operon. Thus, the biotin regulator might work as a switch between carboxylating enzymes containing biotin and lysine biosynthesis through tetrahydropicolinate. This is probably needed to regulate the concentration of oxaloacetate in the cells.
The two operons in the \(\text{kdgRKAT-kdnDI}\) divergon resemble the classical \(\text{lac}\) operon case from Gram-negative bacteria (Jacob & Monod, 1961). The putative regulator \(\text{KdgR}\) is highly homologous to the CcpA catabolite regulator of \(B. \text{subtilis}\) and therefore belongs to the LacI family of regulators. Preliminary data indicated that \(\text{KdgR}\) and CepA are both negative regulators of transcription of the \(\text{kdgRKAT}\) operon (P. Pujic, personal communication). It would be interesting to identify the structural features which specify the difference in the interaction of \(\text{KdgR}\) and CepA with the \(\text{Kdg}\) and other regulatory regions.

**Conclusions**

We have described the application of a sequencing strategy of a YAC insert by direct subcloning into M13 and by using random and directed selection of the templates to be sequenced. Since the \(B. \text{subtilis}\) genome contains sequences which are difficult or impossible to clone in \(E. \text{coli}\), the direct sequencing of PCR products appears to be the most efficient directed sequencing approach. An optimal practical strategy can be proposed: (i) random sequencing, which becomes ineffective when the redundancy of accumulated raw sequence data approaches 3-4; and (ii) directed sequencing of PCR products corresponding to the gaps between contigs, ordered by combinatorial PCR strategies. This second step allows us to overcome the bias of the M13 libraries due to the unclonability of some genome fragments in bacteriophage M13 (Glas et al., 1993; Sorokin et al., 1993; Ogasawara et al., 1994). We do not see any obvious arguments against the application of this approach to the sequencing of a whole small genome (less than 10 Mb). Long-target PCR (Barnes, 1994; Cheng et al., 1994) greatly extends the possibilities for the use of combinatorial PCR. The feasibility of such an approach for the sequencing of a complete bacterial genome is supported by the data presented on YAC15-6B and by the sequencing of the \(B. \text{subtilis}\) genome regions contained in the YAC10-9 and YAC15-132 clones of the same collection, which is now being carried out in this laboratory. Our results also demonstrate that YAC-based collections of large genomes, for example the human genome, can be used directly for sequencing of regions of interest.

**ACKNOWLEDGEMENTS**

We thank C. Anagnostopoulos for his continuous interest and advice. We also thank P. Pujic for useful discussions and communicating unpublished data. The advice of Dr L. Bourgueleret and D. Catena (CEPH, Paris) on using the \(x\)-hap program and of Ph. Bessier, F. Samson and V. Biaudet on computing are highly appreciated. The part-time participation of students F. Bourguoin, G. Taylor and S. Darreau in the YAC15-6B sequencing cannot be overstated. V.A. was a recipient of a Brazilian government fellowship. This work was supported, in part, by grants from GREG (décision 21) and EEC (BIO2-CT93-0272).

**Fig. 4.** Regulatory region of the \(\text{ypjD}\) gene. Structures resembling \(\alpha\) promoter regions (TGAGCA and TATAAT) and homologous to the apparent BirA-operator site of \(B. \text{sphaericus}\) (AATGTGTAACTTAAAACATATAGTTGTTAAC) are shown in capital letters. Numbering as for Fig. 1.
REFERENCES


Barnes, W. M. (1994). PCR amplification of up to 35-kb DNA with high fidelity and high yield from *E. coli* bacteriophage templates. *Proc Natl Acad Sci USA* 91, 2216–2220.


Kimura, K. & Goto, T. (1977). Dihydropicolinate reductases from...


A. SOROKIN and OTHERS


Received 16 January 1996; revised 29 February 1996; accepted 5 March 1996.