Integrated mapping and sequencing of a 115 kb DNA fragment from Bacillus subtilis: sequence analysis of a 21 kb segment containing the sigL locus

C. Fabret, Y. Quentin, N. Chapal, A. Guiseppi, J. Haiech and F. Denizot

Author for correspondence: François Denizot. Tel: +33 91 16 43 87. Fax: +33 91 71 89 14. e-mail: denizot@lcb.cnrs-mrs.fr

INTRODUCTION

Recent improvements in DNA sequencing technology have paved the way for large-scale sequencing projects, such as whole genome analyses (Fleischmann et al., 1995; Fraser et al., 1995). The latter have generally followed the ‘shotgun’ or random sequencing approach. The strict application of this method, most convenient for obtaining a first coverage of the DNA fragment to be sequenced, leads to considerable cost and time investment at later stages. Indeed, as the coverage increases, redundancy increases and the chances of gathering new data and filling the gaps decreases (Edwards & Caskey, 1991). Therefore, completing a partial sequence obtained by random sequencing with a directed approach such as chromosomal walking or PCR is likely to be more efficient. The concept of an integrated sequencing approach was presented in a preliminary way previously (Fabret et al., 1995).

A sequence strategy which combines a low redundancy shotgun approach and directed sequencing has been elaborated. Essentially, the sequences, as well as the size of the fragments utilized for a low coverage shotgun approach, were exploited for the construction of a physical map of the region to be sequenced. The latter considerably simplified the subsequent directed sequencing steps. We report the physical mapping of a 115 kb segment which covers nearly 100 kb of the hisA-cysB region of the Bacillus subtilis chromosome and contains previously sequenced genes sigL and sacB. Sequencing and analysis of a 21305 bp segment, which includes the sigL locus, revealed 21 ORFs, apparently belonging to at least seven transcription units. This segment has a G+C content greater than 47%, compared to 43% characteristic of the flanking regions, and mainly consists of genes whose products seem to be involved in the synthesis of an exopolysaccharide. These observations leave open the possibility that the analysed fragment has been acquired through horizontal transfer.

Keywords: Bacillus subtilis, genome sequencing, sigL, sacB, exopolysaccharide synthesis

METHODS

Bacterial strains and yeast artificial chromosome (YAC). The B. subtilis chromosomal region to be sequenced extended from hisA to cysB and was almost entirely covered by the YAC 11521 clone (Azevedo et al., 1993). Analyses by Southern blotting of the reference B. subtilis 168 repC2 strain and YAC 11521 revealed identical hybridization patterns (data not shown), suggesting the integrity of the DNA fragment cloned into YAC 11521.

YAC DNA purification. Plugs of yeast recombinant cells were

INTRODUCTION

Recent improvements in DNA sequencing technology have paved the way for large-scale sequencing projects, such as whole genome analyses (Fleischmann et al., 1995; Fraser et al., 1995). The latter have generally followed the ‘shotgun’ or random sequencing approach. The strict application of this method, most convenient for obtaining a first coverage of the DNA fragment to be sequenced, leads to considerable cost and time investment at later stages. Indeed, as the coverage increases, redundancy increases and the chances of gathering new data and filling the gaps decreases (Edwards & Caskey, 1991). Therefore, completing a partial sequence obtained by random sequencing with a directed approach such as chromosomal walking or PCR is likely to be more efficient. The concept of an integrated sequencing approach was presented in a preliminary way previously (Fabret et al., 1995).

A sequence strategy which combines a low redundancy shotgun approach and directed sequencing has been elaborated. Essentially, the sequences, as well as the size of the fragments utilized for a low coverage shotgun approach, were exploited for the construction of a physical map of the region to be sequenced. The latter considerably simplified the subsequent directed sequencing steps. We report the physical mapping of a 115 kb segment which covers nearly 100 kb of the hisA-cysB region of the Bacillus subtilis chromosome and contains previously sequenced genes sigL and sacB. Sequencing and analysis of a 21305 bp segment, which includes the sigL locus, revealed 21 ORFs, apparently belonging to at least seven transcription units. This segment has a G+C content greater than 47%, compared to 43% characteristic of the flanking regions, and mainly consists of genes whose products seem to be involved in the synthesis of an exopolysaccharide. These observations leave open the possibility that the analysed fragment has been acquired through horizontal transfer.

Keywords: Bacillus subtilis, genome sequencing, sigL, sacB, exopolysaccharide synthesis

METHODS

Bacterial strains and yeast artificial chromosome (YAC). The B. subtilis chromosomal region to be sequenced extended from hisA to cysB and was almost entirely covered by the YAC 11521 clone (Azevedo et al., 1993). Analyses by Southern blotting of the reference B. subtilis 168 repC2 strain and YAC 11521 revealed identical hybridization patterns (data not shown), suggesting the integrity of the DNA fragment cloned into YAC 11521.

YAC DNA purification. Plugs of yeast recombinant cells were
obtained as described by Schwartz & Cantor (1984). PFGE, performed with a Bio-Rad CHEF-DR II apparatus in 0.5 x TBE buffer (45 mM Tris/Borate, 1 mM EDTA, pH 8.3) at 14 °C and a constant voltage of 200 V, consisted of 30 s pulses for 14 h followed by 60 s pulses for 6 h. After staining with ethidium bromide, the well-separated band corresponding to the YAC was excised, transferred in a dialysis bag and submitted to electrophoresis in 1 x TBE buffer at a constant voltage of 150 V for 3 h. After a 30 s change in current direction, the DNA-containing supernatant was recovered from the bag and concentrated with 2-butanol. YAC DNA was extracted twice with phenol/chloroform/isooamyl alcohol (25:24:1, by vol.), ethanol-precipitated and resuspended in TE buffer (10 mM Tris, 1 mM EDTA, pH 7.4); DNA concentration was determined by A260. The mean recovery was 10 ng DNA per initial plug.

**Shotgun library construction.** YAC DNA, solubilized in 20 µl distilled water and sonicated for 20 s at maximum power with a Heat System Ultrasonics W-375 cup horn, yielded fragments with a mean size of 1-15 kb. End repair of DNA fragments was done in one step, i.e. by simultaneous addition of Klenow enzyme [50 units (µg DNA)] to T4 DNA polymerase [50 units (µg DNA)] (Fitzgerald et al., 1992). DNA was size-fractionated by agarose electrophoresis. Two size classes of fragments ranging from 1 to 3 kb and 3 to 6 kb, respectively, were separated and purified with Geneclean II (Bio101). Two corresponding libraries were built: a major and a minor one consisting of small and large class fragments, respectively. Fragments from each size class were ligated overnight (16 °C, molar ratio 20 - 1) to Smal-linearized and dephosphorylated pUC18 vector (Pharmacia). After 30 min dialysis against water on a Millipore filter (0.025 µm), the ligation mixture (20 - 50 µg DNA ml⁻¹) was used to transform *Escherichia coli* DH5α by electroporation (Bio-Rad apparatus, 25 µF, 2.5 kV and 200 Ω). Bacteria were plated on media containing 50 µg ampicillin ml⁻¹, 100 µg IPTG ml⁻¹ and 40 µg X-Gal ml⁻¹. White recombinant clones were selected on LB medium plates and screened with a β²P-labelled total *B. subtilis* genomic DNA probe. Positive clones were organized on microtiter plates and frozen (Gergen et al., 1979), as well as plated on square (10 x 10 cm) Petri dishes with the same pattern (12 x 8 clones). Thus, when required, either the microtiter tray or the Petri dishes could be easily duplicated with a 96-tooth comb.

**Plasmid preparation and sequencing.** Random sequencing was done exclusively using radiolabeled DNA. Plasmid preparation and radioactive sequencing reactions were as described by Fabret et al. (1995). To check plasmid integrity and to deduce the approximate size of inserts, plasmids were cut with *EcoR1* and *HindIII* and analysed by electrophoresis on 1% agarose gels. All sequencing reactions were done using the universal (U, GTAAGAAGCGGGCCGAT) and the reverse (R, AACAGCTATGACCATG) primers from Pharmacia. When specific primers were designed, either plasmid or PCR-amplified *B. subtilis* DNA was used (see below) and the sequencing was performed with a Model 373A DNA Sequencer (Perkin-Elmer) and fluorescent dyes (Dye Terminator Cycle Sequencing Kit with AmpliTaq DNA Polymerase, Perkin Elmer).

**PCR reactions.** Normal PCR reactions were performed in a volume of 20 µl, containing 1 ng *B. subtilis* DNA, 12 pmol each primer (*Tm* > 62 °C, G+C content > 50%), 0.1 mM dNTP, 0.2 units Goldstar Taq polymerase (Eurogentec) and recommended buffer containing 2 mM Mg²⁺. The reaction was as follows: 2 min at 94 °C, 30 cycles of 30 s at 94 °C, 30 s at 63 °C, 1 min at 72 °C, and 5 min at 72 °C. Long Range (LR) PCR was performed as indicated by the manufacturer (Perkin Elmer and Boehringer kits) using 50 ng *B. subtilis* DNA.

**DNA analysis.** DNA was extracted from cells growing in mid to late exponential phase in Luria broth (LB) or CGC medium (Msadek et al., 1990). Fifty millilitres of culture was centrifuged and the pellet resuspended in 0.2 ml 12.5 mM Tris/HCl, pH 7.5, containing 5 mM EDTA, 10% (w/v) glucose and 20 mg lysosome ml⁻¹, supplemented with 3 ml extraction buffer (50 mM Tris/HCl, pH 6.8, 2 mM EDTA, 1%, w/v, SDS) and incubated for 2 min at 100 °C. Following addition of 0.14 ml 3 M sodium acetate, pH 5.2, the solution was extracted three times with 2 ml acid phenol (pH 4.3, Amresco) at 65 °C for 10 min. RNA was precipitated, solubilized in 50 µl formamide (Clontech) and stored at -20 °C.

Electrophoresis of RNA through formaldehyde gels and RNA transfer (Hybond-N+ membrane from Amersham) were done under standard conditions (Sambrook et al., 1989). Previously sequenced pUC clone inserts were prepared by PCR, labelled with [α-32P]dCTP (Prime-a-Gene kit, Promega) and used as probes (2 x 10⁶ c.p.m. ml⁻¹). Pre-hybridization and hybridization were done at 65 °C in sodium phosphate buffer (0.5 M, pH 7.2) containing 7% SDS and 200 µg herring sperm DNA ml⁻¹. Blots were washed at 65 °C, once with 2 X SSC, 1% SDS, and twice in 0.5 X SSC, 1% SDS.

**Data handling and computer analysis.** DNA sequence handling and assembly were performed with the GAP program of the Staden Package (Staden, 1982). Contig ordering was done with a new computer program developed for this purpose by Alain Guénoche and Jean-Marie Bienfait (Laboratoire d’Informatique de Marseille, unpublished).

The probability of coding was estimated with the GENMARK program with optimal parameters: matrix order, 4; window length, 96 bp; step length, 12 bp (Borodovsky et al., 1994). Sequence motifs such as Shine–Dalgarno and σ consensus patterns were searched with the ANREP program (Mehlau & Myers, 1993). Frame-shift error detection was done with the RISE program with optimal parameters: tuple length, 6 bp; window length, 60 bp; step length, 6 bp (Fichant & Quentin, 1995). The prediction of σ-independent transcription terminators was achieved with the program proposed by d’Aubenton Carafa et al. (1990). Putative ORFs were compared to data banks using the BLAST package (Altschul et al., 1990). Sequences and information related to our project were managed with the ACEDB database manager (Durbin & Thierry-Mieg, 1991).

**RESULTS**

The rationale of the approach

The mean redundancy (r) of a fully or partly sequenced DNA segment is defined as:

\[ r = \frac{n}{L} \tag{1} \]

where L is the length of the relevant DNA segment, s the mean gel reading length and n the number of sequences. Reported sequences, obtained by a shotgun approach, are characterized by mean redundancies of above 6. To reduce this relatively high redundancy, we have devised a method combining a shotgun and a directed sequencing approach.

Assuming a random distribution of sequenced fragments, theoretical considerations (Clarke & Carbon, 1976; Lander & Waterman, 1988; Edwards & Caskey, 1991) allow the calculation of the number of contigs:

\[ n_c = \frac{r^{-(1-T/s)} - 1}{r^{-(1-T/s)} - 1} \tag{2} \]
The **sigL** region of the *B. subtilis* chromosome

![Fig. 1. Physical map of the 19 scaffolds belonging to the 115 kb DNA fragment. Positions of **sigL**, **sacB**, the NotI and SfiI restriction sites and several ORFs with high similarities to known genes, are indicated. The vertical bar separates the domains sequenced by the D. Karamata/C. Mauel (Lausanne, Switzerland) group and the J. Haiech/F. Denizot group (this group). The height of the empty boxes (scale on the right) indicates the mean number of sequences per kb for a given scaffold.](image)

as well as of the fraction \( f_L \) of the sequence covered by thus obtained contigs:

\[
f_L = 1 - \left(1 - \frac{s}{L}\right)^n
\]

where \( T \) is the minimum number of overlapping base pairs that defines two contiguous sequences.

From expression (1), it follows that obtaining a mean redundancy \( r = 2 \) of a segment \( L = 115000 \) bp (see below) with a mean gel reading length \( s = 250 \) bp requires \( n = 950 \) sequence readings (i.e. 425 clones from which inserts are sequenced at each extremity). Fixing \( T \) to 15 bp and introducing these parameters in expressions (2) and (3), we calculate that the 950 sequences will yield 123 contigs and cover about 86\% of the 115 kb sequence. Completion of this partial sequence by directed sequencing would be most uneconomical and time consuming. For comparison, shotgun approaches with redundancies of 4 and 6 should yield only 44 and 10 contigs, respectively. Thus, to facilitate the following direct sequencing steps, the contigs were positioned relative to each other on the basis of the size of any given cloned fragment whose two extremities belong to different contigs. With the parameters obtained above, computer simulations (A. Guénoche, unpublished) suggest that a near total ordering of the contigs could be achieved with inserts ranging from 1 to 4 kb. When total ordering is not reached, independent scaffolds, the name given to groups of linked contigs by Roach *et al.* (1995), can be ordered in a complete physical map with LR PCR.

**Random library construction and sequencing of the clones**

The shotgun library of the 115 kb DNA fragment for random sequencing was constructed according to parameters defined above. Clones were obtained with a minimal amount (1-5 \( \mu \)g) of YAC DNA. To maximize the randomness, DNA fragments were generated by mechanical shearing and cloned according to the procedure described in Methods. Out of the 2400 white colonies, 1300 were eliminated following screening with a total *B. subtilis* genomic probe. None of the remaining 1100 clones had an insert exceeding 3 kb: 490 contained a 1--3 kb insert, 140 had an insert of < 400 bp (they were not retained for further analysis), while the others contained rearranged DNA (see Discussion).

Plasmids, purified and sequenced with the U and R primers, gave rise to over 900 sequences which were assembled into 260 contigs. Comparison of the latter with sequences contained in the database revealed that 30 and 70 of these contigs corresponded to the pYAC4 vector and yeast sequences, respectively. In conclusion, it appears that about 160 contigs are likely to correspond to *B. subtilis* DNA, i.e. more than the calculated figure of 123 (see above), suggesting a cloning bias (see below).

To establish the connections between contigs, the size of each of the retained 490 clones was determined. A search with a computer program developed by J.-M. Bienfait (unpublished) for sequences of the two extremities of a clone belonging each to a different contig, allowed the ordering and the grouping of the latter into 19 sets of physically linked contigs or scaffolds, ranging in size from 2 to 20 kb. That all of the contigs were not ordered is not surprising, since cloned inserts did not exceed 3 kb.

**Ordering of the scaffolds by PCR**

To order the scaffolds, primers were obtained for the extremities of each of them. Starting from a primer of a given scaffold, LR PCR was performed with each of the primers of the other scaffolds. A positive response, together with the size of the PCR product, allows orientation and positioning of the scaffolds relative to each other. As amplified products of over 15 kb were routinely obtained, a single PCR frequently led to the positioning of more than two scaffolds. With the extremities of thus obtained scaffolds, the procedure was repeated until the 19 original scaffolds were ordered on a physical map of the 115 kb region (Fig. 1).
A preliminary search for restriction sites and known \textit{B. subtilis} genes allowed us to (i) position the \textit{SfiI} and the relatively frequent \textit{NotI} restriction sites and thus complete the Itaya \& Tanaka (1991) map, and (ii) identify the previously sequenced loci \textit{sacB} (Steinmetz et al., 1985) and \textit{sigL} (Debarbouillé \textit{et al.}, 1991) (Fig. 1). An overall inspection of the 115 kb region reveals a 20 kb fragment denoted S1 (Fig. 1), for which the mean number of readings is 3.4, i.e. a significantly higher figure than the theoretical mean redundancy of 2 (see above). Screening the S1 fragment with a 50 bases window revealed that the mean number of readings, within the window, varies between 3 and 15, i.e. fragment S1 is over-represented.

### Identification of putative coding regions and transcriptional units of fragment S1

The apparent cloning bias of fragment S1 prompted us to complete its sequence. Appropriate oligonucleotides were synthesized and the sequence on both strands determined directly on relevant PCR products. The computer analysis of this 21305 bp sequence revealed 20 complete ORFs and 1 truncated ORF (Fig. 2), named according to the convention of the \textit{B. subtilis} genome sequencing consortium. Only four of these ORFs (\textit{pnbA, yveJ, yveG} and \textit{yvfH}) have an orientation opposite to that of DNA replication. In 18 cases, the initiation codon is ATG, while TTG or GTG each occurs once. The low frequency of the latter is in agreement with published \textit{B. subtilis} sequences (Glaser \textit{et al.}, 1993). At the expected distance, each start codon is preceded by a Shine–Dalgarno sequence (Table 1). A few bases overlap of several contiguous ORFs suggests their translational coupling.

A search for −35 and −10 promoter motifs suggested the presence of at least seven transcriptional units. Two contiguous $\sigma^{A}$-like promoter sequences in an opposite orientation, found between genes \textit{yveJ} and \textit{yveK}, revealed two divergently transcribed operons which might form a divergon (Fig. 2). That \textit{yveJ} and \textit{pnbA}, and \textit{yveK} and \textit{yveL} are likely to each represent an operon has been established by Northern blotting (Fig. 2). A large non-coding region, separating ORFs \textit{yveL} and \textit{yvfA}, seems to encompass one $\sigma^{A}$-like promoter sequence from which an operon consisting of nine ORFs, \textit{yveM}–\textit{yvfA}, might be transcribed. Interestingly, the presence at the level of \textit{yveQ} of a putative \textit{sigA}-dependent promoter and a $\rho$-independent terminator suggests a possible transcriptional modulation of the downstream genes. \textit{yvfA} is followed by a \textit{ρ}-independent terminator. The downstream five ORFs, \textit{yvfB}–\textit{yvfF}, seem to form a \textit{sigA}-controlled operon, with a $\rho$-independent terminator. So far, no regulatory elements were identified in the 138 bp non-coding region located between \textit{yvfE} and \textit{yvfF}. The latter gene is followed by a short ORF of 219 bp, transcribed in the opposite orientation from a putative $\sigma^{A}$ promoter. ORFs \textit{yvfL} and \textit{yvfH}, separated by a putative $\rho$-independent terminator, are transcribed in opposite directions. Northern blot analyses confirmed that \textit{sigL} and \textit{yvfH} are monocistronic operons (Fig. 2).

Finally, under our experimental conditions, no RNA
The sigL region of the B. subtilis chromosome

![Diagram of ORF organization in the 51 segment](image)

**Fig. 2.** ORF organization in the 51 segment (see Fig. 1). ORFs are shown by arrows indicating their direction of transcription and translation. DNA replication proceeds from right to left. The letters below each arrow represent the name of the ORF (only the last two letters are indicated, **y** being common to all of them except sigL and pnbA). NotI restriction sites, putative promoters (r) and p-independent terminators (o) are indicated. All indicated promoters are recognized by sigA except **yveM**, which is likely to be recognized by sigC (*). Wavy arrows indicate transcriptional units detected by Northern blotting.

**Table 2.** Comparison of the ORF products to proteins in databases

Sets of genes separated by horizontal lines correspond to putative transcription units.

<table>
<thead>
<tr>
<th>ORF*</th>
<th>Similarity (BLAST score &gt; 90)$^\dagger$</th>
<th>Putative function</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>pnbA</em></td>
<td>PNBA_BACSU</td>
<td>Paranitrobenzyl esterase (100% identity)</td>
<td>Zock <em>et al.</em> (1994)</td>
</tr>
<tr>
<td><em>yveJ</em></td>
<td>SINR_BACSU</td>
<td>Transcriptional regulator</td>
<td>Mandic-Mulec <em>et al.</em> (1995)</td>
</tr>
<tr>
<td><em>yveK</em></td>
<td>CAPA_STAAU</td>
<td>Exopolysaccharide synthesis</td>
<td>Lin <em>et al.</em> (1994)</td>
</tr>
<tr>
<td><em>yveL</em></td>
<td>CAPB_STAAU</td>
<td>Exopolysaccharide synthesis</td>
<td>Lin <em>et al.</em> (1994)</td>
</tr>
<tr>
<td><em>yveM</em></td>
<td>CAPD_STAAU</td>
<td>Exopolysaccharide synthesis</td>
<td>Lin <em>et al.</em> (1994)</td>
</tr>
<tr>
<td><em>yveN</em></td>
<td>RFAG_ECOLI</td>
<td>Exopolysaccharide synthesis</td>
<td>Parkez <em>et al.</em> (1992)</td>
</tr>
<tr>
<td><em>yveO</em></td>
<td>VIPC_SALT</td>
<td>Exopolysaccharide synthesis</td>
<td>Hashimoto <em>et al.</em> (1993)</td>
</tr>
<tr>
<td><em>yveP</em></td>
<td>VIPC_SALT</td>
<td>Exopolysaccharide synthesis</td>
<td>Hashimoto <em>et al.</em> (1993)</td>
</tr>
<tr>
<td><em>yveR</em></td>
<td>EXOO_RHIME</td>
<td>Exopolysaccharide synthesis</td>
<td>Becker <em>et al.</em> (1993)</td>
</tr>
<tr>
<td><em>yveT</em></td>
<td>GGAB_BACSU</td>
<td>Exopolysaccharide synthesis</td>
<td>Lazarevic <em>et al.</em> (1995)</td>
</tr>
<tr>
<td><em>yveB</em></td>
<td>?</td>
<td>Exopolysaccharide synthesis</td>
<td>Muller <em>et al.</em> (1993)</td>
</tr>
<tr>
<td><em>yveC</em></td>
<td>EXOY_RHIME</td>
<td>Exopolysaccharide synthesis</td>
<td>Rubens <em>et al.</em> (1993)</td>
</tr>
<tr>
<td><em>yveD</em></td>
<td>CPSD_STRAG</td>
<td>Exopolysaccharide synthesis</td>
<td>Lin <em>et al.</em> (1994)</td>
</tr>
<tr>
<td><em>yveE</em></td>
<td>CAPG_STAAU</td>
<td>Exopolysaccharide synthesis</td>
<td>Coleman <em>et al.</em> (1988)</td>
</tr>
<tr>
<td><em>yveF</em></td>
<td>SPSC_BACSU</td>
<td>Exopolysaccharide synthesis</td>
<td>Glaser <em>et al.</em> (1993)</td>
</tr>
<tr>
<td><em>sigL</em></td>
<td>RP54_BACSU</td>
<td>σ factor L (100% identity)</td>
<td>Débarbouillé <em>et al.</em> (1991)</td>
</tr>
<tr>
<td><em>yveH</em></td>
<td>LLDN_ECOLI</td>
<td>Putative lactate permease</td>
<td>Dong <em>et al.</em> (1993)</td>
</tr>
</tbody>
</table>

$^*$ ORF indicated in Fig. 2.

$^\dagger$ Accession name from SwissProt data bank.
transcripts corresponding to the putative yveM–yveA and yvfB–yvfF operons were detected.

Protein sequence analysis

The results of the database screening (Altschul et al., 1990) for sequences homologous to the ORFs identified above are summarized in Table 2, where only BLAST scores > 90 are reported. It appears that pabA, which encodes a paranitrobenzyl esterase (Zock et al., 1994), and sigL, the structural gene of the L σ factor (Debarbouille et al., 1991), have already been sequenced. The high level of similarity (73.7% over 453 aa) between a putative E. coli lactate permease (Dong et al., 1993) and YvfH suggests that the latter might also be a lactate permease. YveJ shares homologies with a B. subtilis transcriptional regulator. Products of ORFs belonging to putative operons yveK–yveL, yveM–yveA and yvfB–yvfF all exhibit relatively high similarities to a set of proteins involved in the synthesis of exopolysaccharides in Staphylococcus aureus, E. coli, Salmonella typhimurium, Rhizobium meliloti, Streptococcus agalactiae, Streptomyces peucetius, B. stearothermophilus and B. subtilis.

DISCUSSION

The sequencing strategy presented above combines a low redundancy shotgun method with a directed method. Its original feature is the exploitation of the size of different clones together with a computer program, which allow the ordering of the contigs into larger entities, called scaffolds, and the establishment of an extensive physical map of the region to be sequenced. The latter provides the necessary data for completing the sequence by the directed approach. According to computer simulations, a complete physical map would require large clones. Nevertheless, should the latter be difficult to obtain (see Results and below), PCR offers a relatively easy way for contig ordering. The feasibility of our strategy has been illustrated on a 115 kb region and we believe that it is applicable to megabase fragments, i.e. to whole bacterial genomes.

Comparison of the sequencing cost of different methods by a formula devised by Burland et al. (1993) reveals that the most expensive is the total directed approach, followed by the totally random approach, the conventional mixed approach, our approach and finally the Janus strategy (Burland et al. 1993). Our approach, the least expensive but one, can be rendered even more efficient. Firstly, the mean reading length can be increased from 250 to 400 bases thanks to automatic fluorescent readers and secondly, the price of synthetic oligonucleotides is steadily decreasing. Thus, the cost of both the random and the directed approach steps can be reduced.

A total of 2400 white colonies were obtained from our libraries, i.e. the 1–3 kb library and the 3–6 kb library. Inspection of a fair sample showed that less than 50% of them contained an insert. This prompted us to screen the white colonies. Hybridization with a B. subtilis total genomic probe revealed that 54% of the clones did not contain any insert or B. subtilis DNA. Another 43% of the remaining clones containing B. subtilis DNA were discarded because of DNA rearrangements. Indeed, the size of the vector, obtained by plasmid cutting with HindIII and EcoRI, was lower than expected. These observations, combined with the apparent absence of inserts larger than 3 kb, is likely to be due to the instability of B. subtilis DNA cloned in E. coli (Glaser et al., 1993, Kunst et al., 1995). Indeed, when a 80 kb human DNA segment cloned in a PI vector was sonicated and subcloned in E. coli using exactly the same procedure, such a high proportion of clones not containing any insert, as well as DNA rearrangements were not observed (J. Haiech & F. Denizot, unpublished).

As mentioned in Results, the 21 kb segment, called S1, was over-represented in the library. Inspection of the contigs in the 115 kb region, as well as the comparison of S1 to subsequently fully sequenced (not presented) flanking regions, revealed that the G+C content of S1 was almost uniformly above 47%, compared to 43% characteristic of the flanking regions and of the B. subtilis genome (Priest, 1993). With an overlapping 5000 bp window which evens local variations, the S1 region exhibits a G+C content greater than 47% all along its length. The cloning bias in favour of this region appears to be well correlated with its relatively high G+C content, close to 51%, characteristic of E. coli (Blake & Earley, 1986). That 17 out of the 21 putative ORFs located on segment S1 seem to encode products involved in the synthesis of an exopolysaccharide(s) leaves open the possibility that S1 has been acquired by horizontal transfer.

ACKNOWLEDGEMENTS

We are grateful to Dr Alain Guénoche and to Jean-Marie Bienfait (Laboratoire d’Infection et de Maladies de Marseille) for their help in computer simulations and for placing their computer program at our disposal. We express our gratitude to Professor Dimitri Karamata for helpful discussions. We also thank Jeaninne Busuttil for her excellent technical support. This work was supported by the Centre National de la Recherche Scientifique (CNRS), by the Ministère de l’Éducation Nationale (MEN), by the Groupement de Recherches et d’Etudes des Génotomes (GREG) and by EC contract no. BIO2-CT94-2011.

REFERENCES


mapping of promoters located on the *exoHKLAMONP* fragment. 


Received 8 May 1996; revised 22 July 1996; accepted 26 July 1996.