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

Laboratoire de Chimie
Bactérienne, 31 Chemin J.
Aiguier, 13402 Marseille
Cédex 20, France

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

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).

Abbreviations: LR PCR, Long Range PCR; YAC, yeast artificial chromosome.

The EMBL accession number for the nucleotide sequence reported in this paper is Z71928.

In this study, we have developed an efficient strategy which combines random and directed methods and optimizes the time of the shift from the former to the latter, the main novelty being the construction of a physical map based on the size of the clones used for shotgun sequencing. Our theoretical considerations were applied to a 115 kb chromosomal fragment sequenced within the *Bacillus subtilis* genome sequencing project. We report the physical map of the whole region and a preliminary analysis of the nucleotide sequence of a 21 kb segment. During the preparation of this manuscript, a theoretical contribution, reaching conclusions almost identical to ours, has been published (Roach *et al.*, 1995).

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 *trpC2* 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 × 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 × 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/isoamyl 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 A_{260} . 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 1–1.5 kb. End repair of DNA fragments was done in one step, i.e. by simultaneous addition of Klenow enzyme [50 units (µg DNA)⁻¹] and T4 DNA polymerase [50 units (µg DNA)⁻¹] (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 *Sma*I-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 transferred onto nylon membranes and screened with a ³²P-labelled total *B. subtilis* genomic DNA probe. Positive clones were ordered on microtitre plates and frozen (Gergen *et al.*, 1979), as well as plated on square (10 × 10 cm) Petri dishes with the same pattern (12 × 8 clones). Thus, when required, either the microtitre tray or the Petri dishes could be easily duplicated with a 96-tooth comb.

Plasmid preparation and sequencing. Random sequencing was done exclusively using radioactive detection. 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 *Eco*RI and *Hind*III and analysed by electrophoresis on 1 % agarose gels. All sequencing reactions were done using the universal (U, GTAAAACGACGGCCAGT) 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 ($T_m > 62$ °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.

RNA analysis. RNA was extracted from cells growing in mid to late exponential phase in Luria broth (LB) or CGCH 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 lysozyme 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 × 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 × SSC, 1 % SDS, and twice in 0.5 × 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 GENEMARK 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 (Mehldau & Myers, 1993). Frame-shift error detection was done with the FSED program with optimal parameters: tuple length, 6 bp; window length, 60 bp; step length, 6 bp (Fichant & Quentin, 1995). The prediction of *p*-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:

$$r = ns/L \quad (1)$$

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:

$$ne^{-r(1-T/s)} - ne^{-2r(1-T/s)} \quad (2)$$

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. Mauël (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.

as well as of the fraction fL of the sequence covered by thus obtained contigs:

$$fL = 1 - (1 - s/L)^n \quad (3)$$

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 µg) 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).

Table 1. General features of the identified ORFs

ORF	Endpoints (nt)	Ribosome-binding site (upper case) and initiation codon (bold)	GENMARK score
<i>pnbA</i>	132 < 1601	AAGGgaGaGaacacaatg	0.82
<i>yveJ</i>	1677 < 2135	gAGGAGaattcatattatg	0.79
<i>yveK</i>	2381 > 3085	AAGGctATGaatacatg	0.79
<i>yveL</i>	3091 > 3774	ggGGAGtgaagcgagtg	0.64
<i>yveM</i>	4033 > 5829	ccGGAGaaGactgtcaatg	0.84
<i>yveN</i>	5841 > 6986	ggGGAGtgaaaatg	0.65
<i>yveO</i>	6983 > 7819	AAGGAGaaaagcgtatg	0.79
<i>yveP</i>	7812 > 8966	AAGGgctTGtcaagcatg	0.59
<i>yveQ</i>	8963 > 10066	AAGGAccataaccgatg	0.52
<i>yveR</i>	10091 > 11125	gAGGAGccGgtcattatg	0.61
<i>yveS</i>	11130 > 12206	cAGagGGTGaggacatg	0.80
<i>yveT</i>	12203 > 13237	tcaGcGGTGaagcgcaatg	0.75
<i>yvfA</i>	13234 > 13563	AAGGcaGTGcgaagcaggcatg	0.52
<i>yvfB</i>	13713 > 14750	cAGGcGGTGcaaatg	0.67
<i>yvfC</i>	14747 > 15355	AAGGAGcTGtgaatcrttg	0.68
<i>yvfD</i>	15352 > 16002	gcGGAGATGtgctcctcatg	0.83
<i>yvfE</i>	16007 > 16912	AAGGATGattctatg	0.74
<i>yvfF</i>	17150 > 18118	gtcGAGGTGaagaaatg	0.76
<i>yvfG</i>	18128 < 18346	AAGGAGcTGccaaagaaatg	0.66
<i>sigL</i>	18425 > 19735	AAaGgGgAGtgagcaagatg	0.81
<i>yvfH*</i>	19762 <		0.73

* Partial ORF.

A preliminary search for restriction sites and known *B. subtilis* genes allowed us to (i) position the *Sfi*I and the relatively frequent *Not*I restriction sites and thus complete the Itaya & Tanaka (1991) map, and (ii) identify the previously sequenced loci *sacB* (Steinmetz *et al.*, 1985) and *sigL* (Débarbouillé *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 21 305 bp sequence revealed 20 complete ORFs and 1 truncated ORF (Fig. 2), named according to the convention of the *B. subtilis* genome sequencing consortium. Only four of these ORFs (*pnbA*, *yveJ*, *yvfG* and *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 *B. subtilis* sequences (Glaser *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 σ^A -like promoter sequences in an opposite orientation, found between genes *yveJ* and *yveK*, revealed two divergently transcribed operons which might form a divergon (Fig. 2). That *yveJ* and *pnbA*, and *yveK* and *yveL* are likely to each represent an operon has been established by Northern blotting (Fig. 2). A large non-coding region, separating ORFs *yveL* and *yveM*, seems to encompass one σ^C -like promoter sequence from which an operon consisting of nine ORFs, *yveM*–*yvfA*, might be transcribed. Interestingly, the presence at the level of *yveQ* of a putative *sigA*-dependent promoter and a p -dependent terminator suggests a possible transcriptional modulation of the downstream genes. *yvfA* is followed by a p -independent terminator. The downstream five ORFs, *yvfB*–*yvfF*, seem to form a *sigA*-controlled operon, with a p -independent terminator. So far, no regulatory elements were identified in the 138 bp non-coding region located between *yvfE* and *yvfF*. The latter gene is followed by a short ORF of 219 bp, transcribed in the opposite orientation from a putative σ^A promoter. ORFs *sigL* and *yvfH*, separated by a putative p -independent terminator, are transcribed in opposite directions. Northern blot analyses confirmed that *sigL* and *yvfH* are monocistronic operons (Fig. 2).

Finally, under our experimental conditions, no RNA

Fig. 2. ORF organization in the S1 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, *yv* being common to all of them except *sigL* and *pnbA*). *NotI* restriction sites, putative promoters (Γ) and p -independent terminators ($\odot$) 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.

ORF*	Similarity (BLAST score > 90)†	Putative function	Reference
<i>pnbA</i>	PNBA_BACSU	Paranitrobenzyl esterase (100 % identity)	Zock <i>et al.</i> (1994)
<i>yveJ</i>	SINR_BACSU	Transcriptional regulator	Mandic-Mulec <i>et al.</i> (1995)
<i>yveK</i>	CAPA_STAAU	Exopolysaccharide synthesis	Lin <i>et al.</i> (1994)
<i>yveL</i>	CAPB_STAAU	Exopolysaccharide synthesis	Lin <i>et al.</i> (1994)
<i>yveM</i>	CAPD_STAAU	Exopolysaccharide synthesis	Lin <i>et al.</i> (1994)
<i>yveN</i>	RFAG_ECOLI	Exopolysaccharide synthesis	Parker <i>et al.</i> (1992)
	VIPC_SALTI		Hashimoto <i>et al.</i> (1993)
<i>yveO</i>	SPSA_BACSU	Exopolysaccharide synthesis	
<i>yveP</i>	VIPC_SALTI	Exopolysaccharide synthesis	Hashimoto <i>et al.</i> (1993)
<i>yveQ</i>	?		
<i>yveR</i>	EXOO_RHIME	Exopolysaccharide synthesis	Becker <i>et al.</i> (1993)
<i>yveS</i>	?		
<i>yveT</i>	GGAB_BACSU	Exopolysaccharide synthesis	Lazarevic <i>et al.</i> (1995)
<i>yvfA</i>	?		
<i>yvfB</i>	?		
<i>yvfC</i>	EXOY_RHIME	Exopolysaccharide synthesis	Muller <i>et al.</i> (1993)
	CPSD_STRAG		Rubens <i>et al.</i> (1993)
<i>yvfD</i>	CAPG_STAAU	Exopolysaccharide synthesis	Lin <i>et al.</i> (1994)
	LXPA_ECOLI		Coleman <i>et al.</i> (1988)
<i>yvfE</i>	SPSC_BACSU	Exopolysaccharide synthesis	Glaser <i>et al.</i> (1993)
	MOSB_RHIME		Murphy <i>et al.</i> (1993)
	DEGT_BACST		Takagi <i>et al.</i> (1990)
	DNRJ_STRPE		Stutzman-Engwall <i>et al.</i> (1992)
<i>yvfF</i>	?		
<i>yvfG</i>	?		
<i>sigL</i>	RP54_BACSU	σ factor L (100 % identity)	Débarbouillé <i>et al.</i> (1991)
<i>yvfH</i>	LLDP_ECOLI	Putative lactate permease	Dong <i>et al.</i> (1993)

* ORF indicated in Fig. 2.

† Accession name from SwissProt data bank.

transcripts corresponding to the putative *yveM-yvfA* 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 *pnbA*, which encodes a paranitrobenzyl esterase (Zock *et al.*, 1994), and *sigL*, the structural gene of the 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-yvfA* 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 P1 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'Informatique 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 Jeannine 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'Études des Génomes (GREG) and by EC contract no. BIO2-CT94-2011.

REFERENCES

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *J Mol Biol* **215**, 403–410.
- d'Aubenton Carafa, Y., Brody, E. & Thermes, C. (1990). Prediction of *p*-independent *Escherichia coli* transcription terminators: a statistical analysis of their RNA stem-loop structures. *J Mol Biol* **216**, 835–858.
- Azevedo, V., Alvarez, E., Zumstein, E., Damiani, G., Sgarbetta, V., Ehrlich, S. D. & Serró, P. (1993). An ordered collection of *Bacillus subtilis* DNA segments cloned in yeast artificial chromosomes. *Proc Natl Acad Sci USA* **90**, 6047–6051.
- Becker, A., Kleickmann, A., Keller, M., Arnold, W. & Puhler, A. (1993). Identification and analysis of the *Rhizobium meliloti* *exoAMONP* genes involved in exopolysaccharide biosynthesis and

- mapping of promoters located on the *exoHKLAMONP* fragment. *Mol Gen Genet* **241**, 367–379.
- Blake, R. D. & Earley, S. (1986). Distribution and evolution of sequence characteristics in the *E. coli* genome. *J Biomol Struct Dyn* **4**, 291–307.
- Borodovsky, M., Rudd, K. E. & Koonin, V. (1994). Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. *Nucleic Acids Res* **22**, 4756–4767.
- Burland, V., Daniels, D. L., Plunkett, G. & Blattner, F. R. (1993). Genome sequencing on both strands: the Janus strategy. *Nucleic Acids Res* **21**, 3385–3390.
- Clarke, L. & Carbon, J. (1976). A colony bank containing synthetic ColEI hybrid plasmids representative of the entire *E. coli* genome. *Cell* **9**, 91–99.
- Coleman, J. & Raetz, C. R. (1988). First committed step of lipid A biosynthesis in *Escherichia coli*: sequence of the *lpxA* gene. *J Bacteriol* **170**, 1268–1274.
- Débarbouillé, M., Martin-Verstaete, I., Kunst, F. & Rapoport, G. (1991). The *Bacillus subtilis sigL* gene encodes an equivalent of sigma 54 from Gram-negative bacteria. *Proc Natl Acad Sci USA* **88**, 9092–9096.
- Dong, J. M., Taylor, J. S., Latour, D. J., Iuchi, S. & Lin, E. C. (1993). Three overlapping *let* genes involved in L-lactate utilization by *Escherichia coli*. *J Bacteriol* **175**, 6671–6678.
- Durbin, R. & Thierry-Mieg, J. (1991). *A C. elegans Database*. Documentation, code and data available from anonymous FTP servers at lirmm.lirmm.fr, cele.mrc-lmb.cam.ac.uk and ncbi.nlm.nih.gov
- Edwards, A. & Caskey, C. T. (1991). Closure strategies for random DNA sequencing. *Methods: Companion Methods Enzymol* **3**, 41–47.
- Fabret, C., Quentin, Y., Guiseppi, A., Busuttil, J., Haiech, J. & Denizot, F. (1995). Analysis of errors in finished DNA sequences: the surfactin operon of *Bacillus subtilis* as an example. *Microbiology* **141**, 345–350.
- Fichant, G. & Quentin, Y. (1995). A frameshift error detection algorithm for DNA sequencing projects. *Nucleic Acids Res* **23**, 2900–2908.
- Fitzgerald, M. C., Skowron, P., Van Etten, J. L., Smith, L. M. & Mead, D. A. (1992). Rapid shotgun cloning utilizing the two base recognition endonuclease *Cvi*J1. *Nucleic Acids Res* **20**, 3753–3762.
- Fleischmann, R. D. and others (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512.
- Fraser, C. M. and others (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403.
- Gergen, J. P., Stern, H. & Wensink, C. (1979). Filter replicas and permanent collections of recombinant DNA plasmids. *Nucleic Acids Res* **7**, 2115–2136.
- Glaser, P., Kunst, F., Arnaud, M., Coudart, M.-P., Gonzales, W., Hullo, M.-F., Ionescu, M., Lubochinsky, B., Marcelino, L., Moszer, I., Presecan, E., Santana, M., Schneider, E., Schweizer, J., Vertes, A., Rapoport, G. & Danchin, A. (1993). *Bacillus subtilis* genome project: cloning and sequencing of the 97 kb region from 325° to 333°. *Mol Microbiol* **10**, 371–384.
- Hashimoto, Y., Li, N., Yokoyama, H. & Ezaki, T. (1993). Complete nucleotide sequence and molecular characterization of *ViaB* region encoding Vi antigen in *Salmonella typhi*. *J Bacteriol* **175**, 4456–4465.
- Itaya, M. & Tanaka, T. (1991). Complete physical map of the *Bacillus subtilis* 168 chromosome constructed by a gene-directed mutagenesis method. *J Mol Biol* **220**, 631–648.
- Kunst, F., Vassarotti, A. & Danchin, A. (1995). Organization of the European *Bacillus subtilis* genome sequencing project. *Microbiology* **141**, 249–255.
- Lander, E. S. & Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**, 231–239.
- Lazarevic, V., Mauël, C., Soldo, B., Freymond, P.-P., Margot, P. & Karamata, D. (1995). Sequence analysis of the 308° to 311° segment of the *Bacillus subtilis* 168 chromosome, a region devoted to cell wall metabolism, containing non-coding grey holes which reveal chromosomal rearrangements. *Microbiology* **141**, 329–335.
- Lin, W. S., Cunneen, T. & Lee, C. Y. (1994). Sequence analysis and molecular characterization of genes required for the biosynthesis of type 1 capsular polysaccharide in *Staphylococcus aureus*. *J Bacteriol* **176**, 7005–7016.
- Mandic-Mulec, I., Doukhan, L. & Smith, I. (1995). The *Bacillus subtilis* SinR protein is a repressor of the key sporulation gene *spo0A*. *J Bacteriol* **177**, 4619–4627.
- Mehldau, G. & Myers, G. (1993). A system for pattern matching applications on biosequences. *Comput Appl Biosci* **9**, 299–314.
- Msadek, T., Kunst, F., Henner, D., Klier, A., Rapoport, G. & Dedonder, R. (1990). Signal transduction pathway controlling synthesis of a class of degradative enzymes in *Bacillus subtilis*: expression of the regulatory genes and analysis of mutations in *degS* and *degU*. *J Bacteriol* **172**, 824–834.
- Muller, P., Keller, M., Weng, W. M., Quandt, J., Arnold, W. & Puhler, A. (1993). Genetic analysis of the *Rhizobium meliloti* *exoYFQ* operon: ExoY is homologous to sugar transferases and ExoQ represents a transmembrane protein. *Mol Plant-Microbe Interact* **6**, 55–65.
- Murphy, P. J., Trenz, S. P., Grzemski, W., De Bruijn, F. J. & Schell, J. (1993). The *Rhizobium meliloti* rhizopine *mos* locus is a mosaic structure facilitating its symbiotic regulation. *J Bacteriol* **175**, 5193–5204.
- Parker, C. T., Pradel, E. & Schnaitman, C. A. (1992). Identification and sequences of the lipopolysaccharide core biosynthetic genes *rfaQ*, *rfaP*, and *rfaG* of *Escherichia coli* K-12. *J Bacteriol* **174**, 930–934.
- Priest, F. G. (1993). Systematics and ecology of *Bacillus*. In *Bacillus subtilis and Other Gram-Positive Bacteria: Biochemistry, Physiology, and Molecular Genetics*, pp. 3–16. Edited by A. L. Sonenshein, J. A. Hoch & R. Losick. Washington, DC: American Society for Microbiology.
- Roach, J. C., Boysen, C., Wang, K. & Hood, L. (1995). Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics* **26**, 345–353.
- Rubens, E., Heggen, L. M., Haft, R. F. & Wessels, M. R. (1993). Identification of *epsD*, a gene essential for type III capsule expression in group B streptococci. *Mol Microbiol* **8**, 843–855.
- Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: a Laboratory Manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.
- Schwartz, D. C. & Cantor, C. R. (1984). Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell* **37**, 67–75.
- Staden, R. (1982). Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. *Nucleic Acids Res* **10**, 4731–4751.
- Steinmetz, M., LeCoq, D., Aymerich, S., Gonzy-Treboul, G. & Gay, P. (1985). The DNA sequence of the gene for secreted *Bacillus subtilis* enzyme levane-sucrase and its genetic control sites. *Mol Gen Genet* **200**, 220–228.
- Stutzman-Engwall, K. J., Otten, S. L. & Hutchinson, C. R. (1992). Regulation of secondary metabolism in *Streptomyces* spp. and

overproduction of daunorubicin in *Streptomyces peucetius*. *J Bacteriol* **174**, 144–154.

Takagi, M., Takada, H. & Imanaka, T. (1990). Nucleotide sequence and cloning in *Bacillus subtilis* of the *Bacillus stearothermophilus* pleiotropic regulatory gene *degT*. *J Bacteriol* **172**, 411–418.

Zock, J., Cantwell, C., Swartling, J., Hodges, R., Pohl, T., Sutton,

K., Rosteck, P., Jr, McGilvray, D. & Queener, S. (1994). The *Bacillus subtilis* *pnbA* gene encoding *p*-nitrobenzyl esterase: cloning, sequence and high-level expression in *Escherichia coli*. *Gene* **151**, 37–43.

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