A 22 kb DNA sequence in the cspB-glpPFKD region at 75° on the Bacillus subtilis chromosome

Noback, Michiel A.; Terpstra, Peter; Holsappel, Siger; Venema, Gerard; Bron, Sierd

Published in:
Default journal

Publication date:
1996

Microbiology (1996), 142, 3021–3026 Printed in Great Britain

A 22 kb DNA sequence in the *cspB–glpPFKD* region at 75° on the *Bacillus subtilis* chromosome

Michiel A. Noback, Peter Terpstra, Siger Holsappel, Gerard Venema and Sierd Bron

Author for correspondence: Sierd Bron. Tel: +31 50 3632105. Fax: +31 50 3632348.
e-mail: S.Bron@biol.rug.nl

Department of Genetics, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Kerklaan 30, 9751NN, Haren, The Netherlands

A 21808 bp nucleotide sequence at 75° on the genetic map of the *Bacillus subtilis* chromosome was determined. The sequence of this region is adjacent to the *glpPFKD* operon involved in glycerol utilization. Twenty-six ORFs were identified, one of which corresponds to the *cspB* gene, encoding a cold-shock protein. Seventeen of the deduced protein sequences of these ORFs displayed significant homology to known proteins in the data banks. One putative operon was identified, consisting of five ORFs, that is probably involved in the uptake and processing of copper. The location of *cspB* in this sequence does not confirm the genetic mapping data, indicating that the gene is closely linked to *comK*, which is located at 80° on the *B. subtilis* chromosome.

Keywords: *Bacillus subtilis*, genome sequencing, glycerol operon, *cspB*

INTRODUCTION

In the framework of the European *Bacillus subtilis* genome sequencing project, our group is responsible for sequencing the region between 74° (*glyB*) and 86° (*addAB*) on the genetic map of *B. subtilis*. This paper deals with the cloning and sequence analysis of the region located counter-clockwise from the *glpPFKD* (glycerol) operon, which has been mapped at 75° on the genetic map (Anagnostopoulos *et al.*, 1993).

METHODS

Bacterial strains and DNA handling procedures. *B. subtilis* 168 (*trpC2*) was used as the standard strain for sequence determinations. DNA fragments for sequencing were obtained by Long Range PCR (LR PCR), or inverse LR PCR techniques using the Gene Amp XL-PCR kit with r*Tth* polymerase (Perkin Elmer). All amplification reactions were performed according to the protocols supplied by the manufacturer. Inverse LR PCR was performed by digestion of *B. subtilis* chromosomal DNA with appropriate restriction enzymes, purification of the digested DNA and subsequent self-ligation at low concentrations of DNA (< 5 μg ml^{-1}). PCR primers used are listed in Table 1 and an overview of the amplified fragments is given in Fig. 1.

Abbreviations: LR PCR, Long Range PCR; RBS, ribosome binding site.

The EMBL accession number for the nucleotide sequence reported in this paper is X96983.

PCR fragments were treated in the following way to obtain a shotgun bank of randomly overlapping clones in phage M13. Amplified DNA fragments were sheared by nebulization under nitrogen gas pressure using a DNA Nebulizer (GATC GmbH), according to the instructions of the supplier. The sheared DNA was treated with Klenow enzyme (Boehringer Mannheim) in the presence of a mixture of the four dNTPs. The DNA mixture was fractionated according to size by agarose gel electrophoresis and segments in the 500–1000 bp range were extracted from the agarose using the JETsorb DNA extraction kit (GENOMED GmbH). The DNA fraction obtained was treated with T4 DNA polymerase and dNTPs (Boehringer Mannheim) to obtain blunt-ended fragments. This DNA mixture was ligated into the M13mp18 phage vector, which had been digested with *Sma*I and treated with alkaline phosphatase (Boehringer Mannheim), and the ligation mixture was used to transform *Escherichia coli* strain XL1-Blue (*supE44 lac hsdR17 recA1* [F′ *proAB lacI*^{q} *lacZ*ΔM15]).

Sequence determination. Single-stranded DNA from recombinant M13 phage was isolated on the Vistra DNA Labstation 625 supplied by Amersham using the automated M13 template preparation kit. The DNA inserts were sequenced by the dideoxy chain-termination method (Sanger *et al.*, 1977) using the Amersham automated Δ*Taq* cycle sequencing kit and the Vistra DNA sequencer 725. The universal forward sequencing primer was used (5′ GTAAAACGACGGCCAGT 3′). Remaining gaps, comprising about 5–10% of the total sequence, between the contiguous sequences obtained from sequencing M13 clones were filled by primer walking on PCR material using the Amersham Sequenase PCR product sequencing kit and [^{35}S]dATP.

Data handling and computer analysis. DNA sequences were assembled using the Staden package (MRC) on a SUN

Table 1. Sequences of the primers used for LR PCR

Primer	Sequence (5′–3′)
SH25	CGG TAT ATA TCT GGC GGA GCT GCA T
xlp02	CTA GTG ATC GCA GGC TAT GGA GGC T
xlp03	GCA GGT CGT CAG AAT CAG CTC TTC C
xlp09	GTC ATA TTC GGC TCT AGC TTC CTG C
xlp11	CTG TTC CAT ATC CTG CGC ATC AAG
xlp12	GAA GCC TTC GCC TTG AAT AGC AGA G
xlp13	TGC CAT CCA CAT ACT GAG TCG AGT C

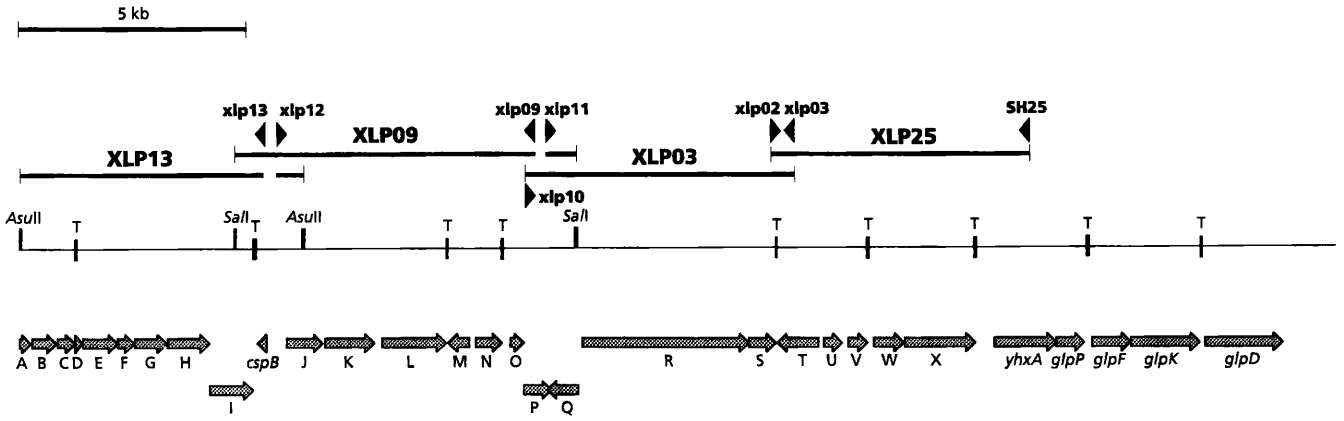

Fig. 1. Map of the 28 kb region around 75° on the *B. subtilis* chromosome. Only relevant restriction sites are indicated. The upper part of this map shows the location of the LR PCR fragments and the primers that were used. Filled arrowheads indicate the positions and orientations of primers. Stippled arrows indicate the positions and orientations of the putative genes. The letters below the arrows correspond to the last letter of the names used for the designation of ORFs referred to in Table 3. The letters T above the map indicate the presence of transcription terminator-like sequences.

SPARCstation LX workstation. A redundancy of four readings per base with a minimum of one reading for each strand was taken as the criterion for a reliable sequence. The compiled sequence was analysed for the presence of ORFs consisting of more than 50 codons using the Staden package. The amino acid sequences of the putative proteins encoded by the ORFs were analysed for similarity to known sequences in data banks using the FASTA program and BLAST e-mail server at the NCBI (retrieve@ncbi.nlm.nih.gov).

Transformation and competence. *B. subtilis* cells were made competent essentially as described by Bron & Venema (1972). *E. coli* cells were made competent and transformed by the method of Mandel & Higa (1970).

Isolation of DNA. *B. subtilis* chromosomal DNA was purified as described by Bron (1990). Plasmid DNA was isolated by the alkaline lysis method of Ish-Horowicz & Burke (1981).

RESULTS AND DISCUSSION

Cloning of the 21·8 kb DNA region

The entire region around 75° on the genetic map of the *B. subtilis* chromosome was obtained by LR PCR. The region was present on four LR PCR products, which are depicted in Fig. 1. Sequence information of a 6 kb DNA fragment from this region, indicated as XLP03, was kindly provided by Dr Patrick Stragier (Institut de Biologie Physico Chimique, Paris, France). Using sequence information from this fragment and the *glpPFKD* operon (Beijer *et al.*, 1993; GenBank accession no. M99611), primers were designed for LR PCR. In addition to the XLP03 fragment, a 6 kb fragment (XLP25) was obtained by linear amplification between the XLP03 fragment and the glycerol operon. Two successive rounds of inverse LR PCR yielded fragments of 9 kb (XLP09) and 5 kb (XLP13). The restriction sites for the generation of the latter products were *Sal*I and *Asu*II, respectively.

These four PCR fragments described above together span a region of 21·8 kb of non-redundant DNA on the *B. subtilis* chromosome.

Assignment of ORFs in the 21808 bp region

ORFs were searched in all six possible translational reading frames. We selected ORFs of at least 50 codons and more using the following criteria. As start codons

Table 2. **ORFs in the 21808 bp sequence around 75° on the *B. subtilis* chromosome**

ORF	Endpoints (nt)*	Size of deduced product		SD consensus sequence (upper case) and initiation codon (bold)
		aa	Mass (kDa)	
yhcA	1–253	84	9·6	
yhcB	279–806	176	19·0	AgGGAGGTttcCtga**ATG**
yhcC	822–1193	124	14·0	aAggaGAGGTtgaa**ATG**
yhcD	1196–1348	51	6·0	AGAAAaagta**ATG**
yhcE	1356–2114	253	29·5	GGAGGTaAagac**ATG**
yhcF	2120–2482	121	14·0	aGAGGTGtaaat**ATG**
yhcG	2487–3182	232	26·5	agAgGGAGGctAaa**ATG**
yhcH	3202–4116	305	34·5	AaAgAGGAGGaatatg**ATG**
yhaI	4112–5050	313	34·9	AaAgAGGAGGTtcagc**ATG**
cspB	5348–5148 C	67	7·4	AGGAGGaaATttc**ATG**
yhcJ	5784–6572	263	29·2	AGGAGtatggtcaca**ATG**
yhcK	7695–6619 C	359	40·7	aagGGTGATaatat**TTG**
yhcL	7868–9256	463	49·0	GAAgGGAGagtttacctgct**TTG**
yhcM	9754–9302 C	151	17·0	AAAGGAGGgatc**ATG**
yhcN	9904–10470	189	21·0	AAAGGAGGaatTCac**ATG**
yhcO	10653–10949	99	11·4	GGAGtccttgtg**ATG**
yhcP	10943–11557	205	24·1	GGAGGcttaCtccggttta**TTG**
yhcQ	12145–11495 C	217	24·8	AAAGGAGGaatTCggt**TTG**
yhcR	12228–15878	1217	132·7	GAAAGGAatTat**ATG**
yhcS	15878–16471	198	22·9	AAAGGAGcgccTCcagaac**GTG**
yhcT	17417–16507	302	33·7	AAAGGAGccatTtaac**ATG**
yhcU	17523–17915	131	15·3	AGGAtaTtcg**ATG**
yhcV	18055–18474	140	14·9	GAAAGGgGtgctgaca**ATG**
yhcW	18604–19263	220	24·6	AAAGGAGtTGtaCcca**GTG**
yhcX	19282–20820	513	60·2	AGAAAGGAGCgagtagg**TTG**
yhxA	21243–21808			AGGGAAcgctaatgaa**ATG**
	(–22593)†	(450)	(49·9)	

* C, complementary strand.

† *yhxA* extends into the unpublished sequence of the glycerol operon.

ATG, GTG or TTG were used if these were preceded within 5–15 bp by a putative Shine–Dalgarno (SD) sequence. The latter should show (partial) complementarity to the 3′ end of the *B. subtilis* 16S rRNA (3′ UCUUUCCUCCACUAG 5′). We also selected ORFs on the basis of high-codon usage statistics, using the bsu.cod table on the EMBL CD-ROM. The bsu_h.cod table, for highly expressed genes, appeared to be not applicable to the region analysed here. In total, 26 putative ORFs were identified and are listed in Table 2. Nomenclature of the ORFs is according to agreements made among the participants in the European *B. subtilis* genome sequencing project.

Similarity analysis of the putative gene products

The similarity of deduced protein products of ORFs in the sequenced DNA region to known protein sequences in the data banks are summarized in Table 3.

YhcD. It is questionable whether this represents a real ORF. It is only 51 codons long and not preceded by an appropriate ribosome binding site (RBS) sequence. Moreover, two large hairpin-like structures are present within the coding region (Fig. 2).

YhcG. This sequence shows similarity to ATP-dependent ABC transporters, or traffic ATPases, like the copper transport protein NosF from *Pseudomonas stutzeri*.

YhcH. Again there is similarity to ATP-dependent ABC transporters, like the copper transport protein NosF from *P. stutzeri*, and also to YhcG of *B. subtilis* (this study), BcrA from *Bacillus licheniformis* and StpC from *Staphylococcus aureus*.

YhcI. This sequence shows similarity to the membrane protein NosY from *P. stutzeri*, involved in copper processing, to BcrB from *B. licheniformis* and to SmpC from *S. aureus*. The flanking genes *yhcH–yhcI* are present in three bacteria other than *B. subtilis*: *P. stutzeri* (*nosF–nosY*), *B. licheniformis* (*bcrA–bcrB*) and *S. aureus* (*stpC–smpC*). The ORFs *yhcE* through *yhcI* conceivably constitute an operon involved in the uptake and processing of copper. No consensus promoter sequence was observed within 200 bp upstream of *yhcE*. Instead, immediately preceding the *yhcE* RBS, two large inverted repeats were found,

Table 3. **Similarity of the deduced ORF products to other proteins**

Product	Similar protein(s) in database	Database accession no.*	Percentage identity (percentage similarity)†
YhcA	None		
YhcB	Trp repressor-binding protein WrbA (*E. coli*) and flavodoxin (*Clostridium acetobutylicum*)	SP: P18855 SP: P304849	37 (58)
YhcC	None		
YhcD	None		
YhcE	None		
YhcF	GntR regulator family, like KorA from *Streptomyces lividans* and FarA from *E. coli*. YhcF is much shorter, spanning only the N-terminal half of these proteins	SP: P22405 SP: P13669	25 (45)
YhcG	ABC transporters: CysA from *Synechococcus* and NosF from *P. stutzeri*	SP: P14788 SP: P19844	31 (56) 31 (57)
YhcH	ABC transporters: NosF (*P. stutzeri*), BcrA (*B. licheniformis*), StpC (*S. aureus*) and YhcG, the preceding ORF on the *B. subtilis* chromosome	SP: P19844 SP: P42332 E: Z30588	35 (60) 28 (51)
YhcI	Membrane protein NosY (*P. stutzeri*), BcrB (*B. licheniformis*) and SmpC (*S. aureus*)	SP: P19845 E: L20573 E: Z305588	23 (53)
CspB	Cold shock protein B	E: X59715	100
YhcJ	Lipoprotein-28 precursor NlpA (*E. coli*)	SP: P04846	28 (51)
YhcK	Hypothetical proteins from *Streptomyces ambofaciens* and *Vibrio anguillarum* (ORF3)	SP: P36892 GB: U17054	25 (47) 36 (54)
YhcL	Proton/sodium-glutamate symport protein GltT from *Bacillus caldotenax*	SP: P24944	25 (55)
YhcM	None		
YhcN	CS3 pili biogenesis protein from *E. coli*	SP: P15487	18 (43)
YhcO	None		
YhcP	None		
YhcQ	Spore coat protein F (CotF) from *B. subtilis*, mainly in the C-terminal half	SP: P23261	20 (44)
YhcR	UDP-sugar hydrolase precursor UshA (*E. coli*), and 5′-nucleotidase precursor (bovine) in the C-terminal half	SP: P07024 SP: Q05927	26 (47)
YhcS	None		
YhcT	DRAP deaminase (*Sacch. cerevisiae*) and a family of hypothetical proteins of which YceC (*E. coli*) is also a member	PIR: S50972 SP: P33643	25 (48) 34 (52)
YhcU	None		
YhcV	IMP dehydrogenase (GuaB), from *B. subtilis* and AcuB (involved in acetoin utilization by *B. subtilis*)	SP: P21879 SP: P39066	29 (55) 24 (52)
YhcW	Phosphoglycolate phosphatase (*Alcaligenes eutrophus*) and a family of hypothetical proteins (YieH from *E. coli*)	SP: P40852 SP: P31467	24 (47)
YhcX	Nitrilase 2 (*Arabidopsis thaliana*) and a hypothetical protein from *Sacch. cerevisiae*	SP: P32962 PIR: S51459	19 (43) 26 (51)
YhxA	DAPA aminotransferase (BioA) from *B. sphaericus*	SP: P22805	37 (55)

* SP, SwissProt; E, EMBL; GB, GenBank.

† Only identities of 18% and greater are mentioned.

comprising nucleotides 1199–1239 and 1268–1340, respectively (depicted in Fig. 2).

CspB. This ORF represents the cold shock gene *cspB* (Willinsky *et al.*, 1992). The nucleotide sequences of the coding regions here and those deposited in EMBL are identical, but some discrepancies were found in the non-coding region. This gene has been mapped, by PBS1 transduction, close to *comK* (Schröder *et al.*, 1993). Since *comK* has been located at 80° on the chromosome (D. Van Sinderen, personal communication), the mapping of *cspB* is probably not very accurate.

YhcM. Although no clear similarity to known proteins was observed, some striking features are worth mention. The codon usage in this ORF is rather biased with respect to

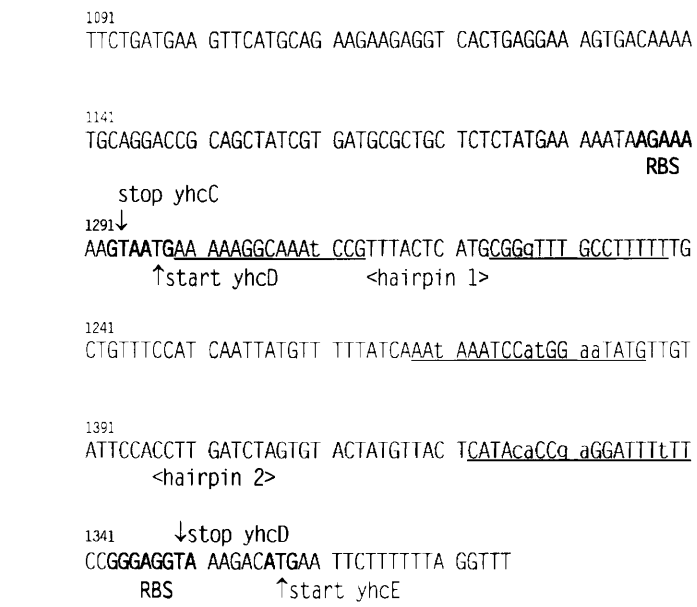

Fig. 2. **Nucleotide sequence from nt 1091 to 1375. Indicated are the start and stop signals of the ORFs in this region, as well as the RBS sequences. Also indicated are the two possible hairpin structures (inverted repeats, underlined) that were found within the *yhcD* coding region. Non-paired bases in these hairpin sequences are indicated by lower-case letters.**

the EMBL bsu.cod table, and the amino acid composition of the deduced protein sequence is abnormal: of the 151 amino acids, 22 are glutamine (14·6 %), 14 serine (9·3 %), 12 arginine (7·9 %), 12 asparagine (7·9 %) and 7 histidine (4·6 %). This means that 35 % of the amino acid residues in this protein have an extra amino group. The calculated pI of this protein is 9·55.

YhcN. This putative protein, like YhcM, has an aberrant amino acid composition: of the 189 amino acids, 31 are asparagine (16·4 %), 20 aspartic acid (10·6 %), 14 lysine (7·4 %) and 12 arginine (6·3 %). In this protein, 30·1 % of the amino acid residues have an extra amino group. Interestingly, the calculated pI of this protein is 5·44. Conceivably, these proteins represent an acidic and basic storage depot for amino groups. The YhcN product has similarity to a protein from *E. coli* (SwissProt P15487) which is essential for the biogenesis of mature CS3 pili.

YhcQ. This sequence shows similarity to spore coat protein F (CotF) from *B. subtilis* (SwissProt P23261). The similarity is mainly located in the C-terminal part of CotF (8 kDa part of processed CotF).

YhcT. This putative protein is a member of a large family of primarily hypothetical proteins found in *Haemophilus influenzae* (SwissProt P44445 and P44433), *E. coli* (SwissProt P33643), *Pseudomonas aeruginosa* (SwissProt P33640), *Mycoplasma genitalium* (GenBank U02214), *Erwinia carotovora* (PIR S45107), *Leishmania donovani* (GenBank U02459), *Schizosaccharomyces pombe* (SwissProt Q09709), *Saccharomyces cerevisiae* (PIR S50972) and *Mycoplasma capricolum* (SwissProt P45614). The function of the *Sacch. cerevisiae* protein has been reported to be DRAP deaminase [DRAP, 2,5-diamino-6-(ribosylamino)-4(3H)-pyrimidinone 5′-phosphate] which is part of the riboflavin biosynthetic pathway (Gonzalez *et al.*, 1995). However, the *B. subtilis* riboflavin operon has already been cloned and is located at 209° on the chromosome; the *ribG* gene encodes the protein responsible for this function (Perkins & Pero, 1993).

YhcV. This sequence shows similarity to IMP dehydrogenase (GuaB or GnaB) of *B. subtilis* (SwissProt P21879) and to AcuB (acetoin utilization) from *B. subtilis* (SwissProt P39066). It is also similar to a family of hypothetical proteins, like Ybp3, from *Desulfurolobus ambivalens* (SwissProt P32987).

YhxA. This is the N-terminal fragment of a hypothetical protein, the C-terminal sequence of which has been published (ORF1 in Beijer *et al.*, 1993; submitted as YhxA_BACSU in SwissProt P33189). It is also similar to DAPA aminotransferase (BioA) from *Bacillus sphaericus* (SwissProt P22805). This BioA homologue is located directly counter-clockwise from the glycerol operon (*glpPFKD*).

ACKNOWLEDGEMENTS

We wish to thank Dr Patrick Stragier and Lena Beijer for communicating unpublished sequence information to us. This work was supported by the Commission of the EU (contract Bio2-CT93-0272) as part of the European *Bacillus* genome sequencing project.

NOTE ADDED IN PROOF

Since this manuscript was accepted for publication, we detected major errors in the existing genetic map (Anagnostopoulos *et al.*, 1993) in the *glyB–addAB* region. As a consequence, the region of the chromosome described here is in fact located 30 kb upstream of the *glyB* locus, in the direction of the origin of replication.

REFERENCES

Anagnostopoulos, C., Piggot, P. J. & Hoch, J. A. (1993). The genetic map of *Bacillus subtilis*. In *Bacillus subtilis and Other Gram-positive Bacteria: Biochemistry, Physiology and Molecular Genetics*, pp. 425–461. Edited by A. L. Sonenshein, J. A. Hoch & R. Losick. Washington, DC: American Society for Microbiology.

Beijer, L., Nilsson, P.-R., Holmberg, C. & Rutberg, L. (1993). The *glpP* and *glpF* genes of the glycerol regulon in *Bacillus subtilis*. *J Gen Microbiol* **139**, 349–359.

Bron, S. (1990). Plasmids. In *Molecular Biological Methods for Bacillus*, pp. 75–174. Edited by C. R. Harwood & S. M. Cutting. Chichester: John Wiley & Sons.

Bron, S. & Venema, G. (1972). Ultraviolet inactivation and excision repair in *Bacillus subtilis*. I. Construction and characterization of a transformable eightfold auxotrophic strain and two ultraviolet-sensitive derivatives. *Mutat Res* **15**, 1–10.

Gonzalez, G. A., Santos, M. A., Garcia-Ramirez, J. J. & Revuelta, J. L. (1995). Cloning and sequencing of the *RIB2* gene from *Saccharomyces cerevisiae*. *J Biol Chem* **270**, 437–444.

Ish-Horowicz, D. & Burke, F. J. (1981). Rapid and efficient cosmid cloning. *Nucleic Acid Res* **9**, 2989–2999.

Mandel, M. & Higa, A. (1970). Calcium-dependent bacteriophage DNA infection. *J Mol Biol* **53**, 159–162.

Perkins, J. B. & Pero, J. G. (1993). Biosynthesis of riboflavin, biotin, folic acid, and cobalamin. In *Bacillus subtilis and Other Gram-positive Bacteria: Biochemistry, Physiology and Molecular Genetics*, pp. 319–334. Edited by A. L. Sonenshein, J. A. Hoch & R. Losick. Washington, DC: American Society for Microbiology.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* **74**, 5463–5467.

Schröder, K., Zuber, P., Willimsky, G., Wagner, B. & Marahiel, M. A. (1993). Mapping of the *Bacillus subtilis cspB* gene and cloning of its homologs in thermophilic, mesophilic and psychrophilic bacilli. *Gene* **136**, 277–280.

Willinsky, G., Bang, H., Fisher, G. & Marahiel, M. A. (1992). Characterization of *cspB*, a *Bacillus subtilis* inducible cold shock gene affecting cell viability at low temperature. *J Bacteriol* **174**, 6326–6335.

Received 18 April 1996; revised 27 June 1996; accepted 8 July 1996.