The 172 kb prkA-addAB region from 83° to 97° of the Bacillus subtilis chromosome contains several dysfunctional genes, the glyB marker, many genes encoding transporter proteins, and the ubiquitous hit gene

Noback, Michiel A.; Holsappel, Siger; Kiewiet, Rense; Terpstra, Peter; Wambutt, Rolf; Wedler, Holger; Venema, Gerard; Bron, Sierd

Published in:
Default journal

Publication date:
1998

Microbiology (1998), 144, 859–875 Printed in Great Britain

The 172 kb *prkA–addAB* region from 83° to 97° of the *Bacillus subtilis* chromosome contains several dysfunctional genes, the *glyB* marker, many genes encoding transporter proteins, and the ubiquitous *hit* gene

Michiel A. Noback,[1] Siger Holsappel,[1] Rense Kiewiet,[1] Peter Terpstra,[2] Rolf Wambutt,[3] Holger Wedler,[3] Gerard Venema[1] and Sierd Bron[1]

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

[1] Department of Genetics, Groningen Biomolecular Sciences and Biotechnology Institute (GBB), University of Groningen, Kerklaan 30, 9751 NN Haren, The Netherlands

[2] BioMedical Technology Centre (BMTC), University of Groningen, Hanzeplein 1, Building 25, 9713 GZ Groningen, The Netherlands

[3] AGON, Glienicker Weg 185, 12489 Berlin, Germany

A 171 812 bp nucleotide sequence between *prkA* and *addAB* (83° to 97°) on the genetic map of the *Bacillus subtilis* 168 chromosome was determined and analysed. An accurate physical/genetic map of this previously poorly described chromosomal region was constructed. One hundred and seventy open reading frames (ORFs) were identified on this DNA fragment. These include the previously described genes *cspB, glpPFKD, spoVR, phoAIV, papQ, citRA, sspB, prsA, hpr, pbpF, hemEHY, aprE, comK* and *addAB*. ORF *yhaF* in this region corresponds to the *glyB* marker. Among the striking features of this region are: an abundance of genes encoding (putative) transporter proteins, several dysfunctional genes, the ubiquitous *hit* gene, and five multidrug-resistance-like genes. These analyses have also revealed the existence of numerous paralogues of ORFs in this region: about two-thirds of the putative genes seem to have at least one paralogue in the *B. subtilis* genome.

Keywords: genome sequencing, functional genomics, paralogous genes, *Bacillus subtilis*

INTRODUCTION

Since 1995, when the Gram-negative bacterium *Haemophilus influenzae* was the first free-living organism to be entirely determined at the DNA level (Fleischmann *et al.*, 1995), the sequences of several other microbial genomes have been elucidated. Among these are the smallest known genome of the bacterium *Mycoplasma genitalium* (Fraser *et al.*, 1995), and the genomes of the archaeon *Methanococcus jannaschii* (Bult *et al.*, 1996), the bacterium *Mycoplasma pneumoniae* (Himmelreich *et al.*, 1996), the bacterium *Escherichia coli* (O'Brien, 1997), the cyanobacterium *Synechocystis* PCC 6803 (Kaneko *et al.*, 1996) and the eukaryote *Saccharomyces cerevisiae* (Goffeau *et al.*, 1996, Mewes *et al.*, 1997). In the framework of the combined European/Japanese *Bacillus subtilis* genome sequencing project that was recently completed (Kunst *et al.*, 1997), a 171 812 bp DNA sequence, representing 4·1% of the genome, was determined and analysed by our group. The sequence spans the region between 83° (*prkA*) and 97° (*addAB*) on the genetic map of the *B. subtilis* chromosome (Anagnostopoulos *et al.*, 1993, Biaudet *et al.*, 1996). The present paper deals with the cloning, sequencing and *in silico* analysis of putative genes in this region. The availability of the entire *B. subtilis* genome sequence enabled us to compare ORFs in the region analysed here with all coding sequences in the genome. These analyses revealed a high frequency of paralogous genes, i.e. genes specifying related (putative) proteins. A correction of the existing genetic (Anagnostopoulos *et al.*, 1993; Biaudet *et al.*, 1996) and physical maps (Itaya & Tanaka, 1991) is also presented.

The sequence of part of this region (22 kb) has been published previously (Noback *et al.*, 1996), but for completeness this fragment has been included in this paper.

Abbreviations: LR PCR, long-range PCR; iLR PCR, inverse long-range PCR.

The EMBL accession numbers for the sequences reported in this paper are X96983 and Y14077 to Y14084 inclusive.

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 mainly by long-range PCR (LR PCR; Cheng *et al.*, 1994; Barnes, 1994), or inverse long-range PCR (i-LR PCR) techniques, using the Gene Amp XL-PCR kit with rTth polymerase (Perkin Elmer). All amplification reactions were performed according to the protocols supplied by the manufacturer. i-LR PCR was performed by digestion of *B. subtilis* chromosomal DNA with appropriate restriction enzymes, followed by 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. An overview of the amplified fragments is presented in Fig. 1.

Some fragments were cloned as phage lambda DNA inserts. The *B. subtilis* lambda EMBL12 library, constructed from a sized partial *Sau*3A digest of the chromosome (kindly provided by Dr C. Harwood, University of Newcastle upon Tyne, UK), was screened by the plaque hybridization method (Sambrook *et al.*, 1989) for the presence of desired sequences. For sequence determinations, phage lambda DNA inserts were amplified by LR PCR and subsequently processed in the same way as other LR PCR fragments (see below).

PCR fragments used for sequencing were treated in one of the following ways.

(i) Shotgun cloning in M13mp18 phage by nebulization, followed by DNA sequencing. This method has been described in a previous paper (Noback *et al.*, 1996).

(ii) Shotgun cloning in pUC18 after limited DNase I digestion in buffer consisting of 500 mM Tris/HCl pH 7·6, 100 mM $MnCl_2$, 1 mg BSA ml^{-1}. Subsequently, the DNA fragments were treated with T4 DNA polymerase and Klenow enzyme (in 10 mM Tris/HCl, pH 8·5, 0·25 mM dNTPs, 5 mM $MgCl_2$) and fractionated by agarose gel electrophoresis. Fragments ranging from 500 to 1500 bp were extracted and ligated into pUC18 which had been digested with *Sma*I and treated with alkaline phosphatase. The ligation mixtures were used to transform *E. coli* XL-1 Blue (*supE*$^+$ *lac hsdR17 recA1* [F′ *proAB*$^+$ *lacI*^{q} *lacZ*ΔM15]) (Stratagene). DNA inserts were sequenced by the method described below.

(iii) Sequencing directly on PCR-generated DNA. To prevent sequencing mistakes that were generated during the PCR reaction, eight separate amplification reactions were performed and they were pooled.

Sequence determination. DNAs were isolated on a Vistra DNA Labstation 625 (Amersham) using either the 'automated M13 template preparation kit' or the 'automated plasmid preparation kit'. DNA inserts were sequenced by the dideoxy chain-termination method (Sanger *et al.*, 1977) using the

Table 1. Primer sequences, position, and type of amplification

Primer	Sequence (5′ → 3′)	Position*	Amplification †
SH25	CGG TAT ATA TCT GGC GGA GCT GCA T	29268 C	+XLP02 LR PCR
XLP01b	TGT AAC GGT TGT CAA AGA ACA GGA AC	35832	+XLP21 i-LR PCR *Ssp*I
XLP02	CTA GTG ATC GCA GGC TAT GGA GGC T	23377	+SH25 LR PCR
XLP03	GCA GGT CGT CAG AAT CAG CTC TTC C	23868 C	+XLP10 LR PCR
XLP04	GTA TAC CGA ACA GCG TGG CTC AGA A	145844	+XLP08 LR PCR
XLP05	CCT GTT CGG TCA GCT CCT TCC TAT T	146021	+XLP07 LR PCR
XLP06	CGG CTC TTC ACT CTC AAG GCT ACA C	133516 C	+XLP36 LR PCR
XLP07	CTG TAG AAC CAG TAG GTC CGC CAA G	133157	+XLP05 LR PCR
XLP08	GCT GAT TAT CTC CGC ACA TCT CTC C	164524 C	+XLP04 LR PCR
XLP09	GTC ATA TTC GGC TCT AGC TTC CTG C	18726 C	+XLP11 i-LR PCR *Sal*I
XLP10	CTG ATC GAG ACT GGC AGG AAG C	18689	+XLP03 LR PCR
XLP11	CTG TTC CAT ATC CTG CGC ATC AAG	19030	+XLP09 i-LR PCR *Sal*I
XLP12	GAA GCC TTC GCC TTG AAT AGC AGA G	12695	+XLP13 i-LR PCR *Asu*II
XLP13	TGC CAT CCA CAT ACT GAG TCA AGT C	12397 C	+XLP12 i-LR PCR *Asu*II
XLP17	GGT GAC AGC CTC AAT CGT ATC CAT C	90063 C	+XLP18 i-LR PCR *Pst*I
XLP18	GAA GGA CCA AGG ATC ACC AAG AAG G	90500	+XLP17 i-LR PCR *Pst*I
XLP20	GGA TCG ACA GAC TTG GCT ACT TGT G	7947	+XLP28 i-LR PCR *Eco*RI
XLP21	GCT TCC TCA CCT TGC TTC GAG ATG T	35360 C	+XLP01b i-LR PCR *Ssp*I
XLP28	GAC ATT GGA ATC GAG TGA TGC GTG	7557 C	+XLP20 i-LR PCR *Eco*RI
XLP35	GAT GAT CCC GCT GAA AGA GTT GAG G	79421 C	+LT7 LR PCR on λ
XLP36	AGA ATA GTT CCG AGC GGC TCA GTT G	109109	+XLP06 LR PCR
XLP38	GCA CAT GTT TTA AGC CGC AAA CCG	41808	+LT7 LR PCR on λ
XLP401	GAC GAT GAA TTG TTT ACT CCG ACC	50328	+XLP402 LR PCR
XLP402	GCG CAC TTG GTG TTC CAG TCA TAG	71296 C	+XLP401 LR PCR
LT7	GCC TAA TAC GAC TCA CTA TAG GGA G		λGEM-11 left arm
LSP6	GGC CAT TTA GGT GAC ACT ATA GAA G		λGEM-11 right arm

* A capital C means that the primer is on the complementary strand.

† In this column the second primer used for the amplification is indicated. In the case of i-LR PCR, the restriction enzyme that was used for digestion of the chromosome is also specified. 'On λ' means that the insert of a recombinant lambda phage was amplified.

Amersham 'automated Delta Taq cycle sequencing kit' and the Amersham Vistra automated DNA sequencer 725. The universal forward sequencing primer was used (5′-GTAAAACGACGGCCAGT-3′). Remaining gaps between the contiguous sequences obtained through shotgun cloning were determined by primer walking on PCR material using the Amersham 'sequenase PCR product sequencing kit' and [^{35}S]dATPαS.

Data handling and computer analysis. DNA sequences were assembled using the Staden package (Dear & Staden, 1991; obtained from MRC, Cambridge, UK). A redundancy of four readings per base, with a minimum of one reading for each strand, was taken as a standard 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 protein products encoded by the ORFs were analysed for similarities to known sequences in databases using the FASTA program (Pearson & Lipman, 1988), and the 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 *prkA–addAB* region

For the cloning of the *prkA–addAB* region we started from two marker regions on the genetic map: the *glpPFKD* operon, which was already cloned and sequenced (Beijer *et al.*, 1993; Holmberg *et al.*, 1990), and the *glyB* marker, which was only genetically mapped (Harford *et al.*, 1976). The cloning and analysis of the *yhcA–glpP* region (22 kb), which is part of the *prkA–addAB* region, has been reported in a previous paper (Noback *et al.*, 1996).

An overview of cloned fragments from this region, and the method by which they were obtained, is shown in Fig. 1. Fragments indicated in this figure as 'formerly known' were partially (at least 10%) resequenced. Other previously known sequences (*cspB*, *sspB*, *prsA*, *hpr*, *hemEHY*, *aprE* and *comK*) were resequenced in their entirety. In a total of about 15 kb of resequenced DNA, less than ten discrepancies were found, and these were present in non-coding areas.

By i-LR PCR, using *Eco*RI from *yhcA* outward in the direction of *prkA* (Fischer *et al.*, 1996), a 5 kb fragment was amplified which spans the region from *yzdC* to *yhcA*. In the other direction, from *glpD* outward in the direction of *addAB*, an i-LR PCR fragment of 7 kb was obtained using *Ssp*I and primers XLP21 and XLP1B. Using a terminal part of this fragment as probe, a lambda DNA clone was isolated containing an additional 3 kb. This fragment unexpectedly proved to contain part of the *spoVR–citA* contig (Beall & Moran, 1994; Hulett *et al.*, 1991; Jin & Sonenshein, 1994a, b), already present in *SubtiList* (the project's central database for *B. subtilis* sequences: Moszer *et al.*, 1995), and mapped outside our region.

A 13·5 kb clone was isolated by screening a lambda-GEM11 genome bank with a 4·5 kb *glyB*$^{+}$ *Sac*I chromosomal fragment (kindly provided by M. Sarvas, Helsinki, Finland). Southern analysis revealed that this clone also contained the *hpr* (Perego & Hoch, 1988) and *prsA* (Kontinen *et al.*, 1991) genes. By plasmid rescue, 'walking' in the direction of *prkA*, two *E. coli* plasmid clones were isolated containing *yhaO–yhaM* (5 kb) and *yhaR–yhaP* (4 kb), respectively. Using the divergent primers XLP17 and XLP18, and *Pst*I-digested chromosomal DNA, a 12 kb DNA fragment was amplified by i-LR PCR (*yhaR* to *yheD*). Using the *yheD* end of this fragment as a probe, a clone was isolated from a lambda-GEM11 genomic bank that contained the *yheD–yheM* region (9 kb). Using a primer from the end of this clone, XLP402, we were able to amplify the region between *yheM* and *citA* (primer XLP401) by LR PCR, yielding a fragment of 21 kb.

Finally, three LR PCR fragments were obtained which together span the region between *glyB* and *addAB*. First, a 26 kb fragment between *yhaA* and *aprE* (Stahl & Ferrari, 1984) was amplified using primers XLP36 and XLP06. Unexpectedly, this fragment contained the *hemEHY* gene cluster (Hansson & Hederstedt, 1992) that was formerly mapped at a different position (94 °). Second, a 12·5 kb fragment was generated between *aprE* and *comK* (primers XLP07 and XLP05). Finally, a PCR fragment was obtained between *comK* and *addB* (primers XLP04 and XLP08), yielding a fragment of 18 kb.

Updating and correction of the genetic map of the *prkA–addAB* region

From our cloning and sequencing data, it became clear that the genetic map of this region (Anagnostopoulos *et al.*, 1993) contained several errors. The corrected genetic/physical map of the region is presented in Fig. 2. The corrected positions of genes are presented in degrees relative to the origin of replication. We calculated the size of a DNA fragment corresponding to one degree on the chromosome by dividing the determined genome size (4214807 bp; Kunst *et al.*, 1997) by 360. According to this calculation, one degree on the chromosome corresponds to 11708 bp.

Assignment of ORFs

ORFs were sought in all six possible reading frames and selected according to the following criteria. A putative ORF should have an ATG, TTG or GTG start codon preceded within 5–15 bp by a Shine–Dalgarno (SD) sequence that is (partly) complementary to the 3′ end of the *B. subtilis* 16S rRNA (3′-UCUUUCCUCCACUAG-5′). We also selected ORFs on the basis of codon usage

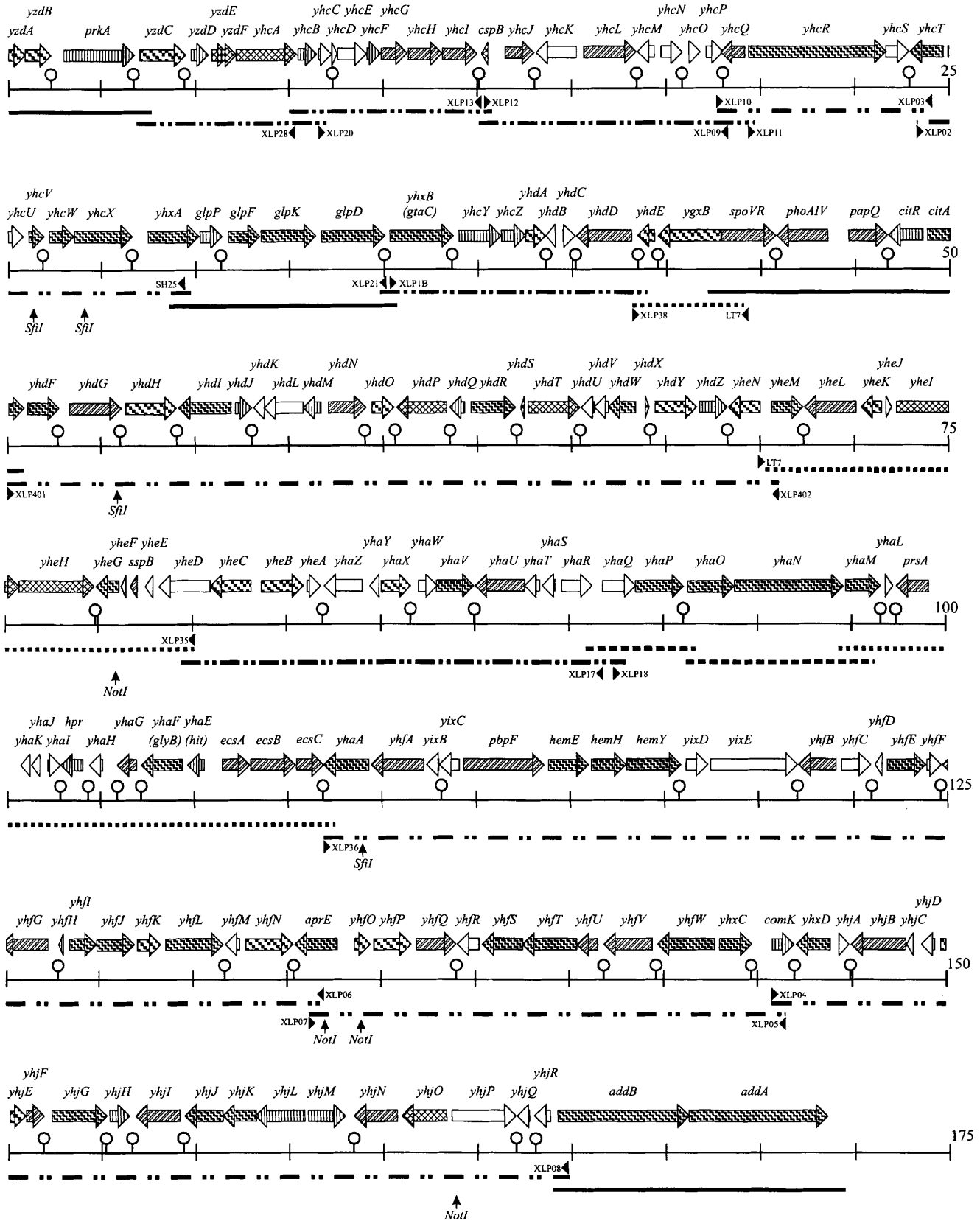

Fig 1. Overview of the *prkA–addAB* region. Below the line representing the map, which is divided into 25 kb fragments, *Not*I and *Sfi*I restriction sites and the method used to obtain the clones are indicated: ——, formerly known sequences spanning more than 5 kb; - - -, plasmid rescue clones; ····, fragments cloned in lambda phage; — ··, fragments cloned by (linear) LR PCR; —·—, fragments cloned by i-LR PCR. Above the line, the ORFs, their classification, and the presence of

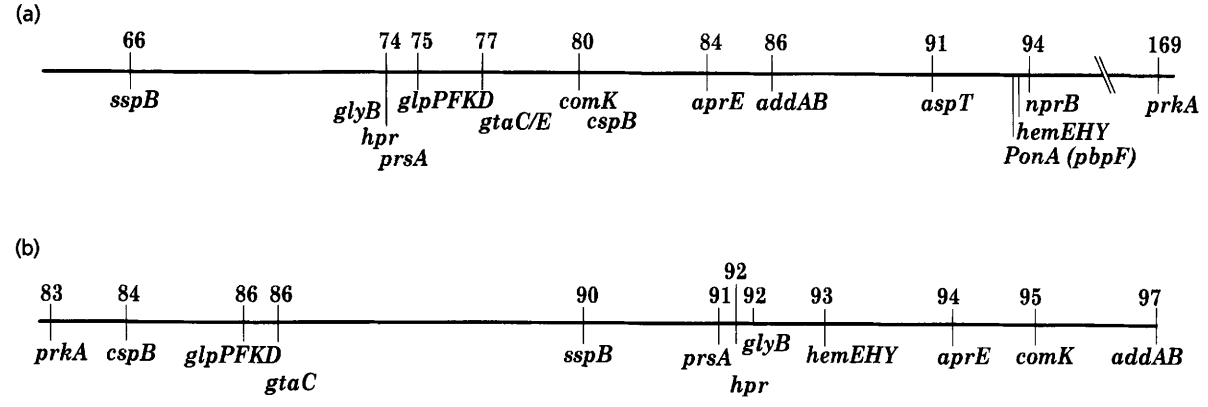

Fig. 2. Update of the genetic map of the *prkA–addAB* region. (a) Part of the genetic map of the *B. subtilis* chromosome according to Anagnostopoulos *et al.* (1993). Numbers above the line representing the map indicate the position, in degrees, relative to the origin of replication. (b) Corrected map of the region based on sequence data. Numbers above the line indicate positions in degrees relative to the origin of replication, as deduced from the total genome sequence, with one degree calculated to be 11708 bp.

statistics, using the Bsu.cod table on the EMBL CD-ROM. In total, 170 ORFs were identified, and these are indicated in Fig. 1. The protein coding density of this region is 90%. Fifty-eight per cent of the putative ORFs are transcribed in the direction of replication fork movement (clockwise); forty-two per cent are transcribed in the counterclockwise direction. The classification of these ORFs according to their putative function (also indicated in Fig. 1) is described in the following section.

Table 2 lists the coordinates of the ORFs relative to the first base in this region, the sizes of the deduced products in amino acids, the calculated molecular masses (kDa) and pIs, and the putative SD sequences. The nomenclature of the ORFs is according to agreements made among the participants in the European/Japanese *B. subtilis* genome sequencing project.

Deduced gene products and similarity analysis

All deduced amino acid sequences from putative genes within this region were compared to known protein sequences in public databases, and to the putative protein products encoded by the *B. subtilis* chromosome.

The similarity of deduced protein products from the sequenced region with known protein sequences in the databases is presented in Table 3. On the basis of similarity to known proteins, we propose that *yhxB* corresponds to the *gtaC* marker and *yhaF* corresponds to the *glyB* marker (see also below).

We classified all ORFs according to their putative function (the results of which are summarized in Fig. 1). The different global classes of functions are mainly as described by Kunst *et al.* (1997). 'Cell envelope and cellular processes' includes proteins involved in cell wall metabolism, transport/binding proteins, lipoproteins, and proteins involved in membrane bioenergetics, mobility, chemotaxis and sporulation. 'Intermediary metabolism' includes proteins involved in the metabolism of carbohydrates, amino acids, nucleotides and nucleic acids, and coenzymes and prosthetic groups. 'Information pathways' includes proteins involved in DNA synthesis, restriction/modification, recombination and repair, RNA synthesis, and protein synthesis. 'Other' includes functions like antibiotic production, drug (-analogue) sensitivity, and adaptation to atypical conditions ('stress proteins').

The availability of the entire *B. subtilis* genome sequence (Kunst *et al.*, 1997), enabled us to search for paralogues on the *B. subtilis* chromosome. For this purpose, paralogues were defined as proteins, encoded by the *B. subtilis* chromosome, showing a minimum of 25% identity over at least three-quarters of their amino acid sequence. The numbers of paralogues for the ORF products in the region analysed here are listed in the second column of Table 3. The frequency distribution of paralogous sequences from the region studied here is summarized in Fig. 3. A considerable number of genes in this region have one or more paralogues: only 38% of the deduced proteins are unique, about 23% have one paralogue, 12% have two paralogues, etc. Some protein families have very many representatives. For instance, more than 60 members of the ABC transporter family are present on the *B. subtilis* chromosome, with six

terminator-like sequences are indicated. The ORFs were classified as follows: ▭, genes of unknown function without homologues in public databases; ▩, genes of unknown function with unknown homologues in public databases; ▥, genes involved in information pathways; ▨, genes involved in intermediary metabolism; ▨, genes for cell envelope and cellular processes; ▩, other. ϙ, Terminator-like sequence. See text for further details.

Table 2. Coordinates of ORFs within the *prkA–addAB* region, the size of their deduced products in amino acids and kDa, the calculated pI, and the putative SD sequence and initiation codon

In the column 'ORF', bold letters represent genes which have already been characterized in other studies. In the column 'endpoints', a right-pointing arrow means that the ORF is transcribed clockwise on the chromosome; left-pointing arrows indicate putative genes that are transcribed counterclockwise. In the column 'SD consensus sequence and initiation codon', bases that are complementary to the 16S rRNA are indicated with capitals; the putative initiation codon is indicated in bold capitals. NP, Not present. When an alternative possible initiation codon was found, it is also indicated in bold.

ORF	Endpoints (nt)	Size of deduced product		Calculated pI	SD consensus sequence (upper case) and initiation codon (bold)
		aa	kDa		
yzdA	1 → 431	141	15·2	5·09	
yzdB	443 → 1150	234	25·1	6·90	GtAAGGAGGatcgtaATG
prkA	1500 → 3395	630	72·9	6·36	AtAgAGGAGGTccTtATG
yzdC	3575 → 4753	391	45·3	5·63	AAGGAGGgGAattcATG
yzdD	4913 → 5377	154	17·6	8·56	AGgAAtGAGGTGAaaaggagTTG
yzdE	5446 → 5850	134	14·9	9·64	GAAAGGAGaaaAcaaaATG
yzdF	5697 → 6077	126	13·7	4·49	NP
yhcA	6118 → 7716	532	58·3	9·30	GAAAGGAGGTtgTCttagATG
yhcB	7739 → 8269	176	19·0	4·74	AgGGGGTttcCtgaATG
yhcC	8282 → 8656	124	14·0	5·43	aAggaGAGGTtgaaATG
yhcD	8656 → 8811	51	6·0	9·75	AGAAAaagtaATG
yhcE	8816 → 9577	253	29·5	9·47	GGAGGTaAagacATG
yhcF	9580 → 9945	121	14·0	5·59	aGAGGTGtaaatATG
yhcG	9947 → 10645	232	26·5	5·52	agAgGGAGGctAaaATG
yhcH	10662 → 11579	305	34·5	6·63	AaAgAGGAGGaatatgATG
yhcI	11572 → 12513	313	34·9	6·61	AaAgAGGAGGTtcagcATG
cspB	12605 ← 12808	67	7·4	4·47	AGGAGGaaATttcATG
yhcJ	13244 → 14035	263	29·2	5·21	AGGAGtatggtcacaATG
yhcK	14076 ← 15155	359	40·7	8·56	aagGGTGATaatatTTG
yhcL	15328 → 16716	463	49·0	9·14	GAAgGGAGagtttacctgctTTG
yhcM	16759 ← 17214	151	17·0	9·55	AAAGGAGGgatcATG
yhcN	17364 → 17933	189	21·0	5·44	AAAGGAGGaatTCacATG
yhcO	18113 → 18412	99	11·4	9·56	GGAGtccttgtgATG
yhcP	18403 → 19020	205	24·1	4·94	GGAGGcttaCtccggtttaTTG
yhcQ	18952 ← 19605	217	24·8	6·00	AAAGGAGGaatTCggtTTG
yhcR	19688 → 23341	1217	132·7	4·79	GAAAGGAatTatATG
yhcS	23338 → 23934	198	22·9	7·31	AAAGGAGcgccTCcagaacGTG
yhcT	23964 ← 24872	302	33·7	9·28	AAAGGAGccatTtaacATG
yhcU	24983 → 25375	131	15·3	8·99	AGGAtaTtcgATG
yhcV	25515 → 25937	140	14·9	5·13	GAAAGGgGtgctgacaATG
yhcW	26064 → 26726	220	24·6	4·74	AAAGGAGtTGtaCccaGTG
yhcX	26742 → 28283	513	60·2	5·51	AGAAAGGAGcgagTaggTTG
yhxA	28703 → 30055	450	49·9	5·87	AGGgaacGcTaatgaaATG
glpP	30083 → 30661	192	21·6	8·08	AAAGGAGcacATG
glpF	30840 → 31664	274	28·7	9·30	AGGAGGaatgtgctATG
glpK	31683 → 33173	496	55·1	5·10	AAaGgGGaGAcatcttATG
glpD	33314 → 34981	555	62·5	7·96	AacAAGGAGGaaAcgtaATG
yhxB	35113 → 36810	565	62·9	5·03	AcAtAGGAGGacgaatATG
yhcY	36959 → 38098	379	42·0	6·90	GGAGtgagaaacGTG
yhcZ	38095 → 38739	214	24·0	6·16	AAAGGAGGgGcggtATG
yhdA	38736 → 39260	174	18·9	6·77	gAAtGgaGaGATCtcaaaATG
yhdB	39275 ← 39517	80	9·8	4·68	AGAAAGGAGaaGcgattcATG
yhdC	39718 → 40041	107	12·3	6·59	AcAGGAGactgaaaaaATG
yhdD	40082 ← 41548	488	51·4	10·03	AaAAAGGAGaactaagATG

Table 2. *(cont.)*

ORF	Endpoints (nt)	Size of deduced product		Calculated pI	SD consensus sequence (upper case) and initiation codon (bold)
		aa	kDa		
yhdE	41701 ← 42142	146	16·6	7·83	GAGGTctTattATG
ygxB	42244 ← 43902	552	60·0	9·85	GGActTatctataATG
spoVR	43933 → 45339	468	55·6	5·62	AgtaGgGGgGATtcggTTG
phoAIV	45369 ← 46754	461	50·3	9·52	AAAGGAGGcATGaaaaaaATG
papQ	47286 → 48317	343	37·3	10·31	GGAGGaaAatATG
citR	48336 ← 49262	308	35·6	8·64	AgGGAGaatAgaaATG
citA	49371 → 50471	366	40·9	6·03	GAtAGGaGGaataCaaATG
yhdF	50545 → 51414	289	31·5	5·31	AGGAGtgatgaatGTG
yhdG	51664 → 53061	465	49·7	9·46	GGAGtTGAaggggaATG
yhdH	53179 → 54534	451	48·9	9·48	GAAAGGAaGTGAcgtttaTTG
yhdI	54569 ← 55978	469	52·8	7·07	AcAAAGGAGacatgagATG
yhdJ	56088 → 56516	142	16·4	8·26	GAAAGGgGaTtgagaagATG
yhdK	56547 ← 56837	96	10·6	8·20	GcgAGGtGGaatTATG
yhdL	56825 ← 57901	358	40·6	6·42	AtAAtaGAGGTGtTaATG
yhdM	57891 ← 58382	163	19·4	7·04	AgaGgGGaGAaaaggcaGTG
yhdN	58579 → 59574	331	37·3	4·87	AAGGAGtgGcaCaATG
yhdO	59709 → 60308	199	21·9	9·58	AcAAAGGAaGTGcgatATG
yhdP	60377 ← 61711	444	49·8	4·50	AGAgtGaAGGTtcTaaTTG
yhdQ	61772 ← 62188	138	15·8	7·95	GttgGGAGGgatATA
yhdR	62360 → 63541	393	43·9	5·13	GGAaGgGAcagATG
yhdS	63681 ← 63791	35	4·1	8·21	AacAttGAGGTacgCgGTG
yhdT	63868 → 65253	461	51·5	4·86	GttgaGAGGatAgggttaaATG
yhdU	65267 ← 65623	118	12·4	9·52	AGAAAGGgGcTGcaggaaaaATG
yhdV	65620 ← 66015	131	13·9	10·04	AtAAAGGAtGgcAaacATG
yhdW	66002 ← 66733	243	27·5	9·24	AGGAGcTGActtagcTTG
yhdX	66967 → 67074	35	4·0	9·70	AaAAAGGAGGcGAgatcATG
yhdY	67223 → 68338	371	42·5	5·98	AgaGgGGaGAcagtcATG
yhdZ	68408 → 69151	247	27·4	5·28	AaAAAGGcGGTGtTgagTTG
yheN	69175 ← 70023	282	31·7	8·46	AAtGGAGagatTgttATG
yheM	70308 → 71156	282	31·2	4·99	AaAAgGGAGGgctTttATG
yheL	71199 ← 72560	453	48·0	7·98	AaAtAtGgGGTGtTattTTG
yheK	72687 ← 73241	184	20·1	4·92	AtAGGAaaaGgTtaaTTG
yheJ	73351 → 73512	53	6·3	10·64	AAGGAatTtgcGTG
yheI	73632 → 75389	585	65·1	6·52	AGGAGaTGgggtagATG
yheH	75386 → 77407	673	76·3	7·31	AagAAGGgGGaGcaggggcATG
yheG	77456 ← 78076	206	22·8	6·04	GcgAGGAGGTttTttaATG
yheF	78115 ← 78240	41	5·0	8·22	AaAAgGGAGGgaATCgggGTG
sspB	78345 ← 78548	67	7·0	4·87	AaAAAGGAGaTttTacacATG
yheE	78757 ← 78975	72	8·5	6·17	GtAAGGAGcGTG
yheD	79125 ← 80486	453	51·4	8·61	AGAAAGGAGtTctTCcgcGTG
yheC	80476 ← 81567	363	41·9	9·03	AAAGGgaGaGtctcaccATG
yheB	81834 → 82967	377	42·9	8·96	AaGGAGGaagatgaataggaATG
yheA	83060 → 83413	117	13·6	4·53	GAAAGGAGcTatTtacaATG
yhaZ	83457 ← 84530	357	41·8	8·86	AAAaGcGGTGtTtatATG
yhaY	84723 ← 84974	83	9·6	9·64	GtgtatAGGatATG
yhaX	85017 → 85814	265	29·2	7·13	AaggAGGgGGacATCtctGTG
yhaW	85994 → 86494	166	19·0	6·34	AtAAAttAGGTGATgaagTTG
yhaV	86458 → 87498	346	39·7	6·07	NP
yhaU	87516 ← 88742	408	43·9	8·97	GGAGagGgcgtGTG
yhaT	88739 ← 89236	165	18·7	5·00	GGAGGgatTttcaTTG
yhaS	89300 ← 89638	112	12·8	7·15	AaAAAGaAGGgatatcttgATG
yhaR	89803 → 90630	275	29·5	6·13	AtGGAGGTGcTtttaATG
yhaQ	90901 → 91797	298	33·8	6·37	GcAAGcAGGaGATtcaGTG

Table 2. *(cont.)*

ORF	Endpoints (nt)	Size of deduced product		Calculated pI	SD consensus sequence (upper case) and initiation codon (bold)
		aa	kDa		
yhaP	91790 → 93049	419	45·4	5·42	AAAGGtGGgGgcCgtct**ATG**
yhaO	93156 → 94382	408	46·8	5·40	GAAAGGAGcaGA**atgTTG**
yhaN	94396 → 97278	963	111·1	6·03	AtAcAtGAGGcG**gTg**acagct**TTG**
yhaM	97352 → 98296	314	35·7	6·12	AcGGAGGgagctttaatagaA**TG**
yhaL	98421 → 98633	70	8·4	5·08	AagggGGAGGaGccC**GTG**
prsA	98674 ← 99552	292	32·5	8·77	AGGAGtgttTgaaaacA**TG**
yhaK	100352 ← 100612	86	9·7	8·93	AGAAAaaAaGTttTacata**TTG**
yhaJ	100630 ← 100869	79	8·9	8·66	AAGGAtGactTtgA**TG**
yhaI	101077 → 101418	113	13·3	4·36	AGAAAGaAGtgGtgtggA**TG**
hpr	101415 ← 102026	203	23·7	5·34	AAGcAGGTGAcgtaA**TG**
yhaH	102204 ← 102560	118	13·1	8·33	AaAAgacgGGTGATtgtaA**TG**
yhaG	102953 ← 103471	172	18·3	10·70	AgAGGAGagcATagttA**TG**
yhaF	103596 ← 104675	359	40·1	5·72	AaAcAGGgaGaGATCataA**TG**
yhaE	104825 ← 105259	145	16·3	6·41	AAGGAGGaaccCtcA**TG**
ecsA	105747 → 106490	247	27·7	5·86	AcAtAaGgGGaGAaactA**TG**
ecsB	106483 → 107709	408	47·3	9·95	AcAAAGGAaGacgctggccA**TG**
ecsC	107729 → 108439	236	26·7	8·77	GAAAaGAGGTaATCaaA**TG**
yhaA	108457 ← 109647	396	43·3	6·14	AAAGGgGGaagcggct**TTG**
yhfA	109720 ← 111111	463	48·8	7·47	AaGgaGTGATCgA**TG**
yixB	111177 ← 111491	104	12·0	9·57	GGAGGaaAgCaaaA**TG**
yixC	111536 ← 112036	166	18·8	5·38	AagcAGGAGGTGgcTgatA**TG**
pbpF	112158 → 114302	714	79·3	7·29	AaAggcGAGGTGAgttcA**TG**
haem	114424 → 115485	353	39·7	5·40	GAAAGGtGGaaATCagA**TG**
hemH	115557 → 116489	310	35·3	4·72	AAAGagGGTGtaaacaG**TG**
hemY	116504 → 117916	470	51·2	8·08	AAAGaAGGcGATgaacA**TG**
yixD	118062 → 118637	191	21·8	7·38	AGtttcGAGGTGAatacaA**TG**
yixE	118708 → 121035	775	84·1	5·08	AGAAtGGAGGcatcaggA**TG**
yhfB	121077 ← 122054	325	35·4	5·90	AAGGAGtgatTCatA**TG**
yhfC	122180 → 122956	258	28·7	8·73	AaAAAGGAGGctgaaaaA**TG**
yhfD	123047 ← 123250	67	8·5	8·11	AgAGGAGGgatTtctA**TG**
yhfE	123369 → 124409	346	38·7	6·16	AAAGGAGGaatgCcctA**TG**
yhfF	124422 → 124829	135	15·3	4·52	AAGGgGGaGgaCcaA**TG**
yhfG	124866 ← 126155	429	45·9	8·82	GGAGGTaATCtA**TG**
yhfH	126426 ← 126560	43	5·1	6·92	GGgAAaGaGGgA**TggttATG**
yhfI	126718 → 127452	244	26·5	5·86	AGAtAGGAGGacATtA**TG**
yhfJ	127465 → 128460	331	38·0	6·09	AtAAAGGAGGaGcaCcA**TG**
yhfK	128525 → 129169	214	22·8	5·30	AGgcAGGAGGgatTCacA**TG**
yhfL	129286 → 130827	513	56·6	5·46	ActtAaGgGGTGggagaA**TG**
yhfM	130866 ← 131261	131	15·0	8·25	GtttGGAGtgatgCaaA**TG**
yhfN	131410 → 132690	426	48·9	6·35	GtgAGGAGtgaggCgttA**TG**
aprE	132729 ← 133874	381	39·5	9·08	AAAGGAGagGgTaaagaG**TG**
yhfO	134309 → 134758	149	16·7	7·99	AGggAGGAaGaaATaagA**TG**
yhfP	134830 → 135822	330	34·8	4·80	AAAGGAGtgtgcCgaA**TG**
yhfQ	135964 → 137010	348	38·6	8·96	AaAtAattGGTGATaA**TG**
yhfR	137042 ← 137623	193	22·0	5·31	AgGaAGGgGATtttA**TG**
yhfS	137694 ← 138788	364	38·4	5·25	GgAAGagaGTGtaCagtataaA**TG**
yhfT	138785 ← 140224	479	52·9	6·20	AAAGGAGGatgaCaatacA**TG**
yhfU	140231 ← 140791	186	20·0	10·32	GGAGGattTCacA**TG**
yhfV	140926 ← 142224	432	48·8	5·62	AAGGgGGatcattgta**TTG**
yhfW	142363 ← 143892	509	57·1	5·90	GAttGGAGGTataacggc**TTG**
yhxC	144004 → 144861	285	30·8	7·42	GAAAaGgaGTGATttca**TTG**
comK	145415 → 145993	192	22·4	7·77	AGgAtGGAGGccATaatA**TG**
yhxD	146040 ← 146939	299	31·9	4·64	AAAGGAGcgttgCtgA**TG**

Table 2. *(cont.)*

ORF	Endpoints (nt)	Size of deduced product		Calculated pI	SD consensus sequence (upper case) and initiation codon (bold)
		aa	kDa		
yhjA	147156 → 147425	89	9·8	9·99	GagaGTGAatcgtcATG
yhjB	147468 ← 148937	489	52·8	9·42	AaAAAGGAGGaagcagaATG
yhjC	148934 ← 149134	66	7·4	7·14	AAGGAGGattctATG
yhjD	149342 ← 149704	120	14·5	5·87	AGAAAGaAGGaGtcaatATG
yhjE	149857 → 150480	207	23·3	10·12	GtAAGGAGtatAaATG
yhjF	150482 → 150988	168	19·0	9·72	AtgAtGGAGGgagaCagtaacATG
yhjG	151171 → 152667	498	54·2	6·96	AAAGGAGtgGtgaatgATG
yhjH	152744 → 153271	175	20·4	7·55	AaAAAGGAtGgaAaacccgATG
yhjI	153429 ← 154634	401	44·9	8·70	AtaGgGGTGtaatgaATG
yhjJ	154706 ← 155758	350	39·3	6·50	AGGAGGaaATaaaaATG
yhjK	155761 ← 156621	286	33·2	5·52	AAtGGAGGgacTgtttcATG
yhjL	156593 ← 157918	441	50·1	6·24	AttGGAGGTacTgttcATG
yhjM	158022 → 159011	329	37·7	6·91	AAGGAaGgGAaaatATG
yhjN	159225 ← 160379	384	41·0	9·64	AGgAAGaAGGgttTtacaTTG
yhjO	160486 ← 161691	401	44·1	9·53	GAAAGGcGGcGATCacATG
yhjP	161805 → 163532	575	66·4	6·40	GtcgGGAGGTGcgggaTTG
yhjQ	163562 ← 163888	108	11·8	5·03	AtAcAGGgGGaatcaaccATG
yhjR	164006 ← 164443	145	17·2	6·21	AGAAtGGAGtTGAatccccTTG
addB	164627 → 168127	1166	134·6	5·56	AagAgaGgGGTctTCtaatTTG
addA	168114 → 171812	1232	141·1	5·26	AaAAAGGAGGcGgatggcaATG

representatives in this region: *yhaD*, *yhaQ*, *yhcG*, *yhcH*, *yheI* and *yheH*. The paralogue frequency distribution observed within this region is globally similar to what is observed for the entire genome (Kunst *et al.*, 1997).

Identification of the *glyB* gene

The homology of the deduced amino acid sequence of ORF *yhaF* to several phosphoserine aminotransferases, or SerC proteins, from other organisms (see Table 3) suggested that *yhaF* might correspond to the *glyB* marker. The SerC protein is involved in the biosynthesis of serine: it catalyses the conversion of 3-phosphohydroxypyruvate to 3-phosphoserine, which is subsequently converted to serine by phosphoserine phosphatase (SerB). Glycine is metabolically derived either from serine by serine hydroxymethyltransferase (GlyA), or from threonine by threonine aldolase (Stauffer, 1983). To confirm that *yhaF* indeed corresponded to the *glyB* marker, we transformed *B. subtilis* strain 1A5 (*glyB133 metC3 tre-12 trpC2*; Dedonder *et al.*, 1977) with a *B. subtilis* plasmid carrying the entire *yhaF* coding sequence under control of a constitutive promoter. This plasmid, and not its parental plasmid without ORF *yhaF*, did complement strain *B. subtilis* 1A5 with respect to glycine auxotrophy when cultured in minimal medium. This indicates that *yhaF* corresponds to the *glyB* marker.

Evidence for non-functional (remnants of) genes

In the region studied, several ORFs were found of which the deduced proteins are almost certainly not functional or not expressed. This is likely to be due to rearrangements and/or deletions within the coding sequence or absence of proper transcriptional or translational signals for expression. Based on repeated sequence analysis of these regions, we feel confident that these findings are not the result of sequencing errors. The first example is the *yzdE-yzdF* pair of partially overlapping ORFs, which code for the N-terminal (*yzdE*) and C-terminal (*yzdF*) fragments, respectively, of a protein that is present in its entirety in *E. coli* (EmrA; Lomovskaya & Lewis, 1992) and *H. influenzae* (EmrA; Fleischmann *et al.*, 1995). When compared to the *E. coli* and *H. influenzae* genes, the middle 350 bp are absent in *B. subtilis*, which also results in a frameshift (Fig. 4). Moreover, *yzdE* is preceded by proper translational start signals (a SD sequence followed by an ATG start codon), but such signals are absent upstream of *yzdF*.

The second example is *yhaV*. Its deduced ORF product displays significant homology to several HemN proteins, or anaerobic coproporphyrinogen III oxidases involved in haem synthesis under anaerobic conditions, from *H. influenzae* (383 aa), *E. coli* (457 aa), *Salmonella typhimurium* (457 aa) and *Rhodobacter sphaeroides* (305 aa). However, no possible translational start site could be found for *yhaV*, and the homology is mainly restricted to the N-terminal two-thirds of the protein.

Another interesting ORF is located upstream of, and partially overlapping with, *yhaE*. ORF *yhaE* encodes a possible *B. subtilis* representative of the ubiquitous Hit-like protein (Seraphin, 1992). The first member of this family of proteins was isolated from bovine tissue and

Table 3. **Deduced ORF products, the number of paralogous sequences, and their similarities with protein sequences in public databases**

Proteins that were previously known are indicated in bold. In the second column, the numbers of paralogous sequences within the *B. subtilis* genome are indicated. Hypo, hypothetical protein (no experimental evidence for its function). SP, Swiss Prot; GB, GenBank; E, EMBL; GP, GenPept. The final column gives the percentage identity, the Smith–Waterman score (S–W score), and the length of the homology, in amino acids.

ORF product	No. of paralogues	Similar protein(s) in databases	Database accession no.	% Identity, S–W score, overlap (aa)
YhbE	2	= YzdA from *Bacillus subtilis*	SP: P39132	100
YhbF	1	= YzdB from *B. subtilis*	SP: P39133	100
PrkA	0	= Protein kinase A PrkA from *B. subtilis*	SP: P39134	100
YhbH	0	= YzdC from *B. subtilis*	SP: P45742	100
		Hypo YzdC from *Escherichia coli*	D90822/g1736412	27, 523, 411
YhbI	4	Multiple antibiotic resistance operon regulatory protein MarR from *Salmonella typhimurium*	U54468/g1293698	30, 189, 138
YhbJ	0	Multidrug resistance protein A (EmrA) from *E. coli*	SP: P27303	29, 121, 75
YzdF	1	Multidrug resistance protein A (EmrA) from *E. coli*	SP: P27303	31, 216, 114
YhcA	5	Multidrug resistance protein B (EmrB) from *E. coli*	SP: P27304	29, 732, 431
YhcB	1	Trp repressor-binding protein WrbA from *E. coli*	SP: P304849	32, 294, 189
		Flavodoxin from *Clostridium acetobutylicum*	SP: P18855	31, 210, 119
YhcC	3	None		
YhcD	1	None		
YhcE	0	None		
YhcF	5	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 (KorA); SP: P13669 (FarA)	28, 161, 881; 39, 156, 7
YhcG	53	ABC transporters:		
		CysA from *Synechococcus* sp.	SP: P14788	34, 369, 212
		NosF from *Pseudomonas stutzeri*	SP: P19844	31, 373, 222
YhcH	29	ABC transporters:		
		NosF from *P. stutzeri*	SP: P19844	34, 544, 307
		BcrA from *Bacillus licheniformis*	SP: P42332	37, 683, 303
		StpC (*Staphylococcus aureus*)	E: Z30588/g459256	37, 535, 226
		YhcG, the preceding ORF on the *B. subtilis* chromosome	This paper	
YhcI	1	Membrane protein NosY from *P. stutzeri*	SP: P19845	25, 123, 231
		BcrB from *B. licheniformis*	SP: P42333	24, 139, 183
		SmpC from *Staphylococcus aureus*	E: Z30588/g459257	26, 242, 219
CspB	4	Cold-shock protein B	U58859/g1336658	100
YhcJ	1	Lipoprotein-28 precursor NlpA from *E. coli*	SP: P04846	30, 374, 257
YhcK	0	Hypothetical proteins from		
		Streptomyces ambofaciens	SP: P36892	29, 166, 162
		Vibrio anguillarum (ORF3)	U17054/g576657	34, 313, 201
YhcL	0	Proton/sodium-glutamate symport protein GltT from *Bacillus caldotenax*	SP: P24944	27, 483, 421
YhcM	0	None		
YhcN	0	CS3 pili biogenesis protein from *E. coli*	SP: P15487	22, 81, 98
YhcO	3	None		
YhcP	2	None		
YhcQ	0	Spore coat protein F (CotF) from *B. subtilis*, mainly in the C-terminal half	SP: P23261	23, 122, 90
YhcR	0	The C-terminal half: UDP-sugar hydrolase precursor UshA from *E. coli*	SP: P07024	28, 490, 572
		5′-Nucleotidase precursor from *Bos taurus* (bovine)	SP: Q05927	22, 383, 546
YhcS	0	None		
YhcT	1	DRAP deaminase from *Saccharomyces cerevisiae*	PIR: S50972	24, 274, 246

Table 3. *(cont.)*

ORF product	No. of paralogues	Similar protein(s) in databases	Database accession no.	% Identity, S–W score, overlap (aa)
		A family of hypothetical proteins of which YceC from *E. coli* is also a member	SP: P33643	39, 529, 254
YhcU	0	None		
YhcV	9	IMP dehydrogenase GuaB from *B. subtilis*	SP: P21879	31, 193, 118
		AcuB (involved in acetoin utilization) from *B. subtilis*	SP: P39066	27, 160, 121
YhcW	3	Phosphoglycolate phosphatase from *Alcaligenes eutrophus*	SP: P40852	25, 179, 186
		A family of hypothetical proteins (like YieH from *E. coli*)	SP: P31467	27, 204, 181
YhcX	0	Nitrilase 2 from *Arabidopsis thaliana*	SP: P32962	34, 156, 103
		A hypothetical protein from *S. cerevisiae*	PIR: S51459	27, 326, 292
YhxA	6	DAPA aminotransferase (BioA) from *Bacillus sphaericus*	SP: P22805	34, 839, 446
GlpP	1	= Glycerol operon regulator GlpP from *B. subtilis*	SP: P30300	100
GlpF	2	= Glycerol uptake facilitator GlpF from *B. subtilis*	SP: P18156	100
GlpK	2	= Glycerol kinase GlpK from *B. subtilis*	SP: P18157	100
GlpD	0	= Glycerol-3-phosphate dehydrogenase GlpD from *B. subtilis*	SP: P18158	100
YhxB	1	Phosphomannomutase or phosphoglucomutase from *Mycoplasma pirum*	PIR: E53312	28, 793, 564
YhcY	0	Sensory transduction kinase DegS from *B. subtilis*	SP: P13799	31, 261, 221
YhcZ	15	Transcriptional regulator DegU from *B. subtilis*	SP: P13800	39, 517, 219
YhdA	2	Hypo YieF from *E. coli*	SP: P31465	26, 174, 136
YhdB	0	None		
YhdC	0	None		
YhdD	2	Phosphatase-associated protein PapQ from *B. subtilis*	GB: U38819	50, 943, 316
YhdE	1	Hypo YjeB from *E. coli*	SP: P40610	44, 393, 142
YgxB	0	= YgxB from *B. subtilis* (partial)	SP: P37874	100
		Hypo from *Synechococcus* sp.	PIR: S20924	28, 248, 173
SpoVR	0	= Stage V sporulation protein SpoVR from *B. subtilis*	SP: P37875	100
PhoAIV	1	= Alkaline phosphatase PhoAIV from *B. subtilis*	SP: P19406	100
PapQ	4	= Phosphatase-associated protein PapQ from *B. subtilis*	E: U38819	100
CitR	2	= Negative regulator for *citA*, CitR, from *B. subtilis*	SP: P39127	100
CitA	2	= Citrate synthase I CitA from *B. subtilis*	SP: P39119	100
YhdF	21	Glucose and ribitol dehydrogenase from *Hordeum vulgare* (barley)	GP: S7226	52, 952, 286
YhdG	5	Hypo from *Mycobacterium tuberculosis*	Z79702/g264157	41, 1269, 464
		Cationic amino acid transporter from *Homo sapiens*	D29990/g849051	36, 893, 435
YhdH	1	Hypo YG90 from *Haemophilus influenzae*	SP: P455320	36, 1064, 457
YhdI	5	Probable rhizopine catabolism regulatory protein MocR from *Rhizobium meliloti*	SP: P49309	34, 897, 481
		Aminotransferase from *Sulfolobus solfataricus*	E283830/g1707790	27, 397, 370
YhdJ	0	Regulator of alkylphosphate uptake PhnO from *E. coli*	SP: P16691	34, 136, 82
YhdK	4	None		
YhdL	0	None		
YhdM	7	Putative RNA polymerase sigma factor YbbL from *B. subtilis*	D84214/g1256141	31, 280, 160
YhdN	5	Hypo YxbF from *B. subtilis*	SP: P46336	37, 704, 311
		Potassium channel $\beta 2$ subunit from *Homo sapiens* (human)	U33429/g995761	30, 402, 334
YhdO	0	Hypo from *Synechocystis* sp.	D90915/g1653690	26, 200, 180
YhdP	4	YhdT from *B. subtilis*	This paper	61, 1687, 430
		Haemolysin from *Synechocystis* sp.	D90914/g1653594	30, 677, 441
YhdQ	2	Hypo HI1623 from *H. influenzae*	SP: P45277	33, 184, 120
		Mercury resistance regulatory protein MerR from *Thiobacillus ferrooxidans*	SP: P22896	35, 154, 87
YhdR	4	Aspartate aminotransferase from *Methanococcus jannaschii*	U67459/g1592252	30, 520, 391
YhdS	0	Hypo from fowlpox virus (small internal fragment)	SP: P21973	44, 63, 25
YhdT	4	YhdP from *B. subtilis*	This paper	61, 1687, 430
		Haemolysin from *Synechocystis* sp.	D90914/g1653594	31, 683, 439

Table 3. *(cont.)*

ORF product	No. of paralogues	Similar protein(s) in databases	Database accession no.	% Identity, S–W score, overlap (aa)
YhdU	2	NADH-plastoquinone oxidoreductase chain 2 (chloroplast) from *Marchantia polymorpha*	SP: P06257	24, 125, 122
YhdV	5	None		
YhdW	2	Glycerol diester phosphodiesterase GlpQ from *B. subtilis*	SP: P37965	38, 575, 252
YhdX	0	Hypo human transposon L1.1 ORF1	M80340/g339770	32, 60, 34
YhdY	0	Hypo MJ1143 from *M. jannaschii*	g1591775	27, 550, 357
YhdZ	0	Lac repressor LacR from *S. aureus*	M32103/g845686	36, 446, 251
YheN	1	Hypo Yfu2 from *B. stearothermophilus*	SP: Q04729	32, 305, 205
YheM	2	D-Amino acid aminotransferase from *B. licheniformis*	U26947/g857561	64, 1179, 275
YheL	1	Na^+/H^+ antiporter from *B. firmus*	SP: P27611	53, 1377, 390
YheK	1	Hypo YxiE from *B. subtilis*	SP: P42297	30, 230, 166
YheJ	0	None		
YheI	11	Multidrug-resistance-like ATP binding protein MDL from *E. coli*	SP: P30751	37, 1134, 507
YheH	9	Multidrug-resistance-like ATP binding protein MDL from *E. coli*	SP: P30751	40, 1341, 519
YheG	2	Flavin reductase FLR from *Bos taurus* (bovine)	SP: P52556	27, 211, 208
YheF	0	None		
SspB	3	= Small, acid-soluble spore protein B, SspB, from *B. subtilis*	SP: P04832	100
YheE	1	None		
YheD	0	None		
YheC	0	Central part of hypo MJ0776 from *M. jannaschii*	U67522/g1499596	32, 142, 123
YheB	0	Hypo orf sll0412 from *Synechocystis* sp.	D64001/g1001108	26, 335, 405
YheA	3	None		
YhaZ	0	None		
YhaY	1	None		
YhaX	2	Hypo YcsE from *B. subtilis*	SP: P42962	27, 266, 257
		Hypo Cof protein from *E. coli*	SP: P46891	26, 234, 251
YhaW	1	None		
YhaV	1	Anaerobic coproporphyrinogen III oxidase HemN from *H. influenzae* (see also text)	SP: P43899	27, 404, 332
YhaU	1	Na^+/H^+ antiporter from *Enterococcus hirae*	SP: P26235	26, 410, 386
YhaT	2	C-terminal part of hypo from *Synechocystis* sp.	D64006/g1001375	29, 138, 84
YhaS	0	None		
YhaR	4	Enoyl-CoA-hydratase from *Rhodobacter capsulatus*	SP: P24162	33, 390, 246
YhaQ	24	ATP-binding transport proteins (ABC transporter) from:		
		B. firmus (hypothetical)	SP: P26946	62, 1168, 266
		M. jannaschii	U67545/g1499865	42, 690, 260
YhaP	0	N-terminal part to methylmalonyl-CoA mutase homologue, MutX from *B. firmus*	SP: P26947	45, 168, 56
		M. jannaschii hypo MJ1024 (full length)	U67545/g1499866	25, 403, 402
YhaO	0	Hypo sll0021 from *Synechocystis* sp.	D64000/g1001554	26, 228, 310
		Hypo MJ1323 from *M. jannaschii*	U67572/g1591963	25, 243, 306
		SbcD from *E. coli*	SP: P13457	25, 154, 276
		SbcD homologue from *B. subtilis*	SP: P23479	24, 141, 277
YhaN	0	Hypo Orf X from *S. aureus* (from aa 600 of YhaN)	U21636/g710421	25, 415, 358
		Exonuclease subunit SbcC from *E. coli*	SP: P13458	20, 313, 856
		Rad50 of multiprotein complex implicated in recombinational DNA repair from *H. sapiens*	U63139/g1518806	21, 234, 821
YhaM	0	Cmp-binding factor 1 from *S. aureus*	U21636/g710422	52, 1137, 300
		Hypo MJ0837 from *M. jannaschii*	U67528/g1499663	32, 196, 144
YhaL	1	None		
PrsA	4	= Protein export protein PrsA from *B. subtilis*	SP: P24327	100
YhaK	1	None		

Table 3. *(cont.)*

ORF product	No. of paralogues	Similar protein(s) in databases	Database accession no.	% Identity, S–W score, overlap (aa)
YhaJ	2	None		
YhaI	0	None		
Hpr	0	= Protease production regulatory protein Hpr from *B. subtilis*	SP: P11065	
YhaH	2	Clone pSJ7 product from *B. subtilis* (from aa 57 of yhaH)	S70232/g547157	79, 229, 42
		Hypo YtxH from *B. subtilis*	SP: P40780	25, 178, 113
		Apolipoprotein A-I (Apo-AI) precursor from *Oryctolagus cuniculus* (rabbit)	SP: P09809	29, 128, 107
YhaG	1	Glycine betaine/L-proline transport system permease protein ProW from *E. coli* (only C-terminal half; see also text)	SP: P14176	20, 86, 148
YhaF	0	Phosphoserine aminotransferases from:		
		B. circulans	gnl: PID: e123178	54, 1329, 357
		Spinacia oleracea (SerC)	SP: P52877	50, 1162, 363
		Arabidopsis thaliana	D88541/g1665831	50, 1156, 362
		H. influenzae (SerC)	SP: P44336	46, 985, 360
		Rabbit (SerC) and	SP: P10658	44, 1000, 362
		E. coli (SerC)	SP: P23721	44, 953, 364
YhaE	1	Member of the HIT family of proteins, with members from:		
		M. jannaschii	U67530/g1499694	50, 372, 128
		Mycoplasma pneumoniae	/g1674261	49, 365, 110
		Borrelia burgdorferi	U49938/g1753229	50, 354, 113
		Mycoplasma genitalium	SP: P47378	46, 352, 134
		S. solfataricus	Y08256/g1707769	42, 300, 105
EcsA	60	= ABC-type transporter ATP-binding protein EcsA from *B. subtilis*	SP: P55339	100
EcsB	0	= Hypothetical integral membrane protein EcsB from *B. subtilis*	SP: P55340	100
EcsC	1	= Protein EcsC from *B. subtilis*	SP: P55341	100
YhaA	4	*N*-Acyl-L-amino acid amidohydrolase from *B. stearothermophilus*	SP: P37112	43, 864, 305
YhfA	0	Anaerobic carrier for dicarboxylates DcuC from *E. coli*	X99112/g252616	24, 194, 476
YixB	0	= Hypo YixB from *B. subtilis* (fragment)	SP: P38048	100, 67
YixC	1	= Hypo YixC from *B. subtilis*	SP: P38049	100
PbpF	3	= Penicillin-binding protein PbpF from *B. subtilis*	SP: P38050	100
Haem	0	= Uroporphyrinogen decarboxylase Haem (= DcuP) from *B. subtilis*	SP: P32395	100
HemH	0	= Ferrochelatase HemH from *B. subtilis*	SP: P32396	100
HemY	0	= Coproporphyrinogen III oxidase HemY from *B. subtilis*	SP: P32397	100
YixD	5	= Hypo YixD from *B. subtilis*	SP: P32398	100
YixE	0	= Hypo protein in HemY 3′ region (orfB; fragment) from *B. subtilis*	SP: P32399	100, 145
		Phage infection protein from:		
		Lactococcus lactis	SP: P49022	23, 742, 885
YhfB	1	β-Ketoacyl-acyl carrier protein (FabH) from *E. coli*	SP: P24249	39, 741, 319
		Porphyra purpurea, and others	SP: P51196	36, 720, 323
YhfC	1	None		
YhfD	1	Part of metallothionein isoform Ia from *Callinectes sapidus*	g1176448	29, 63 31
YhfE	1	Endoglucanase CelM from *Clostridium thermocellum*	g1097207	26, 304, 345
YhfF	1	Late embryogenesis abundant protein group 3 from *Tritium aestivum* (wheat); partial	PIR: S33616	29, 99, 96
YhfG	2	Proton/sodium-glutamate symport protein from:		
		B. stearothermophilus (GltT)	SP: P24943	64, 1489, 344
		B. caldolyticus (GltT)	SP: P24944	63, 1478, 344
		E. coli (GltP)	SP: P21345	57, 1272, 341
		B. subtilis (GltP)	SP: P39817	46, 1037, 349

Table 3. *(cont.)*

ORF product	No. of paralogues	Similar protein(s) in databases	Database accession no.	% Identity, S–W score, overlap (aa)
YhfH	0	Small toxin SCXI from *Mesobuthus tamulus sindicus* scorpion, and low similarity to many zinc-finger proteins; this ORF contains the zinc finger motif CXXC...CXXC	SP: P15229	52, 71, 23
YhfI	1	Arylsulfatase precursor from *Mycobacterium leprae*	U00014/g466916	29, 337, 249
YhfJ	0	Lipoate protein ligase from:		
		M. pneumoniae (LplA)	U00089/g1674137	34, 758, 327
		M. genitalium (LplA)	SP: P47512	34, 700, 336
		E. coli (LplA)	SP: P32099	35, 596, 315
YhfK	3	Hypo YM9582.15 from *S. cerevisiae*	PIR: S54466	38, 462, 225
YhfL	19	Long-chain-fatty-acid CoA ligase LcfA from:		
		E. coli	SP: P29212	40, 1173, 533
		H. influenzae	SP: P46450	36, 1040, 532
YhfM	0	None		
YhfN	0	Hypo YzoA from *B. subtilis* (= fragment of YhfN),	SP: P40769	100, 42
		Hypo YJ87 from *S. cerevisiae*	SP: P47154	25, 382, 419
AprE	5	= Subtilisin (extracellular alkaline serine protease) from *B. subtilis*	SP: P04189	100
YhfO	3	Hypo Y677 from *H. influenzae*	SP: P44036	32, 234, 135
YhfP	3	Hypo YhdH from *E. coli*	SP: P26646	47, 976, 325
YhfQ	7	Iron(III) dicitrate transport protein from:		
		E. coli (FecB)	PIR: S56515	32, 486, 282
		Synechocystis sp.	D90899/g1651665	28, 434, 328
YhfR	0	Hypo o215b from *E. coli*	PIR: S56619	32, 307, 189
		Probable phosphoglycerate mutase (Pgm) from *E. coli*	SP: P36942	32, 303, 189
		Pgm from *Treponema pallidum*	U55214/g1777938	38, 221, 100
YhfS	2	Acetyl-CoA acetyltransferase ThiL from:		
		Thiocystis violacea	SP: P45363	39, 790, 392
		Chromatium vinosum	SP: P45369	38, 788, 394
		Alcaligenes eutrophus	SP: P14611	40, 773, 392
		B. subtilis	SP: P45855	38, 729, 391
YhfT	8	Long-chain-acyl-CoA synthetase from *B. subtilis*	Z75208/g1770038	29, 590, 539
		Bile acid-CoA ligase from *Eubacterium* sp.	SP: P19409	28, 546, 487
		Long-chain-fatty-acid-CoA ligase (LcfA) from *E. coli*	SP: P29212	26, 455, 479
YhfU	4	BioY (biotin synthesis) from *B. sphaericus*	SP: P22819	31, 250, 186
YhfV	0	Methyl-accepting chemotaxis protein from:		
		Halobacterium salinarium (HtB)	U75436/g1654420	26, 496, 454
		B. subtilis (TlpC)	SP: P39209	30, 383, 288
		B. subtilis (TlpB)	SP: P39217	30, 377, 289
		B. subtilis (TlpA)	SP: P39216	30, 366, 250
YhfW	0	Oxidoreductase OrdL from *E. coli*	U38543/g1054921	20, 308, 431
YhxC	22	= YhxC from *B. subtilis* (fragment)	SP: P40397	100, 114
		Glucose and ribitol dehydrogenase homologue from *Hordeum vulgare* (barley)	GB: S72926	56, 1002, 295
ComK	0	= Competence protein K from *B. subtilis*	SP: P40396	100
YhxD	23	= YhxD from *B. subtilis* (fragment)	SP: P40398	100, 140
		Hypo ORF_o294 *E. coli*	U26377/g882532	64, 1281, 292
		Glucose and ribitol dehydrogenase homologue from *H. vulgare* (barley)	GB: S72926	43, 691, 288
YhjA	3	None		
YhjB	1	Proline permease PutP from *S. typhimurium*	GB: S72926	25, 400, 495
YhjC	1	None		
YhjD	1	None		
YhjE	0	Hypo YqeD from *B. subtilis*	D84432/g1303784	22, 225, 190

Table 3. *(cont.)*

ORF product	No. of paralogues	Similar protein(s) in databases	Database accession no.	% Identity, S–W score, overlap (aa)
YhjF	4	Type I signal peptidase from:		
		B. caldolyticus (SipC)	SP: P41027	50, 497, 159
		B. subtilis (SipT)	U45883/g1518930	42, 394, 161
YhjG	0	Tetracycline 6-hydroxylase from *Streptomyces aureofaciens*	PIR: JC4098	40, 1080, 493
		Pentachlorophenol 4-monooxygenase from *Flavobacterium* sp.	SP: P42535	32, 893, 476
YhjH	1	Hypo YzhA from *B. subtilis*	SP: P40762	42, 362, 143
		Multidrug resistance operon repressor MexR from *Pseudomonas aeruginosa*	U23763/g886021	24, 103, 71
YhjI	0	Hypo YOL173w from *S. cerevisiae*	EMBL: Z74879	25, 315, 375
		Glucose and galactose transporter from *Brucella abortus*	U43785/g1171339	22, 227, 365
YhjJ	2	*myo*-Inositol 2-dehydrogenase MI2D from *B. subtilis*	SP: P26935	26, 237, 262
		Glucose:fructose oxidoreductase Gfo from *Zymomonas mobilis*	Z80356/g1657416	23, 200, 307
YhjK	0	Hypo YpdA from *B. stearothermophilus*	SP: P21878	37, 173, 82
		Phosphoserine phosphatase SerB from *H. influenzae*	SP: P44997	23, 102, 230
YhjL	1	Pleiotropic regulatory protein DegT from *B. stearothermophilus*	SP: P15263	37, 695, 369
		Spore coat polysaccharide biosynthesis protein SpsC from *B. subtilis*	SP: P39623	33, 676, 392
YhjM	10	Transcriptional repressor CytR from *E. coli*	SP: P06964	33, 609, 330
		Degradation activator DegA from *B. subtilis*	SP: P37947	31, 568, 331
		Catabolite control protein CcpA from *B. subtilis*	SP: P25144	30, 581, 332
YhjN	0	Hypo f363 from *E. coli*	gi1786933	27, 333, 297
		Proton antiporter efflux protein from *Mycobacterium smegmatis*	U40487/g1110518	23, 95, 271
YhjO	1	Hypo YqjV from *B. subtilis*	D84432/g1303973	23, 423, 392
		Multidrug resistance protein 1 (BMR1) from *B. subtilis*	SP: P33449	25, 307, 381
		Multidrug resistance protein 2 (BMR2) from *B. subtilis*	SP: P39843	24, 274, 385
YhjP	0	Hypo YabN from *E. coli*	SP: P33595	25, 551, 586
		Oligopeptide-binding protein AppA from *B. subtilis*	SP: P42061	26, 223, 298
YhjQ	1	Polyferredoxin from *M. jannaschii*	U67560/g1591821	24, 115, 78
YhjR	0	Nigerythrin from *Desulfovibrio vulgaris*	U71215/g1616801	25, 112, 128
AddB	0	= ATP-dependent deoxyribonuclease subunit B from *B. subtilis*	SP: P23477	100
AddA	0	= ATP-dependent deoxyribonuclease subunit A from *B. subtilis*	SP: P23478	100

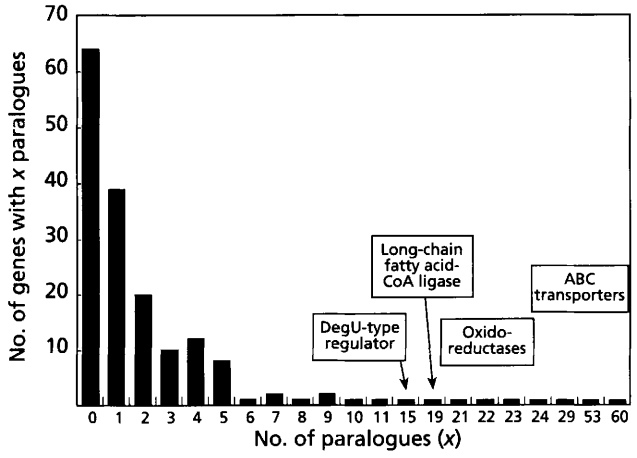

Fig. 3. Paralogue frequency distribution of ORFs in the *prkA–addAB* region compared to all *B. subtilis* protein sequences. On the *x*-axis, the number of paralogues for a given protein sequence is indicated, and on the *y*-axis the number of proteins encoded within the *prkA–addAB* region for which this number of paralogues is found.

identified as being a protein kinase C inhibitor (Pearson *et al.*, 1990). The ORF in front of *yhaE* (results not shown) is 120 codons long, and its deduced amino acid sequence displays blocks of similarity with the catalytic subunit of human DNA-dependent protein kinase (database reference PIR: A57099). However, the latter protein is 4096 aa long. This may be due to 'background noise', but the coincidence of finding these blocks of similarity to a protein kinase together with a gene encoding a putative protein kinase C inhibitor is striking.

A similar situation was found downstream of *yhaG*. The deduced YhaG product displays similarity to ProW from *E. coli*, which is involved in a multicomponent binding-protein-dependent transport system for glycine betaine/L-proline (Gowrishankar, 1989). Downstream of *yhaG*, a small ORF was found that shows some similarity to glycine receptor beta subunits from mouse (database reference GP: MMGRBMRA_1), rat (SP: GRB_RAT) and human (SP: GRB_HUMAN) and an unknown ORF product from *Arabidopsis thaliana*. The deduced ORF product is only 63 aa long, while the

```
EmrA  H.influenzae  MTQIATENPSTKSVSNKTDRKKGLSIFILLLLIIGIACALYWFFFLKDFEETEDAYVGGN   60
EmrA  E.coli        MSANAETQTPQQPVKKSGKRKRLLLLLTLLFIIIAVAIGIYWFLVLRHFEETDDAYVAGN   60
YzdE  B.subtilis                  MNRGRLILTNIIGLIVVLAIIAGGAYYYYQSTNYVKTDEAKVAGD   45
                                        *    . *. .. .*  *..    . *..* * *.

EmrA  H.influenzae  QVMVSSQVAGNVAKINADNMDKVHAGDILVELDDTNAKLSFEQAKSNLANAVRQVEQLGF  120
EmrA  E.coli        QMQIMSQVSGSVTKVWADNTDFVKEGDVLVTLDPTDARQAFEKAKTALASSVRQTHQLMI  120
YzdE  B.subtilis    MAAITAPAAGKVSDWDLDEGKTVKKGDTVAKIKGEQTVDVKSIMDGTIVKNEVKTDKPYK  105
                      . . .* *..   *.  *. ** . ..  .       . .

EmrA  H.influenzae  TVQQLQSAVHANEISLAQAQGNLARRVQLEKMGAIDKESFQEAKEAVELAKANLNASKNQ  180
EmrA  E.coli        NSKQLQANIEVQKIALAQAQSDYNRRVPLGNANLIGREELQHARDAVTSAQAQLDVAIQQ  180
YzdE  B.subtilis    LVQQLHKRLTWTTYTSQQILKKQILRILK                               134
                     .**. .   .  *      *.

EmrA  H.influenzae  LAANQALLRNVPLREQPQIQNAINSLKQAWLNLQRTKIRSPIDGYVARRNVQVGQAVSVG  240
EmrA  E.coli        YNANQAMILGTKLEDQPAVQQAATEVRNAWLALERTRIISPMTGYVSRRAVQPGAQISPT  240
YzdF  B.subtilis                                      RKIHYGWHNCEKRSENGQTVQAG   23
                                                                . . * .

EmrA  H.influenzae  GALMAVVSNEQMWLEANFKETQLTNMRIGQPVKIHFDLYGKNKEFDGVINGIEMGTGNAF  300
EmrA  E.coli        TPLMAVVPATNMWVDANFKETQIANMRIGQPVTITTDIYGDDVKYTGKVVGLDMGTGSAF  300
YzdF  B.subtilis    TTIAQTIDMDNLYITANIKETDIADIEVGNSVDVVVDGDP-DTTFDGTVEEIGYATNSTF   82
                      . .  .... ** ***.......*. * . * . .  * . .  * .*

EmrA  H.influenzae  SLLPSQNATGNWIKVVQRVPVRIKLDPQQFTETPLRIGLSATAKVRISDSSGAMLREKTE  360
EmrA  E.coli        SLLPAQNATGNWIKVVQRLPVRIELDQKQLEQYPLRIGLSTLVSVNTTNRDGQVLANKVR  360
YzdF  B.subtilis    DMLPSTNSSGNYTKVTQKVPVKISIKNPSDKVLPGMNASVKISE                  126
                     .**. *..**. ** *..**.* ..       *

EmrA  H.influenzae  PKTLFSTDTLKYDESAVENLIESIIQQNSHD                               391
EmrA  E.coli        STPVAVSTAREISLAPVNKLIDDIVKANAG                               390
```

Fig. 4. Homology comparison of EmrA amino acid sequences (multidrug resistance protein A) from *H. influenzae* and *E. coli*, and the deduced protein products of *yzdE* and *yzdF*. Amino acid residues that are conserved between *E. coli* and *H. influenzae* are indicated in bold; amino acid residues that are identical in all three organisms are indicated with an asterisk below the three sequences; amino acid residues that are conserved are indicated with a dot.

glycine receptor beta subunits are 484, 496 and 497 aa long, respectively, and the similarity is restricted to three small blocks of amino acids. However, a proper SD sequence with accompanying start codon is present in front of this ORF (AAAGGAGGgagaaggTTG). Functional analysis will hopefully reveal the biological relevance of the above-mentioned features.

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.

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

Beall, B. & Moran, C. P., Jr (1994). Cloning and characterization of *spoVR*, a gene from *Bacillus subtilis* involved in spore cortex formation. *J Bacteriol* **176**, 2003–2012.

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

Biaudet, V., Samson, F., Anagnostopoulos, C., Ehrlich, S. D. & Bessières, Ph. (1996). Computerized genetic map of *Bacillus subtilis*. *Microbiology* **142**, 2669–2729.

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

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

Bult, C. J., White, O., Olsen, G. J. & 37 other authors (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273**, 1058–1073.

Cheng, S., Chang, S.-Y., Gravitt, P. & Respess, R. (1994). Long PCR. *Nature* **369**, 684–685.

Dear, S. & Staden, R. (1991). A sequence assembly and editing program for efficient management of large projects. *Nucleic Acids Res* **19**, 3907–3911.

Dedonder, R. A., Lepesant, J.-A., Lepesant-Kejzlarovà, J., Billault, A., Steinmetz, M. & Kunst, F. (1977). Construction of a kit of reference strains for rapid genetic mapping in *Bacillus subtilis* 168. *Appl Environ Microbiol* **33**, 989–993.

Fischer, C., Geourjon, C., Bourson, C. & Deutscher, J. (1996). Cloning and characterization of the *Bacillus subtilis prkA* gene encoding a novel serine protein kinase. *Gene* **168**, 55–60.

Fleischmann, R. D., Adams, M. D., White, O. & 37 other authors (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512.

Fraser, C. M., Gocayne, J. D., White, O. & 26 other authors (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403.

Goffeau, A., Barrell, B. G., Bussey, H. & 13 other authors (1996). Life with 6000 genes. *Science* **274**, 546–567.

Gowrishankar, J. (1989). Nucleotide sequence of the osmoregulatory *proU* operon of *Escherichia coli*. *J Bacteriol* **171**, 1923–1931.

Hansson, M. & Hederstedt, L. (1992). Cloning and characterization of the *Bacillus subtilis hemEHY* gene cluster, which encodes protoheme IX biosynthetic enzymes. *J Bacteriol* **174**, 8081–8093.

Harford, N., Lepesant-Kejzlarova, J., Lepesant, J.-A., Hamers, R. & Dedonder, R. (1976). Genetic circularity and mapping of the replication origin region of the *Bacillus subtilis* chromosome. In *Microbiology – 1976*, pp. 28–34. Edited by D. Schlesinger. Washington, DC: American Society for Microbiology.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.-C. & Herrmann, R. (1996). Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res* **24**, 4420–4449.

Holmberg, C., Beijer, L., Rutberg, B. & Rutberg, L. (1990). Glycerol catabolism in *Bacillus subtilis*: nucleotide sequence of the genes encoding glycerol kinase (*glpK*) and glycerol-3-phosphate dehydrogenase (*glpD*). *J Gen Microbiol* **136**, 2367–2375.

Hulett, F. M., Kim, E. E., Bookstein, C., Kapp, N. V., Edwards, C. W. & Wyckoff, H. W. (1991). *Bacillus subtilis* alkaline

phosphatase III and IV. Cloning, sequencing, and comparisons of deduced amino acid squence with *Escherichia coli* alkaline phosphatase three-dimensional structure. *J Biol Chem* **266**, 1077–1084.

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

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.

Jin, S. & Sonenshein, A. L. (1994a). Identification of two distinct *Bacillus subtilis* citrate synthase genes. *J Bacteriol* **176**, 4669–4679.

Jin, S. & Sonenshein, A. L. (1994b). Transcriptional regulation of *Bacillus subtilis* citrate synthase genes. *J Bacteriol* **176**, 4680–4690.

Kaneko, T., Sato, S., Kotani, H. & 21 other authors (1996). Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res* **3**, 109–136.

Kontinen, V. P., Saris, P. & Sarvas, M. (1991). A gene (*prsA*) of *Bacillus subtilis* involved in a novel late stage of protein export. *Mol Microbiol* **5**, 1273–1283.

Kunst, F., Ogasawara, N., Moszer, I. & 148 other authors (1997). The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249–256.

Lomovskaya, O. & Lewis, K. (1992). Emr, an *Escherichia coli* locus for multidrug resistance. *Proc Natl Acad Sci USA* **89**, 8938–8942.

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

Mewes, H. W., Albermann, K., Bahr, M. & 9 other authors (1997). Overview of the yeast genome. *Nature* **387**, 7–9.

Moszer, I., Glaser, P. & Danchin, A. (1995). SubtiList: a relational database for the *Bacillus subtilis* genome. *Microbiology* **141**, 261–268.

Noback, M. A., Terpstra, P., Holsappel, S., Venema, G. & Bron, S. (1996). A 22 kb DNA sequence in the *cspB–glpPFKD* region at 75 ° on the *Bacillus subtilis* chromosome. *Microbiology* **142**, 3021–3026.

O'Brien, C. (1997). Entire *E. coli* genome sequenced – at last. *Nature* **385**, 472.

Pearson, J. D., DeWald, D. B., Mathews, W. R. & 10 other authors (1990). Amino acid sequence and characterization of a protein inhibitor of protein kinase C. *J Biol Chem* **265**, 4583–4591.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* **85**, 2444–2448.

Perego, M. & Hoch, J. A. (1988). Sequence analysis and regulation of the *hpr* locus, a regulatory gene for protease production and sporulation in *Bacillus subtilis*. *J Bacteriol* **170**, 2560–2567.

Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: a Laboratory Manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.

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

Seraphin, B. (1992). The HIT protein family: a new family of proteins present in prokaryotes, yeast and mammals. *DNA Seq* **3**, 177–179.

Stahl, M. L. & Ferrari, E. (1984). Replacement of the *Bacillus subtilis* subtilisin structural gene with an in vitro-derived deletion mutation. *J Bacteriol* **158**, 411–418.

Stauffer, G. V. (1983). Regulation of serine, glycine, and one-carbon biosynthesis. In *Amino Acids: Biosynthesis and Genetic Regulation*, pp. 103–113. Edited by K. M. Herrman & R. L. Somerville. Reading, MA: Addison-Wesley.

Received 20 October 1997; revised 8 December 1997; accepted 12 December 1997.