Cloning and sequencing of a 40·6 kb segment in the 73°–76° region of the *Bacillus subtilis* chromosome containing genes for trehalose metabolism and acetoin utilization

Hiroki Yamamoto, Shigeki Uchiyama and Junichi Sekiguchi

Author for correspondence: Junichi Sekiguchi. Tel: +81 268 21 5344. Fax: +81 268 21 5331.
e-mail: jsekigu@giptc.shinshu-u.ac.jp

Department of Applied Biology, Faculty of Textile Science and Technology, Shinshu University, 3-15-1 Tokida, Ueda-shi, Nagano 386, Japan

In the framework of the international project aimed at the sequencing of the *Bacillus subtilis* genome, a 40·6 kb chromosome segment, which contains the *tre* locus, has been cloned and sequenced. This region (40601 bp; 73°–76° on the genetic map) contains 38 complete ORFs and one partial one. Three ORFs, the closest to the *hsdC* locus, correspond to the *treP*, *treA* and *treR* genes encoding enzyme II_{Tre}, trehalose-6-phosphate hydrolase and the repressor of the *tre* operon, respectively. A homology search for the products deduced from the 39 ORFs revealed that 23 exhibit significant similarity to known proteins, e.g. proteins involved in acetoin utilization, deoxyribonuclease, methyladenine glycosidase, hydroxyisobutyrate dehydrogenase, multidrug resistance proteins, protein phosphatase, cyclic-nucleotide phosphodiesterase, 5'-nucleotidase and NADP(H)-flavin oxidoreductase. Based on the gene organization and the results of the homology search, it is predicted that YfjG, YfjH, YfjI, YfjJ and YfjK form an acetoin dehydrogenase system (acetoin regulatory protein, and acetoin dehydrogenase components/subunits E3, E2, E1 β and E1 α , respectively). *yfkN*, an extremely large ORF comprising 4386 nucleotides, seems to correspond to the fusion of the genes for 2',3'-cyclic-nucleotide 2'-phosphodiesterase and 5'-nucleotidase precursor.

Keywords: *Bacillus subtilis*, genome sequencing, *tre* locus, acetoin utilization

INTRODUCTION

As participants in the international *Bacillus subtilis* genome project, we are responsible for sequencing an approximately 144 kb chromosomal region between 69° (NEXT11) and 81° (*sspE*) containing the *hsdC* and *tre* loci. In our first communication on this project (Yamamoto *et al.*, 1996), we reported the sequencing and genetic features of a 12·4 kb region, from the *NotI* site of NEXT23 (Itaya & Tanaka, 1991), containing the deduced operons of 6-phospho- β -glucosidase, the phosphoenolpyruvate:sugar phosphotransferase system and an ABC transport system. In this communication, we describe the sequencing and genetic features of a 40·6 kb region containing the *tre* locus (Helfert *et al.*, 1995; Schöck & Dahl, 1996). In particular, we describe a putative operon for the acetoin dehydrogenase enzyme system, and a novel fused gene

comprising the 2',3'-cyclic-nucleotide 2'-phosphodiesterase and 5'-nucleotidase precursor genes.

METHODS

Bacterial strains, phages, plasmids and growth media. *B. subtilis* AC327 (*purB his-1 smo-1*) (Sekiguchi *et al.*, 1988) was used as the source of chromosomal DNA to construct a λ library. pNEXT23 (Itaya & Tanaka, 1991) was also used as the source of DNA for sequencing and as a probe. *Escherichia coli* XL1-Blue MRA(P2) [Δ (*mcrA*)183 Δ (*mcrCB-hsdSMR-mrr*)173 *endA1 supE44 thi-1 gyrA96 relA1 lac* (P2 lysogen)] (Stratagene) and JM109 (*relA1 supE44 endA1 hsdR17 gryA96 relA1 thi* Δ (*lac-proAB*) F' [*traD36 proAB lac^r lacZ* Δ M15]) (Takara Shuzo) were used as cloning hosts for λ DASH II (Stratagene) and pUC118 (Takara Shuzo) clones, respectively. L-broth was used for the growth of *B. subtilis* and *E. coli*. If necessary, ampicillin (50 μ g ml⁻¹) was added to the medium.

Cloning and sequencing of a 40·6 kb region between 73° and 76°. A 5·4 kb insert of pNEXT23 was sequenced as described previously (Yamamoto *et al.*, 1996) and the sequence from the *NotI* site in the direction of NEXT11 (69°) is described in that

The GSD, DDBJ, EMBL and NCBI accession number for the nucleotide sequence reported in this paper is D83967.

paper. λ DASH II DNA digested with *Bam*HI was ligated to 12–23 kb fragments of *B. subtilis* chromosomal DNA partially digested with *Sau*3AI. The phage DNAs were packaged *in vitro*, according to the manufacturer's instructions (Stratagene) and then used for transduction of *E. coli*. Clones in the library were selected by plaque hybridization (Sambrook *et al.*, 1989) with 32 P-labelled pNEXT23 DNA as probe. The insert (16 kb) in the selected clone (phage 18) was partially digested with DNase I, blunt-ended with T4 polymerase and the fragments were separated by agarose (1.2%, w/v) gel electrophoresis to obtain 700–1500 bp fragments. These were ligated to *Sma*I-digested and alkaline-phosphatase-treated pUC118 DNA. *E. coli* JM109 cells were transformed with the ligation mixture to obtain a randomly overlapping library. To avoid rearrangements and reconfirm the sequence, we made primers ORF9-1R (a 38-mer: 5' CAGGAAACAGCTATGACCACAGTCTCAAACTGTCTCC 3'; the underlining indicates the M13RP sequence; the 3' end of the primer corresponds to position 10734 with respect to the *No*I site in Fig. 1) and N23-3F (a 38-mer: 5' TGTAAAACGACGGCCAGTCACTTCATGCTCCTCTTGCG 3'; the underlining indicates the –21M13 sequence; the 3' end corresponds to position 3877). We then performed PCR with phage 18 or *B. subtilis* chromosomal DNA as template.

Chromosome walking was performed by the inverse PCR method as follows. A 0.8 kb fragment, about 10 kb from the *No*I site of NEXT23, which had been labelled with the DIG DNA Labeling and Detection Kit (Boehringer Mannheim), was hybridized with a 13.4 kb fragment of *B. subtilis* chromosomal DNA digested with *Mlu*I. Inverse PCR with primers *Mlu*I-1R (a 37-mer: 5' CAGGAAACAGCTATGACCCTGACGGATTGCAGCTGG 3') and F-Uchi3 (a 38-mer: 5' TGTAAAACGACGGCCAGTCTTTGGTACATCCACAGCG 3') about 6.5 kb and 10.6 kb from the *No*I site of NEXT23, respectively, and with the self-ligated DNA of the 13.4 kb *Mlu*I fragment as template, allowed DNA amplification of a 9.3 kb fragment. After partial sequencing of the termini of the amplified fragment with the M13RP primer by the Dye Primer method (Applied Biosystems), we obtained primer MSI-R (a 38-mer: 5' CAGGAAACAGCTATGACCATCATTCCTTTCACCATCCG 3'). After Long Range PCR with primers F-Uchi3 and MSI-R, and the *B. subtilis* chromosome as template, the resultant 9.1 kb amplified product (*Mlu*I-inv) was used to construct a randomly overlapping library for sequencing.

Furthermore, Southern hybridization of *Apa*I-digested fragments of the *B. subtilis* chromosome was performed with a 0.7 kb fragment, as probe, from the terminal region of the cloned 9.1 kb fragment described above. The probe hybridized with a 22.9 kb *Apa*I fragment. Inverse PCR with primers R-*Apa*I-1 (a 38-mer: 5' CAGGAAACAGCTATGACCGTTCATCAGCGTTCCATACG 3') and F-Uchi4 (a 38-mer: 5' TGTAAAACGACGGCCAGTGTGAAGCAGGTAAATGGCTC 3') in the terminal 2.0 kb region of the 9.1 kb fragment, and with the self-ligated DNA of the 22.9 kb *Apa*I fragment as template, allowed DNA amplification of a 21.3 kb fragment. After partial sequencing of the termini of the amplified fragment with the M13RP primer using the Dye Primer method, we obtained primer ASI-R (38-mer, 5' CAGGAAACAGCTATGACCGAGAACTGACGGTCTGG 3'). After Long Range PCR with primers F-Uchi4 and ASI-R, and the *B. subtilis* chromosome as template, the resultant 21.2 kb amplified product (*Apa*I-inv) was used to construct a randomly overlapping library.

Computer analysis. The DNA sequences were determined using an Inherit Auto Assembler (Applied Biosystems), and further analysed for the location of possible ORFs using GENEWORKS (IntelliGenetics). The amino acid sequences of the

putative products of identified ORFs were examined as to their similarity to sequences reported in a non-redundant protein sequence data bank using the FASTA e-mail server at the Human Genome Center, Institute of Medical Science, University of Tokyo, Japan. According to the nomenclature proposed by A. Danchin (Institut Pasteur, Paris, France), we used *yjf* and *yjg* for ORFs whose functions are unknown.

RESULTS AND DISCUSSION

Cloning and physical mapping of the 73°–76° region of the *B. subtilis* genome

Cloning and sequencing were performed from the *No*I site of NEXT23 (76°) in the direction of the NEXT11 site (69°) (Itaya & Tanaka, 1991). pNEXT23 DNA was digested with *Pst*I and the resultant 5.4 kb fragment (the insert of *B. subtilis* CU741 DNA) was used for the construction of a randomly overlapping library, followed by sequencing. Moreover, λ phage clones containing a part of the insert of pNEXT23 were selected by plaque hybridization. Among 19 independent clones isolated, 2 (phages 18 and 1-1) were selected and their fragments were analysed with restriction enzymes. The digestion patterns indicated common 2.8 kb *Bam*HI and 3.3 kb *Eco*RI inserts. The inserts correspond to the regions from position 4121 (with respect to the *No*I site) to position 6903 and from 6184 to 9451, respectively, in Fig. 1. Phage 18 contained a 16 kb insert and one of its terminal ends is at position 3716 in Fig. 1. Then we sequenced the insert of phage 18. Downstream of position 9451, there was no overlapping fragment. To avoid rearrangements and reconfirm the sequence, we synthesized primer ORF9-1R (corresponding to positions 10753–10734) and primer N23-3F (3858–3877), followed by amplification with phage 18 DNA or *B. subtilis* chromosomal DNA as template. The two amplified DNAs gave identical fragment patterns with *Eco*RI, *Hind*III, *Pst*I and *Eco*RI + *Pst*I (data not shown). Based on the sequence information around 6.5 kb and 10.6 kb from the *No*I site of NEXT23, two differently oriented primers were made, and an inverse PCR product, *Mlu*I-inv (9.1 kb), was amplified as described in Methods. The inverse PCR method was further used to amplify a product, *Apa*I-inv, which covers the 21.2 kb region containing the *treA* locus (Fig. 1). Therefore, the phages and inverse PCR products covered a 40.6 kb region from the *No*I site of NEXT23, corresponding to the 73°–76° region of the *B. subtilis* chromosome (Fig. 1).

Sequencing of the 40.6 kb region between 73° and 76°, and genes found in the region and their features

The inserted fragments covering the 40.6 kb region were digested with DNase I, followed by ligation to blunt-ended pUC118, to construct a randomly overlapping library. Then inserted fragments in the pUC118 derivatives were amplified by PCR from the lysates of transformants, followed by sequencing with a *Taq* Dye Primer Cycle Sequencing Kit (Applied Biosystems).

Putative ORFs were searched for based on the criteria

Fig. 1. Physical map and gene organization of the 40.6 kb segment of the 73° (tre)–76° (NEXT23) region of the *B. subtilis* chromosome. Numbers over the thick line indicate distances with respect to the *NotI* site of NEXT23, and the direction is opposite to that of the movement of the replication fork. Only the sequenced regions for pNEXT23 and phage 18 are shown, and the PCR products for *MluI*-inv and *ApaI*-inv; they are fully described in the text. The 39 deduced ORFs, and their transcriptional and translational directions are indicated by arrows. The deduced *p*-independent terminators are denoted by: *1, $\Delta G = -21.4 \text{ kcal mol}^{-1}$; *2, -31.8 ; *3, -21.3 ; *4, -20.6 ; *5, -18.0 ; *6, -20.8 ; *7, -16.2 ; *8, -24.3 ; *9, -23.8 ; *10, -27.4 ; *11, -27.0 . B, E and P indicate *Bam*HI, *Eco*RI and *Pst*I sites, respectively. In the case of the 5' region of *yfjA* (broken arrow), results are also based on those of Yamamoto *et al.* (1996).

described previously (Yamamoto *et al.*, 1996), and 38 complete and one incomplete ORFs (the 3' region of *yfjA*) were identified (Table 1). Among them, 31 start with ATG, 4 with GTG and 4 with TTG. Sixteen ORFs are transcribed and translated in the same orientation from 76° to 73°, and the other 23 in the opposite orientation, which is identical to that of the movement of the replication fork (Fig. 1).

Functions of the five ORFs supposedly belonging to the acetoin dehydrogenase enzyme system (Table 2)

The amino acid sequences of the putative products of the identified ORFs were examined as to their similarity to sequences in a non-redundant protein sequence data bank. The sequences of YfjG, YfjH, YfjI, YfjJ and YfjK exhibit extremely high similarity to those of the regulatory

proteins of the acetoin dehydrogenase system of *Clostridium magnum* and *E. coli* (AcoR and AtoC, respectively), component E3 of the acetoin dehydrogenase from *C. magnum* (AcoL), component E2 from *C. magnum* (AcoC), and subunits E1 β and E1 α from various bacteria, respectively. The acetoin dehydrogenase enzyme system consists of acetoin dehydrogenase subunits (E1 α and E1 β), dihydrolipoamide acetyltransferase (E2), and dihydrolipoamide dehydrogenase (E3), and the gene order of those of *Pelobacter carbinolicus* and *C. magnum* are *acoA* (encoding E1 α)-*acoB* (E1 β)-*acoC* (E2)-*acoS*-*acoL* (E3) and *acoR*-*acoA* (E1 α)-*acoB* (E1 β)-*acoX*-*acoC* (E2)-*acoL* (E3), respectively. Judging from the gene organization of *yfjK*-*yfjJ*-*yfjI*-*yfjH*-*yfjG* showing sequence similarity to *acoA*-*acoB*-*acoC*-*acoL*-*acoR*, they probably form an acetoin dehydrogenase enzyme system (Fig. 1 and Table 1). Moreover, these five ORFs are transcribed and translated

Table 1. Putative ORFs in the 40601 bp sequence from the *NorI* site of NEXT23

ORF	Endpoints (nt)	Size of product (aa/kDa)	SD consensus sequence (upper case) and initial codon (bold)
<i>yjfA</i> *	→0 > 264	104/12.0	gaAAAGGcaGgGgaCaaaagttg
<i>yjfB</i>	273 > 1493	407/43.7	ActcgGGAGGTGAaCggtgtg
<i>yjfC</i>	1508 > 2272	255/29.2	gGAtAGGAGaatAaacaatg
<i>yjfD</i>	2307 > 2861	185/21.7	caAAAGGAGaaaAgCattg
<i>yjfE</i>	2958 > 3413	152/17.7	ttgtgGGAGtctggCtggtg
<i>yjfF</i>	3466 > 3792	109/11.9	caAAAtGAGGaGtgCrcatg
<i>yjfG</i>	5881 > 4067	605/67.2	gtAAgGGAGGaGcgatcaaggaaatg
<i>yjfH</i>	7373 > 6000	458/48.9	gaAAAGcAGGTGAaaacgacatg
<i>yjfI</i>	8590 > 7397	398/42.9	ttAAAGGgGGgacaCaaatg
<i>yjfJ</i>	9632 > 8607	342/36.8	AaAAAGGAGGaatgtaagagatg
<i>yjfK</i>	10637 > 9639	333/36.1	tagAAGGAGGcGcaCaaaatg
<i>yjfL</i>	11558 > 10872	229/25.3	AcAtAtGAGGaaAagagtgggaacatg
<i>yjfM</i>	12040 > 11591	150/17.2	gGgggaGAGGaaAatg
<i>yjfN</i>	13213 > 12239	325/37.0	tGAAAGGAtTGATttttgttatg
<i>yjfO</i>	16237 > 14840	466/52.1	gaAAcGGAGagaAcaacgtg
<i>yjfP</i>	17214 > 16354	287/33.2	AaAAAGGgtGgatctcatg
<i>yjfQ</i>	17333 > 18289	319/37.8	caggtGGAGGgtgttttatg
<i>yjfR</i>	18466 > 19248	261/27.9	ttcAAGcAGGctAcaaatg
<i>yjfS</i>	20080 > 19292	263/30.1	cagAAGGAGcgtgtgccaatg
<i>yjfT</i>	20366 > 20184	61/7.3	tGAAtGGAGGaacagaatg
<i>yfkA</i>	20511 > 20972	154/17.8	AtAAAGGAGtTagatagtcatg
<i>yfkB</i>	21170 > 21628	153/17.5	cGtcAcGAGGTccaCccaatg
<i>yfkC</i>	21632 > 22471	280/32.2	AGccAGGAGtTagatg
<i>yfkD</i>	23307 > 22516	264/29.6	AGAAAGagtGatag
<i>yfkE</i>	24433 > 23381	351/37.5	AaAAAGGAGcTatTtcaccatg
<i>yfkF</i>	24601 > 25773	391/42.7	tatAAGGAGaactgacacgatg
<i>yfkG</i>	25813 > 25947	45/5.0	tGtcAGGAGGcttTtatg
<i>yfkH</i>	26802 > 25978	275/30.5	AcAtgcGAGGTGagacacatg
<i>yfkI</i>	27146 > 26829	106/12.1	tGAAAGGAGaattTgcatg
<i>yfkJ</i>	27634 > 27167	156/17.3	AatAgGGAGGTgcatcatg
<i>yfkK</i>	27837 > 28049	71/8.1	AGAAAGGAtGgaAcCctttg
<i>yfkL</i>	28145 > 29332	396/42.2	ttgAAGGAGtacAagcatg
<i>yfkM</i>	29893 > 29378	172/18.9	AagcgGGAGGTtATgaatcatg
<i>yfkN</i>	30136 > 34521	1462/159.7	AaggtGGAGaactggtg
<i>yfkO</i>	35226 > 34564	221/25.6	ctAAAGGAGagGtggttttacatg
<i>yfkP</i>	35369 > 35217	51/6.0	tGcgcGGcGGgGgaaatg
<i>treR</i>	36082 > 35369	238/27.8	AGctgGGtGGTccataatg
<i>treA</i>	37788 > 36106	561/65.2	AtggtGGtGGaaAcagatg
<i>treP</i>	39271 > 37862	470/50.0	gatAcGGAGGgGtTggcatg

* The *NorI* site of NEXT23 is located in the middle of the *yjfA* gene. The sequence of the 5' terminal region of *yjfA* is based on Yamamoto *et al.* (1996).

in the same direction, and the three deduced *p*-independent terminators are located between *yjfL* and *yjfK*, *yjfH* and *yjfG*, and *yjfG* and *yjfF* (Fig. 1). Therefore, *yjfK*, *yjfJ*, *yjfI* and *yjfH* very likely form an operon (and we therefore designated them as the *acoABCL* operon). The lipoyl domains in the E2 components of *E. coli*, *P. carbinolicus*, *Alcaligenes eutrophus* and *C. magnum*, and in E3 of *C. magnum* are also found in the N-terminal region of the E2 component (YfjI) of *B. subtilis* (Krüger *et al.*, 1994; data not shown). Recently, the *acuABC* genes involved in acetoin utilization were reported in *B. subtilis* (Grundy *et al.*, 1993). However, these genes exhibit no similarity to

the *acoXABC* genes of *A. eutrophus*, or the putative *acoABCL* (*yjfKJIH*) genes of *B. subtilis* (data not shown). These results support the previous findings that *B. subtilis* has two different pathways for acetoin utilization (Lopez *et al.*, 1975; Grundy *et al.*, 1993; Krüger *et al.*, 1994).

Homology of the other ORFs in the 73°–76° region (Table 2)

The genes for two DNA hydrolysing/repair enzymes (*yjfO* and *yjfP*) were found upstream of *yjfN* in the same orientation (Fig. 1). The amino acid sequences of YfjO

Table 2. Similarity of predicted ORF products to known proteins

Product	Similar protein in database*	Identity observed	FASTA score	Reference
YfjF	Hypothetical protein of <i>Synechocystis</i> sp. [GP: SYCSLRG_45, GI: 1001824]	52.3 % in 109 residues	369	
YfjG	Acetoin dehydrogenase (<i>acoR</i>) of <i>Clostridium magnum</i> [PRF: 2014293A]	40.6 % in the C-terminal region 320 residues	639	Krüger <i>et al.</i> (1994)
	<i>E. coli</i> chromosomal region [GP: ECU28375_19, GI: 887819]	42.1 % in the C-terminal region 323 residues	635	
	Acetoacetate metabolism regulatory protein (<i>atoC</i>) of <i>E. coli</i> [SP: ATOC_ECOLI]	36.9 % in the C-terminal region 388 residues	602	
	Signal transduction and transcriptional control protein (fragment) of <i>Clostridium beijerinckii</i> [SP: STC_CLOBE]	42.8 % in the C-terminal region 318 residues	601	
YfjH	Acetoin dehydrogenase (<i>acoL</i>) of <i>Clostridium magnum</i> [PRF: 2014293F]	40.6 % in 458 residues	927	Krüger <i>et al.</i> (1994)
	Lipoamide dehydrogenase component E3 of <i>B. subtilis</i> [SP: DLDH_BACSU]	37.9 % in 457 residues	815	Hemilä <i>et al.</i> (1990)
	Lipoamide dehydrogenase component E3 of <i>B. stearothermophilus</i> [SP: DLDH_BACST]	36.6 % in 456 residues	811	Borges <i>et al.</i> (1990)
	Lipoamide dehydrogenase component E3 of <i>Staphylococcus aureus</i> [PIR: S19723]	35.4 % in 457 residues	788	Hemilä (1991)
YfjI	Dihydrolipoamide succinyltransferase component E2 (<i>sucB</i>) of <i>Coxiella burnetii</i> [GP: CBU07789_2, GI: 641969]	31.0 % in 403 residues	652	
	<i>tufA</i> gene product of <i>Neisseria gonorrhoeae</i> [GP: NGOFUSAA_2, GI: 790863]	33.3 % in 402 residues	612	
	Dihydrolipoamide succinyltransferase E2 (<i>sucB</i>) of <i>E. coli</i> [SP: ODO2_ECOLI]	29.7 % in 407 residues	572	Spencer <i>et al.</i> (1984)
	Acetoin dehydrogenase (<i>acoC</i>) of <i>Clostridium magnum</i> [PRF: 2014293E]	33.4 % in 377 residues	499	Krüger <i>et al.</i> (1994)
YfjJ	Acetoin dichlorophenolindophenol oxidoreductase (<i>acoB</i>) of <i>Alcaligenes eutrophus</i> [SP: ACOB_ALCEU]	58.0 % in 331 residues	977	Priefert <i>et al.</i> (1991)
	Acetoin dichlorophenolindophenol oxidoreductase (<i>acoB</i>) of <i>Pelobacter carbinolicus</i> [PRF: 2004329B]	56.2 % in 329 residues	976	Oppermann & Steinbüchel (1994)
	<i>acoB</i> gene product of <i>Klebsiella pneumoniae</i> [PRF: 2014286B]	51.6 % in 337 residues	902	Deng <i>et al.</i> (1994)
	Acetoin dehydrogenase subunit (<i>acoB</i>) of <i>Pseudomonas putida</i> [PRF: 2104227C]	54.0 % in 339 residues	874	
	Acetoin dehydrogenase subunit (<i>acoB</i>) of <i>Clostridium magnum</i> [PRF: 2014293C]	49.4 % in 338 residues	739	Krüger <i>et al.</i> (1994)
YfjK	Acetoin dichlorophenolindophenol oxidoreductase (<i>acoA</i>) of <i>Alcaligenes eutrophus</i> [SP: ACOA_ALCEU]	56.7 % in 323 residues	917	Priefert <i>et al.</i> (1991)
	Acetoin dehydrogenase subunit (<i>acoA</i>) of <i>Clostridium magnum</i> [PRF: 2014293B]	54.6 % in 326 residues	914	Krüger <i>et al.</i> (1994)
	Acetoin dichlorophenolindophenol oxidoreductase (<i>acoA</i>) of <i>Pelobacter carbinolicus</i> [PRF: 2004329A]	57.5 % in 322 residues	909	Oppermann & Steinbüchel (1994)
	Acetoin dehydrogenase subunit (<i>acoA</i>) of <i>Pseudomonas putida</i> [PRF: 2104227B]	58.8 % in 320 residues	902	
	<i>acoA</i> gene product of <i>Klebsiella pneumoniae</i> [PRF: 2014286A]	54.1 % in 318 residues	882	Deng <i>et al.</i> (1994)

Table 2 (cont.)

Product	Similar protein in database*	Identity observed	FASTA score	Reference
YfjN	Hypothetical protein of <i>Synechocystis</i> sp. [GP: SYCSLRD_22, GI: 1001634]	28.4 % in 310 residues	294	Machado <i>et al.</i> (1995)
	Hypothetical protein of <i>B. subtilis</i> [SP: YACF_BACSU]	27.1 % in 314 residues	273	
	<i>nifR3</i> protein of <i>Azospirillum brasilense</i> [PIR: S49971]	26.9 % in 312 residues	271	
YfjO	Endo-exodeoxyribonuclease of <i>Saccharomyces cerevisiae</i> [SP: RNC1_YEAST]	25.2 % in 412 residues	294	Van Vliet-Reedijk & Planta (1993)
	Hypothetical protein of <i>Haemophilus influenzae</i> [SP: Y958_HAEIN]	25.3 % in 396 residues	363	Fleischmann <i>et al.</i> (1995)
	<i>relA</i> -upstream ORF A of <i>Vibrio</i> sp. [PRF: 2020349A]	31.9 % in the C-terminal region 188 residues	306	
	<i>E. coli</i> chromosomal region [GP: ECU29580_10, GI: 882679]	32.2 % in the C-terminal region 177 residues	291	
YfjP	DNA-3-methyladenine glycosylase (<i>alkA</i>) of <i>B. subtilis</i> [SP: 3MGA_BACSU]	29.7 % in 291 residues	403	Morohoshi <i>et al.</i> (1993)
YfjQ	Hypothetical protein of <i>Synechocystis</i> sp. [GP: SYCSLLH_50, GI: 1001341]	34.0 % in 297 residues	600	
	Hypothetical protein of <i>Synechocystis</i> sp. [GP: SYCSLRG_80, GI: 1006592]	30.6 % in 310 residues	494	
	<i>corA</i> gene product of <i>Mycobacterium leprae</i> [GP: MLU15180_39]	30.1 % in 296 residues	443	
YfjR	3-Hydroxyisobutyrate dehydrogenase (<i>mmsB</i>) of <i>Pseudomonas aeruginosa</i> [SP: MMSB_PSEAE]	31.1 % in 267 residues	330	Steele <i>et al.</i> (1992)
	Hypothetical protein of <i>E. coli</i> [SP: YHAE_ECOLI]	27.2 % in 257 residues	293	Rougraff <i>et al.</i> (1989)
	Hypothetical protein of <i>E. coli</i> [SP: YIHU_ECOLI]	25.9 % in 263 residues	285	
	3-Hydroxyisobutyrate dehydrogenase of <i>Rattus norvegicus</i> [SP: D3HL_RAT]	24.5 % in 269 residues	230	
YfjS	Nodulation protein homologue of <i>B. stearothermophilus</i> [PIR: B47692]	59.8 % in 256 residues	876	Reaney <i>et al.</i> (1993)
	Nodulation protein B (<i>nodB</i>) of <i>Rhizobium leguminosarum</i> [SP: NODB_RHILV]	34.7 % in 199 residues	327	Rossen <i>et al.</i> (1984)
YfkE	Hypothetical protein of <i>Saccharomyces cerevisiae</i> [PIR: S51293]	34.3 % in the N-terminal region 137 residues	216	Ivey <i>et al.</i> (1993)
	Calcium/protein antiporter of <i>E. coli</i> [SP: CHAA_ECOLI]	19.8 % in 354 residues	197	
YfkF	Multidrug resistance protein (<i>bmr2</i>) of <i>B. subtilis</i> [SP: BMR2_BACSU]	20.9 % in 378 residues	318	Ahmed <i>et al.</i> (1995)
	Multidrug resistance protein (<i>bmr1</i>) of <i>B. subtilis</i> [SP: BMR1_BACSU]	22.0 % in 368 residues	308	Ahmed <i>et al.</i> (1994)
	Quinolone resistance <i>norA</i> protein of <i>Staphylococcus aureus</i> [SP: NORA_STAAU]	19.5 % in 379 residues	252	
	Hypothetical protein of <i>E. coli</i> [SP: YCAD_ECOLI]	21.0 % in 271 residues	229	
YfkH	Hypothetical protein of <i>Mycobacterium leprae</i> [GP: MLU15181_4, GI: 699178]	27.1 % in 269 residues	261	Fleischmann <i>et al.</i> (1995)
	Hypothetical protein of <i>Haemophilus influenzae</i> [SP: YIHY_HAEIN]	22.5 % in 262 residues	223	
	Hypothetical protein of <i>E. coli</i> [SP: YIHY_ECOLI]	24.2 % in 273 residues	221	Fernandez & Weiss (1994)
	Virulence factor of <i>Bordetella pertussis</i> [PRF: 2102300A]	19.1 % in 273 residues	213	
YfkJ	Protein-tyrosine phosphatase of <i>Schizosaccharomyces pombe</i> [PIR: A55446]	40.8 % in 152 residues	306	Mondesert <i>et al.</i> (1994)
	Hypothetical protein of <i>Synechocystis</i> sp. [GP: SYCSLRA_12, GI: 1001408]	40.6 % in 155 residues	298	

Table 2 (cont.)

Product	Similar protein in database*	Identity observed	FASTA score	Reference
YfkL	Protein phosphatase 1 of <i>Rattus norvegicus</i> [PRF: 1817363A]	39.2 % in 153 residues	286	Manao <i>et al.</i> (1992)
	Acid phosphatase Af of <i>Homo sapiens</i> [PRF: 1814272A]	39.9 % in 153 residues	286	Dissing & Johnsen (1992)
	Multidrug resistance protein (<i>bmr1</i>) of <i>B. subtilis</i> [SP: BMR1_BACSU]	21.0 % in 376 residues	239	Ahmed <i>et al.</i> (1994)
	Multidrug resistance protein (<i>bmr2</i>) of <i>B. subtilis</i> [SP: BMR2_BACSU]	19.7 % in 375 residues	213	Ahmed <i>et al.</i> (1995)
YfkM	<i>E. coli</i> chromosomal region [GP: ECOUW67_288, GI: 606298]	18.0 % in 377 residues	212	
	<i>E. coli</i> chromosomal region [GP: ECOUW67_83, GI: 606093]	64.1 % in 170 residues	540	
YfkN	<i>lytA</i> -downstream ORF of <i>Staphylococcus aureus</i> [PRF: 2005198A]	48.6 % in 173 residues	416	Borchardt <i>et al.</i> (1993)
	2', 3'-Cyclic-nucleotide 2'-phosphodiesterase of <i>Haemophilus influenzae</i> [SP: CN16_HAEIN]	45.0 % in the N-terminal region 647 residues	1445	Fleishmann <i>et al.</i> (1995)
	2', 3'-Cyclic-nucleotide 2'-phosphodiesterase of <i>E. coli</i> [SP: CN16_ECOLI]	44.2 % in the N-terminal region 634 residues	1429	
	2', 3'-Cyclic-nucleotide 2'-phosphodiesterase of <i>Yersinia enterocolitica</i> [PIR: S52695]	43.0 % in the N-terminal region 611 residues	1399	Liu <i>et al.</i> (1986)
	2', 3'-Cyclic-nucleotide 2'-phosphodiesterase of <i>Salmonella typhimurium</i> (fragment) [PIR: S11915]	43.8 % in the N-terminal region 240 residues	605	
	5'-Nucleotidase precursor of <i>Rattus norvegicus</i> [SP: 5NTD_RAT]	26.8 % in the C-terminal region 527 residues	466	Misumi <i>et al.</i> (1990a)
	5'-Nucleotidase precursor of <i>Homo sapiens</i> [SP: 5NTD_HUMAN]	26.1 % in the C-terminal region 522 residues	457	Misumi <i>et al.</i> (1990b)
	5'-Nucleotidase precursor of <i>Bos primigenius taurus</i> [SP: 5NTD_BOVIN]	26.4 % in the C-terminal region 507 residues	452	Suzuki <i>et al.</i> (1993)
	5'-Nucleotidase precursor of <i>Discopyge ommata</i> [SP: 5NTD_DISOM]	24.2 % in the C-terminal region 509 residues	412	Volkandt <i>et al.</i> (1991)
	NAD(P)H-flavin oxidoreductase homologue of <i>Haemophilus influenzae</i> [PIR: B64114]	48.2 % in 222 residues	634	Fleishmann <i>et al.</i> (1995)
TreR	TreR of <i>B. subtilis</i> [GP: BSTREAPR_3, GI: 1000453]	100 % in 238 residues	1206	Schöck & Dahl (1996)
TreA	TreA of <i>B. subtilis</i> [SP: TREC_BACSU, GI: 1000452]	97.3 % in 547 residues	2767	Helfelt <i>et al.</i> (1995)
TreP	TreP of <i>B. subtilis</i> [GP: BSTREAPR_1, GI: 1000451]	99.4 % in 470 residues	2270	Schöck & Dahl (1996)

* GI, NCBI; GP, GenBank Protein; SP, SwissProt.

and YfjP exhibit high similarity to those of the endonucleoxyribonuclease of *Saccharomyces cerevisiae* and 3-methyladenine glycosylase of *B. subtilis*, respectively. The deduced amino acid sequences of *yfkF* and *yfkL*, located approximately 4 and 8 kb upstream of *yfiS*, respectively, exhibit 20–22 % identity to those of the multidrug-efflux transporter genes, *bmr* and *blt*, sharing 51 % sequence identity (Table 2). YfkF and YfkL share 21 % sequence

identity over 319 aa and have multiple transmembrane regions (data not shown). They may use a transmembrane proton gradient as an energy source, as predicted for Bmr and Blt. Upstream of *yfkF*, an ORF (*yfkE*) exhibits deduced amino acid sequence similarity to a hypothetical protein of *S. cerevisiae* and a calcium/proton antiporter of *E. coli*. The antiporter, whose overexpression leads to resistance to growth inhibition by both calcium

and sodium, is an extremely hydrophobic protein with 11 membrane-spanning regions, and YfkE is also an extremely hydrophobic protein with multiple transmembrane-spanning regions (data not shown). YfkJ exhibits high sequence similarity to the protein-tyrosine phosphatase of *Schizosaccharomyces pombe*, protein phosphatase 1 of *Rattus norvegicus* and acid phosphatase Af of *Homo sapiens*. A protein sequence motif, Cys-X5-Arg, which is present in the active sites of all known types of protein-tyrosine phosphatases, is also conserved in YfkJ.

yfkN, which is an extremely large ORF comprising 4386 nucleotides, seems to be a fused gene (Fig. 1 and Table 2). The amino acid sequence of the 5'-terminal region of *yfkN* exhibits extremely high similarity to the 2', 3'-cyclic-nucleotide 2'-phosphodiesterases of *Haemophilus influenzae* (CpdB), *E. coli* (CpdB) and *Yersinia enterocolitica* (CpdB). 2', 3'-Cyclic-nucleotide 2'-phosphodiesterase is a bifunctional enzyme that catalyses two consecutive reactions, converting a 2', 3'-cyclic-nucleotide to a 3' nucleotide and then the 3' nucleotide to a nucleic acid and phosphate. Moreover, the amino acid sequence of the 3'-terminal region of *yfkN* exhibits high similarity to those of 5'-nucleotidase precursors of *Rattus norvegicus*, *Homo sapiens*, *Bos primigenius taurus* and *Discopyge ommata*. Although the N-terminal and C-terminal regions of YfkN may be separated on processing at the post-translational stage, as far as we know, this type of gene fusion is the first example in a living organism.

Three ORFs, the closest to the *hsdC* locus, were found to be the *treP-treA-treR* genes. *treP*, *treA* and *treR* encode enzyme IITre, trehalose-6-phosphate hydrolase and the repressor of the *tre* operon, respectively (Helfert *et al.*, 1995; Schöck & Dahl, 1996; Table 2). The deduced amino acid sequence of *treR* was completely identical to that reported, but those of *treP* and *treA* were not (Helfert *et al.*, 1995; Schöck & Dahl, 1996; Table 2). We consider that this discrepancy is not due to sequencing error, because we always used more than three independent inverse PCR products to avoid PCR error, followed by the construction of a randomly overlapping library. We determined both sequences using the results for only regions of well-separated fluorescence peaks. For TreP, Phe at position 140 and Ala at position 465 in our data were different from Ser and Gly, respectively, at identical positions in the previous report. For TreA, there are several differences between our sequence and the reported one. A strain difference (AC327 and 168) may be one of the reasons for them.

Among the 39 ORFs in the 40.6 kb region, 16 did not exhibit any significant amino acid sequence similarity to ones in the non-redundant protein sequence data bank. Therefore, our research is now directed toward the isolation of gene-disrupted mutants in a systematic way under the international co-operative project.

ACKNOWLEDGEMENTS

We would like to thank M. Itaya for providing the *NotI* linking clone, pNEXT23, and Mr Aji N. Fajar for his technical

assistance. This research was supported by a Grant-in-Aid for Creative Research (Human Genome Project) from the Ministry of Education, Science and Culture of Japan.

REFERENCES

- Ahmed, M., Borsch, C. M., Taylor, S. S., Vázquez-Laslop, N. & Neyfakh, A. A. (1994). A protein that activates expression of a multidrug efflux transporter upon binding the transporter substrates. *J Biol Chem* **269**, 28506–28513.
- Ahmed, M., Lyass, L., Markham, P. N., Taylor, S. S., Vázquez-Laslop, N. & Neyfakh, A. A. (1995). Two highly similar multidrug transporters of *Bacillus subtilis* whose expression is differentially regulated. *J Bacteriol* **177**, 3904–3910.
- Borchardt, S. A., Babwah, A. V. & Jayaswal, R. K. (1993). Sequence analysis of the region downstream from a peptidoglycan hydrolase-encoding gene from *Staphylococcus aureus* NCTC8325. *Gene* **137**, 253–258.
- Borges, A., Hawkins, C. F., Packman, L. C. & Perham, R. N. (1990). Cloning and sequence analysis of the genes encoding the dihydro-lipoamide acetyltransferase and dihydrolipoamide dehydrogenase components of the pyruvate dehydrogenase multienzyme complex of *Bacillus stearothermophilus*. *Eur J Biochem* **194**, 95–102.
- Deng, W. L., Chang, H. Y. & Peng, H. L. (1994). Acetoin catabolic system of *Klebsiella pneumoniae* CG43: sequence, expression and organization of the *aco* operon. *J Bacteriol* **176**, 3527–3535.
- Dissing, J. & Johnsen, A. H. (1992). Human red cell acid phosphatase (ACP1): the primary structure of the two pairs of isozymes encoded by the ACP1*A and ACP1*C alleles. *Biochim Biophys Acta* **1121**, 261–268.
- Fernandez, R. C. & Weiss, A. A. (1994). Cloning and sequencing of a *Bordetella pertussis* serum resistance locus. *Infect Immun* **62**, 4727–4738.
- Fleischmann, R. D. and others (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512.
- Grundy, F. J., Waters, D. A., Takova, T. Y. & Henkin, T. M. (1993). Identification of genes involved in utilization of acetate and acetoin in *Bacillus subtilis*. *Mol Microbiol* **10**, 259–271.
- Helfert, C., Gotsche, S. & Dahl, M. K. (1995). Cleavage of trehalose-phosphate in *Bacillus subtilis* is catalysed by a phospho- α -(1-1)-glucosidase encoded by the *treA* gene. *Mol Microbiol* **16**, 111–120.
- Hemilä, H. (1991). Lipoamide dehydrogenase of *Staphylococcus aureus*: nucleotide sequence and sequence analysis. *Biochim Biophys Acta* **1129**, 119–123.
- Hemilä, H., Palva, A., Paulin, L., Arvidson, S. & Palva, I. (1990). Secretory S complex of *Bacillus subtilis*: sequence analysis and identity to pyruvate dehydrogenase. *J Bacteriol* **172**, 5052–5063.
- 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.
- Ivey, D. M., Guffanti, A. A., Zemsky, J., Pinner, E., Karpel, R., Padan, E., Schuldiner, S. & Krulwich, T. A. (1993). Cloning and characterization of a putative $\text{Ca}^{2+}/\text{H}^{+}$ antiporter gene from *Escherichia coli* upon functional complementation of $\text{Na}^{+}/\text{H}^{+}$ antiporter-deficient strains by the overexpressed gene. *J Biol Chem* **268**, 11296–11303.
- Krüger, N., Oppermann, F. B., Lorenzl, H. & Steinbüchel, A. (1994). Biochemical and molecular characterization of the *Clostridium magnum* acetoin dehydrogenase enzyme system. *J Bacteriol* **176**, 3614–3630.

- Liu, J., Burns, D. M. & Beacham, I. R. (1986). Isolation and sequence analysis of the gene (*cpdB*) encoding periplasmic 2', 3'-cyclic phosphodiesterase. *J Bacteriol* **165**, 1002–1010.
- Lopez, J. M., Thoms, B. & Rehbein, H. (1975). Acetoin degradation in *Bacillus subtilis* by direct oxidative cleavage. *Eur J Biochem* **57**, 425–430.
- Machado, H. B., Yates, M. G., Funayama, S., Rigo, L. U., Steffens, M. B. R., Souza, E. M. & Pedrosa, F. O. (1995). The *ntfBC* genes of *Azospirillum brasilense* are part of a *nifR* 3-like-*ntfB-ntfC* operon and are negatively regulated. *Can J Microbiol* **41**, 674–684.
- Manao, G., Pazzagli, L., Cirri, P., Caselli, A., Camici, G., Cappugi, G., Saeed, A. & Ramponi, G. (1992). Rat liver low M_r phosphotyrosine protein phosphatase isoenzymes: purification and amino acid sequences. *J Protein Chem* **11**, 333–345.
- Misumi, Y., Ogata, S., Hirose, S. & Ikehara, Y. (1990a). Primary structure of rat liver 5'-nucleotidase deduced from the cDNA. Presence of the COOH-terminal hydrophobic domain for possible post-translational modification by glycopospholipid. *J Biol Chem* **265**, 2178–2183.
- Misumi, Y., Ogata, S., Ohkubo, K., Hirose, S. & Ikehara, Y. (1990b). Primary structure of human placental 5'-nucleotidase and identification of the glycolipid anchor in the mature form. *Eur J Biochem* **191**, 563–569.
- Mondesert, O., Moreno, S. & Russell, P. (1994). Low molecular weight protein-tyrosine phosphatases are highly conserved between fission yeast and man. *J Biol Chem* **269**, 27996–27999.
- Morohoshi, F., Hayashi, K. & Munakata, N. (1993). *Bacillus subtilis alkA* gene encoding inducible 3-methyladenine DNA glycosylase is adjacent to the *ada* operon. *J Bacteriol* **175**, 6010–6017.
- Oppermann, F. B. & Steinbüchel, A. (1994). Identification and molecular characterization of the *aco* genes encoding the *Pelobacter carbinolicus* acetoin dehydrogenase enzyme system. *J Bacteriol* **176**, 469–485.
- Priefert, H., Hein, S., Krüger, N., Zeh, K., Schmidt, B. & Steinbüchel, A. (1991). Identification and molecular characterization of the *Alcaligenes eutrophus* H16 *aco* operon genes involved in acetoin catabolism. *J Bacteriol* **173**, 4056–4071.
- Reaney, S. K., Bungard, S. J. & Guest, J. R. (1993). Molecular and enzymological evidence for two classes of fumarase in *Bacillus stearothermophilus* (var. *non-diastaticus*). *J Gen Microbiol* **139**, 403–416.
- Rossen, L., Johnston, A. W. B. & Downie, J. A. (1984). DNA sequence of the *Rhizobium leguminosarum* nodulation genes *nodAB* and *C* required for root hair curling. *Nucleic Acids Res* **12**, 9497–9508.
- Rougraff, P. M., Zhang, B., Kuntz, M. J., Harris, R. A. & Crabb, D. W. (1989). Cloning and sequence analysis of a cDNA for 3-hydroxyisobutyrate dehydrogenase. Evidence for its evolutionary relationship to other pyridine nucleotide-dependent dehydrogenases. *J Biol Chem* **264**, 5899–5903.
- Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: a Laboratory Manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.
- Schöck, F. & Dahl, M. K. (1996). Analysis of DNA flanking the *treA* gene of *Bacillus subtilis* reveals genes encoding a putative specific enzyme II^{Tr} and a potential regulator of the trehalose operon. *Gene* (in press).
- Sekiguchi, J., Ezaki, B., Kodama, K. & Akamatsu, T. (1988). Molecular cloning of a gene affecting the autolysin level and flagellation in *Bacillus subtilis*. *J Gen Microbiol* **134**, 1611–1621.
- Spencer, M. E., Darlison, M. G., Stephens, P. E., Duckenfield, I. K. & Guest, J. R. (1984). Nucleotide sequence of the *sucB* gene encoding the dihydrolipoamide succinyltransferase of *Escherichia coli* K12 and homology with the corresponding acetyltransferase. *Eur J Biochem* **141**, 361–374.
- Steele, M. I., Lorenz, D., Hatter, K., Park, A. & Sokatch, J. R. (1992). Characterization of the *mmsAB* operon of *Pseudomonas aeruginosa* PAO encoding methylmalonate-semialdehyde dehydrogenase and 3-hydroxyisobutyrate dehydrogenase. *J Biol Chem* **267**, 13585–13592.
- Suzuki, K., Furukawa, Y., Tamura, H., Ejiri, N., Suematsu, H., Taguchi, R., Nakamura, S., Suzuki, Y. & Ikezawa, H. (1993). Purification and cDNA cloning of bovine liver 5'-nucleotidase, a GPI-anchored protein and its expression in COS cells. *J Biochem* **113**, 607–613.
- Van Vliet-Reedijk, J. C. & Planta, R. J. (1993). The RHO4a and NUDI genes on *Saccharomyces cerevisiae* chromosome XI. *Yeast* **9**, 1139–1147.
- Volkandt, W., Vogel, M., Pevsner, J., Misumi, Y., Ikehara, Y. & Zimmermann, H. (1991). 5'-Nucleotidase from the electric ray electric lobe. Primary structure and relation to mammalian and procaryotic enzymes. *Eur J Biochem* **202**, 855–861.
- Yamamoto, H., Uchiyama, S., Fajar, A. N., Ogasawara, N. & Sekiguchi, J. (1996). Determination of a 12 kb nucleotide sequence around the 76° region of the *Bacillus subtilis* chromosome. *Microbiology* **142**, 1417–1421.

Received 18 April 1996; revised 13 June 1996; accepted 24 June 1996.