Determination of a 21548 bp nucleotide sequence around the 24° region of the *Bacillus subtilis* chromosome

Kou-ichiro Ogawa, Eiji Akagawa, Kouji Nakamura and Kunio Yamane

Author for correspondence: Kunio Yamane. Tel./Fax: +81 298 53 6680.

Institute of Biological Sciences, University of Tsukuba, Tsukuba-shi, Ibaraki 305, Japan

A 21548 bp nucleotide sequence around the 24° region of the *Bacillus subtilis* chromosome, located 306 kb downstream from the zero point of the physical map, was determined. Twenty-one putative ORFs were identified: two ORFs (Orf1 and Orf20) were consistent with the nucleotide and amino acid sequence of *B. subtilis* OrfJ, whose function is not known, and Pcp, respectively; four were found to display significant similarities to known proteins in data banks, i.e. 5-keto-3-deoxyglucarate dehydratase (Orf2), aldehyde dehydrogenase (Orf3) and two glucarate dehydratases (Orf4 and 5); three had considerable similarity to the sensor-regulator proteins of bacterial two-component signal transduction systems (Orf1, 11 and 12); two had considerably high levels of similarity to parts of known proteins (Orf13 and 19); and five showed low levels of similarity to known proteins in the data banks. The remaining six ORFs showed no similarity to known proteins.

Keywords: *Bacillus subtilis*, genome sequencing, 24° region

INTRODUCTION

Itaya & Tanaka (1991) have constructed a map of *NotI* and *SfiI* restriction sites of the entire circular *Bacillus subtilis* chromosome and determined the total size of the genome to be 4188 kb. As part of the *B. subtilis* genome cooperative project between European and Japanese laboratories to determine the entire *B. subtilis* genome nucleotide sequence, we have been analysing the 230 kb region from 20° to 40°, of which the α -amylase gene (*amyE*) (Yang *et al.*, 1983; Yamazaki *et al.*, 1983) is part. A *NotI* linking clone, pNEXT9, is located around 24° in the chromosome. A 21548 bp sequence of the pNEXT9 DNA region was determined and analysed.

METHODS

Bacterial strains, phage and plasmids. *B. subtilis* 168 (*trpC2*), from our laboratory stock, was used as the standard strain for sequencing. Its chromosomal DNA was extracted by the phenol-pH 9 method (Saito & Miura, 1963) and was partially digested by the restriction enzyme *Sau3AI*. DNA fragments of 15–20 kb were prepared by agarose gel electrophoresis using ion-exchange chromatography DE81 paper (Whatman) and libraries of chromosomal DNA were prepared by the *in vitro*

packaging method (Sambrook *et al.*, 1989) using λ DASH II as the vector. The libraries were not amplified. For library screening, the plaque hybridization method (Sambrook *et al.*, 1989) was applied using *Escherichia coli* 2392 as the host. Plasmid pUC18 and *E. coli* JM109 were used to prepare randomly overlapping libraries for shotgun sequencing.

Sequence determination. All sequences were determined on both strands using a 373A DNA Sequencer from Applied Biosystems using the Taq Dye Primer Cycle Sequencing Kit and the Taq DyeDeoxy Terminator Cycle Sequencing Kit.

Randomly overlapping libraries for shotgun sequencing were prepared according to the method of Sambrook *et al.* (1989) after treatment by sonic oscillation (30 W for 30 s) of the DNA fragments to be sequenced. Inserts of pUC18 derivatives were amplified by PCR from the cell-lysates of the transformants using the forward and reverse sequencing primers (5' GTTT-TCCCAGTCACGACG 3' and 5' GAATTGTGAGCGGAT-AAC 3'). The amplified PCR products were used as templates for sequencing by the kits. Sequencing information from random inserts corresponding to approximately four times the length of the fragments to be sequenced covered most of the fragments. The remaining gap regions were filled using region-specific primers and the Taq DyeDeoxy Terminator Cycle Sequencing Kit.

Data handling and computer analysis. DNA sequences obtained were compiled using the ARSQ program from Software Development (Tokyo). To avoid errors in sequencing analysis, we checked every chart and used sequence outputs only from the regions where every base was well separated (approximately 400 bp in one run). The compiled sequence was further analysed for location of possible ORFs using GENETYX-MAC. The amino

Abbreviations: aa, amino acid; SD, Shine–Dalgarno.

The GSD, EMBL, DDBJ and NCBI accession number for the nucleotide sequence reported in this paper is D30808.

acid sequences of the products of identified ORFs were searched for their similarity to sequences reported previously in non-redundant protein sequence data banks using the FASTA and BLAST e-mail servers in the Human Genome Center, Institute of Medical Science, University of Tokyo.

RESULTS AND DISCUSSION

Cloning of the pNEXT9 region (24°) of the *B. subtilis* chromosome

Clones containing the pNEXT9 DNA region were isolated by the plaque hybridization method (Sambrook *et al.*, 1989), using ³²P-labelled pNEXT9 (provided by

M. Itaya) as the probe, from the λ DASH II libraries of *B. subtilis* chromosomal DNA. Thirty-one independent clones were isolated from approximately 7000 plaques. Among them, 10 clones were selected and their inserted DNA fragments were analysed. The relative location and rough restriction enzyme maps of clones 9-5, 9-9 and 9-22 are shown in Fig. 1. The inserted DNA fragments in the seven other clones tested overlapped (data not shown).

Assignment of ORFs in the pNEXT9 region

Putative ORFs were searched for all six possible translation frames. We considered valid the longest possible

Fig. 1. Restriction map and ORF organization of the *B. subtilis* chromosome around the 24° region, located 306 kb downstream from the zero point of the physical map of Itaya & Tanaka (1991). The upper part of this map shows the location of the inserted DNA fragments in the λ DASH II clones (9-5, 9-9 and 9-25) that were hybridized with pNEXT9. The scale (in kb) starts at the A residue of an *EcoRI* site of clone 9-5. The restriction sites *EcoRI* (E) and *BglII* (B), and the junction site of *Sau3AI* and *BamHI* (S/Bm) are indicated. ORFs are shown by thick arrows and the arrowheads show the direction of their transcription and translation. Numbers on the arrows indicate each ORF. *orf20* corresponds to *B. subtilis pcg*.

Table 1. Putative ORFs in the 21 548 bp sequence around 24° of the *B. subtilis* chromosome

ORF	Endpoints (nucleotides)	Size of product (aa/kDa)	SD consensus sequence (upper case) and initiation codon (bold)
<i>orf1</i>	247 > 1188	314/35.7	AGAAGGgtGaaccaaaagct atg
<i>orf2</i>	1418 > 2341	308/34.0	AcAGGAGGTtatgttg atg
<i>orf3</i>	2374 > 3837	488/52.4	GAAAGGgGcaatgggc atg
<i>orf4</i>	3924 > 5288	455/49.3	AAAGGAaagatact atg
<i>orf5</i>	5328 > 6692	455/50.8	AAAGGgAGGaaaatgc atg
<i>orf6</i>	6765 > 7463	233/26.6	gAAGGGAGagttgttc gtg
<i>orf7</i>	7557 > 8015	153/17.1	aaatatgAGGAGtgattttc atg
<i>orf8</i>	8173 > 9084	304/32.4	cGGcGGTGAagta atg
<i>orf9</i>	9657 > 10283	209/23.6	AcAccGGAatTGATCgcttacc atg
<i>orf10</i>	10870 > 11805	312/33.9	AAAGGAaaaaagctgagctgag atg
<i>orf11</i>	11905 > 12582	226/25.8	AGGAGatGcagcgcata ttg
<i>orf12</i>	12587 > 13060	158/17.9	AGGAGagtttttaga atg
<i>orf13</i>	13513 > 14055	181/20.6	AAGaAGtttaga atg
<i>orf14</i>	14195 > 14617	141/15.7	ttAtGGGtttt atg
<i>orf15</i>	14850 < 15233	128/14.0	AAAGGAGGaattgta atg
<i>orf16</i>	15547 > 15972	142/16.4	GGtGGTGAtgtt atg
<i>orf17</i>	16081 > 16809	243/28.0	GatGGTGAtgatc atg
<i>orf18</i>	17376 > 18761	462/52.7	AAtGGtGGtggt atg
<i>orf19</i>	19324 > 19866	181/20.6	AtGGGtttt atg
<i>orf20</i>	19869 < 20513	215/23.8	AAAGGtGatggtt atg
<i>orf21</i>	20592 > 21548→	319/35.1	AAAGGAGtgaatcata atg

Table 2. Similarity of predicted ORF products to other proteins

Product	Similar protein in database	Identity observed	BLAST score
Orf1	<i>B. subtilis</i> OrfJ	100 % for the entire length	1573
	High affinity potassium transport system (KdpE) of <i>E. coli</i>	23 % in the N-terminal region 112 residues	72
	Nitrogen fixation regulatory protein (NtrC) of <i>K. pneumoniae</i>	23 % in the N-terminal region 122 residues	67
Orf2	5-Keto-4-deoxyglucarate dehydrogenase of <i>P. putida</i>	35.5 % in 290 residues	296
Orf3	Aldehyde dehydrogenase of <i>B. stearothermophilus</i>	67 % in 485 residues	1719
	Aldehyde dehydrogenase of human liver	41.5 % in 480 residues	–
	Betaine aldehyde dehydrogenase of <i>E. coli</i>	37 % in 486 residues	402
	Betaine aldehyde dehydrogenase of spinach	36 % in 477 residues	–
Orf4	Glucarate dehydratase of <i>P. putida</i>	60 % in 456 residues	774
Orf5	Glucarate dehydratase of <i>P. putida</i>	73 % in 451 residues	1782
Orf6	Protein A of the gene which is located upstream of the pyruvate dehydrogenase gene cluster of <i>E. coli</i>	25 % in 236 residues	133
Orf9	Orf4 of genes of macrolide 2'-phosphotransferase of <i>E. coli</i>	40 % in the middle region 174 residues	359
	Viomycin phosphotransferase of <i>S. vinaceus</i>	19.4 % in the C-terminal region 186 residues	78
Orf10	Glutathione-regulated potassium efflux system (KefC) of <i>E. coli</i>	23 % in 168 residues in the N-terminal half	–
Orf11	Vancomycin resistance gene product (VanR) of <i>Enterococcus faecium</i>	39 % in 224 residues	352
	PhoP of <i>B. subtilis</i>	29 % in 231 residues	173
	OmpR of <i>E. coli</i>	35 % in 223 residues	167
Orf12	PhoR of <i>B. subtilis</i>	41 % in the C-terminal region 82 residues	149
	Sensor protein BarA of <i>E. coli</i>	37 % in the middle region 86 residues	–
Orf13	<i>bcrA</i> gene product of <i>B. licheniformis</i>	44 % in the middle region 163 residues	381
	Branched-chain amino acid transport protein of <i>P. aeruginosa</i>	34 % in the C-terminal region 107 residues	159
	Molybdenum transport protein (ChlD) of <i>E. coli</i>	32 % in the middle region 110 residues	154
Orf15	NADH-Ubiquinone oxidoreductase chain 5 of <i>Anopheles gambiae</i>	20 % in the N-terminal region 94 residues	42
Orf19	Hypothetical protein o131 of <i>E. coli</i>	36 % in the N-terminal region 122 residues	171
Orf20	Pcp of <i>B. subtilis</i>	100 % for the entire length	1123
Orf21	NifS-like protein of <i>B. subtilis</i>	37 % in the N-terminal region 89 residues	130

ORF starting with ATG, GTG or TTG and preceded by a Shine–Dalgarno (SD) sequence complementary to the 3' end of the 16S rRNA of *B. subtilis* (3' UCUUCCUC-CACUAG---). Based on these criteria, 21 putative ORFs were identified and are listed in Table 1.

The initiation codon for 19 ORFs was ATG, GTG for *orf6* and TTG for *orf11*. The direction of transcription and translation of 19 ORFs in the region was the same, but in the opposite direction for two of the ORFs (*orf15* and *orf20*). We did not determine the orientation of the

movement of the replication fork in this region but believe it should be identical to that of the transcription and translation of the 19 ORFs.

Similarity analysis of the gene products to known proteins (Table 2)

Orf1. The nucleotide sequence of *orf1* and the deduced amino acid (aa) sequence of its product (Orf1) coincide with those of *B. subtilis* OrfJ, whose function is not known (P. G. Quirk and others, unpublished; EMBL accession no. L22503). Also the aa sequence in the N-terminal region of Orf1 has similarity (23–26 % identity) to the N-terminal regions (aa 110–120) of KdpE (Walderhaug *et al.*, 1992) and NtrC (Miranda-Rios *et al.*, 1987) of *E. coli*, NtrC from *Klebsiella pneumoniae* (Drummond *et al.*, 1986) and a phosphate acceptor regulatory protein in *Pseudomonas fluorescens* (Starnbach & Lory, 1992). KdpE and NtrC are parts of the two-component signal transduction systems of bacteria. Therefore, Orf1 may have some part to play in a similar two-component signal transduction system of *B. subtilis*. Its N-terminal region also contains a motif for the phosphorylation site.

Orf2. The aa sequence of Orf2 showed relatively high similarity (35.5 % identity) to 5-keto-4-deoxyglucarate dehydratase of *Pseudomonas putida* (R. P. Burlingame and others, unpublished; EMBL accession no. M69160). Orf2 could be its counterpart in *B. subtilis*.

Orf3. The aa sequence of Orf3 showed high similarity (67 % identity) to aldehyde dehydrogenase from *B. stearothermophilus* (Imanaka *et al.*, 1993) and relatively high similarities (35–40 % identity) to those of betaine-aldehyde dehydrogenases of *E. coli*, beet and spinach, and of aldehyde dehydrogenases (NAD) of *E. coli*, *Emerella nidulans*, mouse, rat, chicken, horse and man (Hidalgo *et al.*, 1991). Orf3 was compared with aldehyde dehydrogenase from *B. stearothermophilus* (Imanaka *et al.*, 1993) and man (Hempel *et al.*, 1984), and betaine-aldehyde dehydrogenases from *E. coli* (Lamark *et al.*, 1991) and spinach (Weretilnyk & Hanson, 1990) (Fig. 2). These sequences all contain the coenzyme binding motif (GXGXXXG) and all except the thermostable enzyme contain the consensus aa sequence (FXXGQXC) of the active centres. Therefore, *orf3* probably encodes an aldehyde dehydrogenase.

Orf4 and Orf5. The aa sequences of Orf4 and Orf5 have high similarity (60 and 73.3 % identity, respectively) to glucarate dehydratase (456 aa) and glucarate dehydratase (451 aa) of *P. putida*, respectively (R. P. Burlingame and others; M69160), suggesting that Orf4 and Orf5 could be their counterparts in *B. subtilis*. However, there is no aa sequence similarity between Orf4 and Orf5 as there is between the two glucarate dehydratases of *P. putida*. The gene order in the *P. putida* chromosome is glucarate dehydratase (456 aa), glucarate dehydratase (451 aa), followed by 5-keto-4-deoxyglucarate dehydratase (R. P. Burlingame and others, unpublished; M69160), whereas the order observed in the *B.*

subtilis genome was 5-keto-4-deoxyglucarate dehydratase, aldehyde dehydrogenase and the two putative glucarate dehydratases corresponding to the 456 and 451 aa proteins.

Orf6. The aa sequence has 25 % identity with the sequence of protein A, the gene of which was found in the pyruvate dehydrogenase gene cluster of *E. coli* (Guest, 1987).

Orf7 and Orf8. No homologous proteins were found.

Orf9. The aa sequence of Orf9 showed relatively high similarity (40 % identity) to the middle region 174 residues of Orf4 of the genes for macrolide 2'-phosphotransferase of *E. coli* (N. Noguchi and others, unpublished; DDBJ accession no. D16251) and low similarity (19.4 % identity) to the C-terminal 186 residues of viomycin phosphotransferase of *Streptomyces vinaceus* (Bibb *et al.*, 1985).

Orf10. Part of Orf10 (aa residues 38–199) showed 23 % identity to the sequence of residues 143–310 of the glutathione-regulated potassium-efflux system (KefC) of *E. coli* (Munro *et al.*, 1991). Orf10 is rich in hydrophobic amino acids (68 %).

Orf11. The aa sequence of Orf11 showed relatively high similarity (30–40 % identity) to PhoP of *B. subtilis* (Seki *et al.*, 1987), VanR of *Enterococcus faecium* (Arthur *et al.*, 1992; Wright *et al.*, 1993), OmpR of *E. coli*, *Salmonella typhimurium* and red algae (Comeau *et al.*, 1985; Russo *et al.*, 1993; Kessler *et al.*, 1992), and PhoB of *E. coli*, *Shigella dysenteriae* and *K. pneumoniae* (Makino *et al.*, 1986). These proteins are members of a family of bacterial transcription regulators, suggesting a similar role for Orf11.

Orf12. The aa sequence of Orf12 showed 37 % identity with an 86 aa sequence in the middle region of the sensor-regulator protein, BarA (918 aa) of *E. coli* (Nagasawa *et al.*, 1992) and 41 % identity with the N-terminal 82 aa sequence of PhoR (579 aa) of *B. subtilis* (Seki *et al.*, 1988).

Orf11 and Orf12 may be sensor and regulator proteins in a two-component signal transduction system of *B. subtilis*. The first codon, ATG, of *orf12* is located 1 bp downstream from the termination codon, TAG, of *orf11*, and the SD sequence of *orf12* is part of the C-terminal coding region of *orf11*, suggesting that the expression of *orf12* could be regulated by *orf11*.

Orf13. The aa sequence Orf13 showed 44 % identity with 163 residues from the middle region of *B. licheniformis* BcrA (Z. Podlesek & A. Comino, unpublished; EMBL accession no. L20573) and 34 % identity with the C-terminal 107 aa of the branched-amino acid transport protein (255 aa, encoded by *braF*) from *Pseudomonas aeruginosa* strain PAO (Hoshino & Kose, 1990). The sequence also showed 32 % identity with 110 residues from the middle region (aa 111–220) of the molybdenum transport protein (ChlD) of *E. coli*. ChlD has strong similarity to other bacterial ATP-binding transport proteins (Johann & Hinton, 1987). Therefore it is possible that Orf13 is a transport protein for the uptake of some unknown substrate.

<i>B. sub</i>	1:-----MSVITEQNTYLNFINGEWVKSQSGDMVKVENPADVNDIVGYVQNSTAED
<i>B. ste</i>	1:-----MKVQTEIKTYFNYINGNWVSSVSNVPSINPANRHIDIVGYVQRSTLED
Human liver	1: SSSGTPDLPVLLTDLKIQYTKIFINNEWHDSVSGKKFPVFNPA-ATEEELCQVEEGDKED
<i>E. coli</i>	1:-----MSRMAEQQLYIHGGYTSATSGRTFETINPA-NGNVLATVQAAGRED
Spinach	1:-----MA-FIPARQLFIDGEWREPIKKNRIPVINP-STEEIIGDIPAATAED
	* ** *
<i>B. sub</i>	50: VERAUTAANE---KTAWRKLGAERGQYLYKTADIMEQRLLEEIAACATREMCKTLPEAK
<i>B. ste</i>	50: VNEAVTAANE---QTSWWKRSQVERGEYLYKAAHILEQLQDIAETMTREMCKTLAEAK
Human liver	59: VDKAVKAARQAFQIGSPWRTMDASERGRLLYKLADLIERDRLLLATMESMNGGKLYSNAY
<i>E. coli</i>	46: VDRVKSAAQGG---QKIWASMTAMERSRIIRRAVDILRERNDLAKLETLDTGKAYSETS
Spinach	47: VEVAVVAARRAFRRNN-WSATSGAHRATYLRAIAAKITEKKDHFVKLETIDSGKPFDEAV
	* ** * * * *
<i>B. sub</i>	107: -GETARGIAILRYYAGEGMRKTG-DVIP-STDKDAL-MFTTRVPLGVGVISPWNPVAI
<i>B. ste</i>	107: -AETMRGVHILRYYAGEGARKIG-DVIP-SSDSEGL-LFTTRVPLGVGVISPWNPVAI
Human liver	119: LNDLAGCIKTLRYCAGWADKIQR-TIPID-G-NFF-TYTRHEPIGVCGQIIPWNPFLVM
<i>E. coli</i>	103: TVDIVTGADVLEYAGLIPALEG-SQIP--LRETSF-VYTRREPLGVVAGIGAWNYPIQI
Spinach	106: L-DIDDVASCFEYFAGQAEALDGKQKAPVTLPMERFKSHVLRQPLGVVGLISPWNPPLLM
	* ** * * * *
<i>B. sub</i>	163: PIWKMAPALVYGNTVVIKPATETAVTCAKIIACFEEAGLPAGVINLVTGPGSVVQGLAE
<i>B. ste</i>	163: PIWKMAPALVYGNTVVLKPASETAVTAAKVICFHEAGFPKGVVNMVCGSGSVVQGLIAN
Human liver	175: LIWKIGPALSCGNTVVVKPAEQPTLTAHVASLIKEAGFPVGVNIVPGYGPAGAAISS
<i>E. coli</i>	159: ALWKSAPALAAGNAMIFKPSEVTPLTALKLAEIYSEAGLPDGVFNVLPGVGAETQYLYTE
Spinach	165: ATWKIAPALAAGCTAVLKPSELASVTCLEFGEVCNEVGLPPGVNLITGLGPDAGAPLVS
	** *** * ** * * * * *
<i>B. sub</i>	223: HDGVNAVTFGTSNQVGKIGQAALARGAKY-QLEMGGKNPVIVADDADLEAAAEAVITGA
<i>B. ste</i>	223: HPDIDGVTFTGTSNTVGKQVGRAAFARGAKY-QLEMGGKNPVIVAKDADLDLAVEGTISGG
Human liver	235: HMDIDKVAFTGSTEVGKLIKEAAGKSNLKRVTLELGGKSPCIVLADADLDNAVEFAHHGV
<i>E. coli</i>	219: HPGIAKVSFTGGVASGKKVMANSAASSLKEVTMELGGKSPLIVFDDADLDLAADIAMMAN
Spinach	225: HPDVDKIAFTGSSATGSKVMSAAQ-LVKPVTLELGGKSPIVVFEDVDIKVVEWTFGC
	* *** * * * * *
<i>B. sub</i>	282: FRSTGQKCITATSRVIVQSGIYERFKEKLLQRTKDITIGDSLKEDVMMGPIASKNQLDNCL
<i>B. ste</i>	282: LRSTGQKCITATSRVIEREVYEPFKAKLLERVKQLKIGNGLDAETWMGPCASESQFHTVL
Human liver	295: FYHQGCCITAAASRIFFVEESIYDEFVRRSVERAKKYILGNPLTPGVTTQGPQIDKEQYDKIL
<i>E. coli</i>	279: FFSGQVCTNGTRVFPKCKAAFEQKILARVERIRAGDVFDPQTNFGPLVSFPHRDNL
Spinach	284: FWTNGOICSATSRLLVHESIAAEFVDKLVKNTKNIKISDPFEEGCRLLGPVISKGYQDKIM
	* *** * *
<i>B. sub</i>	342: SYIEKGKQEGASLLIGGEK-LENGKYQNGYYVQPAIFDNVTSEMTIAQEEIFGPVIALIK
<i>B. ste</i>	342: SYIEKGKSEGAKLIYGGNRCLG-LENGKFFVEPTIFEDVDLQMTIAREEIFGPVIALIQ
Human liver	355: DLIESGKKEGAKLECGGG--P---WGNKGYFVQPTVFSNVTDEMRIAKEEIFGPVQIMK
<i>E. coli</i>	339: RYIAKGKEGARVLCCGDDV-LKGDGFDNGAWVAPTFTDCSDMTIVREEIFGPVMSILT
Spinach	344: KFISTAKSEGATILYGGSR-PE--HLKKGYIEPTIVTDISTSMQIWKEEVFGPVLCKVT
	* * *** ** * *
<i>B. sub</i>	401: VDSIEEALNIANDVKFGLSASIFTENIGRMLSFIDEIDAGLVRINAESAGVELQAPFGGM
<i>B. ste</i>	401: VDSIEEAIKLANDTEYGLSASIYTKNIGNALEFIKDIEAGLIKVNAETAGVEFQAPFGGM
Human liver	410: FKSLLDDVIKRANNTFYGLSAGVFTKIDKAITISSALQAGTVWVNC-YGVVSAQCFFGGF
<i>E. coli</i>	398: YESEDEVIRRANDTDYGLAAGIVTADLNRAHRVHQLAEGICWIN-TWGESPAEMPVGGY
Spinach	401: FSSDEAIALANDTEYGLAAVFSNDLERCERITKALEVGAVWVNC-SQPCFVQAPWGGI
	* ** *** * *
<i>B. sub</i>	461: KQSSSHSREQGEAAKDFFTAICTVFKP-----
<i>B. ste</i>	461: KQSSSHSREQQAIEFFTSIKTVFKA-----
Human liver	469: K-MSGNGRELGEYGFHEYTEVKTIVTKISQKNS-----
<i>E. coli</i>	457: KH-SGIGRENGVMTLQSYTQVKSIVEMAKFQSI-F---
Spinach	460: K-RSGFGRELGEWGIQNYLNKQVTQDISDEPWGMYKSP
	* * *** *

Fig. 2. Comparison of the amino acid sequences of a putative aldehyde dehydrogenase of *B. subtilis* (*B. sub*), the aldehyde dehydrogenase of *B. stearothermophilus* (*B. ste*), the cytoplasmic aldehyde dehydrogenase of human liver (Human liver), and the betaine-aldehyde dehydrogenases from *E. coli* (*E. coli*) and *Spinacia oleracea* (Spinach). Dashes indicate spaces inserted in the sequences to give the best alignment. The consensus sequence for the co-enzyme binding site (A) and active centres of aldehyde dehydrogenases (B) are boxed.

Orf14, 16, 17 and 18. No homologous proteins were identified.

Orf15. The aa sequence of Orf15 showed low similarity (20% identity) to the N-terminal 94 aa of the 580 aa NADH-ubiquinone oxidoreductase chain 5 of *Anopheles gambiae* (C. B. Beard and others, unpublished; EMBL accession no. L20934).

Orf19. The aa sequence of Orf19 showed 36% identity to the N-terminal 122 aa of a hypothetical 131 aa protein, o131, of *E. coli* (Daniels *et al.*, 1992).

Orf20. The nucleotide sequence of *orf20* coincides completely with that of *B. subtilis* *pcp*, the gene for pyrrolidone carboxyl peptidase (Awade *et al.*, 1992).

Orf21. Since the region of *orf21* we sequenced did not contain a termination codon, the aa sequence of its product would continue. However, the incomplete aa sequence showed 37% identity with the middle region 89 residues of a 395 aa NifS-like protein of *B. subtilis* (Sun & Setlow, 1993).

ACKNOWLEDGEMENTS

The authors thank M. Itaya for a *NotI* linking clone, pNEXT9, and Joel Smith for critically reading the manuscript. 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

- Arthur, M., Molinas, C. & Courvalin, P. (1992). The VanS-VanR two-component regulatory system controls synthesis of depsi-peptide peptidoglycan precursors in *Enterococcus faecium* BM4147. *J Bacteriol* **174**, 2582–2591.
- Awade, A., Cleuziat, P., Gonzales, T. & Robert-Baudouy, J. (1992). Characterization of the *pcp* gene encoding the pyrrolidone carboxyl peptidase of *Bacillus subtilis*. *FEBS Lett* **305**, 67–73.
- Bibb, M. J., Bibb, M. J., Ward, J. M. & Cohen, S. N. (1985). Nucleotide sequences encoding and promoting expression of three antibiotic resistance genes indigenous to *Streptomyces*. *Mol & Gen Genet* **199**, 26–36.
- Comeau, D. E., Ikenaka, K., Tsung, K. & Inouye, M. (1985). Primary characterization of the protein products of the *Escherichia coli* *ompB* locus: structure and regulation of synthesis of the OmpR and EnvZ proteins. *J Bacteriol* **164**, 578–584.
- Daniels, D. L., Plunkett, G., III, Burland, V. & Blattner, F. R. (1992). Analysis of the *Escherichia coli* genome: DNA sequence of the region from 84.5 to 86.5 minutes. *Science* **257**, 771–778.
- Drummond, M., Whitty, P. & Wootton, J. (1986). Sequence and domain relationships of *nirC* and *nifA* from *Klebsiella pneumoniae*: homologies to other regulatory proteins. *EMBO J* **5**, 441–447.
- Guest, J. R. (1987). Functional implications of structural homologies between chloramphenicol acetyltransferase and dihydro-lipoamide acetyltransferase. *FEMS Microbiol Lett* **44**, 417–422.
- Hempel, J., Bahr-Lindstrom, H. & Jornvall, H. (1984). Aldehyde dehydrogenase from human liver. Primary structure of the cytoplasmic isoenzyme. *Eur J Biochem* **141**, 21–35.
- Hidalgo, E., Chen, Y. M., Lin, E. & Aguilar, J. (1991). Molecular cloning and DNA sequencing of the *Escherichia coli* K-12 *ald* gene encoding aldehyde dehydrogenase. *J Bacteriol* **173**, 6118–6123.
- Hoshino, T. & Kose, K. (1990). Cloning, nucleotide sequences, and identification of products of the *Pseudomonas aeruginosa* PAO *bra* genes, which encode the high-affinity branched-chain amino acid transport system. *J Bacteriol* **172**, 5531–5539.
- Imanaka, T., Ohta, T., Sakoda, H., Widhyastuti, N. & Matsuoka, M. (1993). Cloning, nucleotide sequence and efficient expression of the gene coding for thermostable aldehyde dehydrogenase from *Bacillus stearothermophilus*, and characterization of the enzyme. *J Ferment Bioeng* **76**, 161–167.
- 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.
- Johann, S. & Hinton, S. M. (1987). Cloning and nucleotide sequence of the *chlD* locus. *J Bacteriol* **169**, 1911–1916.
- Kessler, U., Maid, U. & Zetsche, K. (1992). An equivalent to bacterial *omp'R* genes is encoded on the plastid genome of red algae. *Plant Mol Biol* **18**, 777–780.
- Lamark, T., Kaasen, I., Eshoo, M. W., Falkenberg, P., McDougall, J. & Strom, A. R. (1991). DNA sequence and analysis of the *bet* genes encoding the osmoregulatory choline-glycine betaine pathway of *Escherichia coli*. *Mol Microbiol* **5**, 1049–1064.
- Makino, K., Shinagawa, H., Amemura, M. & Nakata, T. (1986). Nucleotide sequence of the *phoB* gene, the positive regulatory gene for the phosphate regulon of *Escherichia coli* K-12. *J Mol Biol* **190**, 37–44.
- Miranda-Rios, J., Sanchez-Pescador, R., Urdea, M. & Cavarrubias, A. A. (1987). The complete nucleotide sequence of the *glnALG* operon of *Escherichia coli* K12. *Nucleic Acids Res* **15**, 2757–2770.
- Munro, A. W., Ritchie, G. Y., Lamb, A. J., Douglas, R. M. & Booth, I. R. (1991). The cloning and DNA sequence of the gene for the glutathione-regulated potassium-efflux system KefC of *Escherichia coli*. *Mol Microbiol* **5**, 607–616.
- Nagasawa, S., Tokishita, S., Aiba, H. & Mizuno, T. (1992). A novel sensor-regulator protein that belongs to the homologous family of signal-transduction proteins involved in adaptive responses in *Escherichia coli*. *Mol Microbiol* **6**, 799–807.
- Russo, F. D., Slauch, J. M. & Silhavy, T. J. (1993). Mutations that affect separate functions of OmpR, the phosphorylated regulator of porin transcription in *Escherichia coli*. *J Mol Biol* **231**, 261–271.
- Saito, H. & Miura, K. (1963). Preparation of transforming deoxy-ribonucleic acid by phenol treatment. *Biochim Biophys Acta* **72**, 619–629.
- Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual*, 2nd edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.
- Seki, T., Yoshikawa, H., Takahashi, H. & Saito, H. (1987). Cloning and nucleotide sequence of *phoP*, the regulatory gene for alkaline phosphatase and phosphodiesterase in *Bacillus subtilis*. *J Bacteriol* **169**, 2913–2916.
- Seki, T., Yoshikawa, H., Takahashi, H. & Saito, H. (1988). Nucleotide sequence of the *Bacillus subtilis* *phoR* gene. *J Bacteriol* **170**, 5935–5938.
- Starnbach, M. N. & Lory, S. (1992). The *flaA* (*rpoF*) gene of *Pseudomonas aeruginosa* encodes an alternative sigma factor required for flagellin synthesis. *Mol Microbiol* **6**, 459–469.
- Sun, D. & Setlow, P. L. (1993). Cloning, nucleotide sequence and regulation of the *Bacillus subtilis* *nadB* gene and a *nifS*-like gene both of which are essential for NAD biosynthesis. *J Bacteriol* **175**, 1423–1432.
- Walderhaug, M. O., Daniel, J. M., Hesse, J. & Epstein, W. (1992). KdpD and KdpE, regulatory proteins for the *kdp* operon of

Escherichia coli, are members of the two component sensor-effector class. *J Bacteriol* **174**, 2152–2159.

Weretilnyk, E. A. & Hanson, A. D. (1990). Molecular cloning of a plant betaine-aldehyde dehydrogenase, an enzyme implicated in adaptation to salinity and drought. *Proc Natl Acad Sci USA* **87**, 2745–2749.

Wright, G. D., Holman, T. R. & Walsh, C. T. (1993). Purification and characterization of VanR and the cytosolic domain of VanS: a two-component regulatory system required for vancomycin resistance in *Enterococcus faecium* BM4147. *Biochemistry* **32**, 5057–5063.

Yamazaki, H., Ohmura, K., Nakayama, A., Takeichi, Y., Otozai, K., Yamasaki, M., Tamura, G. & Yamane, K. (1983). α -Amylase genes (*amyR2* and *amyE*) from an α -amylase hyperproducing *Bacillus subtilis*: molecular cloning and nucleotide sequence. *J Bacteriol* **156**, 327–337.

Yang, M., Galizzi, A. & Henner, D. (1983). Nucleotide sequence of the amylase gene from *Bacillus subtilis*. *Nucleic Acids Res* **11**, 237–249.

Received 13 July 1994; revised 13 September 1994; accepted 27 September 1994.