Sequence around the 159° region of the *Bacillus subtilis* genome: the *pkcX* locus spans 33.6 kb

Alessandra M. Albertini, Tiziana Caramori, Francesco Scoffone,[†]
Claudio Scotti[‡] and Alessandro Galizzi

Author for correspondence: Alessandra M. Albertini. Tel: +39 382 505549. Fax: +39 382 528496.
e-mail: albertini@ipvgen.unipv.it

Dipartimento di Genetica e
Microbiologia, Università di
Pavia, 27100 Pavia, Italy

The nucleotide sequence of 20 kb contiguous to the *pkcX* locus of *Bacillus subtilis* was determined. Six ORFs were recognized, one of which extended for 13341 nucleotides. Their predicted products have significant similarities to proteins with known functions involved in the synthesis of polypeptides and polyketides or in fatty acid metabolism. At the nucleotide level, three regions with a high level of sequence identity (49–54%) to the *Aspergillus nidulans* *wA* gene, responsible for the synthesis of a polyketide pigment, were recognized. The observed similarities suggest that the 20 kb region and the previously reported 13.6 kb region containing *pkcX* are part of the same locus, possibly involved in secondary metabolism.

Keywords: *Bacillus subtilis*, *pkcX*, genome sequencing, polyketide synthases, polypeptide synthases

INTRODUCTION

The project to systematically sequence the *Bacillus subtilis* genome is a cooperative effort between 18 European Community (EC) and seven Japanese research groups. It is hoped that the whole genome sequence will be available by the end of 1997. This concerted effort has produced a total of more than 347 kb of published sequence since it started in 1990: the largest contig (180 kb) is near the replication origin (Ogasawara *et al.*, 1994); two other large contigs of 97 and 35 kb are located at 325–333° and at 210–220° of the physical map respectively, (Glaser *et al.*, 1993; Sorokin *et al.*, 1993).

In all these cases, only 50% of the identified coding regions have been correlated to known functions or showed significant similarity to identified functions;

similar results have been obtained in *Escherichia coli* for a 325 kb contig (Plunkett *et al.*, 1993).

In our laboratory we determined and published the sequence of 8.3 kb of the *flaA* locus (Albertini *et al.*, 1991) and 13.6 kb of the *pkcX* locus (Scotti *et al.*, 1993), both located in the chromosomal region between 150° and 165° and assigned to our group by the EC *B. subtilis* genome sequencing project, as part of the SCIENCE and BIOTECHNOLOGY programmes (Kunst & Devine, 1991).

The 8.3 kb of the *flaA* locus represents part of the complex of genes involved in flagella synthesis and motility, encoding proteins identified by their similarity to the functionally equivalent proteins of *E. coli* or *Salmonella typhimurium*.

Sequence analysis of 13.6 kb of the locus named *pkcX* revealed that this region of the chromosome, very poor in genetic markers, is, in contrast, rich in informational content. A unique long ORF for a predicted product of 4427 amino acids (aa) occupies 97.6% of the sequence. Similarity analysis of the deduced product to known proteins suggested that it may function as a type I polyketide synthase (PKS I). This type of enzyme, highly similar to the eukaryotic fatty acid synthases, is usually involved in the synthesis of complex polyketides, long alkyl carboxylates more or less reduced (Katz & Donadio,

[†] **Present address:** School of Pharmacy, University of Wisconsin, Madison, WI 53706, USA.

[‡] Deceased.

Abbreviations: AAS, ATPase site; ABS, amino-acid-binding site; ACP, acyl carrier protein; DH, dehydratase; ECH, enoyl-CoA hydratase; Fop, frequency of optimal codons; HCS, hydroxymethylglutaryl-CoA synthase; KR, β -ketoreductase; KS, β -ketoacyl-ACP synthase; PKS, polyketide synthase; RAC, racemization (domain); SU, synthesis units.

The GenBank accession number for the nucleotide sequence reported in this paper is U11039.

1993). Preliminary data obtained with deletion mutants suggested that the locus is not essential for the viability of *B. subtilis*, at least under standard laboratory conditions (Scotti *et al.*, 1993).

In this paper we report the sequence determination and analysis of an additional 20 kb contiguous to the 5' end of the 13.6 kb region of the *pksX* locus (BACPKSX).

METHODS

Bacterial strains, phages and plasmids. *B. subtilis* 168 *trpC2* was used as the source of DNA in the construction of the two λ genomic libraries.

The *E. coli* strain used as host for λ Charon 4A recombinant derivatives (Ferrari *et al.*, 1981) was DP50 *supF* (*tonA53 glnV44 tyrT58 dapD8 lacY1 supE44 supF58 Δ gal wvr47 gyrA29 thyA57 hsdS3*). For λ FixII (Stratagene) (Kunst & Devine, 1991) the host was *E. coli* P2392 F⁻ [*hds* R574 ($r_k^- m_k^+$) *supE44 supF58 lacY1* or Δ (*lacIZY*)6 *galK2 galT22 metB1 trpR55*]. *E. coli* DH5 α (Gibco/BRL) was used for plasmid sub-cloning.

Plasmid vectors used for sub-cloning were pGEM7z (Promega) and the integrative pJM103 (Perego, 1993).

Media. LB medium was used for standard *E. coli* cultures, and TY (tryptone–yeast extract medium) for λ phage propagation. For *E. coli* DP50, TY was supplemented with thymidine (50 μ g ml⁻¹) and diaminopimelic acid (100 μ g ml⁻¹). Nutrient broth (Difco) was used for standard *B. subtilis* cultures.

Antibiotics were added at the following concentrations: ampicillin, 100 μ g ml⁻¹; chloramphenicol, 20 μ g ml⁻¹ for *E. coli* and 5 μ g ml⁻¹ for *B. subtilis*.

Transformation. *E. coli* transformation was performed following the procedure of Hanahan (1985).

DNA manipulations. Phage and plasmid DNA preparations were performed by standard procedures according to Sambrook *et al.* (1989).

Chromosomal DNA preparations obtained as described by Albertini & Galizzi (1985) and digested with suitable restriction enzymes, were analysed by Southern hybridization, following the standard procedures described in Sambrook *et al.* (1989).

Sequencing. Nucleotide sequencing was performed on both strands of plasmid subclones with Sequenase (USB) or T7 (Pharmacia) sequencing kits, and custom primers synthesized with a Cyclon Plus DNA Synthesizer (Millipore). The sequence of the junction between λ NF2 and λ BA7 clones was verified by direct sequencing of the asymmetric PCR product obtained with suitable primers on the *B. subtilis* chromosomal DNA.

Data handling and computer analysis. DNA sequence compilation, assembly and a preliminary search for putative ORFs were performed with the IBI/Pustell Sequence Analysis Program (Pustell & Kafatos, 1984). Sequences were compared to data banks using FASTA (Pearson, 1990) and BLAST packages (Altschul *et al.*, 1990). In particular we used BLASTX to ascertain the coding potential in all six potential reading frames, BLASTN to compare the nucleotide sequence and BLASTP to compare the deduced protein product to the non-redundant nucleotide and protein data banks, respectively. Multiple alignments of the putative products showing significant similarity scores to well known proteins, have been obtained with CLUSTAL V (Higgins *et al.*, 1992) and MACAW (Schuler *et al.*, 1991). Codon usage analysis was performed with the CODONS program of Lloyd & Sharp (1992).

RESULTS AND DISCUSSION

Nucleotide sequence analysis

Five partially overlapping λ clones (λ FIXB14, λ U6, λ NF2, λ IG9, λ BA7) were obtained by plaque hybridization (Fig. 1). The previously published sequence of the *pksX* locus (BACPKSX; Scotti *et al.*, 1993) is present in clones λ IG9 and λ BA7. To extend the sequence upstream we made use of λ NF2, λ FIXB14 and plasmid subclones derived from them. This work reports the new sequence of 20 kb (at present called SDPVFST), contiguous to BACPKSX whose main features are reported in Fig. 1 for the sake of completeness. Nucleotide sequence comparison using BLASTN, revealed, starting about 10 kb from the 5' end, a sequence organization strikingly similar to that characteristic of the *pksX* gene (Scotti *et al.*, 1993): the DNA similarity matrix in Fig. 2 shows the presence of a long, albeit interrupted symmetry diagonal and three additional shorter diagonals. This type of pattern is typical of genomic regions characterized by the presence of long stretches of repeated, similar sequences. The long diagonal represents the high overall similarity of the two sequences, with identity scores ranging from 51% to 71%. The interruption or shift up of the diagonal at approximately 8000 nt on the BACPKSX axis, can be interpreted as being caused by a duplication in BACPKSX or a deletion of about 2500 nucleotides (nt) in SDPVFST. The three additional diagonals on the abscissa represent three of the four repetitive sequences present in BACPKSX (Scotti *et al.*, 1993), the fourth being represented by the main diagonal. Intrasequence (self) comparison of SDPVFST showed the presence of three repetitions (data not shown), which can be deduced from the matrix of Fig. 2 as diagonals on the vertical axis. Therefore, the sequence analysis of an additional 10 kb, to the 5' *pksX* locus, reveals the presence of internal repetitions as already shown by Scotti *et al.* (1993). This organization again confirms the structural similarity of this locus to the *eryA* locus of *Saccharopolyspora erythraea* (Donadio *et al.*, 1991).

Even more striking was the finding of a significantly high identity score (BLASTN probability of 6.9×10^{-8}) between SDPVFST and the *wA* gene of *Aspergillus nidulans*, a eukaryotic gene encoding a polyketide or fatty acid synthase (Mayorga & Timberlake, 1992). As reported in Table 1, three long stretches of SDPVFST, corresponding to the aforementioned three repeated sequences, have high identity to a coding region, exon 3, of *wA*. Finally, it is noteworthy that high identity scores (BLASTN probability $< 5.6 \times 10^{-14}$) were observed with *B. subtilis* *srfA* (Fuma *et al.*, 1993; Cosmina *et al.*, 1993) and with *Bacillus brevis* *grsB* (Turgay *et al.*, 1992), both encoding polypeptide synthases involved in the synthesis of peptide antibiotics.

Identification of the coding regions

Following the three criteria proposed by Glaser *et al.* (1993) for the identification of coding regions, we recognized only six coding regions spanning 93% of the 20 kb of the SDPVFST sequence (Table 2 and Fig. 1). All

Fig. 1. Structural organization of the *pksX* locus. Dashed arrows represent the overlapping λ clones. The large bar represents the portion of the clones that has been sequenced: the two shaded regions indicate the sequenced portions described in this paper (dark shading) and by Scotti *et al.* (1993) (light shading). Some of the restriction sites used for sub-cloning and for Southern blotting to compare the sequence data to the chromosomal organization are indicated (E, *EcoRI*; H, *HindIII*; Hc, *HindI*; P, *PstI*; Pv, *PvuII*; B, *BamHI*; N, *NotI*). The position of two putative terminators is indicated by hairpins. The ORFs are indicated by arrows. Circles indicate the positions of the different domains or regions of similarity as deduced by the amino acid sequence comparison to known protein functions. Circles with a dashed outline indicate similarity devoid of consensus PROSITE signatures. All other circles refer to similar sequences with consensus PROSITE signatures; shaded circles indicate similarities to polypeptide synthases; the KS domains are hatched since they help in defining the SU of PKSs. Domain abbreviations are defined in the text.

ORFs are on the same strand and have the same orientation as the ORFs present in BACPKSX. In addition, no significant similarities were found among the translation products of the complementary strand and

protein data banks using BLASTX. All ORFs, starting with one of the three initiation codons ATG, TTG or GTG, were preceded by a consensus sequence for binding to 16S rRNA. Two potential transcription terminators were

Fig. 2. DNA similarity matrix comparing SDPVFST to BACPKSX. The Pustell & Kafatos (1984) method was used with the following parameters: hash level, 6; window, 30; minimal score, 65%; minimal match shown, 66%. Numbers refer to nt residues; letters represent different degrees of sequence similarity (from B = 99–98% to R = 67–66%).

Table 1. Similarity of SDPVFST to exon3 of *wA* of *A. nidulans*

	Endpoints of compared regions (nt)	Identity scores (%)
<i>wA</i>	2810–3950	49.5
SDPVFST	10157–11294	
<i>wA</i>	2760–3760	54.4
SDPVFST	14847–15848	
<i>wA</i>	3270–3903	51.2
SDPVFST	19050–19672	

identified at 323–369 nt and 1961–1988 nt (Fig. 1); the second hairpin is followed by a sequence with a high level of identity to the consensus sequence of σ H-dependent promoters.

Codon usage

We evaluated several parameters for the six identified ORFs. Their G+C content was between 0.44 and 0.48, and the G+C₃ values (the G+C content evaluated for synonymous codons with a different base in the third position) varied between 0.39 and 0.48. Codon usage and some derivative parameters, for example CAI (codon adaptation index) varying between 0.35 and 0.41, Fop (frequency of optimal codons) varying between 0.16 and 0.29, and N_e (effective number of codons used) varying

between 52.3 and 58.6, indicated that the six putative genes belong to class 1 genes of *B. subtilis*, a class with a very low bias in codon usage (Shields & Sharp, 1987).

The unexpected significant similarity at the nucleotide level, observed with the *wA* gene of *A. nidulans*, corresponded to the long ORF named *pksorfx6*. The three regions of similarity to *wA* (Table 1) correspond, at the protein level, to three β -ketoacyl-ACP (acyl carrier protein) synthase (KS) domains, particularly in the portion where the KS signature (DTACSSS) could be identified (Figs 1 and 4). Codon usage analysis of these sequences clarified some aspects of this unexpected finding. (1) The G+C content of these portions of *pksorfx6* is 0.5, higher than the mean value of the entire ORF (0.47). (2) The internal codon usage of *pksorfx6* changed in these stretches of sequence. Fop, evaluated from a table of optimal and non-optimal codons of *B. subtilis* (Sharp *et al.*, 1990) decreased in all instances and for one stretch the decrease was highly significant (from 0.25 to 0.08). (3) The Fop of the *wA* sequence also decreases from 0.37 (mean value of exon 3) to 0.29. (4) Using the *A. nidulans* table of optimal codons, the Fop value of the *pksorfx6* sequences, increased (from 0.08 to 0.26 in the aforementioned case). We interpret these observations to indicate that in the regions of reciprocal similarity, *pksorfx6* and *wA* have a codon usage different from that characteristic of the entire gene. This is obtained, at least in the case of *B. subtilis*, by the reduction of optimal codons and an increased use of non-optimal codons, with the effect of significant variation in the local G+C content, which is not due to the use of synonymous codons with a different base in the third position. This

Table 2. General features of the seven ORFs of SDPVFST

ORF	Endpoints	Size (codons)	Translational start	Similarity (BLAST probabilities)
<i>pksorfx1</i>	795–2042	416	tGAGGTgcttt ATG ACA	PKS II (4.7×10^{-18})
<i>pksorfx2</i>	2269–3306	345	aGAAGGAccgattgaa TTG CTG	HCS acetylation site (7.8×10^{-10})
<i>pksorfx3</i>	3293–4073	259	tAGAAGAAtgcttgcaa GTG TTA AAC	FadB and MenB of <i>E. coli</i> , ECH (1.5×10^{-7})
<i>pksorfx4</i>	4113–4862	249	tAGGAGGAcattc ATG ACG CAT	FadB and MenB of <i>E. coli</i> , ECH (4.6×10^{-12}). PROSITE signature PS00166.
<i>pksorfx5</i>	4907–6580	558	AAAtGGAGGcagcgtg ATG AGA AAT	ATP-dependent AMP-binding enzymes (1.1×10^{-11}). PROSITE signature PS00455.
<i>pksorfx6</i>	6657–20000	4447	tAAGGgaAGaaatagaagcattc TTG CTC ACC	1–1300 aa polypeptide synthases (3.2×10^{-113}). PROSITE signature PS00455. 1150–4447 aa PKS I (0 with BACPKSX, 3.2×10^{-131} with EryA). PROSITE signatures PS00012 and PS00606. Conserved and repeated ACP, KS, DH and KR domains.
<i>pksorfx7</i>	20316–21043	> 250	AAGAAGcttcagaagaagctctat ATG AAG GCT	ORF of BACPKSX. Amino end of putative PksX (PKS I).

RAC

pkx6rac	WILTLARETHYYTIERIASLSEHedlsslsadkkgtdtyktepRSRTFQ-----LSEVVRglwTQFMSPAKKA	74
bssrf1s1	W-KLEMDINSHITIEHTQYGERA-----EGKADQG-VEGEVILTIPIQR--WFFEHNFNTNEHH	57
bssrf1s2	W-KLEMDIDFHITIEHTQYGERA-----EGKADQG-VEGEVILTIPIQR--WFFEHNFNTNEHH	57
pkx6rac	YHVPICFFPSSHHHTFLAFGLVIANPILPHMIPKKGvplk---NEPALSISjktenisslkesdipafIRKKI	150
bssrf1s1	WNQSVMLHAKKHDPDPRVHTLQALIEHIDALEMYYRGGQEDVIQYNAGLEAASAQLEVIQIEGgaady-----EDR	130
bssrf1s2	WNQSVMLHAKKHDPDPRVHTLQALIEHIDALEMYYRGGQEDVIQYNAGLEAASAQLEVIQIEGgaady-----EDR	131
pkx6rac	K-----EPYVKENSPLVPMSPSRSEQEHLVVIHHLIFDGVSVTFRSDFDTQOLLGQQPEKAVSPAITYHCPA	223
bssrf1s1	EREAERLQSSIDLQEGGLLAAGLQAEDCEHLLIAIHHLVVDGVSWRILLDFSAVVTQLEGNPEVLPQKTHSBAEYA	209
bssrf1s2	KLLANKLQSSLDLRNGPILKAAYRTEACDEHLLIAVHHLVVDGVSWRILLDFASGMQAEKEESLVFPQKTNSEKQWA	210
pkx6rac	AWEKNMLAGKDGKHKRTWQqlsgtlpnql	255
bssrf1s1	ERLQDANSKAFIKEKEKWSLSEQAaaaklp	241
bssrf1s2	EELAAESQSAHLQQAEEKWSLAAEQYslp-	241

ATP-binding site

pkx6abs	SDRKVLTRTVTPENLAYVIYTSGETGKPKGVMIHKALTNFLVSMGETPGLTA-- --ENKMLAVTIVCT--SIAALELH	73
bssrf1s1	GAEESLNADVARNLAYIYTSGETGKPKGVMIHROHHLVESLQQTIIYQSPtq-- --TLPMFLPPT--HPSASVKQIP	75
bssrf1s2	ADARNLALPLDSAANLAYIYTSGETGKPKGVMIHVTIANLETVKETNYLSITE-----QITILGLSNVY--SAFMFDM	72
bssrf1s3	ENPENPETAVTAENLAYMIYTSGETGKPKGVMEHHAIVMLCFWHDFAFSmtA-----ETRSAKYAGFG--SASIWEM	73
bssrf2s4	EAADNLNLPSPSDLAYIMYTSGETGKPKGVMIHKSLLHLVKNAGYVPVTE-----ETAMAQTGA--SISAGTFEV	72
bssrf2s5	ESKERPSSISGRNLAYVIYTSGETGKPKGVMIHRTNINYSWVFSEAGLTKrradgNKTVLLSSAII--LGYTCMI	78
bssrf2s6	NDEGNPETAVTAENLAYMIYTSGETGKPKGVMEHHAIVMLCFWHDFAFSmtA-----ETRSAKYAGFG--SASIWEM	73
bssrf3s7	EQASDPATAIDPNDLAYIMYTSGETGKPKGVMIHVTIANQLVVKHVDYMAFSD-----QITFLSVSNIA--SAFTDFD	72
bbgrsbP	ELTEEIDYINESEDLAYIYTSGETGKPKGVMLEHKNVLLHFTFEKTNInF-----SKVLQYTNA--VLCVTKKF	73
bbgrsbV	GDSSNLAPVNQAGLAYIMYTSGETGKPKGVMEHRTNINYSWVFSEAGLTKrradgNKTVLLSSAII--LGYTCMI	72
bbgrsbO	EEASNLEPIARPHYIAYIYTSGETGKPKGVMIHQSIVNVVAMAWKDAYRLdtfpv--- --RLLQMASBAAFVVSAGDF	76
bbgrsbL	EDGTNLPMNQSTDLAYIYTSGETGKPKGVMIHQSIVNCLQWRKEEYFgp--- --GDTALQVFSBAF--DGFVASLF	73

AAS

pkx6aas	VARGYVYKERLIDSKFDNPHFSGSKLYRTGDMARELTSGRIEYIGHIINQVKIRGFRIELGDIHSRSEHPGLECVV	80
bssrf1s1	VCRGLHLPLLEHFLQPFVPGDRMYRTGDVVRMLPDGTIEYLGHIDQVKVGYRIELGEIHAVVQAPDAAAVVL	80
bssrf1s2	-VKGYNRPFLLEHKEFVPHPTSGERMVYRTGDLARLPLDGTIEYIGHIINQVKIRGFRIELGDIHSRSEHPGLECVV	79
bssrf1s3	-ARGNLNRPFLLEHKEFVPHPTSGERMVYRTGDLARLPLDGTIEYIGHIINQVKVGYRIELGEIHAVVQAPDAAAVVL	78
bssrf2s4	VARGYVYKERLIDSKFDNPHFSGSKLYRTGDMARELTSGRIEYIGHIINQVKIRGFRIELGDIHSRSEHPGLECVV	80
bssrf2s5	VARGYVYKERLIDSKFDNPHFSGSKLYRTGDMARELTSGRIEYIGHIINQVKIRGFRIELGDIHSRSEHPGLECVV	80
bssrf2s6	-ARGNLNRPFLLEHKEFVPHPTSGERMVYRTGDLARLPLDGTIEYIGHIINQVKVGYRIELGEIHAVVQAPDAAAVVL	78
bssrf3s7	VSKGYVNRALLEHKEFVPHPTSGERMVYRTGDLARLPLDGTIEYIGHIINQVKIRGFRIELGDIHSRSEHPGLECVV	80
bbgrsbP	VARGYVYKERLIDSKFDNPHFSGSKLYRTGDMARELTSGRIEYIGHIINQVKIRGFRIELGDIHSRSEHPGLECVV	80
bbgrsbV	VARGYVYKERLIDSKFDNPHFSGSKLYRTGDMARELTSGRIEYIGHIINQVKIRGFRIELGDIHSRSEHPGLECVV	80
bbgrsbO	VARGYVYKERLIDSKFDNPHFSGSKLYRTGDMARELTSGRIEYIGHIINQVKIRGFRIELGDIHSRSEHPGLECVV	80
bbgrsbL	VARGYVYKERLIDSKFDNPHFSGSKLYRTGDMARELTSGRIEYIGHIINQVKIRGFRIELGDIHSRSEHPGLECVV	80

Pantetheine-binding motif

pkx6pan	DTVFTWQBVLYTS-DIEWDGFHDVKNDSLLA-VTVADRIKHELSCEFSVTDIFRYST-IMK	61
bssrf1s1	KAAALWQDVNLVE-KAGIFDNHETIGHSLKA-MTLTKHKETGIEIHQQFIFDHPT-ITAI	61
bssrf1s2	QKQADIWAQVLAQVQAYDFEII-IGHSLAMKMPALVHQLGVELSLKDLAQNPPT-EGLE	61
bssrf1s3	ELIACIWDVNLVE-SEVGVTDNHSLEPDSIEI-IQMASRLNQ-HGWKLEMKDLNHTTEEL	60
bssrf2s4	KKLAELWEGILTV-KAGVTDNPHMI-IGHSLKAMMTAKIQEH-FHKEVPIKVLFEKPTIQEL	60
bssrf3s7	ETIACIWSVLLER-KQIGIHDFHALGHSLEA-MTAVPHQE-LGIDLAKLLFPAFTI-AG	60
bbgrsbP	EKIAKIEWEVLIS-QIGIQNFHSLGHSLEA-ITLSRMNKECNVDILRLLEFPAFTI-IOE	61
bbgrsbV	MKLAELWNNVLVN-KIGVLNHFELGHSLEA-MTMISQVHKEFDVELLKVLEFTIT-IS--	59
bbgrsbO	gKLEELWQDVLLQ-RVGIHDFETIGHSLKA-MAVISQVHKECQTEVLRVLEFTIT-IQ--	59
bbgrsbL	aQVLVWQBVLLIE-LIGITNHFELGHSLEA-TLLVAKIYEYMQIEMPLNVVFEFTI-IMK	61

Fig. 3. For legend see facing page.

peculiar codon usage appears to be typical of the KS domain and is conserved in two evolutionarily distant organisms: the prokaryote *B. subtilis* and the eukaryote *A. nidulans*.

A possible explanation of this occurrence could be that these sequences have been acquired by both organisms recently, in evolutionary terms, by horizontal transfer and have maintained some of the codon usage characteristic of the organism of origin.

Protein sequence analysis

The six ORFs identified in SDPVFST have been compared with those in data banks using the BLASTP and FASTA programs. The results reported in Table 2 refer to similarities with the lowest BLASTP Poisson probability ($< 10^{-7}$). None of the putative products corresponded to known proteins, but they all showed significant similarity to protein functions well known in other biological systems. We could distinguish three classes of similarity. Class I, with low but significant similarity scores to well identified functions, even though devoid of any consensus PROSITE signature. The PROSITE signature is defined as the short primary amino acid sequence pattern representing, in certain cases, the demonstrated or putative active site of an enzyme, and in all cases as a stretch of amino acids conserved in a set of related biological functions (Bairoch, 1991). We classified Pksorfx1 and Pksorfx3 in this class. Class II, with high similarity scores to well identified functions, reinforced by the presence of a consensus PROSITE signature characterizing the reference function, or a determined active site: Pksorfx2, Pksorfx4 and Pksorfx5 pertained to this class. Class III, with the highest similarity scores and, in particular, similarity to functions encoded by other *B. subtilis* loci, reinforced again by the presence of PROSITE signatures: this is the case for the large ORF, *pksorfx6*, encoding a 4427 aa polypeptide (M_r 496025), highly similar to the product of BACPKSX.

The similarity of Pksorfx1 to the type II polyketide or fatty acid synthases is low and centred in the KS amino acid sequence, products of *nodE* of *Rhizobium leguminosarum* and of *fabB* of *E. coli*, but devoid of the consensus signature sequence around the active site Cys residue, an example of which is well represented in Fig. 3 for Pksorfx6.

Pksorfx2 showed similarity to a defined class of enzymes involved in ketogenesis in eukaryotes, i.e. mammalian hydroxymethylglutaryl-CoA (HCS) synthases of cytoplasmic or mitochondrial origin. The region of similarity was centred around the ACYS/GTA sequence in which the Cys residue represents the acetylation site. From this observation we inferred that a protein encoded by *pksorfx2* should be an acyl-enzyme. It may be relevant to this assignment that the same signature can be recognized in chloramphenicol acetyltransferase.

Pksorfx3 and Pksorfx4 represent an example of two hypothetical products having the same function; this was not totally unexpected since weak similarity was observed between the two products. However, only Pksorfx4 showed a clear consensus (LAAMQGHgiGGGfvmg-LfADI) to the signature pattern of the enoyl-CoA hydratases (ECH) of bacterial or mitochondrial origin (conserved region rich in Gly and hydrophobic residues). These enzymes, e.g. FadB of *E. coli*, are involved in fatty acid metabolism and are in turn similar to MenB of *E. coli*, the dihydroxy-naphtoic acid synthase responsible for the synthesis of menaquinone.

Pksorfx5 showed similarity to a large family of prokaryotic and eukaryotic enzymes, which all probably act via ATP-dependent covalent binding of AMP to activate their substrate. Long-chain fatty-acid-CoA ligases (LCFA, e.g. FadD in *E. coli*), enterobactin synthase, polypeptide synthases (GrsA, GrsB, TycA of *Bacillus* spp.) and insect luciferases belong to this family. The region of similarity is very rich in Gly, Ser and Thr, followed by a conserved Lys PROSITE PS00455; Bairoch, 1991). Pksorfx5 is characterized from residues 192–204 by the presence of the signature LLITSGS-TGtPK, the putative ATP-binding site, and from residues 433–437 by the presence of the signature eTGDLgF, the putative ATPase motif. These two motifs are also present in the portion of Pksorfx6 similar to polypeptide synthases (see below) but the similarity between these two ORFs of *pksX* is restricted to these two domains only.

Due to the peculiarities in the nucleotide sequence (see Fig. 1) and codon usage, the *pksorfx6* ORF has been subjected to a deeper analysis with the aim of producing a detailed description of the molecular organization of this large multifunctional polypeptide. This kind of investigation seemed particularly interesting since the similarity of Pksorfx6 to PKS I and to the downstream BACPKSX

Fig. 3. Multiple alignment of the amino acid sequence of Pksorfx6 with *B. subtilis* SrfAORF1, SrfAORF2 and SrfAORF3, and *B. brevis* GrsB. In the RAC alignment, Pksorfx6 from residue 46 to 300 (pksx6rac) was compared to SrfAORF1 (bssrf1s1, 3089–3329) and to SrfAORF2 (bssrf2s1, 3088–3328). In the ATP-binding site alignment, Pksorfx6 from residue 691 to 769 (pksx6abs), was aligned with SrfAORF1 (bssrf1s1, 587–668; bssrf1s2, 1634–1712; bssrf1s3, 2680–2758), SrfAORF2 (bssrf2s4, 594–671; bssrf2s5, 1625–1705; bssrf2s6, 2679–2754), SrfAORF3 (bssrf3s7, 595–670) and to GrsB (bbgrsbP, 594–670; bbgrsbV, 1635–1711; bbgrsbO, 2671–2752; bbgrsbL, 3751–3791). In the AAS alignment, Pksorfx6 from residue 904 to 986 (pksx6aas) was compared to SrfAORF1 (bssrf1s1, 813–895; bssrf1s2, 1853–1934; bssrf1s3, 2887–2967), SrfAORF2 (bssrf2s4, 810–892; bssrf2s5, 1855–1937; bssrf2s6, 2886–2966), SrfAORF3 (bssrf3s7, 813–897) and to GrsB (bbgrsbP, 815–897; bbgrsbV, 1851–1933; bbgrsbO, 2898–2980; bbgrsbL, 3935–4017). In the pantetheine-binding motif, Pksorfx6 from 1065 to 1125 (pksx6pan) was aligned with SrfAORF1 (bssrf1s1, 978–1039; bssrf1s2, 2015–2076; bssrf1s3, 3046–3106), SrfAORF2 (bssrf2s4, 972–1032), SrfAORF3 (bssrf3s7, 975–1035) and to GrsB (bbgrsbP, 978–1039; bbgrsbV, 2014–2073; bbgrsbO, 3060–3119; bbgrsbL, 4098–4159). For GrsB, bbgrsbP, -V, -O and -L, refer to the activating domains for proline, valine, ornithine and leucine respectively. Shaded boxes indicate positions at which amino acids are related on the basis of identity, similarity and chemical properties.

KS

pksx6ks1	ISVAIIIGISNCFE--PAKNHDEWENIRCKESAFFMKERLFRFGIS-KEIAE---NADYVPAKASIIIGKDRFDISFHQ-	75
pksx6ks2	EPVAIVGIGNGE---LAMDIDEWQNEEKKDSITEVVKDRADWREHIGNPD-TDVNKTIIKWGFIDGVAEFDILFHG-	77
pksx6ks3	EGIAVVGLNCFE--LAETLESFWSLSSEIRSSGPIDAEHFGCKTPY-----YAEVIDGVSYFQNDIFLL-	64
pksx7ks3	DAIAIVGMSGRYIGARN-VREYWDNVHARNARDIHTSRADVDKYADPVL-NKKCKVYCKSMMLDIEHFOPLFFN-	77
erya2ks3	DPAAIVSMACPLFGVNTPORLWELIREIGETSIGFTTNGDLARL--HPDPNPITSYVDKGFLLDAAAGDAEFG-	79
pksx6ks1	SKKDAEFMDPQQLMLITHSWFAIEDAYAAQIPQTSVITsasnnysyrallpsdttesletpdgyvswvlaqsgtiPTM-	155
pksx6ks2	SPREADYVDQQLMLITYVWPALEDAVCSPOSISGTGTNIFTTGNTGYKdlfhranlpieghaatghMIPSVG--PNRM	155
pksx6ks3	HEBDVRAMPICAHVVEECVPLLTHAYTPEEKKGKPVWYIgrsqhkpdedslidhaknp--IVT--VGQNYL--AANI	138
pksx7ks3	PPSEAEIMDIQHIIFIQEGYHAFDAVYARTINEKKCVYLIMSNEYGvmlnrqsrn---ATG--NSFAIA--AAR-	150
erya2ks3	SPREAAAMDQQLMLIETSNELVNAKIDPHSIRGTATVFLVAKFGYGedtaaaedvegysVTG--VAPAVA--SGR-	155
pksx6ks1	SKKLGRGPNYFVHANSSSLIGLSAYKSLSDSDVALVGTATLHTESNIGYVHQPLNFSSPHIANDASADIMIG	235
pksx6ks2	SYFLNIEGPNSEVETASSSLVAIHRATVAKQNGCEMAIAGVNTILTEEAHISYSKAGMISTDGRCTASADANFYVR	235
pksx6ks3	SGFFDVRGPNVVDIASSALVYAMMAICARGLIQSAIVGVSLLLSSDASHRLFDRRGILSKHSSPIVDERADVVL	218
pksx7ks3	FYFLNKGIAPIEDASSSLVITLARCAIINIKIDMALVGVSLYLTPESYMCMCEAGMLSPGQCQANDNCANFVP	230
erya2ks3	SYTMGLEGPNISVDIASSSLVALHVAVERKRGSSMAVVGAAVMATPGVFVDFSRQRALAADGRSKAGACADFFGF	235
pksx6ks1	GEGVAVVLLKKAADAVKDDHIYALLREIGVNMGDQKVFYAPSVKGLADVQVQVMNQTKVQESICVYEAHGTGIKL-	315
pksx6ks2	GEGVGMVMLKLEDAERDGNHLYGVIRET-ABNHCRANTLTSNPKAJADLVRAYQADIDASTVTYIEAHGTGTIEL-	314
pksx6ks3	GEGVGMVMLKTVKQALEKCDIHYAVVKAASVNNDERTAPATPNLEAKKEMKDALFKSGKKEDISYLEANGSSIVT	297
pksx7ks3	GEGAGALVLRKQAEADRDIYGLIIESGINQDQKTNQITAPSAKSMMDHERDIYETYGIHESISYVEMHGTGIKQ-	309
erya2ks3	SEGVTLVLLERLSEARRNGHEVLAVVRESALNQDQASNELSAHSGPAQRRIROALESGLBEGDVAEAGTGTIAL-	314
pksx6ks1	DPIELAAITNVIRQYTNKTQ-----PFGIGSVFVNIGHLDTAAGLACIKVVMSTLYQELAPSUNYKEPNPMYLA	386
pksx6ks2	DPIELNGIKAAKELSNMRGEIqpdpvphrRGIGSVFVNIGHLELAAGISRLIKVLQKIKKTIVKSHCETLNPYLLT	394
pksx6ks3	DLLELKATQSVIRs-----GHSS-----PISLGSIFENIGHPLCAEGIASFIKVVMKERRRVPFISGEKENAFQQ	366
pksx7ks3	DPIELBALSTVQEKTDKKQ-----PQAGISVFSNIGHTSAAAGVAIVQVVLQNMKKTIVPTINFTTPNEBFPE	380
erya2ks3	DPIELANALDITGRDRIADR-----PFWLGSVFSNIGHTQAAGVITLLVVLALRNGEIPATIHVEEPTPFWWS	385
pksx6ks1	SSPFYVVDQKKTLSREIKTHR-----AALSIFG-LGTTNTHAF-	425
pksx6ks2	DSPFYIVQEKQDWSVtdrdgnelprAGISSF-XIGSVNAHIMIE	439
pksx6ks3	KANITFSRALPKWTDSQPTAA-----INCRAGG--GTNAHIVE	403
pksx7ks3	HSPLYVNTLKPWETADGKPR-----RACVSSFSYSTNAHIMIE	420
erya2ks3	SGGVALLAGNCPWRRGERTRR-----ARVSAFG-ISGTNAHIVE	424

DH

pksx6dh	AQISG--NQTGigsHPLLHQNTSDFEQKFSSVF----TGDEFTIRAEVVRKPLLEAVYIEAYAINQAAAGSEIG	73
pksx7dh	AEkkTDRSKqerhiHPLLHQNVSDISGVRFRSAF----TGRFPIIKIEIVKIEHELAAALDEVRRAVERAAADQFP	75
ratfasdh	----DHSQTwdip--VAEDFPngsSssatvynidassSSHHIVVACDQRVVFRTGYIYVWKILARSlsIsLE	72
erya2dh	ADVSAIgvrgaehp11AAVDVPghgavftgrls----TDEPPIIAEHVVGRTIVVSVLMDALAAAGEDvgipvLE	75
pksx6dh	QDVRIRlnHTVQVPVVDHRSAQVDISLPEEDKIPFIYSTQEDGI--DPVI---HSQSAELASAAETPVADLTEMS	149
pksx7dh	TGFRLR--NIWVRPFAVTEQQKDIDVRLYPHENIEIFICRDPQSAeSPIV---YGOEAVLCEAGENPVINIEELK	150
ratfasdh	EtpvvfenvtflqatILPRTGTvplev-----RFLASAHAFVSLSGNLIVgkvvweDpdsklfDHpeVpiPae	143
erya2dh	Elvlqrpl-----VLGAGAGALlrmvsgapQSRRIIVHAABVALIADaqwqHATVTLAQGVAAgPrdtEqwPp-	147
pksx6dh	RRCGKGKSPDQFTEGRSRMPFHAIIGKNNVIGNREEL	191
pksx7dh	ASYNGRISPPICEAYTEM-IHT--SLAEDSLYAGENG-	191
ratfasdh	sESV-SRTQGVVKKELRLRYD--HILGGV-----	173
erya2dh	-EDA-VREPLDHHDLAEQCYEY--SSLAARAWRKDDSY	187

Fig. 4. For legend see facing page.

KR		
pkx6kr1	eGVYITITAGLILFLAEIENRTGRSTITGSEVLSE-----DENELEALRSIGREIVYREAD	62
pkx6kr2	dHVLITIGTRHILLCARHPRECYGVKKLVITGHEQLPPEERWARFE-TINTSLR--EPIQAVRELEAKGQEMSLT	77
pkx7kr1	kGVYITITAGLILFIFAEIENQINDAVVITGSEPLDE-----RKKKLKALQKLQIYRQAD	62
pkx7kr2	dKVLITITGIRHILLCARHPKTYGVKRLVITGHEELPPEDWNSVVIS--LAEKIPAVQSELDMGAGQQLSLDLT	77
erya1kr2	gGTILITVITIAHILAEVAEWLGRGAELHLSRSGPDTE-----GVGILTAELTRLGARSISACD--	62
erya3kr1	rGTVLITGSGTHIAHVAPWLRSGAEHLVLRGREGADAP-----GASLREELTALGTGCTAACD--	62
pkx6kr1	VSLQHAIRHLLEIKERYGTINIIH--SAGSSKDRFIHKTNELFQELQIRSLHLHDECSKAF--D-DHIFFS	137
pkx6kr2	LSNDAQMEQTQHIKRTLGPAGVIRKALTDMDTLAFIRKTSDIQRELEFSLTLTYHVCSE--QIVLFS	154
pkx7kr1	LAKQTYDALKETQNVYGDIDHIE--SAGLIKDNFIMKKKEVQTLALPSALIHDDATKII--D-DHIFFS	137
pkx7kr2	--HRAVEQSUKTIHETMGALGAVIRKAMVNQNPAFIRKSLEIGQLEIFNEQLTFELQLSE--LAFPLS	151
erya1kr2	VSSREPARELHGLIQGDVVRVVAALPL--QVAVINDMEAFDEVAAYAGAVHDELCS--ALHILFS	136
erya3kr1	VAIRARAEAVLAERAEGRTSAVMAA--V--TSTPLDDLTEAFTEADDERRTVNDLCP--LAFVFS	136
pkx6kr1	VSGC--LG--NAGQADYAAANSEMLAE-----RRSLAASERFIS-TIRNHWLWEEGGM-QV--Gede	197
pkx6kr2	VSAI--IEISAGQADYAA-----NS-----MDYFAEAKKHAP-IIPYQWNNKETNGEVNQA--	209
pkx7kr1	GAGA--Vg--SAGQADYAMAFMMAE-----RNGQAELEFRYK-TIRNHWLWDCGM-QI--DA--	194
pkx7kr2	SVSA--AIALAGQADYAMAFMYAE-----AKDK-CIVTQWNNKETELge-----	202
erya1kr2	AGVW--G--SARQAYAA-----NAFLDA-----RRHREGRLAATSAWELWAAGQVGTDE-----	186
erya3kr1	AGVW--G--SPGLASVAA-----NAFLDG-----RRRRESEAVTSAWELWAGQNMAGDE-----	186
ACP		
pkx6acp1	Rctspkniqdtvftiwgev--LKTSDHEWDDGFVVGNNLLAVTVADREK-HELSCFESVTDLFESSIKNISQYIT	75
pkx6acp2	QggqqrKLEAALIQMGAILKVNTHDQVSTERSHYFDFVTFTVFNKIN-KFQL-ELTPTIFFEYSQSLAEY--	76
pkx6acp3	S-----ETQSWLIDLFTTEERIDREDPEIIGLAQYVVDNIIAQLQRN-RKLE-AALDPSILYEPIQRFADWLI	72
pkx7acp1	S-----EAEQFVSQLIAEKINKPVEQVKGIVGYYMFLNNSGLLEVVE--SKIGE--SLSPDLLFETIAELSAFLA	72
pkx7acp2	DdeilreKAENYFKQVLSSVKLPAGQIDASAPPEYIINSIMIMHVTGQ--EKVFG-SLSKTLFFEYIDRSLTRY-FI	77
pkx7acp3	E-----ALCDELTAGLAEVYMDQNEIDPSEALITIMDSITGLEWIKAN-KQYG-TSLNVTKVYDEPTRDFAVVLA	72
pkx7acp4	E-----TLHENLRESIADVYMEPYEVDIEAGIITIMDSITGLEWIKAN-KQYG-TSFTVTRVYDPIIRDFAEMLK	72
pkx7acp5	P-----ETVDWLVLTFSDPKIAAEDFETEPQCYIDNIIAQLVQGN-QQLN-GDIDPSILFEPHIESFAHWLI	72
erya2acp3	QNNLELVANAAEVLGH--ESAAENVIRAGSRLDLSLNAMELRKRIS-ASTG-LRLPASLVDFPIVTLAQHLR	74
erya3acp5	RHHLAHLIRAEVAAVLGH--GDDAALDRFRAIRTLFDSMTAVDLNRRA-AVTG-VREAATVVFDEPIITRLADHYL	74

Fig. 4. Multiple alignment of Pksorfx6 with Pksorfx7, EryA products and rat FAS (fatty acid synthase). In the KS alignment, the three putative KS domains of Pksorfx6 (pkx6ks1, 1165–1590; pkx6ks2, 2744–3183; pkx6ks3, 3993–4396) were compared to one of the four putative KS domains of Pksorfx7 (pkx7ks3, 2763–3187) and to one of the KS domains of EryAORF2 (erya2ks3, 30–455). In the DH alignment, Pksorfx6 residues 1765–1956 (pkx6dh), were aligned with the putative dehydratase domain of Pksorfx7 (pkx7dh, 934–1116), of rat FAS (ratfasdh, 837–1110) and of EryAORF2 (erya2dh, 2365–2552). In the KR alignment, two stretches of Pksorfx6 (pkx6kr1, 2253–2449; pkx6kr2, 3575–3783) were compared to the two putative KR domains of Pksorfx7 (pkx7kr1, 1233–1415; pkx7kr2, 3575–3776) and to EryAORF1 (erya1kr2, 3056–3241) and EryAORF3 (erya3kr1, 1115–1300) KR domains. In the ACP alignment, three putative Pksorfx6 ACP domains (pkx6acp1, 1019–1133; pkx6acp2, 2462–2588; pkx6acp3, 3840–3940) were compared to the five putative ACP domains of Pksorfx7 (pkx7acp1, 159–233; pkx7acp2, 1647–1726; pkx7acp3, 2448–2520; pkx7acp4, 2588–2662; pkx7acp5, 3816–3890) and to EryAORF2 (erya2acp3, 1376–1471) and EryAORF3 (erya3acp5, 1338–1416) ACP domains.

ORFs, suggested a modular organization in synthesis units (SU), according to the model first proposed by Donadio *et al.* (1991) for the *eryA* genes.

The putative product of *pkx6* can be divided into two regions of similarity: the 1300 aa of the amino end of the polypeptide are significantly similar to sequences of the large family of enzymes encoding polypeptide synthases (in *Bacillus* spp. the best known are GrsA and B, and TycA and B, encoding peptide antibiotics, gramicidin and tyrocidine, respectively, and SrfA and B, encoding the lipopeptide surfactin; Zuber *et al.*, 1993); and the 3300 aa at the carboxy end, highly similar to PKS I as described above.

Fig. 3 shows some of the multiple alignments between Pksorfx6 and *B. subtilis* SrfAORF1, SrfAORF2 and *B.*

brevis GrsB. In particular, we analysed four key domains of these enzymes. (1) RAC (amino acid racemization domain) represents the multiple alignment between residues 43–300 of Pksorfx6 and the polypeptide sequences at the amino end of SrfAORF1 and SrfAORF2. The latter two are reported by Cosmina *et al.* (1993) as domains involved in racemization of the D-amino acid bound to the nearby amino-acid-binding (ABS) domain. (2) The ATP-binding site refers to the multiple alignment between Pksorfx6, from residues 691–769, the seven putative ATP-binding sites of SrfAORF1 and SrfAORF2, and the four putative ATP-binding sites of GrsB. The same type of similarity extends to the proposed ATP-binding site of TycA (data not shown). In this case site-directed mutagenesis of the consensus sequence SGTGKPGPKG has been demonstrated to reduce the ATP-PP_i amino-acid-

dependent exchange (Gocht & Marahiel, 1994). (3) The ATPase site (AAS) shows the multiple alignment of residues 904–986 of Pksorfx6, the seven putative AASs of SrfAORF1 and SrfAORF2, and the four sites of GrsB. The role of the TGDLA motif in ATPase activity has been demonstrated experimentally for the related TycA enzyme (Gocht & Marahiel, 1994). (4) Finally, the pantetheine-binding motif shows the multiple alignment of Pksorfx6 residues 1065–1125 with five pantetheine-binding domains of SrfAORF1 and SrfAORF2 and the homologous domains of GrsB. Substitution of the Ser residue in each of the three GGH/DS motifs of SrfAORF1 and in the first motif of SrfAORF2, affected surfactin production (D'Souza *et al.*, 1993). The pantetheine-binding motif of Pksorfx6 is reported in Fig. 1 as an ACP motif, since it has similarity to such a motif as well as to the pantetheine-binding site (Fig. 3). This putative site of attachment of an activated molecule is on the boundary between the region of homology to polypeptide synthases and the region of homology to PKSs.

The last three-quarters of Pksorfx6 at the carboxy end, showed striking similarity to the product of the BACPKSX ORF (here referred to as Pksorfx7) and the already mentioned *eryA* gene products. Fig. 4 shows the multiple alignments of the repeated domains of Pksorfx6 with similar domains of Pksorfx7 and with one of the EryA subunits. The following features can be recognized: three putative KS domains, all with good similarity to the DTACSSS motif with the Cys of the active site; two putative dehydratase domains (DH) with similarity to the canonical HxxxGxxxxP active site signature; two putative β -ketoreductase (KR) domains with similarity to the LVTGGTGGLG NADPH-binding site; and three putative ACP domains with similarity to the GGH/DS motif, the first one of which has already been mentioned as the pantetheine-binding motif.

The distribution of these structural regions in Pksorfx6 and Pksorfx7, shown in Fig. 1, suggests strong similarity to the organization of the EryA PKS I. The two deduced polypeptides of 4447 and 4427 aa, respectively, could be organized in synthetic units (centred around the KS domains) in a fashion similar to that reported for the long ORFs of *eryA*. Both polypeptides could be multifunctional subunits of a PKS, able to perform at least seven condensation steps. If valid, the hypothesis that the primary sequence organization of type I PKS reflects the succession of the synthetic steps, Pksorfx6 could be responsible for the first steps of synthesis and Pksorfx7 for the intermediate steps, since no thioesterase domain (TE) has been found in these putative products. The first putative ACP site of Pksorfx6 could be the attachment site of an unusual 'starter' molecule of the polyketide: from the similarity data for the N-terminal of the polypeptide we propose that the starter is an amino acid.

Pksorfx1 to Pksorfx5 suggest the presence of regions encoding functions more or less related to the synthesis or modification of molecules of a polyketidic or fatty-acid nature. For this reason, the locus denomination *pksX*, can be extended to the entire 33.6 kb of the contig described

by Scotti *et al.* (1993) and in this paper. Preliminary sequencing and similarity analysis data indicate that the locus is even broader, expanding in the 5' direction for 10 kb and in the 3' direction for more than 35 kb.

ACKNOWLEDGEMENTS

Supported in part by Consiglio Nazionale delle Ricerche (Grants 92.02183.CT14 and 93.02009.CT14), by P. F. Ingegneria genetica CNR and by the EC (contract no. SCI* 0211-C-EDB and no. BIO2-CT93-0272).

REFERENCES

- Albertini, A. M. & Galizzi, A. (1985). Amplification of a chromosomal region in *Bacillus subtilis*. *J Bacteriol* **162**, 1203–1211.
- Albertini, A. M., Caramori, T., Crabb, W. D., Scoffone, F. & Galizzi, A. (1991). The *flaA* locus of *Bacillus subtilis* is part of a large operon coding for flagellar structures, motility functions, and ATPase-like polypeptide. *J Bacteriol* **173**, 3573–3579.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *J Mol Biol* **215**, 403–410.
- Bairoch, A. (1991). PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res* **19**, 2241–2245.
- Cosmina, P., Rodriguez, F., de Ferra, F., Grandi, G., Perego, M., Venema, G. & van Sinderen, V. (1993). Sequence and analysis of the genetic locus responsible for surfactin synthesis in *Bacillus subtilis*. *Mol Microbiol* **8**, 821–831.
- Donadio, S., Staver, M. J., McAlpine, J. B., Swanson, S. J. & Katz, L. (1991). Modular organization of genes required for complex polyketide biosynthesis. *Science* **252**, 675–679.
- D'Souza, C., Makano, M. M., Corbell, N. & Zuber, P. (1993). Amino-acylation site mutations in amino-acid activating domains of surfactin synthetase: effects on surfactin production and competence development in *Bacillus subtilis*. *J Bacteriol* **175**, 3502–3510.
- Ferrari, E., Henner, D. J. & Hoch, J. (1981). Isolation of *Bacillus subtilis* genes from a Charon 4A library. *J Bacteriol* **146**, 430–432.
- Fuma, S., Fujishima, Y., Corbell, N., D'Souza, C., Nakano, M. M., Zuber, P. & Yamane, K. (1993). Nucleotide sequence of the 5' portion of *srfA* that contains the region required for competence establishment in *Bacillus subtilis*. *Nucleic Acids Res* **21**, 93–97.
- Glaser, P., Kunst, F., Arnaud, M., Coudart, M.-P., Gonzales, W., Hullo, M.-F., Ionescu, M., Lubochinsky, B., Marcelino, L., Moszer, I., Presecan, E., Santana, M., Schneider, E., Schweizer, J., Vertès, A., Rapoport, G. & Danchin, A. (1993). *Bacillus subtilis* genome project: cloning and sequencing of the 97 kb region from 325° to 333°. *Mol Microbiol* **10**, 371–384.
- Gocht, M. & Marahiel, M. (1994). Analysis of core sequences in the D-Phe activating domain of the multifunctional peptide synthetase TycA by site directed mutagenesis. *J Bacteriol* **176**, 2654–2662.
- Hanahan, D. (1985). Techniques for transformation of *E. coli*. In *DNA Cloning: a Practical Approach*, vol. 1, pp. 109–132. Edited by D. M. Glover. Oxford: IRL Press.
- Higgins, D. G., Bleasby, A. J. & Fuchs, R. (1992). CLUSTAL V: improved software for multiple sequence alignment. *CABIOS* **8**, 189–191.
- Katz, L. & Donadio, S. (1993). Polyketide synthesis: prospects for hybrid antibiotics. *Annu Rev Microbiol* **47**, 875–912.
- Kunst, F. & Devine, K. (1991). The project of sequencing the entire *Bacillus subtilis* genome. *Res Microbiology* **142**, 905–912.

- Lloyd, A. T. & Sharp, P. M. (1992). CODONS: a microcomputer program for codon usage analysis. *J Hered* **83**, 239–240.
- Mayorga, M. E. & Timberlake, W. E. (1992). The developmentally regulated *Aspergillus nidulans* *wA* gene encodes a polypeptide homologous to polyketide and fatty acid synthases. *Mol & Gen Genet* **235**, 205–212.
- Ogasawara, N., Nakai, S. & Yoshikawa, H. (1994). Systematic sequencing of the 180 kilobase region of the *Bacillus subtilis* chromosome containing the replication origin. *DNA Res* **1**, 1–14.
- Pearson, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol* **183**, 63–98.
- Perego, M. (1993). Integrational vectors for genetic manipulation in *Bacillus subtilis*. In *Bacillus subtilis and Other Gram-positive Bacteria*, pp. 615–624. Edited by A. L. Sonenshein, J. A. Hoch & R. Losick. Washington, DC: American Society for Microbiology.
- Plunkett, G., III, Burland, V., Daniels, D. & Blattner, R. (1993). Analysis of the *Escherichia coli* genome. III. DNA sequence of the region from 87.2 to 89.2 minutes. *Nucleic Acids Res* **21**, 3391–3398.
- Pustell, J. & Kafatos, F. C. (1984). A convenient and adaptable package of computer programs for DNA and protein sequence management, analysis and homology determination. *Nucleic Acids Res* **2**, 643–655.
- Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: a Laboratory Manual*, 2nd edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.
- Schuler, G. D., Altschul, S. F. & Lipman, D. J. (1991). A workbench for Multiple Alignment construction and analysis. *Proteins Struct Funct Genet* **9**, 180–190.
- Scotti, C., Piatti, M., Cuzzoni, A., Perani, P., Tognoni, A., Grandi, G., Galizzi, A. & Albertini, A. M. (1993). A *Bacillus subtilis* large ORF coding for a polypeptide highly similar to polyketide synthases. *Gene* **130**, 65–71.
- Sharp, P. M., Higgins, D. G., Shields, D. C., Devine, K. M. & Hoch, J. A. (1990). *Bacillus subtilis* gene sequences. In *Genetics and Biotechnology of Bacilli*, vol. 3, pp. 89–98. Edited by M. M. Zukowski, A. T. Ganesan & J. A. Hoch. San Diego: Academic Press.
- Shields, D. C. & Sharp, P. M. (1987). Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. *Nucleic Acids Res* **15**, 8023–8040.
- Sorokin, A., Zumnstein, E., Azevedo, V., Ehrlich, D. & Serror, P. (1993). The organization of the *Bacillus subtilis* 168 chromosome region between the *spoVA* and *serA* genetic loci, based on sequence data. *Mol Microbiol* **10**, 385–395.
- Turgay, K., Krause, M. & Marahiel, M. A. (1992). Four homology domains in the primary structure of GrsB are related to domains in the super family of adenylate-forming enzymes. *Mol Microbiol* **6**, 529–546.
- Zuber, P., Nakano, M. M. & Marahiel, M. A. (1993). Peptide antibiotics. In *Bacillus subtilis and Other Gram-positive Bacteria*, pp. 897–916. Edited by A. L. Sonenshein, J. A. Hoch & R. Losick. Washington, DC: American Society for Microbiology.

Received 13 July 1994; revised 22 September 1994; accepted 11 October 1994.