VIROLOGY 246, 329–340 (1998)
ARTICLE NO. VY989197

Genes and Regulatory Sites of the "Host-Takeover Module" in the Terminal Redundancy of *Bacillus subtilis* Bacteriophage SPO1

Charles R. Stewart,*[1] Irphan Gaslightwala,*[2] Kaede Hinata,*[3] Katherine A. Krolikowski,*[4] David S. Needleman,† Angela Shu-Yuen Peng,* Mark A. Peterman,*[5] Angela Tobias,*[6] and Ping Wei*[7]

Department of Biochemistry and Cell Biology, Rice University, P.O. Box 1892, Houston, Texas 77251-1892; and †Molecular Genetics Core Facility, Department of Microbiology and Molecular Genetics, University of Texas-Houston Medical School, Houston, Texas 77030

Received February 3, 1998; returned to author for revision March 26, 1998; accepted April 16, 1998

Early in infection of *Bacillus subtilis* by bacteriophage SPO1, the synthesis of most host-specific macromolecules is replaced by the corresponding phage-specific biosyntheses. It is believed that this subversion of the host biosynthetic machinery is accomplished primarily by a cluster of early genes in the SPO1 terminal redundancy. Here we analyze the nucleotide sequence of this 11.5-kb "host-takover module," which appears to be designed for particularly efficient expression. Promoters, ribosome-binding sites, and codon usage statistics all show characteristics known to be associated with efficient function in *B. subtilis.* The promoters and ribosome-binding sites have additional conserved features which are not characteristic of their host counterparts and which may be important for competition with host genes for the cellular biosynthetic machinery. The module includes 24 genes, tightly packed into 12 operons driven by the previously identified early promoters P_E1 to P_E12. The genes are smaller than average, with half of them having fewer than 100 codons. Most of their inferred products show little similarity to known proteins, although zinc finger, trans-membrane, and RNA polymerase-binding domains were identified. Transcription–termination and RNase III cleavage sites were found at appropriate locations.

INTRODUCTION

The genes of *Bacillus subtilis* bacteriophage SPO1 have been categorized as early, middle, or late. Early genes are transcribed by the unmodified host RNA polymerase from promoters recognized by *B. subtilis* σ^A. Transcription of the middle and late genes requires replacement of σ^A by a succession of phage-specified σ factors, with different promoter specificities. In general, the middle genes specify the DNA replication machinery, while the late genes specify the structural and morphogenetic proteins (Gage and Geiduschek, 1971; Fujita *et al.,* 1971; Talkington and Pero, 1977; Losick and Pero, 1981; Stewart, 1993).

The primary role of the *early* genes appears to be the subversion of the host's biosynthetic machinery to the purposes of the infecting bacteriophage (Wei and Stewart, 1993). Host mRNA, protein, and DNA syntheses are rapidly shut off and are replaced by the highly efficient synthesis of the corresponding phage-specific macromolecules (Shub, 1966; Gage and Geiduschek, 1971; Reeve *et al.,* 1978; Heintz and Shub, 1982). The large majority of the early genes are clustered in the rightmost 11.5 kb of the 12.4-kb terminal redundancy (Pero *et al.,* 1979; Romeo *et al.,* 1981; Brennan *et al.,* 1981; Perkus and Shub, 1985). Many of these genes are cytotoxic when expressed individually in uninfected cells (Curran and Stewart, 1985; Wei and Stewart, 1993; unpublished observations, this laboratory), so a major part of their role is undoubtedly to specify proteins that shut off essential host functions. However, it also seemed likely that the genes themselves would be designed to compete effectively with host genes for the cellular machinery for transcription and translation. In fact, this region was known to include a large number of unusually active early promoters (Lee *et al.,* 1980; Romeo *et al.,* 1981).

Here we analyze the nucleotide sequence of this "host-takeover module" and confirm that its genes are indeed designed for efficient expression. The promoters, ribosome-binding sites (RBSs), and codon-usage statistics show features that are known to be important for highly efficient function in *B. subtilis.* The promoters and RBSs show other conserved features which are *not* typical of their host counterparts and which therefore might

[1] To whom correspondence and reprint requests should be addressed. Fax: (713) 285-5154. E-mail: crs@bioc.rice.edu.

[2] Present address: Baylor College of Medicine, Houston, TX 77030.

[3] Present address: Department of Dermatology, Stanford University School of Medicine, Stanford, CA 94305.

[4] Present address: Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138.

[5] Present address: University of Texas Southwestern Medical Center, Dallas, TX 75235.

[6] Present address: Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030.

[7] Present address: The Salk Institute for Biological Studies, 10010 N. Torrey Pines Road, La Jolla CA, 92037.

0042-6822/98 $25.00

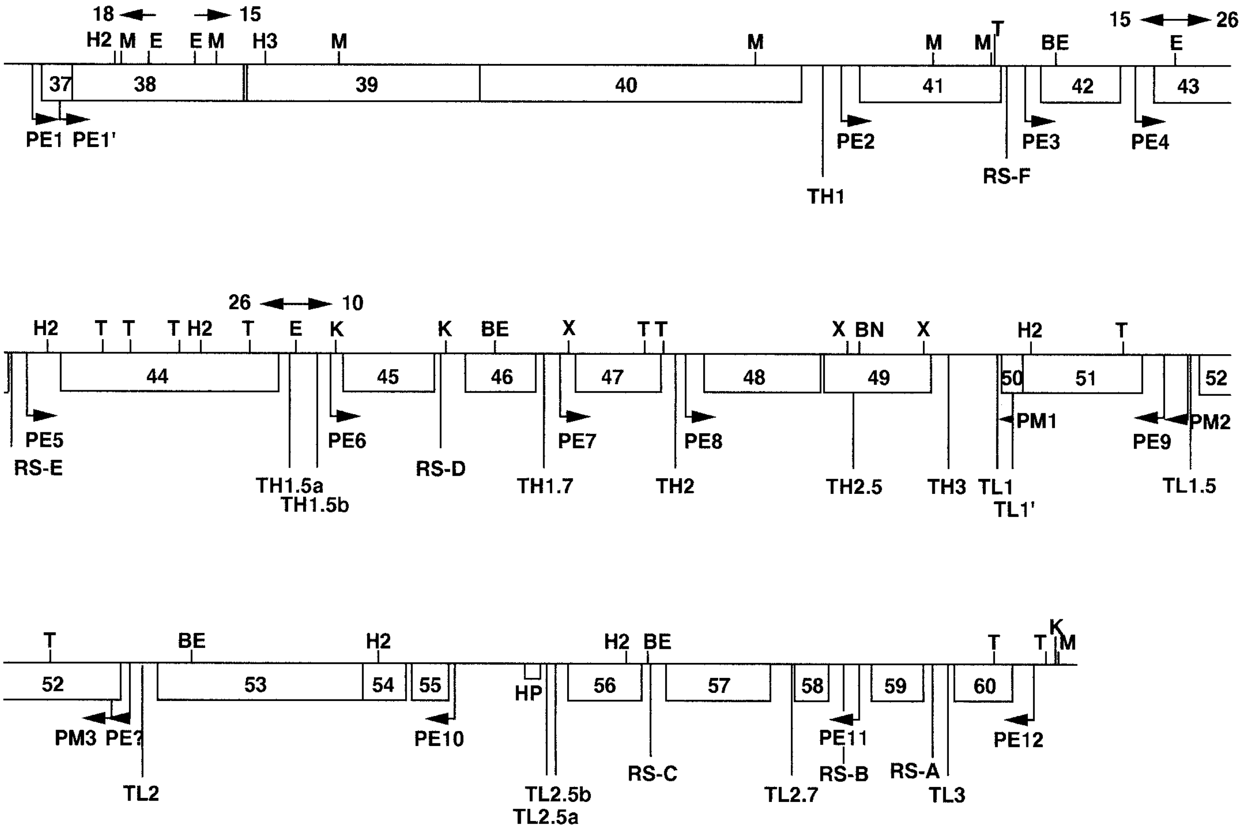


FIG. 1. Map of the early gene region of the SPO1 terminal redundancy (the "host-takeover module"). The three lines show, respectively, nucleotides 1–4000, 4001–8000, and 8001–11,500, with nucleotide 11,500 being the right end of the SPO1 genome. Genes 37 to 60 are represented by boxes below the line. P_E1 to P_E12 are early promoters; P_M1 to P_M3 reprsent the possible middle promoters P_MIII-1 to P_MIII-3. For each promoter, the arrow shows the direction of transcription, and the vertical line is at the start site. For each transcription terminator (T . . .) or RNase III site (RS . . .), the vertical line is at the position of the end of the resulting mRNA molecule. HP represents the homopolymer. Restriction sites are indicated above the line. BE, *Bst*EII; BN, *Bst*N1; E, *Eco*RI; H2, *Hae*II; H3, *Hae*III; K, *Kpn*I; M, *Msp*I (=*Hpa*II); T, *Taq*I; X, *Xba*I. Arrows starting above the *Eco*RI sites point to the numbers of the resulting *Eco*RI* fragments. With a few exceptions, all previously reported restriction sites (Pero *et al.,* 1979; Lee *et al.,* 1980; Brennan *et al.,* 1981; Romeo *et al.,* 1981; Lee and Pero, 1981; Brennan and Geiduschek, 1983; Curran and Stewart, 1985; Perkus and Shub, 1985; Greene and Geiduschek, 1985; Stewart, 1993) were located approximately at their expected positions. There are no *Bam*HI, *Bgl*II, *Cla*I, *Pst*I, *Sal*I, or *Sma*I sites in this sequence.

have evolved for a specific competitive advantage. Since hydroxymethyluracil (hmUra) replaces thymine in the SPO1 genome, some of the promoter features may have been conserved for their usefulness in that context.

The 24 genes identified in this sequence are remarkable for their relatively small size and for their lack of homology to other known genes.

RESULTS AND DISCUSSION

Sequence

The sequence of 11,500 nucleotides has been submitted to the GenBank nucleotide sequence database and has been assigned Accession No. AF031901. Several portions of the sequence had been reported previously (Lee *et al.,* 1980; Lee and Pero, 1981; Brennan and Geiduschek, 1983; Yansura and Henner, 1984; Greene *et al.,* 1986; Wei and Stewart, 1993), and a few corrections of those earlier versions are included. Figure 1 shows a map of this 11.5-kb region, and Tables 1 and 2 describe the genes and other sites included therein. Twenty-four genes were identified, arranged in 12 operons driven by early promoters P_E1 through P_E12 (Romeo *et al.,* 1981). Apparent transcription–termination sites were located after 8 of the operons and apparent RNase III-cleavage sites after 3 others. Other such sites were located between genes of multigene operons. Some of these genes and sites will be identified below as those whose activity has been demonstrated previously. For others, there is no direct evidence of functionality.

Identification of genes

Table 1 lists the 24 putative genes, numbered 37 to 60 in extension of the numbering begun by Okubo *et al.* (1972). These all have strong and properly located ribosome binding sites (Mountain, 1989) and more than 20 codons. All are located completely within 1 of the 12 operons defined above, and they almost completely fill the space in each of those operons, with

TABLE 1

Genes

Gene[a]	No. codons[b]	ΔG[c]	CAI[d]	Protein[e]
37	32	−21.4	0.29	*
38	191	−19.0	0.48	e9++++
39	256	−21.0	0.42	e6++++
40	351	−23.8	0.42	e4++++
41	155	−18.0	0.37	e15++++
42	89	−18.2	0.46	e18+++
43	90	−18.0	0.38	e20′+
44	238	−16.0	0.55	e3++++
45	100	−18.4	0.45	e17+
46	78	−12.0	0.48	e21′+
47	95	−18.0	0.38	e19+
48	127	−16.6	0.52	
49	118	−19.0	0.37	
50	24	−17.8	0.28	*
51	132	−16.2	0.51	e16+++
52	159	−18.0	0.47	e12+++
53	222	−17.6	0.43	e7++++
54	47	−21.0	0.50	*
55	41	−16.2	0.53	*
56	80	−18.4	0.59	e21++
57	114	−19.8	0.48	
58	36	−17.4	0.56	*
59	57	−18.0	0.47	*
60	74	−12.8	0.54	e20++

[a] SPO1 genes 1 to 36 were mapped previously (Okubo *et al.,* 1972; Cregg and Stewart, 1978; Stewart, 1993). The genes in the present sequence, located at the right end of the genome, were assigned numbers 37 to 60, in order from left to right. Genes 43 and 44 were previously called e22 and e3 (Wei and Stewart, 1993). Promoters P_E1 through P_E8 direct transcription rightward through genes 37–49. P_E9–P_E12 direct transcription leftward through genes 60–50. The precise position of each gene and promoter in the sequence is noted in the GenBank file, Accession No. AF031901.

[b] Number of codons, including initiation and termination codons.

[c] Free energy of binding between the RBS for each gene and the 3′ end of *B. subtilis* 16S rRNA.

[d] Codon adaptation index (Sharp and Li, 1987). The values were calculated by Dr. Paul Sharp.

[e] This column indicates which protein, of those defined by a band on SDS–PAGE (Heintz and Shub, 1982), could be identified from the data of Perkus and Shub (1985) as being specified by that particular gene. (++++) This assignment is definitive. (+++) The preponderance of the evidence favors this assignment, but an alternative interpretation of one element of the data from Perkus and Shub (1985) was required. (++) This is the only gene of appropriate size in the region in which the gene specifying this protein must be located, but the data on restriction inactivation were equivocal. (+) This assignment is based only on the correspondence between the length of the gene and the size of the protein as estimated from its SDS–PAGE position. (*) The product of this gene would be too small to account for any of the defined bands; however, it could be in the smear of undifferentiated bands near the bottom of the gel. e20′ and e21′ are bands that can be seen adjacent to e20 and e21 but to which no name was assigned (Heintz and Shub, 1982).

little or no overlapping. ORFs beginning at bases 1077, 1347, 9547, 10,246, and 10,625 were not included because they have weak ribosome-binding sites which are located either in the middle of another gene or outside of any transcription unit. ORFs beginning at 1807 and 5216 were not included despite strong ribosome-binding sites, because they constitute in-frame subsets of longer ORFs.

TABLE 2

Characteristics of Transcription–Termination and RNase III Cleavage Sites

Site[a]	ΔG[b]	Stem length[c]	Loop sequence[d]	3′ Tail sequence[e]
Region of rightward transcription				
Terminator TH1	−14.6	7	GUACU	UAUUUUUUU
RNase III site RS-F	−24.6	19	UAAAA	
RNase III site RS-E	−13.8	15	CCUGAAA	
Terminator TH1.5a	−22.8	9	UCAA	UAGUUU
Terminator TH1.5b	−15.6	7	CUACA	UUUGU
RNase III site RS-D	−24.8	22	UCAAA	
Terminator TH1.7	−14.6	7	AUAAA	AUUAUUUUUU
Terminator TH2	−28.2	8	UCUU	ACUUUUU
Terminator TH2.5	−13.6	9	CUAG	AGUUU
Terminator TH3	−26.0	13	UUAGUU	UUUUU
Region of leftward transcription				
Terminator TL1	−19.2	9	GAAA	UUUU
Terminator TL1′	−14.0	8	CUAU	UUUCUUUU
Terminator TL1.5	−22.8	12	AUAG	GUUUU
Terminator TL2	−23.8	9	AGCUUAC	AAUCUUUUU
Terminator TL2.5a	−3.0	8	AGUCU	UCUUUU
Terminator TL2.5b	−13.4	7	GUUUC	UUUU
RNase III site RS-C	−22.8	19	GCUAC	
Terminator TL2.7	−12.6	8	GUAGG	UUU
RNase III site RS-B	−17.4	19	UACUA	
RNase III site RS-A	−18.2	15	AAGC	
Terminator TL3	−16.0	9	AGUAUA	CCCGCUU
SP82 RNase III sites[f]				
A	−18.2	15	AAGC	
B	−15.0	19	UACUA	
C	−15.2	19	UGAGC	

[a] Each transcription termination site (T . . .) and RNase III cleavage site (RS . . .) is listed in order of its position in the sequence. Most of the terminators have the names given when originally described (Romeo *et al.,* 1981; Brennan *et al.,* 1981; Brennan and Geiduschek, 1983). Terminator T_L2.5 is actually two adjacent sequences, which we have named T_L2.5a and T_L2.5b. Five additional sites, whose termination activity has not been demonstrated, were named T_H1.5a, T_H1.5b, T_H1.7, T_H2.5, and T_L2.7, maintaining consistency with the earlier nomenclature. The putative RNase III cleavage sites were named RS-A through RS-F, with RS-A, -B, and -C being homologous to SP82 sites A, B, and C, respectively (Panganiban and Whiteley, 1983a). The precise position of each site in the sequence is noted in the GenBank file, Accession No. AF031901.

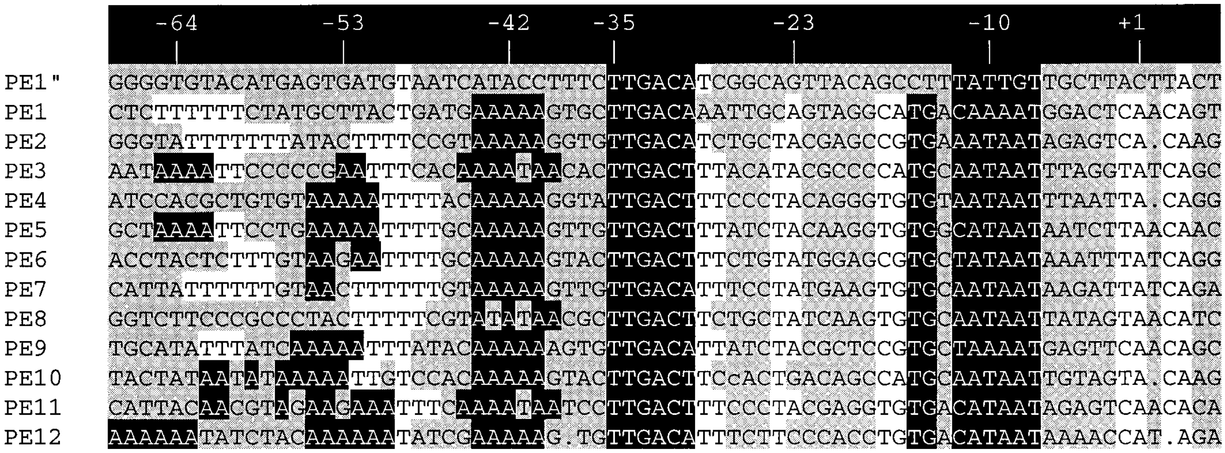
[b] Free energy of formation of the hairpin in each of the termination or RNase III cleavage sites.

[c] The number of base pairs in the stem of the hairpin. In each of the terminators, all of the base pairs in the stem are consecutive, but all of the RNase III site stems include a number of unpaired bases that were not counted in the stem length.

[d] 5′→3′ sequence of the hairpin loop.

[e] 5′→3′ sequence of the unpaired bases immediately adjacent to the 3′ end of the stem.

[f] The parameters of the known SP82 RNase III sites (Panganiban and Whiteley, 1983a) are shown for comparison.



```
          -64        -53        -42    -35         -23          -10      +1
PE1"  GGGGTGTACATGAGTGATGTAATCATACCTTTCTTGACATCGGCAGTTACAGCCTTTATTGTTGCTTACTTACT
PE1   CTCTTTTTTCTATGCTTACTGATGAAAAAGTGCTTGACAAATTGCAGTAGGCATGACAAAATGGACTCAACAGT
PE2   GGGTATTTTTTTATACTTTTCCGTAAAAAGGTGTTGACATCTGCTACGAGCCGTGAAATAATAGAGTCA.CAAG
PE3   AATAAAATTCCCCCGAATTTCACAAAATAACACTTGACTTTACATACGCCCCATGCAATAATTTAGGTATCAGC
PE4   ATCCACGCTGTGTAAAAATTTTACAAAAAGGTATTGACTTTCCCTACAGGGTGTGTAATAATTTAATTA.CAGG
PE5   GCTAAAATTCCTGAAAAATTTTGCAAAAAGTTGTTGACTTTATCTACAAGGTGTGGCATAATAATCTTAACAAC
PE6   ACCTACTCTTTGTAAGAATTTTGCAAAAAGTACTTGACTTTCTGTATGGAGCGTGCTATAATAAATTTATCAGG
PE7   CATTATTTTTTGTAACTTTTTTGTAAAAAGTTGTTGACATTTCCTATGAAGTGTGCAATAATAAGATTATCAGA
PE8   GGTCTTCCCGCCCTACTTTTTCGTATATAACGCTTGACTTCTGCTATCAAGTGTGCAATAATTATAGTAACATC
PE9   TGCATATTTATCAAAAATTTATACAAAAAAGTGTTGACATTATCTACGCTCCGTGCTAAAATGAGTTCAACAGC
PE10  TACTATAATATAAAAATTGTCCACAAAAAGTACTTGACTTCcACTGACAGCCATGCAATAATTGTAGTA.CAAG
PE11  CATTACAACGTAGAAGAAATTTCAAAATAATCCTTGACTTTCCCTACGAGGTGTGACATAATAGAGTCAACACA
PE12  AAAAAATATCTACAAAAAATATCGAAAAAG.TGTTGACATTTCTTCCCACCTGTGACATAATAAAACCAT.AGA
```

FIG. 2. Early promoter sequences. The sequences of the 13 demonstrably functional early promoters are shown. T represents hmUra. The sequences known to be important for efficient expression (−10, −35, upstream poly(A)s, and TG at −14, −15) are highlighted in black; the other conserved sites discussed in the text are highlighted in white. The transcription start site, labeled +1, is assumed to be the A located 7 bases downstream of the −10 region. The validity of this assumption has been demonstrated only for P_E4 and P_E5 (Lee *et al.,* 1980; Lee and Pero, 1981).

It is likely that the protein products of all of genes 37–60 are synthesized during infection, although this was demonstrated definitively only for genes 38, 39, 40, 41, 44, and 53. As indicated in Table 1, those six genes could be identified, from the data of Perkus and Shub (1985), as specifying proteins whose *in vivo* synthesis was demonstrated by their appearance as specific bands on SDS–PAGE (Heintz and Shub, 1982). For others, the identification was less certain, and for some, there was no experimental evidence of functionality. Three seemed particularly doubtful: genes 46 and 60, because of the relative weakness of their ribosome-binding sites, and gene 50, because it is so short and spans a transcription–termination signal. We have nevertheless included genes 46 and 60, because each fills about half of the space in its two-gene operon and because there is fairly good evidence that gene 60 specifies a protein whose expression *in vivo* has been documented (Table 1). We have included gene 50 because it *does* have a strong ribosome-binding site, because it is located in a position where it could be coupled translationally with gene 51, because the transcription terminator is only conditionally effective, and another more efficient terminator is located just at the end of gene 50, and because of its similarity to portions of known genes (see below). It appears unlikely that expression of any of the genes would be blocked by secondary structure in the mRNA. Although several of the mRNAs include sequences that theoretically could hybridize to certain RBSs, the hairpins formed would be much less stable than the binding of those RBSs to 16S rRNA, and two of the genes involved are among those definitively shown to be highly expressed.

The products of genes 39 and 40 show 47% identity among 191 aligned amino acids, suggesting an evolutionary relationship between these adjacent genes.

Features suggesting design for efficient expression

Promoters. Figure 2 shows the sequences of the 13 early promoters, which were found at the positions expected from their *in vitro* characterization (Romeo *et al.,* 1981). P_E1' is by far the weakest *in vitro,* which can be accounted for by two major deviations from consensus in the −10 region, as well as the absence of nearly all of the conserved accessory sites to be discussed below. Located just downstream of P_E1, its functional role probably is negligible, and it will be excluded from the following discussion.

The sequences of the other 12 promoters show features that presumably account for their observed high activity. Their −35 and −10 regions are much more concordant with the consensus for *B. subtilis* σ^A promoters (−35 TTGACA; −10 TATAAT) than are those of the average *B. subtilis* σ^A promoter (Helmann, 1995). All have at least one poly(A) tract, spanning position −42. Some have a second and some a third, spanning −53 and −64, and some have poly(T) tracts between. Such regularly spaced poly(A) tracts were found in a significant fraction of *B. subtilis* promoters and have been correlated with increased expression (Helmann, 1995). All 12 promoters have the TG at −15 and −14, which is conserved significantly among *B. subtilis* promoters (Helmann, 1995) and whose mutation decreased expression of two promoters in *B. subtilis* (Voskuil *et al.,* 1995). All 12 had exactly 17 spaces between the −35 and −10 regions and exactly 6 spaces between the −10 region and the A at the putative start site. (These features are also characteristic of the strongest T4 early promoters (Wilkens and Ruger, 1996). The TG is also conserved in certain *Escherichia coli* promoters and is responsible for the increased activity of their "extended minus 10" regions (Keilty and Rosenberg, 1987).)

```
RBS37         ~AGAGAACAAGGAGGGGUGU.ACAUGAGUGAU~~..........................
RBS38         ~AUAUUAAGAGGAGGAGAUAAUCGUGUCAGUG~~..........................
RBS39         ~AGAAGAUGAGGAGGAAU.CAAUAUGGAACUA~~..........................
RBS40         ~CCUAAGAAAGG.GGU.AU.AAUAUGCAUAUU~~..........................
RBS41         ~AUUGUAAAAGGAG.AGGUUAUAUAUGGAAAAG~..........................
RBS42         ~AUUGUAAAAGGAG.AGGUUCACAAUGAGAAAA~..........................
RBS43         ~AGCGUAAAAGGAG.CGGUUAACAUGAUCCAA~~..........................
RBS44         ~AAUGUAAAAGGAG.AGAUUUCAAAUGGCUAAA~..........................
RBS45         ~GCUCCACUAGGGGGAAUUGGACAUGAUGAUG~~..........................
RBS46         ~UAGGCGAGAGGAG.AAUUUAUUAUGAUGACA~~..........................
RBS47         ~AAUGAAAAAGGAG.AGGUUAGUAUGGAUUGG~~..........................
RBS48         ~AUAAACAAGGGAG.CGGUUGUAAUGCCAUAU~~..........................
RBS49         ~UACUGCAAAGGAGGAAUAAUACAUGAUUAAA~~..........................
RBS50         ~GGCGCUUAAGGAGGCAGGUAAGCUUGGAUAAA~..........................
RBS51         ~UAGGCAGAAGGAG.AAGAUAACGAUGGCAAAA~..........................
RBS52         ~UACAUAUAAGGAGGAGAAAAAUAAUGACACAC~..........................
RBS53         ~CCUAUUAAGGGAGGAAGAUAAGCAUGAGAACA~..........................
RBS54         ~UAUGUAAAAGGAG.AGUAUCACACAUGGUAAUC..........................
RBS55         ~UAGUACAAGGGAGGAGAACAUCUUGUUCAAA~~..........................
RBS56         ~ACAAUCACAGGGGGAAUAUACAUAUGUUUAAA~..........................
RBS57         ~CAAGUAAAAGGAG.AGGUCACAAUGACAUUA~~..........................
RBS58         ~CAGGCAGAAGGAG.AAUGUACGAUGACAUUA~~..........................
RBS59         ~AUUGUAAAAGGAG.AGGUUAACAUGAAGAAA~~..........................
RBS60         ~CAAAUGAAUGGAG.AGAUUGAUAUGCUGAAU~~..........................
"Ideal_RBS"   ~~~~~AGAAAGGAGGUGAUC~~~AUG~~~~~~~~~..........................
```

FIG. 3. Ribosome-binding sites. The sequences of the 24 RBSs are shown. The [GGAGG] core and the initiation codon are highlighted in black. Other conserved features discussed in the text are highlighted in gray.

Differences among the 12 promoters in observed *in vitro* activity can also be explained by specific differences in their sequences. P_E1, 2, and 9 are the weakest of the 12 promoters (Romeo *et al.,* 1981). Two of those three have an A instead of the consensus T at −10. Otherwise, all 12 conform exactly to the *B. subtilis* consensus for the −10 and −35 regions, except that 7 of the strongest promoters have a T instead of the consensus A at −30, and all but 2 of the promoters lack the consensus T at −12. Thus, the latter two deviations from consensus may cause *increased* efficiency, perhaps specifically for hmUra-containing DNA.

Several other sites, not noticeably conserved among *B. subtilis* promoters (Helmann, 1995), were strongly conserved among the SPO1 early promoters. They include: TT at −29 and −28; TA at −24 and −23; YR at −17 and −16; YA at −1 and +1; and CA at +2 and +3 or +3 and +4. The conservation of these elements suggests that they may contribute to promoter efficiency, perhaps specifically in the context of hmUra-containing DNA. $\underline{T}$ $\underline{A}$ (actually $\underline{\text{hmUra}}$ $\underline{A}$) dinucleotides have been inferred to be necessary for certain bends in hmUra-containing DNA (Grove *et al.,* 1997). Thus, the conserved TA at −24, −23 might contribute to the bending necessary for optimal wrapping of the promoter around the RNA polymerase (Buc, 1986; Travers, 1993).

Ribosome-binding sites. The sequences of the 24 RBSs are shown in Fig. 3. On the basis of parameters of known significance, these RBSs appear to be particularly efficient. Their sequences are equal to those of the seven highly expressed *B. subtilis* genes analyzed by Mountain (1989) with respect to: the average number of bases identical to the central 12 bases of the "ideal RBS"; the proportion of A and U residues in the "AU-rich spacer;" and the spacing between the [GGAGG] core and the initiation codon. The average of their ΔG values for binding to the complementary 16S rRNA is −18.0, compared with −17.0 for a set of 12 well-documented *B. subtilis* RBSs (Hager and Rabinowitz, 1985).

Some of these ribosome-binding sites share other features that distinguish them even from the strong *B. subtilis* RBSs. To assess the significance of these features, we prepared databases A to D for comparison, as described under Materials and Methods: (A: highly expressed *B. subtilis* RBSs. B and B′: random *B. subtilis* RBSs. C: other SPO1 RBSs. D: non-RBS sequences.)

The SPO1 early RBSs have several conserved deviations from complementarity to the 3′ end of 16S rRNA: (i) 19 of the 24 have an A immediately following the [GGAGG] core; only one has the U expected for the "ideal" RBS. Among the 217 RBSs in databases A, B, and C, there are 99 As and 78 Us in that position. (ii) Nine of the 24 include the sequence 5′GGAGAGGUU3′ (or a sequence that differs from that by only one base), instead of the "ideal" GGAGGUGAU. Only four sequences within one mismatch of GGAGAGGUU were found in databases A, B, and C. (iii) Seven of the 24 include the sequence GUAAAA at the 5′ end instead of the "ideal" AGAAA. Only 2 GUAs in that position were found in databases A, B, and C. The conservation of these deviations suggests that they may provide a competitive advantage that compensates for their cost in stability of binding to 16S rRNA.

Fifteen of the 24 RBSs (63%) have sequences capable of forming a hairpin, with the [AAAGGAGG] portion of the RBS in the loop and the initiation codon at the end of the stem. Table 3 shows several examples. Hairpins also occur in the other databases, but are not located as precisely as those for the SPO1 early RBSs, as shown in Table 4. The existing data are consistent with a functional role for these hairpins. Of the nine SPO1 early proteins whose expression was observed by Heintz and Shub (1982), and which reasonably can be identified with specific genes (++++ or +++ in Table 1), the five most actively expressed (38, 39, 41, 42, and 44) are among those whose RBSs can form hairpins, and the two with the lowest expression (51 and 52) are not, even though genes 38 and 39 are expressed from P_E1, the promoter with the lowest activity *in vitro* (Romeo *et al.,* 1981). Most of the predicted hairpins are so weak that they would not be stable in solution, so any functional role remains speculative. However, it is conceivable that hairpin formation permits or stabilizes an optimal interaction with the ribosome, such as optimal placement of the initiation codon.

TABLE 3

Structures of Hairpins from SPO1 Early RBSs

Gene	Structure[a]	Position[b]
37	CAAGAG A ACAUG A ||||| G UGUACAUG GAGGGG	106 133
38	GAGA G AUUA A |||| G UAAUCGUG GAGA	209 230
39	GAAGAUG A UUGG U |||| G AAUCAAUAUG AGGAGG	765 794
41	GAA G AAUGU A ||||| G UUAUAUAUG AGG	2776 2798
42	GGAAAA A UGUUA G ||||| A ACAAUG GGUUC	3365 3388
44	GUAAGACA U CUAAA A ||||| A GAUUUCAAAUG AAGGAGA	4167 4200
49	AACGUC A AUUGUG G |||||| G UAAUACAUG AGGAA	6661 6689
53	GGAA G UUAUCCG A |||| || G GAUAAGCAUG GAA	9200 9174

[a] Hairpin structures that could be formed by the sequences around the RBSs of the indicated SPO1 early genes. In each case, the [AAAGGAGG] sequence (in the loop) and the initiation codon (at the right end of the stem) are shown in bold letters, underlined.

[b] The position of the ends of each hairpin within the 115-kb sequence.

In 9 of our 24 genes the third codon specifies lysine. However, lysine also appears with high frequency as the third (and also as the second or fourth) codon in databases A, B, and C. The lysine codon AAA caused a moderate increase in translational efficiency in *E. coli,* when substituted for either CAG or GCU as the second codon in a hybrid gene (Ringquist *et al.,* 1992). We are not aware of any information on the effect of AAA in the third position, and we suppose that this feature is as likely to be meeting the needs of protein structure as of translational efficiency.

Codon usage bias. Codon adaptation index (CAI) values suggest that many of the genes in this cluster have evolved codon usage patterns consistent with efficient expression and that this evolution has proceeded further than it has for *E. coli* phages T4 and T7. The CAI measures how closely the bias in codon usage for any particular gene approximates the bias shown by a set of highly expressed genes of that species. High CAI values have been correlated with high levels of expression in *B. subtilis* and other species (Sharp and Li, 1987; Shields and Sharp, 1987). Among the *B. subtilis* genes analyzed, 11 of 56 (<20%), including the 7 highly expressed genes used as the reference set, had CAI values ≥0.5 (Shields and Sharp, 1987). Of the SPO1 early genes 33% (8/24) have CAI values ≥0.5 (see Table 1). Sixty-three percent (15/24) have CAI values ≥0.45, compared with 39% (22/56) for *B. subtilis* genes. By contrast, only 2 of 174 T4 genes and 3 of 60 T7 genes had CAI values ≥0.5, although the distribution of CAI values for *E. coli* genes

TABLE 4

Proportion of Ribosome-Binding Sites with Precisely Located Hairpins[a]

SPO1 early RBSs	*B. subtilis* RBSs (databases A and B′)	Other SPO1 RBSs (database C)	Non-RBS sequences (database D)
37.5% (9/24)	16.7% (5/30)	18.2% (4/22)	5.0% (1/20)

[a] The percentage of the ribosome-binding sites in each database that were able to form a hairpin in which all or all but one of the [AAAGGAGG] nucleotides were located in the loop and in which the distance between the end of the stem and the first nucleotide of the initiation codon was less than or equal to one nucleotide. For the non-RBS sequences in database D, the [AAAGGAGG] region was considered 9–16 nucleotides upstream from the initiation codon.

was similar to that for *B. subtilis* (Sharp and Li, 1987; Cowe and Sharp, 1991). (The T4 dataset used was incomplete, and T4-coded tRNAs could have distorted some of the CAI values.)

Other regulatory sites

Transcription terminators. Fifteen stem-loop structures similar to ρ-independent transcription terminators (Roberts, 1996) are identified in Fig. 1 and described in Table 2. They include each of the sites previously shown to cause termination of *in vitro* transcription (Brennan *et al.,* 1981; Brennan and Geiduschek, 1983) and five additional sites. T_H3 and T_L1, the only ones that were unconditionally efficient *in vitro* (Brennan *et al.,* 1981), are also the only ones that have both a large negative free energy of formation of the hairpin and Us in at least the first four positions of the 3′ tail. However, other demonstrably active terminators have relatively weak stems ($T_L2.5$ a and b) or a paucity of Us in those positions (T_L2). Each of the sites is probably functional *in vivo,* since most are located immediately after the last gene in an operon, and the others are downstream of at least one gene in their respective operons. The latter presumably terminate a fraction of their operons' transcripts, as a way of regulating the relative expression of different genes in the operon. (T_L3 may be an exception since its 3′ tail is so deficient that its observed *in vitro* activity might depend upon the adjacent hairpin structure of RNase III site RS-A.) All four of the terminators tested *in vivo* were functional, including two that are located within operons (Brennan and Geuduschek, 1983).

RNA processing sites. Early RNAs of SPO1 (and of the closely related phage SP82) are processed by cleavage at specific sites by *B. subtilis* RNase III (Downard and Whiteley, 1981; Panganiban and Whiteley, 1983a,b; Hue *et al.,* 1995). The three SP82 sites that were characterized consist of stable hairpins, with stems having 15–19 base pairs, but interrupted by 7–13 unpaired bases. The counterparts of the three SP82 sites are identifiable in the SPO1 sequence, with stems that are either identical to, or more stable than, those of their SP82 homologs. Sequences yielding similar hairpins are located at three other sites in the SPO1 sequence. The six putative RNase III cleavage sites are identified in Fig. 1 and described in Table 2. Although there is no direct evidence that these are the sites that are cleaved *in vivo,* their locations are appropriate for meaningful function. Three are at the end of operons that lack evident transcription–termination sequences, and the others are between genes in multigene operons. (RNase III sites RS-A, -B, -C, and -D, but not -E and -F, share extensive similarity in the sequence of the stem (Panganiban and Whiteley, 1983a). There is no evidence that that particular sequence is important for the activity of *B. subtilis* RNase III, which has been shown to cleave effectively an RNA molecule with a very different sequence (Mitra and Bechhofer, 1994). However, sites E and F also have strings of three Us within one or two bases of the 3′ end of the stem. It is conceivable that they act as transcription–termination sites instead of, or in addition to, RNase III cleavage sites, although that would be extremely unusual, considering the many unpaired bases within each stem.)

Other promoters. Perkus and Shub (1985) showed that e16, the apparent product of gene 51 (Table 1), was also expressed from a middle promoter. Of the three sequences in this region that come close to the middle promoter consensus sequence (Lee and Pero, 1981; Scarlato *et al.,* 1991), two of which were noted by Brennan and Geiduschek (1983), P_MIII-2 is located most appropriately to be the functional promoter. There is also a weak early promoter sequence (labeled PE? in Fig. 1) after T_L2 and before gene 52. It is conceivable that T_L2 stops all transcription from P_E10 and that this promoter defines a separate operon for gene 52.

Analysis of amino acid sequences

The amino acid sequences of the 24 gene products, as well as the complete nucleotide sequence, were used to probe the complete NCBI database and the database having the complete sequence of bacteriophage T4. Very few significant similarities were found, including none to T4 or to any other bacteriophage except the closely related *B. subtilis* phages SP82 and 2C. In view of the similarities in structure and life cycle between SPO1 and T4 (Stewart, 1993; Mathews, 1994), the extensive homologies between their respective hosts (Sonenshein *et al.,* 1993), their similarly large numbers of active early genes apparently involved in host shutoff (Kutter *et al.,* 1994), and the fact that several SPO1 early genes had effects on *E. coli* that were similar to their effects on uninfected *B. subtilis* (Curran and Stewart, 1985; Wei and Stewart, 1993; this laboratory, unpublished observations), the lack of similarity between any of the SPO1 and T4 genes

```
GENE 53,2C         1 ...RLVHYVCVPIISIHHAEDTINMTRKELSYLAQTIAKYIIADVEDTYLTFK 50
                        |||||,||||||||||||||.|||||. ,||,||| ,|| |,  || ||
GENE 53,SPO1      16 RFERLVHYICVPIISIHHAEDTISMTRKEVGHLAETIANHIILDINGTYRTFS 68

                  51 VDDIVHCSLENVIILDGAVTNEFKDRLQVFVNKEVQGEKSTQQS... 94
                     |.|||||||| || |,| ||||| ||||, ||||||| .|||||
                  69 VNDIVHCSLEKVITLEGDVTNEFIDRLQILVNKEVQGSQSTQQSLSS 115


GENE 56,SP82       1 MFKSTDRSVRQCI....... 13
                     ||| ||||||| |
GENE 56,SPO1       1 MFKYTDRSVRQYIERQQRSA 20


GENE 57,SP82       1 .......VVNYVWIAGTATTAFTYTAAIRHYCM* 27
                              .   ,  ||||||||||||||||||
GENE 57,SP01      81 DGFEGEIPDSSEDLRRTATTAFTYTAAIRHYCM* 114


GENE 58,SP82       1 MTLAGYRVDSCNGCGKAYLVGESHDRKKCAEC....
                     ||||||||||||||||||||||||||||||||
GENE 58,SPO1       1 MTLAGYRVDSCNGCGKAYLVGESHDRKKCAECASK* 36


GENE 59,SP82       1 MKKPLYKQQHYLRIIHHNIQVGNFSSPTNA*CTAMRNLTPGGIARVQHYN 50
                     |||       ,       , |||||||||| | ||||||| |||||||||
GENE 59,SPO1       1 MKKRYKVTALFEDGTSQCLVVGNFSSPTNAWCAAMRNLTPEGIARVQHYN 50

                  51 VEEISK* 57
                     |||||||
                  51 VEEISK* 57


GENE 60,SP82       1 M.NRVVGFHVECMLKVMSSNVETQPSNAPVIEVFTEDNLEEGIIPEYVTA 49
                     | |.|     | .   .       ||||||||||||||||||||||||||
GENE 60,SPO1       1 MLNQVEVLREEYVEGYVVQMWRRNPSNAPVIEVFTEDNLEEGIIPEYVTA 50

                  50 NDDTFDRIVYAVEFGYLEVLELV* 73
                     ||||||||| |||||||| |||||
                  51 NDDTFDRIVDAVEFGYLEELELV* 74
```

FIG. 4. Homologies to related phages. Alignments are shown between the products of SPO1 genes 53, 56, 57, 58, 59, and 60 and the homologous products of the related phages SP82 and 2C (Panganiban and Whiteley, 1983a; Daxhelet *et al.*, 1996). The published sequences included only fractions of the 2C and SP82 homologs of genes 53, 56, 57, and 58. These fractions are shown in their entirety in the figure, numbered beginning with the first AA in the known sequence. Some of the apparent divergence may be due to small discrepancies in the SP82 sequence. In gene 60, a single frameshift would generate perfect alignment among about half of the mismatched AAs, and a frameshift has been introduced arbitrarily into the early portion of SP82 gene 59 to produce the alignment shown.

suggests a remarkably independent evolution of mechanisms with similar functions.

Hints about possible protein function were obtained from the few similarities that *were* revealed. As discussed previously, GP44 (E3) acts on host RNA polymerase, presumably via an acidic/hydrophobic domain (Wei and Stewart, 1993, 1995). GP51 shows substantial similarity both to this domain of GP44 and to the carboxy-terminal region of the product of SPO1 gene 27, which apparently plays a role both in SPO1 DNA replication and in the expression of SPO1's late genes (Greene *et al.,* 1982; Heintz and Shub, 1982; Stewart, 1984). GP38 also includes several short sequences that are similar to the acidic/hydrophobic region of GP44.

GP58 consists of an almost perfect consensus sequence for a single Cys_2/Cys_2 zinc finger (Lewin, 1997), with a few additional amino acids on each end. Each of the Cys_2 segments is similar to the Cys_2 segments of several Cys_2/His_2 zinc finger proteins. Thus, it seems likely that GP58 could form a zinc finger, although the role of a single zinc finger in such a small protein is uncertain. The complete identity between its first 32 amino acids and those of the known segment of its SP82 homolog (Fig. 4) argues that it has a selectively significant function.

GP50 and GP56, respectively, include 18- and 21-amino-acid sequences that approximate known transmembrane domains. GP53 has several short regions of similarity to the *Clostridium perfringens* ϵ toxin (Shone and Hambleton, 1989; Hunter *et al.,* 1992) and to the *diphtheria* toxin (Greenfield *et al.,* 1983), which inhibits a specific step in eukaryotic protein synthesis (Collier, 1990).

Homology to related phages

Published sequences from the related phages SP82 (Panganiban and Whiteley, 1983a) and 2C (Daxhelet *et al.,* 1996) include the SP82 homologs of SPO1 genes 59 and 60, part of the SP82 homologs of genes 56, 57, and 58, and part of the 2C homolog of gene 53. Alignments of their amino acid sequences are shown in Fig. 4. The sequenced portion of SP82 includes its terminal *Hae*II fragment. Since this sequence is nearly identical to the right end of our sequence, base pair 11,500 can be identified as the right end of the SPO1 genome.

Tight packing of genetic information

As in the other known regions of the SPO1 genome (Costanzo *et al.,* 1983, 1984; Greene *et al.,* 1984; Goodrich-Blair *et al.,* 1990; Scarlato and Sayre, 1992; Scarlato and Gargano, 1992; Wilhelm and Ruger, 1992), these 24 genes are packed tightly together. In most cases, the last element of one operon overlaps at least the upstream activation sequence of the promoter for the next. Within each operon, there is little distance between the promoter and the first RBS and little or no distance between the last gene and the transcription–termination site. In the seven cases in which 2 genes are not separated by a regulatory element, they are close enough together for translational coupling (Vellanoweth, 1993; Draper, 1996), including two pairs (38/39 and 51/50) that overlap by 10 and 8 bp, respectively. In operons P_E1 and P_E10, there could be translational coupling of 4 or 3 consecutive genes. The only significant amounts of apparently unused sequence are about 230 bp between $T_L2.5$ and P_E10, which includes a homopolymeric region, and 160 bp between T_H3 and T_L1, the endpoints of converging rightward and leftward transcription.

Identity of opposite ends

Since both copies of the terminal redundancy were equally represented in most of the sequencing templates, and since consistent ambiguities in the sequencing data were not observed, we conclude that there has been no evolutionary divergence, and thus that there must be a mechanism of mismatch correction, between the two copies of the terminal redundancy. Such a mechanism is implied by the model for formation and resolution of concatemers proposed by Watson (1972) and supported for SPO1 by our earlier experiments (Glassberg *et al.,* 1977; Cregg and Stewart, 1978).

MATERIALS AND METHODS

SPO1 DNA preparation

Phage from 300 ml lysates was concentrated 100-fold by centrifugation (150 min at 13,200 *g*) and resuspension. They were centrifuged twice through CsCl step gradients (1.7 and 1.5 gm/ml; 2 h at 35,000 rpm). The phage band that formed at the interphase was dialyzed and treated with 0.8% SDS (2 min at 65°C) and 0.2 mg/ml proteinase K (30 min at 37°C), extracted gently with an equal volume of phenol (30 min at room temperature), precipitated with ethanol, and redissolved in TE. For use in automated sequencing, aliquots were diluted 1:20 with water, and the macromolecules were reconcentrated with a Microcon 100 Microconcentrator (Amicon).

Sequencing

*Sau*3A or *Alu*I digests of SPO1 *Eco*RI* fragments 10 and 15 were shotgun cloned into plasmids pPW110 or pEMBL19$^+$, as described (Wei and Stewart, 1993). Twenty-nine such cloned fragments were sequenced. Contigs were formed, using the Fragment Assembly programs of the Wisconsin Package Versions 8 and 9 of the Genetics Computer Group (GCG) (Madison, WI) and were connected by PCR, using SPO1 genomic DNA as template (Wei and Stewart, 1993). To extend the sequence beyond the outermost clones, to fill in gaps in the sequence, and to confirm the sequences determined from other templates, SPO1 genomic DNA also was used as a sequencing template. Most of the cloned fragments were sequenced manually, as described previously (Wei and Stewart, 1993). Most sequencing of PCR products and genomic DNA was done at the Molecular Genetics Core Facility of the University of Texas–Houston Medical School, using a PE/ABI Prism 377 DNA sequencer and an ABI Prism dye termination kit. Twenty-five to 27 base primers were needed to give the same quality sequences from SPO1 genomic DNA as those obtained with 20 to 22 base primers and thymine-containing templates, presumably because the presence of hmUra reduces the T_m by about 10°C (Marmur and Doty, 1962; Okubo *et al.,* 1964; Truffaut *et al.* 1970). Sequencing with PCR product templates always used the complete product of the PCR reaction, and the sequences determined showed no higher proportion of ambiguity than those from other templates.

With the two exceptions described below, every sequence was determined by at least one sequencing reaction on each strand. Approximately 92% of the nucleotides were determined from three or more independent sequences, and 77% from four or more. There were no disagreements between sequences determined from different types of templates. Because the initial sequence determined for bases 7116 to 7342 was identical to a published sequence (Brennan and Geiduschek, 1983), no further sequences were obtained for that region. The only uncertainty in the sequence is in and around a homopolymer that begins at nucleotide 9713. Every sequencing reaction that entered this region ended in a string of 24 to 28 Ts or As, depending on the direction. A primer consisting of 25 Ts followed by the 5 nucleotides to the right of the homopolymer (i.e., nucleotides 9732–9761) generated the correct sequence for bases 9762–10,413. A primer consisting of the 20 nucleotides to the left of the homopolymer followed by 22 Ts (i.e., nucleotides 9693–9734) generated a sequence of approximately 23 Ts, suggesting a minimum of 45 consecutive A:hmUra base pairs. Similar experiments on the opposite strand were unproductive, presumably because the weak bonding of A:hmUra pairs prevented primers with long strings of As from annealing effectively. For this reason, there are no leftward sequence data for the region just to the left of the homopolymer. Since this entire region is between a transcription terminator and the next promoter, and contains no significant ORFs, any sequence uncertainties are probably inconsequential.

Database analyses

Searches for sequence similarities used both the complete GenBank nucleotide and amino acid sequence database maintained by the National Center for Biotechnology Information (NCBI), accessible at http://www.ncbi.nlm.nih.gov, and the separate database for bacteriophage T4, accessible at http://www.evergreen.edu/user/T4/home.html. They used the BLAST sequence comparison program (Altschul *et al.,* 1990) of the GCG package, version 9.0, and the NCBI BLAST network service.

In the databases constructed for comparison with RBSs, each element consisted of the sequence from 40 bases upstream to 30 bases downstream of an initiation codon. Database A was made from *B. subtilis* genes rpmH, rpmA, rpsK, sspA, sspB, rpoA, rpoD, dnaH, infA, gyrB, rpmJ, rplQ, rpsM, gyrA, and dnaG, the first 7 of which are the strongly expressed genes whose RBSs were analyzed by Mountain (1989). Database B utilized 180 RBSs from the database of *B. subtilis* initiation sequences maintained by Dalphin *et al.* (1997), randomized by choosing the first 5 sequences on every third page that had at least a GAGG, GGAG, or GGGG at an appropriate distance upstream of the initiation codon. Database B′ consisted of 15 sequences chosen randomly from database B. Database C contained the other 22 SPO1 RBSs whose sequences are known, including 10 complete and 4 partial gene sequences (Costanzo *et al.,* 1983, 1984; Costanzo and Pero, 1983; Greene *et al.,* 1984; Scarlato and Sayre, 1992; Scarlato and Gargano, 1992; Wilhelm and Ruger, 1992; Goodrich-Blair *et al.,* 1990) and 10 RBSs whose sequences were published in connection with adjacent promoters (Costanzo *et al.*, 1984; Scarlato *et al.,* 1991). Database D was made from similar regions surrounding ATG codons not associated with ribosome-binding sites. For 18 of our 24 SPO1 genes, we selected the first in-frame ATG whose surrounding region did not overlap a ribosome-binding site. For 2 of the 6 that had no such ATG, we used the region surrounding a TTG or a GTG.

Calculation of ΔG values

The free energy of binding between an RBS and the 3′ end of *B. subtilis* 16S RNA (3′-UCUUUCCUCCACUAG-5′) (Green *et al.,* 1985; Shine and Dalgarno, 1975) was calculated according to Tinoco *et al.* (1973). For hairpins, the lowest energy structure predicted for each sequence was determined with the help of the FoldRNA program (Zuker, 1989), which is part of the GCG Version 9 package, and the ΔG values were calculated using the parameters of Tinoco *et al.* (1973). For comparison purposes, the same parameters were used to calculate the ΔG values given for structures published elsewhere, if those values had not been calculated or had been calculated with different parameters.

Terminology

(a) For convenience of exposition, operon means the gene or genes that are in position to be transcribed from 1 of the 12 major promoters and that are located between that promoter and the next major promoter downstream. Since some operons lack transcription–termination sites, the term is not synonymous with transcription unit. (b) A consensus sequence enclosed in brackets, such as [GGAGG], refers to the sequence that occupies that *position* in each of the RBSs, although individual sequences may deviate from the consensus. (c) For convenience of presentation, hmUra residues will be represented by Ts. (d) GPX is the protein specified by gene X.

ACKNOWLEDGMENTS

We are grateful to Paul Sharp for calculating the CAI values and to Mark Dalphin for personal communication of his database of *B. subtilis* initiation sequences. The very early portion of this work was supported by NSF Grant DMB-8801703.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Brennan, S. M., Chelm, B. K, Romeo, J. M., and Geiduschek, E. P. (1981). A transcriptional map of the bacteriophage SPO1 genome. II. The major early transcription units. *Virology* 111, 604–628.

Brennan, S. M., and Geiduschek, E. P. (1983). Regions specifying transcriptional termination and pausing in the bacteriophage SPO1 terminal repeat. *Nucleic Acids Res.* 11, 4157–4175.

Buc, H. (1986). Mechanism of activation of transcription by the complex formed between cyclic AMP and its receptor in *Escherichia coli. Biochem. Soc. Trans.* 14, 196–199.

Collier, R. J. (1990). Diphtheria toxin, structure and function of a cytocidal protein. *In* “ADP-Ribosylating Toxins and G. Proteins” (J. Moss, and M. Vaughan, Eds.), pp. 3–19. Am. Soc. for Microbiol., Washington, DC.

Costanzo, M., and Pero, J. (1983). Structure of a *Bacillus subtilis* bacteriophage SPO1 gene encoding a RNA polymerase σ factor. *Proc. Natl. Acad. Sci. USA* 80, 1236–1240.

Costanzo, M., Hannett, N., Brzustowicz, L., and Pero, J. (1983). Bacteriophage SPO1 gene 27, location and nucleotide sequence. *J. Virol.* 48, 555–560.

Costanzo, M., Brzustowicz, L., Hannet, N., and Pero, J. (1984). Bacteriophage SPO1 genes 33 and 34. Location and primary structure of genes encoding regulatory subunits of *Bacillus subtilis* RNA polymerase. *J. Mol. Biol.* 180, 533–547.

Cowe, E., and Sharp, P. M. (1991). Molecular evolution of bacteriophages, discrete patterns of codon usage in T4 genes are related to the time of gene expression. *J. Mol. Evol.* 33, 13–22.

Cregg, J. M., and Stewart, C. R. (1978). Terminal redundancy of ‘high frequency of recombination‘ markers of *Bacillus subtilis* phage SPO1. *Virology* 86, 530–541.

Curran, J. F., and Stewart, C. R. (1985). Cloning and mapping of the SPO1 genome. *Virology* 142, 78–97.

Dalphin, M. E., Brown, C. M., Stockwell, P. A., and Tate, W. P. (1997). The translational signal database, TransTerm, more organisms, complete genomes. *Nucleic Acids Res.* 25, 246–247.

Daxhelet, G., Gilot, P., and Hoet, P. (1996). Cloning and characterization of transcriptional promoters from *Bacillus subtilis* phage 2C. *Can. J. Microbiol.* 42, 919–926.

Downard, J. S., and Whiteley, H. R. (1981). Early RNAs in SP82- and SPO1-infected *B. subtilis* may be processed. *J. Virol.* 37, 1075–1078.

Draper, D. E. (1996). Translational initiation. *In* "*Escherichia coli* and *Salmonella*" (F. C. Neidhardt *et al.,* Eds.), 2nd ed., Vol. 1, pp. 902–908. ASM Press, Washington, DC.

Fujita, D. J., Ohlsson-Wilhelm, B. M., and Geiduschek, E. P. (1971). Transcription during bacteriophage SPO1 development: Mutations affecting the program of viral transcription. *J. Mol. Biol.* 57, 301–317.

Gage, L. P., and Geiduschek, E. P. (1971). RNA synthesis during bacteriophage SPO1 development, six classes of SPO1 DNA. *J. Mol. Biol.* 57, 279–300.

Glassberg, J., Franck, M., and Stewart, C. R. (1977). Initiation and termination mutants of *Bacillus subtilis* bacteriophage SPO1. *J. Virol.* 21, 147–152.

Goodrich-Blair, H., Scarlato, V., Gott, J. M., Xu, M.-Q., and Shub, D. A. (1990). A self-splicing group I intron in the DNA polymerase gene of *Bacillus subtilis* bacteriophage SPO1. *Cell* 63, 417–424.

Green, C. J., Stewart, G. C., Hollis, M. A., Vold, B. S., and Bott, K. F. (1985). Nucleotide sequence of the *Bacillus subtilis* ribosomal RNA operon, *rrnB. Gene* 37, 261–266.

Greene, J. R., Chelm, B. K., and Geiduschek, E. P. (1982). SPO1 gene 27 is required for viral late transcription. *J. Virol.* 41, 715–720.

Greene, J. R., Brennan, S. M., Andrew, D. J., Thompson, C. C., Richards, S. H., Heinrikson, R. L., and Geiduschek, E. P. (1984). Sequence of bacteriophage SPO1 gene coding for transcription factor 1, a viral homologue of the bacterial type II DNA-binding proteins. *Proc. Natl. Acad. Sci. USA* 81, 7031–7035.

Greene, J. R., and Geiduschek, E. P. (1985). Site-specific DNA binding by the bacteriophage SPO1-encoded type II DNA binding protein. *EMBO J.* 4, 1345–1349.

Greene, J. R., Morrissey, L. M., Foster, L. M., and Geiduschek, E. P. (1986). DNA binding by the bacteriophage SPO1-encoded type II DNA-binding protein, transcription factor 1. *J. Biol. Chem.* 261, 12820–12827.

Greenfield, L., Bjorn, M. J., Horn, G., Fong, D., Buck, G. A., Collier, R. J., and Kaplan, D. A. (1983). Nucleotide sequence of the structural gene for diphtheria toxin carried by corynebacteriophage beta. *Proc. Natl. Acad. Sci. USA* 80, 6853–6857.

Grove, A., Figueiredo, M. L., Galeone, A., Mayol, L., and Geiduschek, E. P. (1997). Twin hydroxymethyluracil-A base pair steps define the binding site for the DNA-bending protein TF1. *J. Biol. Chem.* 272, 13084–13087.

Hager, P. W., and Rabinowitz, J. C. (1985). Translational specificity in *Bacillus subtilis. In* "The Molecular Biology of the Bacilli" (D. A. Dubnau, Ed.), pp. 1–32. Academic Press, New York.

Heintz, N., and Shub, D. A. (1982). Transcriptional regulation of bacteriophage SPO1 protein synthesis *in vivo* and *in vitro. J. Virol.* 42, 951–962.

Helmann, J. D. (1995). Compilation and analysis of *Bacillus subtilis* σ^A-dependent promoter sequences, evidence for extended contact between RNA polymerase and upstream promoter DNA. *Nucleic Acids Res.* 23, 2351–2360.

Hue, K. K., Cohen, S. D., and Bechhofer, D. H. (1995). A polypurine sequence that acts as a 5′ mRNA stabilizer in *Bacillus subtilis. J. Bacteriol.* 177, 3465–3471.

Hunter, S. E. C., Clarke, I. N., Kelley, D. C., and Titball, R. W. (1992). Cloning and sequencing of the *Clostridium perfringens* epsilon-toxin gene and its expression in *E. coli. Infect. Immun.* 60, 102–112.

Keilty, S., and Rosenberg, M. (1987). Constitutive function of a positively regulated promoter reveals new sequences essential for activity. *J. Biol. Chem.* 262, 6389–6395.

Kutter, E., White, T., Kashlev, M., Uzan, M., McKinney, J., and Guttman, B. (1994). Effects on host genome structure and expression. *In* "Molecular Biology of Bacteriophage T4" (J. D. Karam, Ed.), pp. 357–368. Am. Soc. for Microbiol., Washington, DC.

Lee, G., Talkington, C., and Pero, J. (1980). Nucleotide sequence of a promoter recognized by *Bacillus subtilis* RNA polymerase. *Mol. Gen. Genet.* 180, 57–65.

Lee, G., and Pero, J. (1981). Conserved nucleotide sequences in temporally controlled bacteriophage promoters. *J. Mol. Biol.* 152, 247–265.

Lewin, B. (1997). "Genes VI." Oxford Univ. Press.

Losick, R., and Pero, J. (1981). Cascades of sigma factor. *Cell* 25, 582–584.

Marmur, J., and Doty, P. (1962). Determination of the base composition of deoxyribonucleic acid from its thermal denaturation temperature. *J. Mol. Biol.* 5, 109–118.

Mathews, C. K. (1994). An overview of the T4 developmental program. *In* "Molecular Biology of Bacteriophage T4" (J. D. Karam, Ed.), pp. 1–8. Am. Soc. for Microbiol., Washington, DC.

Mitra, S., and Bechhofer, D. H. (1994). Substrate specificity of an RNase III-like activity from *Bacillus subtilis. J. Biol. Chem.* 269, 31450–31456.

Mountain, A. (1989). Gene expression systems for *Bacillus subtilis. In* "Biotechnology Handbook. 2. Bacillus" (C. R. Harwood, Ed.), pp. 73–114. Plenum, New York.

Okubo, S., Strauss, B., and Stodolsky, M. (1964). The possible role of recombination in the infection of competent *Bacillus subtilis* by bacteriophage deoxyribonucleic acid. *Virology* 24, 552–562.

Okubo, S., Yanagida, T., Fujita, D. J., and Ohlsson-Wilhem, B. M. (1972). The genetics of bacteriophage SPO1. *Biken. J.* 15, 81–97.

Panganiban, A. T., and Whiteley, H. R. (1983a). *B. subtilis* RNase III cleavage sites in phage SP82 early mRNA. *Cell* 33, 907–913.

Panganiban, A. T., and Whiteley, H. R. (1983b). Purification and properties of a new *Bacillus subtilis* RNA processing enzyme. *J. Biol. Chem.* 258, 12487–12493.

Perkus, M. E., and Shub, D. A. (1985). Mapping the genes in the terminal redundancy of bacteriophage SPO1 with restriction endonucleases. *J. Virol.* 56, 40–48.

Pero, J., Hannett, N. M., and Talkington, C. (1979). Restriction cleavage map of SPO1 DNA, General location of early, middle and late genes. *J. Virol.* 31, 156–171.

Reeve, J. N., Mertens, G., and Amann, E. (1978). Early development of bacteriophages SPO1 and SP82G in minicells of *Bacillus subtilis. J. Mol. Biol.* 120, 183–207.

Ringquist, S., Shinedling, S., Barrick, D., Green, L., Binkley, J., Stormo, G. D., and Gold, L. (1992). Translation initiation in *Escherichia coli,* sequences within the ribosome-binding site. *Mol. Microbiol.* 6, 1219–1229.

Roberts, J. W. (1996). Transcription termination and its control. *In* "Regulation of Gene Expression in *Excherichia coli*" (E. C. C. Lin and A. S. Lynch, Eds.), pp. 27–45. Landes, Austin.

Romeo, J. M., Brennan, S. M., Chelm, B. K., and Geiduschek, E. P. (1981). A transcriptional map of the bacteriophage SPO1 genome. I. The major early promoters. *Virology* 111, 588–603.

Scarlato, V., Greene, J. R., and Geiduschek, E. P. (1991). Bacteriophage SPO1 middle transcripts. *Virology* 180, 716–728.

Scarlato, V., and Sayre, M. H. (1992). The structure of the bacteriophage SPO1 gene 30. *Gene* 114, 115–119.

Scarlato, V., and Gargano, S. (1992). The DNA polymerase-encoding gene of *Bacillus subtilis* bacteriophage SPO1. *Gene* 118, 109–113.

Sharp, P. M. and Li, W-H. (1987). The codon adaptation index—A measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15, 1281–1295.

Shields, D. C., and Sharp, P. M. (1987). Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. *Nucleic Acids Res.* 15, 8023–8040.

Shine, J., and Dalgarno, L. (1975). Determinant of cistron specificity in bacterial ribosomes. *Nature* 254, 34–38.

Shone, C. C., and Hambleton, P. (1989). Toxigenic clostridia. *In* "Biotechnology Handbooks, Vol. 3, *Clostridium*" (N. P. Minton and O. J. Clarke, Eds.), pp. 265–292. Plenum, New York.

Shub, D. A. (1966). "Functional Stability of Messenger RNA during

Bacteriophage Development." Ph.D. thesis, Massachusetts Institute of Technology, Cambridge.

Sonenshein, A. L., Hoch, J. A., and Losick, R. (Eds.). (1993). "*Bacillus subtilis* and Other Gram-Positive Bacteria, Biochemistry, Physiology, and Molecular Genetics." Am. Soc. for Microbiol., Washington, DC.

Stewart, C. R. (1984). Dissection of HA20, a double mutant of bacteriophage SPO1. *J. Virol.* 49, 300–301.

Stewart, C. R. (1993). SPO1 and related bacteriophages. *In* "*Bacillus subtilis* and Other Gram-Positive Bacteria, Biochemistry, Physiology, and Molecular Genetics" (A. L. Sonenshein *et al.,* Eds.), pp. 813–829. Am. Soc. for Microbiol., Washington, DC.

Talkington, C., and Pero, J. (1977). Restriction fragment analysis of temporal program of bacteriophage SPO1 transcription and its control by phage-modified RNA polymerases. *Virology* 83, 365–379.

Tinoco, I., Borer, P. N., Dengler, B., Levine, M. D., Uhlenbeck, O. C., Crothers, D. M., and Gralla, J. (1973). Improved estimation of secondary structure in RNA. *Nature New Biol.* 246, 40–41.

Travers, A. (1993). "DNA-Protein Interactions." Chapman & Hall, London.

Truffaut, N., Revet, B., and Soulie, M. O. (1970). Etude comparative des DNA de phages 2C, SP8*, SP82, ϕe, SPO1 et SP50. *J. Biochem.* 15, 391–400.

Vellanoweth, R. L. (1993). Translation and its regulation. *In* "*Bacillus subtilis* and Other Gram-Positive Bacteria, Biochemistry, Physiology, and Molecular Genetics" (A. L. Sonenshein *et al.,* Eds.), pp. 99–711. Am. Soc. for Microbiol., Washington, DC.

Voskuil, M. I., Voepel, K., and Chambliss, G. H. (1995). The −16 region, a vital sequence for the utilization of a promoter in *Bacillus subtilis* and *Escherichia coli. Mol. Microbiol.* 17, 271–279.

Watson, J. D. (1972). Origin of concatemeric T7 DNA. *Nature New Biol.* 239, 197–201.

Wei, P., and Stewart, C. R. (1993). A cytotoxic early gene of *Bacillus subtilis* bacteriophage SPO1. *J. Bacteriol.* 175, 7887–7900.

Wei, P., and Stewart, C. R. (1995). Genes that protect against the host-killing activity of the E3 protein of *B. subtilis* bacteriophage SPO1. *J. Bacteriol.* 177, 2933–2937.

Wilhelm, K., and Rüger, W. (1992). Deoxyuridylate-hydroxymethylae of bacteriophage SPO1. *Virology* 189, 640–646.

Wilkens, K., and Rüger, W. (1996). Characterization of bacteriophage T4 early promoters *in vivo* with a new promoter probe vector. *Plasmid* 35, 108–120.

Yansura, D., and Henner, D. J. (1984). Use of the *E. coli* lac repressor and operator to control gene expression in *Bacillus subtilis. Proc. Natl. Acad. Sci. USA* 81, 439–443.

Zuker, M. (1989). Computer prediction of RNA structure. *Methods Enzymol.* 180, 262–288.