Volume 14 Number 10 1986 Nucleic Acids Research

Nucleotide sequence and organization of *Bacillus subtilis* RNA polymerase major sigma (σ^{43}) operon

Lin-Fa Wang and Roy H.Doi*

Department of Biochemistry and Biophysics, University of California, Davis, CA 95616, USA

Received 6 March 1986; Revised and Accepted 14 April 1986

ABSTRACT

The gene coding for Bacillus subtilis RNA polymerase major σ^{43}, rpoD, was cloned together with its neighboring genes in a 7 kb EcoRI fragment. The complete nucleotide sequence of a 5 kb fragment including the entire rpoD gene revealed the presence of two other genes preceding rpoD in the order P23-dnaE-rpoD. The dnaE codes for DNA primase while the function of P23 remains unknown. The three genes reside in an operon that is similar in organization to the E.coli RNA polymerase major σ^{70} operon, which is composed of genes encoding small ribosome protein S21 (rpsU), DNA primase (dnaG), and RNA polymerase σ^{70} (rpoD). There is a relatively high degree of base and amino acid homology between the DNA primase and σ genes. The most significant differences between the two operons are observed in the molecular size of the first genes (P23 and rpsU), the complete lack of amino acid homology between P23 and S21, the molecular weights of the two rpoD genes, the size of the intercistronic region between the first two genes, and the regulatory elements of the operon.

INTRODUCTION

The existence of multiple RNA polymerase σ factors in B. subtilis has been well documented (1,2), but little is known about their genetic properties, the regulation of their synthesis, and the factors that govern their interactions with the RNA polymerase core. An analysis of their molecular organization and the parameters which regulate their genetic expression should provide a initial basis for determining their roles in the physiology of this Gram positive sporulating bacterium.

Our laboratory has been particularly interested in the study of the σ^{43} gene (rpoD), whose product is known to play a major role during vegetative growth, and the early stationary and sporulation phases (3). We have been able to clone (4), genetically map (5), and sequence (6) the σ^{43} gene (rpoD), and show that its derived amino acid sequence had a very high degree of homology with that of the E.coli major σ^{70} (7). By genetic mapping (5) and DNA sequencing (8), we also showed that immediately upstream of the rpoD gene was located the dnaE gene, which encodes the B. subtilis DNA primase, whose product is very homologous to the E.coli dnaG DNA primase (9,10). No promoter region was observed in the intercistronic region between rpoD and dnaE, nor in the region immediately upstream of dnaE (6,8).

Recently, we have determined the nucleotide sequence of the region upstream of dnaE including the operon regulatory region, which provided support for our previous suggestion (6,8) that dnaE and rpoD were coordinately regulated with one or more unknown genes in an operon. The DNA sequence analysis of the region upstream of dnaE revealed an open reading frame capable of coding for a protein of molecular weight 22,540. The function of this protein is unknown, and hence the designation P23 is being used for this gene until a physiological role can be assigned to it.

In this paper we will discuss the similarities and differences of the structure and organization of the major sigma operons of B. subtilis and E. coli, the transcriptional and translational regulatory features of the operon, and the codon usage frequency encountered in the operon.

MATERIALS AND METHODS

Strains, Phages and Plasmids

E. coli JM101 was used as host for the sequencing phage vectors M13mp8, M13mp9, M13mp10, and M13mp11 (11,12), and the plasmid pCPS1 (5). E. coli BNN45 (13) was used to prepare the phage lysate of λgtWES-σ82 (4). Plasmid pSB was provided by Sui-Lam Wong (unpublished data).

DNA Manipulations

Standard procedures of Maniatis et al. (14) were followed exactly as described.

DNA Sequencing

DNA sequencing was conducted by the dideoxy chain termination method of Sanger *et al*. (15) using the sequencing kit purchased from Amersham Corporation.

Computer Analysis

Routine analysis of DNA or protein sequences were carried out using either the Delaney (16) or the Pustell (17) program, while the homology search against the NBRF Data Bank was made using the Microgenie Sequence Analysis Program developed by Queen and Korn (18).

RESULTS

Nucleotide sequence of the Entire Operon

The nucleotide sequences and the sequencing strategies of *dnaE* and *rpoD* genes have been reported previously (6,8). The sequencing strategy for the upstream 1.5 kb fragment is shown in Fig. 1 (bottom) along with the physical map of the σ^{43} operon (upper). As indicated, the nucleotide sequence has been determined for both strands of virtually the entire region except for the 100 bp at the extreme 5' end. The sequence was determined across the junctions of all the restriction sites used for subcloning during sequencing, as well as for the *Eco*RI site between the *dnaE* and *rpoD* genes (not shown here). In our previous reports, the sequences for these two genes were determined separately (6,8). Although unlikely, the possibility existed that a small *Eco*RI fragment may have been left out during the subcloning of the *Eco*RI fragments into plasmids from the original phage λgtWES-σ82 (4,5). Therefore we sequenced the 0.9 kb HindIII fragment containing the EcoRI junction region, which was subcloned into M13mp10 directly from λgtWES-σ82, and the possibility mentioned above has been experimentally excluded. Now, the entire *Eco*RI-*Sph*I fragment has been sequenced, including all the junctions of restriction sites used for sequencing. The nucleotide sequence of the entire operon and its flanking regions, and the deduced amino acid sequence of each gene are given in Fig. 2 with the first base of the 5' end *Eco*RI site labeled as number 1.

Features of the First Gene of the Operon

When the sequence of the region upstream of *dnaE* was analyzed by computer, only one large open reading frame was

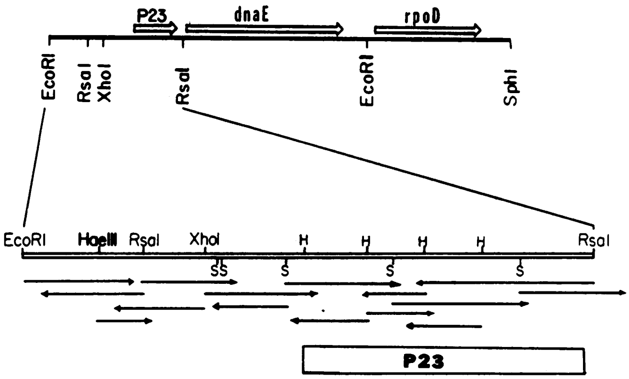

Figure 1. Sequencing strategy for the first gene and regulatory region of the σ^{43} operon.
The upper part represents the physical location of the genes in the operon. The lower part indicates the restriction sites used for subcloning and sequencing. The bar indicates the location of the cryptic P23 protein. Abbreviations: H, HpaII; S, Sau3A.

discovered. But unlike the case for the other two genes in the operon, we could not identify any strong ribosomal binding site by sequence analysis within the open reading frame. We found instead several weak ones preceeding the potential initiation codons ATG (855), TTG (930), ATG (951), and GTG (1,200), which were able to code for proteins of molecular weights 22,540, 19,734, 18,934 and 9,312, respectively. However, our previous maxicell data showed that a protein of molecular weight around 23,000 was encoded within the upstream region of dnaE (4,5), which corresponded very well with the largest open reading frame identified here by sequencing, i.e., the open reading frame starting from the initiation codon ATG at nt 855.

Although the ribosomal binding site was very weak as predicted from its calculated free energy of binding ($\Delta G = -9.2$ kcal/mol), our assignment was further strengthened by our recent protein fusion studies. When the N-terminal two thirds of the P23 was fused to the *E. coli* β-galactosidase in frame, a functional hybrid protein was expressed in both *E. coli* and *B. subtilis* with the expected size as determined by Western blot analysis using anti-β-galactosidase antibody (data not shown). Functional P23-β-galactosidase (P23-gal) fusion protein was expressed even when only the first 8 amino acid residues of P23

```
        10         20         30         40         50         60
GAATTCTTAT GTTGAAGATA TAGGCACTAT CAATGAAGTG ATTTCACTTG CAAAGGCAGA

        70         80         90        100        110        120
CGGCGGCATT ATCTGTTTTA CACTCGTGGT GCCGGAAATC AGAGAATATT TGATAGCCGA

       130        140        150        160        170        180
AGCGGAAAAA GCAAATGTTT TATATTATGA TATTATCGGC CCGTTGATTG ATAAAATGGA

       190        200        210        220        230        240
AACAGCCTAC GGTTTAACAG CGAAATACGA ACCGGGGCGG GTGCGCCAGC TTGATGAAGA

       250        260        270        280        290        300
TTATTTCAAA AAAGTGGAGG CCATCGAGTT TGCAGTTAAA TACGATGATG GACGTGATCC

       310        320        330        340        350        360
AAGAGGGATT TTAAAAGCTG ATATCGTTTT GATCGGCGTG TCAAGAACGT CTAAAACACC

       370        380        390        400        410        420
GCTGTCTCAA TATCTCGCAC ACAAACGCCT GAAGGTTGCC AATGTTCCGA TTGTACCGGA

       430        440        450        460        470        480
GGTTGATCCG CCGGAAGAAC TCTTTAACGT TGATCCGAAA AAATGCATCG GTTTAAAGAT

       490        500        510        520        530        540
TAGCCCTGAT AAACTGAATC ATATCAGAAA AGAACGTTTA AAATCACTCG GGCTTAATGA

       550        560        570        580        590        600
TAAAGCGATT TATGCAAATA TCAACAGAAT CAAAGAGGAA CTCGAGTATT TCGAAAAGAT

       610        620        630        640        650        660
TGTGGATCGG ATCGGCTGCC AGGTTGTTGA TGTTTCAAAT AAAGCGGTTG AGGAAACAGC

       670        680        690        700        710        720
AAATATTATC CATCATCTCA AAACAAAAAA CATATAACTC AGGACGCTCT ATCCTGGGTT

       730        740        750        760        770        780
TTTGGCTGTG CCAAAAGGGA ATAATGAAAA ACAATAGCAT CTTTGTGAAG TTTGTATTAT
                    P1                    P2
       790        800        810        820        830        840
AATAAAAAAT TGTGATAAAA TGATTAATTT TAGGTTTAAG GATCGTGTGA TACGAATAAA

       850
CTATTATGGG TAAG
```

P23

```
                                         884                                     914
ATG TCA AGA ATT TCT CCC GGA AAT TTT TCG ACA AAT TCA TAT ACA TCC ACA ATA ATA AAG
Met Ser Arg Ile Ser Pro Gly Asn Phe Ser Thr Asn Ser Tyr Thr Ser Thr Ile Ile Lys

                                         944                                     974
GAT GTG CGA TTT TGC TTG CTT TTA TGC AGG AGT TTA ATG GAG GGA TGG AGA ATT ACT CTT
Asp Val Arg Phe Cys Leu Leu Leu Cys Arg Ser Leu Met Glu Gly Trp Arg Ile Thr Leu
                                                      P3
                                        1004                                    1034
CTT AAT GAA CAA GAA AAG ACG ATT TTT GTC GAT GCT GAT GCT TGT CCG GTA AAA GAT GAA
Leu Asn Glu Gln Glu Lys Thr Ile Phe Val Asp Ala Asp Ala Cys Pro Val Lys Asp Glu

                                        1064                                    1094
ATT TTA CAA ACA GCA TCC GAG TAT GAA GTT CAA GTT CTT TTT GTC GCT TCA TTT GAA CAT
Ile Leu Gln Thr Ala Ser Glu Tyr Glu Val Gln Val Leu Phe Val Ala Ser Phe Glu His

                                        1124                                    1154
TAT CAG CTT TCC AGA AGC AAT GAA GAA AAA TGG AAG TAT GTT GAT CCT CAT AAA GAA GCT
Tyr Gln Leu Ser Arg Ser Asn Glu Glu Lys Trp Lys Tyr Val Asp Pro His Lys Glu Ala
```

```
                                        1184                                    1214
GCT GAT TTA TAT ATC GCA AAT CAC GTG AAA CCG GGA GAT ATT GTT GTG ACG CAG GAC ATC
Ala Asp Leu Tyr Ile Ala Asn His Val Lys Pro Gly Asp Ile Val Val Thr Gln Asp Ile

                                        1244                                    1274
GGA TTA GCA TCT CTG CTG TTG AAC AGA AAT GTC TCT GTT ATG TCG GAA AGA GGT CGT CTT
Gly Leu Ala Ser Leu Leu Leu Asn Arg Asn Val Ser Val Met Ser Glu Arg Gly Arg Leu

                                        1304                                    1334
TAC AAG GAA GAC ACG ATT GAT TTT GCC CTA GAG GGC CGT CAT TTT TCC GGC AAA CAA AGA
Tyr Lys Glu Asp Thr Ile Asp Phe Ala Leu Glu Gly Arg His Phe Ser Gly Lys Gln Arg

                                        1364                                    1394
AGA AAA GGC GTA TAT GCC AAA GGG CCT AAA AAA TTG AAT AAA GAA GAT CGA GAA CGA TTT
Arg Lys Gly Val Tyr Ala Lys Gly Pro Lys Lys Leu Asn Lys Glu Asp Arg Glu Arg Phe

                                        1424                                    1456
ATT ACA CTG CTG CAA AAA ATC CTG TCG AAC GAT GAA GGG ATT TTG CAC TAA AGCATCGAATA
Ile Thr Leu Leu Gln Lys Ile Leu Ser Asn Asp Glu Gly Ile Leu His End

ATGTACGACG GAGTGTTATA AG
```

dnaE

```
                                        1508                                    1538
ATG GGA AAT CGG ATA CCA GAT GAA ATT GTG GAT CAG GTG CAA AAG TCG GCA GAT ATC GTT
Met Gly Asn Arg Ile Pro Asp Glu Ile Val Asp Gln Val Gln Lys Ser Ala Asp Ile Val

                                        1568                                    1598
GAA GTC ATA GGT GAT TAT GTT CAA TTA AAG AAG CAA GGC CGA AAC TAC TTT GGA CTC TGT
Glu Val Ile Gly Asp Tyr Val Gln Leu Lys Lys Gln Gly Arg Asn Tyr Phe Gly Leu Cys

                                        1628                                    1658
CCT TTT CAT GGA GAA AGC ACA CCT TCG TTT TCC GTA TCG CCC GAC AAA CAG ATT TTT CAT
Pro Phe His Gly Glu Ser Thr Pro Ser Phe Ser Val Ser Pro Asp Lys Gln Ile Phe His

                                        1688                                    1718
TGC TTT GGC TGC GGA GCG GGC GGC AAT GTT TTC TCT TTT TTA AGG CAG ATG GAA GGC TAT
Cys Phe Gly Cys Gly Ala Gly Gly Asn Val Phe Ser Phe Leu Arg Gln Met Glu Gly Tyr

                                        1748                                    1778
TCT TTT GCC GAG TCG GTT TCT CAC CTT GCT GAC AAA TAC CAA ATT GAT TTT CCA GAT GAT
Ser Phe Ala Glu Ser Val Ser His Leu Ala Asp Lys Tyr Gln Ile Asp Phe Pro Asp Asp

                                        1808                                    1838
ATA ACA GTC CAT TCC GGA GCC CGG CCA GAG TCT TCT GGA GAA CAA AAA ATG GCT GAG GCA
Ile Thr Val His Ser Gly Ala Arg Pro Glu Ser Ser Gly Glu Gln Lys Met Ala Glu Ala

                                        1868                                    1898
CAT GAG CTC CTG AAG AAA TTT TAC CAT CAT TTG TTA ATA AAT ACA AAA GAA GGT CAA GAG
His Glu Leu Leu Lys Lys Phe Tyr His His Leu Leu Ile Asn Thr Lys Glu Gly Gln Glu

                                        1928                                    1958
GCA CTG GAT TAT CTG CTT TCT AGG GGC TTT ACG AAA GAG CTG ATT AAT GAA TTT CAG ATT
Ala Leu Asp Tyr Leu Leu Ser Arg Gly Phe Thr Lys Glu Leu Ile Asn Glu Phe Gln Ile

                                        1988                                    2018
GGC TAT GCT CTT GAT TCT TGG GAC TTT ATC ACG AAA TTC CTT GTA AAG AGG GGA TTT AGT
Gly Tyr Ala Leu Asp Ser Trp Asp Phe Ile Thr Lys Phe Leu Val Lys.Arg Gly Phe Ser

                                        2048                                    2078
GAG GCG CAA ATG GAA AAA GCG GGT CTC CTG ATC AGA CGC GAA GAC GGA AGC GGA TAT TTC
Glu Ala Gln Met Glu Lys Ala Gly Leu Leu Ile Arg Arg Glu Asp Gly Ser Gly Tyr Phe
```

```
                                   2108                                    2138
GAC CGC TTC AGA AAC CGT GTC ATG TTT CCG ATC CAT GAT CAT CAC GGG GCT GTT GTT GCT
Asp Arg Phe Arg Asn Arg Val Met Phe Pro Ile His Asp His His Gly Ala Val Val Ala

                                   2168                                    2198
TTC TCA GGC AGG GCT CTT GGC AGC CAG CAG CCT AAG TAT ATG AAC AGT CCT GAA ACC CCG
Phe Ser Gly Arg Ala Leu Gly Ser Gln Gln Pro Lys Tyr Met Asn Ser Pro Glu Thr Pro

                                   2228                                    2258
CTC TTT CAT AAA AGC AAA CTG CTT TAC AAT TTT TAT AAG GCC CGC CTT CAT ATC AGA AAG
Leu Phe His Lys Ser Lys Leu Leu Tyr Asn Phe Tyr Lys Ala Arg Leu His Ile Arg Lys

                                   2288                                    2318
CAG GAA AGA GCA GTC TTA TTT GAA GGG TTT GCT GAT GTC TAT ACG GCC GTA AGC TCG GAT
Gln Glu Arg Ala Val Leu Phe Glu Gly Phe Ala Asp Val Tyr Thr Ala Val Ser Ser Asp

                                   2348                                    2378
GTA AAG GAA AGC ATA GCC ACG ATG GGA ACG TCT CTT ACA GAT GAT CAT GTC AAG ATC CTG
Val Lys Glu Ser Ile Ala Thr Met Gly Thr Ser Leu Thr Asp Asp His Val Lys Ile Leu

                                   2408                                    2438
AGA AGA AAC GTC GAA GAA ATC ATT CTT TGC TAT GAC TCT GAT AAA GCC GGT TAT GAA GCC
Arg Arg Asn Val Glu Glu Ile Ile Leu Cys Tyr Asp Ser Asp Lys Ala Gly Tyr Glu Ala

                                   2468                                    2498
ACC TTA AAA GCT TCG GAG CTT CTG CAA AAA AAA GGC TGC AAA GTC AGA GTT GCA ATG ATT
Thr Leu Lys Ala Ser Glu Leu Leu Gln Lys Lys Gly Cys Lys Val Arg Val Ala Met Ile

                                   2528                                    2558
CCT GAC GGA TTG GAC CCT GAT GAT TAC ATC AAA AAA TTC GGC GGG GAA AAA TTT AAA AAC
Pro Asp Gly Leu Asp Pro Asp Asp Tyr Ile Lys Lys Phe Gly Gly Glu Lys Phe Lys Asn

                                   2588                                    2618
GAC ATT ATT GAC GCA AGT GTC ACC GTA ATG GCG TTC AAA ATG CAA TAT TTC CGA AAA GGA
Asp Ile Ile Asp Ala Ser Val Thr Val Met Ala Phe Lys Met Gln Tyr Phe Arg Lys Gly

                                   2648                                    2678
AAG AAC CTG TCC GAT GAA GGC GAC CGC CTA GCT TAC ATT AAA GAC GTA CTG AAA GAA ATC
Lys Asn Leu Ser Asp Glu Gly Asp Arg Leu Ala Tyr Ile Lys Asp Val Leu Lys Glu Ile

                                   2708                                    2738
AGC ACG CTT TCA GGG TCT CTA GAG CAG GAA GTC TAT GTA AAG CAG CTT GCT TCA GAG TTT
Ser Thr Leu Ser Gly Ser Leu Glu Gln Glu Val Tyr Val Lys Gln Leu Ala Ser Glu Phe

                                   2768                                    2798
TCG CTT TCA CAG GAG TCT TTA ACT GAG CAG CTG TCT GTT TTC AGC AAG CAA AAC AAA CCT
Ser Leu Ser Gln Glu Ser Leu Thr Glu Gln Leu Ser Val Phe Ser Lys Gln Asn Lys Pro

                                   2828                                    2858
GCT GAC AAT AGC GGT GAA ACT AAA ACG CGG CGA GCG CAT CTG ACG ACA AAA GCA AGG CAA
Ala Asp Asn Ser Gly Glu Thr Lys Thr Arg Arg Ala His Leu Thr Thr Lys Ala Arg Gln

                                   2888                                    2918
AAA CGT TTG CGT CCG GCG TAT GAA AAT GCA GAA AGG CTG TTA CTC GCT CAC ATG CTT CGA
Lys Arg Leu Arg Pro Ala Tyr Glu Asn Ala Glu Arg Leu Leu Leu Ala His Met Leu Arg

                                   2948                                    2978
GAT CGG AGC GTC ATC AAA AAA GTG ATT GAC CGG GTA GGG TTT CAA TTT AAT ATT GAT GAG
Asp Arg Ser Val Ile Lys Lys Val Ile Asp Arg Val Gly Phe Gln Phe Asn Ile Asp Glu

                                   3008                                    3038
CAC CGG GCA TTA GCC GCT TAT CTT TAT GCT TTT TAT GAA GAG GGA GCC GAG CTG ACG CCT
His Arg Ala Leu Ala Ala Tyr Leu Tyr Ala Phe Tyr Glu Glu Gly Ala Glu Leu Thr Pro

                                   3068                                    3098
CAG CAT CTG ATG GCC AGG GTG ACG GAT GAT CAT ATA AGC CAG CTC TTG TCC GAT ATA TTA
Gln His Leu Met Ala Arg Val Thr Asp Asp His Ile Ser Gln Leu Leu Ser Asp Ile Leu

                                   3128                                    3158
ATG CTT CAG GTT AAT CAA GAG CTT AGC GAA GCC GAG TTA TCA GAT TAT GTA AAA AAA GTG
Met Leu Gln Val Asn Gln Glu Leu Ser Glu Ala Glu Leu Ser Asp Tyr Val Lys Lys Val

                                   3188                                    3218
TTG AAT CAA AGA AAT TGG TCA ATG ATA AAA GAA AAA GAG GCG GAA AGA GCC GAA GCA GAA
Leu Asn Gln Arg Asn Trp Ser Met Ile Lys Glu Lys Glu Ala Glu Arg Ala Glu Ala Glu

                                   3248                                    3278
AGG CAA AAA GAT TTT TTA AGA GCT GCT TCT TTG GCT CAA GAA ATC GTT ACA TTG AAC CGA
Arg Gln Lys Asp Phe Leu Arg Ala Ala Ser Leu Ala Gln Glu Ile Val Thr Leu Asn Arg

TCT TTA AAA TAA
Ser Leu Lys End


      3300       3310       3320       3330       3340       3350
CTGGAGAACT GATGAGGAGC ATTTATTGGC AATGATTCCT TGCGGAGGAG CAAATAGATC

      3360       3370       3380       3390       3400       3410
GCTTAACCTC ATCATGAATT GTCATTTCAT TATTCGCACA TTGTTAAAGG CAGTTCACAT

      3420       3430       3440       3450       3460       3470
AGAAAACGCC TGAATGGACC GAATAAGAAT CATACCGCTT ATAGAATTCG TTGCAAGCTT

      3480
TGGAAGGAGG GATCCATA
```

rpoD

```
                                   3518                                    3548
ATG GCT GAT AAA CAA ACC CAC GAG ACA GAA TTA ACA TTC GAC CAA GTA AAA GAG CAA TTA
Met Ala Asp Lys Gln Thr His Glu Thr Glu Leu Thr Phe Asp Gln Val Lys Glu Gln Leu

                                   3578                                    3608
ACA GAG TCT GGT AAA AAA CGT GGC GTT TTG ACA TAT GAA GAA ATT GCT GAG CGT ATG TCC
Thr Glu Ser Gly Lys Lys Arg Gly Val Leu Thr Tyr Glu Glu Ile Ala Glu Arg Met Ser

                                   3638                                    3668
AGC TTT GAA ATT GAA TCA GAC CAA ATG GAT GAG TAT TAT GAA TTT TTA GGT GAA CAA GGT
Ser Phe Glu Ile Glu Ser Asp Gln Met Asp Glu Tyr Tyr Glu Phe Leu Gly Glu Gln Gly

                                   3698                                    3728
GTT GAA TTA ATT AGT GAG AAT GAA GAA ACA GAA GAT CCT AAT ATT CAG CAG CTT GCC AAA
Val Glu Leu Ile Ser Glu Asn Glu Glu Thr Glu Asp Pro Asn Ile Gln Gln Leu Ala Lys

                                   3758                                    3788
GCC GAA GAA GAA TTT GAC CTT AAT GAC CTA AGT GTA CCG CCT GGC GTT AAA ATC AAT GAC
Ala Glu Glu Glu Phe Asp Leu Asn Asp Leu Ser Val Pro Pro Gly Val Lys Ile Asn Asp

                                   3818                                    3848
CCA GTT CGT ATG TAT TTA AAG GAA ATC GGT CGG GTT AAC CTT CTT TCT GCA AAA GAA GAA
Pro Val Arg Met Tyr Leu Lys Glu Ile Gly Arg Val Asn Leu Leu Ser Ala Lys Glu Glu

                                   3878                                    3908
ATC GCC TAC GCT CAA AAG ATT GAA GAA GGT GAC GAA GAA TCT AAA CGC AGA TTG GCT GAA
Ile Ala Tyr Ala Gln Lys Ile Glu Glu Gly Asp Glu Glu Ser Lys Arg Arg Leu Ala Glu
```

```
                                   3938                                    3968
GCG AAC CTG CGG CTT GTT GTC AGT ATC GCA AAA CGG TAT GTC GGA CGC GGT ATG CTG TTC
Ala Asn Leu Arg Leu Val Val Ser Ile Ala Lys Arg Tyr Val Gly Arg Gly Met Leu Phe

                                   3998                                    4028
CTT GAT CTG ATC CAT GAA GGA AAC ATG GGC CTG ATG AAA GCC GTT GAA AAA TTT GAT TAT
Leu Asp Leu Ile His Glu Gly Asn Met Gly Leu Met Lys Ala Val Glu Lys Phe Asp Tyr

                                   4058                                    4088
CGC AAA GGT TAT AAA TTC AGT ACG TAT GCT ACG TGG TGG ATC AGA CAG GCG ATT ACA CGC
Arg Lys Gly Tyr Lys Phe Ser Thr Tyr Ala Thr Trp Trp Ile Arg Gln Ala Ile Thr Arg

                                   4118                                    4148
GCC ATT GCC GAT CAG GCG AGA ACG ATC CGG ATT CCC GTT CAT ATG GTT GAA ACC ATT AAT
Ala Ile Ala Asp Gln Ala Arg Thr Ile Arg Ile Pro Val His Met Val Glu Thr Ile Asn

                                   4178                                    4208
AAA TTA ATC CGT GTG CAG CGT CAA TTA CTG CAA GAC TTA GGC AGA GAA CCA ACA CCT GAA
Lys Leu Ile Arg Val Gln Arg Gln Leu Leu Gln Asp Leu Gly Arg Glu Pro Thr Pro Glu

                                   4238                                    4268
GAA ATT GCG GAA GAT ATG GAT TTA ACG CCT GAA AAA GTA CGC GAA ATC TTA AAG ATT GCT
Glu Ile Ala Glu Asp Met Asp Leu Thr Pro Glu Lys Val Arg Glu Ile Leu Lys Ile Ala

                                   4298                                    4328
CAA GAG CCG GTA TCT CTG GAA ACA CCG ATC GGT GAA GAG GAT GAC TCG CAC CTT GGT GAT
Gln Glu Pro Val Ser Leu Glu Thr Pro Ile Gly Glu Glu Asp Asp Ser His Leu Gly Asp

                                   4358                                    4388
TTC ATT GAA GAC CAA GAA GCA ACT TCA CCT TCT GAC CAC GCC GCA TAC GAG CTA TTG AAA
Phe Ile Glu Asp Gln Glu Ala Thr Ser Pro Ser Asp His Ala Ala Tyr Glu Leu Leu Lys

                                   4418                                    4448
GAG CAG CTG GAA GAT GTG CTT GAT ACG TTA ACT GAT CGT GAA GAA AAT GTA TTG CGT CTT
Glu Gln Leu Glu Asp Val Leu Asp Thr Leu Thr Asp Arg Glu Glu Asn Val Leu Arg Leu

                                   4478                                    4508
CGA TTC GGT CTT GAT GAC GGC CGT ACA AGA ACA TTA GAA GAG GTC GGC AAA GTA TTT GGA
Arg Phe Gly Leu Asp Asp Gly Arg Thr Arg Thr Leu Glu Glu Val Gly Lys Val Phe Gly

                                   4538                                    4568
GTA ACG AGA GAG CGT ATT CGA CAA ATC GAA GCC AAA GCG TTG CGG AAA CTA AGA CAT CCT
Val Thr Arg Glu Arg Ile Arg Gln Ile Glu Ala Lys Ala Leu Arg Lys Leu Arg His Pro

                                   4598
AGC AGA AGT AAA CGT TTG AAA GAT TTC CTT GAA TAA
Ser Arg Ser Lys Arg Leu Lys Asp Phe Leu Glu End


      4614       4624       4634       4644       4654       4664
GATGGAACGG GTCTTGAAGA TCCGTTCTTC TTTTTTTAAA AAGATATATG GATAATATGC

      4674       4684       4694       4704       4714       4724
CTTTATTTTA CTGAAAAATG ATGTCATTTG CAAATGAACA TTGTGGTGAA AAATTTCAAA

      4734       4744       4754       4764       4774       4784
ATCTAATTCC ATATTTTCTA TTGTAAGCGT ATACAATACA TTATACAATA GAATAAAAAG

      4794       4804       4814       4824       4834       4844
GATATTAGAG ATTTAGGCAT GTTTCTATTT CAATTTTGTA TAAAGTGTGA ATAAAAAACT

      4854       4864       4874       4884       4894       4904
TTTGTATAGC AATCCATTTA CTTTTTGTAA AAATAAGTTA GAATTAGAAG TGTTTACATA

      4914       4924       4934       4944       4954       4964
GGGGGAAGGA TCAAAAAGGG GGAAGGGAAA TGAAATGGAA CCCGCTTATT CCATTTTTGC

      4974       4984       4994       5004       5014       5024
TGATCGCTGT TTTAGGAATG GGTCTAACTT TCTTTTTATC AGTAAAAGGA CTTGATGACT

      5034       5044       5054       5064       5074       5084
CTCGGGAGGT TGCGAGCGGA GGAGAAAGCA AATCTGCTGA AAAGAAAGAT GCAAACGCTT

      5094       5104       5114       5124
CACCAGAAGA AATTTACAAG GCAAATCGCA TCGCATCGCA TGC
```

Figure 2. Nucleotide sequence of B. subtilis σ^{43} operon. The DNA sequence of the upper strand is given in the 5' to the 3' direction, numbered from nucleotide 1 at the 5' end EcoRI site. The predicted amino acid sequence for each open reading frame is given below the corresponding DNA sequence. Sequences for promoters, ribosomal binding sites, and terminator are underlined.

Table 1. Amino Acid Composition (mol%) Analysis

Group	*P23*	*dnaE*	*rpoD*	Ave. B.s. Proteins*
Small aliphatic (A+G)	10.2	13.3	11.2	15.0
Hydroxyl (S+T)	12.2	11.0	9.6	13.4
Acidic (D+E)	14.3	14.3	21.1	14.2
Acidic + acid amide (D+E+N+Q)	22.4	22.3	28.0	23.1
Basic (K+R+H)	14.3	16.7	16.0	14.8
Hydrophobic (L+I+V+M)	22.9	22.8	24.9	26.3
Aromatic (F+Y+W)	8.7	9.5	6.0	7.7
Charged (D+E+K+R+H)	28.6	31.0	37.1	29.0

* The amino acid composition of average *B. subtilis* proteins is calculated from 35 sequenced genes published up to 1985.

were fused to the 8th amino acid residue of β-galactosidase, which led us to the conclusion that ATG (855) was functionally active *in vivo*. Hence, we designated the first gene of the operon as *P23* from these data and for the reason that its physiological function is still unknown.

The deduced amino acid sequence of P23 was examined for homology against the NBRF Protein Data Bank using the Microgenie Program (18), but no significant homology was found to any of the known proteins in the bank, indicating to us that P23 was not homologous with *E. coli* S21, the first gene in the σ^{70} operon (also their size difference is significant), and that P23 might be unique to *B. subtilis* , or that its counterpart in *E. coli* has not been characterized as yet. The latter case is a possibility, since a reasonable degree of homology has been found between many *B. subtilis* and *E. coli* proteins.

The deduced amino acid composition of P23 is shown in Table 1, together with those of DNA primase, σ^{43} and an average of *B. subtilis* proteins for comparison. One difference noted from σ^{43}, which is a highly acidic protein typical of most transcription factors (7,19,20), is that P23 is more like the average composition of *B. subtilis* proteins. Thus it is difficult to categorize this protein based on its amino acid composition. We are currently raising antibody against P23 using a P23-gal fusion protein as antigen, hoping that this will provide us a tool to determine the location and possibly the function of P23 in B. subtilis.

Regulatory Features

Previously, we reported that there was no promoter activity detected within the intercistronic regions of the operon except for a weak heat shock promoter activity located at the C-termianl end of the dnaE gene (8). So we concluded that a promoter(s) should exist in front of P23 if the operon was composed of three genes as in the case of E.coli σ^{70} operon (10). By sequence analysis we did find at least two potential promoters with significant homology to the consensus sequence of B. subtilis σ^{43} promoters (1,2), which were then confirmed to function in vivo by fusing the 211 bp Sau3A fragment (609-821) to the subtilisin gene (aprA) in a promoter-probe plasmid pSB (Wong and Doi, unpublished data). These sequences are underlined in Fig. 2, and designated as P1 and P2. To our surprise, one additional promoter activity was detected when the 316 bp Sau3A fragment (829-1136) downstream of P1 and P2 was cloned in pSB. This promoter (P3) was temporally regulated in that it was not expressed until the culture reached the sporulation phase, while P1 and P2 were expressed efficiently mainly during growth. More detailed mapping and functional characterization of these promoters are in progress, and will be published elsewhere.

Earlier sequence analysis (6) allowed us to identify a sequence typical of rho independent terminator (21) immediately following the TAA stop codon of rpoD gene (underlined in Fig. 2). Recently, we confirmed its termination activity in vivo by subcloning the PvuII-AhaIII fragment (4394-4641) into a B. subtilis terminator-probe plasmid pST19 constructed in our laboratory (Wang and Doi, unpublished data). We were able to show that introduction of this 247 bp fragment reduced the activity of the indicator enzyme in the terminator probe (subtilisin, in this case) by more than 90% compared to the control (vector alone), indicating that this was a relatively strong terminator (data not shown).

Thus we have determined the presence of three genes in the

Table 2. Codon Usage of B.subtilis Sigma-43 Operon*

AA Codon	P23	dnaE	rpoD	B.s.	E.c.	AA Codon	P23	dnaE	rpoD	B.s.	E.c.
Phe UUU	1.00	.73	.45	.67	.37	Tyr UAU	.86	.74	.80	.62	.40
Phe UUC	.00	.27	.55	.33	.63	Tyr UAC	.14	.26	.20	.38	.60
Leu UUA	.24	.20	.31	.23	.07	OCH UAA	1.00	1.00	1.00	.57	.75
Leu UUG	.19	.12	.15	.13	.09	AMB UAG	.00	.00	.00	.06	.08
Leu CUU	.29	.29	.28	.24	.07	His CAU	.60	.78	.50	.68	.54
Leu CUC	.00	.10	.00	.11	.07	His CAC	.40	.22	.50	.32	.46
Leu CUA	.05	.03	.08	.05	.02	Gln CAA	.71	.53	.65	.56	.24
Leu CUG	.24	.26	.18	.23	.68	Gln CAG	.29	.47	.35	.44	.76
Ile AUU	.62	.39	.54	.50	.36	Asn AAU	.78	.58	.67	.53	.26
Ile AUC	.23	.35	.40	.40	.61	Asn AAC	.22	.42	.33	.47	.74
Ile AUA	.15	.26	.00	.10	.03	Lys AAA	.73	.73	.88	.75	.76
Met AUG	1.00	1.00	1.00	1.00	1.00	Lys AAG	.27	.27	.12	.25	.24
Val GUU	.39	.29	.43	.20	.36	Asp GAU	.83	.64	.58	.63	.46
Val GUC	.23	.31	.14	.24	.15	Asp GAC	.17	.36	.42	.37	.54
Val GUA	.15	.26	.33	.23	.22	Glu GAA	.81	.63	.77	.69	.73
Val GUG	.23	.14	.10	.23	.27	Glu GAG	.19	.37	.23	.31	.27
Ser UCU	.20	.30	.31	.25	.23	Cys UGU	.33	.20	.00	.50	.43
Ser UCC	.27	.09	.06	.12	.27	Cys UGC	.67	.80	.00	.50	.57
Ser UCA	.20	.13	.13	.14	.07	OPL UGA	.00	.00	.00	.37	.17
Ser UCG	.20	.15	.06	.11	.11	Trp UGG	1.00	1.00	1.00	1.00	1.00
Pro CCU	.40	.53	.50	.33	.12	Arg CGU	.15	.08	.33	.24	.56
Pro CCC	.20	.07	.08	.09	.07	Arg CGC	.00	.11	.17	.18	.36
Pro CCA	.00	.20	.17	.18	.16	Arg CGA	.23	.14	.07	.09	.03
Pro CCG	.40	.20	.25	.40	.65	Arg CGG	.00	.17	.17	.11	.03
Thr ACU	.11	.10	.10	.17	.25	Ser AGU	.07	.07	.31	.10	.06
Thr ACC	.00	.14	.10	.14	.50	Ser AGC	.07	.26	.13	.23	.26
Thr ACA	.56	.29	.30	.42	.07	Arg AGA	.54	.28	.27	.28	.01
Thr ACG	.33	.47	.50	.27	.18	ARG AGG	.08	.22	.00	.10	.01
Ala GCU	.50	.37	.26	.26	.26	Gly GGU	.10	.14	.53	.21	.48
Ala GCC	.20	.26	.35	.19	.21	Gly GGC	.30	.35	.32	.33	.39
Ala GCA	.30	.22	.17	.31	.22	Gly GGA	.40	.37	.16	.32	.05
Ala GCG	.00	.15	.22	.24	.31	Gly GGG	.20	.14	.00	.14	.08

* 1) B.s. codon usage frequency of average B. subtilis proteins, compiled from 10,919 codons of 35 sequenced genes.
2) E.c. codon usage frequency of average E. coli proteins, from reference 24.
3) The number tabulated is the fraction usage of each codon compared with total for identical amino acid.

RNA polymerase major σ^{43} operon in the order P23, dnaE and rpoD and have physically and functionally located the promoter and terminator regions for the operon.

Codon Usage

It has been well established in E. coli that there is a correlation between expression level of a gene and its codon usage pattern. The more highly expressed genes show a very non-random pattern of codon usage, utilizing a restricted set of codons which are recognized by major species of isoacceptor tRNAs, while genes which are expressed at very low levels show an almost random pattern of codon usage (22,23). The analysis of this kind of correlation in *B. subtilis* has been limited due to the small number of sequenced genes, and the lack of knowledge concerning the expression levels of these genes *in vivo*. Recently, the rapid advance of cloning and sequencing of *B. subtilis* genes have allowed us to compile a codon usage table for average *B. subtilis* proteins and compare this with the codon usage of each gene in the σ^{43} operon, and also with that of average *E. coli* proteins (24) (see Table 2). Comparison of the codon usages between *B. subtilis* and *E. coli* led us to the general conclusion that *B. subtilis* tends to more evenly or randomly distribute the codons for its amino acids than *E. coli*. Nevertheless, the rarely used codons in *E. coli*, CUA (Leu), AUA (Ile), CCC (Pro), AGG (Arg) and GGG (Gly), were also used least in *B. subtilis*, although the bias is not as dramatic as that in *E. coli*. When the usage frequencies of codons AUA, AGG and GGG in the three genes of the operon were carefully examined, striking differences were found. In *rpoD*, a relatively highly expressed gene in *B. subtilis* [2,000-10,000 molecules/cell during growth (6)] just as its counterpart in *E. coli* (10), these codons were not used at all, while in *P23* and *dnaE* they were used quite frequently. Especially in *dnaE*, the usage frequencies for codons AUA and AGG were 0.26 and 0.22, which was much higher than those for the average *B. subtilis* proteins, which were 0.10 and 0.10, respectively. This preliminary analysis suggests to us that *P23* and *dnaE* are expressed at lower levels than *rpoD*, which was not unexpected for *dnaE* since the DNA primase is required only in small amounts during DNA replication (25). Also consistent with this idea were the relative strengths of the ribosome binding sites for the three genes, which were found in an increasing order of $\Delta G'$ = -9.2 kcal/mol for *P23*, -13.8 kcal/mol for *dnaE* and -18.8 Kcal/mol for *rpoD*.

Since we do not know the function of P23, we can only speculate that it may have some role in translation or act as a regulatory protein present in relatively low concentrations in the cell.

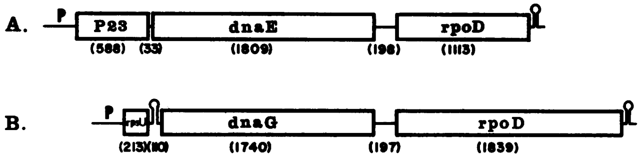

Figure 3. Schematic representation of the organization of *B. subtilis* (A) and *E. coli* (B) major σ operons.
The number of base pairs is shown in parentheses for each structural unit. The organization of the *E. coli* σ^{70} operon is based on results from Burton *et al.* (10) and Lupski *et al.* (26). For simplicity, internal promoters and RNA processing sites are not shown.

Operon Organization

A comparison of the two major σ operons from *B. subtilis* and *E. coli* is illustrated in Fig. 3. Both operons contain three genes including the DNA primase and major σ genes, which are ordered in the same way. The operons are all under control of multiple promoters, indicating a complex transcription regulation system. However, significant differences do exist between them, of which the most striking is the first gene in the operon. P23 is more than twice as large as S21, and there is no sequence homology at all between them, while the middle gene products (DNA primases) are 31% homologous (8), and the last gene products (major σ factors) more than 50% homologous (6). Also different are the sizes of the intercistronic regions between the first and second genes, 33 bp in *B. subtilis* and 110 bp in *E. coli*, while that between the second and third genes are very similar in size (8,10). In *E. coli*, there is a mRNA processing site immediately following the *dnaG* gene (10), which was not found in *B. subtilis*.

DISCUSSION

We determined the nucleotide sequence of a 5 kb fragment in the *dnaE*-*rpoD* region of the *B. subtilis* chromosome and found three open reading frames transcribed in the same counterclockwise direction, two of which were identified as genes for the DNA primase and RNA polymerase major σ^{43} factor. The function of the first gene is still unknown. Discovery of the promoters in front of *P23* and a terminator 3' to *rpoD*, and the absence of promoters in the intercistronic regions between these genes provided strong evidence that P23, dnaE, and rpoD comprised a three gene operon. The operon was named RNA polymerase major sigma (σ^{43}) operon in analogy to that in *E. coli*.

The structure and organization of the σ^{43} operon resemble those of the *E. coli* σ^{70} operon except for the first gene. The size of P23 and its lack of amino acid sequence homology with S21 represent the most significant differences between the two operons at the molecular level. Since a reasonably high degree of homology exists between the DNA primase and σ genes, one might have expected some homology between the first genes P23 and rpsU. We are currently attempting to identify P23 by use of immunological and cell fractionation techniques to see whether P23 might be associated with the ribosome fraction of *B. subtilis*.

The regulation of the *E. coli* σ^{70} operon is very complex since it is an important operon controlling not only translation, but also DNA replication and transcription (10,26). One of its interesting features is the control mechanism(s) to keep the expression of the dnaG gene lower than its adjacent genes, rpsU and rpoD. At least four mechanisms have been proposed including an internal terminator between the first and second genes (10), a weak ribosomal binding site for dnaG (10), frequent use of rare codons (27), and a mRNA processing site between the second and third genes (10). Although no experimental data are available concerning the expression level of dnaE *in vivo*, a low expression is expected from its function, and its counterpart in *E. coli*. The results of our codon usage analysis and the comparison of ribosomal binding sites are also in good agreement with the notion that dnaE represents a weakly expressed gene, and rpoD a fairly highly expressed gene. However, besides the possible regulatory mechanisms at the translational level, it is very likely that there are also control mechanism(s) involved at the transcriptional level.

In general, *B. subtilis* requires a more stringent Shine-Dalgarno sequence for gene expression than *E. coli* (28,29). The calculated free energies of interaction of the Shine-Dalgarno regions of *B. subtilis* mRNAs with the 3' end of its 16s rRNA have an average value of -17 kcal/mol (30), contrasted with an average of -11 kcal/mol for that in *E. coli* (31). However, the calculated free energy value for *P23* gene, -9.2 kcal/mol, is far below that of the average for *B. subtilis*. Considering this and the codon analysis data, it is tempting to propose that *P23* encodes a regulatory protein which is weakly expressed, but physiologically important. Also, since *B. subtilis* cells undergo differentiation and can form spores, it is possible that the cell may have evolved a unique regulation system that is absent in *E. coli*, and that P23 may be one of the members in that system. The possibility also exists that *P23* encodes an unidentified component of the *B. subtilis* translation machinery which is absent or has not been identified as yet in *E. coli*, since it has been reported that sequences other than the Shine-Dalgarno region can affect the translation efficiency of a gene (29,32); it thus is possible that *P23* might still be expressed efficiently *in vivo*. More experimental data are required before we can say anything conclusive about this cryptic gene.

ACKNOWLEDGEMENTS

We thank Sui-Lam Wong for use of plasmid pSB. This research was supported in part by National Institute of General Medical Sciences grant GM19673 and National Science Foundation grant PCM82-18304.

*To whom correspondence should be addressed

REFERENCES

1. Doi, R.H. (1982) Arch. Biochem. Biophys. 214, 772-781
2. Losick, R. and Pero, J. (1981) Cell 25, 582-584
3. Doi, R.H., Gitt, M., Wang, L.-F., Price, C.W. and Kawamura, F. (1984) In: Molecular Biololgy of Microbial differentiation (J.A. Hoch and P. Setlow, eds.), pp147-161, American Society for Microbiology, Washington, D.C.
4. Price, C.W., Gitt, M.A. and Doi, R.H. (1983) Proc. Natl. Acad. Sci. USA 72, 1589-1593
5. Price, C.W. and Doi, R.H. (1985) Mol. Gen. Genet. 201,88-95
6. Gitt, M.A., Wang, L.-F. and Doi, R.H. (1985) J. Biol. Chem. 260, 7178-7185
7. Burton, Z., Burgess, R., Lin, J., Moore, D., Holder, S., and Gross, C. (1981) Nucl. Acids Res. 9, 2889-2903
8. Wang, L.-F., Price, C.W. and Doi, R.H. (1985) J. Biol. Chem. 260, 3368-3372
9. Smiley, B.L., Lupski, J.R., Svec, P.S., McMacken, R. and

Godson, G.N. (1982) Proc. Natl. Acad. Sci. USA 79, 4550-4554
10. Burton, Z.F., Gross, C.A., Watanabe, K.K. and Burgess, R.R. (1983) Cell 32, 335-349
11. Messing, J., Crea, R. and Seeberg, P.H. (1981) Nucl. Acids Res. 9, 309-321
12. Messing, J. and Vieira, J. (1982) Gene 19, 269-276
13. Davis, R.W., Botstein, D. and Roth, J.R. (1980) Advanced Genetics: A Manual for Genetic Engineering, Cold Spring Harbor Laboratory, New York.
14. Maniatis, R., Fritsch, E.f., and Sanbrook, J. (1982) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, New York.
15. Sanger, F., Nicklen, S. and Coulsen, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467
16. Delaney, A.D. (1982) Nucl. Acids Res. 10, 61-67
17. Pustell, J. and Kafatos, F. (1982) Nucl. Acids Res. 10,51-59
18. Queen, C. and Korn, L. (1984) Nucl. Acids Res. 12, 581-599
19. Ishii, S., Ihara, M., Maekawa, T., Nakamura, Y., Uchida, H., and Imamoto, R. (1984) Nucl. Acids Res. 12, 3333-3342
20. Hunt, T.P. and Magasanik, B. (1985) Proc. Natl. Acad. Sci. USA 82, 8453-8457
21. Platt, T. and Bear, D.G. (1983) In: Gene Function in Prokaryotes (J. Beckwith, J. Davis, and J.A. Gallant, eds.), Cold Spring Harbor Laboratory, New York.
22. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. (1981) Nucl. Acids Res. 9, r43-r74
23. Ikemura, T. and Ozeki, H. (1982) Cold Spring Harbor Symp. Quant. Biol. 47, 1087-1097
24. Alff-Steinberger, C. (1984) Nucl. Acids Res. 12, 2235-2241
25. Kornberg, A. (1981) DNA Replication, Freeman & Co., San Francisco.
26. Lupski, J.R., Smiley, B.L., Blattner, F.R. and Godson, G.N. (1982) Mol. Gen. Genet. 185, 120-128
27. Konigsberg, W. and Godson, G.N. (1983) Proc. Natl. Acad. Sci. USA 80, 687-691
28. McLaughlin, J.R., Murray, C.L. and Rabinowitz, J.C. (1981) J. Biol. Chem. 256, 11283-11291
29. Band, L. and Henner, D.J. (1984) DNA 5, 17-21
30. Murray, C.L. and Rabinowitz, J.c. (1982) J. Biol. Chem. 257, 1053-1062
31. Gold, L., Pribnow, d., Schneider, T., Shinedling, S., Singer, B.S. and Stormo, G. (1981) Ann. Rev. Microbiol. 35, 365-403
32. Stanssens, P., Remaut, E. and Fiers, W. (1985) Gene 36,211-223