Volume 9 Number 24 1981 Nucleic Acids Research

Nucleotide sequence at the end of the gene for the RNA polymerase β' subunit (rpoC)

C.Squires, A.Krainer, G.Barry, W.-F.Shen and C.L.Squires

Department of Biological Sciences, Columbia University, New York, NY 10027, USA

Received 14 September 1981

ABSTRACT

We have determined the DNA sequence surrounding the transcription terminator following *rpoC*, the gene that codes for the β' subunit of RNA polymerase in *E. coli* K12. The 2044 bp sequence obtained contains the distal 335 codons of *rpoC* followed by a 212 bp non-coding region and a second open reading frame (ORFa) of 179 codons. The final 181 nucleotides of the sequence form the 5' end of a third open reading frame (ORFb). The *in vivo* 3' end of the *rpoC* mRNA was located by analysis of RNA/DNA hybrids cleaved with nuclease S1 (S1 mapping). These results indicated that the major transcription termination of the *rplJL-rpoBC* transcription unit occurs a short distance past the translation stop codon for *rpoC*. Four regions of symmetry, suggesting secondary structure in the mRNA, were found in the DNA sequence near the *rpoC* translation termination codon. The last of these hairpin structures is similar to other rho-independent transcription terminators and its 3' end coincides with the end of the *rpoC* mRNA as predicted by S1-mapping. Inspection of the open reading frames indicates that *rpoC* uses a high percentage of codons that are recognized by the major tRNA species of *E. coli* while ORFa and ORFb contain many codons recognized by minor tRNA species. ORFa specifies a very basic peptide.

INTRODUCTION

The genes that encode the β and β' subunits of RNA polymerase (*rpoB* and *rpoC*) are found in a large transcription unit which also contains the genes for the ribosomal proteins L10 and L7/12 (*rplJ* and *rplL*). This unit is preceded by another transcription unit containing genes for ribosomal proteins L11 and L1 (*rplK* and *rplA*)(Fig. 1A)(1,2,3). We are interested in defining the ends of these transcription units and determining the role they play in the regulation of surrounding genes. We have previously identified a strong terminator of transcription on a 2.0 kbp DNA fragment that spans the end of the *rpoC* gene (4). It is the purpose of this paper to precisely locate this terminator and to describe its genetic environment.

The regulation of gene expression within the *rplJL-rpoBC* transcription unit is complex, involving transcriptional and translational control elements and possibly mRNA processing as well. There are strong promoters before *rplK* and *rplJ* (PK and P1)(2) and weak promoters have been identified before *rplL*, *rpoB* and *rpoC* (P2, P3 and P4)(4). While *rplJL* and *rpoBC* are promoted from P1 and *rplKA* are promoted from PK, no terminator-like secondary structure is evident in the sequence following *rplA* (5) and preliminary S1 mapping results indicate that some transcripts do

not have ends between *rplA* and *rplJ* (CL Squires, unpublished). Therefore, the PK promoter may also be involved in *rplJL-rpoBC* expression. An attenuator (*atn*, Fig. 1A), which causes 80% transcription termination under conditions of normal growth, is located in the intercistronic region between *rplL* and *rpoB* (6). Whether *atn* is modulated under other growth conditions which disrupt coordinate expression of *rplJL* and *rpoBC* (7) is not known. Other mechanisms besides initiation and termination of transcription are also involved in the modulation of gene expression in this region. Translational regulation has been demonstrated in the case of *rplJL* and *rplKA* expression (8,9,10). In addition we have described an RNaseIII processing site (*rps*, Fig 1A) in the mRNA of this transcription unit (6). Clones in which *rps* has been deleted produce less of the following gene products (S Hastrup, personal communication). This suggests that *rps* and RNaseIII may also be involved in *rpoBC* control.

Much of the *rplKAJL-rpoBC* region has already been sequenced (Fig. 1C)(5,11,12). In this paper, we have used shotgun cloning in the single-stranded DNA bacteriophage, M13mp7 (13), to

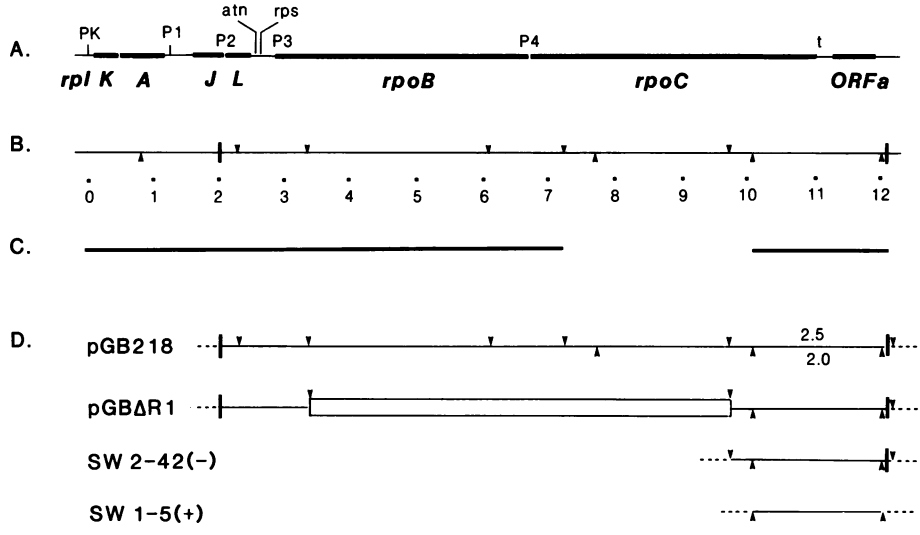

Figure 1. A. Genetic map. The heavy lines indicate the location and extent of genes and the open reading frame, ORFa. Control elements (PK, P1, P2, P3, P4, *atn*, *rps* and *t*) are identified and discussed in the text. B. Restriction map. Sites cleaved by EcoRI (↓), BglII (↑) and HindIII (|) are shown. Numbers under the map indicate the distance in kbp. All maps in this figure are to the same scale. C. Regions for which the DNA sequence is known (refs. 5,11,12 and this paper). D. Diagram of hybrid plasmids and phage used in this work. Solid lines indicate *E. coli* sequences, dashed lines represent vector sequences. pGB218 is a clone of the 10 kbp HindIII fragment from λ*drif*^{d}18 in the vector pBR322. Note that the EcoRI site to the right of the right-hand HindIII site is in pBR322. pGBΔRI is the same HindIII fragment, from which three EcoRI fragments have been deleted, cloned in the vector pMC81. SW2-42(-) and SW1-5(+) are clones of the 2.5 kbp EcoRI and 2.0 kbp BglII fragments from pGB218 in the vector M13mp7.

determine the DNA sequence of the 2.0 kbp region that spans the end of the *rpoC* gene and we have demonstrated that the strong terminator activity associated with this region is located near the end of *rpoC*. We have identified a structure in the nucleic acid sequence that is similar to other strong rho-independent terminators. The sequence has also revealed the distal 335 codons of *rpoC* followed by two additional open reading frames, ORFa and ORFb.

MATERIALS AND METHODS

Origin of DNA sequenced

The ultimate source of the DNA sequenced in this work was the bacteriophage λ*drif*^{d}18 which carries genes from the 88-89 minute region of the *E. coli* chromosome. We have previously described the subcloning, in pBR322, of the 10 kbp HindIII fragment (Fig. 1B) from λ*drif*^{d}18. The resulting plasmid, pGB218 (4), was the source of DNA for the isolation of the 2.5 kbp EcoRI and the 2.0 kbp BglII fragments which span the end of the *rpoC* gene. The subcloning of these fragments as well as the shotgun cloning of Sau3A, HpaII, TaqI and AluI fragments derived from the 2.0 kbp BglII fragment in the vector M13mp7 are described in this paper.

Strains, phage and plasmids

Strain MC1000 (Δ*lacIZY galKU* Δ*ara-leu strA*) (14) was obtained from M. Casadaban. Strain 71.18 (Δ*lac-proAB supE thi* /F'*lacI*^{q}ZΔM15 *proA*$^{+}$*B*$^{+}$) (15) and phage M13mp7 (13) were obtained from J. Messing. M13mp7 produces blue plaques when plated with strain 71.18 on special media (X-gal media) which contains IPTG and 5-bromo-4-chloro-3-indolyl-β-D-galactoside (obtained from Bachem). Insertion of foreign DNA into one of the cloning sites in M13mp7 RF DNA usually causes the transformants to make colorless plaques on X-gal media. Our isolation and characterization of plasmids pGB218 and pGBΔR1 (Fig.1D) is reported elsewhere (4). The bacteriophage λT4*lig* (16) and a method for the purification of the T4-DNA ligase were obtained from S. Silverstein.

Cloning in M13mp7

a. Routine methods. General methods for the propagation, isolation and purification of DNA from phage, as well as the method for transfection with M13 RF DNA were obtained from J. Messing (13,17, personal communication). A lysozyme-detergent method was used for the isolation of plasmid and RF DNA (18). Analytical DNA isolation was done by the alkaline-SDS method (19). Cohesive DNA ends were ligated by overnight incubation in a 20 μl solution of 66mM TrisHCl pH 7.4, 6.6mM $MgCl_2$, 60 μM rATP, 10mM DTT and 0.25 units of T4 DNA ligase. Flush ended ligations (FEL) were performed under similar conditions except incubation was at room temperature for four hours in the presence of 1 unit ligase and 400 μM rATP. A rapid hybridization method (20) was used to identify the orientation of the sequences cloned in M13. Clones are designated plus (+) if they hybridize with *rpoC* mRNA or if they are derived from the same strand that codes for *rpoC*.

b. Preparation of vector DNAs. The RF DNA was adjusted to 0.1 mg/ml and digested with EcoRI, BamHI, AccI or HincII for the minimum time required to give total cleavage (digested DNAs were designated: M13mp7/E, /B, /A and /H respectively). Vector DNAs were tested for their ability to transfect strain 71.18 to give blue plaques on X-gal media. Uncleaved, cleaved and re-ligated-cleaved DNAs were compared in these tests. Typical results showed that cleavage reduced the transfection frequency 100-fold. Re-ligation of cleaved DNA restored transfections to 30-50 percent. In these control experiments the presence of colorless plaques indicates that the enzymes used for digestion have exonuclease contamination. Many commercially-obtained restriction endonucleases produce an unacceptably large number of such plaques. We used preparations that gave fewer than three percent colorless plaques.

c. Subcloning the rpoC terminator region. The 2.5 kbp EcoRI and 2.0 kbp BglII fragments were prepared from pGB218 DNA (Fig. 1D), purified on Sea Plaque agarose (Marine Colloids) and ligated into M13mp7/E and /B respectively. Ligated mixtures were used to transfect 71.18 cells which were plated on X-gal media. Colorless plaques were isolated and single-stranded and RF DNAs were prepared. Digests of the RF DNAs were compared with the isolated 2.5 kbp EcoRI and 2.0 kbp BglII fragments by agarose electrophoresis. Single-stranded DNA was used in DNA/DNA hybridizations to identify clones having complementary sequences (20). Two clones were chosen: SW1-5(+) is a 2.0 kbp BglII clone and SW2-42(-) is a 2.5 kbp EcoRI clone. The relationship of these clones to the plasmid pGB218 is shown in Fig. 1D. The extreme left-hand sequences of SW2-42(-) (31 nucleotides between HindIII and EcoRI sites) are derived from pBR322. Otherwise, the inserted sequences in SW1-5(+) and SW2-42(-) are from $\lambda drif^{d}18$.

d. Shotgun cloning in M13mp7. Shotgunned clones were derived from the 2.0 kbp BglII fragment isolated from pGB218 DNA. Aliquots were digested with Sau3A, HpaII, TaqI and AluI and mixed with the appropriate vector DNA: Sau3A fragments with M13mp7/B, HpaII and TaqI fragments with M13mp7/A and AluI fragments with M13mp7/H. AluI fragments were ligated using FEL conditions and the other fragments were ligated under cohesive-end conditions. Strain 71.18 was transfected with the ligation mixtures and plated on X-gal media. Between 60 and 120 colorless plaques were isolated from each preparation. Infected cells were stored in 40% glycerol at -20°C.

DNA sequence analysis by chain termination

Single-stranded DNA preparations of the clones were first screened using the dideoxythymidine triphosphate (ddTTP) chain terminator sequencing reaction (21). All of the clones which produced unique T-sequence ladders were then sequenced using all four ddNTP chain termination reactions. The method of Heideker *et al* (17) was used except for the following: Sequencing reactions were primed with 1.5 μM TCACGACGTTGT (Collaborative Research). Primer was annealed to template by mixing 1 μl of primer mix (7.5 μM primer, 0.3 M NaCl, 0.1 M TrisHCl pH 7.6, 0.1 mM EDTA) and 1 μl template (0.5 mg/ml) per each 5 μl reaction. The mixture was heated at 70°C for 5 minutes and cooled to room temperature. The chain-terminated samples

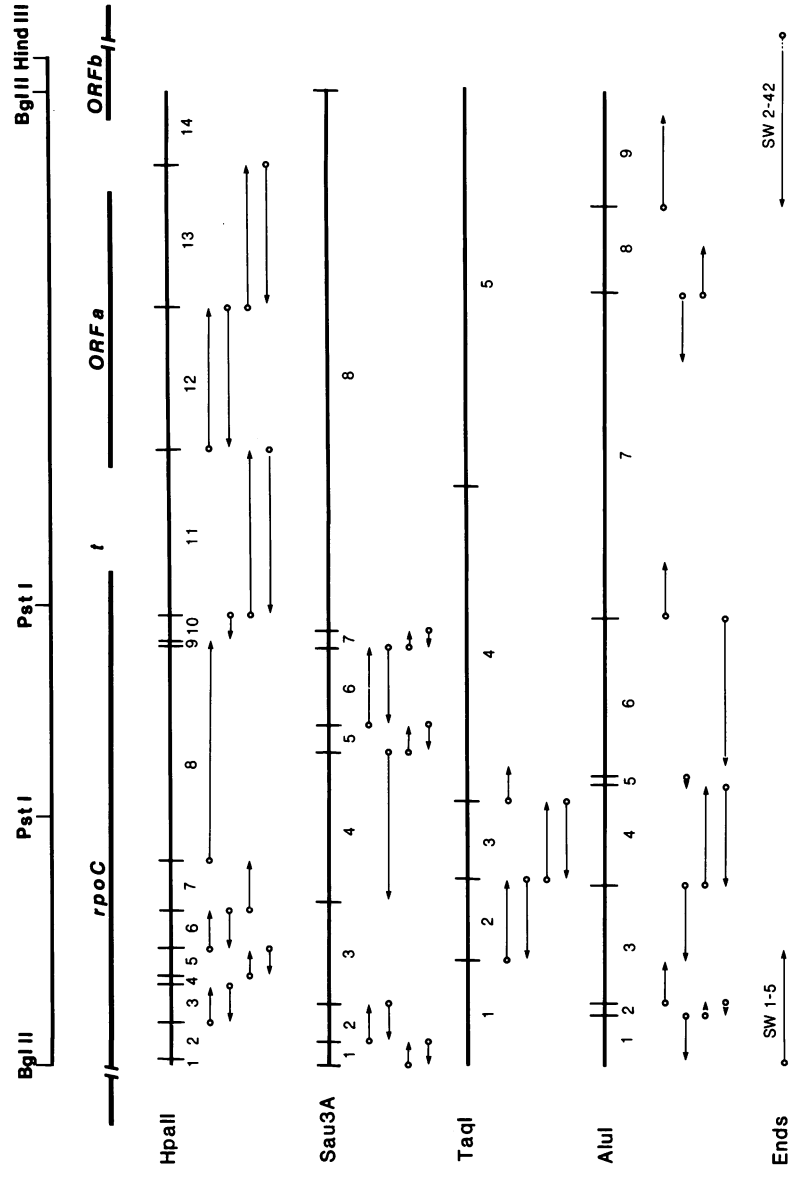

Figure 2. DNA sequences obtained.

Figure 3. DNA sequence encoding the end of *rpoC* and two open reading frames (ORFa and ORFb). The corresponding amino acid sequences are written above the DNA sequence. Underlined features are discussed in the text.


1021: CAAATAACGTAAAAACCCGCTTCGGCGGGTTTTTTTATGGGGGGAGTTTAGGGAAAGAGC
1081: ATTTGTCAGAATATTTAAGGAATTTCTGAATACTCATAATCAATGTAGAGATGACTAATA
1141: TCCTGAAACTGACTGAACTAATTGAGTCAAACTCGGCAAGGATTCGATACTATTCCTGTG
METLysGlnGluValGluLysTrpArgProPheGlyHis
1201: TAACTTTCTTAAGGAACGAGAATGAAACAGGAAGTGGAAAAGTGGCGACCTTTTGGACAT
ProAspGlyAspIleArgAspLeuSerPheLeuAspAlaHisGlnAlaValTyrValGln
1261: CCGGATGGTGATATTCGTGATTTATCATTTCTTGATGCTCATCAGGCTGTCTACGTTCAG
HisHisGluGlyLysGluProLeuGluTyrArgPheTrpValThrTyrSerLeuHisSer
1321: CATCATGAGGGCAAAGAGCCTTTAGAGTATCGCTTTTGGGTTACCTACTCTCTTCACTCG
PheThrLysAspTyrGluHisGlnThrAsnGluGluLysGlnSerLeuMetTyrHisAla
1321: TTCACAAAAGATTATGAACATCAGACGAACGAAGAAAAACAATCGTTAATGTACCACGCG
ProLysGluSerArgProPheCysGlnHisArgTyrAsnLeuAlaArgThrHisLeuLys
1441: CCTAAAGAATCTCGTCCCTTCTGCCAGCACCGTTATAACTTAGCGCGCACACACTTAAAA
ArgThrIleLeuArgCysGlnLysAlaThrLeuPheMetProGlyMetValAlaMetPro
1501: AGAACTATTTTGCGCTGCCAGAAAGCAACGTTATTCATGCCGGGTATGGTAGCTATGCCC
ValIleGluValAspLeuAspGlyGlyAspLysAlaPheTyrPheValAlaPheArgAla
1561: GTGATTGAGGTGGACTTAGACGGAGGAGATAAGGCATTTTACTTTGTTGCGTTCAGGGCT
PheArgGluLysLysLysLeuArgLeuHisValThrSerAlaTyrProIleSerGluLys
1621: TTCAGGGAAAAGAAAAAACTCCGTTTGCATGTAACTAGCGCTTATCCCATTTCTGAAAAA
GlnLysGlyLysSerValLysPhePheThrIleAlaTyrAsnLeuLeuArgAsnLysGln
1681: CAGAAAGGTAAATCAGTGAAATTTTTCACCATTGCCTACAACTTATTGAGAAATAAGCAG
LeuProGlnProSerLys***
1741: CTTCCTCAGCCCTCAAAATAACAAAACCCACCTTAAGGTGGGTTTCGCCAGAGAATTATC
1801: TCTGGTATTCAGAACGCCATTACCGGACTTTGCCTTGACCTTGCGATAATCGCAGGTTGC
METSerGluPheLeuGlnSerAlaAlaSerTrpLysMETArgThrCysValLeuIle
1861: GGGATGTCTGAATTTCTTCAGTCTGCTGCATCCTGGAAGATGAGAACATGTGTTCTTATT
PheValSerIleIleValGluTyrLeuLeuSerTyrAsnGlnIleSerPheIleArgGln
1921: TTCGTCTCTATCATAGTTGAGTATTTACTCTCTTACAATCAGATCTCTTTCATTCGTCAA
GlnHisAlaSerAspPheAspThrGluPheLeuArgLysAlaGlyArgAsnGluGluGlu
1981: CAGCATGCTTCAGACTTCGATACGGAATTTTTAAGAAAGGCAGGGCGAAACGAGGAAGAA
Ala
2041: GCTT

Fig. 3 *cont'd*

were run on 8% polyacrylamide gels (0.02 cross-linked) containing 8M urea, 50 mM Tris-borate, pH 8.0, and 0.05 mM EDTA. The sequence information obtained was analyzed using the computer programs of Staden (22).

RNA/DNA hybridization and S1 mapping

RNA was purified by hot phenol extraction (23) from early logrithmic phase cells (6). The cells were grown in preconditioned (24) L-broth containing glycerol (0.4%) and, if required, ampicillin (20 μg/ml). When required, L-Arabinose (0.4%) was added to the growth medium one doubling time before the cells were harvested. RNA (100 μg) and single-stranded M13 probe DNA (10-20 μg) were mixed together in 100 μl of hybridization buffer (10 mM TrisHCl pH 8.0, 0.15 M NaCl, 10 mM $MgCl_2$, 0.1 mM EDTA) and incubated at 68°C for 30 minutes. The mixtures were adjusted to a final concentration of 30 mM NaOAc pH 4.5, 1 mM $ZnSO_4$ and incubated with nuclease S1 (3000 units) at 45°C for 30 minutes. These digestions were stopped with 300 μl of 0.3 M NaOAc and 1mM EDTA. One ml of absolute ethanol was added to precipitate the hybrids. The pellets were washed twice with 80% ethanol containing 1 mM EDTA and dried in a vacuum. Samples were run onto 5% polyacrylamide gel (0.02 cross-linked) and the hybrids were identified by staining with ethidium bromide.

RESULTS AND DISCUSSION

Determination of DNA sequence

The DNA sequence of the fragment spanning the end of *rpoC* was determined from a number of fragments cloned in the vector M13mp7. The 2044 nucleotide sequence which was obtained (Figs. 2 & 3) includes the entire sequence of the 2.0 kbp BglII fragment (actually 1961 bp) and 182 additional nucleotides leftward to the HindIII site. The left and right ends of the sequence were determined using SW1-5(+) and SW2-42(-) DNA templates. The remainder of the sequence was obtained from clones of HpaII, Sau3A, TaqI and AluI fragments derived from the 2.0 kbp BglII fragment from pGB218. These clones are designated by a letter representing the restriction endonuclease (H, S, T and A for HpaII, Sau3A, TaqI and AluI), a number representing the position of the fragment (the fragments are numbered from left to right in Fig. 2) and a (+) or (-) designating the strand orientation of the clone. Thus, the HpaII clone which hybridizes with the 3'-end of *rpoC* mRNA is H11(+). Sequences were obtained which span all of the restriction sites shown in Figure 2. It was necessary in one case to prime SW1-5(+) template with the 427 bp PstI fragment to show that HpaII fragments 11 and 12 are contiguous. Entire sequences for both the (+) and (-) orientations were obtained except for H4(-) (20 bases), a portion of H10(+) (49 bases) and 103 bases of the (+) orientation at the 5'-end of the sequence. Examination of a large number of clones in M13mp7 revealed that not all of the subfragments were cloned. Some of the missing clones [H2(+), H2(-), H7(-) & H8(-)] would contain open reading frames that are in phase with the $lacZ_\alpha$ sequence of M13mp7 and consequently might make an active β-galactosidase

α-subunit. It has been shown that fusion of some foreign genetic sequences to the begining of the intact *lacZ* gene does not destroy the β-galactosidase activity of the fusion protein (25). The possibility is being investigated that some of the missing clones produce an active β-galactosidase α-subunit and consequently make blue plaques.

Open reading frames

Calculations based on the size of the β' subunit suggest that the 2.0 kbp BglII fragment contains approximately 1000 bases of the *rpoC* sequence. The DNA sequence was examined for potential amino acid coding sequences (open reading frames). The three most extensive open reading frames were observed on the (+) strand (Figure 3). The first open reading frame of 1009 nucleotides is assigned to the distal end of *rpoC*. Significant aspects of the codon usage, secondary structure at the 3' end and more precise location of the termination point of *rpoC* are discussed below. Additional open reading frames of 537 bases (ORFa; nucleotides *1222-1758) and 180 bases (ORFb; nucleotides *1864-2043) occupy much of the remainder of the DNA sequence. While insufficient information exists to be certain that ORFb is, in fact, a gene, there are two possible ATG start codons in this region. Both have rather weak ribosome binding sites. The ATG starting at nucleotide *1900 is preceded by a GGA (*1895) while the ATG at *1864 is preceded by an AGG (*1854). Speculation about possible promoters for ORFb is gratuitous in the absence of a demonstration of the mRNA 5' end. However, the sequence TAAGGTG (*1774) which is located in a palindrome following ORFa suggests a plausable -10 sequence (26) for ORFb. ORFa, in contrast, possesses several additional features which indicate that it may be a gene. These include the greater extent of the open reading frame, a strong ribosome binding site (TAAGGA; *1210) preceding the first ATG (*1222) and a palindrome, AAACCCACCTTAAGGTGGGTTT (*1764-1785), following the translation termination codon. This secondary structure could be a terminator as it contains a short string of Ts at its 3' end. There is no obvious promoter sequence in the non-coding region between the end of *rpoC* and ORFa; however, promoter sequences are sufficiently variable (26) that in the absence of additional information we cannot exclude the possibility that this region contains a promoter. The amino acid composition of the peptide predicted from the ORFa DNA sequence has a large number of charged residues (18 Lys, 12 Arg, 8 Asp & 12 Glu) with a net charge of +10. Thus, ORFa specifies a very basic protein.

It is interesting to note that Taylor and Burgess have demonstrated that RNA polymerase binds to a 1890 bp HaeIII fragment from λ*drif*^{d}18 that spans this region (27). The left end of this HaeIII fragment is at position *821 near the end of *rpoC*. The fragment contains the regions before ORFa and ORFb as well as sequences 600-700 nucleotides beyond the HindIII site at the end of the sequence we present in this paper. The junction between *E. coli* and λ sequences occurs somewhere in the region following the HindIII site (27). Thus ORFa and ORFb are *E. coli* sequences.

3' end of the *rpl-rpo* transcription unit

The presence of two open reading frames following *rpoC* raises the question of precisely

where the *rplJL-rpoBC* transcript ends. We have shown that there is a strong terminator on the 2.0 kbp BlgII fragment (4). A more precise location of the transcription termination point within that fragment was undertaken using the S1 mapping technique (28). In the experiments reported here, single-stranded DNA probes derived from M13mp7 phage clones were also used to improve the efficiency of the hybridization reaction. Messenger RNA enriched for *rpoC* terminator transcripts was isolated from cells that carried the plasmid pGBΔR1 (6) and control mRNA was isolated from a strain (MC1000) which did not contain a plasmid. Messenger RNA from pGBΔR1 was hybridized with three different single-stranded DNA probes, all of which span the 3' end of *rpoC*. When the probe was SW1-5(+) SS DNA, an RNA/DNA hybrid approximately 1,100 bp was obtained (Fig. 4, lane 2). When mRNA was hybridized with T4(+) SS DNA, a 550 bp hybrid was observed (lane 4). T4(+) is a 640 base clone that starts at nucleotide *543 in the DNA sequence. Hybridization of pGBΔR1 mRNA with H11(+) SS DNA probe yielded a 140 bp band (lane 5). The left-hand end of the H11(+) insert corresponds to nucleotide *920. The weak 1100 bp band observed when MC1000 mRNA is probed with SW1-5(+) SS DNA (lane 3) indicates the amount of *rpoC* mRNA contributed by the chromosomal copy of the gene. Comparison of lanes 2 and 3 suggests that the plasmid, when induced with arabinose, produces about 20 times as much distal *rpoC* mRNA. The three hybrids observed in figure 4 indicate that there is only one major transcript end and that it terminates shortly after the end of *rpoC* (*1009) in the vicinity of

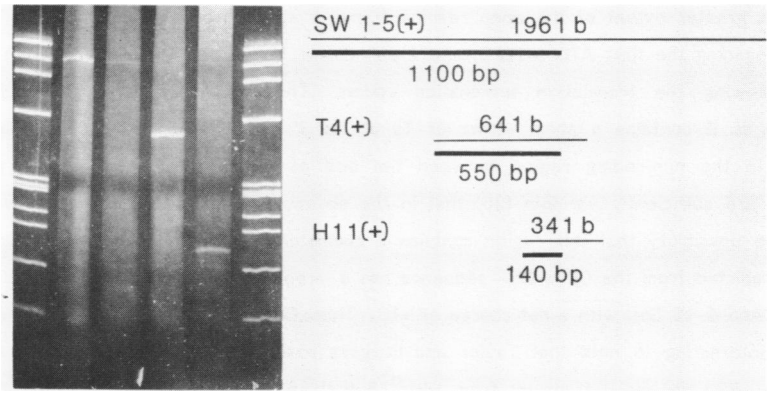

Figure 4. RNA/DNA hybrids. Lanes 1 and 6 contain ϕX174 RF DNA digested with HaeIII as a size standard. The bands are 1342, 1078, 872, 606, 310, 278, 271, 234, 194, 118 and 72 bp. Lanes 2, 4 and 5 are hybrids obtained with RNA isolated from arabinose-induced MC1000 cells containing the plasmid pGBΔR1. Lane 3 was made with RNA isolated from MC1000 cells which did not contain the plasmid. Lanes 2 and 3 used SW1-5(+) probe. Lanes 4 and 5 used T4(+) and H11(+) probes. The diagram on the right shows the extent of the *E. coli* DNA in each M13 clone (light lines) and the approximate size of the RNA/DNA hybrid detected with each probe (heavy lines). T4(+) and H11(+) are positioned in accordance with their locations within the SW1-5(+) sequence.

*1060. It should be noted that no other RNA/DNA hybrid bands can be detected in lane 2. If a major transcript continued beyond *rpoC*, a larger hybrid(s) would be present. Shorter hybrids might be formed with SW1-5(+) SS DNA if there were major transcripts for ORFa and/or ORFb. We intend to use a more sensitive method to probe for transcripts from the region following *rpoC*. ORFa may have its own promoter. Alternatively, the putative ORFa gene product may be made from the rare transcripts which survive the *rpoC* terminator. If, in fact, ORFa is expressed as part of the *rplJL-rpoBC* transcription unit and its transcript is "trickled" through the *rpoC* terminator, the level of expression of ORFa might be on the order of one percent, or less, of the *rpo* expression. These possibilities are under investigation.

Codon usage

The codons used by *rpoC*, ORFa and ORFb are compared in Table I. Data in Table I for the

Table I. Codon usage.

codons	rpoC 5'	rpoC 3'	ORFa	ORFb	codons	rpoC 5'	rpoC 3'	ORFa	ORFb
phe UUU	4	1	6	2	tyr UAU	2	0	4	1
phe UUC	6	5	6	3	tyr UAC	4	8	5	1
leu UUA	2	1	8	2	*** UAA	0	1	1	0
leu UUG	0	1	3	0	*** UAG	0	0	0	0
leu CUU	1	0	3	2	his CAU	0	1	6	1
leu CUC	1	1	1	1	his CAC	3	4	4	0
leu CUA	0	0	0	0	gln CAA	0	1	1	1
leu CUG	15	24	0	0	gln CAG	5	13	9	3
ile AUU	1	6	5	2	asn AAU	0	0	1	1
ile AUC	11	12	0	2	asn AAC	2	11	3	1
ile AUA	0	0	0	1	lys AAA	10	11	14	0
met AUG	4	4	5	2	lys AAG	5	7	4	2
val GUU	3	13	3	2	asp GAU	5	9	6	1
val GUC	1	6	1	1	asp GAC	4	10	2	1
val GUA	3	9	2	0	glu GAA	13	19	8	4
val GUG	3	5	4	0	glu GAG	6	5	4	2
ser UCU	1	6	3	5	cys UGU	1	1	0	1
ser UCC	3	7	0	1	cys UGC	4	1	2	0
ser UCA	1	2	3	1	*** UGA	0	0	0	0
ser UCG	2	4	2	0	trp UGG	2	3	2	1
pro CCU	0	1	4	0	arg CGU	9	15	4	1
pro CCC	0	0	4	0	arg CGC	4	8	3	0
pro CCA	2	3	0	0	arg CGA	0	1	1	1
pro CCG	5	12	2	0	arg CGG	0	2	0	0
thr ACU	4	4	2	0	ser AGU	0	0	0	0
thr ACC	4	11	2	0	ser AGC	0	5	1	0
thr ACA	0	0	2	1	arg AGA	0	0	2	2
thr ACG	1	0	2	1	arg AGG	0	0	2	0
ala GCU	3	6	5	3	gly GGU	5	20	3	0
ala GCC	1	2	1	0	gly GGC	5	8	1	0
ala GCA	0	13	2	2	gly GGA	0	0	3	0
ala GCG	4	13	3	0	gly GGG	1	0	0	1
					total	176	336	180	60

176 codons at the 5' end of *rpoC* are from (11,12). It has been noted by several workers that codons used by other RNA polymerase and ribosomal protein genes (11,29,30) make use of the most abundant tRNA species of *E. coli*. This has been interpreted as a means of insuring the most rapid and faithful translation of these genes. As expected, *rpoC* also uses codons recognized by major tRNA species. By contrast, ORFa and ORFb use a greater percentage of minor tRNAs. This is evident in the use of Leu and Pro as well as other codons. The major *E. coli* leucine tRNA species recognizes CUG (50-80% of the total leucine accepting activity)(29). In *rpoC*, 39/46 of the Leu codons are CUG while in ORFa + ORFb, the frequency is 0/20. CCA/G are the codons recognized by the major proline tRNA species (29). The frequencies of occurance of CCA/G are 22/23 in *rpoC* and 2/10 in ORFa. ORFb contains no proline codons. The decreased use of major tRNA species by the codons of ORFa and ORFb may suggest that these putative genes are not highly expressed. Alternatively, this observation might be used to argue that the open reading frames are not genes. Initial experiments to demonstrate ORFa and/or ORFb mRNA (see above) and peptides (Barry and Calhoun, data not shown) also suggest that these putative genes are not highly expressed if they are, in fact, expressed at all.

Secondary structure at the end of *rpoC*

The DNA sequence presented in Figure 3 does not contain much secondary structure. In addition to the palindrome following ORFa (see above), there is one other significant region of secondary structure at the 3' end of *rpoC* (Fig. 5). Because of the demonstrated relationship of this region to transcription termination the structures shown in Figure 5 are represented in the

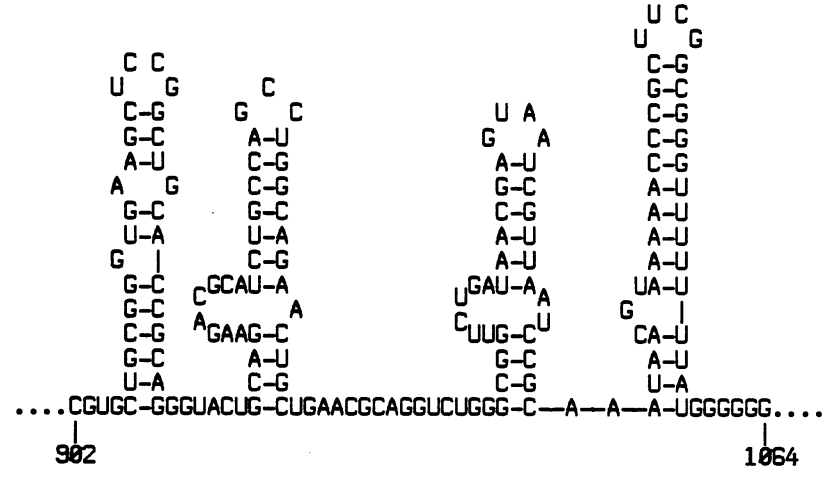

Figure 5. Proposed RNA secondary structure at the end of *rpoC*. Numbers indicate the positions as presented in the DNA sequence (Fig. 2). The nonsense codon (UAA) for translation termination of *rpoC* is in the loop of the third hairpin. The fourth hairpin structure corresponds to the predicted end of the mRNA transcript.

RNA form. The first two hairpins are within the *rpoC* gene and the third hairpin contains the translation termination codon in its loop. The fourth hairpin has a structure similar to strong rho-independent transcription terminators (26) and to ribosomal RNA terminators. Furthermore, the 3' end of this structure occurs at position *1058 in the DNA sequence. This position compares favorably with the 3' end predicted by S1 mapping with H11(+) probe (*1060) and conclusively identifies this structure as the *rpoC* terminator. In addition, this structure contains a GC rich region close to its hairpin loop and a string of oligo-U at its 3' end (actually, eight Us and one A). This string is entirely base paired with an interupted sequence in the 5' arm of the structure. Occurence of a 3' oligo-U string that can base pair with a 5' oligo-A string has been observed in several bacterial attenuators (in the *thr, his* and *phe* operons)(31,32,33) and for bacterial ribosomal RNA terminators (rrnB and rrnD)(34,35). A unique feature of the of the *rpoC* terminator is the string of six Gs immediately following the hairpin structure. *In vitro* experiments to determine the precise 3' end(s) of the transcript are in progress.

The significance of the first three hairpin structures in Figure 5 is unclear. One possibility is that the high concentration of secondary structure serves to protect the 3' end of the mRNA from exonucleolytic cleavage. However, another intriguing observation is that a variety of secondary structures (not shown) can be drawn which involve the sequence which forms the first three hairpins presented in Figure 5. It is of interest to speculate that competition between different secondary structures might somehow be involved in modulating the efficiency of termination at the end of *rpoC*. At present, there seems to be no question that the major termination event occurs immediately following *rpoC*; any modulation would be expected to involve relatively minor distal gene expression. The identity of the putative genes represented by ORFa and ORFb, as well as the mechanisms involved in their expression, are unanswered questions.

ACKNOWLEDGEMENTS

The authors wish to thank A. Giammarinaro, E. Saederup and C. Flaherty for their help with sequencing and S. Li and S. Beychok for helpful comments concerning the manuscript. Strains, phage and helpful advice were obtained from M. Casadaban, J. Messing, and S. Silverstein. Computer programs and generous assistance were provided by B. Bush, R. Staden and D. Yarmush. Computer analysis was supported by a NIH grant (RR00442) to the Columbia Biology Department Computer Graphics Facility. The remainder of the work was supported by NIH grants (GM24751 & GM25178) to CLS.

REFERENCES

1. Lindahl,L., Yamamoto,M., Nomura,M., Kirschbaum,J., Allet,B. and Rochaix, J-D. (1977) *J. Mol. Biol.* **109**, 23-47.
2. Yamamoto,M. and Nomura,M. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 3891-3895.
3. Yamamoto,M. and Nomura,M. (1979) *J. Bacteriol.* **137**, 584-594.
4. Barry,G., Squires,C.L. and Squires,C (1979) *Proc. Natl. Acad. Sci. USA* **76**, 4922-4926.
5. Post,L., Strycharz,G., Nomura,M., Lewis,H. and Dennis,P. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 1697-1701.

6. Barry,G., Squires,C. and Squires,C.L. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 3331-3335.
7. Little,R. and Dennis,P. (1979) *J. Bacteriol.* **137**, 115-123.
8. Yates,J., Arfsten,A. and Nomura,M. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 1837-1841.
9. Fukuda,R. (1980) Molec. gen. Genet. **178**, 483-486.
10. Fiil,N., Friesen,J., Downing,W. and Dennis,P. (1980) *Cell* **19**, 837-844.
11. Ovchinnikov,Y., Monastyrskaya,G., Gubanov,V., Guryev,S., Chertov,O., Modyanov,N., Grinkevich,V., Makarova,I., Marchenko,T., Polovnikova,I., Lipkin,V. and Sverdlov,E. (1981) *Eur. J. Biochem.* **116**, 621-629.
12. Delcuve,G., Downing,W., Lewis,H. and Dennis,P. (1980) *Gene* **11** 367-373.
13. Messing,J., Crea,R. and Seeburg,P. (1981) *Nucl. Acids Res.* **9**, 309-321.
14. Casadaban,M. and Cohen,S. (1980) *J. Mol. Biol.* **138**, 179-207.
15. Messing,J., Gronenborn,B., Muller-Hill,B. and Hofschneider,P. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 3642-3646.
16. Murray,N., Bruce,S. and Murray,K. (1979) *J. Mol. Biol.* **132**, 493-505.
17. Heidecker,G., Messing,J. and Gronenborn,B. (1980) *Gene* **10**, 69-73.
18. Meagher,R., Tait,R., Betlach,M. and Boyer,H. (1977) *Cell* **10**, 521-536.
19. Birnboim,H. and Doly,J. (1979) *Nuc. Acids Res.* **7**, 1513-1523.
20. Herrman,R., Neugebauer,K., Pirkl,E., Zentgraf,H. and Schaller,H. (1980) *Molec. gen. Genet.* **177**, 231-242.
21. Sanger,F., Coulson,A., Barrell,B., Smith,A. and Roe,B. (1980) *J. Mol. Biol.* **143**, 161-178.
22. Staden,R. (1979) *Nuc. Acids Res.* **6**, 2601-2610.
23. Salser,W., Gestland,R. and Bolle,A. (1967) *Nature New Biol.* **215**, 588-591.
24. Revel,H. (1965) *J. Mol. Biol.* **11**, 23-34.
25. Guarente,L., Lauer,G., Roberts,T. and Ptashne,M. (1980) *Cell* **20**, 543-553.
26. Rosenberg,M. and Court,D. (1979) *Ann. Rev. Gen.* **13**, 319-353.
27. Taylor,W. and Burgess,R. (1979) *Gene* **6**, 331-365.
28. Berk,A. and Sharp,P. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 1274-1278.
29. Post,L. and Nomura,M. (1980) *J. Biol. Chem.* **255**, 4660-4666.
30. Burton,Z., Burgess,R., Lin,L., Moore,D., Holder,S. and Gross,C. (1981) *Nuc. Acids Res.* **9**, 2889-2903.
31. Gardner,J. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 1706-1710.
32. Barnes,W. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 4281-4285.
33. Zurawski,G., Brown,K., Killingly,D. and Yanofsky,C. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 4271-4275.
34. Brosius,J., Dull,T., Sleeter,D. and Noller,H. (1981) *J. Mol. Biol.* **148**, 107-127.
35. Duester,G. and Holmes,W. (1980) *Nuc. Acids Res.* **8**, 3793-3807.