Volume 11 Number 11 1983 Nucleic Acids Research

The nucleotide sequence of the *rho* gene of *E. coli* K-12

Jennifer L.Pinkham and Terry Platt

Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06510, USA

Received 31 March 1983; Revised and Accepted 3 May 1983

ABSTRACT

We have determined the nucleotide sequence of the rho gene which encodes the E. coli K-12 transcription termination factor. The structural gene was located on a cloned 3.6 kilobase BglII-HindIII restriction fragment by the introduction of the insertion element $\gamma\delta$ and analysis of the recombinant plasmids by restriction analysis and in maxicells. The coding region consists of 1260 nucleotides directing the synthesis of a polypeptide 419 amino acids in length with a calculated molecular weight of 46,094. The deduced amino acid composition, amino-terminal protein sequence and calculated molecular weight are consistent with the data from the analysis of purified rho protein (16). We have shown that the rho genes from E. coli K-12, B and C strains are located on PvuII-HindIII fragments of the same size by hybridization to the rho (K-12) coding sequences.

INTRODUCTION

The E. coli transcription termination factor rho was first identified by Roberts (1). Rho causes termination of transcription at specific sites and catalyzes the release of the nascent RNA from the transcription complex (1, 2). Mutants of rho are known and alleles have been referred to as suA, psu, and nitA as well as rho (3-7). Rho is an essential protein, and it has been mapped to 84 minutes on the E. coli chromosome (8).

Rho has an RNA-dependent ATPase activity which appears to be required for the release of the nascent mRNA from the transcription complex (9), and the ATPase is stimulated in the absence of transcription by specific T7 mRNAs (10), the RNA encoding the rho-dependent terminator (t') at the end of the trp operon (J. Sharp and T. Platt, submitted) and synthetic RNA such as poly(C) (11, 12). Electron microscopy (13) and crosslinking studies (14) suggest that rho can exist as a hexamer. Studies of rho in solution indicate that although it dissociates readily at dilute concentrations, the hexameric form may be stabilized by the addition of poly(C) under these conditions. Since rho hexamers appear most stable under conditions of ATP hydrolysis, it has been inferred that the

active form of the enzyme *in vivo* is hexameric (14). It is not known if rho monomers exhibit ATPase activity.

Rho is a moderately abundant protein of *E. coli*, approximately 1000 hexamers per cell, as determined by immunoprecipitation (15) and by yields obtained from the rho purification procedures (16, 17). The polypeptide has an estimated molecular weight of 48,000 determined by electrophoretic mobility in SDS-polyacrylamide gels (16) and it is a basic protein with a pI between 8.0 and 9.0 (18, T. Platt, unpublished results). The amino acid sequence of the amino terminus of rho has been reported previously (19), as well as the amino acid composition, extinction coefficient and circular dichroism spectrum (16).

To aid in correlating enzymatic and physical properties of the rho polypeptide and to begin to understand the genetics and regulation of *rho* expression, we have determined the DNA sequence of the *rho* gene and its regulatory regions. The DNA sequence will permit a comparison of *rho* mutants on the nucleotide level, and the deduced amino acid sequence provides a framework for the assignment of structural and functional domains of the rho molecule. Both will be required for a detailed study of the mechanisms by which rho interacts with the RNA polymerase elongation complex and RNA during transcription termination.

METHODS

Isolation of plasmids with γδ insertions and preparation of maxicells

The plasmid containing the *rho* coding region was a generous gift of Stanley Brown. p39 is a pBR322 derivative which has a chromosomal *Bgl*II-*Hind*III (3.6 kilobases) insertion replacing the 346 base pair *Hind*III-*Bam*HI fragment. We used the protocols described in Sancar and Rupp (20) to isolate γδ insertions in the p39 plasmid with the following modifications: the donor strain was MG1063 ($F^+(\gamma\delta)$, *recA56*) transformed with p39(Ap^R) and the recipient strain was NG136 (*recA1*, *galΔS165*, str^R, F^-). Cells which had received the Ap^R determinant by transfer of a cointegrant formed with the $F^+(\gamma\delta)$ (21) were selected first in Luria broth with 200 μg/ml ampicillin and 250 μg/ml streptomycin for two hours. Then the cells were washed to remove β-lactamase and grown on MacConkey base agar (Difco) with 1% galactose, ampicillin (200 μg/ml) streptomycin (250 μg/ml) plates. White colonies which were ampicillin and streptomycin resistant were picked for plasmid analysis by mini DNA preparations (22), and recombinant plasmids containing γδ sequences within the p39 insert DNA were identified by restriction enzyme analysis.

The maxicell procedure was modified in the following manner: CSR603

(uvrA6, recA1, phr-1) cells transformed with p39::$\gamma\delta$ were grown in Luria broth with ampicillin (200 µg/ml) to a cell density of $2x10^8$, centrifuged, resuspended in minimal salts, irradiated as described, centrifuged again and resuspended in Luria broth for overnight incubation. All subsequent steps were performed as described (23).

Restriction mapping and DNA sequence analysis

Restriction enzymes were obtained from commercial suppliers (Biolabs, Boehringer-Mannheim). DNA fragments were treated with alkaline phosphatase and labeled at their 5' termini by incubation with [γ-^{32}P] ATP, >9000 Ci/mmol [synthesized by the method of Johnson and Walseth (24) as modified by I. Kennedy and O. Uhlenbeck] and polynucleotide kinase (Boehringer-Mannheim). DNA fragments were labeled at their 3' termini by incubation with 0.5 U Polymerase I large fragment (Boehringer-Mannheim), 50-100 µCi [α-^{32}P] dCTP (3000 Ci/mmol, Amersham), 10 mM Tris-HCl, pH 7.6, 10 mM $MgCl_2$, 1 mM dithiothreitol for 20 minutes at 15°C. Restriction mapping was carried out by partial digestion of end-labeled fragments (25) and the DNA sequence was determined by the methods of Maxam and Gilbert (26) and Sanger et al. (27). Computer analyses of the DNA sequence were performed using programs from Queen and Korn (28) and Staden (29).

Isolation of M13 phage DNA and dideoxy sequencing

M13 clones were constructed by digesting the 920 base pair BclI-ClaI restriction fragment internal to the rho gene with HpaII, Sau3A or RsaI and ligating each mixture of restriction fragments into the replicative form (RF) M13mp9 (30) digested with AccI, BamHI or SmaI, respectively. Purified ClaI fragments were ligated into M13mp9 RF digested with AccI. Ligated DNA was transfected into the JM101 strain and plaques were screened as described (31). White plaques were picked and single-stranded DNA was prepared in the following way: JM101 cells infected with phage from a single plaque were grown in Luria broth for 5-7 hours at 37°C. The cells were removed by filtering the cultures through polysulfone membranes (Gelman Tuffryn Membranes, 0.2µm). Phage particles in 1.2 mls of supernatant were precipitated by the addition of 250 µl 2.5 M NaCl, 20% PEG 6000, allowed to stand at room temperature for 15 minutes and centrifuged for 5 minutes at room temperature. The supernatant was removed by aspiration with a finely drawn capillary. The phage pellet was dissolved in 10 mM Tris-HCl, pH 7.6, 0.1 mM EDTA, extracted with phenol, chloroform and ether and ethanol precipitated. These volumes yield enough DNA template for 40 sequencing tracks or 10 complete sequencing experiments.

The sequencing reactions contained 0.34 pmole single-stranded DNA tem-

plate, 0.1 pmole 17 nucleotide primer (Collaborative Research), 0.6 U Polymerase I large fragment (Boehringer-Mannheim), 40 μM of three deoxynucleotide triphosphates, 0.5 μM [α-^{32}P] dATP (400 Ci/mmol, Amersham), 12-200 μM of one deoxynucleotide and 2 μM of the corresponding dideoxynucleotide triphosphate per sequencing track. The ratios of dideoxynucleotides to deoxynucleotides were varied according to the desired DNA elongation.

Genome blot

The genomic DNA from strains C117, W3110 and E. coli B was prepared from 2 ml overnight cultures. The cells were centrifuged and resuspended in 50 mM Tris-HCl, pH 7.6, 10 mM EDTA, 0.1% SDS, 1 mg/ml proteinase K (Beckman) and incubated at 50°C for 30 minutes. The DNA solution was sequentially extracted with phenol, chloroform and ether and ethanol precipitated. The DNA was digested, electrophoresed on a 1% agarose gel and transferred to nitrocellulose by the method of Southern (32). The blots were hybridized at 37° and 49°C with nick-translated (33) 920 base pair BclI-ClaI fragment in 50% formamide, 5x SSC, 50 mM $NaPO_4$ buffer pH 6.5, 1x Denhardt's solution and 0.1 mg/ml sonicated salmon sperm DNA.

RESULTS AND DISCUSSION

Identification of the rho gene

The rho gene has been localized to 3.6 kilobase BglII-HindIII restriction fragment subcloned into pBR322 from a λdrho transducing phage (19). The resulting plasmid, p39, complements rho mutants (19), and in maxicells, the production of a polypeptide which co-migrates with purified rho protein in SDS-polyacrylamide gels is observed (Figure 1). A small polypeptide of approximately 12,000-14,000 molecular weight is made from this plasmid as well. It has been observed previously by other workers (S. Brown, personal communication), but the identity of this protein is not known.

The position of the rho gene on the insert DNA of p39 was determined using F-mediated transfer of p39 by random insertion of a γδ insertion element (21, 22) (see Methods). Recombinant plasmids were analyzed by restriction patterns to determine the position of the γδ element within the BglII-HindIII insert of p39. Nine different recombinant plasmids were found, 6 in the γδ orientation, 3 in the δγ orientation. The positions of the insertions in the γδ orientation are shown in Figure 2. From the DNA sequence of γδ (34, R. Reed, personal communication) we knew that the γ arm introduces termination codons in all translational reading frames within 110 base pairs of the insertion junction. We expected that insertion of γδ within the rho gene would

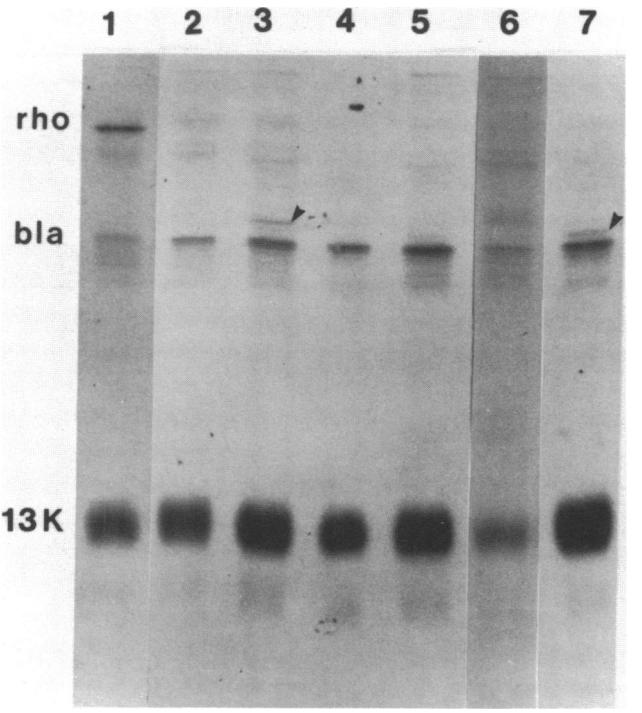

Figure 1. [^{35}S] methionine-labeled proteins produced in maxicells by p39 and its derivatives containing γδ inserts in the rho gene. Cell extracts were prepared as described (23) and electrophoresed on a 15% SDS-polyacrylamide gel. Lane 1 shows the protein products of the rho and bla (β-lactamase) genes with M_r values of 40,094 and 31,000, respectively as well as the 13,000 protein synthesized from p39. The rho polypeptide co-migrates with purified rho protein. The arrow in lane 3 marks the probable restart protein product (34,000) of p39::γδ28, and the arrow in lane 7 marks the truncated protein (32,000) made from p39::γδ36. Lanes 2, 4, 5 and 6 show the maxicell products of p36::γδ34, p39::γδ11, p39::γδ27 and p39::γδ10, respectively. No full size rho is made from any of the recombinant plasmids.

cause premature cessation of polypeptide synthesis, and provided that the shortened proteins were stable, the truncated rho proteins could be seen by maxicell analysis.

The six recombinant plasmids in the γδ orientation were transformed into CSR603 for maxicell visualization of their protein products. No recombinant plasmids made full size rho polypeptide, and truncated proteins were detectable with only 2 of these plasmids (Figure 1). We had the ambiguous result that p39::γδ36 made a peptide of 32,000 and p39::γδ28 made a peptide of 34,000 molecular weight. The paradox was resolved when the direction of rho tran-

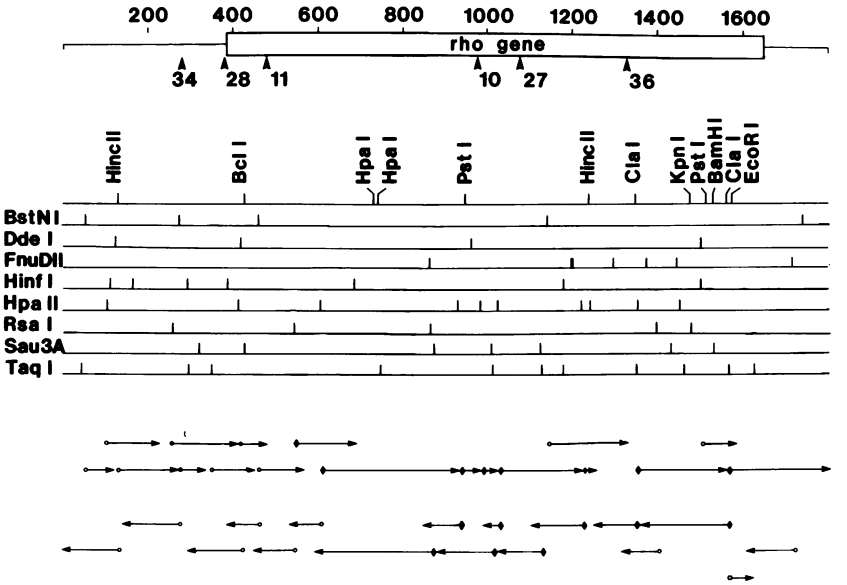

Figure 2. Restriction map and sequencing strategy for the E. coli K-12 rho gene. The upper part of the figure shows the position of the rho gene with respect to the sequenced DNA. The positions of the γδ insertions within the gene are indicated by the heavy arrows. The middle section illustrates the restriction map of the rho gene and includes restriction data for the enzymes used to generate the sequenced DNA fragments. The lower section shows the sequenced DNA fragments. Open circles indicate 5' end-labeled DNA sequenced by the Maxam and Gilbert method. The open square indicates a 3' end-labeled fragment. The closed diamonds denote DNA sequenced by the dideoxy method.

scription was known (19, S. Brown and S. Pedersen, personal communication). It was determined that p39::γδ36 makes a true truncated peptide and p39::γδ28 directs the synthesis of a probable restart protein. The recombinant plasmids p39::γδ34 and p39::γδ11 make no detectable peptides because these insertions are close to the initiating ATG. p39::γδ28 should be in this group. p39::γδ10 and p39::γδ27 should direct the synthesis of peptides 22,000 and 25,000, respectively, but these truncated products are not detectable. This observation may be due to instability of these truncated proteins, since unstable mutant rho proteins (15) as well as restart rho peptides have been reported (19).

DNA sequence of the rho gene

The restriction map and the sequencing strategy employed is presented in Figure 2. The Maxam and Gilbert (26) technique was used for the 5' region of the gene as shown, and the dideoxy sequencing method of Sanger et al. (27) was used for most of the gene. The sequence of 1800 nucleotides of the coding

strand is given along with the corresponding amino acid sequence it predicts in Figure 3. The sequenced region includes the 1260 nucleotides encoding a 419 amino acid protein, the 387 nucleotides preceding the initiating ATG codon, and the 153 nucleotides following the TAA termination codon. The GC content of the rho coding region is 50% compared to 51% for the total E. coli genome (35). The GC content is 45% for the 387 base pairs preceding the gene and 41% for the 400 base pairs following the gene.

The direction of transcription and the DNA sequence of the promoter region of rho including the first 45 nucleotides of the structural gene have been reported by Brown et al. (19). The sequence presented in this paper differs at four positions from that of Brown et al. At position 55 we report a G rather than a T consistent with a BstNI recognition site (CCTGG) from which we have sequenced. However, in E. coli the internal C residue is methylated and is resistent to modification by hydrazine. Consequently, cleavage does not occur at this residue and a blank space appears at that position in the sequencing gel. The other three differences occur in the coding region of the gene close to the BclI site. We sequenced across the BclI site from restriction sites 5' and 3' to BclI rather than from the BclI site only as Brown et al. did. We find a T rather than a C at nucleotide 420, a T instead of an A at nucleotide 423, and we assign a C at nucleotide 427 where Brown et al. report an unknown N. Our predicted amino terminal sequence agrees completely with the 16 amino acid residues reported by Brown et al. (19).

An open translational reading frame extends from the ATG beginning at nucleotide 388 to a TAA stop codon that ends at nucleotide 1648. We have judged the validity of this open reading frame by several criteria: 1) the molecular weight for rho calculated from this open reading frame is very close to that estimated by SDS-polyacrylamide gel electrophoresis, 2) the predicted amino-terminal peptide sequence is exactly the experimentally determined sequence, and 3) the amino acid composition predicted by the open reading frame correlates closely to the experimentally determined composition for rho from E. coli B. In addition, partial tryptic digestion of rho in the presence of poly(C) and ATP gives two peptide fragments: the larger one has the amino terminus of intact rho, and the smaller of these has the amino-terminal sequence Val-Leu-Thr-Gly (D. Bear and T. Platt, in preparation) which follows a Lys residue at amino acid 283. Not only is this unique in the deduced protein sequence, but the sizes of the peptide fragments determined by SDS-polyacrylamide gel electrophoresis are consistent with a cleavage at Lys 283.

```
                                        30                                                  60                                                  90
CAA AGT GGG TGC ACT GTC TAA AGG TCA GTT GAA AGA GTT CCT CGA CGC TAA CCT GGC GTA AGG GAA TTT CAT GTT CGG GTG CCC CGT CGC

                                       120                                                 150                                                 180
TAA AAA CTG GAC GCC CGG CGT GAG TCA TGC TAA CTT AGT GTT GAC TTC GTA TTA AAC ATA CCT TAT TAA GTT TGA ATC TTG TAA TTT CCA

                                       210                                                 240                                                 270
ACG CTT CCC GTT TTA TCT TAA ATG CGA AGT GAA CAG ATT TCT GGC TCG TCA CTC AAT CCG TCT TGT CGT TTC AGT TCT GCG TAC TCT CCT
                           (MET ARG SER GLU GLN ILE SER GLY SER SER LEU ASN PRO SER CYS ARG PHE SER SER ALA TYR SER PRO

                                       300                                                 330                                                 360
GTG ACC AGG CAG CGA AAA GAC ATG AGT CGA TGA CCG TAA ACA GGC ATG GAT GAT CCT GCC ATA CCA TTC ACA ACA TTA AGT TCG AGA TTT
VAL THR ARG GLN ARG LYS ASP MET SER ARG END)

                                       390                                                 420                                                 450
ACC CCA AGT TTA AGA ACT CAC ACC ACT ATG AAT CTT ACC GAA TTA AAG AAT ACG CCG GTT TCT GAG CTG ATC ACT CTC GGC GAA AAT ATG
                                    MET ASN LEU THR GLU LEU LYS ASN THR PRO VAL SER GLU LEU ILE THR LEU GLY GLU ASN MET

                                       480                                                 510                                                 540
GGG CTG GAA AAC CTG GCT CGT ATG CGT AAG CAG GAC ATT ATT TTT GCC ATC CTG AAG CAG CAC GCA AAG AGT GGC GAA GAT ATC TTT GGT
GLY LEU GLU ASN LEU ALA ARG MET ARG LYS GLN ASP ILE ILE PHE ALA ILE LEU LYS GLN HIS ALA LYS SER GLY GLU ASP ILE PHE GLY

                                       570                                                 600                                                 630
GAT GGC GTA CTG GAG ATA TTG CAG GAT GGA TTT GGT TTC CTC CGT TCC GCA GAC AGC TCC TAC CTC GCC GGT CCT GAT GAC ATC TAC GTT
ASP GLY VAL LEU GLU ILE LEU GLN ASP GLY PHE GLY PHE LEU ARG SER ALA ASP SER SER TYR LEU ALA GLY PRO ASP ASP ILE TYR VAL

                                       660                                                 690                                                 720
TCC CCT AGC CAA ATC CGC CGT TTC AAC CTC CGC ACT GGT GAT ACC ATC TCT GGT AAG ATT CGC CCG CCG AAA GAA GGT GAA CGC TAT TTT
SER PRO SER GLN ILE ARG ARG PHE ASN LEU ARG THR GLY ASP THR ILE SER GLY LYS ILE ARG PRO PRO LYS GLU GLY GLU ARG TYR PHE

                                       750                                                 780                                                 810
GCG CTG CTG AAA GTT AAC GAA GTT AAC TTC GAC AAA CCT GAA AAC GCC CGC AAC AAA ATC CTC TTT GAG AAC TTA ACC CCG CTG CAC GCA
ALA LEU LEU LYS VAL ASN GLU VAL ASN PHE ASP LYS PRO GLU ASN ALA ARG ASN LYS ILE LEU PHE GLU ASN LEU THR PRO LEU HIS ALA

                                       840                                                 870                                                 900
AAC TCT CGT CTG CGT ATG GAA CGT GGT AAC GGT TCT ACT GAA GAT TTA ACT GCT CGC GTA CTG GAT CTG GCA TCA CCT ATC GGT CGT GGT
ASN SER ARG LEU ARG MET GLU ARG GLY ASN GLY SER THR GLU ASP LEU THR ALA ARG VAL LEU ASP LEU ALA SER PRO ILE GLY ARG GLY

                                       930                                                 960                                                 990
CAG CGT GGT CTG ATT GTG GCA CCG CCG AAA GCC GGT AAA ACC ATG CTG CTG CAG AAC ATT GCT CAG AGC ATT GCT TAC AAC CAC CCG GAT
GLN ARG GLY LEU ILE VAL ALA PRO PRO LYS ALA GLY LYS THR MET LEU LEU GLN ASN ILE ALA GLN SER ILE ALA TYR ASN HIS PRO ASP
```

```
                                    1020                                    1050                                    1080
TGT GTG CTG ATG GTT CTG CTG ATC GAC GAA CGT CCG GAA GAA GTA ACC GAG ATG CAG CGT CTG GTA AAA GGT GAA GTT GTT GCT TCT ACC
CYS VAL LEU MET VAL LEU LEU ILE ASP GLU ARG PRO GLU GLU VAL THR GLU MET GLN ARG LEU VAL LYS GLY GLU VAL VAL ALA SER THR

                                    1110                                    1140                                    1170
TTT GAC GAA CCC GCA TCT CGC CAC GTT CAG GTT GCG GAA ATG GTG ATC GAG AAG GCC AAA CGC CTG GTT GAG CAC AAG AAA GAC GTT ATC
PHE ASP GLU PRO ALA SER ARG HIS VAL GLN VAL ALA GLU MET VAL ILE GLU LYS ALA LYS ARG LEU VAL GLU HIS LYS LYS ASP VAL ILE

                                    1200                                    1230                                    1260
ATT CTG CTC GAC TCC ATC ACT CGT CTG GCG CGC GCT TAC AAC ACC GTT GTT CCG GCG TCA GGT AAA GTG TTG ACC GGT GGT GTG GAT GCC
ILE LEU LEU ASP SER ILE THR ARG LEU ALA ARG ALA TYR ASN THR VAL VAL PRO ALA SER GLY LYS VAL LEU THR GLY GLY VAL ASP ALA

                                    1290                                    1320                                    1350
AAC GCC CTG CAT CGT CCG AAA CGC TTC TTT GGT GCG GCG CGT AAC GTG GAA GAG GGC GGC AGC CTG ACC ATT ATC GCG ACG GCG CTT ATC
ASN ALA LEU HIS ARG PRO LYS ARG PHE PHE GLY ALA ALA ARG ASN VAL GLU GLU GLY GLY SER LEU THR ILE ILE ALA THR ALA LEU ILE

                                    1380                                    1410                                    1440
GAT ACC GGT TCT AAA ATG GAC GAA GTT ATC TAC GAA GAG TTT AAA GGT ACA GGC AAC ATG GAA CTG CAC CTC TCT CGT AAG ATC GCT GAA
ASP THR GLY SER LYS MET ASP GLU VAL ILE TYR GLU GLU PHE LYS GLY THR GLY ASN MET GLU LEU HIS LEU SER ARG LYS ILE ALA GLU

                                    1470                                    1500                                    1530
AAA CGC GTC TTC CCG GCT ATC GAC TAC AAC CGT TCT GGT ACC CGT AAA GAA GAG CTG CTC ACG ACT CAG GAA GAA CTG CAG AAA ATG TGG
LYS ARG VAL PHE PRO ALA ILE ASP TYR ASN ARG SER GLY THR ARG LYS GLU GLU LEU LEU THR THR GLN GLU GLU LEU GLN LYS MET TRP

                                    1560                                    1590                                    1620
ATC CTG CGC AAA ATC ATT CAC CCG ATG GGC GAA ATC GAT GCA ATG GAA TTC CTC ATT AAT AAA CTG GCA ATG ACC AAG ACC AAT GAC GAT
ILE LEU ARG LYS ILE ILE HIS PRO MET GLY GLU ILE ASP ALA MET GLU PHE LEU ILE ASN LYS LEU ALA MET THR LYS THR ASN ASP ASP

                                    1650                                    1680                                    1710
TTC TTC GAA ATG ATG AAA CGC TCA TAA ATT TGT CTT ATG CCA AAA ACG CCA CGT GTT TAC GTG GCG TTT TGC TTT TAT ATC TGT AAT CTT
PHE PHE GLU MET MET LYS ARG SER END

                                    1740                                    1770                                    1800
AAT GCC GCG CTG GCG ATG TTA GGA AAA TTC CTG GAA TTT GCT GGC ATG TTA TGC AAT TTG CAT ATC AAA TGG TTA ATT TTT GCA CAG GAC
```

Figure 3. The nucleotide sequence of the E. coli K-12 rho gene. The nucleotide sequence of the coding strand of the DNA is given from 5' to 3' and is numbered above the nucleotides according to the numbering system used in Figure 2. The 5' end of the rho mRNA (19) is shown by an arrow at nucleotide 132. The open reading frame in the "leader" mRNA along with the amino acid sequence it predicts is shown in parentheses. The Shine-Dalgarno sequence for the rho ATG is underlined.

Table I Codon Usage in E. coli K-12 rho Gene

Phe	UUU	8	Ser	UCU	9	Tyr	UAU	1	Cys	UGU	1
Phe	UUC	8	Ser	UCC	4	Tyr	UAC	6	Cys	UGC	0
Leu	UUA	3	Ser	UCA	3	End	UAA	1	End	UGA	0
Leu	UUG	2	Ser	UCG	0	End	UAG	0	Trp	UGG	1
Leu	CUU	2	Pro	CCU	4	His	CAU	1	Arg	CGU	17
Leu	CUC	9	Pro	CCC	1	His	CAC	7	Arg	CGC	13
Leu	CUA	0	Pro	CCA	0	Gln	CAA	1	Arg	CGA	0
Leu	CUG	28	Pro	CCG	12	Gln	CAG	10	Arg	CGG	0
Ile	AUU	10	Thr	ACU	6	Asn	AAU	5	Ser	AGU	1
Ile	AUC	20	Thr	ACC	13	Asn	AAC	16	Ser	AGC	4
Ile	AUA	1	Thr	ACA	1	Lys	AAA	19	Arg	AGA	0
Met	AUG	16	Thr	ACG	3	Lys	AAG	9	Arg	AGG	0
Val	GUU	14	Ala	GCU	8	Asp	GAU	12	Gly	GGU	20
Val	GUC	1	Ala	GCC	7	Asp	GAC	11	Gly	GGC	7
Val	GUA	4	Ala	GCA	8	Glu	GAA	27	Gly	GGA	1
Val	GUG	6	Ala	GCG	8	Glu	GAG	9	Gly	GGG	1

Operon structure

Brown et al. (19) report that the 5' end of the rho mRNA is located 256 (nucleotide 132 in Figure 3) nucleotides from the start of the structural gene. A short translational reading frame occurs within this "leader" mRNA and extends from an ATG at position 202 to a TGA codon at 303 (Figure 3). If this open reading frame is translated, the putative peptide is 33 amino acids in length, has a calculated molecular weight of 3662 and is serine rich (9/33). This putative peptide has a weak Shine-Dalgarno (38) sequence, and it is in the same reading frame as the rho protein.

The rho ATG codon is preceded by a ribosome binding site identified by a Shine-Dalgarno sequence (underlined in Figure 3). It consists of 4 (possibly 5) nucleotides complementary to the 3' end of the 16S RNA and centered 11 nucleotides 5' to the ATG. The average ribosome binding site has a 4.8 residue complementarity centered 9.8 nucleotides 5' to the AUG codon (38, 39).

A region of GC-rich dyad symmetry followed by a series of T residues, typical of prokaryotic terminator structures has been found downstream of the TAA termination codon. The base of the stem and loop is 15 nucleotides 3' from the UAA stop codon. Functional analysis of this structure in vivo and in vitro is in progress (J. Pinkham and T. Platt, in preparation).

Codon usage

The codon usage is given for the rho gene in Table I, and it is similiar to other essential, moderately expressed genes (37, 40), especially the genes

Table II Amino Acid Composition of Rho

Amino acid	Residues, observed *E. coli* B (16)	Residues, predicted *E. coli* K-12
Lys	29	28
His	8	8
Arg	30	30
Asx	43	44
Asp		23
Asn		21
Thr	22	23
Ser	23	21
Glx	51	47
Glu		36
Gln		11
Pro	15	17
Gly	31	29
Ala	31	31
Cys	+	1
Val	26	25
Met	17	16
Ile	28	31
Leu	43	44
Tyr	7	7
Phe	15	16
Trp	2	1
total	421	419

for the beta (*rpoB*) and sigma (*rpoD*) subunits of RNA polymerase (36). In general, codons are used which are decoded by the most abundant tRNA species in *E. coli* (37). Using assignments of optimal and non-optimal codons of Ikemura (37), *rho* contains 77.2% optimal codons and 22.8% non-optimal codons. However, there appears to be a clustering of non-optimal codons in the first 40 codons where the frequency of non-optimal codon use is 47.7%.

Amino acid composition

The amino acid composition of rho from *E. coli* K-12 deduced from the DNA sequence is compared to the experimentally determined amino acid composition for rho from *E. coli* B in Table II, and they agree very well. Some features of the composition include one Cys residue at position 202, and a single Trp at residue 381. Rho is particularly poor in aromatic residues but contains a slightly higher percent of charged amino acids (29.8%) than does the "average" protein (25.1%) (41). This number represents 14.1% acidic residues and 15.7% basic residues and there is approximate agreement between calculated and ob-

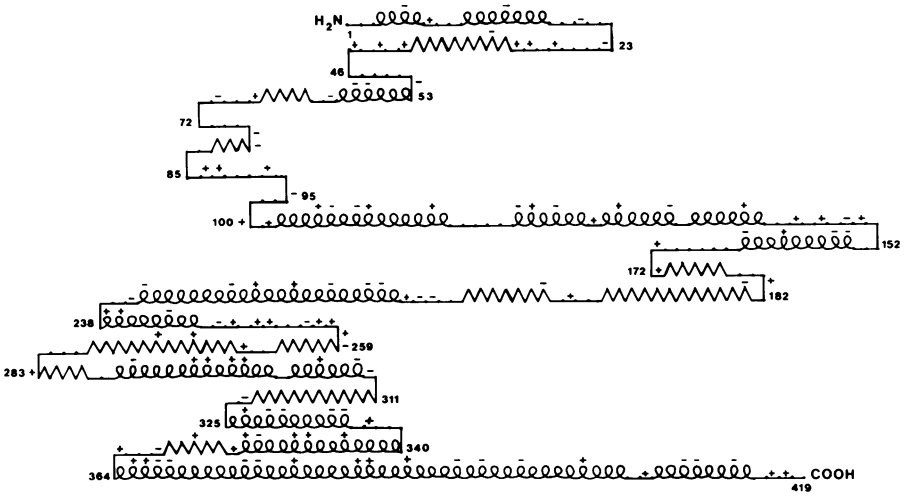

Figure 4. Secondary structure of rho protein predicted by Chou and Fasman rules. The numbers refer to the residues at the β turns, + and - refer to charged residues, α helix; β sheet; β turns; random coil.

served isoelectric points.

Secondary structure of rho

Figure 4 is a representation of the Chou and Fasman (42) prediction for the secondary structure of the rho monomer. The distribution of amino acids is without any notable clustering of hydrophobic, acidic or basic residues. This structure predicts 42% α helix, 18% β sheet and 40% random coil. Although it is not in agreement with secondary structure predictions from the circular dichroism spectrum (24% α helix, 22% β sheet and 54% random coil), Finger and Richardson suggest that discrepancies may arise from the use of inappropriate reference proteins to interpret the CD spectrum (16). The Chou and Fasman analysis predicts a rather disordered amino-terminal third and a tightly ordered structure for the remainder of the protein. The trypsin-sensitive Lys residue at 283 is located at a β turn and could conceivably be on the surface of the hexameric enzyme where it would be accessible to tryptic digestion.

Comparison of rho genes from E. coli K-12, B and C

Figure 5 shows that the structural gene for rho in K-12, B, and C strains occurs on 3.2 kilobase PvuII-HindIII fragments. Stringent hybridization conditions indicate that the B and C strains hybridized the K-12 rho sequences as efficiently as the K-12 control; thus, the rho genes from these strains are closely related. Rho proteins from K-12 and B strains co-migrate in SDS-poly-

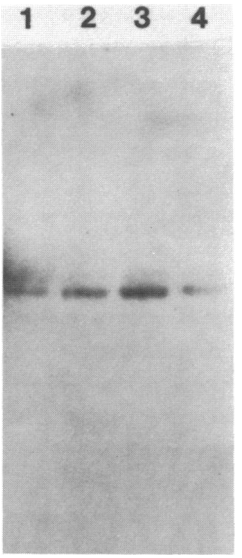

Figure 5. Genomic blot of the rho genes from E. coli K-12, B and C strains. The probe is nick-translated 920 base pair BclI-ClaI restriction fragment, and all lanes are PvuII-HindIII digests of the DNA. Lane 1, single copy reconstruction and control hybridization to rho K-12 3.2 kilobase PvuII-HindIII fragment of p39; lane 2, W3110 DNA; lane 3, E. coli B; lane 4, C117.

acrylamide gels (16) and in isoelectric focusing gels (T. Platt, unpublished results), in contrast to previous observations (18, 43). The amino acid composition of rho from E. coli B correlates well with that deduced from the DNA sequence of the K-12 rho gene. Since the rho proteins are indistinguishable by other criteria as well (D. Bear and T. Platt, in preparation), it is likely that rho is identical in B and K-12 strains.

CONCLUSIONS

Because rho$^-$ phenotypes are varied, it would not be surprising to find that the ATPase activity, RNA binding properties and the subunit interactions involved in hexamer formation can be altered to produce several different classes of rho mutants. Deletion analysis of rho mutants is now feasible and characterization of these mutants at the nucleotide level is possible. It would be interesting to know if the rho mutants isolated for readthrough of lambda terminators involve different functions of the rho protein than mutants isolated as polarity suppressors of bacterial operons, and if any known rho mutants are multiple mutations.

It has been suggested that rho is autogenously regulated because mutant *rho* strains consistently give higher yields of purified rho protein (15, 43-45). To better understand the regulation of *rho* expression, we are presently investigating *rho* transcription *in vitro* and *in vivo* and attempting to obtain overproduction by directed transcription from the the lambda P_L promoter (J. Mott *et al.*, in preparation). The gene sequence presented here will provide a solid framework for future genetic and biochemical studies of the properties of rho and the mechanism of its action.

ACKNOWLEDGEMENTS

We thank S. Brown for his generous gift of the p39 plasmid, strains and advice, S. Pedersen and D. Bear for sharing results prior to publication, H. Liebke and I. Eperon for assistance with the dideoxy sequencing and the M13 DNA isolation, R. Reed and N. Grindley for strains, A. Perlo for assistance in running the Chou and Fasman program, and J. Mott and R. Grant for valuable discussions and critical reading of the manuscript. This work was supported by a grant from the National Institutes of Health (GM 22830) to T. Platt.

REFERENCES

1. Roberts, J. (1969) Nature (London) 224, 1168-1174.
2. Rosenberg, M. and Court, D. (1979) Ann. Rev. Genet. 13, 319-353.
3. Richardson, J.P., Grimley, C. and Lowery, C. (1975) Proc. Nat. Acad. Sci. USA 72, 1725-1728.
4. Korn, L.J. and Yanofsky, C. (1976) J. Mol. Biol. 103, 395-409.
5. Inoko, H., Shigesada, K. and Imai, M. (1977) Proc. Nat. Acad. Sci. USA 74, 1162-1166.
6. Das, A., Court, D. and Adhya, S. (1976) Proc. Nat. Acad. Sci. USA 73, 1959-1963.
7. Guarente, L.P., Mitchell, D.H. and Beckwith, J. (1977) J. Mol. Biol. 112, 423-436.
8. Bachmann, B.J. and Low, K.B. (1980) Microbiol. Rev. 44, 1-56.
9. Galluppi, G., Lowery, C. and Richardson, J.P. (1976) in RNA Polymerase, R. Losick, M. Chamberlin, eds. Cold Spring Harbor Laboratory, pp. 657-665.
10. Richardson, J.P. and Macy, M.R. (1981) Biochem. 20, 113-1139.
11. Richardson, J.P. and Conaway, R. (1980) Biochem. 19, 4293-4299.
12. Galluppi, G. and Richardson J.P. (1980) J. Mol. Biol. 138, 513-539.
13. Oda, T. and Takanami, M. (1972) J. Mol. Biol. 71, 799-802.
14. Finger, L.R. and Richardson, J.P. (1982) J. Mol. Biol. 156, 203-219.
15. Imai, M. and Shigesada, K. (1978) J. Mol. Biol. 120, 451-466.
16. Finger, L.R. and Richardson, J.P. (1981) Biochem. 20, 1640-1645.
17. Sharp, J., Galloway, J.L. and Platt, T. (1983) J. Biol. Chem. 258, 3482-3486.
18. Blumenthal, R.M., Reeh, S. and Pedersen, S. (1976) Proc. Nat. Acad. Sci. USA 73, 2285-2288.
19. Brown, S., Albrechtsen, B., Pedersen, S. and Klemm, P. (1982) J. Mol. Biol. 162, 283-298.

20. Sancar, A. and Rupp, W.D. (1979) Biochem. Biophys. Res. Commun. 90, 123-129.
21. Guyer, M.S. (1978) J. Mol. Biol. 126, 347-365.
22. Holmes, D.S. and Quigley, M. (1981) Anal. Biochem. 114, 193-197.
23. Sancar, A., Wharton, R.P., Seltzer, S., Kacinski, B.M., Clarke, N.D. and Rupp, D.W. (1981) J. Mol. Biol. 148, 45-62.
24. Johnson, R.A. and Walseth, T.F. (1979) Adv. Cyclic Nucleotide Res. 10, 135-168.
25. Smith, O.H. and Birnstiel, M.L. (1976) Nuc. Acids Res. 3, 2387-2398.
26. Maxam, A. and Gilbert, W. (1980) Methods Enzymol. 65, 499-560.
27. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Nat. Acad. Sci. USA 74, 5463-5467.
28. Queen, C.L. and Korn, L.J. (1980) Methods Enzymol. 65, 595-609.
29. Staden, R. (1977) Nuc. Acids Res. 4, 4037-4051.
30. Gronenborn, B. and Messing, J. (1978) Nature (London) 272, 375-377.
31. Messing, J. and Vieira, J. (1982) Gene 19, 269-276.
32. Southern, E.M. (1975) J. Mol. Biol. 98, 503-517.
33. Rigby, P.W.J., Dieckmann, M., Rhodes, C. and Berg, P. (1977) J. Mol. Biol. 113, 237-251.
34. Reed, R.R., Young R.A., Steitz, J.A., Grindley, N.D.F. and Guyer, M.S. (1979) Proc. Nat. Acad. Sci. USA 76, 4882-4886.
35. Sober, H.A., ed. (1970) in Handbook of Biochemistry, CRC Press, Cleveland, Ohio.
36. Burton, Z., Burgess, R.R., Lin, J., Moore, D., Holder, S. and Gross, C.A. (1981) Nuc. Acids Res. 9, 2889-2903.
37. Ikemura, T. (1981) J. Mol. Biol. 151, 389-409.
38. Shine, J. and Dalgarno, L. (1974) Proc. Nat. Acad. Sci. USA 71, 1342-1346.
39. Steitz, J.A. (1979) in Ribosomes, G. Chambliss, G.R. Crowen, J. Davies, K. Davis, L. Kahan and M. Nomura, eds., University Park Press, pp. 479-495.
40. Gouy, M. and Gautier, C. (1982) Nuc. Acids Res. 10, 7055-7074.
41. Dayhoff, M.O., Hunt, L.T. and Hurst-Calderone, S. (1978) in Atlas of Protein Sequence and Structure, M. Dayhoff, ed., Vol. 5, Suppl. 3, 363-369.
42. Chou, P.Y. and Fasman, G.D. (1978) Adv. in Enzmol. 47, 45-148.
43. Ratner, D. (1976) in RNA Polymerase, R. Losick, M. Chamberlin, eds., Cold Spring Harbor Laboratory, pp. 645-655.
44. Das, A., Merril, C. and Adhya, S. (1978) Proc. Nat. Acad. Sci. USA 75, 4828-4832.
45. Richardson, J.P. and Carey, J.L. (1982) J. Biol. Chem. 257, 5767-5771.