Volume 9 Number 12 1981 Nucleic Acids Research

The nucleotide sequence of the cloned rpoD gene for the RNA polymerase sigma subunit from E. coli K12

Zachary Burton, Richard R. Burgess, Judy Lin[1], David Moore*[1], Sarah Holder[1] and Carol A. Gross

McArdle Laboratory for Cancer Research, and *Department of Genetics, University of Wisconsin, Madison, WI 53706, USA

Received 15 April 1981

ABSTRACT

We have determined the nucleotide sequence of the rpoD gene which codes for the sigma subunit of RNA polymerase from E. coli K12. The gene, which we formerly cloned as a HindIII restriction fragment in the transducing phage, Charon 25, was recloned into several plasmids. We have determined a 2600 base pair DNA sequence which includes the entire structural gene for sigma. The resulting amino acid sequence agrees with previous information obtained about sigma including the amino acid composition, partial sequence data for the N-terminus, the highly acidic nature of the polypeptide, and the cleavage pattern at cysteines. The molecular weight of 70,263 daltons calculated for the 613 amino acid polypeptide is significantly lower than had been determined previously by SDS polyacrylamide gel analysis.

INTRODUCTION

The sigma subunit of E. coli RNA polymerase has been shown to play an important role both in selective binding of polymerase to promoters and in the efficient initiation of transcription (2-3). The rpoD gene, coding for sigma, has been mapped to about 66 minutes on the E. coli genetic map (4-6). Several mutants affecting sigma have been isolated (6-11), and for some of them, alterations in the sigma polypeptide have been observed (12). Utilizing a temperature sensitive sigma mutant, rpoD800, we were able to isolate a transducing phage carrying the sigma gene rpoD (13). Other transducing phages carrying the Salmonella typhimurium and E. coli sigma genes have also been isolated (14,15).

Although physical and chemical studies of sigma have been hampered by the difficulty of obtaining sufficient quantities of pure material, some properties have been determined. These include amino acid composition (16-19), N-terminal amino acid sequence (17,19), isoelectric point (19,20), molecular weight (19,21), α-helical content (20), thermal inactivation behavior (22), and molecular-dimension estimated by small-angle X-ray

studies (23).

It is known from crosslinking studies that sigma interacts with several subunits of the core polymerase (24,25), with DNA in the promoter region (26,27), with short nascent RNA (28), and perhaps weakly with rifampicin bound to the beta (β) subunit (29). In order to obtain a more detailed knowledge of the structure of the sigma polypeptide, we have determined the DNA sequence of the sigma gene and deduced its amino acid sequence. This sequence information will aid studies to locate the structural and functional domains of sigma. This will allow us to better determine the mechanism by which sigma functions in the regulation of transcription initiation.

METHODS

1) Subcloning *rpoD* gene into plasmids

a. *Preparation of pRRB1*. Ch25sig-39H DNA (13) and pBR322 DNA (30) were digested with restriction endonuclease HindIII, adjusted to 2 μg/ml, mixed together and ligated overnight at 12°C. The ligated mixture was used to transform CAG384 (an *E. coli* K12 C600 derivative which cannot grow at 42° because it contains the *ts* sigma allele *rpoD800*) to ampicillin^R (amp^R) ts$^+$. Such transformants should contain both the plasmid pBR322 which confers amp^R and the 9.2 kb HindIII piece from Ch25sig-39H (see Fig. 1A) which codes for the wild type sigma gene. Putative transformants were picked and grown to stationary phase in LB broth supplemented with 25 μg/ml ampicillin. DNA was prepared for restriction enzyme analysis by the method of Birnboim (31). Restriction enzyme analysis using HindIII, BamHI and AvaI confirmed that the recombinant plasmid, termed pRRB1, contained pBR322 and the 9.2 kb HindIII piece from Ch25sig-39H.

b. *Preparation of pRRB2*. The largest fragment from a HaeIII restriction endonuclease digest of pRRB1 is a 2 kb piece which includes 1000 bases of the sigma gene itself, about 1000 bases preceding the sigma gene, and the sigma promoter (W. Taylor, Z. Burton, R. Burgess, C. Gross, unpublished results). This fragment was separated from other HaeIII fragments on a 6% polyacrylamide gel, eluted (Maxam & Gilbert, 32) and purified on a DEAE-cellulose column. The resulting fragment was ligated into the SmaI site of pK03, a plasmid which contains the galactokinase gene without a promoter. pK03 is essentially the same as pK01 [described in great detail by McKinney *et al.* (33)] except the former has two additional bases, C and G, at positions 10 and 11 downstream

from the SmaI site of pK01. Ligated plasmid was used to transform C600 galK$^-$ to amp^R galK$^+$. Transformants were picked and restriction enzyme analysis of the plasmid DNA with HaeIII, HindIII, SacI and PvuII confirmed that the 2 kb HaeIII piece was inserted at the SmaI site of pK03 in the proper orientation for transcription to start at the sigma promoter and read through the sigma fragment into the galactose kinase gene.

c. Preparation of pRRB3. The 3 kb PvuII fragment, containing most of the sigma gene and no promoter (see Fig. 1), was cloned into the SmaI site of pK03 and is termed pRRB3. To construct this plasmid, pRRB1 was digested with PvuII, mixed with pK03 digested with SmaI, and ligated overnight at 12^o. Cells were transformed with the plasmid mixture and selected for amp^R. Thirty independent transformants were grown up in LB both supplemented with 25 µg/ml ampicillin. DNA was prepared (31) and subjected to restriction enzyme analysis to find a plasmid in which the 3 kb PvuII piece was inserted.

2) Preparation of labeled fragments and DNA sequencing

Preparation of labeled DNA fragments was by the methods described by Maxam and Gilbert (32) except that calf intestine alkaline phosphatase (Boehringer-Mannheim) was substituted for bacterial alkaline phosphatase in preparing DNA fragments for 5' end labeling. Some fragments were labeled at their 3' end using the Klenow fragment of DNA polymerase I (Boehringer-Mannheim). For 3' labeling, DNA fragments were incubated in 10 mM Tris, pH 7.4, 6.6 mM $MgCl_2$, 1 mM DTT, 50 mM NaCl, 50-100 µCi of [α-^{32}P]deoxyribonucleoside triphosphate (400 Ci/mmol.) and 2.0 units of Klenow fragment for 2 hr at 25^o (adapted from Maniatis et al. (34)). All fragments for sequencing were prepared from pRRB2 or pRRB3. DNA sequencing of fragments was by the method of Maxam and Gilbert (32).

RESULTS AND DISCUSSION

1) DNA sequence of the rpoD gene

We have previously reported the isolation of two sigma transducing phage carrying overlapping fragments of E. coli DNA. In order to more easily prepare DNA to be used for sequence analysis, the region of the DNA containing sigma was subcloned from Ch25sig-39H into plasmids pRRB1, pRRB2, and pRRB3 as described in Methods. The sequencing strategy employed is presented in Fig. 1. The resulting sequence of 2600 nucleotides of the coding strand is given in Fig. 2 along with the corresponding amino acid sequence which it predicts. The sequenced region includes the 1839

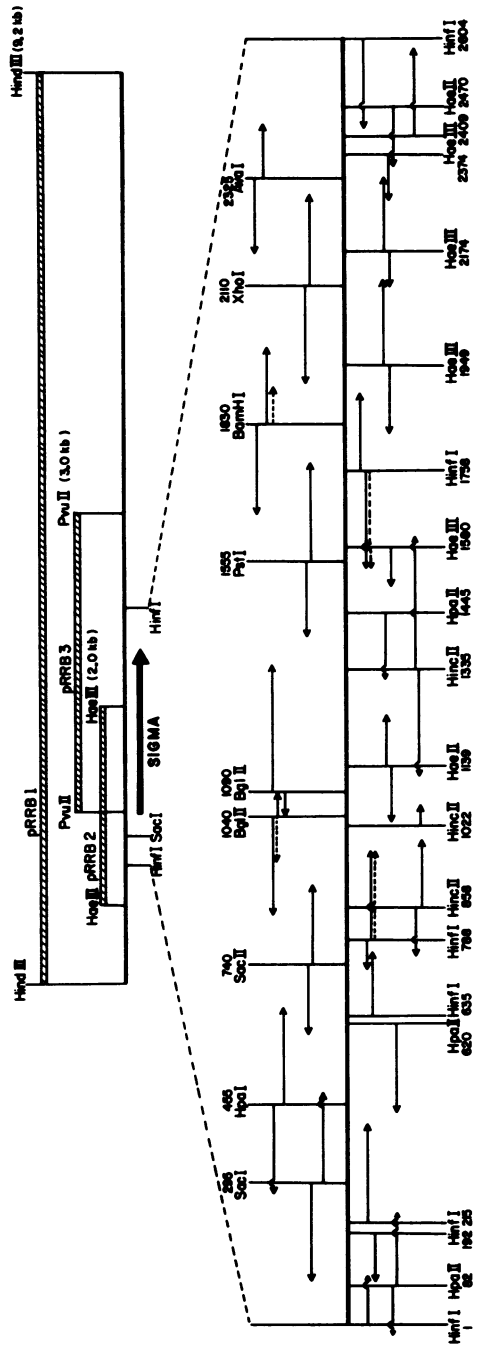

FIGURE 1. Sequencing strategy for the E. coli K12 rpoD gene. The upper part of the figure shows the relative position of the E. coli DNA cloned into the three plasmids used in the sequencing, with an indication of where within this region the rpoD gene is located. In the expanded region in the lower part of the figure we indicate by vertical lines restriction sites used in sequencing. The infrequent restriction sites are above the heavy horizontal line, the more frequent sites below the line. Nucleotide base positions are numbered in the direction from 5' to 3' in the coding strand, starting with number one at the left hand HinfI site. The extent and direction of each fragment sequenced is shown by horizontal arrows. Dotted arrows indicate fragments which were labeled at the 3' end. All regions were sequenced on both DNA strands.

nucleotides coding for a 613 amino acid protein, the 524 nucleotides preceding the initiation codon AUG, and the 234 nucleotides following the termination codon UAA.

Our previous studies (13) had identified the direction of sigma transcription and had located the beginning of the sigma polypeptide to within several hundred bases of the SacI restriction site (see Fig. 1). When the sequence of the coding strand in this region was examined we found an initiating methionine codon AUG followed by a continuous reading frame for 1836 base pairs. This AUG codon (doubly underlined below) is preceded by a ribosome binding site identified by a Shine-Dalgarno sequence (underlined below) located the expected distance from the AUG (35). When the predicted N-terminal amino acid sequence was compared to the published N-terminal amino acid sequence of Lowe et al. (23), the agreement was perfect.

DNA sequence: GTGTGGATTACCGTCTTATGGAGCAAAAGCCGCAGTCACAGCTGAAACTTCTT...

Protein sequence: MetGluGlnAsnProGlnSerGlnLeuLysLeuLeu...

Lowe et al., 1979: MetGlxGlxAsxProGlx$^{Ser}_{Cys}$GlxLeuLysLeuLeu...

2) Amino acid composition

The amino acid composition of sigma deduced from DNA sequence studies is shown in Table I (columns 3 and 4) along with previous composition determinations based on amino acid analysis (17,19) (columns 1 and 2). The values for Cys and Met based on DNA sequence determination are in very good agreement with the determinations made previously by a different technique (16). There is reasonable agreement with our previous amino acid analysis (19), except that Trp and Thr were significantly lower and Arg and Tyr significantly higher than expected.

Also shown for comparison in Table I are the amino acid compositions of the alpha (36) and beta (37) subunits of RNA polymerase deduced from published sequence data (columns 5 and 6), and an "average" protein composition based on 314 sequenced proteins (38) (column 7). When the composition of sigma is compared with the other proteins, several features are clear. First, sigma has a very high content of charged amino acid residues (34.9%) compared to an "average" protein (25.1%). Sigma has 20.4% acidic residues compared to 11.6% for an average protein and 16-17% for beta and alpha which are also acidic proteins. Sigma is also quite low in Pro and Gly and high in Met. All of the sequenced RNA polymerase subunits are relatively high in Arg and Ile and low in

```
ATCTGGATGAAAATAAGCTCCTTGGACTTGGCTTATTCAGAGAACTGGTCAACACTTGTCTCTCCCAGCAGGTCTGACCACCGGGCAACTTTTAGAGCAC     1-100

TATCGTGGTACAAATAATGCTGCCACCCTTGAAAAACTGTCGATGTGGGACGATATAGCAGATAAGAATATTGCTGAGCAAACCTTCACCGACTCACTCA    101-200

ACCATATGTTTGATTCGCTGCTTGAACTGCGCCAGGAAGAGTTAATCGCTCGTGAGCGCACGCATGGTTTAAGCAACGAAGAACGCTGGAGCTCTGGACA    201-300

TTAAACCAGGAGCTGGCGAAAAAGTGATTTAACGGCTTAAGTGCCGAAGAGCGATCGGGAAGCCCCCGACAGCCGCACTGAGAGGCAGCGGCAAATATAT    301-400

AAGTACGCCCTCGTAATTATCGTTGGCGGTAAACAACCGTTGGATTTCAGCGTTAACGGCTGAAGGACATCGGGTCAATCGCCCAACACCAACCTCATGA    401-500

AATAAGTGTGTGGATTACCGTCTTATGGAGCAAAACCCGCAGTCACAGCTGAAACTTCTTGTCACCCGTGGTAAGGAGCAAGGCTATCTGACCTATGCCG    501-600
                                MetGluGlnAsnProGlnSerGlnLeuLysLeuLeuValThrArgGlyLysGluGlnGlyTyrLeuThrTyrAlaG      1-25

AGGTCAATGACCATCTGCCGGAAGATATCGTCGATTCAGATCAGATCGAAGACATCATCCAAATGATCAACGACATGGGCATTCAGGTGATGGAAGAAGC    601-700
luValAsnAspHisLeuProGluAspIleValAspSerAspGlnIleGluAspIleIleGlnMetIleAsnAspMetGlyIleGlnValMetGluGluAl     26-59

ACCGGATGCCGATGATCTGATGCTGGCTGAAAACACCGCGGACGAAGATGCTGCCGAAGCCGCCGCGCAGGTGCTTTCCAGCGTGGAATCTGAAATCGGG    701-800
aProAspAlaAspAspLeuMetLeuAlaGluAsnThrAlaAspGluAspAlaAlaGluAlaAlaAlaGlnValLeuSerSerValGluSerGluIleGly     60-92

CGCACGACTGACCCGGTACGCATGTACATGCGTGAAATGGGCACCGTTGAACTGTTGACCCGCGAAGGCGAAATTGACATCGCTAAGCGTATTGAAGACG    801-900
ArgThrThrAspProValArgMetTyrMetArgGluMetGlyThrValGluLeuLeuThrArgGluGlyGluIleAspIleAlaLysArgIleGluAspG     93-125

GGATCAACCAGGTTCAATGCTCCGTTGCTGAATATCCGGAAGCGATCACCTATCTGCTGGAACAGTACAATCGTGTTGAAGCAGAAGAAGCGCGTCTGTC    901-1000
lyIleAsnGlnValGlnCysSerValAlaGluTyrProGluAlaIleThrTyrLeuLeuGluGlnTyrAsnArgValGluAlaGluGluAlaArgLeuSe    126-159

CGATCTGATCACCGGCTTTGTTGACCCGAACGCAGAAGAAGATCTGGCACCTACCGCCACTCACGTCGGTTCTGAGCTTTCCCAGGAAGATCTGGACGAT    1001-1100
rAspLeuIleThrGlyPheValAspProAsnAlaGluGluAspLeuAlaProThrAlaThrHisValGlySerGluLeuSerGlnGluAspLeuAspAsp    160-192

GACGAAGATGAAGACGAAGAAGATGGCGATGACGACAGCGCCGATGATGACAACAGCATCGACCCGGAACTGGCTCGCGAAAAATTTGCGGAACTACGCG    1101-1200
AspGluAspGluAspGluGluAspGlyAspAspAspSerAlaAspAspAspAsnSerIleAspProGluLeuAlaArgGluLysPheAlaGluLeuArgA    193-225

CTCAGTACGTTGTAACGCGTGACACCATCAAAGCGAAAGGTCGCAGTCACGCTACCGCTCAGGAAGAGATCCTGAAACTGTCTGAAGTATTCAAACAGTT    1201-1300
laGlnTyrValValThrArgAspThrIleLysAlaLysGlyArgSerHisAlaThrAlaGlnGluGluIleLeuLysLeuSerGluValPheLysGlnPh    226-259

CCGCCTGGTGCCGAAGCAGTTTGACTACCTGGTCAACAGCATGCGCGTCATGATGGACCGCGTTCGTACGCAAGAACGTCTGATCATGAAGCTCTGCGTT    1301-1400
eArgLeuValProLysGlnPheAspTyrLeuValAsnSerMetArgValMetMetAspArgValArgThrGlnGluArgLeuIleMetLysLeuCysVal    260-292

GAGCAGTGCAAAATGCCGAAGAAAAACTTCATTACCCTGTTTACCGGCAACGAAACCAGCGATACCTGGTTCAACGCGGCAATTGCGATGAACAAGCCGT    1401-1500
GluGlnCysLysMetProLysLysAsnPheIleThrLeuPheThrGlyAsnGluThrSerAspThrTrpPheAsnAlaAlaIleAlaMetAsnLysProT    293-325
```

```
GGTCGGAAAAACTGCACGATGTCTCTGAAGAAGTGCATCGCGCCCTGCAAAAACTGCAGCAGATTGAAGAAGAAACCGGCCTGACCATCGAGCAGGTTAA  1501-1600
rpSerGluLysLeuHisGluValSerGluGluValHisArgAlaLeuGlnLysLeuGlnGlnIleGluGluGluThrGlyLeuThrIleGluGlnValLy   326-359

AGATATCAACCGTCGTATGTCCATCGGTGAAGCGAAAGCCCGCCGTGCGAAGAAAGAGATGGTTGAAGCGAACTTACGTCTGGTTATTTCTATCGCTAAG  1601-1700
sAspIleAsnArgArgMetSerIleGlyGluAlaLysAlaArgArgAlaLysLysGluMetValGluAlaAsnLeuArgLeuValIleSerIleAlaLys   360-392

AAATACACCAACCGTGGCTTGCAGTTCCTTGACCTGATTCAGGAAGGCAACATCGGTCTGATGAAAGCGGTTGATAAATTCGAATACCGCCGTGGTTACA  1701-1800
LysTyrThrAsnArgGlyLeuGlnPheLeuAspLeuIleGlnGluGlyAsnIleGlyLeuMetLysAlaValAspLysPheGluTyrArgArgGlyTyrL   393-425

AGTTCTCCACCTACGCAACCTGGTGGATCCGTCAGGCGATCACCCGCTCTATCGCGGATCAGGCGCGCACCATCCGTATTCCGGTGCATATGATTGAGAC  1801-1900
ysPheSerThrTyrAlaThrTrpTrpIleArgGlnAlaIleThrArgSerIleAlaAspGlnAlaArgThrIleArgIleProValHisMetIleGluTh   426-459

CATCAACAAGCTCAACCGTATTTCTCGCCAGATGCTGCAAGAGATGGGCCGTGAACCGACGCCGGAAGAACTGGCTGAACGTATGCTGATGCCGGAAGAC  1001-2000
rIleAsnLysLeuAsnArgIleSerArgGlnMetLeuGlnGluMetGlyArgGluProThrProGluGluLeuAlaGluArgMetLeuMetProGluAsp   460-492

AAGATCCGCAAAGTGCTGAAGATCGCCAAAGAGCCAATCTCCATGGAAACGCCGATCGGTGATGATGAAGATTCGCATCTGGGGGGATTTCATCGAGGATA  2001-2100
LysIleArgLysValLeuLysIleAlaLysGluProIleSerMetGluThrProIleGlyAspAspGluAspSerHisLeuGlyAspPheIleGluAspT   493-525

CCACCCTCGAGCTGCCGCTGGATTCTGCGACCACCGAAAGCCTGCGTGCGGCAACGCACGACGTGCTGGCTGGCCTGACCGCGCGTGAAGCAAAAGTTCT  2101-2200
hrThrLeuGluLeuProLeuAspSerAlaThrThrGluSerLeuArgAlaAlaThrHisAspValLeuAlaGlyLeuThrAlaArgGluAlaLysValLe   526-559

GCGTATGCGTTTCGGTATCGATATGAACACCGACTACACGCTGGAAGAAGTGGGTAAACAGTTCGACGTTACCCGCGAACGTATCCGTCAGATCGAAGCG  2201-2300
uArgMetArgPheGlyIleAspMetAsnThrAspTyrThrLeuGluGluValGlyLysGlnPheAspValThrArgGluArgIleArgGlnIleGluAla   560-592

AAGGCGCTGCGCAAACTGCGTCACCCGAGCCGTTCTGAAGTGCTGCGTAGCTTCCTGGACGATTAATCGGTAGGCCGGATCAGGCGTTACGCCGCACCCG  2301-2400
LysAlaLeuArgLysLeuArgHisProSerArgSerGluValLeuArgSerPheLeuAspAspTer                                        593-613

GCACTAGGCCCTCTGCACAAACGCCACCTTTTCGGTGGCGTTTTTTATCGCCCACGCACTACCAGCGCCTGGTCCAGCTCGCGATACGCTTCAACCAGTT  2401-2500

TCTCCAGTGAAACGCGACTTAAACCGCTGGGGTTTGGCAGCACCCAAATCTGCGTCGAACCAATGGTGAGCGTTTGTTTCCCCCACTGTGCACCGCGCTG  2501-2600
```

FIGURE 2. The nucleotide sequence of the E. coli K12 rpoD gene. The nucleotide sequence of the coding strand of the DNA is given from 5' to 3' and is numbered at the right with the numbering system used in Fig. 1. The predicted complete amino acid sequence is shown below the DNA sequence. Amino acids are numbered at the right starting with the N-terminal methionine as number one; every 10th amino acid is indicated by a dot below it.

Table I Amino Acid Composition of E. coli K12 Sigma Compared to Other Proteins

		1.	2.	3.	4.	5.	6.	7.
		Amino acid analysis		From sequence data				
		rpoD (sigma)		rpoD (sigma)		rpoA (alpha)	rpoB (beta)	Average protein
		(ref 19)	(ref 17)	(this paper)		(ref 36)	(ref 37)	(ref 38)
Amino	Acid	mole%	mole%	residues	mole%	mole%	mole%	mole%
Ala	A	7.9	5.3	49	8.0	7.0	5.9	8.6
Asx	B(D+N)	11.5	13.2	(73)	(11.9)	(9.1)	(10.6)	(9.8)
Cys	C	0.6	0.9	3	0.5	1.2	0.5	2.9
Asp	D	-	-	54	8.8	6.4	6.8	5.5
Glu	E	-	-	71	11.6	10.9	9.1	6.0
Phe	F	2.5	2.5	15	2.45	1.2	3.3	3.6
Gly	G	4.2	4.5	24	3.9	6.1	7.9	8.4
His	H	1.7	2.0	9	1.5	2.4	1.4	2.0
Ile	I	6.9	6.3	43	7.0	7.3	6.3	4.5
Lys	K	6.0	5.6	34	5.55	4.9	6.0	6.6
Leu	L	9.0	8.0	54	8.8	11.5	9.5	7.4
Met	M	4.0	4.9	25	4.1	1.5	2.8	1.7
Asn	N	-	-	19	3.1	2.7	3.8	4.3
Pro	P	3.1	3.5	19	3.1	4.9	4.2	5.2
Gln	Q	-	-	30	4.9	3.0	4.3	3.9
Arg	R	6.5	6.5	46	7.5	7.0	6.7	4.9
Ser	S	5.1	4.3	29	4.7	5.2	5.5	7.0
Thr	T	7.2	6.3	38	6.2	5.8	4.5	6.1
Val	V	5.6	5.2	34	5.55	9.1	8.2	6.6
Trp	W	1.1	0.7	4	0.65	0.3	0.3	1.3
Tyr	Y	1.8	2.1	13	2.1	1.5	3.2	3.4
Glx	Z(E+Q)	15.3	18.4	(101)	(16.5)	(13.9)	(13.4)	(9.9)
Small aliphatic		(A+G)			11.9	13.1	13.8	16.9
hydroxyl		(S+T)			10.9	11.0	10.0	13.1
acidic		(D+E)			20.4	17.3	15.9	11.6
acidic+acid amide		(D+B+N+E+Z+Q)			28.4	23.0	24.0	19.8
basic		(K+R+H)			14.5	14.3	14.1	13.5
hydrophobic		(L+V+I+M)			25.4	29.4	26.8	20.2
aromatic		(F+Y+W)			5.2	3.0	6.8	8.3
charged		(D+E+K+R+H)			34.9	31.6	30.0	25.1
residues					613	329	1342	
unmodified molecular weight (daltons)					70,263	36,511	150,543	
mean residue molecular weight					114.6	111.0	112.2	
calculated ϵ_M ($M^{-1}cm^{-1}$)					39,040	12,000	77,400	
calculated $E^{1\%}_{280nm}$					5.6	3.3	5.2	

Trp and Cys. Groupings of similar amino acids (see ref. 38) are given at the bottom of Table I to aid in these comparisons.

Based on the molecular weight of 70,263 daltons, the content of four Trp and thirteen Tyr, and molar extinction coefficients at 280 nm

of 5600 and 1280 for Trp and Tyr, respectively (39), one can estimate a molar extinction coefficient for sigma of 39,040 M^{-1} cm^{-1} which gives an $E^{1\%}_{280nm}$ of approximately 5.6. This is much lower than the values of 8.4 (19) and 11.6 (20) previously reported. The most likely reason for discrepancy is that it is difficult to prepare sigma in sufficient quantity and purity to do the extinction coefficient measurements accurately.

3) Codon usage

The frequency of usage of the various codons in E. coli is given in Table II for the rpoD and rpoB (37) genes of RNA polymerase, the N-terminal 159 amino acids of the rpoA gene (40), the sum of several ribosomal protein genes (41), the sum of the tufA (42) and tufB (43) genes, and the sum of the trpA and trpB genes (44). The codon usage in the sigma gene is highly nonrandom and similar in many cases to the patterns observed with many other E. coli proteins. This pattern reflects the abundance of the various tRNAs in E. coli (45). For sigma 17/19 Pro codons are CCA, 42/54 Leu codons are CUG, 18/19 Asn codons are AAC, 46/47 Arg codons are CGU or CGC, and 15/61 amino acid codons are used only once or not at all. The Thr codon ACU and the Val codon GUA are used much less frequently and the Asp codon GAU more frequently than with the other E. coli proteins listed. The GC content of the sigma coding region is 53% compared to the overall GC content of E. coli of 51% (39). The GC content was 48.5% for the 524 base pairs preceding the gene and was 59.4% for the 234 base pairs following the gene. G or C is found in the third position of the 32 quartet codons 66.5% of the time and of the 61 codons 61.5% of the time. Grantham has classified the codon usage of 119 genes and concluded that highly expressed bacterial mRNAs have lower GC contents in the third position of the 32 quartet codons than do most bacterial genes (46). The rpoB (54.1%), r-proteins (42.5%), and tufA+B (52.8%) are highly expressed and clearly have lower GC contents than rpoD (66.5%) or trpA+B (67.1%) which are less highly expressed.

4) Amino acid distribution

The distribution of charged amino acid residues in the sigma polypeptide is shown in Fig. 3. There are several small basic regions with no intervening acidic residues at regions 371-377, 493-502 and 593-603. However, the major contribution to charge comes from a preponderance of acidic residues clustered at the N-terminus. The first 215 residues (about one third of sigma) carry a net charge of -50 whereas the entire

Table II Codon Usage in *E. coli* K12 *rpoD* Gene Compared with Other *E. coli* Proteins

Amino Acid	Codon	*rpoD* (sigma)	*rpoA* (alpha)	*rpoB* (beta)	ribosomal proteins	*tufA+B*	*trpA+B*	Amino Acid	Codon	*rpoD* (sigma)	*rpoA* (alpha)	*rpoB* (beta)	ribosomal proteins	*tufA+B*	*trpA+B*
Phe	UUU	4	2	11	9	2	14	Tyr	UAU	4	1	14	4	3	11
Phe	UUC	11	0	33	17	26	11	Tyr	UAC	9	1	29	13	17	8
Leu	UUA	1	1	1	4	0	4	Ter	UAA	1	0	1	7	2	1
Leu	UUG	2	1	6	3	0	11	Ter	UAG	0	0	0	0	0	0
Leu	CUU	5	2	6	4	2	4	His	CAU	4	2	1	4	3	8
Leu	CUC	3	2	17	3	1	4	His	CAC	5	4	18	9	19	10
Leu	CUA	1	2	0	0	0	3	Gln	CAA	7	1	8	8	0	8
Leu	CUG	42	9	97	67	53	40	Gln	CAG	23	5	50	26	16	20
Ile	AUU	11	6	18	16	6	22	Asn	AAU	1	1	3	4	0	6
Ile	AUC	32	7	66	37	52	20	Asn	AAC	18	3	48	32	14	17
Ile	AUA	0	0	0	1	0	1	Lys	AAA	21	6	56	77	35	27
Met	AUG	25	3	37	24	20	20	Lys	AAG	13	2	24	28	11	4
Val	GUU	14	6	41	47	46	12	Asp	GAU	30	7	30	14	4	21
Val	GUC	7	2	15	8	1	7	Asp	GAC	24	1	62	31	41	9
Val	GUA	3	1	31	42	21	2	Glu	GAA	58	5	89	58	60	32
Val	GUG	10	8	22	17	6	15	Glu	GAG	13	8	32	15	13	15
Ser	UCU	9	4	23	24	14	6	Cys	UGU	0	0	5	1	2	3
Ser	UCC	7	0	32	21	6	10	Cys	UGC	3	2	2	6	4	5
Ser	UCA	2	1	0	1	0	2	Ter	UGA	0	0	0	1	0	1
Ser	UCG	2	2	2	2	0	5	Trp	UGG	4	0	4	4	2	1
Pro	CCU	1	2	8	4	0	7	Arg	CGU	28	4	60	46	41	7
Pro	CCC	0	0	0	1	0	5	Arg	CGC	18	3	28	24	5	17
Pro	CCA	1	0	10	6	2	9	Arg	CGA	0	1	1	0	0	2
Pro	CCG	17	4	37	32	38	16	Arg	CGG	0	0	0	1	0	2
Thr	ACU	2	2	17	31	25	5	Ser	AGU	1	1	2	4	0	3
Thr	ACC	29	7	34	21	31	16	Ser	AGC	8	2	15	7	1	3
Thr	ACA	0	1	3	3	3	3	Arg	AGA	0	1	0	1	0	1
Thr	ACG	7	1	6	2	1	8	Arg	AGG	0	0	0	0	0	0
Ala	GCU	11	1	18	75	24	9	Gly	GGU	9	7	69	44	38	26
Ala	GCC	10	2	9	13	2	30	Gly	GGC	12	6	36	35	41	26
Ala	GCA	8	2	23	42	11	14	Gly	GGA	0	0	0	1	0	4
Ala	GCG	20	3	29	27	17	28	Gly	GGG	3	1	2	0	2	6
residues used in analysis		613	1st 159	1342	1109	784	667	GC content of position III of quartet codons		66.5%	58.1%	54.1%	42.5%	52.8%	67.1%

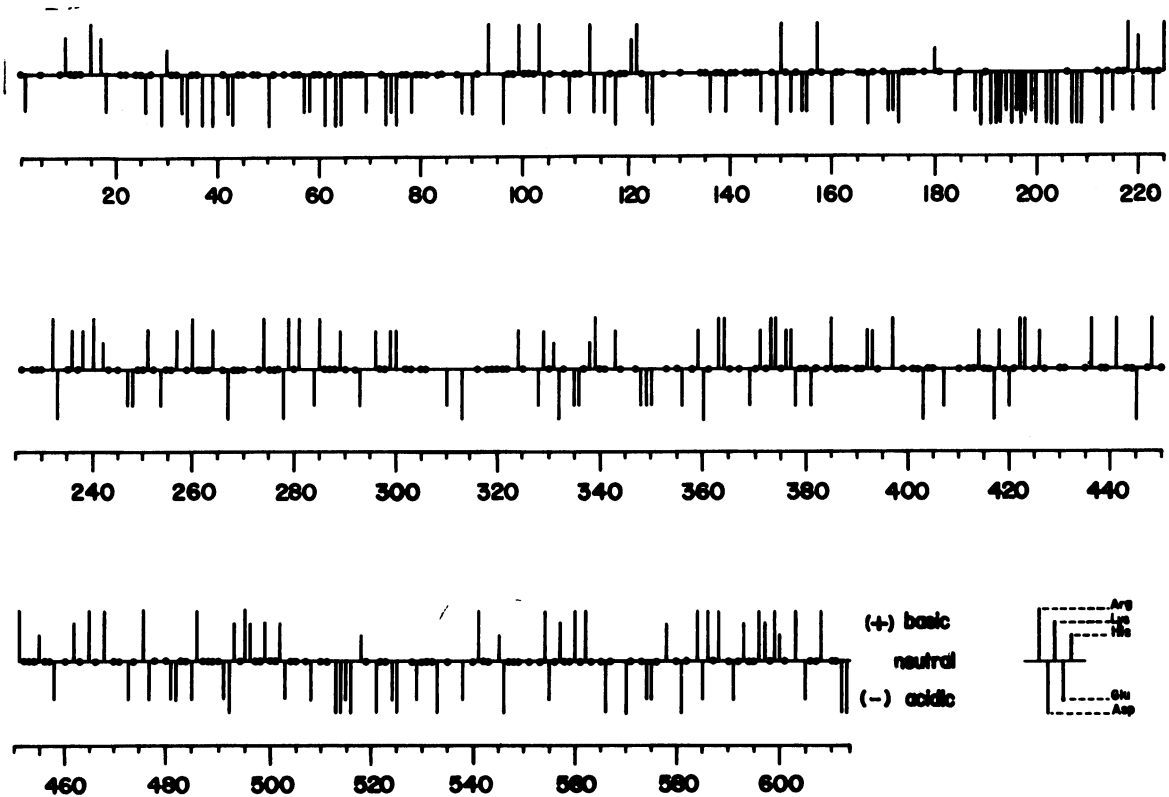

FIGURE 3. The distribution of charged amino acids in the sigma polypeptide of E. coli K12 RNA polymerase. The numbers indicate the amino acid residues from the N-terminus. Basic residues are indicated by upward lines, while acidic residues are indicated by downward lines. The lengths of these lines identify these residues as indicated. Hydrophobic residues (L+V+I+M+A+P+F+W) are indicated by filled circles.

polypeptide has a net charge of -36. Of the three regions with unusually high concentrations of acidic residues, two fall in the N-terminal third of the protein. Between residues number 33 and 90 there are 19 acidic residues and no basic residues. The region between number 184 and 215 has 21 acidic residues out of 32 residues with no basic residues. This includes a stretch of 18 acidic residues out of 22! These long acidic regions would give rise to large tryptic peptides of 76 and 61 residues, respectively. The only other major acidic region in the protein falls between residues 503 and 540 and contains 12 acidic residues with the only basic residue being a single histidine. This region would give rise to a tryptic peptide of 39 residues.

It is not surprising that sigma is one of the most acidic proteins in E. coli, with an isoelectric point, pI, estimated at 4.8-5.1 (19) or 4.40 (20) by isoelectric focusing gel analysis. The high concentration of negative charge in the N-terminus of sigma leads one to predict that an N-terminal sigma fragment would be more acidic than whole sigma. Cells containing pRRB2 (see Methods), synthesize a fusion protein consisting of the N-terminal 351 2/3 amino acids of sigma and 20 1/3 amino acids from the pK03 vector. This fusion protein is observable on two-dimensional gels and has an isoelectric point even lower than that of sigma (C. Gross, W. Walter, and R. Burgess, unpublished results).

The region between residues 419 and 434 is rich in aromatic residues, containing 2 Phe, 3 Tyr and 2 adjacent Trp. The three Cys residues are located at positions 132, 291, and 295. Partial cleavage of sigma at cysteines, with nitrothiocyanobenzoic acid (NTCB) (47) gives a pattern of peptides completely consistent with these positions (R. Burgess, W. Walter, unpublished results).

5) Secondary structure of sigma

The secondary structure of the sigma polypeptide was estimated from the amino acid sequence by the method of Chou and Fasman (48). This is a statistical method for predicting regions likely to form α-helical, β-strands, and reverse turns and is subject to a certain amount of uncertainty in assigning structure. We estimate that sigma contains 55-60% α-helix, 10-15% β-sheet, and 13-15% reverse turn. This preliminary estimate of high α-helical content is somewhat lower than an estimate, based on far-UV circular dichroism spectroscopy, published recently by Levine et al. (20). In that paper they calculated that sigma contains 75% of its residues in α-helical segments and less than 10% in β-sheet.

However, using the molecular weight and extinction coefficient for sigma reported here they have revised their figures for sigma to 55-62% α-helix and less than 10% β-sheet (S. Beychok, personal communication). In contrast, they found that core RNA polymerase is 33% in α-helix and 32% in β-sheet.

6) Molecular weight of sigma

The 613 amino acids coded for by the rpoD gene give an unmodified molecular weight for the sigma polypeptide of 70,263 daltons using the amino acid molecular weights given by Hunt et al. (49). This corresponds to a mean residue molecular weight of 114.6 (see Table I). This molecular weight is significantly less than the values of 82,000 daltons and 90,000 daltons determined by SDS polyacrylamide gel electrophoresis in non-stacking and stacking buffer systems, respectively (19). It seems that sigma is one of a number of proteins which exhibit anomalous electrophoretic migration on SDS gels. The reason for this anomalous behavior for sigma has not yet been determined but may be the result of its unusually high negative charge.

7) Operon Structure

A detailed analysis of the non-coding regions flanking the sigma structural gene is underway and will be presented elsewhere. The promoter for the sigma operon has been determined to lie between coordinates 1 and 191 of our DNA sequence. This determination was based on subcloning DNA fragments to the left of the structural gene into the promoter cloning vector pKO3 (described in Methods). A strong rho-independent terminator, containing a stable self-complementary stem followed by six uridine residues, is predicted by the DNA sequence to lie between co-ordinates 2418-2450. We have confirmed that the in vivo RNA for the operon ends at this point about 80 nucleotides past the end of the coding region (W. Taylor, Z. Burton, R. Burgess, and C. Gross, manuscript in preparation).

Since the sigma polypeptide coding region begins at co-ordinate 524, at least 333 nucleotides are transcribed which are not translated to make sigma. It seems likely that sigma, unlike the genes for core polymerase subunits, is the only gene in its operon. The function of the unusually long leader region remains to be elucidated.

ACKNOWLEDGEMENTS

We thank Dr. Fred Blattner and the following people in his laboratory for providing valuable advice, assistance, and facilities for DNA sequencing;

Nannette Newell, Gregory Goldberg, Donna Daniels, Julia Richards, and John Schroeder. We also thank Elio Vanin for his advice and assistance in DNA sequencing, Michael Gribskov for preparing Fig. 3, Marty Rosenberg and Keith McKenney for providing the promoter cloning vectors and helpful information prior to its publication, and Dr. Walter Fitch for his help in running the Chou-Fasman computer programs. This work was supported by NSF grants PCM77-25099 and PCM79-24915 (to R.R.B.), NIH Core Grant CA-07175 to McArdle Laboratory, NIH Postdoctoral Training Grant CA-09230 (to Z.B.), and NIH Grant GM-21812 (to F. Blattner).

REFERENCES

1. Present addresses: Judy Lin, Stanford Medical School, Stanford, CA 94305; David Moore, Dept. of Biochemistry, University of California, San Francisco, CA 94143; Sarah Holder, Monsanto Chemical Corp., St. Louis, MO.
2. Burgess, R. R., Travers, A. A., Dunn, J. J. & Bautz, F. K. F. (1969) Nature 221, 43-46
3. Chamberlin, M. (1974) Annu. Rev. Biochem. 43, 721-775
4. Nakamura, Y., Osawa, T. & Yura, T. (1977) Proc. Natl. Acad. Sci. USA 74, 1831-1835
5. Harris, J. D., Martinez, I. I. & Calendar, R. (1977) Proc. Natl. Acad. Sci. USA 74, 1836-1840
6. Gross, C., Hoffman, J., Ward, C., Hager, D., Burdick, G., Berger, H. & Burgess, R. (1978) Proc. Natl. Acad. Sci. USA 75, 427-431
7. Nakamura, Y. (1978) Molec. Gen. Genet. 105, 1-6
8. Harris, J. D., Heilig, J. S., Martinez, I. I., Calendar, R. & Isaksson, L. A. (1978) Proc. Natl. Acad. Sci. USA 75, 6177-6181
9. Travers, A. A. Buckland, R., Goman, M., LeGrice, S. S. G. & Scaife, J. G. (1978) Nature 273, 354-358
10. Nakamura, Y., Kurihara, T., Saito, H. and Uchida, H. (1979) Proc. Natl. Acad. Sci. USA 76, 4593-4597
11. Liebke, H., Gross, C., Walter, W. and Burgess, R. (1980) Molec. Gen. Genet. 177, 277-282
12. Burgess, R. R., Gross, C. A., Walter, W. & Lowe, P. A. (1979) Molec. Gen. Genet. 175, 251-257
13. Gross, C. A., Blattner, F. R., Taylor, W. E., Lowe, P. A. and Burgess, R. R. (1979) Proc. Natl. Acad. Sci. USA 76, 5789-5793
14. Scaife, J. G., Heilig, J. S., Rowen, L. and Calendar, R. (1979) Proc. Natl. Acad. Sci. USA 76, 6510-6514
15. Nakamura, Y. (1980) Molec. Gen. Genet. 178, 487-497
16. Burgess, R. R., Gross, C. and Engbaek, F. (1976) J. Bacteriol. 128, 390-393
17. Fujiki, H. and Zurek, G. (1975) FEBS Letters 55, 242-248
18. Burgess, R. R. (1976) in RNA Polymerase, R. Losick and M. Chamberlin, Eds, pp 69-100, Cold Spring Harbor Press, New York
19. Lowe, P. A., Hager, D. A. and Burgess, R. R. (1979) Biochemistry 18, 1344-1352
20. Levine, B. J., Orphanos, P. D., Fischmann, B. S. and Beychok, S. (1980) Biochemistry 19, 4808-4814
21. Berg, D., Barrett, K. and Chamberlin, M. J. (1971) Methods Enzymol. 21D, 506-519

22. Lowe, P. A., Aebi, U., Gross, C. A. and Burgess, R. R. (1981) J. Biol. Chem. 256, 2010-2015
23. Meissenberger, O., Pilz, I. and Heumann, H. (1980) FEBS Letters 112, 39-41
24. Coggins, J. R., Lumsden, J. and Malcolm, A. (1977) Biochemistry 16, 1111-1116
25. Hillel, Z. and Wu, C.-W. (1977) Biochemistry 16, 3334-3342
26. Hillel, Z. and Wu, C.-W. (1978) Biochemistry 17, 2954-2960
27. Simpson, R. B. (1979) Cell 18, 277-285
28. Sverdlov, E. D., Tsarev, S. A. and Begar, V. A. (1980) FEBS Letters 114, 111-114
29. Stender, W., Stütz, A. A. and Scheit, K. H. (1975) Eur. J. Biochem. 56, 129-136
30. Sutcliffe, J. G. (1978) Cold Spring Harbor Symposium, 43, 77-90
31. Birnboim, H. and Doly, J. (1979) Nuc. Acid. Res. 7, 1513-1523
32. Maxam, A. and Gilbert, W. (1980) in Methods in Enzymol., L. Grossman and K. Moldave, Eds, 65, 499-560, Academic Press, New York, NY
33. McKenney, K., Shimatake, H., Court, D., Schmeissner, U., Brady, C. and Rosenberg, M. (1981) in Gene Amplification and Analysis, Vol II; J. Chirikjian and T. Papas, Eds, Elsevier-North Holland, in press
34. Maniatis, T., Jeffrey, A. and Kleid, D. G. (1975) Proc. Natl. Acad. Sci. USA 72, 1184-1188.
35. Shine, J. and Dalgarno, L. (1974) Proc. Natl. Acad. Sci. USA 71, 1342-1346
36. Ovchinnikov, Yu A., Lipkin, V. A., Modyanov, N. N., Chertov, O., and Smirnov, Yu V. (1977) FEBS Letters 76, 108-111
37. Ovchinnikov, Yu A., Monastyrskaia, G. S., Gubanov, V., Guriev, S., Chertov, O., Modyanov, N., Grinkevich, V., Markarova, I., Marchenko, T., Polovnikova, I., Lipkin, V. and Sverdlov, E. (1980) Dokl. Acad. Nauk USSR 253, 994-998
38. Dayhoff, M. O., Hunt, L. T. and Hurst-Calderone, S. (1978) in Atlas of Protein Sequence and Structure, M. O. Dayhoff, Ed. Vol. 5, Suppl. 3, 363-369
39. Sober, H. A., Ed. (1970) in Handbook of Biochemistry, CRC Press, Cleveland, Ohio
40. Post, L. E. and Nomura, M. (1979) J. Biol. Chem. 254, 10604-10606
41. Post, L. E. and Nomura, M. (1980) J. Biol. Chem. 255, 4660-4666
42. Yokota, T., Sugisaki, H., Takanami, M., and Kaziro, Y. (1980) Gene 12, 25-31
43. An, G. and Friesen, J. D. (1980) Gene 12, 33-39
44. Crawford, I. P., Nichols, B. P. and Yanofsky, C. (1980) J. Mol. Biol. 142, 489-502
45. Ikemura, T. (1981) J. Mol. Biol. 146, 1-21
46. Grantham, R., Gautier, C. and Gouy, M. (1980) Nucleic Acids Res. 8, 1893-1912
47. Stark, G. R. (1977) in Methods in Enzymol., C. Hirs, and S. Timasheff, Eds, 47, 129-132, Academic Press, New York, NY
48. Chou, P. Y. and Fasman, G. D. (1978) Adv. in Enzymol., 47, 45-148
49. Hunt, L. T., Barker, W. G. and Dayhoff, M. O. (1978) in Atlas of Protein Sequence and Structure, M. O. Dayhoff, Ed., Vol 5, Suppl. 3, 25-27