Volume 10 Number 13 1982 Nucleic Acids Research

The primary structure of *E. coli* RNA polymerase. Nucleotide sequence of the rpoC gene and amino acid sequence of the β'-subunit

Yu.A.Ovchinnikov, G.S.Monastyrskaya, V.V.Gubanov, S.O.Guryev, I.S.Salomatina, T.M.Shuvaeva, V.M.Lipkin and E.D.Sverdlov

M.M.Shemyakin Institute of Bioorganic Chemistry, USSR Academy of Sciences, Moscow, USSR

Received 12 May 1982; Accepted 6 June 1982

ABSTRACT

The primary structure of the E.coli rpoC gene (5321 base pairs) coding the β'-subunit of RNA polymerase as well as its adjacent segment have been determined. The structure analysis of the peptides obtained by cleavage of the protein with cyanogen bromide and trypsin has confirmed the amino acid sequence of the β'-subunit deduced from the nucleotide sequence analysis. The β'-subunit of E.coli RNA polymerase contains 1407 amino acid residues. Its translation is initiated by codon GUG and terminated by codon TAA. It has been detected that the sequence following the terminating codon is strikingly homologous to known sequences of ρ-independent terminators.

INTRODUCTION

The primary structure determination of E.coli DNA-dependent RNA polymerase is necessary for understanding the mechanism of its activity. Recently we determined the complete amino acid sequences of its α-/1/ and β-subunits /2/. The primary structure of the β-subunit containing 1342 amino acid residues was established by using parallel research of the protein amino acid sequence and the nucleotide sequence of its structural gene. Combination of protein and nucleotide chemistry methods greatly enhanced the reliability of the analysis. At present this approach was successfully applied in the sequencing of the β'-subunit of RNA polymerase. The DNA fragments containing the rpoC gene fragments and adjacent sequences were cloned in pBR-322. Their sequences were determined from both complementary chains by the modified Maxam-Gilbert procedure /3/. The amino acid sequence of the β'-subunit deduced from the nucleotide sequence was compared and appeared to be in complete accord with structures of the peptides obtained by cleavage with cyanogen bromide and trypsin. The β'-subunit of E.coli RNA polymerase comprises 1407 amino acid residues. The β'-subunit sequence determination completes the study of the primary structure of the E.coli RNA polymerase core-enzyme. Recently the sequence of the rpoD gene was also determined /4/. Thus, the primary structure of the whole RNA polymerase holoenzyme is now

0305-1048/82/1013-4035$2.00/0

available permitting investigation of its function.

MATERIALS AND METHODS

The EcoR I, Taq I, Bsp I, Sal I restriction endonucleases were isolated according to /5/. Endonucleases Hpa II, Alu I, Sau3A I, Hinf I were purchased from "P.L. Biochemicals". Polynucleotide kinase was isolated according to /6/. The T4-DNA ligase was from the Institute of Biochemistry and Physiology of Micro-organisms, USSR Academy of Sciences. The E.coli strain containing pJC 703 plasmid /7/ was the gift of Dr. J. Collins (FRG). The phage λrif^{d} 47 was provided by Prof.R.B. Khesin (IMG, USSR Academy of Sciences, Moscow).

The EcoR I DNA fragment of E.coli containing the middle part of the rpoC gene was generated by EcoR I endonuclease from $\lambda rif^{d}47$ phage DNA and isolated by preparative electrophoresis in agarose gel /8/. EcoRI-SalI DNA fragment containing the C-terminal part of the rpoC gene was obtained similarly from the pJC703 plasmid. Cloning of the fragments was carried out with plasmid pBR-322 as an acceptor and E.coli HB 101 as a host /9/. Recombinants were selected by restriction analysis of isolated plasmids. The first fragment was inserted in pBR-322 split with EcoRI, the second one was inserted in a large fragment of pBR-322 split with EcoRI and SalI. In this case the $Ap^{r}Tc^{s}$ clones were selected and their plasmids were characterized by splitting with EcoRI and SalI. The recovery of DNA fragments from the corresponding recombinant plasmids, their preparation for structural analysis and the analysis itself were described earlier. The conditions for β'-subunit cleavage with cyanogen bromide, isolation and analysis of the corresponding peptides have also been given /10/. Tryptic hydrolysis of the citraconylated and carboxymethylated β'-subunit and establishment of the peptide structure were carried out by the methods we used for the β-subunit analysis /2,11/.

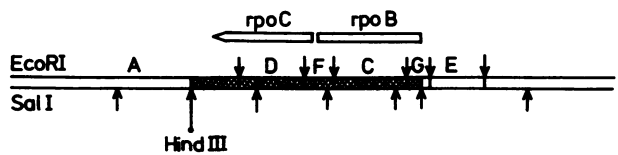

Fig. 1. EcoRI and SalI restriction endonuclease cleavage map of the E.coli DNA region containing the structural genes /rpoB and rpoC / of the β and β' subunits of RNA polymerase. The HindIII cleavage region is also denoted in fragment EcoRI. The fragments, for which the primary structure is established, are hatched.

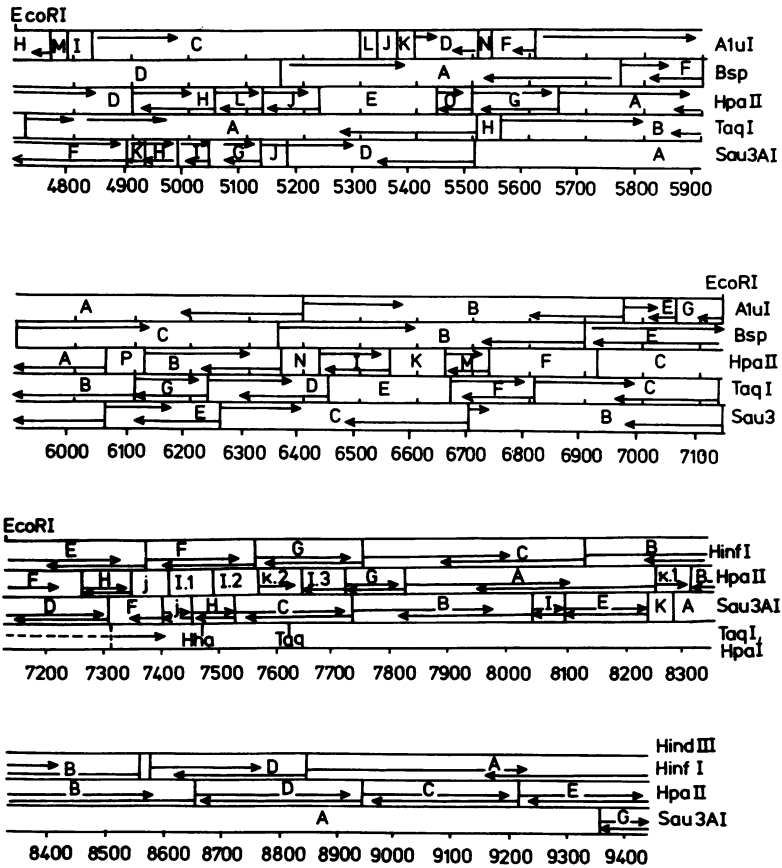

Fig. 2. The scheme for determining the sequences of fragments EcoRI- and EcoRI-A - HindIII. The restriction subfragments are represented by rectangles. The arrows designate lengths of the determined subfragments of complementary chains.

RESULTS AND DISCUSSION

The rpoBC operon coding for β - and β'-subunits of E.coli RNA polymerase is located near 88 min on the genetic map of the bacterium. The positions of the splitting sites of several restriction endonucleases on the DNA in this region were established and given in Fig 1. We determined the primary structures of the EcoRI-G, EcoRI-C and EcoRI-F fragments containing the complete rpoB (β-subunit) gene and a proximal part of the rpoC (β'-subunit) gene /2/. For sequencing the rest of the operon, the fragments EcoRI-A - SalI and EcoRI-D were cloned in pBR-322 plasmid. The recombinant plasmid, containing the first of them, was split with EcoRI and HindIII to obtain

4130-4210
1-8
CTG CTG TCG GGT TAA AAC CCG GCA GCG GAT TGT GCT AAC TCC GAC GGG AGC AAA TCC GTG AAA GAT TTA TTA AAG TTT CTG
Met-Lys-Asp-Leu-Leu-Lys-Phe-Leu-

4211-4291
9-35
AAA GCG CAG ACT AAA ACC GAA GAG TTT GAT GCG ATC AAA ATT GCT CTG GCT TCG CCA GAC ATG ATC CGT TCA TGG TCT TTC
Lys-Ala-Gln-Thr-Lys-Thr-Glu-Glu-Phe-Asp-Ala-Ile-Lys-Ile-Ala-Leu-Ala-Ser-Pro-Asp-Met-Ile-Arg-Ser-Trp-Ser-Phe-

4292-4372
36-62
GGT GAA GTT AAA AAG CCG GAA ACC ATC AAC TAC CGT ACG TTC AAA CCA GAA CGT GAC GGC CTT TTC TGC GCC CGT ATC TTT
Gly-Glu-Val-Lys-Lys-Pro-Glu-Thr-Ile-Asn-Tyr-Arg-Thr-Phe-Lys-Pro-Glu-Arg-Asp-Gly-Leu-Phe-Cys-Ala-Arg-Ile-Phe-

4373-4453
63-89
GGG CCG GTA AAA GAT TAC GAG TGC CTG TGC GGT AAG TAC AAG CGC CTG AAA CAC CGT GGC GTC ATC TGT GAG AAG TGC GGC
Gly-Pro-Val-Lys-Asp-Tyr-Glu-Cys-Leu-Cys-Gly-Lys-Tyr-Lys-Arg-Leu-Lys-His-Arg-Gly-Val-Ile-Cys-Glu-Lys-Cys-Gly-

4454-4534
90-116
GTT GAA GTG ACC CAG ACT AAA GTA CGC CGT GAG CGT ATG GGC CAC ATC GAA CTG GCT TCC CCG ACT GCG CAC ATC TGG TTC
Val-Glu-Val-Thr-Gln-Thr-Lys-Val-Arg-Arg-Glu-Arg-Met-Gly-His-Ile-Glu-Leu-Ala-Ser-Pro-Thr-Ala-His-Ile-Trp-Phe-

4535-4615
117-143
CTG AAA TCG CTG CCG TCC CGT ATC GGT CTG CTG CTC GAT ATG CCG CTG CGC GAT ATC GAA CGC GTA CTG TAC TTT GAA TCC
Leu-Lys-Ser-Leu-Pro-Ser-Arg-Ile-Gly-Leu-Leu-Leu-Asp-Met-Pro-Leu-Arg-Asp-Ile-Glu-Arg-Val-Leu-Tyr-Phe-Glu-Ser-

4616-4696
144-170
TAT GTG GTT ATC GAA GGC GGT ATG ACC AAC CTG GAA CGT CAG CAG ATC CTG ACT GAA GAG CAG TAT CTG GAC GCG CTG GAA
Tyr-Val-Val-Ile-Glu-Gly-Gly-Met-Thr-Asn-Leu-Glu-Arg-Gln-Gln-Ile-Leu-Thr-Glu-Glu-Gln-Tyr-Leu-Asp-Ala-Leu-Glu-

4697-4777
171-197
GAG TTC GGT GAC GAA TTC GAC GCG AAG ATG GGG GCG GAA GCA ATC CAG GCT CTG CTG AAG AGC ATG GAT CTG GAG CAA GAG
Glu-Phe-Gly-Asp-Glu-Phe-Asp-Ala-Lys-Met-Gly-Ala-Glu-Ala-Ile-Gln-Ala-Leu-Leu-Lys-Ser-Met-Asp-Leu-Glu-Gln-Glu-

4778-4858
198-224
TGC GAA CAG CTG CGT GAA GAG CTG AAC GAA ACC AAC TCC GAA ACC AAG CGT AAA AAG CTG ACC AAG CGT ATC AAA CTG CTG
Cys-Glu-Gln-Leu-Arg-Glu-Glu-Leu-Asn-Glu-Thr-Asn-Ser-Glu-Thr-Lys-Arg-Lys-Lys-Leu-Thr-Lys-Arg-Ile-Lys-Leu-Leu-

4859-4939
225-251
GAA GCG TTC GTT CAG TCT GGT AAC AAA CCA GAG TGG ATG ATC CTG ACC GTT CTG CCG GTA CTG CCG CCA GAT CTG CGT CCG
Glu-Ala-Phe-Val-Gln-Ser-Gly-Asn-Lys-Pro-Glu-Trp-Met-Ile-Leu-Thr-Val-Leu-Pro-Val-Leu-Pro-Pro-Asp-Leu-Arg-Pro-

4940-5020
252-278
CTG GTT CCG CTG GAT GGT GGT CGT TTC GCG ACT TCT GAC CTG AAC GAT CTG TAT CGT CGC GTC ATT AAC CGT AAC AAC CGT
Leu-Val-Pro-Leu-Asp-Gly-Gly-Arg-Phe-Ala-Thr-Ser-Asp-Leu-Asn-Asp-Leu-Tyr-Arg-Arg-Val-Ile-Asn-Arg-Asn-Asn-Arg-

5021-5101
279-305
CTG AAA CGT CTG CTG GAT CTG GCT GCG CCG GAC ATC ATC GTA CGT AAC GAA AAA CGT ATG CTG CAG GAA GCG GTA GAC GCC
Leu-Lys-Arg-Leu-Leu-Asp-Leu-Ala-Ala-Pro-Asp-Ile-Ile-Val-Arg-Asn-Gly-Lys-Arg-Met-Leu-Gln-Glu-Ala-Val-Asp-Ala-

5102-5182
306-332
CTG CTG GAT AAC GGT CGT CGC GGT CGT GCG ATC ACC GGT TCT AAC AAG CGT CCT CTG AAA TCT TTG GCC GAC ATG ATC AAA
Leu-Leu-Asp-Asn-Gly-Arg-Arg-Gly-Arg-Ala-Ile-Thr-Gly-Ser-Asn-Lys-Arg-Pro-Leu-Lys-Ser-Leu-Ala-Asp-Met-Ile-Lys-

5183-5263
333-359
GGT AAA CAG GGT CGT TTC CGT CAG AAC CTG CTC GGT AAG CGT GTT GAC TAC TCC GGT CGT TCT GTA ATC ACC GTA GGT CCA
Gly-Lys-Gln-Gly-Arg-Phe-Arg-Gln-Asn-Leu-Leu-Gly-Lys-Arg-Val-Asp-Tyr-Ser-Gly-Arg-Ser-Val-Ile-Thr-Val-Gly-Pro-

5264-5344
360-386
TAC CTG CGT CTG CAT CAG TGC GGT CTG CCG AAG AAA ATG GCA CTG GAG CTG TTC AAA CCG TTC ATC TAC GGC AAG CTG GAA
Tyr-Leu-Arg-Leu-His-Gln-Cys-Gly-Leu-Pro-Lys-Lys-Met-Ala-Leu-Glu-Leu-Phe-Lys-Pro-Phe-Ile-Tyr-Gly-Lys-Leu-Glu-

5345-5425
387-413
CTG CGT GGT CTT GCT ACC ACC ATT AAA GCT GCG AAG AAA ATG GTT GAG CGC GAA GAA GCT GTC GTT TGG GAT ATC CTG GAC
Leu-Arg-Gly-Leu-Ala-Thr-Thr-Ile-Lys-Ala-Ala-Lys-Lys-Met-Val-Glu-Arg-Glu-Glu-Ala-Val-Val-Trp-Asp-Ile-Leu-Asp-

5426-5506
414-440
GAA GTT ATC CGC GAA CAC CCG GTA CTG CTG AAC CGT GCA CCG ACT CTG CAC CGT CTG GGT ATC CAG GCA TTT GAA CCG GTA
Glu-Val-Ile-Arg-Glu-His-Pro-Val-Leu-Leu-Asn-Arg-Ala-Pro-Thr-Leu-His-Arg-Leu-Gly-Ile-Gln-Ala-Phe-Glu-Pro-Val-

5507-5587
441-467
CTG ATC GAA GGT AAA GCT ATC CAG CTG CAC CCG CTG GTT TGT GCG GCA TAT AAC GCC GAC TTC GAT GGT GAC CAG ATG GCT
Leu-Ile-Glu-Gly-Lys-Ala-Ile-Gln-Leu-His-Pro-Leu-Val-Cys-Ala-Ala-Tyr-Asn-Ala-Asp-Phe-Asp-Gly-Asp-Gln-Met-Ala-

5588-5668
468-494
GTT CAC GTA CCG CTG ACG CTG GAA GCC CAG CTG GAA GCG CGT GCG CTG ATG ATG TCT ACC AAC AAC ATC CTG TCC CCG GCG
Val-His-Val-Pro-Leu-Thr-Leu-Glu-Ala-Gln-Leu-Glu-Ala-Arg-Ala-Leu-Met-Met-Ser-Thr-Asn-Asn-Ile-Leu-Ser-Pro-Ala-

5669-5749
495-521
AAC GGC GAA CCA ATC ATC GTT CCG TCT CAG GAC GTT GTA CTG GGT CTG TAC TAC ATG ACC CGT GAC TGT GTT AAC GCC AAA
Asn-Gly-Glu-Pro-Ile-Ile-Val-Pro-Ser-Gln-Asp-Val-Val-Leu-Gly-Leu-Tyr-Tyr-Met-Thr-Arg-Asp-Cys-Val-Asn-Ala-Lys-

5750-5830
522-548
GGC GAA GGC ATG GTG CTG ACT GGC CCG AAA GAA GCA GAA CGT CTG TAT CGC TCT GGT CTG GCT TCT CTG CAT GCG CGC GTT
Gly-Glu-Gly-Met-Val-Leu-Thr-Gly-Pro-Lys-Glu-Ala-Glu-Arg-Leu-Tyr-Arg-Ser-Gly-Leu-Ala-Ser-Leu-His-Ala-Arg-Val-

5831-5911
549-575
AAA GTG CGT ATC ACC GAG TAT GAA AAA GAT GCT AAC GGT GAA TTA GTA GCG AAA ACC AGC CTG AAA GAC ACG ACT GTT GGC
Lys-Val-Arg-Ile-Thr-Glu-Tyr-Glu-Lys-Asp-Ala-Asn-Gly-Glu-Leu-Val-Ala-Lys-Thr-Ser-Leu-Lys-Asp-Thr-Thr-Val-Gly-

5912-5992
576-602
CGT GCC ATT CTG TGG ATG ATT GTA CCG AAA GGT CTG CCT TAC TCC ATC GTC AAC CAG GCG CTG GGT AAA AAA GCA ATC TCC
Arg-Ala-Ile-Leu-Trp-Met-Ile-Val-Pro-Lys-Gly-Leu-Pro-Tyr-Ser-Ile-Val-Asn-Gln-Ala-Leu-Gly-Lys-Lys-Ala-Ile-Ser-

5993-6073
603-629
AAA ATG CTG AAC ACC TGC TAC CGC ATT CTC GGT CTG AAA CCG ACC GTT ATT TTT GCG GAC GAG ATC ATG TAC ACC GGC TTC
Lys-Met-Leu-Asn-Thr-Cys-Tyr-Arg-Ile-Leu-Gly-Leu-Lys-Pro-Thr-Val-Ile-Phe-Ala-Asp-Gln-Ile-Met-Tyr-Thr-Gly-Phe-

6074-6154
630-656
GCC TAT GCA GCG CGT TCT GGT GCA TCT GTT GGT ATC GAT GAC ATG GTC ATC CCG GAG AAG AAA CAC GAA ATC ATC TCC GAG
Ala-Tyr-Ala-Ala-Arg-Ser-Gly-Ala-Ser-Val-Gly-Ile-Asp-Asp-Met-Val-Ile-Pro-Glu-Lys-Lys-His-Glu-Ile-Ile-Ser-Glu-

6155-6235
657-683
GCA GAA GCA GAA GTT GCT GAA ATT CAG GAG CAG TTC CAG TCT GGT CTG GTA ACT GCG GGC GAA CGC TAC AAC AAA GTT ATC
Ala-Glu-Ala-Glu-Val-Ala-Glu-Ile-Gln-Glu-Gln-Phe-Gln-Ser-Gly-Leu-Val-Thr-Ala-Gly-Glu-Arg-Tyr-Asn-Lys-Val-Ile-

6236-6316
684-710
GAT ATC TGG GCT GCG GCG AAC GAT CGT GTA TCC AAA GCG ATG ATG GAT AAC CTG CAA ACT GAA ACC GTG ATT AAC CGT GAC
Asp-Ile-Trp-Ala-Ala-Ala-Asn-Asp-Arg-Val-Ser-Lys-Ala-Met-Met-Asp-Asn-Leu-Gln-Thr-Glu-Thr-Val-Ile-Asn-Arg-Asp-

6317-6397
711-737
GGT CAG GAA GAG AAG CAG GTT TCC TTC AAC AGC ATC TAC ATG ATG GCC GAC TCC GGT GCG CGT GGT TCT GCG GCA CAG ATT
Gly-Gln-Glu-Glu-Lys-Gln-Val-Ser-Phe-Asn-Ser-Ile-Tyr-Met-Met-Ala-Asp-Ser-Gly-Ala-Arg-Gly-Ser-Ala-Ala-Gln-Ile-

6398-6478
738-764
CGT CAG CTT GCT GGT ATG CGT GGT CTG ATG GCG AAG CCG GAT GGC TCC ATC ATC GAA ACG CCA ATC ACC GCG AAC TTC CGT
Arg-Gln-Leu-Ala-Gly-Met-Arg-Gly-Leu-Met-Ala-Lys-Pro-Asp-Gly-Ser-Ile-Ile-Glu-Thr-Pro-Ile-Thr-Ala-Asn-Phe-Arg-

6479-6559
765-791
GAA GGT CTG AAC GTA CTC CAG TAC TTC ATC TCC ACC CAC GGT GCT CGT AAA GGT CTG GCG GAT ACC GCA CTG AAA ACT GCG
Glu-Gly-Leu-Asn-Val-Leu-Gln-Tyr-Phe-Ile-Ser-Thr-His-Gly-Ala-Arg-Lys-Gly-Leu-Ala-Asp-Thr-Ala-Leu-Lys-Thr-Ala-

6560-6640
792-818
AAC TCC GGT TAC CTG ACT CGT CGT CTG GTT GAC GTG GCG CAG GAC CTG GTG GTT ACC GAA GAC GAT TGT GGT ACC CAT GAA
Asn-Ser-Gly-Tyr-Leu-Thr-Arg-Arg-Leu-Val-Asp-Val-Ala-Gln-Asp-Leu-Val-Val-Thr-Glu-Asp-Asp-Cys-Gly-Thr-His-Glu-

6641-6721
819-845
GGT ATC ATG ATG ACT CCG GTT ATC GAG GGT GGT GAC GTT AAA GAG CCG CTG CGC GAT CGC GTA CTG GGT CGT GTA ACT GCT
Gly-Ile-Met-Met-Thr-Pro-Val-Ile-Glu-Gly-Gly-Asp-Val-Lys-Glu-Pro-Leu-Arg-Asp-Arg-Val-Leu-Gly-Arg-Val-Thr-Ala-

6722-6802
846-872
GAA GAC GTT CTG AAG CCG GGT ACT GCT GAT ATC CTC GTT CCG CGC AAC ACG CTG CTG CAC GAA CAG TGG TGT GAC CTG CTG
Glu-Asp-Val-Leu-Lys-Pro-Gly-Thr-Ala-Asp-Ile-Leu-Val-Pro-Arg-Asn-Thr-Leu-Leu-His-Glu-Gln-Trp-Cys-Asp-Leu-Leu-

6803-6883 GAA GAG AAC TCT GTC GAC GCG GTT AAA GTA CGT TCT GTT GTA TCT TGT GAC ACC GAC TTT GGT GTA TGT GCG CAC TGC TAC
873-899 Glu-Glu-Asn-Ser-Val-Asp-Ala-Val-Lys-Val-Arg-Ser-Val-Val-Ser-Cys-Asp-Thr-Asp-Phe-Gly-Val-Cys-Ala-His-Cys-Tyr-

6884-6964 GGT CGT GAC CTG GCG CGT GGC CAC ATC ATC AAC AAG GGT GAA GCA ATC GGT GTT ATC GCG GCA CAG TCC ATC GGT GAA CCG
900-926 Gly-Arg-Asp-Leu-Ala-Arg-Gly-His-Ile-Ile-Asn-Lys-Gly-Glu-Ala-Ile-Gly-Val-Ile-Ala-Ala-Gln-Ser-Ile-Gly-Glu-Pro-

6965-7045 GGT ACA CAG CTG ACC ATG CGT ACG TTC CAC ATC GGT GGT GCG GCA TCT CGT GCG GCT GCT GAA TCC AGC ATC CAA GTG AAA
927-953 Gly-Thr-Gln-Leu-Thr-Met-Arg-Thr-Phe-His-Ile-Gly-Gly-Ala-Ala-Ser-Arg-Ala-Ala-Ala-Glu-Ser-Ser-Ile-Gln-Val-Lys-

7046-7126 AAC AAA GGT AGC ATC AAG CTC AGC AAC GTG AAG TCG GTT GTG AAC TCC AGC GGT AAA CTG GTT ATC ACT TCC CGT AAT ACT
954-980 Asn-Lys-Gly-Ser-Ile-Lys-Leu-Ser-Asn-Val-Lys-Ser-Val-Val-Asn-Ser-Ser-Gly-Lys-Leu-Val-Ile-Thr-Ser-Arg-Asn-Thr-

7127-7207 GAA CTG AAA CTG ATC GAC GAA TTC GGT CGT ACT AAA GAA AGC TAC AAA GTA CCT TAC GGT GCG GTA CTG GCG AAA GGC GAT
981-1007 Glu-Leu-Lys-Leu-Ile-Asp-Glu-Phe-Gly-Arg-Thr-Lys-Glu-Ser-Tyr-Lys-Val-Pro-Tyr-Gly-Ala-Val-Leu-Ala-Lys-Gly-Asp-

7208-7288 GGC GAA CAG GTT GCT GGC GGC GAA ACC GTT GCA AAC TGG GAC CCG CAC ACC ATG CCG GTT ATC ACC GAA GTA AGC GGT TTT
1008-1034 Gly-Glu-Gln-Val-Ala-Gly-Gly-Glu-Thr-Val-Ala-Asn-Trp-Asp-Pro-His-Thr-Met-Pro-Val-Ile-Thr-Glu-Val-Ser-Gly-Phe-

7289-7369 GTA CGC TTT ACT GAC ATG ATC GAC GGC CAG ACC ATT ACG CGT CAG ACC GAC GAA CTG ACC GGT CTG TCT TCG CTG GTG GTT
1035-1061 Val-Arg-Phe-Thr-Asp-Met-Ile-Asp-Gly-Gln-Thr-Ile-Thr-Arg-Gln-Thr-Asp-Glu-Leu-Thr-Gly-Leu-Ser-Ser-Leu-Val-Val-

7370-7450 CTG GAT TCC GCA GAA CGT ACC GCA GGT GGT AAA GAT CTG CGT CCG GCA CTG AAA ATC GTT GAT GCT CAG GGT AAC GAC GTT
1062-1088 Leu-Asp-Ser-Ala-Glu-Arg-Thr-Ala-Gly-Gly-Lys-Asp-Leu-Arg-Pro-Ala-Leu-Lys-Ile-Val-Asp-Ala-Gln-Gly-Asn-Asp-Val-

7451-7531 CTG ATC CCA GGT ACC GAT ATG CCA GCG CAG TAC TTC CTG CCG GGT AAA GCG ATT GTT CAG CTG GAA GAT GGC GTA CAG ATC
1089-1115 Leu-Ile-Pro-Gly-Thr-Asp-Met-Pro-Ala-Gln-Tyr-Phe-Leu-Pro-Gly-Lys-Ala-Ile-Val-Gln-Leu-Glu-Asp-Gly-Val-Gln-Ile-

7532-7612 AGC TCT GGT GAC ACC CTG GCG CGT ATT CCG CAG GAA TCC GGC GGT ACC AAG GAC ATC ACC GGT GGT CTG CCG CGC GTT GCG
1116-1142 Ser-Ser-Gly-Asp-Thr-Leu-Ala-Arg-Ile-Pro-Gln-Glu-Ser-Gly-Gly-Thr-Lys-Asp-Ile-Thr-Gly-Gly-Leu-Pro-Arg-Val-Ala-

7613-7693 GAC CTG TTC GAA GCA CGT CGT CCG AAA GAG CCG GCA ATC CTG GCT GAA ATC AGC GGT ATC GTT TCC TTC GGT AAA GAA ACC
1143-1169 Asp-Leu-Phe-Glu-Ala-Arg-Arg-Pro-Lys-Glu-Pro-Ala-Ile-Leu-Ala-Glu-Ile-Ser-Gly-Ile-Val-Ser-Phe-Gly-Lys-Glu-Thr-

7694-7774 AAA GGT AAA CGT CGT CTG GTT ATC ACC CCG GTA GAC GGT AGC GAT CCG TAC GAA GAG ATG ATT CCG AAA TGG CGT CAG CTC
1170-1196 Lys-Gly-Lys-Arg-Arg-Leu-Val-Ile-Thr-Pro-Val-Asp-Gly-Ser-Asp-Pro-Tyr-Glu-Glu-Met-Ile-Pro-Lys-Trp-Arg-Gln-Leu-

7775-7855 AAC GTG TTC GAA GGT GAA CGT GTA GAA CGT GGT GAC GTA ATT TCC GAC GGT CCG GAA GCG CCG CAC GAC ATT CTG CGT CTG
1197-1223 Asn-Val-Phe-Glu-Gly-Glu-Arg-Val-Glu-Arg-Gly-Asp-Val-Ile-Ser-Asp-Gly-Pro-Glu-Ala-Pro-His-Asp-Ile-Leu-Arg-Leu-

7856-7936 CGT GGT GTT CAT GCT GTT ACT CGT TAC ATC GTT AAC GAA GTA CAG GAC GTA TAC CGT CTG CAG GGC GTT AAG ATT AAC GAT
1224-1250 Arg-Gly-Val-His-Ala-Val-Thr-Arg-Tyr-Ile-Val-Asn-Glu-Val-Gln-Asp-Val-Tyr-Arg-Leu-Gln-Gly-Val-Lys-Ile-Asn-Asp-

7937-8017 AAA CAC ATC GAA GTT ATC GTT CGT CAG ATG CTG CGT AAA GCT ACC ATC GTT AAC GCG GGT AGC TCC GAC TTC CTG GAA GGC
1251-1277 Lys-His-Ile-Glu-Val-Ile-Val-Arg-Gln-Met-Leu-Arg-Lys-Ala-Thr-Ile-Val-Asn-Ala-Gly-Ser-Ser-Asp-Phe-Leu-Glu-Gly-

8018-8098 GAA CAG GTT GAA TAC TCT CGC GTC AAG ATC GCA AAC CGC GAA CTG GAA GCG AAC GGC AAA GTG GGT CCA ACT TAC TCC CGC
1278-1304 Glu-Gln-Val-Glu-Tyr-Ser-Arg-Val-Lys-Ile-Ala-Asn-Arg-Glu-Leu-Glu-Ala-Asn-Gly-Lys-Val-Gly-Ala-Thr-Tyr-Ser-Arg-

8099-8179 GAT CTG CTG GGT ATC ACC AAA GCG TCT CTG GCA ACC GAG TCC TTC ATC TCC GCG GCA TCG TTC CAG GAG ACC ACT CGC GTG
1305-1331 Asp-Leu-Leu-Gly-Ile-Thr-Lys-Ala-Ser-Leu-Ala-Thr-Glu-Ser-Phe-Ile-Ser-Ala-Ala-Ser-Phe-Gln-Glu-Thr-Thr-Arg-Val-

8180-8260 CTG ACC GAA GCA GCC GTT GCG GGC AAA CGC GAC GAA CTG CGC GGC CTG AAA GAG AAC GTT ATC GTG GGT CGT CTG ATC CCG
1332-1358 Leu-Thr-Glu-Ala-Ala-Val-Ala-Gly-Lys-Arg-Asp-Glu-Leu-Arg-Gly-Leu-Lys-Glu-Asn-Val-Ile-Val-Gly-Arg-Leu-Ile-Pro-

8261-8341 GCA GGT ACC GGT TAC GCG TAC CAC CAG GAT CGT ATG CGT CGC CGT GCT GCG GGT GAA GCT CCG GCT GCA CCG CAG GTG ACT
1359-1385 Ala-Gly-Thr-Gly-Tyr-Ala-Tyr-His-Gln-Asp-Arg-Met-Arg-Arg-Arg-Ala-Ala-Gly-Glu-Ala-Pro-Ala-Ala-Pro-Gln-Val-Thr-

8342-8426 GCA GAA GAC GCA TCT GCC AGC CTG GCA GAA CTG CTG AAC GCA GGT CTG GGC GGT TCT GAT AAC GAG TAA TCGTTAATCCGCAAA
1386-1407 Ala-Glu-Asp-Ala-Ser-Ala-Ser-Leu-Ala-Glu-Leu-Leu-Asn-Ala-Gly-Leu-Gly-Gly-Ser-Asp-Asn-Glu-Ter

8427-8532 TAACGTAAAAACCCGCTTCGGCGGGTTTTTTTATGGGGGGAGTTTAGGGAAAGAGCATTTGTCAGAATATTTAAGGAATTTCTGAATACTCATAATCAATGTAGAGA

8533-8639 TGACTAATATCCTGAAACTGACTGAACTAATTGAGTCAAACTCGGCAAGGATTCGATACTATTCCTGTGTAACTTTCTTAAGGAACGAGAATGAAACAGGAAGTGGA

8640-8746 AAAGTGGCGACCTTTTGGACATCCGGATGGTGATATTCGTGATTTATCATTTCTTGATGCTCATCAGGCTGTCTACGTTCAGCATCATGAGGGCAAAGAGCCTTTAG

8747-8853 AGTATCGCTTTTGGGTTACCTACTCTCTTCACTGCTTCACAAAAGATTATGAACATCAGACGAACGAAGAAAAACAATCGTTAATGTACCACGCGCCTAAAGAATCT

8854-8960 CGTCCCTTCTGCCAGCACCGTTATAACTTAGCGCGCACACACTTAAAAAGAACTATTTTGGCGCTGCCAGAAAGCAACGTTATTCATGCCGGGTATGGTAGCTATGC

8961-9067 CGTGATTGAGGTGGACTTAGACGGAGGAGATAAGGCATTTTACTTTGTTGCGTTCAGGGCTTTCAGGGAAAAGAAAAAACTCCGTTTGCATGTAACTAGCGCTTATC

9068-9174 CCATTTCTGAAAAACAGAAAGGTAAATCAGTGAAATTTTTCACCATTGCCTACAACTTATTGAGAAATAAGCAGCTTCCTCAGCCCTCAAAATAACAAAACCCACCT

9175-9281 TAAGGTGGGTTTCGCCAGAGAATTATCTCTGGTATTCAGAACGCCATTACCGGACTTTGCCTTGACCTTGCGATAATCGCAGGTTGCGGGATGTCTGAATTTCTTCA

9282-9388 GTCTGCTGCATCCTGGAAGATGAGAACATGTGTTCTTATTTTCGTCTCTATCATAGTTGAGTATTTACTCTCTTACAATCAGATCTCTTTCATTGCTCAACAGGCGA

9389-9450 TGGCTTCAGACTTTGCATTACGGAATTTTTAAGAAAGGCAGGGCGAAACGAGGAAGAAGCTT

Fig. 3. The nucleotide sequence of the rpoBC operon segment containing the total structural rpoC gene and the total amino acid sequence of the β'-subunit of E.coli RNA polymerase. The nucleotide sequence of the complementary DNA chain, equivalent to the sequence of mRNA, is given. Numeration of nucleotides corresponds to the complete rpoBC operon. The restriction EcoRI cleavage sites dividing fragments EcoRI-G, EcoRI-D and EcoRI-A - HindIII are situated between nucleotides 4709-4710 and 7145-7146. The underlined amino acid sequences are those the structures of which have been determined from analysis of corresponding peptides. C* = 5-Methyl-cytosine. Inverted sequence repetitions entering into the proposed transcription terminator, are framed.

Table 1 Frequencies of codon usage in translation of rpoB and rpoC genes.

Amino acid	Codon	Frequency of codon usage	
		rpoC	rpoB
Arg	CGA	0	1
	CGC	24	28
	CGG	0	0
	CGU	75	61
	AGA	0	0
	AGG	0	0
Leu	CUA	0	0
	CUC	7	15
	CUG	125	99
	CUU	3	6
	UUA	3	1
	UUG	1	6
Ser	UCA	1	0
	UCC	27	31
	UCG	5	3
	UCU	24	23
	AGC	14	15
	AGU	0	2
Thr	ACA	1	3
	ACC	46	34
	ACG	7	6
	ACU	23	17
Pro	CCA	9	10
	CCC	0	0
	CCG	45	38
	CCU	3	8
Ala	GCA	33	22
	GCC	11	9
	GCG	52	29
	GCU	28	19
Gly	GGA	0	0
	GGC	28	35
	GGG	2	2
	GGU	85	69
Val	GUA	32	32
	GUC	7	14
	GUG	16	24
	GUU	58	41

Table 1 /continuation/

Amino acid	Codon	Frequency of codon usage	
		rpoC	rpoB
Lys	AAA	62	56
	AAG	25	24
Asn	AAC	47	48
	AAU	1	3
Gln	CAA	3	8
	CAG	47	50
His	CAC	17	18
	CAU	4	1
Glu	GAA	83	89
	GAG	26	33
Asp	GAC	47	61
	GAU	34	30
Tyr	UAC	27	29
	UAU	7	14
Cys	UGC	8	2
	UGU	7	5
Phe	UUC	26	33
	UUU	9	11
Ile	AUA	0	0
	AUC	75	66
	AUU	17	18
Met	AUG	35	37
	GUG	1	0
Trp	UGG	9	4

an EcoRI-A-HindIII fragment. After isolation of the fragment and an EcoRI-D fragment from the second recombinant plasmid, each of them was split with several restriction endonucleases listed in Fig 2. The resulting mixtures of comparatively small subfragments were terminally labelled with ^{32}P using T4 polynucleotide kinase and γ-^{32}P ATP. Isolated individual subfragments were subjected to separation and then single strands were sequenced by the modified /12/ Maxam-Gilbert technique /13/. Practically the whole structure was determined from both complementary strands. The complete

structures of the fragments were deduced on the basis of overlapping of the subfragment sequences*).

For the protein structure analysis the isolated β'-subunit was cleaved with cyanogen bromide /10/. It was also digested with trypsin after the citraconic anhydride modification of lysine residues /14/. The determined peptide sequence comprises as much as 40% of the β'-subunit whole structure.

The N-terminal amino acid sequence Met-Lys-Asp-Leu-Leu-Lys-Phe-Leu of the β'-subunit was determined by means of an automatic sequencing technique /10/.

The determined primary structure of the nucleic acid and the protein is presented in Fig 3. The β'-subunit consists of 1407 amino acid residues, that corresponds to a molecular weight of 155,162,5. Its translation is initiated by GUG (4187-4189) codon. The Shine-Dalgarno sequence, GGAG (4176-4179), which is complementary to the 3'-end of 16S ribosomal RNA, is located at a distance of 9 nucleotides from the codon. Termination of the translation is brought about by the amber codon TAA.

It is noteworthy that in the interval 8433-8458 close to the terminating codon there is a thymidine cluster which is preceded by a sequence of hyphenated dyad symmetry. The structure of the site is strikingly similar to the known structures of the ρ-independent transcription terminator /15/. It was established recently /16/ that termination, in the case of the rpoBC operon transcription, really occurs close to the end of the rpoC gene. Therefore this sequence could possibly be the transcription terminator of the rpoBC operon.

The amino acid composition of the β-subunit is Asp 81, Asn 48, Thr 77, Ser 71, Glu 109, Gln 50, Pro 57, Gly 115, Ala 124, Cys 15, Val 108, Met 36, Ile 92, Leu 139, Tyr 34, Phe 35, His 21, Lys 87, Arg 99, Trp 9. Thus the β'-subunit is a basic protein containing 207 basic and only 190 acidic amino acids. In the polypeptide strands of the β'-as well as β-subunits one can observe regions of strong clustering of basic amino acids. In the '-subunit these sequences are located in intervals: 74-81, 213-222, 1167-1174 and 1366-1377 (Fig 3). It is of interest that the α-subunit does not

*) It should be noted that in the case of restriction endonuclease Taq I not all of the potential sites of splitting are effective. There is the sequence TCGATC (7308-7311 b.p.) in the fragment which could not be digested with the enzyme. Similar examples have already been described /2/. Methylation of A residues in this sequence has been proposed to be responsible for the effect.

contain similar clusters. Probably at least a few such sequences participate directly in the contacts between RNA polymerase and DNA, and/or RNA.

Table 1 represents the frequencies of the codon usage in the β'-subunit structure. They are analogous to those of the rpoB gene and other bacterial genes.

After publication of our data on the study of the structure of the EcoRI-HindIII fragment containing the C-terminal part of the rpoC gene /17/ a paper by Squires et al. /18/ appeared which also described the sequence of a considerable part of the fragment. The comparison of these two structures shows some differences and among them deletion of A (7584) and insertion of T between C (7717) and A (7718) are the most important. They lead to a complete change of the amino acid sequence between Lys (1132) and Thr (1178). It should be noted that the analysis of the corresponding peptides confirmed our data.

Another difference, leading to substitution of Val (1384) in the presented sequence for Gly in the sequence determined by Squires et al., is an inversion of dinucleotide TG (8337-8338). The other differences concern the residues situated after the translation terminator.

Determination of the nucleotide sequence of the rpoC gene and amino acid sequence of the E.coli RNA polymerase β'-subunit completes the analysis of the core-enzyme primary structure.

REFERENCES

1. Ovchinnikov, Yu.A., Lipkin, V.M., Modyanov, N.N., Chertov, O.Yu, Smirnov, Yu.V. (1977) FEBS Lett., 76, 108-111.
2. Ovchinnikov, Yu.A., Monastyrskaya, G.S., Gubanov, V.V., Guryev, S.O., Chertov, O.Yu., Modyanov, N.N., Grinkevich, V.A., Makarova, I.A., Marchenko, T.V., Polovnikova, I.N., Lipkin, V.M., Sverdlov, E.D. (1980) Eur. J. Biochem., 116, 621-629.
3. Monastyrskaya, G.S., Guryev, S.O., Kalinina, N.F., Sorokin, A.V., Salomatina, I.S., Shuvaeva, T.M., Lipkin, V.M., Sverdlov, E.D. Ovchinnikov, Yu.A. (1982) Bioorgan.Khim., 8, 130-134.
4. Burton, Z., Burgess, R.R., Lin, J., Moore, D., Holder, S., Gross, C.A., (1981) Nucleic Acids Res., 9, 2889-2898.
5. Bickle, T.A., Pirotta, V., Imber, R. (1977), Nucleic Acids Res., 5, 2561-2572.
6. Richardson, C.C. (1971) in Procedures in Nucleic Acids Research (Cantoni, G.L. and Davies, D.R., Eds) vol. 2, pp.815-828, Harper and Row, New York.
7. Collins, J. (1979) Mol. Gen. Genet., 173, 217-220.
8. McDonell, M.W., Simon, M.N., Studier, F.W. (1977) J.Mol. Biol., 110, 119-146.
9. Bolivar F., Rodriguez, R.L., Betlach, M.C., Boyer, H.W. (1977) Gene, 2, 75-93.

10. Shuvaeva, T.M., Lipkin, V.M., Nazimov, I.V., Modyanov, N.N., Ovchinnikov, Yu.A. (1981), Bioorgan. Khim., 7, 1765-1777.
11. Lipkin, V.M., Marchenko, T.V., Khokhryakov, V.S., Polovnikova, I.N., Potapenko, N.A., Modyanov, N.N., Ovchinnikov, Yu.A. (1980) Bioorgan. Khim., 6, 332-347.
12. Ovchinnikov, Yu.A., Guryev, S.O., Krayev, A.S. Monastyrskaya, G.S., Skryabin, K.G., Sverdlov, E.D., Zakharyev, V.M., Bayev, A.A., (1979) Gene, 6, 235-249.
13. Maxam, A., Gilbert, W., (1977) Proc. Nat.Acad.Sci., USA, 74, 560-564.
14. Atassi, M.Z., Habub, R.F.S.A., Methods in Enzymology, S.P., Golovick, N.O., Kaplan, Eds. N.Y, London: Academic Press, 1972, vol. 25B, p. 546-553.
15. Rosenberg, M., Court, D. (1979) Ann.Rev. Genet., 13, 319-353.
16. An, G., Friesen, J.D. (1980), J. Bacteriol., 144, 904-916.
17. Ovchinnikov, Yu. A., Monastyrskaya, G.S., Gubanov, V.V., Salomatina, I.S. Shuvaeva, T.M., Lipkin, V.M., Sverdlov, E.D. (1981) Bioorgan. Khim., 7, 1107-1112.
18. Squires, C., Krainer, A., Barry, G., Shen, W.F., Squires, C.L. (1981) Nucleic Acids Res., 9, 6827-6840.