JOURNAL OF BACTERIOLOGY, Jan. 1989, p. 198–204
0021-9193/89/010198-07$02.00/0

Vol. 171, No. 1

Identification and Sequence Analysis of *Escherichia coli purE* and *purK* Genes Encoding 5′-Phosphoribosyl-5-Amino-4-Imidazole Carboxylase for De Novo Purine Biosynthesis†

WAKAKO WATANABE, GEN-ICHI SAMPEI, ATSU AIBA, AND KIYOSHI MIZOBUCHI*

Department of Biophysics and Biochemistry, Faculty of Science, The University of Tokyo, Hongo, Tokyo 113, Japan

Received 11 July 1988/Accepted 21 September 1988

It has been shown that the *Escherichia coli purE* locus specifying 5′-phosphoribosyl-5-amino-4-imidazole carboxylase in de novo purine nucleotide synthesis is divided into two cistrons. We cloned and determined a 2,449-nucleotide sequence including the *purE* locus. This sequence contains two overlapped open reading frames, ORF-18 and ORF-39, encoding proteins with molecular weights of 18,000 and 39,000, respectively. The *purE* mutations of CSH57A and DCSP22 were complemented by plasmids carrying ORF-18, while that of NK6051 was complemented by plasmids carrying ORF-39. Thus, the *purE* locus consists of two distinct genes, designated *purE* and *purK* for ORF-18 and ORF-39, respectively. These genes constitute a single operon. A highly conserved 16-nucleotide sequence, termed the PUR box, was found in the upstream region of *purE* by comparing the sequences of the *purF* and *purMN* operons. We also found three entire and one partial repetitive extragenic palindromic (REP) sequences in the downstream region of *purK*. Roles of the PUR box and REP sequences are discussed in relation to the genesis of the *purEK* operon.

5′-Phosphoribosyl-5-amino-4-imidazole (AIR) carboxylase (EC 4.1.1.21) catalyzes the conversion of AIR to carboxyl AIR in de novo purine biosynthesis (17). In *Escherichia coli*, AIR carboxylase is encoded by the *purE* locus, which is located at 12 min on the chromosome (1). Genetic studies showed that the *purE* locus was divided into two complementation groups, *purE1* and *purE2* (8, 12). However, it is not clear whether these two cistrons specify two distinct polypeptides or two functional domains in a single polypeptide.

Hamilton and Reeve (9, 10) determined the nucleotide sequences of DNA fragments from *Methanobrevibacter smithii* and *Methanobacterium thermoautotrophicum* that were able to complement both *purE1* and *purE2* mutations of *E. coli*. These sequences encoded a single polypeptide chain whose structure appeared to arise from the fusion of tandemly duplicated polypeptide chains. They also reported that a small deletion that occurred in the 3′ domain of the *M. thermoautotrophicum* gene did not complement either the *purE1* or the *purE2* mutation (10). On the other hand, sequence analysis of a 12-purine gene cluster from *Bacillus subtilis* showed that two distinct genes, *purE* and *purK*, were responsible for the activity of AIR carboxylase (5). These data suggest a considerable variation in organization of the genes and operons of AIR carboxylase from one organism to another.

The *E. coli* DNA fragment that complements *purE1* and *purE2* mutations has been cloned (13; J. M. Smith, cited in reference 5) and sequenced (J. M. Smith, cited in reference 5). However, the details of the sequencing have not been reported. To explore the organization of the *E. coli purE* locus, we also cloned independently the chromosomal fragment of this region and determined the nucleotide sequence. In this report, we demonstrate that the *E. coli purE* locus consists of two overlapped genes in a single operon. We designate these genes *purE* and *purK*, as proposed for the genes in *B. subtilis*. We also describe several characteristic sequences existing in the *purEK* locus. One characteristic sequence, which is 16 nucleotides long and resides in the upstream region of *purE*, is highly conserved in several *E. coli pur* operons and is therefore designated the PUR box. Three repetitive palindromic sequences 35 to 38 nucleotides long were found to be present in the 3′ flanking region of *purK*. Such sequences have been frequently observed in functionally unrelated operons of *E. coli* and *Salmonella typhimurium* (7, 30). On the basis of these observations, we discuss possible roles of the PUR box and of repetitive sequences in relation to the genesis of the *purEK* locus.

MATERIALS AND METHODS

Bacterial strains, plasmids, and media. *E. coli* NK6051 [*purE1*::Tn*10 thi* Δ(*pro-lac*)XIII *spoT relA*] (10, 13), CSH57A (*purE pro ara leu lacY gal trp his argG malA rpsL xyl mtl ilv met thi*) (23), and DCSP22 (*purE argG leu his uvrA rpsL*), obtained from the Stock Center of the National Institute of Genetics, Japan, were used as *purE* mutants. Strain MC4100 (2) carrying plasmid pPE184 was employed for the determination of the 5′ end of the *purEK* operon mRNA. Strain JM103 (22) was used as a host for the subcloning of DNA fragments containing *purE*. Strain MV1184 (35) was employed for the preparation of single stranded DNA (ssDNA) from a series of deletion plasmids. Strain PR13 (24) was used for the preparation of fraction I in in vitro protein synthesis (6). Plasmid pLC8-25, from the library of Clarke and Carbon (3), was used as a source of the wild-type *purE* locus. Plasmids pTZ18R and pTZ19R (20) were used as vectors for the subcloning of DNA fragments and preparation of ssDNA. Plasmids pPE181 and pPE182 contained the same 2.9-kilobase (kb) *Bgl*II restriction fragment of pLC8-25 in the *Bam*HI site of pTZ18R but in opposite orientations (see Fig. 1). Plasmid pPE184 was constructed from pPE181 by the removal of a 0.45-kb *Pst*I restriction fragment. Plasmid pPE183 was prepared by the

* Corresponding author.

† Publication of this paper was delayed to permit publication with the paper that immediately follows.

subcloning of a 2.5-kb *Kpn*I-*Bgl*II restriction fragment of pLC8-25 in the *Kpn*I and *Bam*HI sites of pTZ18R. LB medium (16), M9 minimal medium (23), and their agar media were used for growth of bacteria. 2 × YT (21) was used for the growth of MV1184 cells in the preparation of ssDNA. The adenine requirement for bacterial growth was tested on M9 agar medium supplemented with thiamine (1 μg/ml) and required amino acids (30 μg of each per ml) in the presence or absence of 10 μg of adenine per ml. The CO_2-conditional phenotype of the *purE2* mutations (8) was determined by plating bacteria onto M9 agar medium supplemented with thiamine and required amino acids in the presence or absence of 5% CO_2. Ampicillin was added to 100 or 150 μg/ml if necessary.

DNA sequence analysis. After complete digestion of plasmids pPE182 and pPE184 with restriction enzymes *Sac*I and *Sma*I, DNA fragments were further digested with exonuclease III followed by mung bean nuclease digestion by the method of Henikoff (11) and were religated for the construction of a series of deletion plasmids. Thirty-seven deletions, twenty from pPE182 and seventeen from pPE184, were isolated. These deletions covered a 2.5-kb region in overlapping. After each deletion was introduced into MV1184 cells, ssDNA was isolated by using infecting bacteriophage M13K07 as described previously (35). Analysis of nucleotide sequences was performed by the dideoxy chain-termination method of Sanger et al. (25) by using a reverse primer (5'-CAGGAAACAGCTATGAC-3') (Takara Shuzo Co.) and [α-^{32}P]dCTP (400 Ci/mmol, Amersham).

Complementation tests. Plasmid DNA containing the entire *purE* locus or only a partial *purE* locus was introduced into the recipients of *purE* mutant cells, and ampicillin-resistant transformants were selected. After single-colony isolation on an LB medium plate containing ampicillin, the adenine requirement of the transformants was tested by streaking on M9 medium without adenine or supplemented with adenine.

Protein synthesis. Polypeptides directed by plasmid DNA were analyzed by the in vitro coupled transcription-translation method (29). Cell extract (fraction I) was prepared from the *E. coli* PR13 cells (6), and the reaction was carried out in the presence of [^{35}S]methionine (1,100 Ci/mmol; Amersham) as described previously (19). The translated products were separated by electrophoresis on a sodium dodecyl sulfate (SDS)–12.5% polyacrylamide gel (15) and subjected to autoradiography.

Primer extension experiment. A 1/100 volume of an overnight culture of MC4100 cells carrying pPE184 grown in M9 medium (supplemented with 0.2% Casamino Acids [Difco] and 1 μg of thiamine per ml in the presence or absence of 30 μg of adenine or guanine per ml) was inoculated into the same (fresh) medium and cultured at 37°C to an optical density at 650 nm (OD_{650}) of 0.6, at which time a 1/50 volume of the culture was reinoculated into the same (fresh) medium and cultured continuously to an OD_{650} of 0.4. The cells were harvested by centrifugation and resuspended in a small volume of a solution of 0.5% SDS, 1 mM disodium EDTA, and 20 mM sodium acetate. RNA was extracted twice with hot phenol saturated with 20 mM sodium acetate, pH 4.8, precipitated with ethanol, and suspended in water. After treatment with RNase-free bovine pancreatic DNase I (0.1 U/μg of RNA; Worthington) in the presence of 50 mM Tris hydrochloride (pH 7.5)–10 mM $MgCl_2$ at 37°C for 10 min, RNA was reextracted with phenol and then precipitated with ethanol. To determine the 5' end of the *purEK* operon mRNA, a mixture containing, in a final volume of 50 μl, 30 μg of RNA; 50 mM Tris hydrochloride, pH 8.2; 50 mM KCl; 10 mM $MgCl_2$; and 0.5 pmol of synthetic primer DNA (5'-CATGGTAGCCCAGTCGC-3') labeled at the 5' end with [γ-^{32}P]ATP (5,200 Ci/mmol; Amersham) by polynucleotide kinase was incubated at 80°C for 2 min and then at 60°C for 45 min for DNA-RNA hybridization. Then, four deoxynucleoside triphosphates (final concentration, 0.5 mM each) and 20 U of reverse transcriptase (RAV-2) were added to the mixture, and the primer extension reaction proceeded at 42°C for 1 h. The DNA synthesized was extracted with phenol-chloroform (1:1), precipitated with ethanol, and suspended in DNA-sequencing solution. DNA was analyzed electrophoretically by the method of Sanger et al. (25). As a standard of sequence ladder, 0.5 pmol of ssDNA fragment derived from pPE184 was hybridized with the ^{32}P-labeled primer mentioned above, and the reaction was carried out as described in DNA sequence analysis, above.

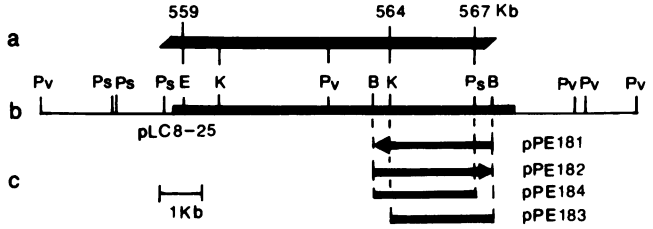

FIG. 1. Restriction maps of the chromosomal insert DNA of plasmid pLC8-25 and of the *E. coli* chromosome. (a) Restriction map of the *E. coli* chromosome at position 559 to 567 kb (14). (b) Restriction map of pLC8-25. The thin line represents ColE1 DNA. (c) DNA fragments of complementation-positive clones. Arrows labeled pPE181 and pPE182 represent the opposite orientation of DNA fragments. B, *Bgl*II; E, *Eco*RI; K, *Kpn*I; Ps, *Pst*I; and Pv, *Pvu*II.

Enzymes. All enzymes except bovine pancreatic DNase I were obtained from Takara Shuzo Co. and used as specified by the manufacturer.

RESULTS

Cloning of the *purE* locus. Plasmid pLC8-25 from the Clarke and Carbon library has been reported to complement *purE* mutations (4, 13). We constructed a restriction map of pLC8-25 for restriction enzymes *Bgl*II, *Eco*RI, *Kpn*I, *Pst*I, and *Pvu*II (Fig. 1). The total size of pLC8-25 was determined to be 14.9 kb, which reflects a chromosomal insert DNA with a size of 8.3 kb. To more precisely localize the *purE* locus, three DNA fragments, *Bgl*II-*Bgl*II, *Bgl*II-*Kpn*I, and *Bgl*II-*Pst*I, were subcloned into pTZ18R. The resulting plasmids, pPE181, pPE182, pPE183, and pPE184, were independently introduced into three *purE* mutant cells, NK6051, CSH57A, and DCSP22. All the plasmids enable the three mutant cells to transform to Pur$^+$, indicating that the *purE* locus locates within a 2.1-kb fragment between the *Kpn*I and *Pst*I sites. These observations were consistent with an earlier result of the cloning of the *purE* locus (13). When the restriction map of the insert DNA of pLC8-25 was compared with that of the *E. coli* chromosome (14), it corresponded exactly to position 559 to 567 kb of the *E. coli* chromosome, and consequently, the *purE* locus is mapped around 566 kb.

Nucleotide sequence. The strategy of DNA sequence analysis of the *purE* region is outlined in Fig. 2. A 2,449-nucleotide sequence from the *Bgl*II site was determined (Fig. 3). Both strands corresponding to this sequence contain nine possible open reading frames (ORFs) which span more than 200 nucleotides.

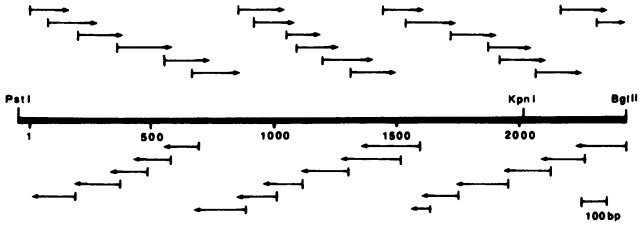

FIG. 2. Sequencing strategy for the *purE* locus. The nucleotide sequence was analyzed by using a series of deletion plasmids and by the method of Sanger et al. (25). Arrows denote the length and direction of sequencing.

As described below, the results of complementation tests and in vitro protein synthesis indicated that two ORFs, ORF-18 and ORF-39, corresponded to the AIR carboxylase genes. ORF-18 locates at position 393 to 899 nucleotides and encodes a protein with the expected molecular weight of 17,780. ORF-18 possesses a potential ribosome-binding site (26), AGGAG, at position 377 to 381 nucleotides. ORF-39 spans from nucleotides 899 to 1963 and overlaps 1 base pair with ORF-18. ORF-39 encodes a protein with the expected molecular weight of 39,460 and possesses a potential ribosome-binding site, GAGGTA, at position 889 to 894 nucleotides. These ORFs are transcribed in the same direction.

Of interest is the finding of three complete and one incomplete repetitive palindromic sequences containing 35 to 38 nucleotides in the 3′ flanking region of ORF-39 (Fig. 3). Such sequences have been frequently observed between genes of a multigenic operon or between functionally unrelated operons in *E. coli* and *S. typhimurium* and termed the repetitive extragenic palindromic (REP) sequence (7, 30). The clustering of the REP sequences in the 3′ flanking region of ORF-39 can form two alternative, stable stem and loop structures of RNA with ΔG's = −109 and −87.7 kcal/mol, respectively, as shown in Fig. 4.

Complementation tests. To establish that both ORF-18 and ORF-39 were responsible for the AIR carboxylase genes, complementation tests were performed between the three chromosomal *purE* mutants (CSH57A, DCSP22, and NK6051) and some deletion plasmids employed for the DNA sequence analysis. A summary of the experiments is given in Fig. 5. Plasmids pPE404, pPE241R, pPE241, and pPE252, those of which contained entire ORF-18 and ORF-39, enabled the three *purE* mutant cells to transform to Pur^+ regardless of the orientation of the chromosomal insert DNA to vector pTZ18R, as pPE184 did. Plasmids pPE415 and pPE428, both of which contained intact ORF-18 but lacked ORF-39 and in which the transcription direction of ORF-18 was opposite to that of the *lacZ′* gene, were capable of complementing CSH57A and DCSP22 cells but not NK6051 cells.

On the other hand, both pPE205 and pPE249, which carried the entire ORF-39 and its potential ribosome-binding site but had partial deletion in ORF-18, were able to com-

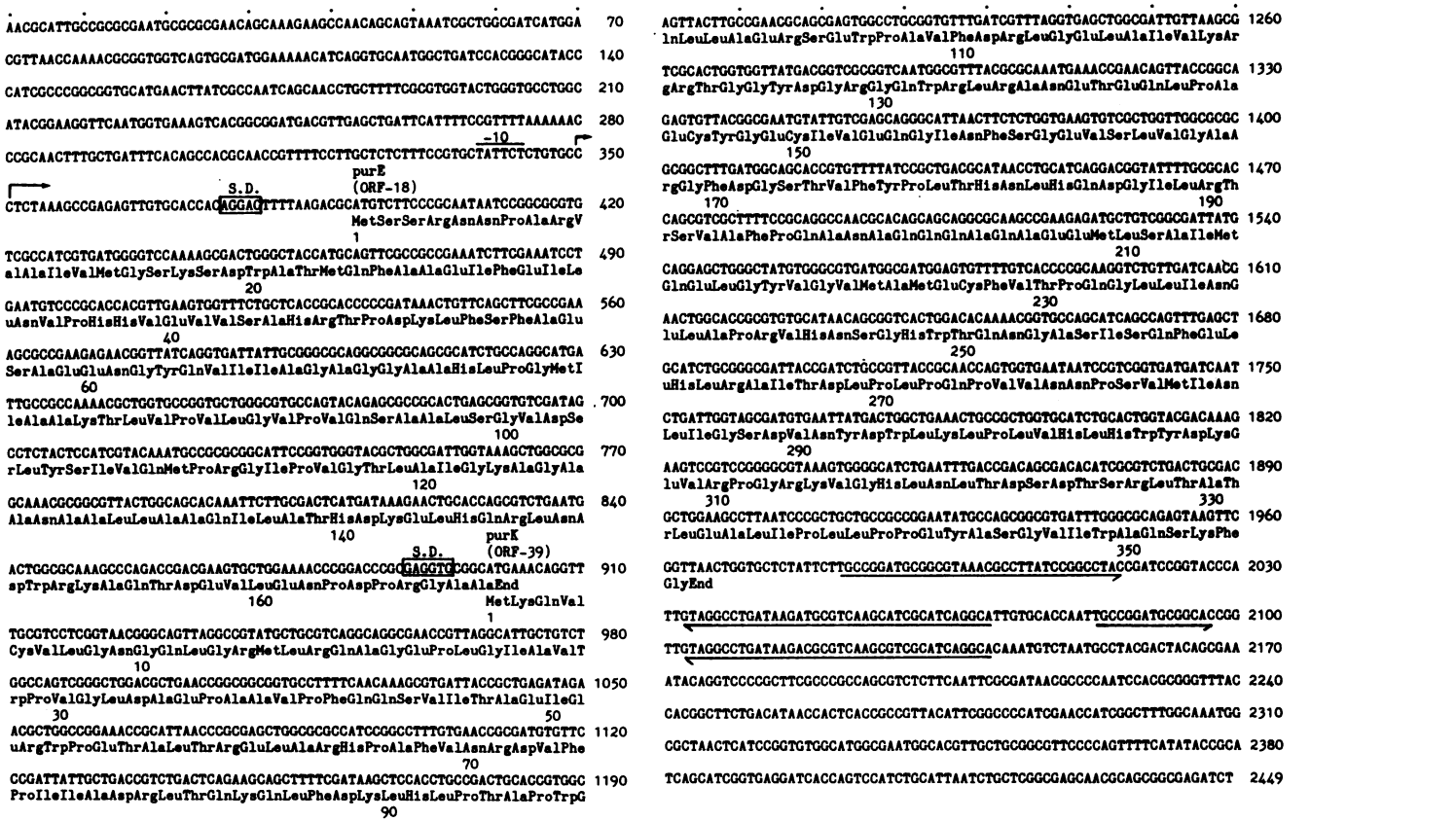

FIG. 3. Nucleotide and amino acid sequences of the *purE* locus. The DNA sequences of the sense strands of *purE* (=ORF-18) and *purK* (=ORF-39) are shown. The potential ribosome-binding sites for *purE* and *purK* are boxed. The −10 region is overlined. Arrows denote transcription initiation sites of the *purEK* operon as determined by the primer extension experiments (see Fig. 7). Three entire and one partial REP sequences are underlined.

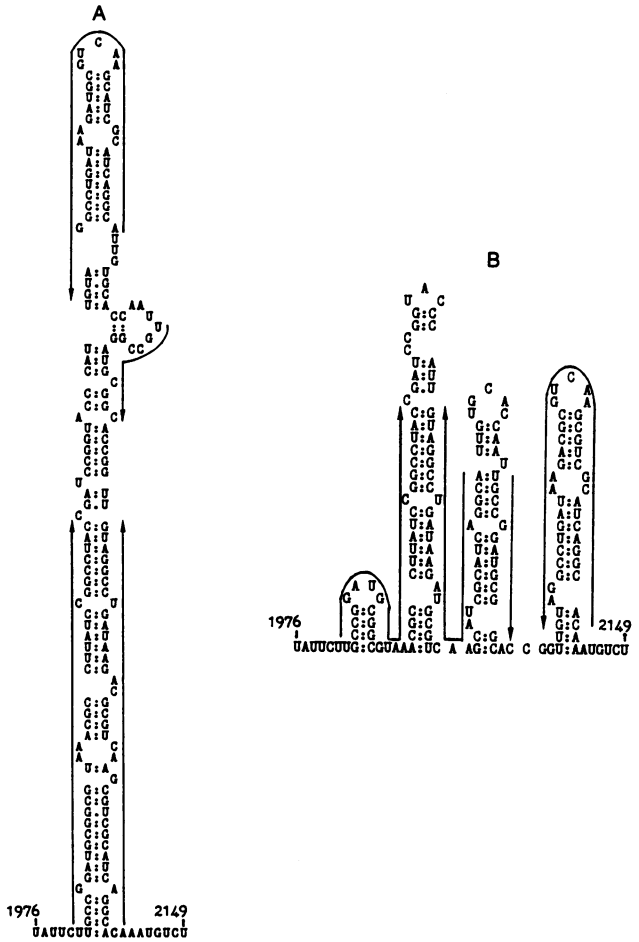

FIG. 4. Predicted secondary structures of mRNA at the REP region of the *purEK* operon. Two alternative, stable secondary structures (A and B) of mRNA corresponding to the 3′ flanking region of *purK* (=ORF-39) were predicted. Arrows denote the REP sequences. Free energies of the structures, calculated by the method of Tinoco et al. (33), are −109 (A) and −87.7 (B) kcal/mol.

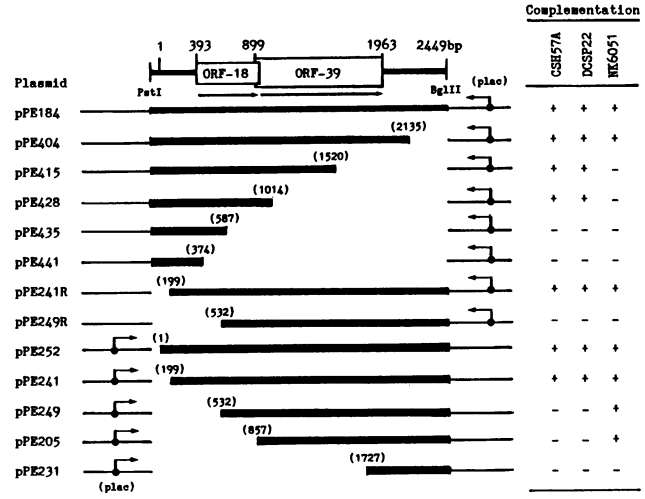

FIG. 5. Complementation analysis of *purE* mutations. Designations of plasmids used in complementation tests are listed on the left side. Plasmids pPE241R and pPE249R were constructed by the subcloning of the *Hin*dIII-*Eco*RI fragments of pPE241 and pPE249 into the *Hin*dIII and *Eco*RI sites, respectively, of pTZ19R. The respective *Hin*dIII-*Eco*RI fragments contain the chromosomal insert DNA of pPE241 and pPE249 and 35 extra nucleotides, 5 to the *Pst*I site and 30 to the *Bgl*II site, from the polylinker of pTZ18R. The other plasmids were constructed by using vector pTZ18R. Thick and thin lines represent DNA of the chromosomal insert and of vector pTZ18R or pTZ19R, respectively. ORF-18 (=*purE*) and ORF-39 (=*purK*) are represented as boxes. Numbers in parentheses represent the ends of nucleotides of the chromosomal insert DNA of deletions. The *Pst*I site locates at about nucleotide −50. Arrows show the transcription direction of *purE*, *purK*, and *lacZ′* genes, respectively. + and −, Positive and negative, respectively, of complementation tests.

plement only NK6051 cells. However, when the insert DNA of pPE249 was recloned into pTZ19R, the resulting plasmid, pPE249R, did not complement NK6051 cells. Note that the transcription direction of ORF-39 of pPE249R was opposite to that of the *lacZ′* gene. These results imply that the complementation of NK6051 by pPE249 and pPE205 may be due to the readthrough expression of ORF-39 from the *lacZ′* gene. Deletion mutants of both ORF-18 and ORF-39 as shown in pPE435, pPE441, and pPE231 did not complement the three *purE* mutations.

These results together indicate that both ORF-18 and ORF-39 are responsible for de novo purine biosynthesis. Furthermore, the lack of complementation of NK6051 by pPE249R shows that the expression of ORF-39 depends on the upstream region of ORF-18. Therefore, it can be concluded that these two ORFs constitute a single operon.

It has been shown that the *purE* locus is divided into two cistrons, *purE1* and *purE2*, and that they map contiguously on the *E. coli* chromosome (12, 13). Mutant NK6051 has been described to have a *purE1*::Tn*10* insertion with a strong polar effect on *purE2* (13). However, this mutant can grow in the presence of increased CO_2 concentrations without supplementation with adenine, indicating that the mutation has a *purE2* phenotype. Conversely, mutants CSH57A and DCSP22 can grow only when adenine is added to the medium. These lines of information and the results of complementation tests indicate that *purE1* and *purE2* correspond to ORF-18 and ORF-39, respectively.

In vitro protein synthesis. The products specified by ORF-18 and ORF-39 were investigated by the in vitro coupled transcription-translation method (29). DNAs of pPE184 and pPE205 were used as templates. Figure 6 shows the patterns of protein synthesis. The insert DNA of pPE184 directed two specific proteins with molecular weights of 18,000 and 39,000 (Fig. 6, lane 4). These values were consistent with those expected from ORF-18 and ORF-39. When pPE205 was used, two specific proteins, one a major product and one a minor product, with molecular weights of 42,000 and 39,000 were synthesized (lane 3). The analysis of the nucleotide sequence of pPE205 indicated that the synthesis of the major, 42,000-molecular-weight polypeptide resulted from the fusion of *lacZ′* derived from vector pTZ18R with the insert DNA at nucleotide 857. Such fusion caused the addition of 29 extra amino acid residues, 14 from *lacZ′* and 15 from the upstream sequence of ORF-39, to the NH_2 terminus of the ORF-39 product. The translation of the fused gene gives the expected molecular weight of 42,548, a value consistent with the observed value on SDS-polyacrylamide gels. On the other hand, the minor, 39,000-molecular-weight polypeptide seemed to be synthesized from the initiation codon of ORF-39.

The identification of the two products indicated that ORF-18 and ORF-39 indeed encoded the respective pro-

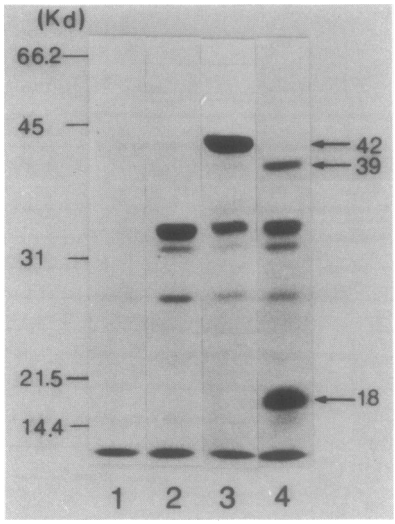

FIG. 6. Autoradiogram of an SDS–12.5% polyacrylamide gel of in vitro-synthesized proteins. Proteins directed by plasmid DNA were synthesized by the in vitro coupled transcription-translation method (29). Lanes: 1, no DNA; 2, pTZ18R; 3, pPE205; and 4, pPE184. (See Fig. 5 for the structures of pPE205 and pPE184). As standard markers, lysozyme, soybean trypsin inhibitor, carbonic anhydrase, ovalbumin, and bovine serum albumin were used.

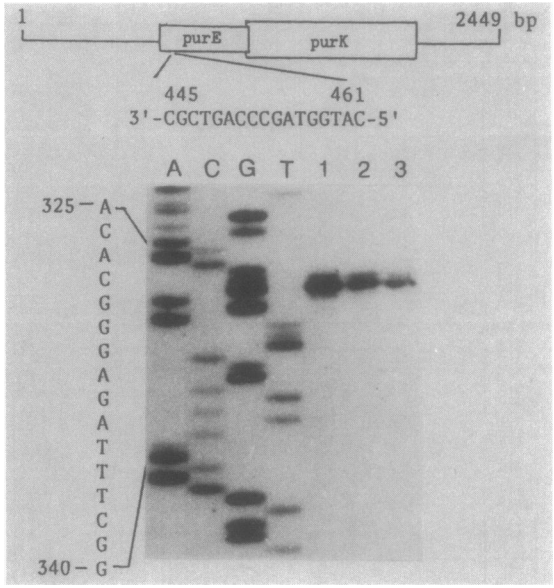

FIG. 7. Primer extension mapping of the 5′ end of the *purEK* operon mRNA. The top of the figure is a schematic illustration of the *purEK* operon and the position and sequence of a synthetic DNA fragment used as a primer. RNA isolated from MC4100 cells carrying pPE184 was hybridized with a synthetic 17-nucleotide primer whose 5′ end was labeled with [γ-^{32}P]ATP by polynucleotide kinase. DNA was synthesized by the reaction of reverse transcriptase. Lanes: 1, RNA from cells grown in M9 medium; 2 and 3, RNA from cells grown in M9 medium supplemented with adenine (lane 2) or guanine (lane 3); A, C, T, and G, a standard of sequence ladder using the same synthetic primer and its complementary ssDNA from pPE184. The nucleotide sequence shown in the figure is complementary to that shown in Fig. 3.

teins. Furthermore, when the deduced amino acid sequences of ORF-18 and ORF-39 were compared with those of the *purE* and *purK* products from *B. subtilis* recently reported (5), 59 and 32% identities were found between ORF-18 and *purE* products and between ORF-39 and *purK* products, respectively. Thus, we designated ORF-18 (=*purE1*) and ORF-39 (=*purE2*) *purE* and *purK*, respectively.

Transcription initiation site of the *purEK* operon. The primer extension method was employed to determine the 5′ end of the *purEK* operon mRNA. RNA was isolated from MC4100 cells carrying pPE184 grown in the presence (repressed condition) or absence (derepressed condition) of adenine or guanine and hybridized with a synthetic primer that corresponded to position 445 to 461 nucleotides (see Fig. 3). DNA was synthesized from this primer by using reverse transcriptase. The autoradiograms of the DNA synthesized are shown in Fig. 7. Under the experimental conditions, two identical DNA bands were detected for all samples. The intensity of DNA bands in the repressed condition was apparently lower than that in the derepressed condition. If this intensity of the DNA bands reflected the extent of the gene expression at the transcriptional level, the addition of purines to the culture medium repressed *purEK* operon expression to approximately 70 to 80% of the level of the derepressed condition (Fig. 7, lanes 2 and 3). Note that this value is comparable to that reported by Kamholz et al. (13).

Comparison of the DNA bands with a DNA sequencing ladder as a standard showed that the 5′ end of the *purEK* operon mRNA corresponded to position 350 (minor) or 351 (major). This indicated that if the 5′ region of the *purEK* mRNA was not processed, the transcription initiation occurred 41 or 42 base pairs upstream from the start codon of the *purE* structural gene. Within this region the sequence TATTCT at position 338 to 343 nucleotides showed resemblance to the consensus sequence for the −10 region. However, the −35 region, TTTTCC at position 315 to 320 nucleotides, did not resemble the consensus sequence for the −35 region. These results and mapping of the *purEK* operon on the *E. coli* restriction map demonstrate that the *purEK* operon is transcribed counterclockwise on the chromosome.

The PUR box. In *E. coli* at least 13 de novo purine biosynthetic genes, including *purK*, have been mapped at nine different loci of the chromosome (1). To date, five genes, *purF*, *purM*, *purN*, *guaA*, and *guaB*, have been sequenced (27, 28, 31, 32, 34). Smith and Daum (27) reported the presence of a conserved sequence (33 of 39 base pairs) in the upstream region of *purF* (18) and *purMN* (27) operons. Comparison of the upstream sequence of the *purEK* operon with those of the *purF* and *purMN* operons showed that

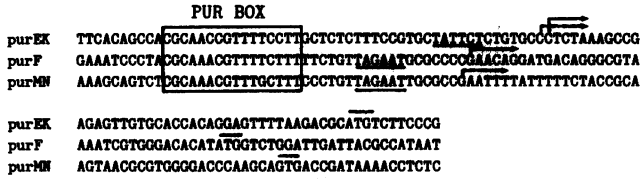

FIG. 8. Comparison of the 5′ flanking region of the *purEK*, *purF*, and *purMN* operons. The nucleotide sequence of the 5′ flanking region of the *purEK* operon is from Fig. 3. The sequences of *purF* and *purMN* operons are from the data of Makaroff and Zalkin (18) and Smith and Daum (27), respectively. Conserved sequences, termed the PUR box, are boxed. Arrows represent the respective transcription initiation sites. The −10 regions are underlined. The respective translation initiation codons are overlined.

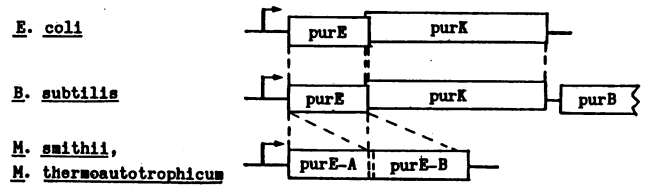

FIG. 9. Schematic illustration of the organization of genes and operons of AIR carboxylase from *E. coli*, *B. subtilis*, *M. smithii*, and *M. thermoautotrophicum*. Genes controlling AIR carboxylase from *B. subtilis* are parts of a 12-purine gene cluster (5). AIR carboxylase genes from *M. smithii* and *M. thermoautotrophicum* (10) have a structure of duplication of the *purE* gene of *E. coli* and *B. subtilis*. Homologous genes between organisms are connected with dashed lines.

there were highly conserved sequences at least 16 nucleotides in length (Fig. 8). These sequences span roughly the −35 region of the three purine operons and have a symmetrical structure. Additionally, we have also found the homologous sequences around the −35 region of the *purL* (Sampei et al., manuscript in preparation) and *purHD* (Aiba et al., manuscript in preparation) operons. Thus, we propose this type of sequence as a PUR box. The function of these highly conserved sequences is not known at present. However, since the *purEK*, *purF*, *purMN*, *purHD*, and *purL* loci were coregulated (8; Sampei et al., in preparation) and since the sequence is located around the −35 region, the PUR box may play an important role in the regulation of the expression of many purine genes at the transcriptional level.

DISCUSSION

The analysis of the *E. coli purE* locus at the nucleotide sequence level disclosed that *purE1* and *purE2* cistrons corresponded to two distinct genes, designated *purE* and *purK*, respectively. The results of complementation tests with pPE249 and pPE249R showed that these two genes constitute a single operon. This notion has been further supported by the preliminary observation that the cells of the prototroph *E. coli* K12W3110 produce one mRNA species with a size of 1.8 to 1.9 kilonucleotides that can be hybridized with a fragment from nucleotides 357 to 2027 but not with a fragment from nucleotides 2118 to 2444. Such mRNA appears to be large enough to encode simultaneously both *purE* and *purK* products.

On the basis of the results of the primer extension experiments, the 5′ end of the *purEK* operon mRNA was determined to be C at position 350 or 351. This information and the result described above indicate that the transcription of the *purEK* operon must be terminated around position 2050. Within this region are present three entire and one partial repetitive palindromic sequences which may form stable stem and loop structures, as shown in Fig. 4. Such structures of RNA may be involved in the transcription termination of the *purEK* operon.

The studies of genes and operons of AIR carboxylase from different organisms have indicated considerable variation in organization, as shown schematically in Fig. 9. This contrasts with the unity of the reaction catalyzed by the enzyme in de novo purine biosynthesis. In *E. coli* and *B. subtilis*, AIR carboxylase is encoded by *purE* and *purK*. Comparison of the deduced amino acid sequences of the *purE*-encoded proteins of *E. coli* and *B. subtilis* shows 59% identity. For *purK*-encoded proteins, the sequence identity is 32%. However, the organization of the operons is different; in *E. coli*, only *purE* and *purK* genes constitute a single operon, while in *B. subtilis* these two genes are parts of the 12-purine gene cluster (5).

In this regard, it may be worth noting that the genomes of *E. coli* and *S. typhimurium* contain many copies of the highly conserved REP sequences, as has been detected in the downstream region of *purK*. Previous studies indicated that the REP sequences might play a role in chromosomal rearrangement and genetic diversity in enteric bacteria (30). This view is attractive for consideration of the genesis of the *E. coli purEK* operon.

Assuming that the purine biosynthetic genes were organized as a cluster in a common ancestor of enteric bacteria, as has been observed in *B. subtilis*, the introduction of the REP sequence into the downstream region of the *purK* gene caused recombination between intergenic REP sequences. As a result of it, *purE* and *purK* genes were translocated from one site to another site to give rise to an operon separate from the rest of the *pur* genes. The translocation of genes required presumably the assurance of maintaining the regulatory mechanism for the gene expression to avoid the imbalance of purine metabolism. Such assurance was obtained by the generation of the PUR box, which may be important in the regulation of the expression of many purine genes. If the scenario described above is correct, it can be expected that both REP and PUR box sequences would be detected even in other purine operons because the *E. coli* purine biosynthetic operons are scattered among the different sites on the chromosome. In this context, it should be noted that the REP sequences are found in the upstream and downstream regions of the *purF* and *purL* operons, respectively, and that the PUR box is also present in these operons, as well as in the *purEK* operon.

AIR carboxylases from *Methanobrevibacter smithii* and *M. thermoautotrophicum* have been shown to be encoded by a single gene (9, 10). The analysis of the deduced amino acid sequences of these gene products indicates that the genes were generated by duplication followed by fusion of an ancestral *purE* gene. Of interest was the observation that a small deletion in the *M. thermoautotrophicum* gene did not complement either the *purE* (=*purE1*) or the *purK* (=*purE2*) mutation of *E. coli* (10). This information indicates that the single duplicated gene is equivalent to *purE* and *purK* in function, although a similar sequence cannot be detected between *purK* and the duplicated gene products.

Strains carrying mutations in the *purK* gene require purines at normal CO_2 concentrations but not at high CO_2 concentrations (8). Although the exact function of the *purK* gene product has not been elucidated yet, it has been suggested that the *purK* gene product helps provide the CO_2 substrate to the *purE* product for the catalysis of the conversion of AIR to carboxyl AIR. If that is the case, one interpretation of the result of complementation of the *E. coli purK* mutation by the *M. thermoautotrophicum* gene is that the AIR carboxylase of methanogenic bacteria evolved to utilize CO_2 efficiently even at low concentrations through duplication and diversity of the *purE*-encoded protein. A similar discussion has been provided by the comparison of amino acid sequences of the *purE* gene products of *B. subtilis-E. coli* and *Methanobrevibacter smithii* (5).

ACKNOWLEDGMENTS

We thank A. Nishimura, T. Nagata, and T. Yura for the gift of *E. coli purE* mutants, T. Sako for the preparation of synthetic oligo-

nucleotide, K. Shiba for discussion, and C. Martin for critical reading of the manuscript.

This work was supported by a grant from the Ministry of Education, Science and Culture of Japan.

LITERATURE CITED

1. **Bachmann, B. J.** 1983. Linkage map of *Escherichia coli* K-12, edition 7. Microbiol. Rev. **47**:180–230.
2. **Casadaban, M.** 1976. Transposition and fusion of the *lac* genes to selected promoters in *Escherichia coli* using bacteriophage lambda and Mu. J. Mol. Biol. **104**:541–555.
3. **Clarke, L., and J. Carbon.** 1976. A colony bank containing synthetic ColE1 hybrid plasmids representative of the entire *E. coli* genome. Cell **9**:91–99.
4. **Clarke, L., and J. Carbon.** 1979. Selection of specific clones from colony banks by suppression or complementation tests. Methods Enzymol. **68**:396–408.
5. **Ebbole, D. J., and D. Zalkin.** 1987. Cloning and characterization of a 12-gene cluster from *Bacillus subtilis* encoding nine enzymes for *de novo* purine nucleotide synthesis. J. Biol. Chem. **262**:8274–8287.
6. **Fuller, R. S., J. M. Kaguni, and A. Kornberg.** 1981. Enzymatic replication of the origin of the *Escherichia coli* chromosome. Proc. Natl. Acad. Sci. USA **78**:7370–7374.
7. **Gilson, E., J.-M. Clement, D. Brutlag, and M. Hofnung.** 1984. A family of dispersed repetitive extragenic palindromic DNA sequences in *E. coli*. EMBO J. **3**:1417–1421.
8. **Gots, J. S., C. E. Benson, B. Jochimsen, and K. R. Koduri.** 1977. Microbial models and regulatory elements in the control of purine metabolism. CIBA Found. Symp. **48**:23–41.
9. **Hamilton, P. T., and J. N. Reeve.** 1985. Structure of genes and an insertion element in the methane producing archaebacterium *Methanobrevibacter smithii*. Mol. Gen. Genet. **200**:47–59.
10. **Hamilton, P. T., and J. N. Reeve.** 1985. Sequence divergence of an archaebacterial gene cloned from a mesophilic and a thermophilic methanogen. J. Mol. Evol. **22**:351–360.
11. **Henikoff, S.** 1984. Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. Gene **28**:351–359.
12. **Jacob, F., A. Ullmann, and J. Monod.** 1965. Délétion fusionnant l'opéron lactose et un operon purine chez *Escherichia coli*. J. Mol. Biol. **31**:704–719.
13. **Kamholz, J., J. Keyhani, and J. S. Gots.** 1986. Molecular cloning and characterization of the *purE* operon of *Escherichia coli*. Gene **44**:55–62.
14. **Kohara, Y., K. Akiyama, and K. Isono.** 1987. The physical map of the whole *E. coli* chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. Cell **50**: 495–508.
15. **Laemmli, U. K.** 1970. Cleavage of structural protein during the assembly of the head of bacteriophage T4. Nature (London) **227**:680–685.
16. **Lennox, E. S.** 1955. Transduction of linked genetic characters of the host by bacteriophage P1. Virology **1**:190–206.
17. **Lukens, L. N., and J. M. Buchanan.** 1959. Biosynthesis of the purines. XXIV. The enzymatic synthesis of 5-amino-1-ribosyl-4-imidazolecarboxylic acid 5′-phosphate from 5-amino-1-ribosylimidazole 5′-phosphate and carbon dioxide. J. Biol. Chem. **234**:1799–1805.
18. **Makaroff, C. A., and H. Zalkin.** 1985. Regulation of *Escherichia coli purF*. Analysis of the control region of a *pur* regulon gene. J. Biol. Chem. **260**:10378–10387.
19. **Masai, H., Y. Kajiro, and K. Arai.** 1983. Definition of *oriR*, the minimum DNA segment essential for initiation of R1 plasmid replication *in vitro*. Proc. Natl. Acad. Sci. USA **80**:6814–6818.
20. **Mead, D. A., E. Szezesna-Skorupa, and B. Kemper.** 1986. Single-stranded DNA blue T7 promoter plasmids: a versatile tandem promoter system for cloning and protein engineering. Protein Engineering **1**:67–74.
21. **Messing, J.** 1983. New M13 vectors for cloning. Methods Enzymol. **101**:20–78.
22. **Messing, J., and J. Vieira.** 1982. A new pair of M13 vectors for selecting either DNA strand of double-digest restriction fragments. Gene **19**:269–276.
23. **Miller, J. H.** 1972. Experiments in molecular genetics. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
24. **Reiner, A. M.** 1969. Characterization of polynucleotide phosphorylase mutants of *Escherichia coli*. J. Bacteriol. **97**:1437–1443.
25. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
26. **Shine, J., and L. Dalgarno.** 1974. The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. Proc. Natl. Acad. Sci. USA **71**:1342–1346.
27. **Smith, J. M., and H. A. Daum III.** 1986. Nucleotide sequence of the *purM* gene encoding 5′-phosphoribosyl-5-aminoimidazole synthetase of *Escherichia coli* K12. J. Biol. Chem. **261**:10632–10636.
28. **Smith, J. M., and H. A. Daum III.** 1987. Identification and nucleotide sequence of a gene encoding 5′-phosphoribosylglycinamide transformylase in *Escherichia coli* K12. J. Biol. Chem. **262**:10565–10569.
29. **Staudenbaur, W. L.** 1976. Replication of small plasmids in extracts of *Escherichia coli*. Mol. Gen. Genet. **145**:273–280.
30. **Stern, M. J., G. F.-L. Ames, N. H. Smith, E. C. Robinson, and C. F. Higgins.** 1984. Repetitive extragenic palindromic sequences: a major component of the bacterial genome. Cell **37**: 1015–1026.
31. **Tiedeman, A. A., and J. M. Smith.** 1985. Nucleotide sequence of the *guaB* locus encoding IMP dehydrogenase of *Escherichia coli* K12. Nucleic Acids Res. **13**:1303–1316.
32. **Tiedeman, A. A., J. M. Smith, and H. Zalkin.** 1985. Nucleotide sequence of the *guaA* gene encoding GMP synthetase of *Escherichia coli* K12. J. Biol. Chem. **260**:8676–8679.
33. **Tinoco, I., P. N. Borer, B. Dengler, M. D. Levin, and J. Graller.** 1973. Improved estimation of secondary structure in ribonucleic acids. Nature (London) New Biol. **246**:40–41.
34. **Tso, J. Y., H. Zalkin, M. van Cleemput, C. Yanofsky, and J. M. Smith.** 1982. Nucleotide sequence of *Escherichia coli purF* and deduced amino acid sequence of glutamine phosphoribosylpyrophosphate amidotransferase. J. Biol. Chem. **257**:3525–3531.
35. **Vieira, J., and J. Messing.** 1987. Production of single-stranded plasmid DNA. Methods Enzymol. **153**:3–11.