THE JOURNAL OF BIOLOGICAL CHEMISTRY
© 1986 by The American Society of Biological Chemists, Inc.

Vol. 261, No. 12, Issue of April 25, pp. 5568–5574, 1986
Printed in U.S.A.

Nucleotide Sequence of *Escherichia coli pyrG* Encoding CTP Synthetase*

(Received for publication, October 28, 1985)

Manli Weng, Christopher A. Makaroff‡, and Howard Zalkin

From the Department of Biochemistry, Purdue University, West Lafayette, Indiana 47907

The amino acid sequence of *Escherichia coli* CTP synthetase was derived from the nucleotide sequence of *pyrG*. The derived amino acid sequence, confirmed at the N terminus by protein sequencing, predicts a subunit of 544 amino acids having a calculated M_r of 60,300 after removal of the initiator methionine. A glutamine amide transfer domain was identified which extends from approximately amino acid residue 300 to the C terminus of the molecule. The CTP synthetase glutamine amide transfer domain contains three conserved regions similar to those in GMP synthetase, anthranilate synthase, *p*-aminobenzoate synthase, and carbamoyl-P synthetase. The CTP synthetase structure supports a model for gene fusion of a *trpG*-related glutamine amide transfer domain to a primitive NH_3-dependent CTP synthetase. The major 5′ end of *pyrG* mRNA was localized to a position approximately 48 base pairs upstream of the translation initiation codon. Translation of the gene *eno*, encoding enolase, is initiated 89 base pairs downstream of *pyrG*. The *pyrG-eno* junction is characterized by multiple mRNA species which are ascribed to monocistronic *pyrG* and/or *eno* mRNAs and a *pyrG eno* polycistronic mRNA.

CTP synthetase is a glutamine amidotransferase that catalyzes the terminal reaction in the *de novo* pathway for pyrimidine nucleotide synthesis: UTP + ATP + glutamine → CTP + ADP + P_i + glutamate. Similar to other glutamine amidotransferases (1), NH_3 can replace glutamine in which case the products are CTP, ADP, and P_i. *Escherichia coli* CTP synthetase is a complex regulatory enzyme that exhibits both positive and negative cooperative effects and is subject to allosteric activation by GTP (2). Activation by GTP results from an increased rate of formation of a covalent glutaminyl enzyme catalytic intermediate. Accordingly, the NH_3-dependent reaction is unaffected by GTP. There is no evidence for regulation of the expression of the gene *pyrG* which encodes CTP synthetase in *E. coli* (3).

The primary structure required for glutamine amide transfer has been obtained for a number of amidotransferases (4–10). Recent sequence analyses of cloned amidotransferase genes have established the existence of two families of glutamine amide transfer domains, each of approximately 190–200 amino acids (8–12). We have employed site-directed mutagenesis to investigate the role of amino acid residues that are important for glutamine amide transfer function (13–15). In the absence of crystallographic data, identification of conserved amino acids in a homologous domain from different enzymes provides a basis for identifying potentially functional residues to be replaced by mutagenesis (15). In addition, analysis of the pattern for fusion of the glutamine amide transfer domain to other functional domains has led to a model for the evolution of amidotransferases having the capacity to utilize NH_3 and glutamine (8). For these reasons, sequences of additional glutamine amidotransferases are of interest.

In this paper, we report the nucleotide sequence of *pyrG* and the derived CTP synthetase amino acid sequence. The CTP synthetase glutamine amide transfer domain contains three conserved regions similar to those in anthranilate synthase, *p*-aminobenzoate synthase, GMP synthetase, and carbamoyl-P synthetase. The CTP synthetase glutamine amide transfer domain is located at the C-terminal end of the molecule, encoded by DNA at the 3′ end of the gene, consistent with a model (8) for evolution by gene fusion to augment the function of an NH_3-dependent enzyme. Experiments to map the 5′ and 3′ ends of the *pyrG* mRNA led to the finding that *eno*, which encodes enolase, is 89 bp[1] downstream of *pyrG*.

EXPERIMENTAL PROCEDURES[2]

RESULTS

Subcloning pyrG—Plasmid pNF1519 contains *pyrG* in a 4.3-kb *Pst*I fragment of *E. coli* DNA cloned in pBR322 (Fig. 1). *pyrG* was subcloned into vector pUC8 as shown in Fig. 1. Plasmid pMW1 was obtained by ligation of a mixture of *Bam*HI fragments from pNF1519 into the *Bam*HI site of pUC8. Selection for *pyrG*$^+$ was by functional complementation of *pyrG* in strain JF646. Plasmid pMW5 was constructed by ligating the 2.6-kb *Sal*I-*Pst*I segment of *E. coli* DNA from pMW1 into the *Sal*I and *Pst*I polylinker sites in pUC8. Further subcloning indicated that DNA at the *Bam*HI and *Kpn*I sites in pMW5 was essential for *pyrG* function.

DNA Sequence—The DNA sequence of *pyrG* was initially determined using fragments isolated from plasmid pMW5. *Rsa*I, *Hpa*II, or *Taq*I digests of plasmid pMW5 or of the *Sal*I-

* This work was supported by United States Public Health Service Grant GM24658. This is Journal Paper No. 10,525 from the Purdue University Agricultural Experiment Station. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ Recipient of National Institutes of Health Predoctoral Training Grant GM07211.

[1] The abbreviations used are: bp, base pair; kb, kilobase pair.

[2] Portions of this paper (including "Experimental Procedures" and Figs. 1, 2, 4, 5, and 6) are presented in miniprint at the end of this paper. Miniprint is easily read with the aid of a standard magnifying glass. Full size photocopies are available from the Journal of Biological Chemistry, 9650 Rockville Pike, Bethesda, MD 20814. Request Document No. 85M-3563, cite the authors, and include a check or money order for $4.40 per set of photocopies. Full size photocopies are also included in the microfilm edition of the Journal that is available from Waverly Press.

*Pst*I insert were ligated into M13mp18 or M13mp19 (Fig. 2*A*). In addition, specific subfragments were isolated from digests that were obtained using restriction enzymes having 6-bp recognition sequences (Fig. 2*B*). Finally, the exonuclease III procedure (26) was employed to obtain a set of overlapping sequences from *Nru*I-*Bam*HI and *Bam*HI-*Pst*I segments of the cloned DNA (Fig. 2*C*). The DNA sequence shown in Fig. 3 extends from 11 bp upstream of the *Nru*I site to the downstream *Pst*I site at nucleotide 2442. The entire sequence was determined on both DNA strands from overlapping fragments.

The derived amino acid sequence of CTP synthetase is shown in Fig. 3. The protein chain of 545 amino acid residues has a calculated molecular weight of 60,450. At nucleotides 2074–2076, an ATG initiates an open reading frame that extends 123 codons to the 3′ end of the cloned *E. coli* DNA. By screening protein data banks, the downstream sequence was found to be homologous with that of yeast enolase. Thus, *E. coli eno* is 89 bp downstream from *pyrG*.

CTP Synthetase—Enzyme was purified to homogeneity from cells bearing plasmid pMW5. A single stained protein band was obtained by sodium dodecyl sulfate-polyacrylamide gel electrophoresis. By comparison with proteins of known molecular weight, the CTP synthetase subunit had an estimated molecular weight of 60,000 (Fig. 4).

The N-terminal amino acid sequence of CTP synthetase was determined by automated Edman degradation. The data for 24 cycles were indicative of a pure protein. In the first cycle, a complex mixture was obtained, apparently containing methionine, threonine, and a residue similar to alanine. From cycles 2 through 24, the sequence was identical with that shown in Fig. 3 from Thr-3 to Leu-25. We conclude that Met-1 was removed by processing and the N terminus of the mature enzyme is threonine 2 or a modified form of threonine.

pyrG mRNA—The 5′ end of *pyrG* mRNA was mapped by the nuclease S1 procedure (29). A 462-nucleotide *Nru*I-*Bam*HI DNA probe was used which extends from nucleotides 12–473 (Figs. 2 and 3). The probe was labeled with [α-^{32}P] dCTP by primer extension or the double-stranded fragment was isolated and 5′ end-labeled with [γ-^{32}P]ATP and polynucleotide kinase. The results of nuclease S1 mapping are shown in Fig. 5. Two protected fragments of coding strand DNA were obtained (Fig. 5*A*, *lane 2*). Noncoding strand DNA did not anneal to RNA and was completely digested (Fig. 5*A*, *lane 4*). To confirm that the two transcripts extend into the *pyrG* coding sequence, nuclease S1 mapping was repeated using the 5′ end-labeled *Nru*I-*Bam*HI probe. Fig. 5*B* shows that the same two protected fragments were obtained. The size of the major protected fragment is about 175 nucleotides, and the minor one approximately 255 nucleotides. More precise mapping was obtained by using a DNA sequencing ladder as a size standard (Fig. 5*C*). These results confirm those obtained with restriction fragment size standards. Corresponding sites for transcription initiation are *overlined* in Fig. 3.

The 3′ end of *pyrG* mRNA was mapped with a *Kpn*I-*Pst*I DNA probe that extends from nucleotides 1663–2442 (Figs. 2 and 3). The results of nuclease S1 mapping are shown in Fig. 6. The major products obtained from the coding strand were undigested probe and a fragment of approximately 400 nucleotides (Fig. 6, *lane 2*). Minor fragments of approximately 420, 530, and 650 nucleotides were also obtained. The same pattern of fragments was obtained when the nuclease S1 concentration was increased 3-fold (data not shown). The noncoding strand DNA probe did not anneal to RNA and was completely digested (Fig. 6, *lane 4*). These results indicate that there are multiple species of *pyrG* and *eno* mRNA. An mRNA that anneals and fully protects the probe is suggestive of polycistronic *pyrG eno* mRNA.

DISCUSSION

The nucleotide sequence of *E. coli pyrG* was determined in order to extend our analysis of the relationship of glutamine amidotransferase structure to function. Recent sequence analyses indicate that different amidotransferases contain one of two distinct glutamine amide transfer domains (8–12). In all amidotransferases, a glutamine amide transfer domain is combined by various arrangements with a domain that catalyzes an NH_3-dependent biosynthetic reaction. This combination endows glutamine amidotransferases with the capacity to catalyze a glutamine-dependent as well as an NH_3-dependent biosynthetic reaction, both *in vitro* and *in vivo* (13, 14). Both types of glutamine amide transfer domain utilize an active site cysteine to form a covalent glutaminyl intermediate for catalysis of amide transfer (13, 14). Amidophosphoribosyltransferase (11) and glucosamine-6-P synthase (12) have a highly conserved glutamine amide transfer domain of approximately 190 amino acids that is characterized by an N-terminal active site cysteine. The second type of glutamine amide transfer domain, in GMP synthetase (30), carbamoyl-P synthetase (7, 10), anthranilate synthase (5), and *p*-aminobenzoate synthase (9), shown in Fig. 7, has three conserved segments. The active site cysteine in segment 2 is usually at a position approximately 80 to 90 amino acids from the N terminus of the domain.

The alignment in Fig. 7 localizes the CTP synthetase glutamine amide transfer domain and establishes its similarity to that in carbamoyl-P synthetase, GMP synthetase, anthranilate synthase, and *p*-aminobenzoate synthase. Using current nomenclature, the CTP synthetase glutamine amide transfer domain is *trpG*-related (8). In GMP synthetase, anthranilate synthase component II, and *p*-aminobenzoate synthase subunit II, the homologous *trpG*-related glutamine amide transfer domain (8) is initiated at the N-terminal residue of the protein chain. By noting that the first block of conserved sequence occurs 46–54 residues from the beginning of the domain in the three preceding enzymes, we estimate that the CTP synthetase glutamine amide transfer domain begins approximately at amino acids 292 to 300. Amino acid residues 1 to approximately 300 should contribute the structure needed for catalyzing the NH_3-dependent reaction. The glutamine amide transfer domain is fused onto the C-terminal end of the enzyme. Likewise, in carbamoyl-P synthetase, the glutamine amide transfer domain is fused onto the C-terminal end of a protein chain, except that the function of the N-terminal 185-amino-acid segment is unknown.

The conservation of amino acids in region 2 is sufficiently high to predict that the conserved cysteine, residue 379 in CTP synthetase, functions to form the covalent glutaminyl intermediate as has been shown for Cys-84 in anthranilate synthase component II (14). Likewise, CTP synthetase His-515 in region 3 is implicated in the proton transfer that is required for ionization of Cys-379 (15). Whereas previous experiments have provided evidence for catalytic roles of cysteinyl and histidyl side chains in regions 2 and 3, respectively, there is no evidence bearing on the possible role of region 1 in glutamine amide transfer.

In a previous analysis of the pattern for fusion of the glutamine amide transfer domain to other protein chains, a model was proposed to explain the evolution of glutamine amidotransferases from primitive NH_3-dependent enzymes (8). According to this model, after duplication, genes encoding

```
        10         20         30         40         50         60         70
GAGCGTCGTT TTCGCGAAGT GGAGCGTATT GTTGCCGCGC GTGGACTGGA AATGACAGGT GTTGACCTCG

        80         90        100        110        120        130        140
AAACAATGGA AGAAGTCTGG CAACAGGTAA AACGGCAGGA AATTGATCTC TAAGGGAATT ACGCGGTCAA

       150        160        170        180        190        200        210
GCGCCATTTG TGTCATTTTT TAAATGACAA GCGCTTGATT TGCGTCAAAA ACATTTACCC CAAAGGGGCT

       220  (2)   230        240        250        260        270        280
ATTTTCTCAC TCCTGATTTC AATAGTGCGC TGGCGAAGAG GAGGGATAAT GAAAGTTTGT GGCACAGGTC

       290        300  (1)   310        320        330        340        350
ATGTTCGGGT ATACTGCTTT CCCGTCTTGG TTATTCCATC GTCTTTTCAA CCTAACTTCT CAGGTTCAGC

                                  380                                        410
ATG ACA ACG AAC TAT ATT TTT GTG ACC GGC GGG GTC GTA TCC TCT CTG GGT AAA GGC ATT
Met Thr Thr Asn Tyr Ile Phe Val Thr Gly Gly Val Val Ser Ser Leu Gly Lys Gly Ile
                                   10                                         20

                                  440                                        470
GCC GCA GCC TCC CTC GCA GCC ATT CTT GAA GCC CGT GGC CTC AAT GTG ACC ATC ATG AAA
Ala Ala Ala Ser Leu Ala Ala Ile Leu Glu Ala Arg Gly Leu Asn Val Thr Ile Met Lys
                                   30                                         40

                                  500                                        530
CTG GAT CCG TAC ATC AAC GTC GAT CCA GGT ACT ATG AGC CCA ATC CAA CAC GGG GAA GTG
Leu Asp Pro Tyr Ile Asn Val Asp Pro Gly Thr Met Ser Pro Ile Gln His Gly Glu Val
                                   50                                         60

                                  560                                        590
TTC GTT ACT GAA GAC GGC GCT GAA ACC GAC CTG GAC CTG GGG CAC TAC GAG CGT TTC ATT
Phe Val Thr Glu Asp Gly Ala Glu Thr Asp Leu Asp Leu Gly His Tyr Glu Arg Phe Ile
                                   70                                         80

                                  620                                        650
CGT ACC AAA ATG AGC CGC CGC AAC AAC TTC ACC ACG GGT CGT ATC TAC TCT GAC GTT CTG
Arg Thr Lys Met Ser Arg Arg Asn Asn Phe Thr Thr Gly Arg Ile Tyr Ser Asp Val Leu
                                   90                                        100

                                  680                                        710
CGT AAA GAA CGC CGC GGT GAC TAC CTC GGC GCA ACC GTG CAG GTT ATT CCG CAC ATC ACT
Arg Lys Glu Arg Arg Gly Asp Tyr Leu Gly Ala Thr Val Gln Val Ile Pro His Ile Thr
                                  110                                        120

                                  740                                        770
AAC GCA ATC AAA GAG CGC GTG CTG GAA GGT GGC GAA GGT CAT GAC GTA GTA CTG GTA GAA
Asn Ala Ile Lys Glu Arg Val Leu Glu Gly Gly Glu Gly His Asp Val Val Leu Val Glu
                                  130                                        140

                                  800                                        830
ATC GGC GGT ACA GTA GGT GAT ATC GAA TCC TTG CCG TTC CTC GAA GCG ATT CGC CAG ATG
Ile Gly Gly Thr Val Gly Asp Ile Glu Ser Leu Pro Phe Leu Glu Ala Ile Arg Gln Met
                                  150                                        160

                                  860                                        890
GCT GTT GAA ATT GGC CGT GAG CAC ACT CTG TTT ATG CAC CTG ACG CTG GTG CCG TAC ATG
Ala Val Glu Ile Gly Arg Glu His Thr Leu Phe Met His Leu Thr Leu Val Pro Tyr Met
                                  170                                        180

                                  920                                        950
GCA GCG TCT GGT GAA GTC AAA ACC AAA CCG ACT CAG CAC TCT GTA AAA GAG CTG CTC TCC
Ala Ala Ser Gly Glu Val Lys Thr Lys Pro Thr Gln His Ser Val Lys Glu Leu Leu Ser
                                  190                                        200

                                  980                                       1010
ATC GGT ATC CAG CCT GAC ATC CTG ATT TGT CGT TCA GAT CGC GCT GTT CCG GCG AAC GAA
Ile Gly Ile Gln Pro Asp Ile Leu Ile Cys Arg Ser Asp Arg Ala Val Pro Ala Asn Glu
                                  210                                        220

                                 1040                                       1070
CGT GCG AAG ATT GCA TTG TTC TGT AAT GTT CCG GAA AAA GCG GTT ATT TCT CTG AAA GAC
Arg Ala Lys Ile Ala Leu Phe Cys Asn Val Pro Glu Lys Ala Val Ile Ser Leu Lys Asp
                                  230                                        240

                                 1100                                       1130
GTC GAT TCC ATC TAT AAA ATT CCG GGC CTG TTG AAA TCT CAG GGG CTG GAC GAT TAT ATT
Val Asp Ser Ile Tyr Lys Ile Pro Gly Leu Leu Lys Ser Gln Gly Leu Asp Asp Tyr Ile
                                  250                                        260

                                 1160                                       1190
TGT AAA CGA TTC AGC TTA AAC TGC CCG GAA GCG AAT CTG TCC GAA TGG GAA CAG GTT ATC
Cys Lys Arg Phe Ser Leu Asn Cys Pro Glu Ala Asn Leu Ser Glu Trp Glu Gln Val Ile
                                  270                                        280

                                 1220                                       1250
TTC GAA GAA GCG AAC CCG GTA AGT GAA GTC ACC ATC GGT ATG GTC GGC AAG TAC ATT GAA
Phe Glu Glu Ala Asn Pro Val Ser Glu Val Thr Ile Gly Met Val Gly Lys Tyr Ile Glu
                                  290                                        300
```

FIG. 3. **Nucleotide sequence of *E. coli pyrG*, flanking regions, and translated amino acid sequence of CTP synthetase.** The translated sequence of an N-terminal segment of enolase is downstream of CTP synthetase. *Underlined* sequences included: *pyrG* ribosome binding site, *eno* ribosome binding site. Regions corresponding to the major and minor mRNA 5′ ends are *overlined* and numbered (*1*) and (*2*), respectively.

```
                                    1280                                    1310
CTG CCG GAT GCT TAT AAA TCA GTG ATC GAA GCA CTG AAA CAC GGT GGG CTG AAG AAT CGT
Leu Pro Asp Ala Tyr Lys Ser Val Ile Glu Ala Leu Lys His Gly Gly Leu Lys Asn Arg
                                     310                                     320

                                    1340                                    1370
GTC AGC GTC AAC ATC AAA CTG ATC GAT TCA CAA GAT GTT GAA ACG CGC GGG CTT GAA ATC
Val Ser Val Asn Ile Lys Leu Ile Asp Ser Gln Asp Val Glu Thr Arg Gly Leu Glu Ile
                                     330                                     340

                                    1400                                    1430
CTT AAA GGT CTG GAC GCA ATC CTC GTA CCT GGC GGT TTC GGC TAT CGT GGC GTA GAA CGC
Leu Lys Gly Leu Asp Ala Ile Leu Val Pro Gly Gly Phe Gly Tyr Arg Gly Val Glu Gly
                                     350                                     360

                                    1460                                    1490
ATG ATT ACG ACC GCG CGT TTT GCG CGT GAG AAC AAT ATT CCT TAT CTG GGC ATT TGC CTG
Met Ile Thr Thr Ala Arg Phe Ala Arg Glu Asn Asn Ile Pro Tyr Leu Gly Ile Cys Leu
                                     370                                     380


                                    1520                                    1550
GGT ATG CAG GTG GCG TTA ATT GAT TAC GCT CGC CAT GTT GCC AAC ATG GAG AAC GCC AAC
Gly Met Gln Val Ala Leu Ile Asp Tyr Ala Arg His Val Ala Asn Met Glu Asn Ala Asn
                                     390                                     400

                                    1580                                    1610
TCT ACG GAA TTT GTG CCA GAC TGT AAG TAC CCG GTT GTG GCG CTG ATT ACC GAG TGG CGC
Ser Thr Glu Phe Val Pro Asp Cys Lys Tyr Pro Val Val Ala Leu Ile Thr Glu Trp Arg
                                     410                                     420

                                    1640                                    1670
GAT GAA AAC GGC AAC GTT GAA GTT CGT AGC GAG AAG AGC GAT CTC GGC GGT ACC ATG CGT
Asp Glu Asn Gly Asn Val Glu Val Arg Ser Glu Lys Ser Asp Leu Gly Gly Thr Met Arg
                                     430                                     440

                                    1700                                    1730
CTC GGC GCA CAG CAG TGC CAG TTG GTT GAC GAT AGC CTG GTT CGC CAG CTG TAC AAT GCG
Leu Gly Ala Gln Gln Cys Gln Leu Val Asp Asp Ser Leu Val Arg Gln Leu Tyr Asn Ala
                                     450                                     460

                                    1760                                    1790
CCG ACA ATT GTT GAG CGT CAT CGT CAC CGT TAC GAA GTC AAC AAC AGT CTG TTG AAA CAG
Pro Thr Ile Val Glu Arg His Arg His Arg Tyr Glu Val Asn Asn Ser Leu Leu Lys Gln
                                     470                                     480

                                    1820                                    1850
ATT GAA GAT GCA GGT CTG CGC GTT CGG GCG CGT TCC GGG GAT GAT CAG TTG GTC GAG ATC
Ile Glu Asp Ala Gly Leu Arg Val Arg Ala Arg Ser Gly Asp Asp Gln Leu Val Glu Ile
                                     490                                     500

                                    1880                                    1910
ATC GAA GTT CCG AAT CAC CCG TGG TTC GTG GCT TGC CAG TTC CAT CCG GAG TTT ACT TCT
Ile Glu Val Pro Asn His Pro Trp Phe Val Ala Cys Gln Phe His Pro Glu Phe Thr Ser
                                     510                                     520

                                    1940                                    1970
ACT CCA CGT GAT GGT CAC CCG CTG TTT GCA GGC TTT GTG AAA GCC GCC AGC GAG TTC CAG
Thr Pro Arg Asp Gly His Pro Leu Phe Ala Gly Phe Val Lys Ala Ala Ser Glu Phe Gln
                                     530                                     540

                                1998        2008        2018        2028        2038
 AAA CGT CAG GCG AAG TAA GTAAAAAAGT TAGAGCGGCA ACGTACCCTG GGTACGCGTT GTTTGTCTGG
 Lys Arg Gln Ala Lys End

        2048        2058        2068                                        2100
   AGTTTCAGTT TAACTAGTGA CTTGAGGAAA ACCTA ATG TCC AAA ATC GTA AAA ATC ATC GGT
                                          Met Ser Lys Ile Val Lys Ile Ile Gly

                                    2130                                    2160
CGT GAA ATC ATC GAC TCC CGT GGT AAC CCG ACT GTT GAA GCC GAA GTA CAT CTG GAG GGT
Arg Glu Ile Ile Asp Ser Arg Gly Asn Pro Thr Val Glu Ala Glu Val His Leu Glu Gly


                                    2190                                    2220
GGT TTC GTC GGT ATG GCA GCT GCT CCG TCA GGT GCT TCT ACT GGT TCC CGT GAA GCT CTG
Gly Phe Val Gly Met Ala Ala Ala Pro Ser Gly Ala Ser Thr Gly Ser Arg Glu Ala Leu

                                    2250                                    2280
GAA CTG CGC GAT GGC GAC AAA TCC CGT TTC CTG GGT AAA GGC GTA ACC AAA GCT GTT GCT
Glu Leu Arg Asp Gly Asp Lys Ser Arg Phe Leu Gly Lys Gly Val Thr Lys Ala Val Ala

                                    2310                                    2340
GCG GTA AAC GGC CCG ATC GCT CAG GCG CTG ATT GGC AAA GAT GCT AAA GAT CAG GCT GGC
Ala Val Asn Gly Pro Ile Ala Gln Ala Leu Ile Gly Lys Asp Ala Lys Asp Gln Ala Gly

                                    2370                                    2400
ATT GAC AAG ATC ATG ATC GAC CTG GAC GGC ACC GAA AAC AAA TCC AAA TTC GGC GCG AAC
Ile Asp Lys Ile Met Ile Asp Leu Asp Gly Thr Glu Asn Lys Ser Lys Phe Gly Ala Asn

                                    2430
GCA ATC CTG GCT GTA TCT CTG GCT AAC GCC AAA GCT GCT GCA
Ala Ile Leu Ala Val Ser Leu Ala Asn Ala Lys Ala Ala Ala
```

FIG. 3—*continued.*

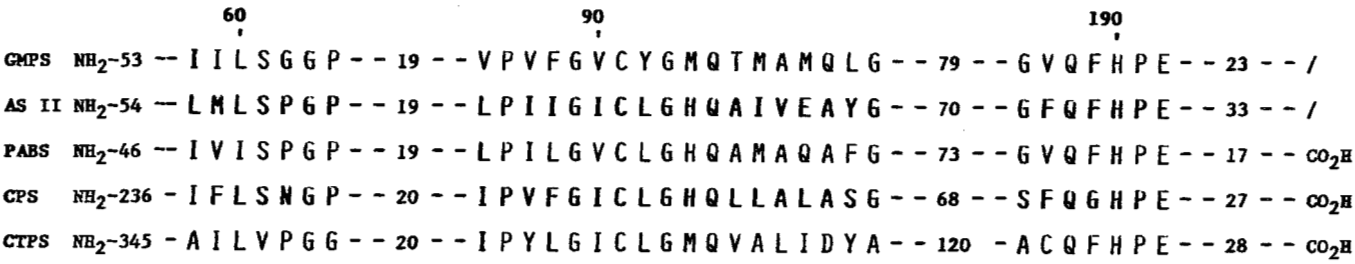


FIG. 7. **Alignment of amino acids in three conserved segments of the glutamine amide transfer domain in *E. coli* GMP synthetase (*GMPS*) (8, 30), anthranilate synthase component II (*AS II*) (5), *p*-aminobenzoate synthase subunit II (*PABS*) (9), carbamoyl-P synthetase (*CPS*) (7), and CTP synthetase (*CTPS*).** The *numbers* between dashes indicate the number of amino acids from the N terminus (NH_2), between segments, and to the CO_2 terminus (CO_2H) or to the end of the domain (/). The numbering system at the *top*, as used previously (8), counts all positions, including gaps, from the start of the domain.

a glutamine amide transfer domain translocated adjacent to operons encoding existing NH_3-dependent enzymes. It was suggested that the patterns of fusion were apparently determined by the inability of a glutamine amide transfer domain to translocate contiguous to a promoter. The arrangement of the glutamine amide transfer domain in CTP synthetase conforms to this model. Nuclease S1 mapping indicates that the major *pyrG* promoter is proximal to the translation initiation codon. No other genes intervene between the promoter and *pyrG*. The position of the glutamine amide transfer domain in CTP synthetase is consistent with translocation and fusion of a *trpG*-related glutamine amide transfer domain to the 3′ promoter distal end of an existing *pyrG* coding sequence of approximately 300 amino acids.

To explain the *trpG* (5, 31, 32) and *pabA* (9) gene fusion pattern in several microorganisms, it was proposed that *trpG*-related gene fusions onto the 3′ end of an existing gene were unfavorable compared to 5′ end fusions (8). It is now apparent that 3′ end *trpG*-related gene fusions occur in carbamoyl-P synthetase (7, 33) and CTP synthetase. It is uncertain why 3′ end *trpG* or *trpG*-related gene fusions did not occur with *trpE* or *pabB*, respectively.

Previously, evidence was reported (34) that *E. coli* CTP synthetase is a dimer (M_r = approximately 105,000) of identical subunits that undergoes aggregation to a tetramer of M_r = approximately 210,000. A subunit M_r = approximately 50,000 was determined by polyacrylamide gel electrophoresis in 8 M urea. The present experiments indicate a calculated M_r = 60,450 for the primary translation product. Removal of the initiator methionine residue should yield a protein chain having a calculated M_r = 60,300. The calculated value of 60,300 is consistent with the results of sodium dodecyl sulfate-polyacrylamide gel electrophoresis. NH_2-terminal protein sequencing verifies that the correct initiator methionine was identified.

Analysis of the *pyrG* 3′ flanking sequence indicates an ATG translation start 86 bp downstream from the *pyrG* TAA translation termination triplet. The ATG, position 2074, is preceded by a ribosome binding site (underlined) and starts an open reading frame that extends to the end of the cloned *E. coli* DNA. The translated sequence exhibits 57% identity with yeast enolase (35) and thus identifies *E. coli eno* immediately downstream of *pyrG*. The gene order on the *E. coli* linkage map (36) at min 59, *pyrG relA relX eno*, must therefore be revised to *eno pyrG relA* (16) based on DNA sequence analysis.

Nuclease S1 mapping of the *pyrG eno* boundary indicates multiple species of mRNA. One of the major products of this mapping experiment was the fully protected probe. The conventional interpretation of this result is that a polycistronic *pyrG eno* mRNA annealed to the probe and protected against nuclease S1 digestion. An alternative possibility, not presently excluded, is that two overlapping monocistronic *pyrG* and *eno* RNA molecules can anneal to the probe forming a tripartate nuclease S1-resistant structure (37). The other major mRNA of approximately 400 nucleotides should correspond either to a *pyrG* transcript having a 3′ end at approximately nucleotide 2059 or an *eno* transcript having a 5′ end at approximately nucleotide 2040. Likewise, the minor mRNA species of 420, 530, and 650 nucleotides either terminate distal to *pyrG* or initiate upstream of *eno*. Further experiments are required to determine whether multiple *pyrG* and *eno* mRNA molecules arise from transcription termination after *pyrG* and transcription initiation prior to *eno* or whether a primary *pyrG eno* mRNA undergoes processing. Since *pyrG* expression appears to be constitutive (3), there are no obvious regulatory barriers to a polycistronic *pyrG eno* transcriptional unit.

Acknowledgments—We thank James Friesen for providing *pyrG* plasmid pNF1519 and *E. coli* strain JF646, Mark Hermodson for sequenator analyses, and Dan Ebbole for insightful discussions.

REFERENCES

1. Buchanan, J. M. (1973) *Adv. Enzymol. Relat. Areas Mol. Biol.* **39,** 91–183
2. Koshland, D. E., Jr., and Levitzki, A. (1974) *The Enzymes* **10,** 539–559
3. O'Donovan, G., and Neuhard, J. (1970) *Bacteriol. Rev.* **34,** 278–343
4. Kawamura, M., Keim, P. S., Goto, Y., Zalkin, H., and Heinrikson, R. L. (1983) *J. Biol. Chem.* **253,** 4659–4668
5. Nichols, B. P., Miozzari, G. F., van Cleemput, M., Bennett, G. N., and Yanofsky, C. (1980) *J. Mol. Biol.* **142,** 503–517
6. Kaplan, J. B., and Nichols, B. P. (1983) *J. Mol. Biol.* **168,** 451–468
7. Piette, J., Nyunoya, H., Lusty, C. J., Cunin, R., Weyens, G., Crabeel, M., Charlier, D., Glansdorff, N., and Piérard, A. (1984) *Proc. Natl. Acad. Sci. U. S. A.* **81,** 4134–4138
8. Zalkin, H., Argos, P., Narayana, S. V. L., Tiedeman, A. A., and Smith, J. M. (1985) *J. Biol. Chem.* **260,** 3350–3354
9. Kaplan, J. B., Merkel, W. K., and Nichols, B. P. (1985) *J. Mol. Biol.* **183,** 327–340
10. Werner, M., Feller, A., and Piérard, A. (1985) *Eur. J. Biochem.* **146,** 371–381
11. Tso, J. Y., Zalkin, H., van Cleemput, M., Yanofsky, C., and Smith, J. M. (1982) *J. Biol. Chem.* **257,** 3525–3531
12. Walker, J. E., Gay, N. J., Saraste, M., and Eberle, A. N. (1984) *Biochem. J.* **224,** 799–815
13. Mäntsälä, P., and Zalkin, H. (1984) *J. Biol. Chem.* **259,** 14230–14236
14. Paluh, J. L., Zalkin, H., Betsch, D., and Weith, H. L. (1985) *J. Biol. Chem.* **260,** 1889–1894
15. Amuro, N., Paluh, J. L., and Zalkin, H. (1985) *J. Biol. Chem.* **260,** 14844–14849
16. Friesen, J. D., An, G., and Fiil, N. P. (1978) *Cell* **15,** 1187–1197

17. Birnboim, H. C., and Doly, J. (1979) *Nucleic Acids Res.* **7,** 1513–1525
18. Vogel, H. J., and Bonner, D. M. (1956) *J. Biol. Chem.* **218,** 97–102
19. Anderson, P. M. (1983) *Biochemistry* **22,** 3285–3292
20. Long, C. W., and Pardee, A. B. (1967) *J. Biol. Chem.* **242,** 4715–4721
21. Layne, E. (1957) *Methods Enzymol.* **3,** 447–454
22. Laemmli, U. K. (1970) *Nature* **227,** 680–685
23. Mahoney, W. C., Hogg, R. W., and Hermodson, M. A. (1981) *J. Biol. Chem.* **256,** 4350–4356
24. Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–5467
25. Biggen, M. D., Gibson, T. J., and Hong, G. F. (1983) *Proc. Natl. Acad. Sci. U. S. A.* **80,** 3963–3965
26. Henikoff, S. (1984) *Gene (Amst.)* **28,** 351–359
27. Queen, C., and Korn, L. J. (1984) *Nucleic Acids Res.* **12,** 581–599
28. Jinks-Robertson, S., Gourse, R. L., and Nomura, M. (1983) *Cell* **33,** 865–867
29. Berk, A. J., and Sharp, P. A. (1978) *Proc. Natl. Acad. Sci. U. S. A.* **75,** 1274–1278
30. Tiedeman, A., Smith, J. M., and Zalkin, H. (1985) *J. Biol. Chem.* **260,** 8676–8679
31. Schechtman, M. G., and Yanofsky, C. (1983) *J. Mol. Appl. Genet.* **2,** 83–99
32. Zalkin, H., Paluh, J. L., van Cleemput, M., Moye, W. S., and Yanofsky, C. (1984) *J. Biol. Chem.* **259,** 3985–3992
33. Nyunoya, H., and Lusty, C. J. (1984) *J. Biol. Chem.* **259,** 9790–9798
34. Levitzki, A., Stallcup, W. B., and Koshland, D. E., Jr. (1971) *Biochemistry* **10,** 3371–3378
35. Holland, M. J., Holland, J. P., Thill, G. P., and Jackson, K. A. (1981) *J. Biol. Chem.* **256,** 1385–1395
36. Bachmann, B. J., and Low, K. B. (1980) *Microbiol. Rev.* **44,** 1–56
37. Lopata, M. A., Sollner-Webb, B., and Cleveland, D. W. (1985) *Mol. Cell. Biol.* **5,** 2842–2846

Supplementary Material to

Nucleotide Sequence of *Escherichia coli pyrG* Encoding CTP Synthetase

Manli Weng, Christopher A. Makaroff, and Howard Zalkin

EXPERIMENTAL PROCEDURES

Plasmids, Strains and Media - *E. coli* strain JF646 (relevant genotype, *pyrE pyrG cdd argE his4 proA thr1 thi1 recA*) and plasmid pNF1519 (16) were provided by James Friesen, University of Toronto. Plasmids were isolated from transformants by the alkaline lysis method (17). This procedure was scaled up for the preparative isolation of plasmid DNA. CsCl/Ethidium bromide centrifugation was included as a final step. For enzyme production strain JF646/pMW5 was grown in media containing salts (18), 0.5% glucose, 0.5% acid hydrolyzed casein, 2 mg/l thiamine, 0.22% uracil. Cells were harvested in late log to early stationary phase.

Subcloning pyrG - Strain JF646 is a pyrimidine auxotroph that requires cytidine for growth. *pyrG*$^+$ was subcloned from pNF1519, a pBR322 derivative, to pUC8. *pyrG*$^+$ plasmids were selected in strain JF646, grown with uracil, by complementation of the chromosomal *pyrG* mutation.

Enzyme Purification - CTP synthetase was purified to electrophoretic homogeneity by the method of Anderson (19). Crude extracts of plasmid-bearing strain JF646 contained approximately 50-fold elevated levels of CTP synthetase (19), thus allowing the phenyl-Sepharose and Sephadex G-200 chromatographic steps to be omitted. CTP synthetase activity was determined at 23°C by the method of Long and Pardee (20) using the assay mixture specified by Anderson (19). The protein content of crude fractions was determined by the Biuret method (21) and for the purified enzyme by amino acid analysis. An extinction coefficient, $E^{1\%}_{1cm}$ at 280 nm of 8.33 was determined. The specific activity of the purified enzyme was 2.3 units/mg protein, assayed at 23°C. A unit of activity corresponds to the production of 1 μmol CTP/min at 23°C. Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (22) was used to assess homogeneity and for estimation of subunit molecular weight.

Amino Acid Analysis and Amino Acid Sequencing - Samples for amino acid analysis were hydrolyzed in 6 N HCl for 22 h, 44 h and 66 h *in vacuo*. Analyses were performed with a Durrum D500 amino acid analyzer. A Beckman 890D sequencer was employed for automated Edman degradation using the procedures of Mahoney *et al.* (23).

DNA Sequencing - DNA sequences were determined by the procedure of Sanger *et al.* (24). The polyacrylamide/urea gel electrophoresis described by Biggen *et al.* (25) was used for resolving the DNA fragments. DNA fragments from digests of pMW5 or the 2.6 Kb SalI-PstI insert were ligated into the appropriate cloning sites in M13mp18 or M13mp19. Digests were obtained using RsaI, TaqI, HpaII and restriction enzymes having six base pair recognition sequences. A set of 13 clones containing overlapping *pyrG* fragments from the BamHI-PstI region and 3 clones from the NruI-BamHI segment were obtained by the exonuclease III procedure of Henikoff (26). M13 recombinants were transformed into JM101. DNA and protein sequences were analyzed by computer (27).

Nuclease S1 Mapping - RNA was isolated (28) from strain W3110 grown to midlog phase in minimal media. A 462 nucleotide NruI-BamHI probe, nucleotides 12-473 (Fig. 3), was used to map the 5' end of *pyrG*. The NruI-BamHI fragment was cloned into M13mp18 and M13mp19. ^{32}P-labeled probes were synthesized by primer extension using 17 base sequencing primer, [α-^{32}P]dCTP, TTP, dATP, dGTP and DNA polymerase Klenow fragment. After primer extension the radioactive probe was cut out of M13 using flanking restriction enzyme sites in the polylinker region. Coding strand hybridization probe, complementary to mRNA was obtained from M13mp19. A noncoding strand control having the same sequence as mRNA was obtained from M13mp18. For a second series of experiments the 462 bp NruI-BamHI fragment was isolated from pMW5 and labeled with [γ-^{32}P]ATP and polynucleotide kinase. DNA probes were hybridized (29) with 79 μg RNA at 54°C for 12 h. Nuclease S1 treatment (1400 units per ml) was at 37°C for 40 min. RNA-protected fragments were resolved on 5% polyacrylamide 7 M urea gels with MspI-digested pBR322 for size standard.

A 780 nucleotide KpnI-PstI probe (nucleotides 1663-2442, Fig. 3) was used to map the 3' end of *pyrG* in a similar way. Primer extension of the M13mp18 recombinant gave a coding strand hybridization probe complementary to mRNA. Primer extension from the M13mp19 recombinant was used to obtain a radioactive noncoding DNA strand to serve as a control.

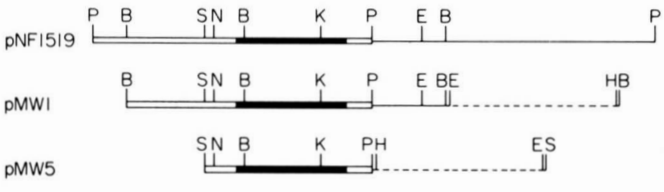


Fig. 1. *E. coli pyrG* plamids. Restriction maps are drawn approximately to scale. The symbols are: filled box, *pyrG* coding DNA; open box, *E. coli* flanking DNA; solid line, pBR322; dashed line, pUC8. Abbreviations for restriction sites are: B, BamHI; E, EcoRI; H, HindIII; K, KpnI; N, NruI; P, PstI; S, SalI.

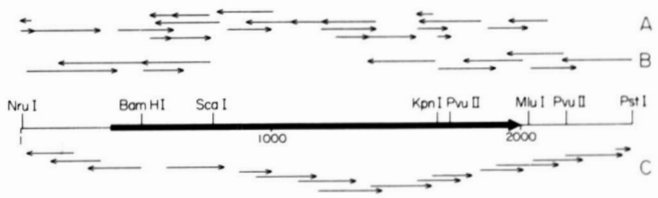


Fig. 2. Sequencing strategy for *pyrG*. The *pyrG* coding sequence is indicated as a solid box. Numbering is from a TaqI site just upstream of the unique NruI site. Arrows indicate the extent of sequence obtained from each cloned fragment. Three sets of clones were used: A, fragments obtained using RsaI, TaqI, HpaII; B, fragments obtained from restriction sites shown on the map; C, obtained by exonuclease III digestion.

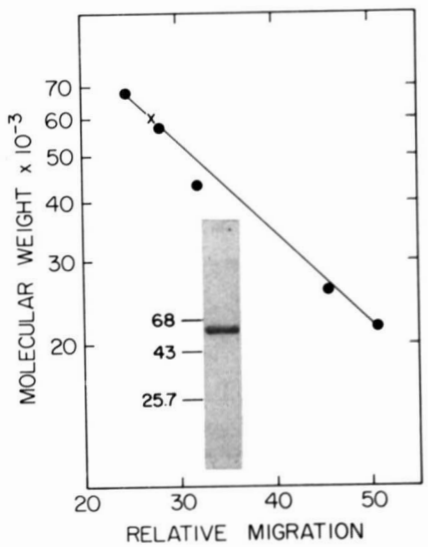
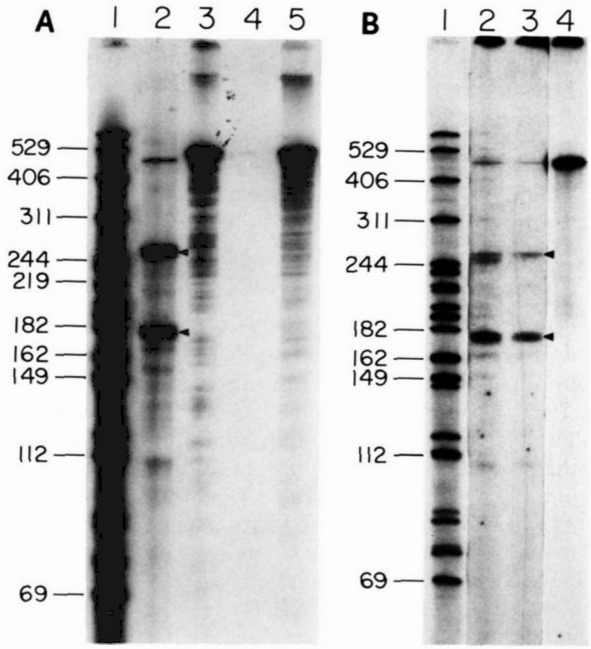
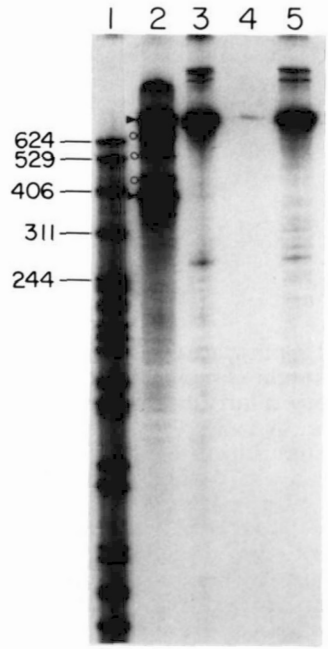


Fig. 4. Estimation of CTP synthetase subunit molecular weight by sodium dodecyl sulfate polyacrylamide gel electrophoresis. Molecular weight reference proteins (●) are bovine serum albumin (68,000), anthranilate synthase component I (57,500), ovalbumin (43,000), α-chymotrypsinogen (25,700), anthranilate synthase component II (21,800); (X), migration of CTP synthetase. The inset shows a photograph of a stained gel. The numbers at the left refer to positions of molecular mass standards in kilodaltons.





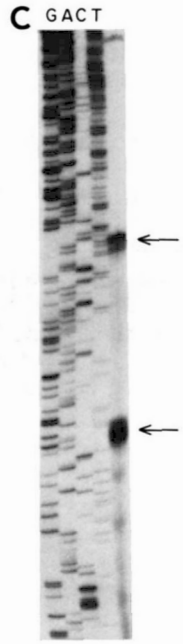


Fig. 6. Nuclease S1 mapping of the 3' end of pyrG mRNA. DNA probes were synthesized by primer extension. The 3'-distal KpnI-PstI fragment was cloned into M13mp18 and M13mp19 to give coding and noncoding strand probes, respectively. Lane 1, MspI-digested pBR322 size standard; lane 2, RNA-protected fragments of coding strand probe; lane 3, undigested coding strand probe; lane 4, non coding strand probe incubated with RNA and digested with nuclease S1; lane 5, undigested non coding strand probe. Arrowheads point to major protected fragments, circles mark minor protected fragments.

Fig. 5. Nuclease S1 mapping of the 5' end of pyrG mRNA. (A) DNA probes synthesized by primer extension. The NruI-BamHI fragment was cloned in M13mp18 and M13mp19. Primer extension using M13mp19 yields coding strand DNA complementary to mRNA; M13mp18 yields noncoding strand DNA. Hybridization reactions contained approximately 0.02 pmol DNA probe and 79 μg RNA. Hybridization was at 54°C for 12 h. Lane 1, MspI-digested pBR322 size standard; lane 2, RNA protected fragments of coding strand probe; lane 3, undigested coding strand probe; lane 4, non coding strand probe incubated with RNA and digested with nuclease S1; lane 5, undigested non coding strand probe.
(B) NruI-BamHI probe 5' end labeled. Hybridization reactions contained approximately 0.3 pmol DNA probe and 79 μg RNA. Lane 1, size standard; lane 2, probe, RNA, 700 units per ml nuclease S1; lane 3, probe, RNA, 1400 units per ml nuclease S1; lane 4, undigested probe. Lanes 1 and 4 were exposed for 4 h at -20°C with an intensifying screen, lanes 2 and 3 for 12 h at -70°C with an intensifying screen.
(C) Nuclease S1-digested fragments from (A) electrophoresed alongside a dideoxy sequencing ladder of the NruI-BamHI fragment. The arrows mark the positions of two nuclease S1-digested fragments. The lower major band corresponds to an mRNA 5' end at approximately 303 in the nucleotide sequence. The upper minor band corresponds to an mRNA 5' end at approximately 222 in the nucleotide sequence.