Proc. Natl. Acad. Sci. USA
Vol. 80, pp. 4629–4633, August 1983
Biochemistry

The *carB* gene of *Escherichia coli*: A duplicated gene coding for the large subunit of carbamoyl-phosphate synthetase

(DNA sequence/primary structure/gene duplication/protein evolution)

HIROSHI NYUNOYA* AND C. J. LUSTY

The Public Health Research Institute of The City of New York, New York, New York 10016

Communicated by Alton Meister, April 13, 1983

ABSTRACT **Previous genetic and biochemical studies indicate that the *carB* gene of *Escherichia coli* codes for the large subunit of carbamoyl-phosphate synthetase (EC 6.3.5.5). We have determined the nucleotide sequence of a 4-kilobase-pair cloned fragment of *E. coli* DNA with genetic determinants for *carB*. The DNA sequence is a 3,219-nucleotide-long reading frame. The polypeptide encoded by this reading frame has been verified to be the large subunit of carbamoyl-phosphate synthetase. The gene product is similar to the large subunit in its molecular weight, amino acid composition and amino-terminal residue, and carboxyl-terminal sequence. The amino acid sequence derived from the nucleotide sequence shows a highly significant homology between the amino- and carboxyl-terminal halves of the protein. We propose that the *carB* gene was formed by an internal duplication of a smaller ancestral gene.**

Carbamoyl phosphate is a common intermediate of the arginine and pyrimidine biosynthetic pathways. In *Escherichia coli* and other bacteria, the synthesis of this compound is catalyzed by a single enzyme, carbamoyl-phosphate synthetase (EC 6.3.5.5). Eukaryotes generally have two different synthetases, one specific for the arginine pathway and the other for the pyrimidine pathway (1). Comparison of the molecular weights and subunit structures of the various carbamoyl-phosphate synthetases suggests their genes could have evolved through gene duplication and gene fusion.

The *E. coli* carbamoyl-phosphate synthetase has been shown to be composed of two different subunits. The smaller subunit ($M_r \approx 42{,}000$) promotes hydrolysis of glutamine, the donor of the amino group (2). Ammonia formed from glutamine is used to synthesize carbamoyl phosphate by the large subunit ($M_r \approx 130{,}000$) (2). The small and large subunits of the *E. coli* enzyme are encoded by the *carA* and *carB* genes, respectively (3, 4). The two genes are genetically linked (4) and have recently been cloned (5, 6).

In studies described in the present communication, a recombinant plasmid with the *car* operon has been used to sequence the gene for the large subunit. The gene consists of a 3,219-nucleotide-long open reading frame coding for a protein of M_r 117,710. The amino acid sequence of the large subunit derived from the gene sequence reveals that the amino and carboxyl halves of the protein are highly homologous. Based on this internal homology, we conclude that the *carB* gene arose as a result of an internal duplication of a smaller ancestral gene.

MATERIALS AND METHODS

Bacterial Strains and DNA. The *car* operon was first isolated in the phage λ*dcarAB37-9* (5, 6). The presence of the genes for the large and small subunits of carbamoyl-phosphate synthetase was confirmed by *in vivo* and *in vitro* complementation (5–7). The recombinant λ phage was also shown to direct synthesis of both carbamoyl-phosphate synthetase subunits in a cell-free-coupled transcription-translation system (8). The *car* operon was subcloned by insertion into the *Hin*dIII site of pBR322 and was provided to us by Marjolaine Crabeel and Nicolas Glansdorff. *E. coli* strain 58.161 (*recA carB8 thr⁻ metB*) carrying recombinant plasmid pMC40 with the adjacent *carAB* genes was grown in LB broth supplemented with 20 μg of ampicillin per ml. Plasmid DNA was isolated by the cleared lysate method (9) and was purified by cesium chloride equilibrium centrifugation.

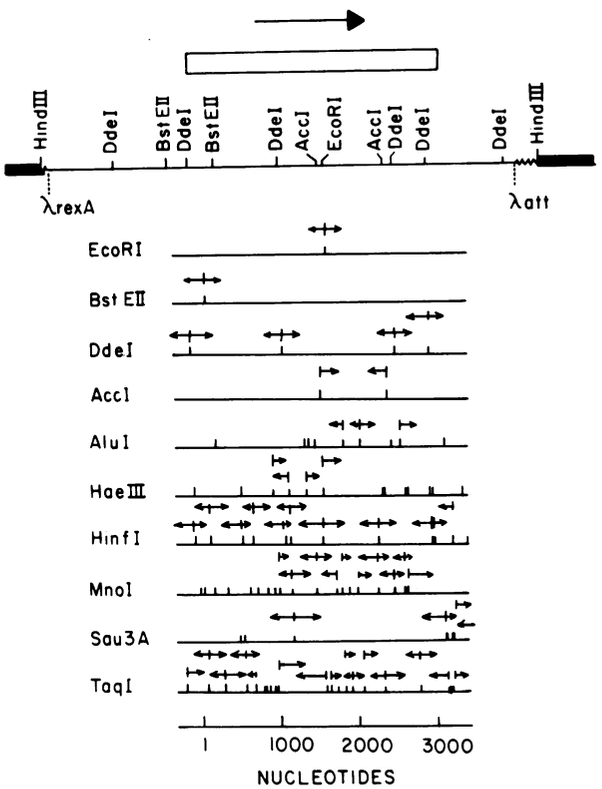

FIG. 1. Restriction map of the *carAB* insert of pMC40. *E. coli* chromosomal DNA is indicated by the thin line (——) and is ≈6 kbp in length, λ DNA is shown by the sawtooth line (〰), and the pBR322 vector is shown by the dark bars (■). The arrows indicate the direction and the lengths of the sequences obtained. The limits of the *carB* gene within the sequenced region are indicated by the open bar (□). The large arrow indicates the direction of transcription.

DNA Sequence Analysis. pMC40 was digested with *Hin*dIII, which cleaves the plasmid into two fragments of 6.3 and 4.4 kilobase pairs (kbp) (Fig. 1). The 6.3-kbp fragment represent-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: kbp, kilobase pair(s).

*On leave from the Dept. of Biology, Okayama Univ., Okayama, Japan.

```
                                   -50      ↓                                                  1
                                                                                                Pro lys arg thr
CACGACGCCGCGCCGTTGTTCGACCACTTTATCGAGTTAATTGAGCAGTACCGTAAAACCGCTAAGTAATCAGGAGTAAAAGAGCC  ATG CCA AAA CGT ACA

asp ile lys ser ile leu ile leu gly ala gly pro ile val ile gly gln ala cys glu phe asp tyr ser gly ala gln
GAT ATA AAA AGT ATC CTG ATT CTG GGT GCG GGC CCG ATT GTT ATC GGT CAG GCG TGT GAG TTT GAC TAC TCT GGC GCG CAA

    100
ala cys lys ala leu arg glu glu gly tyr arg val ile leu val asn ser asn pro ala thr ile met thr asp pro glu
GCG TGT AAA GCC CTG CGT GAA GAG GGT TAC CGC GTC ATT CTG GTG AAC TCC AAC CCG GCG ACC ATC ATG ACC GAC CCG GAA

                            200
met ala asp ala thr tyr ile glu pro ile his trp glu val val arg lys ile ile glu lys glu arg pro asp ala val
ATG GCT GAT GCA ACC TAC ATC GAG CCG ATT CAC TGG GAA GTT GTA CGC AAG ATT ATT GAA AAA GAG CGC CCG GAC GCG GTG

                                                300
leu pro thr met gly gly gln thr ala leu asn cys ala leu glu leu glu arg gln gly val leu glu glu phe gly val
CTG CCA ACG ATG GGC GGT CAG ACG GCG CTG AAC TGC GCG CTG GAG CTG GAA CGT CAG GGC GTG TTG GAA GAG TTC GGT GTC

                                                                                400
thr met ile gly ala thr ala asp ala ile asp lys ala glu asp arg arg arg phe asp val ala met lys lys ile gly
ACC ATG ATT GGT GCC ACT GCC GAT GCG ATT GAT AAA GCA GAA GAC CGC CGT CGT TTC GAC GTA GCG ATG AAG AAA ATT GGT

                                                                                                        500
leu glu thr ala arg ser gly ile ala his thr met glu glu ala leu ala val ala ala asp val gly phe pro cys ile
CTG GAA ACC GCG CGT TCC GGT ATC GCA CAC ACG ATG GAA GAA GCG CTG GCG GTT GCC GCT GAC GTG GGC TTC CCG TGC ATT

ile arg pro ser phe thr met gly gly ser gly gly gly ile ala tyr asn arg glu glu phe glu glu ile cys ala arg
ATT CGC CCA TCC TTT ACC ATG GGC GGT AGC GGC GGC GGT ATC GCT TAT AAC CGT GAA GAG TTT GAA GAA ATT TGC GCC CGC

                     600
gly leu asp leu ser  pro thr lys glu leu leu ile asp glu ser leu ile gly trp lys glu tyr glu met glu val val
GGT CTG GAT CTC TCT  CCG ACC AAA GAG TTG CTG ATT GAT GAG TCG CTG ATC GGC TGG AAA GAG TAC GAG ATG GAA GTG GTG

                                                700
arg asp lys asn asp asn cys ile ile val cys ser ile glu asn phe asp ala met gly ile his thr gly asp ser ile
CGT GAT AAA AAC GAC AAC TGC ATC ATC GTC TGC TCT ATC GAA AAC TTC GAT GCG ATG GGC ATC CAC ACC GGT GAC TCC ATC

                                                                        800
thr val ala pro ala gln thr leu thr asp lys glu tyr gln ile met arg asn ala ser met ala val leu arg glu ile
ACT GTC GCG CCA GCC CAA ACG CTG ACC GAC AAA GAA TAT CAA ATC ATG CGT AAC GCC TCG ATG GCG GTG CTG CGT GAA ATC

                                                                                                900
gly val glu thr gly gly ser asn val gln phe ala val asn pro lys asn gly arg leu ile val ile glu met asn pro
GGC GTT GAA ACC GGT GGT TCC AAC GTT CAG TTT GCG GTG AAC CCG AAA AAC GGT CGT CTG ATT GTT ATC GAA ATG AAC CCA

arg val ser arg ser ser ala leu ala ser lys ala thr gly phe pro ile ala lys val ala ala lys leu ala val gly
CGC GTG TCC CGT TCT TCG GCG CTG GCG TCG AAA GCG ACC GGT TTC CCG ATT GCT AAA GTG GCG GCG AAA CTG GCG GTG GGT

                1000
tyr thr leu asp glu leu met asn asp ile thr gly gly arg thr pro ala ser phe glu pro ser ile asp tyr val val
TAC ACC CTC GAC GAA CTG ATG AAC GAC ATC ACT GGC GGA CGT ACT CCG GCC TCC TTC GAG CCG TCC ATC GAC TAT GTG GTT

                                        1100
thr lys ile pro arg phe asn phe glu lys phe ala gly ala asn asp arg leu thr thr gln met lys ser val gly glu
ACT AAA ATT CCT CGC TTC AAC TTC GAA AAA TTC GCC GGT GCT AAC GAC CGT CTG ACC ACT CAG ATG AAA TCG GTT GGC GAA

                                                                1200
val met ala ile gly arg thr gln gln glu ser leu gln lys ala leu arg gly leu glu val gly ala thr gly phe asp
GTG ATG GCG ATT GGT CGC ACG CAG CAG GAA TCC CTG CAA AAA GCG CTG CGC GGC CTG GAA GTC GGT GCG ACT GGA TTC GAC

                                                                                        1300
pro lys val ser leu asp asp pro glu ala leu thr lys ile arg arg glu leu lys asp ala gly ala asp arg ile trp
CCG AAA GTG AGC  TG GAT GAC CCG GAA GCG TTA ACC AAA ATC CGT CGC GAA CTG AAA GAC GCA GGC GCA GAT CGT ATC TGG

tyr ile ala asp ala phe arg ala gly leu ser val asp gly val phe asn leu thr asn ile asp arg trp phe leu val
TAC ATC GCC GAT GCG TTC CGT GCG GGC CTG TCT GTG GAC GGC GTC TTC AAC CTG ACC AAC ATT GAC CGC TGG TTC  TG GTA

        1400
gln ile glu glu leu val arg leu glu glu lys val ala glu val gly ile thr gly leu asn ala asp phe leu arg gln
CAG ATT GAA GAG CTG GTG CGT CTG GAA GAG AAA GTG GCG GAA GTG GGC ATC ACT GGC CTG AAC GCT GAC TTC CTG CGC CAG

                                1500
leu lys arg lys gly phe ala asp ala arg leu ala lys leu ala gly val arg glu ala glu ile arg lys leu arg asp
CTG AAA CGC AAA GGC TTT GCC GAT GCG CGC TTG GCA AAA CTG GCG GGC GTA CGC GAA GCG GAA ATC CGT AAG CTG CGT GAC

                                                            1600
gln tyr asp leu his pro val tyr lys arg val asp thr cys ala ala glu phe ala thr asp thr ala tyr met tyr ser
CAG TAT GAC CTG CAC CCG GTT TAT AAG CGC GTG GAT ACC TGT GCG GCA GAG TTC GCC ACC GAC ACC GCT TAC ATG TAC TCC

                                                                                    1700
thr tyr glu glu glu cys glu ala asn pro ser thr asp arg glu lys ile met val leu gly gly gly pro asn arg ile
ACT TAT GAA GAA GAG TGC GAA GCG AAT CCG TCT ACC GAC CGT GAA AAA ATC ATG GTG CTT GGC GGC GGC CCG AAC CGT ATC

gly gln gly ile glu phe asp tyr cys cys val his ala ser leu ala leu arg glu asp gly tyr glu thr ile met val
GGT CAG GGT ATC GAA TTC GAC TAC TGT TGC GTA CAC GCC TCG CTG GCG CTG CGC GAA GAC GGT TAC GAA ACC ATT ATG GTT

1800
asn cys asn pro glu thr val ser thr asp tyr asp thr ser asp arg leu tyr phe glu pro val thr leu glu asp val
AAC TGT AAC CCG GAA ACC GTC TCC ACC GAC TAC GAC ACT TCC GAC CGC CTC TAC TTC GAG CCG GTA ACT CTG GAA GAT GTG

                            1900
leu glu ile val arg ile glu lys pro lys gly val ile val gln tyr gly gly gln thr pro leu lys leu ala arg ala
CTG GAA ATC GTG CGT ATC GAG AAG CCG AAA GGC GTT ATC GTC CAG TAC GGC GGT CAG ACC CCG CTG AAA CTG GCG CGC GCG

                                                    2000
leu glu ala ala gly val pro val ile gly thr ser pro asp ala ile asp arg ala glu asp arg glu arg phe gln his
CTG GAA GCT GCT GGC GTA CCG GTT ATC GGC ACC AGC CCG GAT GCT ATC GAC CGT GCA GAA GAC CGT GAA CGC TTC CAG CAT

                                                                                2100
ala val glu arg leu lys leu lys gln pro ala asn ala thr val thr ala ile glu met ala val glu lys ala lys glu
GCG GTT GAG CGT CTG AAA CTG AAA CAA CCG GCG AAC GCC ACC GTT ACC GCT ATT GAA ATG GCG GTA GAG AAG GCG AAA GAG

                                                                                                        2200
ile gly tyr pro leu val val arg pro ser tyr val leu gly gly arg ala met glu ile val tyr asp glu ala asp leu
ATT GGC TAC CCG CTG GTG GTA CGT CCG TCT TAC GTT CTC GGC GGT CGG GCG ATG GAA ATC GTC TAT GAC GAA GCT GAC CTG

arg arg tyr phe gln thr ala val ser val ser asn asp ala pro val leu leu asp his phe leu asp asp ala val glu
CGT CGC TAC TTC CAG ACG GCG GTC AGC GTG TCT AAC GAT GCG CCA GTG TTG CTG GAC CAC TTC CTC GAT GAC GCG GTA GAA

                    2300
val asp val asp ala ile cys asp gly glu met val leu ile gly gly ile met glu his ile glu gln ala gly val his
GTT GAC GTG GAT GCC ATC TGC GAC GGC GAA ATG GTG CTG ATT GGC GGC ATC ATG GAG CAT ATT GAG CAG GCG GGC GTG CAC

                                            2400
ser gly asp ser ala cys ser leu pro ala tyr thr leu ser gln glu ile gln asp val met arg gln gln val gln lys
TCC GGT GAC TCC GCA TGT TCT CTG CCA GCC TAC ACC TTA AGT CAG GAA ATT CAG GAT GTG ATG CGC CAG CAG GTG CAG AAA

                                                                        2500
leu ala phe glu leu gln val arg gly leu met asn val gln phe ala val lys asn asn glu val tyr leu ile glu val
CTG GCC TTC GAA TTG CAG GTG CGC GGC CTG ATG AAC GTG CAG TTT GCG GTG AAA AAC AAC GAA GTC TAC CTG ATT GAA GTT

                                                                                                2600
asn pro arg ala ala arg thr val pro phe val ser lys ala thr gly val pro leu ala lys val ala ala arg val met
AAC CCG CGT GCG GCG CGT ACC GTT CCG TTC GTC TCC AAA GCC ACC GGC GTA CCG CTG GCA AAA GTG GCG GCG CGC GTG ATG

ala gly lys ser leu ala glu gln gly val thr lys glu val ile pro pro tyr tyr ser val lys glu val val leu pro
GCT GGC AAA TCG CTG GCT GAG CAG GGC GTA ACC AAA GAA GTT ATC CCG CCG TAC TAC TCG GTG AAA GAA GTG GTG CTG CCG
```

(*Fig. 2 continues on next page.*)

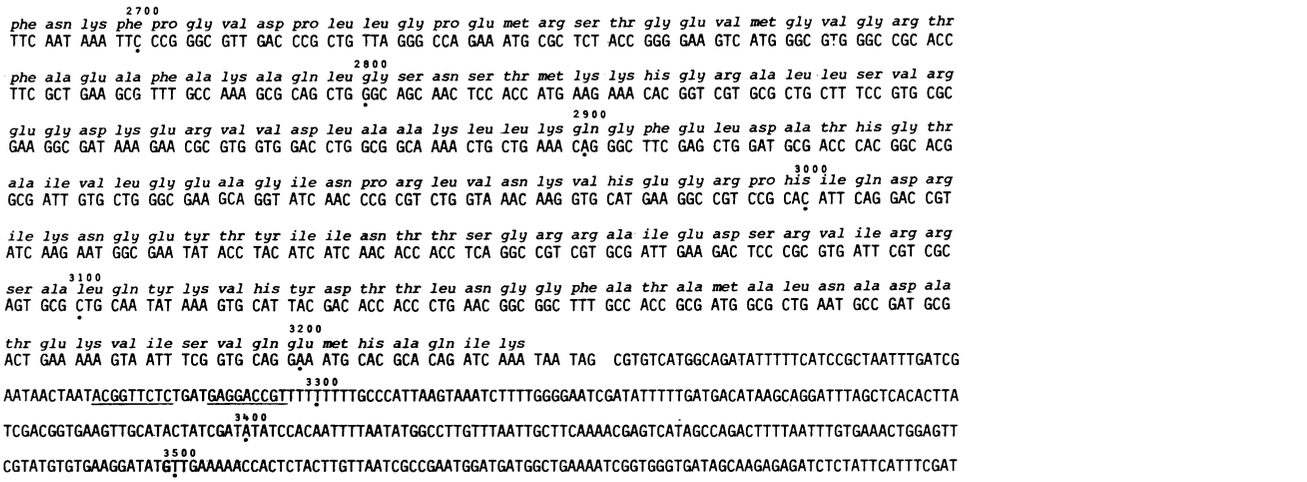

FIG. 2. Nucleotide sequence of the *carB* gene. The sequence is that of the nontranscribed strand. The *carB* gene begins at nucleotide +1 and ends at nucleotide +3,219. A Shine–Dalgarno sequence (A-G-G-A-G), Pribnow box (T-C-G-A-G-T-T), and RNA polymerase binding site (C-G-T-T-G-T-T-C-G-A-C-C) upstream of the ATG initiation codon are underlined. The arrow (↓) denotes the start site of the message. At nucleotide +3,300, a sequence with dyad symmetry (underlined) indicates a site with features of transcriptional termination.

ing the entire insert in the pBR322 vector was separated and isolated on preparative agarose gels. The purified fragment was further digested with either *Eco*RI, *Dde* I, or *Acc* I with *Bst*EII. The resultant fragments were also purified by electrophoresis on preparative agarose gels. The fragments obtained from these digests were cleaved with restriction endonucleases and were 5′-end labeled and the single strands were obtained by electrophoresis on polyacrylamide gels (10). Sequence analysis of the isolated single strands was performed according to Maxam and Gilbert (10).

Isolation of the Large Subunit of Carbamoyl-Phosphate Synthetase. Purified *E. coli* carbamoyl-phosphate synthetase was dissociated with 1 M KSCN, and the large and small subunits were separated on Sephacryl S-300 (2).

Amino- and Carboxyl-Terminal Analyses. The NH_2-terminal residue of the large subunit was dansylated according to the method of Gros and Labouesse (11). After hydrolysis, the amino acid derivatives were identified by chromatography on polyamide sheets as described by Woods and Wang (12). The carboxyl-terminal sequence was determined by digestion of the protein with carboxypeptidase B, phenylmethylsulfonyl fluoride-treated carboxypeptidase A, and a mixture of the two (13). Products released from autodigestion were corrected with a control containing all components except the large subunit.

RESULTS

DNA Sequence of the *carB* Gene. The nucleotide sequence of the *carB* gene was derived from the 6-kbp insert of pMC40 (Fig. 1). Almost the entire sequence of both strands was obtained by using the restriction sites shown in Fig. 1 for 5′-end labeling. All of the labeled sites were crossed from distal sites. The nucleotide sequence presented in Fig. 2 represents approximately two-thirds of the cloned fragment. This sequence has a continuous reading frame of 3,219 nucleotides. None of the other five registers of the sequence revealed reading frames longer than about 300 base pairs. The 3,219-nucleotide reading frame starts with an ATG initiation codon at nucleotide +1 and ends with two consecutive termination signals (TAA TAG). A Shine–Dalgarno (14) ribosome binding site (A-G-G-A-G) is found at −15 to −11 of the initiation codon. The upstream region also has a RNA polymerase binding site (C-G-T-T-G-T-T-C-G-A-C-C) at −73 to −62, followed by a Pribnow box (T-C-G-A-G-T-T) at −55 to −49. Because transcription in *E. coli* generally starts at a purine ≈6–9 nucleotides downstream of the Pribnow box (15), these promoter-like sequences suggested a transcriptional start site at the A or G at position −43 or −42. S1 nuclease mapping of total *E. coli* RNA hybridized to a single-stranded probe extending beyond the proposed start of the message has confirmed that the 5′ end is at the G, 42 nucleotides upstream of the initiation codon (data not shown). Downstream of the gene centered at nucleotide position +3,300 is a run of nine Ts preceded by a sequence with dyad symmetry (A-C-G-G-T-T-C-T-C-T-G-A-T-G-A-G-G-A-C-C-G-T). Both features have been implicated in termination of transcription in *E. coli* (15).

Characterization of the *carB* Structural Gene. The open reading frame reported in Fig. 2 codes for a protein of 1,072 amino acid residues with a calculated M_r of 117,710. The amino acid composition of the encoded polypeptide is almost identical to two independently published (2, 16) amino acid compositions of the large subunit of *E. coli* carbamoyl-phosphate synthetase.

To verify that the protein encoded in the long open reading frame is the large subunit, we have determined the amino-terminal residue and a partial carboxyl-terminal sequence of the purified subunit. The amino-terminal residue was found to be proline, corresponding to the first amino acid after the ATG initiation codon of the gene. Presumably, the formylmethionine residue is cleaved post-translationally. Digestion of the large subunit with carboxypeptidase A failed to cleave the carboxyl-terminal residue. A similar digestion with carboxypeptidase B (specific for arginine and lysine) released lysine alone. A combination of the two enzymes yielded the sequence of the terminal seven residues. The established sequence was Glu-Met-His-Ala-$\frac{\text{Gln}}{\text{Asn}}$-Ile-Lys-COOH (Fig. 3). This sequence matches the last seven residues encoded by the gene.

The agreement in the amino acid composition, the amino-terminal residue, the carboxyl-terminal sequence, and the molecular weights provides strong evidence that the open reading frame is the *carB* gene coding for the large subunit of carbamoyl-phosphate synthetase.

Codon Utilization in the *carB* Gene. A list of codons and the frequency of their usage in the *carB* gene is presented in Table 1. The codon bias is typical of other *E. coli* genes coding for highly expressed proteins and, as previously noted (17, 18), re-

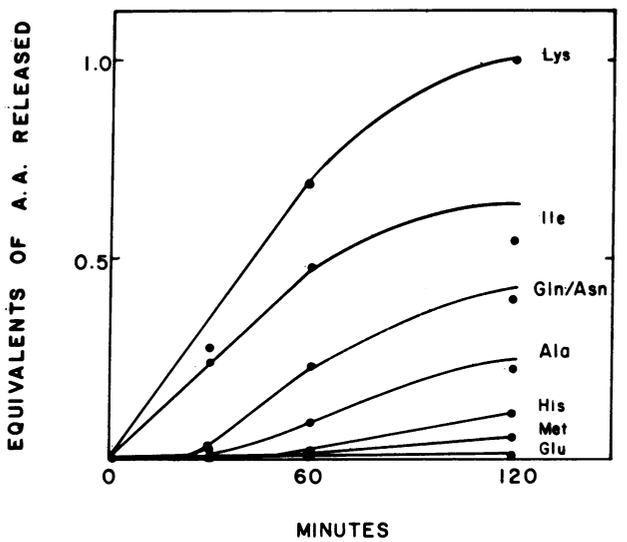

FIG. 3. Carboxyl-terminal sequence of the large subunit of carbamoyl-phosphate synthetase. The third residue from the carboxyl terminus is either glutamine or asparagine. However, the rate of release is consistent with its being glutamine. A.A., amino acid.

flects the abundance of the cognate tRNAs in this organism. The codon usage further supports a correct assignment of the reading frame.

Sequence Homology of the Amino- and Carboxyl-Terminal Halves of the Large Subunit of Carbamoyl-Phosphate Synthetase. A dot matrix (19) of the amino acid sequence encoded in the *carB* gene revealed an internal homology easily discernible to the eye. The best homology based both on absolute identities and conservative amino acid replacements is found between residues 1–400 and residues 553–933 (Fig. 4). These two regions of the protein exhibit 39% identical residues and an additional 25% conservative replacements. These values were obtained with only four minor adjustments for deletions and insertions. One of the deletions is in a region (residues 101–112) with a short duplication of six amino acids (L-E-$^{R-Q}_{E-F}$-G-V). The sequences included from residues 401 to 552 and 933 to 1,072 also appear to be homologous: 25% of the amino acids are identical and 20% represent conserved replacements. This homology was not evident in the dot matrix and required seven deletions or insertions over a considerably shorter sequence. In this region, also, two of the deletions occurred over a short-internal

Table 1. Frequency of codons in the *carB* gene

UUU Phe 8	UCU Ser 10	UAU Tyr 9	UGU Cys 6
UUC Phe 26	UCC Ser 18	UAC Tyr 22	UGC Cys 8
UUA Leu 3	UCA Ser 1	UAA Term 1	UGA Term 0
UUG Leu 5	UCG Ser 9	UAG Term 1	UGG Trp 4
CUU Leu 2	CCU Pro 1	CAU His 4	CGU Arg 37
CUC Leu 5	CCC Pro 0	CAC His 11	CGC Arg 31
CUA Leu 0	CCA Pro 8	CAA Gln 6	CGA Arg 0
CUG Leu 68	CCG Pro 35	CAG Gln 29	CGG Arg 1
AUU Ile 31	ACU Thr 12	AAU Asn 4	AGU Ser 3
AUC Ile 39	ACC Thr 40	AAC Asn 32	AGC Ser 5
AUA Ile 1	ACA Thr 1	AAA Lys 47	AGA Arg 0
AUG Met 30	ACG Thr 7	AAG Lys 9	AGG Arg 0
GUU Val 20	GCU Ala 15	GAU Asp 22	GGU Gly 28
GUC Val 13	GCC Ala 21	GAC Asp 42	GGC Gly 51
GUA Val 14	GCA Ala 13	GAA Glu 66	GGA Gly 2
GUG Val 47	GCG Ala 63	GAG Glu 25	GGG Gly 2

Term, termination. Initiation codon not included.

FIG. 4. Comparison of the deduced amino acid sequences of the amino-terminal half (residues 1–552) and the carboxyl-terminal half (residues 553–1,072) of the large subunit of carbamoyl-phosphate synthetase. The two sequences have been aligned for maximal homology.

duplication (L-V-$\overset{Q\text{-}I}{\underset{R\text{-}L}{}}$-E-E) (residues 462–473). The alignment of the last 150 residues shown in Fig. 4 achieves the best homology, although other alignments are possible.

The homology between the two halves of the protein was confirmed when the amino acid sequence was analyzed by using the ALIGN program (20). The computer analysis showed that the sequences 1–400 and 553–933 have an align score of 26.51. With the exception of human plasminogen and the α-2 chain of human haptoglobin (20), this is, to the best of our knowledge, the most significant score for internal homology yet reported.

There is also significant homology in the DNA sequence of the gene coding for the first and second halves of the protein. When the gene sequences were aligned to match the alignment used for the protein (Fig. 4), the overall homology between the two halves of the sequence was 47%.

DISCUSSION

We have used a recombinant plasmid with the *car* operon of *E. coli* K-12 to sequence the *carB* gene for the large subunit of carbamoyl-phosphate synthetase. The size, amino acid composition, and carboxyl- and amino-terminal features of the polypeptide encoded by *carB* match those of the isolated large subunit. The upstream and downstream regions of the gene have conventional sequences for transcription- and translation-related recognition sites and transcriptional termination. Both the upstream sequences and the results of S1 nuclease mapping indicate that *carB* can be separately transcribed and that the message size is 123–127 nucleotides longer than the reading frame. In addition to a transcript corresponding in size to the *carAB* operon, a smaller transcript of ≈3.4 kb was also detected in RNA transfer blots of total *E. coli* (RR1) RNA probed with a fragment of DNA from the *carB* gene (data not shown).

Analysis of the deduced amino acid sequence of the large subunit shows that the two halves of the protein are highly homologous. The homology is especially strong in the regions defined by residues 1–400 and 553–933. The sum of these two regions constitutes ≈73% of the total sequence. This finding provides strong evidence that the present form of the gene resulted from a single duplication of an ancestral gene half the size of *carB*. Although there are known examples of prokaryotic genes that have evolved by duplication, the *carB* gene is unusual in several respects. The duplication in *carB* is at least 10 times longer than the internal duplications found in bacterial ferredoxin, rubredoxin, and cytochrome c_3 (20). Secondly, the internal homology in *carB* is significantly higher than in the above proteins, suggesting either a more recent duplication or a higher degree of constraint on the primary structure. There are no data bearing on the identity of an ancestral gene. A possibility is that the precursor is related to the gene for carbamate kinase. This enzyme normally catalyzes ATP synthesis from ADP and carbamoyl phosphate formed in the arginine dihydrolase pathway of organisms such as *Mycoplasma* (21), *Streptococcus* (21), and *Pseudomonas* (22).

The homology of the two halves of the large subunit of carbamoyl-phosphate synthetase could explain some previous findings on the catalytic properties of the enzyme. Even though the enzyme is active as a monomer, it nonetheless exhibits allosteric behavior (23). Binding studies have shown one site for each of the ligands UMP, IMP, and ornithine (24). If the two halves of the protein fold as separate domains, these domains will be equivalent to two subunits hinged together (25). Binding of an allosteric ligand to one half could then induce a conformational change in the second half of the protein. Furthermore, there is evidence for two separate but nonidentical ATP binding sites (26, 27). The two sites could reflect the duplicated structure. It will be of interest to determine whether these sites are located in the two halves of the protein and have diverged with time, resulting in a concomitant change in function. The availability of the protein sequence and recognition of the duplicated structure should help in devising experiments to evaluate the function of each half of the protein in the catalysis of carbamoyl-phosphate synthesis and its regulation.

We thank Dr. Marjolaine Crabeel, Nicolas Glansdorff, and Andre Pierard for generously providing us with the *E. coli* clone, pMC40, Dr. Paul M. Anderson for a purified preparation of carbamoyl-phosphate synthetase, and Esther E. Widgren for her excellent assistance. We also thank Dr. Lois T. Hunt of the National Biomedical Research Foundation and Roy Smith for the computer analysis of the sequence. This work was supported by Grant GM 25846 from the National Institutes of Health.

1. Jones, M. E. (1972) in *Current Topics in Cellular Regulation*, eds. Horecker, B. L. & Stadtman, E. R. (Academic, New York), Vol. 6, pp. 227–265.
2. Trotta, P. P., Pinkus, L. M., Haschemeyer, R. H. & Meister, A. (1974) *J. Biol. Chem.* **249**, 492–499.
3. Pierard, A., Glansdorff, N., Mergeay, M. & Wiame, J. M. (1965) *J. Mol. Biol.* **14**, 23–36.
4. Mergeay, M., Gigot, D., Beckmann, J., Glansdorff, N. & Pierard, A. (1974) *Mol. Gen. Genet.* **133**, 299–316.
5. Glansdorff, N., Dambly, C., Palchaudhuri, S., Crabeel, M., Pierard, A. & Halleux, P. (1976) *J. Bacteriol.* **127**, 302–308.
6. Crabeel, M., Charlier, D., Weyens, G., Feller, A., Pierard, A. & Glansdorff, N. (1980) *J. Bacteriol.* **143**, 921–925.
7. Gigot, D., Crabeel, M., Feller, A., Charlier, D., Lissens, W., Glansdorff, N. & Pierard, A. (1980) *J. Bacteriol.* **143**, 914–920.
8. Lissens, W., Cunin, R., Kelker, N., Glansdorff, N. & Pierard, A. (1980) *J. Bacteriol.* **141**, 58–66.
9. Clewell, D. B. & Helinski, D. R. (1969) *Proc. Natl. Acad. Sci. USA* **62**, 1159–1166.
10. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
11. Gros, C. & Labouesse, B. (1969) *Eur. J. Biochem.* **7**, 463–470.
12. Woods, K. R. & Wang, K.-T. (1967) *Biochim. Biophys. Acta* **133**, 369–370.
13. Ambler, R. P. (1967) *Methods Enzymol.* **11**, 155–166.
14. Shine, J. & Dalgarno, L. (1974) *Proc. Natl. Acad. Sci. USA* **71**, 1342–1346.
15. Rosenberg, M. & Court, D. (1979) *Annu. Rev. Genet.* **13**, 319–353.
16. Matthews, S. L. & Anderson, P. M. (1972) *Biochemistry* **11**, 1176–1183.
17. Post, L. E., Strycharz, G. D., Nomura, M., Lewis, H. & Dennis, P. P. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 1697–1701.
18. Ikemura, T. (1981) *J. Mol. Biol.* **151**, 389–409.
19. Konkel, D. A., Maizel, J. V., Jr., & Leder, P. (1979) *Cell* **18**, 865–873.
20. Dayhoff, M. O., ed. (1978) in *Atlas of Protein Sequence and Structure* (Natl. Biomed. Res. Found., Washington, DC), Vol. 5, Suppl. 3, pp. 6, 359–362.
21. Raijman, L. & Jones, M. E. (1973) in *The Enzymes*, ed. Boyer, P. D. (Academic, New York), Vol. 9, pp. 97–119.
22. Abdelal, A. T., Bibb, W. F. & Nainan, O. (1982) *J. Bacteriol.* **151**, 1411–1419.
23. Anderson, P. M. (1977) *Biochemistry* **16**, 583–586.
24. Anderson, P. M. (1977) *Biochemistry* **16**, 587–593.
25. Richardson, J. S. (1981) *Adv. Protein Chem.* **34**, 251.
26. Powers, S. G. & Meister, A. (1978) *J. Biol. Chem.* **253**, 800–803.
27. Boettcher, B. R. & Meister, A. (1980) *J. Biol. Chem.* **255**, 7129–7133.