Characterization and Derivation of the Gene Coding for Mitochondrial Carbamyl Phosphate Synthetase I of Rat*

Hiroshi Nyunoya, Karen E. Brogle, Esther E. Widgren, and C. J. Lusty

From the Molecular Genetics Laboratory, The Public Health Research Institute of the City of New York, Inc., New York, New York 10016

The nucleotide sequence of rat carbamyl phosphate synthetase I mRNA has been determined from the complementary DNA. The mRNA comprises minimally 5,645 nucleotides and codes for a polypeptide of 164,564 Da corresponding to the precursor form of the rat liver enzyme. The primary sequence of mature rat carbamyl phosphate synthetase I indicates that the precursor is cleaved at one of two leucines at residues 38 or 39. The derived amino acid sequence of carbamyl phosphate synthetase I is homologous to the sequences of carbamyl phosphate synthetase of Escherichia coli and yeast. The sequence homology extends along the entire length of the rat polypeptide and encompasses the entire sequences of both the small and large subunits of the E. coli and yeast enzymes. The protein sequence data provide strong evidence that the carbamyl phosphate synthetase I gene of rat, the carAB gene of E. coli, and the CPA1 and CPA2 genes of yeast were derived from common ancestral genes.

Part of the rat carbamyl phosphate synthetase I gene has been characterized with two nonoverlapping phage clones spanning 28.7 kilobases of rat chromosomal DNA. This region contains 13 exons ranging in size from 68 to 195 base pairs and encodes the 453 carboxyl-terminal amino acids of the rat protein. Southern hybridization analysis of rat genomic DNA indicates the carbamyl phosphate synthetase I gene to be present in single copy.

Carbamyl phosphate synthetase I catalyzes the synthesis of carbamyl phosphate from NH₃, HCO₃⁻, and 2ATP and requires N-acetylglutamate as an allosteric activator (1). NH₃-dependent carbamyl phosphate synthetase is found only in ureotelic animals primarily in the liver and small intestine (2), where the enzyme plays an important role in removing excess NH₃ from the cell. The ability of carbamyl phosphate synthetase I to function efficiently at low concentrations of NH₃ is recognized to be an important step in the evolution of ureotelism (3, 4).

In recent studies, we have described the isolation of cDNA clones complementary to rat carbamyl phosphate synthetase I mRNA (5). Partial DNA sequence analysis of the cDNA has shown the gene encoding carbamyl phosphate synthetase I to be evolutionarily related to the genes encoding the glutamine-dependent carbamyl phosphate synthetases of Escherichia coli (6, 7) and yeast (8, 9). We have proposed that the gene encoding the NH₃-dependent carbamyl phosphate synthetase I evolved from the genes coding for the glutamine-dependent carbamyl phosphate synthetases by gene fusion and subsequent mutation of the glutamine binding and catalytic domain (5).

In this communication, we present the entire nucleotide sequence of carbamyl phosphate synthetase I mRNA. The set of the mRNA comprises more than 5,000 nucleotides, more than 50% of which the mitochondrial enzyme is a hybrid polypeptide encoded by a gene derived from two separate genes: one gene coding for a subunit that catalyzes glutamine amidine transfer (glutamine subunit) and the second coding for a synthetase subunit. Analysis of the rat gene shows it to occur in single copy and to be composed of multiple exons separated by long intervening sequences.

MATERIALS AND METHODS

DNA and Enzymes—Enzymes were purchased from P-L Biochemicals, New England Biolabs, and Amersham Corp. Plasmid DNA was isolated by the method of Birnboim and Doly (10) and purified by chromatography on Sepharose 6B.

Carbamyl Phosphate Synthetase cDNA Clones—The construction of rat (Wislar strain) liver cDNA libraries and the isolation of a set of recombinant plasmids (pKB4, pKB21, pHN107, and pHN234) carrying cDNA inserts complementary to rat liver carbamyl phosphate synthetase I mRNA have been described previously (5, 11). The cDNA insert (2.1 kb) in the plasmid pKB4 was confirmed by hybrid-selected in vitro translation to direct the synthesis of a 155-kDa polypeptide corresponding to the precursor form of rat liver carbamyl phosphate synthetase I (11). A near full-length CDN insert (5.3 kb) was carried in plasmid pHN234. Based on the size of carbamyl phosphate synthetase mRNA, the cDNA insert was only 300–400 nucleotides shorter than the expected length of the message. To select for clones containing the 5′-terminal sequences, we screened our Okayama-Berg rat liver cDNA library [1.8 × 10⁶ transfectants] by colony hybridization using as a probe an 800-bp PstI-XhoI fragment from the 5′ end of pHN234. This screen yielded four new clones, pHN291, pHN292, pHN293, and pHN295, with identical cDNA inserts of 1.5 kb. Restriction analysis of the four new clones showed that the cDNA inserts overlapped by 600 bp the 5′ region of the cDNA insert in pHN234. The four clones, however, contained 500 bp of a sequence not homologous to carbamyl phosphate synthetase mRNA. A complete screen of the cDNA library [6 × 10⁶ transformants] with a 175-bp RsaI-KpnI fragment from the 5′ proximal sequence of the cDNA insert of pHN234 yielded another 68 confirmed positive clones. Of 28 clones analyzed by restriction mapping, four contained cDNA inserts identical to the insert in pHN234, and 24 contained cDNA inserts identical to the inserts of pHN291, 292, 293, and 295.

DNA Sequence Analysis—DNA sequence analysis was performed by the method of Maxam and Gilbert (12). Restriction endonuclease fragments were isolated from agarose gels, 5′ end labeled with [γ-³²P] ATP (6000 Ci/mmol, Amersham Corp.), and polynucleotide kinase

*These studies were supported by Grant GM 25846 from the National Institutes of Health. 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.

1The abbreviations used are: kb, kilobase pairs; bp, base pairs; RF, replicative form; SDS, sodium dodecyl sulfate; PTH, phenylthiohydanto; dansyl, 5-dimethylaminonaphthalene-1-sulfonic.
and single strands were separated by electrophoresis on polyacrylamide gels. The nucleotide sequences of both isolated single strands were determined.

**Northern Blot Hybridization**—The size of carbamyl phosphate synthetase mRNA was estimated by Northern analysis of rat liver poly(A+) RNA after denaturation in 2.2 M formaldehyde and separation on agarose gels containing 1.1 M formaldehyde (13). RNA was transferred to nitrocellulose and hybridized as described by Thomas (14) with a radioactive probe prepared by nick translation. The gels were calibrated with a mixture of E. coli and yeast rRNAs as described (8), dX174 RF DNA (6386 bp) and mp13 RF DNA (~7200 bp).

**Amino Acid Composition, NH2-terminal sequence, and carboxyl-terminal amino acids of mature carbamyl phosphate synthetase** were determined on the rat liver enzyme judged by SDS-polyacrylamide gel electrophoresis to be 98% pure. Lyophilized salt-free samples (154 μg) of the protein were hydrolyzed with 6 N HCl (containing a crystal of phenol) in evacuated sealed tubes for 22 h at 110 °C. The amino acids were analyzed on an automatic amino acid analyzer (Beckman, model 119CL) which had been calibrated with 10 nmol of a standard mixture. The average of three determinations was used in the calculation. The NH2 terminus of the enzyme was identified according to the method of Gros and Labouesse (15). Dansyl-amino acids released by SDS-polyacrylamide gels were identified on polyamide thin layers (16). The NH2-terminal sequence of the protein was determined by automated Edman degradation with a Beckman Sequencer, model 890C, using the protein sequencing method of Blin and Stafford (22). Genomic DNA (20 μg) was digested with appropriate restriction endonucleases, separated by gel electrophoresis, and blotted on nitrocellulose sheets. The blots were exposed to Kodak XAR-5 film with an intensifying screen for 24-48 h.

**RESULTS**

**Carbamyl Phosphate Synthetase cDNA Clones**—Recombinant plasmids carrying cDNA inserts complementary to the entire sequence of rat liver carbamyl phosphate synthetase mRNA were isolated from two different cDNA libraries (5). The cDNA inserts of representative plasmids isolated from the two libraries are presented in Fig. 1. pHN234 had the longest cDNA insert starting with a poly(A) tract and including 4215 nucleotides of the coding sequence. This clone, however, lacked the 5' nontranslated leader and part of the NH2-terminal coding region. The 5' region missing from pHN234 was isolated on plasmid pHN291. The latter is one of four recombinant plasmids, all of which had identical cDNA inserts overlapping with the 5' proximal region of pHN234 but extending an additional 400 nucleotides further upstream. This upstream region included the entire NH2-terminal coding region of the mRNA as well as 139 nucleotides of 5' nontranslated leader. pHN291 as well as the other three
Rat Carbamyl Phosphate Synthetase I Gene

Fig. 3. cDNA sequence of carbamyl phosphate synthetase I mRNA. The sequence is that of the sense strand. The coding sequence of the mRNA begins at nucleotide +1 and ends at nucleotide +4500. The derived amino acid sequence is shown above the cDNA sequence. An inverted repeat centered over an upstream ATG codon in the 5′ leader is underlined. The arrows (↑) denote the experimentally determined NHB termini of mature carbamyl phosphate synthetase I. Downstream of the termination codon, the sequence AATAAA signaling polyadenylation is underlined. pA indicates the poly(A) addition site. As indicates the poly(A) tract which varied from 41–110 adenines in the different clones.
| Position | Backbone Amino Acid | Backbone One Letter Code | Position | Backbone Amino Acid | Backbone One Letter Code |
|----------|---------------------|--------------------------|----------|---------------------|--------------------------|
| 1        | Asp                 | D                        | 100      | Val                 | V                        |
| 2        | Asp                 | D                        | 200      | Asp                 | D                        |
| 3        | Cys                 | C                        | 300      | Val                 | V                        |
| 4        | Tyr                 | Y                        | 400      | Cys                 | C                        |
| 5        | Val                 | V                        | 500      | Asp                 | D                        |
| 6        | Val                 | V                        | 600      | Tyr                 | Y                        |
| 7        | Tyr                 | Y                        | 700      | Tyr                 | Y                        |
| 8        | Tyr                 | Y                        | 800      | Tyr                 | Y                        |
| 9        | Tyr                 | Y                        | 900      | Tyr                 | Y                        |
| 10       | Tyr                 | Y                        | 1000     | Tyr                 | Y                        |
| 11       | Tyr                 | Y                        | 1100     | Tyr                 | Y                        |
| 12       | Tyr                 | Y                        | 1200     | Tyr                 | Y                        |
| 13       | Tyr                 | Y                        | 1300     | Tyr                 | Y                        |

FIG. 3—continued.
identical plasmids contained another sequence unrelated to carbamyl phosphate synthetase. This sequence consisting of 528 nucleotides was found to have a reading frame encoding the sequence of low molecular weight kininogen (data not shown). The mechanism by which the hybrid cDNA was formed during reverse transcription is not clear at present.

Nucleotide Sequence of Carbamyl Phosphate Synthetase I mRNA—The nucleotide sequence of carbamyl phosphate synthetase I mRNA was derived from the cDNA inserts of the five recombinant clones, pKB4, pKB21, pHN107, pHN234, and pHN291, according to the strategy outlined in Fig. 2. The entire nucleotide sequence was confirmed from the complementary strands. All of the labeled sites were crossed with a set of overlapping fragments. The cDNA sequence from pKB4; a tract of 100-110 adenines was found in the cDNA sequence of the clone sequenced from the cDNA insert of pHN234.

The coding sequence begins with the ATG at nucleotide +1 and ends with a termination codon (TAG) at nucleotide +5406, followed at nucleotide +5419 with a second termination codon (TAG). The 4,500-nucleotide-long reading frame initiated with the ATG at nucleotide +1 codes for a polypeptide of 1,500 amino acids with a calculated molecular weight of 164,564. This value is in excellent agreement with the molecular weight (165,000) of the precursor form of rat liver carbamyl phosphate synthetase I (26). The ATG at nucleotide +1 is preceded by a purine (A) at −3 and a C at −4, as found in translational start sites of most eukaryotic genes (27, 28). In contrast, an upstream ATG located at nucleotide −4, as found in translational start sites of most eukaryotic genes (27, 28), is preceded by a pyrimidine (C) at −3 and 12 bp downstream is followed by a termination codon. Both features are common to other eukaryotic genes that contain ATG codons upstream of the AUG start codon (29).
**Table I**

| Codon usage in the carbamyl phosphate synthetase I mRNA

| The initiation codon is included in the tabulation. |

| Codon | Amino Acid |
|-------|------------|
| UUU   | Phe        |
| UUC   | Phe        |
| UUA   | Leu        |
| UUG   | Leu        |
| CUC   | Leu        |
| CCA   | Leu        |
| CUG   | Leu        |
| AUU   | Ile        |
| AUC   | Ile        |
| AUA   | Ile        |
| AUG   | Met        |
| GUC   | Val        |
| GCA   | Val        |
| GUA   | Val        |
| GUG   | Val        |

**Fig. 4.** Size of carbamyl phosphate synthetase I transcript.

Rat liver poly(A') RNA (10 μg) was denatured in 2.2 M formaldehyde, separated on 1% agarose gel under denaturing conditions, and transferred to nitrocellulose. The RNA was hybridized with a nick-translated cDNA probe internal to the coding region.

**Fig. 5.** NH2-terminal sequence of mature rat liver carbamyl phosphate synthetase I. Automated Edman degradation from the NH2-terminus of the purified protein (25 μg, 1 nmol) was performed in a Beckman Sequencer, model 890C.PTH derivatives were identified by high performance liquid chromatography, and the peak areas were quantitated by automatic integration and comparison to a mixture of known standards. The yield (nmol) of the PTH derivatives released in each cycle is indicated in the lines below each sequence. PTH-serine and PTH-threonine were qualitatively identified. The yield of PTH-serine was low (3.9 and 1.6 nmol were recovered in cycles 1 and 2, respectively). The yield of glutamine was derived from the sum of PTH-glutamine and PTH-glutamic acid.

**Table II**

| Amino acid composition of mature carbamyl phosphate synthetase I |
|---------------------------------------------------------------|
| Values are given as mol of residue/mol of enzyme. |

| Amino acid | Amino acid analysis | Predicted from DNA sequence |
|------------|---------------------|-----------------------------|
| Lysine     | 83 (90)             | 95                          |
| Histidine  | 25                  | 28                          |
| Arginine   | 50                  | 56                          |
| Aspartic acid | 148            | 72                          |
| Asparagine | 74                  |                             |
| Threonine  | 84†                 | 84                          |
| Serine     | 103†                | 102                         |
| Glutamic acid | 148             | 96                          |
| Glutamine  | 51                  |                             |
| Proline    | 67                  | 67                          |
| Glycine    | 111                 | 110                         |
| Alanine    | 116                 | 116                         |
| 1/2 Cys    | ND                  | 20                          |
| Valine     | 114                 | 120                         |
| Methionine | 37                  | 39                          |
| Isoleucine | 82                  | 95                          |
| Leucine    | 128                 | 128                         |
| Tyrosine   | 32                  | 36                          |
| Phenylalanine | 58             | 59                          |
| Tryptophan | ND                  | 14                          |

Total amino acid residues 1462

a The values (nearest integer) are averages of 3 determinations of a 22-h hydrolysate.
b Value obtained from different hydrolysate (22 h).
c Corrected for 5% decomposition.
d Corrected for 10% decomposition.
e ND, not determined.

**Codon Usage**—Codon usage in the carbamyl phosphate synthetase I mRNA is summarized in Table I. All of the possible 61 codons are utilized in the sequence. Some codons are used less frequently, particularly those codons with CG and UA pairs. GUA (Val), for example, is used only in four of 125 valine codons, and UGC (Ser) is used in six of 105 serine codons. The preferred codons are GUG (Val), AAG (Lys), CAG (Gln), and CUG (Leu).

**Northern Analysis**—The size of carbamyl phosphate synthetase I mRNA was estimated by Northern blot hybridization of rat liver poly(A') RNA with a nick-translated probe from the coding region of the cDNA. As shown in Fig. 4, radioautographs revealed a major radioactive band whose length was estimated to be 5.7 ± 0.3 kb. The size of the transcript is consistent with sequence analysis of the corre-
Fig. 6. Homology of the derived amino acid sequences of rat carbamyl phosphate synthetase I and the small and large subunits of E. coli carbamyl phosphate synthetase (CPS). The sequences of the rat carbamyl phosphate synthetase I precursor (top line) and of E. coli carbamyl phosphate synthetase (bottom line) were manually aligned for maximal amino acid identities, indicated by the vertical lines. Gaps in the sequence represent deletions or insertions. The arrows (1) denote the NH2 termini of the mature rat enzyme. The COOH terminus of the small subunit and the NH2 terminus of the large subunit of the E. coli enzyme are also shown in the figure. The reactive cysteine (Cys 269) in the glutamine-active site of the E. coli small subunit is marked with a large dot (2).
Fig. 7. Dot matrix analyses of the homology of the amino acid sequences of carbamyl phosphate synthetases of rat, yeast, and \textit{E. coli}. \textit{A}, matrix of the rat carbamyl phosphate synthetase I sequence versus the sequences of the glutamine synthetase subunits of \textit{E. coli} carbamyl phosphate synthetase. \textit{B}, matrix of the rat carbamyl phosphate synthetase I sequence versus the sequences of the small and large subunits of yeast carbamyl phosphate synthetase C, matrix plot of the internal homology in the synthetase component of rat carbamyl phosphate synthetase I. The main diagonal represents the homology between the sequences of the two proteins. The shorter diagonal above and below the main line represent the reciprocal internal homologies in the synthetase components of the three enzymes. \textit{E. coli} CPS, \textit{E. coli} carbamyl phosphate synthetase (6, 7); yeast CPS, yeast carbamyl phosphate synthetase (8, 9); rat CPSI, rat carbamyl phosphate synthetase I (this work).

Fig. 8. Alignment of previously proposed nucleotide-binding domain in NH\textsubscript{2}- and carboxyl-terminal halves of carbamyl phosphate synthetases (CPS) of \textit{E. coli}, yeast, and rat. The sequences shown in the upper three lines of the figure are located in the NH\textsubscript{2}-half of the various synthetase subunits; the sequences in the lower three lines of the figure are located in the COOH-half of the synthetases. The alignment of the \textit{E. coli} and yeast sequences is the same as proposed in Ref. 9. Identical and functionally conserved residues in each domain are connected by the vertical lines. Identical and conserved residues between the two domains are marked by the dots (\textbullet) at the bottom of the figure.

responding cDNA. Assuming a poly(A) tract of 100 adenosines and a 5' leader of 139 nucleotides, the minimal full-length transcript is calculated to be 5645 nucleotides.

Identification of Mature Carbamyl Phosphate Synthetase \textit{I}—Rat liver carbamyl phosphate synthetase I has previously been shown to be synthesized as a precursor 5 kDa larger than the mature enzyme (26). The NH\textsubscript{2}-terminal start of the mature protein was determined in the present study by analysis of the amino-terminal residue and also by a partial sequence of the NH\textsubscript{2}-terminal region of the mature protein. Dansylation and hydrolysis of the protein yielded dansyl-leucine and dansyl-serine, although serine appeared to be a minor component. A leucine NH\textsubscript{2} terminus has also been reported by Clarke (29) for rat liver carbamyl phosphate synthetase I prepared by a different procedure.

Two unambiguous NH\textsubscript{2}-terminal sequences displaced by a single amino acid were obtained by automated Edman degradation (Fig. 5). The first sequence started at Leu 39 and the second at Ser 40. Both sequences matched the sequence encoded by the mRNA for the 10 steps analyzed (compare Figs. 3 and 5). The average yield calculated from the stable amino acid derivatives released in cycles 3–5 indicated an approximately 1:1 ratio of the two NH\textsubscript{2}-terminal sequences. While these results suggest two adjacent processing sites involving cleavage on the carboxyl side of Leu 38 and/or Leu 39, the possibility of proteolytic removal of the NH\textsubscript{2}-terminal leucine during the isolation of the enzyme cannot be excluded.

Digestion of purified carbamyl phosphate synthetase with carboxypeptidase A released 1.6 mol of alanine/mol of protein and 0.37–0.43 mol/mol of (tyrosine, lysine, glycine, and serine). Since lysine is known to be a poor substrate for carboxypeptidase A, these data indicate the COOH-terminal sequence to start with two alanines followed by lysine, followed by glycine, tyrosine, and serine; the order of the latter three being indeterminate. These data are consistent with the sequence of carbamyl phosphate synthetase encoded in the cloned mRNA (Fig. 3).

The correct identification of the coding sequence is also supported by the experimentally determined amino acid composition and the molecular weight of mature carbamyl phosphate synthetase. The amino acid composition of the purified rat enzyme is in excellent agreement with the composition of the mature protein predicted from the mRNA sequence (Table II). Assuming that the mature protein starts at Leu 39 of the cDNA sequence, its overall length of 1462 amino acid residues corresponds to a molecular weight of 160,304. This agrees well with the molecular weight of 158,700 previously determined by physical methods (30).

Homology of Carbamyl Phosphate Synthetase \textit{I} and Glutamine-dependent Carbamyl Phosphate Synthetases of Yeast and \textit{E. coli}—The derived amino acid sequence of carbamyl phosphate synthetase \textit{I} is homologous to the amino acid sequences of both yeast and \textit{E. coli} carbamyl phosphate synthetases. The amino acid sequences of the rat and \textit{E. coli} proteins are shown in Fig. 6. The sequence homology between
Fig. 10. Location of exons in cDNA sequence and nucleotide sequence of exon-intron boundaries. The exons are numbered 1–13 in the direction of transcription of the gene. The nucleotide sequences of the exons and the intron-exon boundaries were determined by sequence analysis of restriction fragments of the rat genomic DNA inserts isolated in the phage clones, λ10.cps and X1.cps. The sequences of exons 9 and 10 were not determined. The sequence of exon 13 extends through the EcoRI site at the 3' end of the cloned DNA.

Fig. 11. Proposed evolutionary derivation of mammalian carbamyl phosphate synthetase I. The dark bar represents the length of the sequences derived from the gene of the ancestral glutamine subunit (8). The stippled bar indicates sequences derived from an unidentified gene (8). The open bar represents sequences derived from the gene of the ancestral kinase (6); the leader sequence.

the rat and bacterial enzymes extends along the entire length of the rat polypeptide and encompasses the entire amino acid sequence of both the small and the large subunits of the E. coli enzyme. The alignment of the two sequences shown in Fig. 6 required an average 1.8 deletions or insertions/100 amino acid residues. Of 1389 possible matches, 582 (41.9%) of the amino acid residues are identical. A similar alignment of the rat and yeast amino acid sequences exhibits 45% identical amino acids and required fewer deletions or insertions.

The evolutionary relationship of the rat, yeast, and E. coli enzymes is even more clearly demonstrated by dot matrix (31) analyses, where the sequence homology between the proteins is scored by using a mutation data matrix (32). As shown in Fig. 7, when the amino acid sequence of carbamyl phosphate synthetase I is compared to those of the E. coli and yeast proteins, three lines of identity are evident. The main diagonal shows the homology of the interspecies sequences. It extends over the entire length of the rat hybrid polypeptide and the small and large subunits of the E. coli and yeast enzymes. The shorter diagonal lines above and below the main diagonal show the conservation of the internal duplication of the synthetase subunit in the three species. The duplicated nature of the synthetase has previously been described for the E. coli and yeast enzymes (6, 9). The present data show the synthetase component of the rat enzyme to have the same duplication. The two halves of the rat synthetase component exhibit 23% identical residues. Comparable values for the yeast and E. coli synthetases are 28.5 and 35%, respectively. Since the greatest sequence conservation among the three species occurs in the NH2-terminal halves of the synthetase subunits, the decreasing homology is due primarily to divergence in the carboxy-terminal halves of the three enzymes.

Functional Domains—We have previously proposed three functional domains in E. coli and yeast carbamyl phosphate synthetases. Two of the domains were suggested to be involved in the binding of ATP (9), and the third domain was identified as the site of glutamine hydrolysis (8). The main feature of the glutamine hydrolytic site was the presence of a reactive cysteine residue, which had previously been shown to be essential for amidotransferase activity (33, 34). The absence of this cysteine residue in the sequence of rat carbamyl phosphate synthetase I led us to conclude that the glutamine site is modified such that it could no longer catalyze the hydrolysis of glutamine (5).

The postulated ATP-binding sites in the E. coli and yeast enzymes (9) are essentially conserved in the amino acid sequence of rat carbamyl phosphate synthetase I. The ATP-binding sites are located in two regions of the NH2-terminal half of the E. coli synthetase subunit. These sites have their counterparts in the carboxyl-terminal half. One of the sites (residues 302–352) exhibits sequence similarities to the ATP-binding site of phosphoglycerate kinase and to the predicted dinucleotide fold of glutamate dehydrogenase (9). In rat carbamyl phosphate synthetase I, the analogous domain is located between residues 718–768 (Fig. 8). This domain is highly conserved between the E. coli and rat sequences. Of 50 amino acid residues, 31 are identical and 10 represent functionally conserved substitutions. The corresponding domain (residues 1259–1302) in the carboxyl-terminal half of the rat synthetase component is also conserved (19 identities and 5 functionally conserved substitutions) (Fig. 8).

The second ATP-binding site (residues 152–210 in the NH2-
half and residues 698–756 in the COOH-half of the E. coli large subunit is based on sequence similarities with the glycine-rich loop of the ATP-binding site of adenylate kinase and the β-subunit of F1,ATPase (9). The analogous domains in the rat sequence (residues 571–626 and 1113–1171) are less conserved. Of 55 amino acid residues, 17 are identical (data not shown).

A fourth functional domain of rat carbamyl phosphate synthetase I is absent in both the E. coli and the yeast enzymes. This domain consists of 38 residues, starting from the NH2-terminal methionine. This sequence contains 8 basic amino acid residues, 1 acidic residue, and a Pro-Gly sequence 4 residues before the start of the mature enzyme, these features being common to signal sequences that direct proteins for import into mitochondria (35).

**Characterization of Rat Carbamyl Phosphate Synthetase I Gene**—A λ phage library of rat chromosomal DNA (18) was screened with a 770-bp cDNA probe containing the carboxy-terminal coding sequence of carbamyl phosphate synthetase I. The recombinant phage isolated from the screen were characterized by restriction analysis and grouped into two sets of clones carrying unique nonoverlapping EcoRI fragments of rat nuclear DNA. Two representative clones λ10,cps and λ1,cps with nuclear DNA inserts of 13.5 and 15.2 kb, respectively, were used to derive the structure of the gene region coding for the carboxy-terminal 453 amino acids of carbamyl phosphate synthetase I.

The exons in both clones were localized by Southern hybridization and by sequence analysis. The two clones contained the sequence of the cDNA starting from nucleotide 3142 through 4529. The coding region comprising 1359 nucleotides was ascertained to be split into 13 separate exons whose positions are shown in Fig. 9. With the exception of exons 9 and 10, all the remaining 11 exons were sequenced and the boundaries identified (Fig. 10). The localization of exons 9 and 10 was determined by restriction mapping and Southern hybridization within the limits indicated by the brackets in the figure.

Even though the two clones did not have overlapping sequences, it was possible to show that they represented contiguous segments of the gene. Southern hybridization analysis of rat genomic DNA digested with HindIII indicated a 2.6-kb fragment that hybridized to a cDNA probe containing the sequence encompassed by exons 3 into 7 (see Fig. 9). Based on the location of the HindIII sites in λ10,cps and λ1,cps, the 2.6-kb HindIII fragment should include the EcoRI site common to the two clones. Of course, we cannot exclude the possibility that the genomic DNA may have another EcoRI site (<100 bp) separating the two cloned fragments.

The sequences of the 11 exons which range from 68 to 195 bp in length showed no discrepancies with the sequence of the cDNA reported in Fig. 3. All the intron-exon boundaries conformed to the GT-AG rule (36) and to the consensus sequence compiled by Mount (37). Assuming the same 20:1 ratio for intron/exon lengths for the rest of the gene, we estimate that the rat carbamyl phosphate synthetase I gene should be included in approximately 110–120 kb of rat chromosomal DNA.

Extensive Southern hybridization analysis of rat genomic DNA indicates that the carbamyl phosphate synthetase I gene is present in single copy (data not shown).

**DISCUSSION**

Carbamyl phosphate synthetase of E. coli and of yeast are each composed of two different subunits encoded by two different genes (38). The smaller subunit catalyzes the transfer of the amide-N from glutamine to a catalytic center for carbamyl phosphate synthesis located on the larger synthetase subunit (39, 40). Previous studies of the E. coli (6, 7) and of the yeast (8, 9) genes have shown that the glutamine subunits are homologous and related to other amidotransferases (8). The genes for the larger synthetase subunits were found to have undergone a gene duplication resulting in a polypeptide with two homologous halves (6, 9). The present studies were undertaken to establish the relation of the mammalian NH2-utilizing enzyme carbamyl phosphate synthetase I to the glutamine-specific carbamyl phosphate synthetases of bacteria and yeast.

The entire coding sequence of the rat liver carbamyl phosphate synthetase I mRNA has been determined from overlapping cDNA clones. The message includes a short 5′ leader of at least 139 nucleotides, 4500 nucleotides of coding sequence, and a long 3′ nontranslated extension of 905 nucleotides. Analysis of two genomic clones selected from a λ library indicates that the gene contains multiple exons. The two clones studied represent a total of 28.7 kb of genomic DNA of which only 1388 bp are present in the mRNA.

The predicted primary translation product of the carbamyl phosphate synthetase I mRNA is a 164,564-Da protein. This precursor is cleaved at Leu 38 and/or Leu 39 to yield a mature carbonic anhydrase synthetase I of 160,304 Da. These molecular weights are in very good agreement with previously published data on the sizes of the precursor and mature forms of the rat enzyme estimated on SDS-polyacrylamide gels (26, 29, 30) and by sedimentation equilibrium in guanidine hydrochloride (30).

The derived amino acid sequence of rat carbamyl phosphate synthetase I has revealed several important facts that bear on its evolutionary origin. First, the complete sequence of the message has confirmed our previous conclusion that the mammalian enzyme is a fusion polypeptide of a glutamine amide transfer component and of a synthetase component (5). The amino acid sequences of the glutamine component located at the NH2 end of the enzyme and of the fused synthetase component are both homologous to the separate subunits of the E. coli and yeast enzymes. The homology is unambiguous and extends across the carboxyl-terminal end of the small subunit and the NH2 terminus of the larger synthetase subunit. This suggests that the mammalian gene was formed by a simple gene fusion event by either a mechanism similar to that proposed for the fusion of the his genes (41) or by nonhomologous (illegitimate) recombination (42, 43). Comparison of the protein sequences suggests the fusion probably occurred some time after the fungi diverged from the animal line, but before the separation of the chordate line, which includes cartilaginous and bony fishes, amphibians, and mammals. In all these organisms, arginine-specific carbamyl phosphate synthetase consists of a single 160-kDa polypeptide, suggesting fusion to be an early evolutionary event. A probable scheme depicting the evolution of the mammalian enzyme is presented in Fig. 11. This scheme incorporates our earlier suggestion that the small glutamine subunit of the glutamine-dependent carbamyl phosphate synthetase was derived from a fusion of an ancestral gene coding for the glutamine subunit with another unidentified gene (8). The synthetase subunit of the prokaryotic and fungal enzymes arose by a tandem duplication of an ancestral kinase (6). The present study provides strong evidence that the single gene of mammalian carbamyl phosphate synthetases was formed by a later fusion of the genes for the glutamine and synthetase subunits. Since all mammalian carbamyl phosphate synthetases are localized in the mitochondrion, their evolution requires the further acquir-
sition of a signal peptide directing the protein for import into the organelle.

Even though the two halves of the synthetase component have diverged to a greater extent in mammalian carbamyl phosphate synthetase (23%) than in yeast (28.5%) and E. coli (35%), certain regions are highly conserved. Among these is a domain which we previously proposed to be a possible nucleotide-binding site of the synthetase. This domain has been conserved in both the NH₂- (residues 718–768) and carboxyl-terminal (residues 1259–1302) halves of rat liver carbamyl phosphate synthetase I. The conservation of these sequences in the context of a general loss of sequence identity between the two halves of the mammalian synthetase provides additional evidence that both domains play an important catalytic function.

Another functional domain previously identified to be involved in the transfer of glutamine amide-N has also been conserved in the glutamine component of mammalian carbamyl phosphate synthetase. As noted previously, a cysteine residue shown in other amidotransferases to be essential for glutamine hydrolysis has been substituted by a serine (6). This substitution accounts for the inability of the mammalian synthetase to derive glutamine by the prokaryotic enzyme but requires additional evidence that both domains play an important catalytic function.

Acknowledgments—We express our appreciation to Dr. Gert Kreibich and E. D. Dharmgrongartaa for the protein sequence analysis, and to Dr. Lois T. Hunt of the National Biomedical Foundation for computer and dot matrix analyses of the protein sequences.

REFERENCES

1. Jones, M. E. (1965) Annu. Rev. Biochem. 34, 381–418
2. Jones, M. E., Anderson, A. D., Anderson, C., and Hodes, S. (1961) Arch. Biochem. Biophys. 95, 499-507
3. Jones, M. E., and Lipmann, F. (1960) Proc. Natl. Acad. Sci. U. S. A. 46, 1194–1205
4. Campbell, J. W. (1965) Nature 206, 1299–1301
5. Nyunoya, H., Broglie, K. E., and Lusty, C. J. (1965) Proc. Natl. Acad. Sci. U. S. A. 82, 2434–2440
6. Nyunoya, H., and Lusty, C. J. (1983) Proc. Natl. Acad. Sci. U. S. A. 80, 4629–4633
7. Piette, J., Nyunoya, H., Lusty, C. J., Cunin, R., Weyens, G., Grabbe, M., Charlier, D., Glansdorff, N., and Pérard, A. (1984) Proc. Natl. Acad. Sci. U. S. A. 81, 4134–4138
8. Nyunoya, H., and Lusty, C. J. (1984) J. Biol. Chem. 259, 9790–9798
9. Lusty, C. J., Widgren, E. E., Broglie, K. E., and Nyunoya, H. (1983) J. Biol. Chem. 258, 14466–14472
10. Birnboim, H. C., and Doly, J. (1979) Nucleic Acids Res. 7, 1513–1525
11. Ryall, J., Rachubinski, R. A., Nguyen, M., Rozen, R., Broglie, K. E., and Shore, G. C. (1984) J. Biol. Chem. 259, 9172–9176
12. Maxam, A. M., and Gilbert, W. (1980) Methods Enzymol. 65, 499–560
13. Maniatis, T., Frisch, E. F., and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual, pp. 292–293, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
14. Thomas, P. S. (1980) Proc. Natl. Acad. Sci. U. S. A. 77, 5201–5205
15. Gros, C., and Labouesse, B. (1969) Eur. J. Biochem. 7, 465–470
16. Woods, K. R., and Wang, K.-T. (1987) Biochim. Biophys. Acta 133, 369–370
17. Ambler, R. P. (1967) Methods Enzymol. 11, 155–166
18. Sargent, T. D., Wu, J.-R., Sahi-Trepát, J. M., Wallace, R. B., Reyes, A. A., and Bonner, J. (1979) Proc. Natl. Acad. Sci. U. S. A. 76, 3256–3260
19. Benton, W. D., and Davis, R. W. (1977) Science 196, 180–182
20. Blattner, F. R., Williams, B. G., Blech, A. E., Denniston-Thompson, K., Faber, H. E., Furlong, L.-A., Grunwald, D. J., Kiefer, D. D., Moore, D. B., Schum, J. W., Sheldon, E. L., and Smithies, O. (1977) Science 196, 161–169
21. Maniatis, T., Frisch, E. F., and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual, pp. 80–85, p. 373, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
22. Shapiro, N., and Stafford, D. W. (1976) Nucleic Acids Res. 3, 2303–2308
23. Southern, E. (1979) Methods Enzymol. 68, 152–176
24. Okayama, H., and Berg, P. (1982) Mol. Cell. Biol. 2, 161–170
25. Proudfoot, N. J., and Brownlee, G. G. (1974) Nature 252, 369–371
26. Shore, G. C., Carigian, P., and Raymond, Y. (1979) J. Biol. Chem. 254, 3141–3144
27. Kozak, M. (1984) Nucleic Acids Res. 12, 857–872
28. Johansen, H., Schümperli, D., and Rosenberg, M. (1984) Proc. Natl. Acad. Sci. U. S. A. 81, 7698–7702
29. Clarke, S. (1970) J. Biol. Chem. 251, 950–961
30. Lusty, C. J. (1978) Eur. J. Biochem. 85, 373–383
31. George, D. G., Yeh, L.-S. L., and Barker, W. C. (1983) Biochem. Biophys. Res. Commun. 115, 1061–1068
32. Dayhoff, M. O., Barker, W. C., and Hunt, L. T. (1983) Methods Enzymol. 191, 524–545
33. Tso, J. Y., Hermodson, M. A., and Zalkin, H. (1980) J. Biol. Chem. 255, 1451–1457
34. Dawid, I. B., French, T. C., and Buchanan, J. M. (1963) J. Biol. Chem. 238, 2178–2185
35. Kaput, J., Gotz, M., and Blobel, G. (1982) J. Biol. Chem. 257, 15054–15058
36. Breathnach, R., Benoist, C., O’Hare, K., Gannon, F., and Chambon, P. (1978) Proc. Natl. Acad. Sci. U. S. A. 75, 4853–4857
37. Mount, S. M. (1982) Nucleic Acids Res. 10, 459–472
38. Pérard, A., Grenson, M., Glansdorff, N., and Wanne, M. (1973) in The Enzymes of Glutamine Metabolism (Prusiner, S., and Stadtman, E. R., eds) pp. 483–503, Academic Press, New York
39. Trotta, P. P., Pinkus, L. M., Haschemeyer, R. H., and Meister, A. (1974) J. Biol. Chem. 249, 492–499
40. Pérard, A., and Schröter, B. (1975) J. Bacteriol. 134, 167–176
41. Yornuo, J., Kohno, T., and Roth, J. R. (1970) Nature 228, 820–824
42. Anderson, R. P., and Roth, J. R. (1977) Annu. Rev. Microbiol. 31, 473–505
43. Tisty, T. D., Albertini, A. M., and Miller, J. H. (1984) Cell 37, 217–224
44. Paluh, J. L., Zalkin, H., Betsch, D., and Weith, H. L. (1985) J. Biol. Chem. 260, 1889–1894
45. Alonso, E., and Rubino, V. (1983) Eur. J. Biochem. 135, 331–337
46. Casey, C. A., and Anderson, P. M. (1983) J. Biol. Chem. 258, 8723–8732
47. Tramell, P. R., and Campbell, J. W. (1971) Comp. Biochem. Physiol. 40B, 395–406
48. Anderson, P. M. (1980) Science 208, 291–293