Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 1754–1758, March 1988
Biochemistry

Molecular cloning and nucleotide sequence for the complete coding region of human UMP synthase

(pyrimidine biosynthesis/orotate phosphoribosyltransferase/orotidine-5'-monophosphate decarboxylase)

D. PARKER SUTTLE*†, BARBARA Y. BUGG†, JANET K. WINKLER*, AND JOHN J. KANALAS‡

*Department of Biochemical and Clinical Pharmacology, St. Jude Children's Research Hospital, Memphis, TN 38101; †Veterans Administration Medical Center, Memphis, TN 38104; and ‡Department of Pediatrics, University of Texas Health Science Center, San Antonio, TX 78284

Communicated by Mary Ellen Jones, November 10, 1987 (received for review August 13, 1987)

ABSTRACT The last two steps in the *de novo* biosynthesis of UMP are catalyzed by orotate phosphoribosyltransferase (OPRT; orotidine-5'-phosphate:pyrophosphate phosphoribosyltransferase; EC 2.4.2.10) and orotidine-5'-monophosphate decarboxylase (orotidine-5'-phosphate carboxy-lyase; EC 4.1.1.23). In mammals these two activities are found in a single, bifunctional protein called UMP synthase. A human T-lymphoblastic cell cDNA library constructed in λgt10 was screened with a UMP synthase-specific rat cDNA probe. Human UMP synthase cDNAs were isolated and then used to select UMP synthase gene fragments. The complete coding sequence of the mRNA for UMP synthase was determined by analysis of overlapping cDNA and genomic fragments. One of the cDNAs appears to have been synthesized from an incompletely or alternatively processed form of the UMP synthase mRNA. This cDNA lacks a poly(A) tail and has an extended 3'-nontranslated region that hybridizes with larger forms of the UMP synthase mRNA. The UMP synthase protein is composed of 480 amino acids with a molecular weight of 52,199. The two activities of UMP synthase reside in distinct domains encoded by the 3' and 5' halves of the mRNA. The COOH-terminal 258 amino acids of the human UMP synthase protein contain the orotidine-5'-monophosphate decarboxylase catalytic domain. This region is highly homologous to the mouse orotidine-5'-monophosphate decarboxylase sequence. The NH_2-terminal 214 amino acids contain the OPRT domain. There is amino acid homology between this protein domain and specific regions of the *Escherichia coli* OPRT. The human OPRT domain also contains the putative catalytic site common to other human phosphoribosyltransferases.

In the last two steps of the *de novo* pyrimidine biosynthetic pathway, ribose 5'-phosphate is added to orotic acid to form orotidine 5'-monophosphate, and orotidine 5'-monophosphate is decarboxylated to yield UMP. The two activities that catalyze these reactions are orotate phosphoribosyltransferase (OPRT; orotidine-5'-phosphate:pyrophosphate phosphoribosyltransferase; EC 2.4.2.10) and orotidine-5'-monophosphate decarboxylase (ODC; orotidine-5'-phosphate carboxy-lyase; EC 4.1.1.23). In bacteria and yeast, these two activities reside on separate proteins (1, 2). Early attempts to isolate and purify the two activities from mammalian sources invariably resulted in the copurification of both activities (3–7). Cells in culture that have developed resistance to 6-azauridine or pyrazofurin, inhibitors of ODC activity, have coordinately increased levels of both OPRT and ODC activities (8, 9), indicating that the activities are closely linked. A single protein of $M_r \approx 51{,}500$ that contains both the OPRT and ODC activities has now been purified from mouse ascites cells (10) and from human placenta (11). This bifunctional protein was designated UMP synthase.

The coordinated relationship of OPRT and ODC is strikingly apparent in the autosomal recessive disease orotic aciduria type I (see ref. 12 for review). Patients with this disease have very low levels of the two activities; hence, they are unable to convert orotic acid to UMP, and they excrete large quantities of orotic acid in their urine. The enzyme activity in deficient cells can be increased to near normal levels if the cells are grown in the presence of certain drugs or nucleotide analogues (13–16). This increased enzyme activity is caused by an increase in the amount of immunoprecipitable UMP synthase protein in cell extracts (16). The UMP synthase mRNA from fibroblast cells of patients with orotic aciduria type I is identical in size and quantity to the UMP synthase mRNA from normal fibroblasts (J.K.W. and D.P.S., unpublished observations). This indicates that the enzyme deficiency in the orotic aciduria patients results from a structural change in the UMP synthase protein. To identify the mutation(s) in the amino acid sequence of the defective protein, we have isolated and sequenced cDNA and genomic fragments containing the complete coding region of the normal UMP synthase protein.§ Analysis of the sequence of UMP synthase revealed a highly conserved protein, with distinct OPRT and ODC catalytic domains, each showing homology with the corresponding monofunctional proteins.

EXPERIMENTAL PROCEDURES

Selection of Human UMP Synthase cDNA and Genomic Clones. The cDNA library was constructed in the λgt10 cloning vector by using human T-lymphoblast (HPB-ALL, human peripheral blood acute lymphoblastic leukemia) mRNA, as described (17, 18). The plaques were screened on duplicate filters by the method of Benton and Davis (19) as described by Maniatis *et al.* (20). Positive plaques were isolated through two rounds of selection. The UMP synthase-specific inserts were isolated from these positive recombinant plaques by *Eco*RI digestion and subcloned into the *Eco*RI site of pBR322 for further sizing and restriction map analysis. For isolation of UMP synthase genomic fragments, a human genomic library prepared in the λ Charon 4A vector (21) was screened with the human UMP synthase-specific plasmid pHUSc22 (17) as described for the cDNA libraries.

DNA Sequencing. Specific restriction fragments of the UMP synthase inserts were isolated from low-melting-temperature agarose and subcloned into M13 cloning-sequencing vectors mp8/mp9 (22) or mp18/mp19 (23). The sequence of the fragments was determined by the dideoxy

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.

Abbreviations: ODC, orotidine-5'-monophosphate decarboxylase; OPRT, orotate phosphoribosyltransferase.

§The sequence reported in this paper is being deposited in the EMBL/GenBank data base (Bolt, Beranek, and Newman Laboratories, Cambridge, MA, and Eur. Mol. Biol. Lab., Heidelberg) (accession no. J03626).

chain-termination method, both as originally described (24) and with a quasi end-labeling adaptation (25). Analysis of the DNA sequence and the assignment of homology was aided by the DNA/protein sequence analysis software by James Pustell available through International Biotechnologies.

RESULTS AND DISCUSSION

Isolation and Determination of Human UMP Synthase Sequences. The UMP synthase cDNAs and genomic fragments that were isolated and the strategy used to determine the nucleotide sequence are illustrated in Fig. 1. The human cDNA library was screened with a UMP synthase-specific probe, pUMPS2, that had been isolated by differential hybridization from a rat hepatoma cDNA library (26). The initial screening of the cDNA library resulted in the isolation of a single plaque, λHUSc22, from which a 1.45-kilobase (kb) *Eco*RI insert was subcloned into pBR322 to produce pHUSc22 (17). The library was rescreened with the pHUSc22 insert as the hybridization probe; eight additional positive plaques were identified, and the isolated inserts were used to produce clones pHUSc32 to pHUSc39. The size of the cDNA inserts ranged from 1 to 1.7 kb. The DNA sequence of pHUSc22, -32, and -33 was determined. All three cDNA inserts had identical sequences in all areas that overlapped, pHUSc33 being the longest [1628 base pairs (bp)] and the only one that contained a poly(A) tail. However, pHUSc33 did not extend in the 5′ direction a sufficient distance to contain the ATG translation initiation site.

The sequence at the 5′ end of the coding region was determined by analysis of UMP synthase genomic fragments. To identify the fragment containing the 5′-most region of the UMP synthase mRNA, we analyzed *Eco*RI digests of the genomic clones by Southern blotting. A 2.7-kb *Eco*RI fragment in the clone λHUSg1 was identified by hybridization to the 500-bp *Eco*RI-*Xho* I fragment from the 5′ end of pHUSc33. This 2.7-kb genomic fragment was subcloned into pUC19 resulting in pHUSg1-5. A 545-bp *Eco*RI-*Pst* I fragment was selected by hybridization to the 5′-terminal 122-bp *Eco*RI-*Dde* I fragment of pHUSc33. Sequence analysis showed that the pHUSg1-5 genomic *Eco*RI-*Pst* I fragment contained an overlap of 131 identical bases (bp 129–260) with the 5′ end of pHUSc33. At bp 260 of the genomic fragment, there is the consensus AT/GT intron splice site, and the identity with the cDNA clone is interrupted by an intron sequence.

The nucleotide sequences of the pHUSc33 and pHUSc39 inserts and of the overlapping exon sequence of pHUSg1-5 with the predicted amino acid sequence for UMP synthase are shown in Fig. 2. The ATG initiation codon is located in the pHUSg1-5 fragment, just 26 bp on the 5′ side of the end of pHUSc33. The nucleotides immediately flanking the ATG initiator codon (CGACA*ATG*G) correspond well with the consensus initiation signal (CCACC*ATG*G) reported by Kozak (27, 28), having the required adenosine at position −3 and guanosine at position +4. The open reading frame, from bp 105 to bp 1544, codes for 480 amino acid residues that give a UMP synthase protein of M_r 52,199. In the mature message, the 3′-untranslated region consists of 210 bp containing a putative polyadenylylation signal sequence, AATAAG, 13 bp from the poly(A)-addition site.

Precursor UMP Synthase mRNA. The clone pHUSc39 contained an insert of approximately the same size as pHUSc33 and was initially presumed to be identical to pHUSc33. However, restriction endonuclease digestion of the two fragments indicated that pHUSc39 contained additional *Hin*dIII and *Bam*HI sites not found in pHUSc33. Complete sequencing of pHUSc39 revealed that it was identical to pHUSc33 from bp 482 to the 3′-poly(A)-addition site. However, pHUSc39 contained no poly(A) sequence and extended 485 bp further in the 3′-noncoding region.

To determine the origin of the extra 3′ sequence of pHUSc39, the *Hin*dIII-*Eco*RI fragment containing this sequence was used as a hybridization probe for UMP synthase

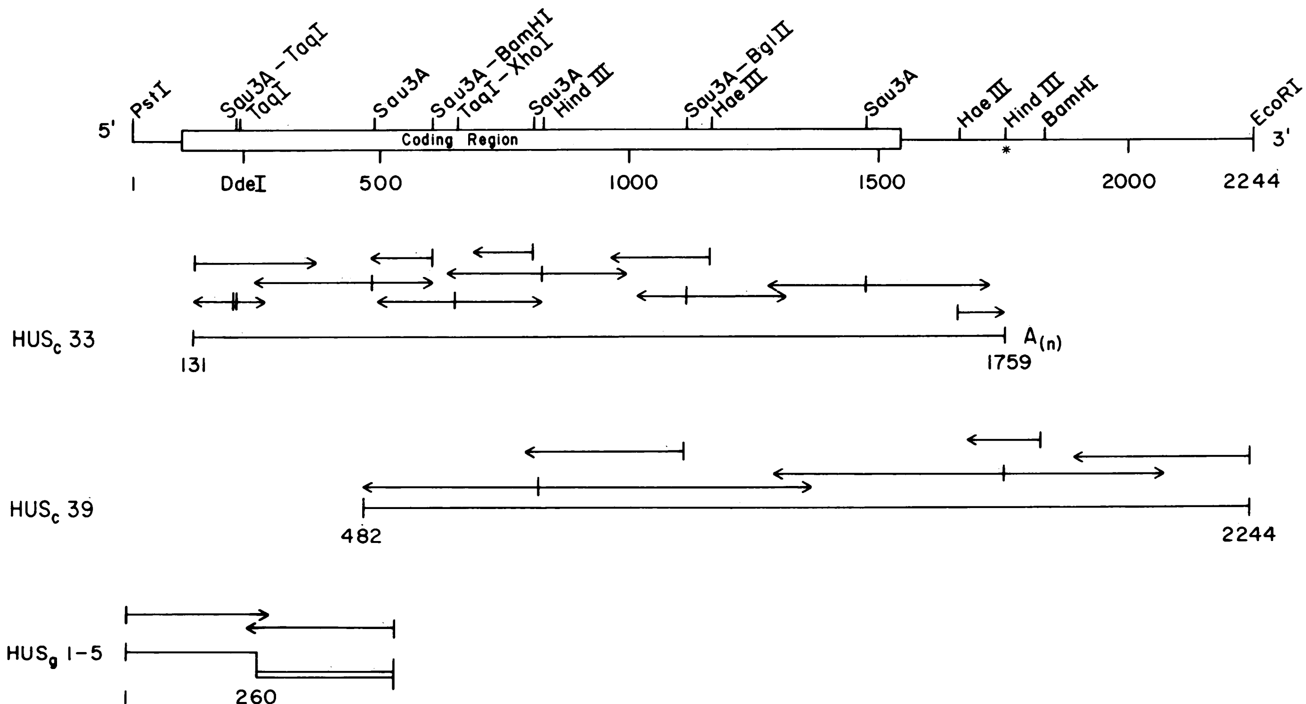

FIG. 1. Restriction endonuclease sites and fragments used in the DNA sequencing of UMP synthase. The open bar represents the UMP synthase protein coding region, and the solid lines are 5′- and 3′-noncoding regions. The asterisk indicates the site of polyadenylylation in the mature mRNA. The sequencing strategy is indicated by the direction and length of the arrows above the sequences. The open bar in HUSg1-5 represents an intron beginning at that site in the gene structure. The two cDNAs and the genomic fragment have identical sequences in the areas of overlap. The numbers at the ends of the sequenced fragments indicate the corresponding bases in the combined structure.

genomic fragments and mRNA. The 3′-*Hin*dIII–*Eco*RI fragment of pHUSc39 hybridized to a subset of the fragments identified by the pHUSc33 sequence on a Southern blot of *Eco*RI-digested UMP synthase genomic clones. These findings indicate that the 3′-noncoding sequence of pHUSc39 is contiguous with the coding region of UMP synthase in the gene structure. This 3′-end fragment also hybridized to mRNA from a human cell line that has amplified the UMP synthase gene ≈60-fold and shows a 40-fold increase in UMP synthase mRNA (17). The 3′ fragment hybridizes only slightly to the mature 2.0-kb UMP synthase mRNA but hybridizes more strongly to higher molecular weight forms of the mRNA. When pHUSc33 was used as the probe, these same larger mRNAs were identified in the amplified line, but the hybridization was weaker than to the mature mRNA. These data indicate that pHUSc39 is derived from a precursor or alternatively processed mRNA for UMP synthase.

A similar cDNA sequence was isolated from a mouse Ehrlich ascites cell cDNA library (29). It contained the ODC domain and a 3′-noncoding region that contained a truncated poly(A)-addition signal but, like pHUSc39, did not include the poly(A) sequence. Unlike HUSc39, hybridization of the mouse 3′ sequence to poly(A) mRNA isolated from the mouse ascites did not detect larger sizes of mRNA. It may be that the level of precursor mRNAs in the ascites sample is too low to be detected. The 3′-noncoding region of pHUSc39 showed detectable hybridization only to mRNA from the cell line that contained the 40-fold increase in UMP synthase mRNA.

UMP Sequence Homology. The amino acid sequence of the ODC enzyme from yeast has been deduced from the nucleotide sequence of the *URA3* gene (30). The ODC domain of the mouse ascites UMP synthase enzyme has also been cloned and sequenced (29). There is a high level of homology between the two sequences (52% identical residues) (29). Fig. 3 illustrates the homology between the ODC domain of the mouse ascites protein and the COOH-terminal half (amino acids 221–480) of the human UMP synthase sequence. These sequences are 90% identical, having only 26 nonidentical amino acids out of 259, with seven of the changes being neutral substitutions (i.e., substitution between alanine, glycine, isoleucine, leucine, and valine).

The NH_2-terminal 214 amino acids of human UMP synthase were compared with the deduced amino acid sequence of the OPRT protein from *Escherichia coli* (31) (Fig. 4). The sequences were aligned with gaps used to maximize the possible homology. The regions between residues 96 and 155 of the human protein and the *E. coli* protein have 43% identical amino acids. If homologies between amino acids with similar hydration potential are taken into account (32, 33), this region of the proteins is 73% identical. A region of 20 amino acids is highly conserved in a series of mammalian and prokaryotic enzymes that share the phosphoribosyltransferase function (34–36). An earlier study by Argos *et al.* (33) predicted that this region may be part of a catalytic domain. Fig. 5 shows the homology of this region in the OPRT domain with human hypoxanthine phosphoribosyltransferase and adenine phosphoribosyltransferase.

UMP Synthase Domain Structure. One characteristic of the structure of multifunctional proteins is that they exist as

```
1    CTGCAGACGAGGCAGGGAGAGGCGGGACTTCGCGGGCGAGACGTCATCGGGGCGCCGGACGCCGGGGCGC

71   CTGGGAGTTTGAAGCAAACAGGCAGCGCGCGACAATG GCG GTC GCT CGT GCA GCT TTG GGG
                                       Met Ala Val Ala Arg Ala Ala Leu Gly
                                       1

132  CCA TTG GTG ACG GGT CTG TAC GAC GTG CAG GCT TTC AAG TTT GGG GAC TTC
     Pro Leu Val Thr Gly Leu Tyr Asp Val Gln Ala Phe Lys Phe Gly Asp Phe
     10                                      20

183  GTG CTG AAG AGC GGG CTT TCC TCC CCC ATC TAC ATC GAT CTG CGG GGC ATC
     Val Leu Lys Ser Gly Leu Ser Ser Pro Ile Tyr Ile Asp Leu Arg Gly Ile
                 30                                      40

234  GTG TCT CGA CCG CGT CTT CTG AGT CAG GTT GCA GAT ATT TTA TTC CAA ACT
     Val Ser Arg Pro Arg Leu Leu Ser Gln Val Ala Asp Ile Leu Phe Gln Thr
                             50                                      60

285  GCC CAA AAT GCA GGC ATC AGT TTT GAC ACC GTG TGT GGA GTG CCT TAT ACA
     Ala Gln Asn Ala Gly Ile Ser Phe Asp Thr Val Cys Gly Val Pro Tyr Thr
                                         70

336  GCT TTG CCA TTG GCT ACA GTT ATC TGT TCA ACC AAT CAA ATT CCA ATG CTT
     Ala Leu Pro Leu Ala Thr Val Ile Cys Ser Thr Asn Gln Ile Pro Met Leu
             80                                      90

387  ATT AGA AGG AAA GAA ACA AAG GAT TAT GGA ACT AAG CGT CTT GTA GAA GGA
     Ile Arg Arg Lys Glu Thr Lys Asp Tyr Gly Thr Lys Arg Leu Val Glu Gly
                         100                                     110

438  ACT ATT AAT CCA GGA GAA ACC TGT TTA ATC ATT GAA GAT GTT GTC ACC AGT
     Thr Ile Asn Pro Gly Glu Thr Cys Leu Ile Ile Glu Asp Val Val Thr Ser
                                     120

489  GGA TCT AGT GTT TTG GAA ACT GTT GAG GTT CTT CAG AAG GAG GGC TTG AAG
     Gly Ser Ser Val Leu Glu Thr Val Glu Val Leu Gln Lys Glu Gly Leu Lys
         130                                     140

540  GTC ACT GAT GCC ATA GTG CTG TTG GAC AGA GAG CAG GGA GGC AAG GAC AAG
     Val Thr Asp Ala Ile Val Leu Leu Asp Arg Glu Gln Gly Gly Lys Asp Lys
                     150                                     160

591  TTG CAG GCG CAC GGG ATC CGC CTC CAC TCA GTG TGT ACA TTG TCC AAA ATG
     Leu Gln Ala His Gly Ile Arg Leu His Ser Val Cys Thr Leu Ser Lys Met
                                 170

642  CTG GAG ATT CTC GAG CAG CAG AAA AAA GTT GAT GCT GAG ACA GTT GGG AGA
     Leu Glu Ile Leu Glu Gln Gln Lys Lys Val Asp Ala Glu Thr Val Gly Arg
     180                                     190

693  GTG AAG AGG TTT ATT CAG GAG AAT GTC TTT GTG GCA GCG AAT CAT AAT GGT
     Val Lys Arg Phe Ile Gln Glu Asn Val Phe Val Ala Ala Asn His Asn Gly
                 200                                     210

744  TCT CCC CTT TCT ATA AAG GAA GCA CCC AAA GAA CTC AGC TTC GGT GCA CGT
     Ser Pro Leu Ser Ile Lys Glu Ala Pro Lys Glu Leu Ser Phe Gly Ala Arg
                             220                                     230

795  GCA GAG CTG CCC AGG ATC CAC CCA GTT GCA TCG AAG CTT CTC AGG CTT ATG
     Ala Glu Leu Pro Arg Ile His Pro Val Ala Ser Lys Leu Leu Arg Leu Met
                                         240

846  CAA AAG AAG GAG ACC AAT CTG TGT CTA TCT GCT GAT GTT TCA CTG GCC AGA
     Gln Lys Lys Glu Thr Asn Leu Cys Leu Ser Ala Asp Val Ser Leu Ala Arg
             250                                     260

897  GAG CTG TTG CAG CTA GCA GAT GCT TTA GGA CCT AGT ATC TGC ATG CTG AAG
     Glu Leu Leu Gln Leu Ala Asp Ala Leu Gly Pro Ser Ile Cys Met Leu Lys
                         270                                     280

948  ACT CAT GTA GAT ATT TTG AAT GAT TTT ACT CTG GAT GTG ATG AAG GAG TTG
     Thr His Val Asp Ile Leu Asn Asp Phe Thr Leu Asp Val Met Lys Glu Leu
                                     290

999  ATA ACT CTG GCA AAA TGC CAT GAG TTC TTG ATA TTT GAA GAC CGG AAG TTT
     Ile Thr Leu Ala Lys Cys His Glu Phe Leu Ile Phe Glu Asp Arg Lys Phe
         300                                     310

1050 GCA GAT ATA GGA AAC ACA GTG AAA AAG CAG TAT GAA GGA GGT ATC TTT AAA
     Ala Asp Ile Gly Asn Thr Val Lys Lys Gln Tyr Glu Gly Gly Ile Phe Lys
                     320                                     330

1101 ATA GCT TCC TGG GCA GAT CTA GTA AAT GCT CAC GTG GTG CCA GGC TCA GGA
     Ile Ala Ser Trp Ala Asp Leu Val Asn Ala His Val Val Pro Gly Ser Gly
                                 340

1152 GTT GTG AAA GGC CTG CAA GAA GTG GGC CTG CCT TTG CAT CGG GGG TGC CTC
     Val Val Lys Gly Leu Gln Glu Val Gly Leu Pro Leu His Arg Gly Cys Leu
     350                                     360

1203 CTT ATT GCG GAA ATG AGC TCC ACC GGC TCC CTG GCC ACT GGG GAC TAC ACT
     Leu Ile Ala Glu Met Ser Ser Thr Gly Ser Leu Ala Thr Gly Asp Tyr Thr
                 370                                     380

1254 AGA GCA GCG GTT AGA ATG GCT GAG GAG CAC TCT GAA TTT GTT GTT GGT TTT
     Arg Ala Ala Val Arg Met Ala Glu Glu His Ser Glu Phe Val Val Gly Phe
                             390                                     400

1305 ATT TCT GGC TCC CGA GTA AGC ATG AAA CCA GAA TTT CTT CAC TTG ACT CCA
     Ile Ser Gly Ser Arg Val Ser Met Lys Pro Glu Phe Leu His Leu Thr Pro
                                         410

1356 GGA GTT CAG TTG GAA GCA GGA GGA GAT AAT CTT GGC CAA CAG TAC AAT AGC
     Gly Val Gln Leu Glu Ala Gly Gly Asp Asn Leu Gly Gln Gln Tyr Asn Ser
             420                                     430

1407 CCA CAA GAA GTT ATT GGC AAA CGA GGT TCC GAT ATC ATC ATT GTA GGT CGT
     Pro Gln Glu Val Ile Gly Lys Arg Gly Ser Asp Ile Ile Ile Val Gly Arg
                         440                                     450

1458 GGC ATA ATC TCA GCA GCT GAT CGT CTG GAA GCA GCA GAG ATG TAC AGA AAA
     Gly Ile Ile Ser Ala Ala Asp Arg Leu Glu Ala Ala Glu Met Tyr Arg Lys
                                     460

1509 GCT GCT TGG GAA GCG TAT TTG AGT AGA CTT GGT GTT TGA GTGCTTCAGATACAT
     Ala Ala Trp Glu Ala Tyr Leu Ser Arg Leu Gly Val ---
         470                                     480

1563 TTTTCAGATACAATGTGAAGACATTGAAGATATGTGGTCCTCCTGAAAGTCACTGGCTGGAAATAAT

1630 CCAATTATTCCTGCTTGGATTCTTCCACAGGGCCTGTGTAAGAATGGGTTCTGGAGTTCTCATGGTC

                                                                   *
1697 TTTAGGAAATATTGAGTAATTTGTAATCACCGCATTGATACTATAATAAGTTCATTCTTAAGCTTGC

1764 TTTTTTTGAGACTGGTGTTTGTTAGACAGCCACAGTCCTGTCTGGGTTAGGGTCTTCCACATTTGAG

1831 GATCCTTCCTATCTCTCCATGGGACTAGACTGCTTTGTTATTCTATTTATTTTTTAATTTTTTTCGA

1898 GACAGGATCTCACTCTGTTGCCCAGGATGGAGTGCAGTGGTGAGATCACGGCTCATTGCAGCCTCGA

1965 CCTCCCAGGTGATCCTCCCACCTCAGCTTCCAGATTAGCTGGTGCTATAGGCATGCACCACCACGTC

2032 CATCTAAATTTCTTTATTATTTGTAGAGATGAGGTCTTGCCATGTTACCCAGGCTGGTCTCAACTCC

2099 TGGGCTCAAGCGATCCTCCTGCCTCAGTCTCTCAAAGTGCTGGGATTACAGGTGTGAGCCACTGTGC

2166 CCAGCCTAATTGCAGTAAGACAAAAATTCTAGGGCACCAAGAGGCTAAAGTCAGCACAGCTTTTCTT

2233 GTGTCCTGTATT
```

FIG. 2. Nucleotide and deduced amino acid sequences of human UMP synthase. The consensus sequence around the initiator ATG (positions 100–108) and the putative poly(A)-addition signal (positions 1741–1745) are underlined. The poly(A) sequence begins at bp 1759 in the HUSc33 cDNA and is marked with an asterisk. The 5′ ends of the HUSc33 cDNA (arrow) and the HUSc39 cDNA (arrowhead) are indicated at bp 131 and 482, respectively. The position of the intron in fragment HUSgl-5 (solid triangle) is indicated at bp 260. The nucleotide numbers are indicated at the beginning of each line, and the amino acid sequence numbers are indicated below the amino acids.

```
221   APKELSFGAR AELPRIHPVA SKLLRLMQKK ETNLCLSADV SLARELLQLA
       *          ** *                            *
4     ACKELSFGAR AELPGTHPLA SKLLRLMQKK ETNLCLSADV SEARELLQLA

271   DALGPSICML KTHVDILNDF TLDVMKELIT LAKCHEFLIF EDRKFADIGN
                                 * **   *
54    DALGPSICML KTHVDILNDF TLDVMEELTA LAKRHEFLIF EDRKFADIGN

321   TVKKQYEGGI FKIASWADLV NAHVVPGSGV VKGLQEVGLP LHRGCLLIAE
             * *         *                             *
104   TVKKQYESGT FKIASWADIV NAHVVPGSGV VKGLQEVGLP LHRACLLIAE

371   MSSTGSLATG DYTRAAVRMA EEHSEFVVGF ISGSRVSMKP EFLHLTPGVQ
         *        *  *  *    *   *
154   MSSAGSLATG NYTKAAVGMA EEHCEFVIGF ISGSRVSMKP EFLHLTPGVQ

421   LEAGGDNLGQ QYNSPQEVIG KRGSDIIIVG RGIISAADRL EAAEMYRKAA
        *   *                     *    ** *
204   LETGGDHLGQ QYNSPQEVIG KRGSDVIIVG RGILAAANRL EAAEMYRKAA

471   WEAYLSRLGV  end
               * *
254   WEAYLSRLAV Qend
```

FIG. 3. Amino acid sequence homology between the human and mouse ODC domains. The COOH-terminal half of the human UMP synthase deduced amino acid sequence (upper sequence) was aligned with the sequence of the ODC domain of the mouse (29). Only the amino acids marked with an asterisk are different in the two sequences. The amino acid numbers for the human sequence correspond to the numbering shown in Fig. 2. The amino acid numbers for the mouse are as shown in figure 7 of ref. 29.

autonomous domains of catalytic activity linked together by a polypeptide tether that may be sensitive to protease action. Therefore, with the use of limited proteolytic treatment, multifunctional proteins may be digested into polypeptide domains that possess a single catalytic activity (37–40). For UMP synthase it has been shown that limited treatment with trypsin or elastase yields protein fragments that contain only OPRT or ODC activity (6, 40). Thus the UMP synthase protein may be composed of three regions: OPRT domain, connector peptide, and ODC domain. The regions of UMP

```
1     mavaraalgpLvTglyDVqaFkFgDfVLKsGls-------SPiYIdlrGIVsrpR--
                          *     *** *          ** *    *     *
1               MkPyqrQFieFaLsKqVLKfGeftlksgrkSP-YFfnaGLFntgRdl

49    -LLsQV-ADILFQTAqNAgIsFdTV-cGVP---YTALpLAtvicstnQIPmLIRRKE
       **    *  *           *     * *    **  **            *   ***
47    aLLgRFyAEALVDSGiEFdLlFgPAykGIPiatTTAVaLAehh--dlDLPyCFNRKE

100   tKDyGtk-RLVeGTInpGEtCLiIEDVVTsGSsVlETVEVLQkEGLkvTdAIVL--L
       ** *    ** *    *    *   ** * *    *   *   *   *    * * ** *
101   aKDhGeggNLV-GSAlqGRvML-VDDVITaGT-ArESMEIIQaNGA--TlAGVLisL

154   DREQgGKdKLqAhgirlH-SvCt-LSkmleI-LEqqkkvdaEtvgrVKRFIqenVfV
      **   *     *            *    *    * *         *              *
153   DRQErGRgEIsAiqeveRdYnCkvISi---ItLKdliayleEkpemAEHLA--aVkA

208   aaNHnGs        214
          *
205   yrEEfGv end    212
```

FIG. 4. Comparison of deduced amino acid sequence for OPRT from human and *E. coli*. The upper sequence is the NH_2-terminal 214 amino acids of human UMP synthase. The lower sequence is for the *E. coli pyrE* gene product (31). The asterisks mark identical residues in the two sequences. Capital letters indicate similar residues based on the hydration potential of each amino acid (32) as reported by Argos *et al.* (33). The amino acids were divided into the following three groups as described by Ohmstede *et al.* (29): strongly water soluble (glycine, alanine, leucine, isoleucine, valine, phenylalanine, cysteine, and methionine); equally water-vapor soluble (threonine, proline, serine, tryptophan, and tyrosine); and vapor soluble (glutamine, asparagine, lysine, histidine, glutamic acid, aspartic acid, and arginine). The dash indicates a gap in the sequence introduced to produce maximum homology.

```
human OPRT   116  [G] E T [C] [L I I E D V V] [T] [S G] [S] [S] [V L E T]
human HGPRT  126  [G] K N  [V  L I V E D I I] [D] [T G] [K] [T M Q] T L
human APRT   120  [G] Q R  [V  V V V D D L L] [A] [T G] [G] [T M N] A A
```

FIG. 5. Sequence homology of human phosphoribosyltransferases. The putative catalytic site is aligned for three human phosphoribosyltransferase enzymes (OPRT; HGPRT, hypoxanthine phosphoribosyltransferase; APRT, adenine phosphoribosyltransferase). Positions of identical or neutral amino acid substitutions are boxed. The numbers indicate the position of the glycine residue in the respective protein sequences.

synthase that exhibit homology with the OPRT and ODC monofunctional proteins of *E. coli* (31) and yeast (30) and with the ODC domain of the mouse UMP synthase (29) are indicated on a hydropathy plot in Fig. 6. The connector peptide would be localized to a central segment (including amino acids 192–237) that appears to be free of significant stretches of either hydrophilic or hydrophobic residues. The size of this proposed polypeptide linker joining the two catalytic domains and its role in UMP synthase structure and function have not been determined. It may be that the connector does not significantly affect the integrity of the

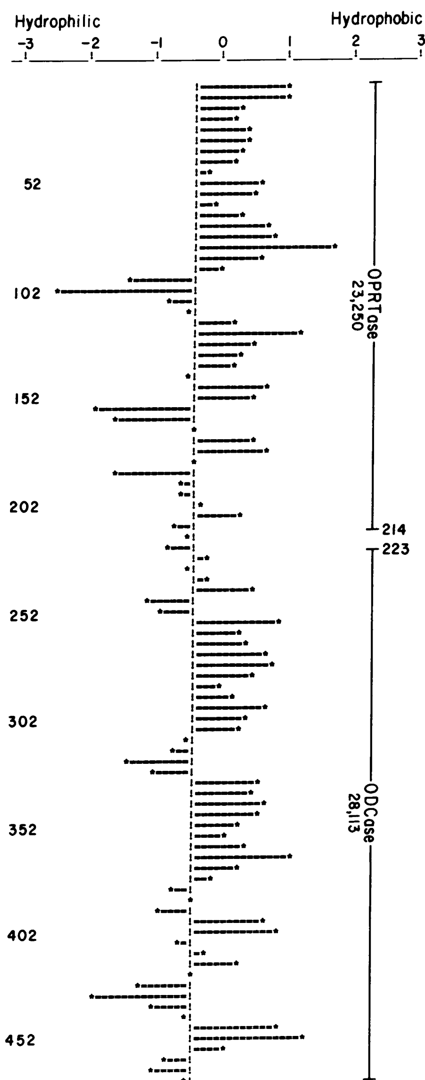

FIG. 6. Hydropathy plot of the UMP synthase amino acid sequence. The plot was derived with the program of J. Pustell available through International Biotechnologies. The hydropathic indices of nine consecutive amino acids are averaged to generate each point. The numbers of the amino acids are indicated at the bottom.

domains. The two domains of UMP synthase appear in many respects to be independent. For example, after proteolytic cleavage, the OPRT and ODC activities can be separated, and an active ODC domain can be isolated (6, 40). A functional mouse ODC domain can also be produced from a bacterial or yeast expression vector containing only the 3′ portion of the UMP synthase cDNA (29, 41). The homology of the putative catalytic site in the OPRT domain with that of other monofunctional phosphoribosyltransferases indicates that a similar catalytic mechanism may be used.

However, it remains intriguing that, in the autosomal recessive disease orotic aciduria type I, both activities of UMP synthase are deficient. The fact that growth in the presence of nucleoside analogues can partially restore both activities of UMP synthase in the deficient cells (13–16) argues against altered catalytic sites, as has been documented with other enzyme deficiencies such as hypoxanthine phosphoribosyltransferase (42). One possibility is that the connector sequence may in fact play a role in the correct folding of the domains, the necessary juxtaposition of the domains within a single polypeptide, or the proper interaction of the domains in the active homodimer form of the protein. A mutation in this region of the protein might destabilize the protein structure and result in increased temperature lability or proteolytic degradation.

Studies of the bifunctional yeast tryptophan synthetase protein indicate a region of 66 amino acids that may form a connector peptide between the α and β subunits (43). Initial experiments show that even a short deletion of 2 amino acids can result in significant loss of activity for both the subunits. Grayson and Evans (38) have isolated the aspartate transcarbamylase domain of the multifunctional protein comprising the first three activities of the *de novo* pyrimidine pathway. Their studies indicate that the aspartate transcarbamylase domain may have similar structural and kinetic properties as the intact multifunctional protein. However, the transcarbamylase activity of the intact protein is approximately six times more stable than that of the isolated domain in the absence of substrates and cryoprotectants, indicating a role for the connector region in the tertiary structure of the protein (38). Since we have determined the complete sequence of the UMP synthase coding region, alterations can be identified in the sequence of UMP synthase cDNAs isolated from orotic aciduria cell lines. The relationship of the alterations and the resulting deficiency of both OPRT and ODC activities in the UMP synthase protein can then be further characterized.

Note Added in Proof. A DNA segment containing the complete coding region of UMP synthase has been constructed from the overlapping cDNA and genomic fragments and inserted into a mammalian expression vector. Urd$^-$C cells [Chinese hamster ovary cells that are deficient in both activities of UMP synthase and require uridine-supplemented medium for growth (44)] were transfected with this vector. Colonies were selected in medium without uridine, and enzyme assays indicated that the transfected cells contained UMP synthase activity at levels up to four times higher than wild-type Chinese hamster lung cells. These results indicate that the selected sequence does encode a functional UMP synthase protein with both OPRT and ODC activity.

This work was supported by Grant DK-36747 from the National Institutes of Health, a Merit Review Award from the Veterans Administration Research Service, and by American Lebanese Syrian Associated Charities.

1. O'Donovan, G. A. & Neuhard, J. (1970) *Bacteriol. Rev.* **34**, 278–343.
2. Yoshimoto, A., Amaya, T., Kobayashi, K. & Tomita, K. (1978) *Methods Enzymol.* **51**, 69–74.
3. Hatfield, D. & Wyngaarden, J. B. (1964) *J. Biol. Chem.* **239**, 2580–2586.
4. Kasbekar, D. K., Nagabhushanam, A. & Greenburg, D. M. (1964) *J. Biol. Chem.* **239**, 4245–4249.
5. Appel, S. H. (1968) *J. Biol. Chem.* **243**, 3924–3929.
6. Reyes, P. & Guganig, M. E. (1975) *J. Biol. Chem.* **250**, 5097–5108.
7. Kavipurapu, P. R. & Jones, M. E. (1976) *J. Biol. Chem.* **251**, 5589–5599.
8. Suttle, D. P. & Stark, G. R. (1979) *J. Biol. Chem.* **254**, 4602–4607.
9. Levinson, B. B., Ullman, B. & Martin, D. W., Jr. (1979) *J. Biol. Chem.* **254**, 4396–4401.
10. McClard, R. W., Black, M. J., Livingstone, L. R. & Jones, M. E. (1980) *Biochemistry* **19**, 4699–4706.
11. Livingstone, L. R. & Jones, M. E. (1987) *J. Biol. Chem.* **262**, 15726–15733.
12. Kelley, W. N. (1983) in *The Metabolic Basis of Inherited Disease*, eds. Stanbury, J. B., Wyngaarden, J. B., Fredrickson, D. S., Goldstein, J. L. & Brown, M. S. (McGraw-Hill, New York), pp. 1227–1248.
13. Pinsky, L. & Krooth, R. S. (1967) *Proc. Natl. Acad. Sci. USA* **57**, 925–932.
14. Pinsky, L. & Krooth, R. S. (1967) *Proc. Natl. Acad. Sci. USA* **57**, 1267–1274.
15. Krooth, R. S., Lam, G. F. M. & Klang, S. Y. C. (1974) *Cell* **3**, 55–57.
16. Perry, M. E. & Jones, M. E. (1987) *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **46**, 309.
17. Kanalas, J. J., Hutton, J. J. & Suttle, D. P. (1985) *Somatic Cell Mol. Genet.* **11**, 359–369.
18. Wiginton, D. A., Adrian, G. S., Friedman, R., Suttle, D. P. & Hutton, J. J. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 7481–7485.
19. Benton, W. D. & Davis, R. W. (1977) *Science* **196**, 180–182.
20. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).
21. Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, G. & Maniatis, T. (1978) *Cell* **15**, 1157–1174.
22. Messing, J. & Vieira, J. (1982) *Gene* **19**, 269–276.
23. Norrander, J., Kempe, T. & Messing, J. (1983) *Gene* **26**, 101–106.
24. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
25. Duncan, C. (1985) *New Engl. Nucl. Prod. News* **4**, 6–7.
26. Kanalas, J. J. & Suttle, D. P. (1984) *J. Biol. Chem.* **259**, 1848–1853.
27. Kozak, M. (1984) *Nucleic Acids Res.* **12**, 857–872.
28. Kozak, M. (1986) *Cell* **44**, 283–292.
29. Ohmstede, C., Langdon, S. D., Chae, C. & Jones, M. E. (1986) *J. Biol. Chem.* **261**, 4276–4282.
30. Rose, M., Grisafi, P. & Botstein, D. (1984) *Gene* **29**, 113–124.
31. Poulsen, P., Jensen, K. F., Valentin-Hansen, P., Carlsson, P. & Lundberg, L. G. (1983) *Eur. J. Biochem.* **135**, 223–229.
32. Wolfenden, R. V., Cullis, P. M. & Southgate, C. C. F. (1979) *Science* **206**, 575–577.
33. Argos, P., Hanei, M., Wilson, J. M. & Kelly, W. N. (1983) *J. Biol. Chem.* **258**, 6450–6457.
34. Dush, M. K., Sikela, J. M., Khan, S. A., Tischfield, J. A. & Stambrook, P. J. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 2731–2735.
35. Hershey, H. V. & Taylor, M. W. (1986) *Gene* **43**, 287–293.
36. Broderick, T. P., Schaff, D. A., Bertino, A. M., Dush, M. K., Tischfield, J. A. & Stambrook, P. J. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 3349–3353.
37. Kirschner, K. & Bisswanger, H. (1976) *Annu. Rev. Biochem.* **45**, 143–166.
38. Grayson, D. R. & Evans, D. R. (1983) *J. Biol. Chem.* **258**, 4123–4129.
39. Rumsby, P. C., Campbell, P. C., Niswander, L. A. & Davidson, J. N. (1984) *Biochem. J.* **217**, 435–440.
40. Floyd, E. E. & Jones, M. E. (1985) *J. Biol. Chem.* **260**, 9443–9451.
41. Langdon, S. D. & Jones, M. E. (1987) *J. Biol. Chem.* **262**, 13359–13365.
42. Wilson, J. M., Young, A. B. & Kelly, W. N. (1983) *New Engl. J. Med.* **309**, 900–910.
43. Crawford, I. P., Clarke, M., van Cleeput, M. & Yanofsky, C. (1987) *J. Biol. Chem.* **262**, 239–244.
44. Patterson, D. (1980) *Somatic Cell Genet.* **6**, 101–114.