De novo purine nucleotide biosynthesis: cloning of human and avian cDNAs encoding the trifunctional glycinamide ribonucleotide synthetase-aminoimidazole ribonucleotide synthetase-glycinamide ribonucleotide transformylase by functional complementation in *E. coli*

Junko Aimi, Hong Qiu, John Williams, Howard Zalkin and Jack E. Dixon*
Purdue University, Department of Biochemistry, West Lafayette, IN 47907, USA

Received May 18, 1990; Revised and Accepted August 8, 1990

EMBL accession nos X54199, X54200

ABSTRACT

The trifunctional enzyme encoding glycinamide ribonucleotide synthetase (GARS)-aminoimidazole ribonucleotide synthetase (AIRS)-glycinamide ribonucleotide transformylase (GART) was cloned by functional complementation of an *E. coli* mutant using an avian liver cDNA expression library. In *E. coli*, genes encoding these separate activities (*purD*, *purM*, and *purN*, respectively) produce three proteins. The avian cDNA, in contrast, encodes a single polypeptide with all three enzyme activities. Using the avian DNA as a probe, a cDNA encoding the complete coding sequence of the trifunctional human enzyme was also isolated and sequenced. The deduced amino acid sequence of the human and avian polypeptides show extensive sequence homologies to the bacterial *purD*, *purM*, and *purN* encoded proteins. Avian and human liver RNAs appear to encode both a trifunctional enzyme (GARS-AIRS-GART) as well as an RNA which encodes only GARS. The trifunctional protein has been implicated in the pathology of Down's Syndrome and molecular tools are now available to explore this hypothesis. Initial efforts to compare the expression of GARS-AIRS-GART between a normal fibroblast cell line and a Down's Syndrome cell line indicate that the levels of RNA are similar.

INTRODUCTION

The biosynthesis of nucleotides plays an integral role in many biochemical processes. Nucleotides and their derivatives serve as precursors for DNA and RNA synthesis, as activated intermediates in biosynthetic pathways, as components of coenzymes, as a universal source of energy, and as metabolic regulators in higher organisms. The genes catalyzing the individual steps in *de novo* purine nucleotide biosynthesis have been analyzed in some detail in bacteria (1-6). Similar detailed studies have not been carried out in higher eucaryotes.

In order to gain a better understanding of the regulation of *de*

novo purine nucleotide biosynthesis and to explore the association of enzymes in this pathway with certain human disease states, we have developed a strategy to clone these enzymes using functional complementation of *E. coli pur* mutants with eucaryotic expression vectors. By design, this approach provides a functional protein which could also be used as a target for drug screening assays or to isolate the eucaryotic enzyme from *E. coli* extracts. This strategy has now been successfully utilized to clone the mRNAs for 8 of the 10 steps leading to the synthesis of IMP. These enzymes include adenylosuccinate lyase (corresponding to *E. coli* gene *purB*) (7), AICAR transformylase-IMP synthetase (*purH*)¹, the bifunctional AIR carboxylase-SAICAR synthetase (*purEK* and *purC*) (8), and here we report the cloning of the trifunctional enzyme encoding glycinamide ribonucleotide synthetase (GARS)-aminoimidazole ribonucleotide synthetase (AIRS)-glycinamide ribonucleotide transformylase (GART) [*purD*, *purM*, and *purN*]. The rate limiting enzyme in the pathway, glutamine phosphoribosylpyrophosphate amidotransferase, has recently been cloned using conserved amino acid sequences within the bacteria and yeast enzymes to generate degenerate oligonucleotides which were used to amplify DNA from the corresponding chicken cDNA.²

The enzymes, GARS, AIRS, and GART, catalyze the second, fifth, and third steps of the *de novo* purine biosynthetic pathway, respectively, as shown in Figure 1. The three enzyme activities have been copurified from chicken liver (9), strongly suggesting that they reside on a single polypeptide. Henikoff *et al.* has also shown that the three enzyme activities are encoded by a single gene in *Drosophila melanogaster* (10). Recently, Schild and coworkers have published a partial sequence of the human GARS-AIRS-GART cDNA which was cloned from a HepG2 cDNA library by functional complementation of yeast mutants (11). Patterson *et al.* have mapped the human gene to chromosome 21 using somatic cell genetics. Trisomy of

¹L.Ni, Z.Chen, J.E.Dixon, and H.Zalkin, unpublished data.

²G.Zhou, J.E.Dixon, and H.Zalkin (submitted for publication).

chromosome 21 is commonly observed in Down's Syndrome, and evidence has also been presented which shows that elevation of purine nucleotide intermediates in Down's Syndrome patients may be associated with the pathological consequences of this disease (12,13). It is important to note, however, that no causal relationship has yet been established between the *de novo* purine biosynthetic pathway and Down's Syndrome. In order to further explore the relationship between *de novo* purine nucleotide biosynthesis and Down's Syndrome, we have cloned and characterized the human cDNA encoding the trifunctional enzyme. The availability of these probes will afford a precise localization of the gene on chromosome 21 as well as providing new molecular tools to analyze the consequences of having extra copies of this gene in Down's Syndrome. Preliminary studies have been carried out by Northern analysis to compare levels of GARS-AIRS-GART and GARS RNA from a Down's Syndrome fibroblast cell line, Detroit 532, and a normal fibroblast cell line, HS27.

METHODS

Bacterial strains, growth conditions, and plasmids

E. coli strain, TX393 [*ara* Δ *lac* Δ (*purM-upp*) *sr1C300:Tn10:recA56*] was used to screen a chicken liver cDNA library. Bacteria were grown in minimal medium (1) supplemented with 0.1% acid hydrolyzed casein and 0.2 μ g/ml thiamine. Recombinant pBluescript plasmids from a Lambda Zap chicken liver cDNA library (Stratagene) were prepared by *in vivo* excision and recircularization as previously described (14). Plasmids which supported growth of strain TX393 on the above described medium were purified and subjected to a second round of transformation prior to further characterization.

Screening the cDNA libraries

The human GARS-AIRS-GART cDNA was isolated from a Lambda Zap HepG2 cDNA library (Stratagene) by screening approximately 500,000 recombinants. The 2.8 kb EcoRI fragment from the avian PM4 (nucleotides 75–2687) was nick-translated and used as a probe at 1×10^7 cpm/ml (specific activity $\sim 2 \times 10^8$ cpm/ μ g). Hybridization was carried out for 20 hours at 42°C in 0.8 M NaCl, 20 mM pipes, pH 6.5, 30% formamide, 0.5% SDS, and 100 μ g/ml salmon sperm DNA. The filters were washed twice for 30 min at 50°C in 0.15 M NaCl/0.015 M NaCitrate, 0.1% SDS, and were then subjected to autoradiography with an intensifying screen.

Additional chicken cDNA clones were isolated by screening the Lambda Zap chicken liver cDNA library as described above. A radiolabeled 586 bp EcoRI fragment (nucleotides 2687–3273) or a synthetic oligonucleotide corresponding to nucleotides 158–176 (labeled at the 5' end) was used as described above.

DNA sequencing

DNA sequence analysis was carried out by the dideoxynucleotide chain termination method using Sequenase (USB) (15). The M13 universal primer and a series of synthetic oligonucleotides were used to sequence both strands of the chicken and human GARS-AIRS-GART cDNA from overlapping clones.

RNA analysis

RNA was extracted from chicken liver, HepG2 cells, Hs27 cells, and Detroit 532 cells by the method of Chirgwin *et al.* (16). Poly(A)⁺ RNA was purified by chromatography on oligo(dT)

cellulose (17). 30 μ g of total RNA and 5 μ g of poly(A)⁺ RNA were fractionated by electrophoresis on a 1% agarose gel containing 2.2 M formaldehyde and were transferred to nylon membranes (18). The chicken mRNA was probed with either the nick-translated 2.15 kb cDNA insert of CD19 or the 2.1 kb insert of PM1. The human mRNAs were probed with a ³²P-labeled 2.8 kb EcoRI fragment of phDMN1. Hybridization and washing conditions were carried out as described by Thomas (19).

The 5' end of the chicken GARS-AIRS-GART mRNA was determined by primer extension mapping (20) with oligonucleotides that hybridized to nucleotides 158–176 and 235–253.

RESULTS AND DISCUSSION

Complementation of *E. coli* purine mutants with avian expression libraries

A plasmid library from chicken liver was prepared by *in vivo* excision of recombinant cDNAs which were cloned in the Lambda Zap vector system. This plasmid expression library was used to screen for cDNAs encoding enzymes in *de novo* purine nucleotide biosynthesis by functional complementation of *E. coli* strains deficient in their ability to catalyze specific steps in the pathway. Several cDNA clones were isolated from colonies grown on selective medium following transformation of the *purM* mutant strain TX393 with the chicken liver expression library. These plasmids were purified and were demonstrated to retransform TX393 at a high efficiency ($\sim 1.8 \times 10^4$ transformants/ μ g). Two of the cDNA clones, which were analyzed in more detail, were denoted as PM1 and PM4 and are schematically shown in Figure 2. Neither PM1 nor PM4 were able to complement *E. coli* strains with mutations in the *de novo* purine nucleotide biosynthetic pathway at step 2 (*purD*), step 6 (*purEK*), step 7 (*purC*), or step 8 (*purB*). This evidence suggested that PM1 and PM4 encoded a eukaryotic protein similar in function to that encoded by the bacterial gene *purM*.

Characterization of the avian cDNAs

Sequence analysis (described below) and a comparison of the deduced amino acid sequence of PM1 and PM4 with sequences from bacteria (*E. coli* and *B. subtilis*), *Drosophila melanogaster*, and *Saccharomyces cerevisiae* genes established that PM1 and PM4 encoded the AIRS activity. In addition, these plasmids contain an open reading frame encoding a protein sequence which showed sequence homologous to GART. The longer clone (PM4) also encoded a major portion of the GARS sequence. The absence of an initiator methionine or a 3' polyadenylation signal indicated that neither clone contained a complete 5' and 3' end.

To obtain sequences which correspond to the 3' end of the chicken GARS-AIRS-GART cDNA, we rescreened the Lambda Zap chicken liver cDNA library with a nick-translated 586 bp EcoRI fragment of PM1 (nucleotide 2687–3273). Initially, 11 positive clones were identified. Characterization of these clones by restriction enzyme analysis indicated that they were identical or were related to the cDNA clones obtained by functional complementation. One of these clones, denoted PME2, possessed additional sequences at the 3' end, including the polyadenylation signal as well as the poly(A) tail. Clones containing longer 5' DNA sequences were identified by rescreening the chicken cDNA library with a synthetic oligonucleotide corresponding to nucleotides 158–176. Several additional cDNA clones were identified and characterized by restriction enzyme analysis.

Sequence analysis of one of the clones, CD19, revealed a nucleotide sequence identical with the first 1348 nucleotides of PM4. This clone, which only encodes amino acid sequences corresponding to GARS, contained sequences encoding the putative initiator methionine and 5'-untranslated sequences. It is of particular interest to note that this clone also contains a poly(A) sequence and a polyadenylation signal at its 3' end, suggesting that it represents another form of the mRNA arising from the same avian gene which also gives rise to the corresponding RNA encoding the trifunctional polypeptide.

Since our cloning strategy either requires synthesis of active β -galactosidase fusion protein or re-initiation of protein synthesis from internal translation sites, we have further examined the expression of these cDNA clones by functional complementation. These results are shown in Table 1. Sequence analysis has revealed that the open reading frames of PM1 and PM4, which were selected by complementation, are not in frame with the initiator methionine of β -galactosidase. This indicates that an active AIR synthetase was synthesized from an internal translation start site utilizing another Shine Dalgarno-like sequence (21). A clone of PM4, denoted PME4, which is in the opposite orientation (3' to 5') with respect to the *lac* promoter is also capable of complementing the *purM* mutant TX393. This indicates that the *lac* promoter is not obligatory for functional complementation. Neither PM1, PM4, nor PME4 complement the *purD* mutant strain TX267, consistent with the fact that these plasmids do not encode the complete GARS sequence. The plasmid CD19 encodes the entire GARS sequence; however, this cDNA is oriented 3' to 5' with respect to the *lac* promoter and does not complement TX267. When the orientation with respect to the *lac* promoter is reversed, the clone, denoted CPURD, is able to complement TX267. Because CPURD is only 8 amino acids longer than PM4 and is not in frame with *lac Z*, this suggests that translation was initiated from GARS Met 1 and that a full length coding sequence for GARS is necessary for enzymatic activity.

Sequence analysis of the avian cDNA

The complete nucleotide sequence of the overlapping cDNA clones encoding the multifunctional protein has been determined from both strands of DNA and is shown in Figure 3. The schematic diagram shown in Figure 2 outlines the regions of the trifunctional cDNA contained within each of the subclones which were selected for sequencing. The cDNA is 3733 nucleotides in length and includes 50 nucleotides in the 5'-untranslated region, 675 nucleotides in the 3'-untranslated region, and 3009 nucleotides coding for the GARS-AIRS-GART enzymatic domains. The putative poly(A) addition signal, AATAAA, precedes the site of polyadenylation by 14 nucleotides. There is only one open reading frame which initiates from the first AUG codon located at position 50 and encodes 1003 amino acids. The nucleotide sequence surrounding the methionine conforms with the consensus eukaryotic translation initiation site (22). The calculated mass of the deduced polypeptide is 106,557 daltons. This value is in good agreement with the molecular weight reported for the purified chicken liver enzyme of approximately 117,000 estimated by SDS polyacrylamide gel electrophoresis (9). The organization of the multifunctional protein from the deduced amino acid sequence was shown to be NH₂-GARS-AIRS-GART-COOH.

The second cDNA (CD19) encodes a protein of 433 amino acids (45,995 calculated mass) and possesses only the GARS

Figure 1. *De novo* purine biosynthetic pathway. GARS-AIRS-GART catalyzes three independent steps (2, 5 and 3, respectively) in purine nucleotide biosynthesis.

Figure 2. Restriction map and isolated clones of the chicken liver GARS-AIRS-GART cDNA. The open boxed area indicates the coding region. The assignment of the enzymatic domains within the polypeptide is shown above and is based upon amino acid sequence identity noted with the bacterial enzymes. The * denotes the site of the stop codon encoded by CD19 following the GARS domain. A sequence of 760 nucleotides follows the * corresponding to the 3' untranslated RNA sequence.

Table 1. Complementation of *E. coli* mutant strains by isolated plasmids

Plasmid	Mutant Strains <i>purM</i> (TX393)	<i>purD</i> (TX267)	Orientation of cDNA with respect to the β -galactosidase promoter
PM1	+	—	5' to 3'
PM4	+	—	5' to 3'
PME4	+	—	reverse orientation (3'–5')*
CD19	—	—	reverse orientation (3'–5')*
CPURD	—	+	5' to 3'

*(i.e., no *lacZ* fusion)

structure. This 2108 nucleotide cDNA sequence is identical to all other clones between nucleotides 1–1348. However, the coding region of the protein terminates at position 1347, following an Arg residue and is succeeded by 760 nucleotides in the 3'-untranslated region. This cDNA harbors its own

Figure 3. Nucleotide sequence of the avian liver GARS-AIRS-GART cDNA and derived amino acid sequence of the polypeptide. The nucleotide sequence number appears above the amino acid sequence number which is shown in brackets.

domain in both RNA species. The smaller RNA species, however, does not hybridize to a probe (PM1 insert) containing only the AIRS-GART domain. The RNA analysis is consistent with two transcripts, one of which gives rise to GARS and the larger transcript corresponding to GARS-AIRS-GART.

A single band corresponding to an mRNA 20 nucleotides longer than the sequence shown in Figure 3 was detected by primer extension mapping (Figure 6B). This observation suggests

that a nearly complete sequence of the GARS-AIRS-GART cDNA was obtained. This result also supports the idea that mRNAs encoding GARS and GARS-AIRS-GART utilize the same transcription start site.

Northern blot analysis of HepG2 RNA also detected two strong RNA signals of 3.7 and 1.9 kb when the human GARS-AIRS-GART cDNA was used as a probe (Figure 6C). These results are

Figure 4. Schematic outline for the proposed transcription and translation of the avian and human GARS-AIRS-GART gene.

Table 2. Identification of domains in multifunction enzymes of *de novo* purine nucleotide synthesis

Organism	Amino Acids (numbers or positions)		
	GARS (<i>purD</i>)	AIRS (<i>purM</i>)	GART (<i>purN</i>)
<i>E. coli</i>	429	350	212
<i>B. subtilis</i>	422	346	195
<i>S. cerevisiae Ade5,7</i>	1-442	443-802	
<i>Ade 8</i>			214
<i>D. melanogaster</i>	1-434	(A)435-789 (B)790-1154	1155-1353
human	1-433	434-809	810-1010
chicken	1-433	434-805	806-1003

acggggcgaagcgggaacccaggtggccaccgggtgctggttttcttttcttttgggaattttctgtttacagcagacaca 78
 ATGCCAGCCGAGTACTTATATGTCAGTGGAGGAAGGAACATACGCTGGGAACTTCACAGTCTCATCTGCAACAAAGTGTGGTGGCCCGAGGAACCGAGCC 192
 MetAlaAlaArgValLeuIleIleGlySerGlyGlyArgGluHisThrLeuAlaTrpLysLeuAlaGlnSerHisHisValLysGlnValLeuValAlaProGlyAsnAlaGly 198
 ACTGCCGTCTGTAAGAAGATTCAATACCCCACTCTCAATCAGTGACCACTGCCCTTCTCAATTCGCAAGAGAGAAGAAATTGAATTTGTAGTTTGGACAGCAAGCA 306
 ThrAlaCysSerGluLysIleSerAsnThrAlaIleSerIleSerAspHisThrAlaLeuAlaGlnPheCysLysGluLysLysIleGluPheValValValGlyProGluAla 306
 CCTCTGGCTGCTGGGATTGTGGAACTTGAGGTCTGAGGAGTGCAATGCTTTGGCCCAACAGCAGAGAGCCGCTCACTAGAGTCCAGCAAGAGGTTTGGCAAGAGTTATG 420
 ProLeuAlaAlaGlyIleValGlyAsnLeuArgSerAlaGlyValGlnCysPheGlyProThrAlaAlaAlaGlnLeuGluSerSerLysArgPheAlaLysGluPheMet 420
 GACAGCATGGAAATCCCAACCCGCAATGGAAGCTTTCACCAAGCTGAAGAGCTGCAGCTTCATTTGAGTGCAGACTTCCCTGCTTTGGTGTGAAGGCCAGTGGCTTT 114
 AspArgHisGlyIleProThrAlaGlnTrpLysAlaPheThrLysProGluGluAlaCysSerPheIleLeuSerAlaAspPheProAlaLeuValLysAlaSerGlyLeu 114
 GCAGCTGGAAGAGGAGTATTGTGCAAGAGCAAGAGAGGCTGCAAGAGTGTCAAGAGAGTATCATCAGGAGAGAGCTTTGGGGCAGCTGGAGAGCAATTTGTCATTGAA 848
 AlaAlaGlyLysGlyValIleValAlaLysSerLysGluGluAlaCysLysAlaValGlnGlnIleMetGlnGluLysAlaPheGlyAlaAlaGlyValIleValGlu 848
 GAATCTTTGAGGAGAGAGGAGTGTGCTGTCTGTGTTTCACTAGTGCAGAGCTGTGGCCCACTGCCCCAGCAGAGAGCAATAGCCATTAAGTGCAGAGAGAGTGTGGCCCT 190
 GLeuLeuAspGlyGluValIleValAlaLysSerLysGluGluAlaCysLysAlaValGlnGlnIleMetGlnGluLysAlaPheGlyAlaAlaGlyValIleValGlu 190
 AACACAGGGGAAATGGAGGCTATTGTCCAGCCCTCAGGTTTCTAATGATCTATTACTAAAAATTAAGATAGTGTCTTCAGAGGACAGTGGATGCATGCAGAGAGG 762
 AsnThrGlyGlyMetGlyAlaLysCysProAlaProGlnValSerAsnAspLeuLeuLysIleLysAspThrValLeuGlnArgThrValAlaAspGlyMetGlnGlnGlu 762
 ACTCCATATACAGGTAATCTCTATGCTGGAATATGCTGACCAAGATGGCCAAAGTCTTAGAGATTAAATGGCTTTTGGTATCCAGAGTGCAGAGTAACTCCCACTT 266
 ThrProLysThrGlyIleLeuLysAlaGlyIleMetLeuThrLysAsnGlyProLysValLeuGluPheAsnCysArgPheGlyAspProGluCysGlnValIleLeuProLeu 266
 CTTAAGATGATCTTTATGAGTGATTCAGTCCACCTTAGATGGACTGCTGCACATCTCTGCTGTTTGGCTAGAAAACACACCGCCCTAACTGTGTGTGTCATGGCAAGTAAA 1104
 LeuLysSerAspLeuTrpGluValIleGlnSerThrLeuAspGlyLeuLeuCysThrSerLeuProValTrpLeuGluAsnHisThrAlaLeuValValValValValVal 1104
 GGTTATCTGGAGACTACACCAAGGTTGAGATTAACAGGTTTCTGAGGCTCAAGCTTAGGAGTGGAGGTTTCCATGCAGGCACTGCCCTCAAAAATGCAAGATAGTA 1218
 GlyTrpProGlyAspTrpThrLysGlyValGluIleThrGlyPheProGluAlaGlnAlaGlyLeuGluValPheHisAlaGlyThrAlaLeuLysAsnGlyLysValVal 1218
 ACTCATGGGGTAGAGTTCTTCAGTACAGCCATCCCGGAAATCTCATATACGCCCTTAGGAGAGCCAGAAAGGACTAGCTGCTATAAGTTTGAAGGAGCAATTTATAGG 1332
 ThrHisGlyGlyArgValLeuAlaValThrAlaAlaArgGluAsnLeuIleSerAlaLeuGluGlnAlaLysLysGlyLeuAlaAlaIleLysPheGluAlaLysIleGlyArg 1332
 AAAGAGCTGGCTTTCTGCTGCAATAGCTTTCTCCAGCAGCCAGGAGTTGACTTACAGGAATCTGGAGTAGATATCCAGCTGGAAATATGCTGGTCAAGAAATTCAGCCT 1446
 LysAspValGlyPheArgAlaIleAlaPheLeuGlnProArgSerLeuThrLysGlyValAlaAlaGlyLeuAlaAlaGlyLeuAlaAlaGlyLeuAlaAlaGlyLeuAla 1446
 TTAGCAAAAGCACTCCAGATCAGCTGTAAGTTGATCTGGAGGTTTTCCTGGCTTTTGAATTTAAAGAGCTGGTTTCAAGATCCCTTCTGGCTCTGGAGAGAT 1560
 LeuAlaLysAlaThrSerArgSerGlyCysLysValAspGlyGlyAlaGlyLeuPheAspLeuLysAlaAlaGlyPheLysAspProLeuAlaAlaSerGlyThrAsp 1560
 GCGCTGGAGACTAACTAAAGATGCCCAGCTATGCAATTAACATGATACCAATGGTCAAGATTTGGTAGCAATTTGGTAGCAATTTGGTAGCAATTTGGTAGCAATTT 1674
 GlyValGlyLeuLysLysIleAlaGlnLeuCysAsnLysHisAspThrIleGlyGlnLeuValAlaLeuMetCysValAsnAspIleLeuAlaGlnGlyAlaGluProLeu 1674
 TTCTCTCTGATTTCTTTCTCTGATTTCTCTGATTTCTCTGATTTCTCTGATTTCTCTGATTTCTCTGATTTCTCTGATTTCTCTGATTTCTCTGATTTCTCTGATTT 1788
 PhePheLeuAspTrpPheSerCysGlyLysLeuAspLeuSerValThrGluAlaValAlaValIleAlaLysAlaCysGlyLysAlaGlyLysAlaGlyLysAlaGlyLys 1788
 ACAAGAGAAATGCCGTGACATGTATCCCGCTGGAGAGTATGACCTAGCTGGGTTTGGCGTTGGTGGCTAGGAGAGATCAGAAACTCCCTCAGCTGGAGAGAACTCACTAGG 1902
 ThrAlaGluMetProAspMetTrpProProGlyGluTrpAspLeuAlaGlyPheAlaValGlyAlaMetGluArgAspGlnLysLeuProHisLeuGluArgIleThrGluGly 1902
 GATGTTGTTGTGGAAATAGCTTCTATGCTCTTATGCAATGGATTAGCTTGTGAGGAAATCGTTGCAAAATCTTCCCTCCAGTACTCTCTCCAGCAGCTGATGGTGT 2018
 AspValValValGlyIleAlaSerSerGlyLeuHisSerAsnGlyPheSerLeuLeuValAlaLysSerSerLeuGlnTrpSerSerProAlaProAspGlyCys 2018
 GGTGACAGACTTTAGGGGACTTACTTCTACGCCCTACCAGAACTACAGCCATTCACTGTTTACCTGCTTACGTTTACGTTTACGTTTACGTTTACGTTTACGTTT 2130
 GlyAspGlnThrLeuGlyAspLeuLeuLeuThrProThrArgIleLysSerHisSerLeuLeuProValLeuArgSerGlyHisValLysAlaPheAlaHisIleThrGlyGly 2130
 GGATTACTAGAGAAATCCCGAGAGTCCCTCCCTGAGAACTTGGGAGATTTAGATGCCAGAGCTGGAGGATCCCGAGGTTTCTCATGGTTGCAGAGCAAGAGCACTC 2244
 GlyLeuLeuGluAsnIleProArgValLeuProGluLysLeuGlyValAspLeuAspAlaGlnThrArgIleProArgValPheSerTrpLeuGlnGlnGluGlyHisLeu 2244
 TCTGAGGAGAGATGGCCAGAACATTTAACTGTGGGTTGGCGCTGTCTTGTGGTATCAAGAGAGCAGAGAGAGATCTGAGGGATATCCAGCAGCAAGGAGAGGCC 2358
 SerGluGluGluMetAlaArgThrPheAsnCysGlyValGlyAlaValLeuValValSerLysGluGlnThrGluGlnIleLeuArgAspIleGlnGlnIleLysGluGluAla 2358
 TGGGTGATGGCAGTGGTTTCCAGAGCTGAAGGTTCCCCACGTGTGAAGTCAAGAATCTGATTGAAGCATGCAAAATAAATGGGTCAAGTGTGAAGAAATGGCTCCCTGACA 2472
 TrpValIleGlySerValValAlaArgAlaGluGlySerProArgValLysValLysAsnLeuIleGluSerMetGlnIleAsnGlySerValLysLysAsnGlySerLeuThr 2472
 AATCATTTCTCTTTTGAAGAAAGGCGCAGAGTGGCTCTTAAATCTTGAAGAGGATCGAAGCTGCAAGCATTATAGACAGTACCGGAAACCAATAGCTGTCACAA 2586
 AsnHisPheSerPheGlnLysLysLysAlaArgValAlaValLeuIleSerGlyThrGlySerAsnLeuGlnAlaLeuIleAspSerThrArgGluProAsnSerSerAlaGln 2586
 ATTGATATTGTTTCTCCCAACAGCCAGTACTGGTTAGATTAAGCGAAGAGCTGGATTTCCCACTAGAGTAATTAATCATAACTGTATATAAAATCTGTGAGAAATT 2700
 IleAspIleValIleSerAsnLysAlaAlaValAlaGlyLeuAspLysAlaGluArgAlaGlyIleProThrArgValIleAsnHisLysLeuTrpAlaAlaGluAlaPhe 2700
 GACAGTCAATTGACCTAGCTTGAAGAGTCTCCATAGACATAGCTGTCTTCCAGGATTCATGAGAAATCTTCTTGGCCCTTTTCCCAAAAGTGAATGGAAAATGCTC 2814
 AspSerAlaIleAspLeuValLeuGluGluPheSerIleAspIleValCysLeuAlaGlyPheArgIleLeuSerGlyProPheValGlnLysTrpAsnGlyLysMetLeu 2814
 AATATCCACCCATCTTCTCTCTCTTTTAAAGGTTCAAAATGCCATGAGCAAGCCCTGGAAACCGGAGTCAAGTACTGGGTGCAGTGTACACTTTGTAGCTGAAGATGTG 2928
 AsnIleHisProSerLeuLeuProSerPheLysGlySerAsnAlaHisGluGlnAlaLeuIleThrGlyValThrValThrGlyCysThrValHisPheValAlaGluAspVal 2928
 GATGCTGGACAGATTTTTCGCAAGAGCTGTTCCCGTGAAGAGGGGTGATCTGTGCACTCTTCTGAAAGAGTAAATAGCAGAACATAAAATTTTCTCCAGCCCTT 3042
 AspAlaGlyGlnIleIleLeuGlnGluAlaValProValLysArgGlyAspThrValLysSerGluArgValLysLeuAlaGluHisLysIlePheProAlaAlaLeu 3042
 CAGCTGGTGGCAGTGGAGCTGTACAGCTTGGAGAAAATGCCAGATCTGTGGGTAAAGAGGAATgaagccttttaattcgaagattggagccttttgaagccttttga 3156
 GlnLeuValAlaSerGlyThrValGlnLeuGlyGluAsnGlyLysIleCysTrpValLysGluGlu 3156
 ctgtttgcatgtgtgtttttatctgagcttggcccaaaagaaaactgttaaaagacaaaagaccctcccttacttctatctatcttttttaaatagagactcaactaa 3270
 aaaaaaaaaaaaaaaaaa 3291

Figure 5. Nucleotide sequence and derived amino acid sequence of the human liver GARS-AIRS-GART cDNA.

consistent with, but do not prove that there is a corresponding human mRNA which encodes for the GARS mRNA.

The presence of a shorter RNA transcript encoding GARS activity has also been observed in *Drosophila* RNA analysis. Henikoff has postulated that the two messages arise by alternate splicing at the GARS-AIRS-GART gene locus to produce mRNA encoding the trifunctional protein or an RNA encoding GARS alone (23). The site of alternate processing of the RNA suggested by Henikoff generates a protein of 434 amino acids. The transcript encoding the avian GARS activity produces a protein of 433 amino acids suggesting that the site of alternate processing in the avian RNA is very similar, although not identical, to that in *Drosophila*.

Preliminary studies of the expression of the GARS-AIRS-GART gene were examined in a normal fibroblast cell line, Hs27, and a fibroblast cell line, Detroit 532, which was derived from a patient with Down's Syndrome. RNA analysis using the phDMN1 cDNA insert as a probe demonstrated a signal of 3.7 kb which is consistent in size with that observed for RNA isolated from human liver and HepG2 cell line (Figure 6D). The steady state RNA level, however, was similar in the two fibroblast cell lines. This is surprising because patients with Down's Syndrome harbor an extra copy of chromosome 21, and elevated expression of the GARS-AIRS-GART mRNA could be expected in the Down's Syndrome cell line. This result suggests that there may be additional factors which compensate for the presence of an extra copy of the gene. In contrast to human liver where both the 3.7 and 1.9 kb RNAs are evident, the fibroblast cell lines show a significantly reduced amount of the shorter RNA transcript encoding GARS. The reason for the different relative levels of the two RNAs in the cell lines is not understood.

Amino acid sequence homology

The chicken and human structural domains of each of the three enzymatic activities have been assigned by amino acid sequence alignment with sequences from the bacterial genes (shown in Figure 7). The *E. coli* and *B. subtilis* bacterial enzymes for GARS (Figure 7A), AIRS (Figure 7B) and GART (Figure 7C) are each encoded by a separate gene. Moreover, the high degree of sequence identity allows us to delineate the boundaries for each of the structural domains in the multifunctional enzymes. The structural organization of each of the enzymes is summarized in Table 2. Several aspects of the amino acid sequence homology are noteworthy. i) The bacterial and higher eucaryotic GARS and AIRS sequences are highly conserved. Using a method developed by Lipman and Pearson (24), the amino acid sequence of two proteins can be compared and the sequences evaluated statistically to determine the probability that such alignments occur by chance. The significance of the alignment is computed for any pair-wise comparison and a Z value calculated. Z values greater than 10 are considered to indicate homology between two sequences. The Z values, calculated from a pair-wise comparisons of GARS sequence, range from 64–251, while a similar comparison for AIRS sequences produce values of 47–132. Comparison of the GART domain indicates, from the smaller Z values, that it has evolved more rapidly than either the GARS or AIRS domain. ii) Both GART and AICAR transformylase utilize N¹⁰ formyl-tetrahydrofolate as a cofactor. Efforts to identify conserved amino acid sequences common to these two folate requiring enzymes using the FASTA alignment program (5, 24) were unsuccessful. In addition, other folate utilizing enzymes including thymidylate synthetase and the trifunctional

Figure 6. RNA analysis of chicken mRNA. (A) Northern blot analysis of chicken liver mRNA. 30 μ g of total (lanes a and c) and 5 μ g of poly(A)⁺ (lanes b and d) RNA were hybridized and washed according to conditions described in methods. Lanes a and b were probed with the 2.1 kb insert of CD19. Lanes c and d were hybridized with the 2.2 kb insert of PM1. The size markers from top to bottom denoted by the horizontal lines are 9.5, 7.5, 4.4, 2.4, and 1.4 kilobases. (B) Primer extension analysis was carried out with five μ g of chicken liver poly(A)⁺ RNA, oligonucleotides corresponding to 158–176 or 235–253 (shown), and AMV reverse transcriptase. Reactions were analyzed on an 8% sequencing gel alongside sequencing ladder obtained from CD19 with the corresponding oligonucleotide. (C) Northern blot analysis of the HepG2 RNA. Five μ g of poly(A)⁺ RNA was hybridized and washed as described. Size markers are as indicated in Figure 6. (D) Northern blot analysis of the Hs27 (lane A) and Detroit 532 (lane B) RNA. 3 μ g of poly(A)⁺ RNA was hybridized and washed as described. Size markers are as indicated.

NADP-dependent dehydrogenase-cyclohydrolase synthetase (25, 26) have shown no significant sequence identity suggesting that 'folate' binding domain cannot be identified within this 'family' of proteins. iii) The second copy of the AIRS domain which is present in *Drosophila* is missing from the higher eucaryotes. Table 2 shows that only a single copy of the AIRS protein (approximately 375 residues) is present in either human or chicken while two copies of 354 and 364 residues are evident in *Drosophila*. This result supports the idea that duplication of the AIRS domain in *Drosophila* occurred after the separation of flies from birds and mammals.

Figure 7. Amino acid alignments. Comparison of the deduced amino acid sequence for the A) GARS, B) AIRS, and C) GART domains of chicken (c), human (h), *D. melanogaster* (dm) [10], *S. cerevisiae* (sc) [30,31], *E. coli* (ec) [2,3,32], and *B. subtilis* (bs) [6], dmGARS-AIRS-AIRS-GART [10] [note that there are two tandem domains in drosophila, denoted dm(A) dm(B)]. Alignment was carried out using the FASTA program (24).

The bacterial genes encode separate proteins for GARS, AIRS, and GART while all the enzyme activities are present within a single polypeptide in eucaryotes. 'Connector sequences' are assumed to have allowed the individual bacterial genes to be joined to form the polypeptide. Examination of Figures 7A–C indicates that these amino acid sequences must be quite short. The connecting sequences between the GARS and AIRS would appear to be less than 10 residues, while the AIRS and GART domain are connected by a sequence of no more than 20 residues. Connector sequences have been suggested to afford conformational flexibility, allowing domains to retain their native structure (27). Connector sequences of approximately 70 residues have been noted in the A and B domains of tryptophan synthetase from *N. crassa* and *S. cerevisiae* (28). Connector sequences of 11 residues fuse *trpG* and *trpC* functions into a bifunctional TRP3 protein chain (27). The minimum size of connector fragments is unknown, however deletions of a connector sequence usually alter the physical and/or catalytic properties of an enzyme (29).

It has also been pointed out that connector sequences may represent remnants of nucleotide sequences involved in transposition or fusion of gene products (27). In *E. coli*, genes for *purD*, *purM*, and *purN* are found in different locations within the chromosome, while in *B. subtilis* a cluster of contiguous genes encodes enzymes for all 10 reactions for *de novo* synthesis of IMP (6). The sequences between the *B. subtilis* genes are often short and some genes have overlapping start-stop codons. Although these genes produce separate proteins, one can envision that simple nucleotide changes could have evolved to produce polypeptides which would be characterized by short connector sequences similar to those noted in the human and avian polypeptide.

ACKNOWLEDGEMENTS

This work was supported by United States Public Health Science Grants GM24658 (HZ) and NIDDKD 18024 (JED). Computer facilities were supported by NIH grant AI27713. This is journal

paper number 12511 from the Purdue University Agricultural Experimentation Station.

REFERENCES

1. Tso, J.Y., Zalkin, H., vanCleemput, M., Yanofsky, C. and Smith, J.M. (1982) *J. Biol. Chem.* **260**, 3525–3531.
2. Smith, J.M. and Daum, H.A. (1986) *J. Biol. Chem.* **261**, 10632–10636.
3. Smith, J.M. and Daum, H.A. (1987) *J. Biol. Chem.* **262**, 10565–10569.
4. Inglese, J., Johnson, D.L., Shiau, A., Smith, J.M. and Benkovic, S.J. (1990) *Biochemistry* **29**, 1436–1443.
5. Aiba, A. and Mizobuchi, K. (1989) *J. Biol. Chem.* **264**, 21239–21246.
6. Ebbole, D.J. and Zalkin, H. (1987) *J. Biol. Chem.* **262**, 8274–8287.
7. Aimi, J., Badylak, J., Williams, J., Chen, Z., Zalkin, H. and Dixon, J.E. (1990) *J. Biol. Chem.* **265**, 9011–9014.
8. Chen, Z., Dixon, J.E. and Zalkin, H. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 3097–3101.
9. Daubner, S.C., Schrimsher, J.L., Schendel, F.J., Young, M., Henikoff, S., Patterson, D., Stubbe, J. and Benkovic, S.J. (1985) *Biochemistry* **24**, 7059–7062.
10. Henikoff, S., Keene, M.A., Sloan, J.S., Bleskan, J., Hards, R. and Patterson, D. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 720–724.
11. Schild, D., Brake, A.J., Kiefer, M.C., Young, D. and Barr, P.J. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 2916–2920.
12. Patterson, D., Graw, S. and Jones, C. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 405–409.
13. Fuller, R.W., Luce, M.W. and Merz, E.P. (1962) *Science* **137**, 868–869.
14. Short, J.M., Fernandez, J.M., Sorge, J.A. and Huse, W.D. (1988) *Nucl. Acids Res.* **16**, 7583–7600.
15. Sanger, F.S., Nicklen, S. and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
16. Chirgwin, J.M., Przybyla, A.E., MacDonald, R.J. and Rutter, W.J. (1979) *Biochemistry* **18**, 5294–5299.
17. Aviv, H. and Leder, P. (1972) *Proc. Natl. Acad. Sci. USA* **69**, 1408–1412.
18. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*, p. 202, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
19. Thomas, P.S. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 5201–5205.
20. Minth, C.D., Andrews, P.C. and Dixon, J.E. (1986) *J. Biol. Chem.* **261**, 11974–11979.
21. Shine, J. and Dalgarno, L. (1974) *Proc. Natl. Acad. Sci. USA* **71**, 1342–1346.
22. Kozak, M. (1986) *Cell* **44**, 283–292.
23. Henikoff, S., Sloan, J., and Kelley, J. (1983) *Cell* **34**, 405–414.
24. Lipman, D.J. and Pearson, W.R. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.
25. Bélanger, C. and MacKenzie, R.E. (1989) *J. Biol. Chem.* **264**, 4837–4843.
26. Ayusawa, D., Shimizu, K., Gotoh, O. and Seno, T. (1985) *Nucl. Acids Res.* **13**, 2035–2043.
27. Zalkin, H., Paluh, J., vanCleemput, M., Meye, S. and Yanofsky, C. (1984) *J. Biol. Chem.* **259**, 3985–3992.
28. Burns, D.M., Horn, V., Paluh, J. and Yanofsky, C. (1990) *J. Biol. Chem.* **265**, 2060–2069.
29. Crawford, I., Clarke, M., vanCleemput, M. and Yanofsky, C. (1987) *J. Biol. Chem.* **262**, 239–244.
30. Henikoff, S. (1986) *J. Mol. Biol.* **190**, 519–528.
31. White, J.H., Lusnak, K. and Fogel, S. (1985) *Nature* **315**, 350–352.
32. Shen, Y., Rudolph, J., Stern, M. and Stubbe, J. (1990) *Biochem.* **29**, 218–227.