The cloning and sequence analysis of the *aspC* and *tyrB* genes from *Escherichia coli* K12

Comparison of the primary structures of the aspartate aminotransferase and aromatic aminotransferase of *E. coli* with those of the pig aspartate aminotransferase isoenzymes

Ian G. FOTHERINGHAM, Sandra A. DACEY, Paul P. TAYLOR, Terry J. SMITH,* Michael G. HUNTER, Mark E. FINLAY, Sandy B. PRIMROSE, David M. PARKER† and R. Mark EDWARDS‡

Department of Applied Microbiology, G. D. Searle Research and Development, P.O. Box 53, Lane End Road, High Wycombe, Bucks. HP12 4HL, U.K.

In this paper we describe the cloning and sequence analysis of the *tyrB* and *aspC* genes from *Escherichia coli* K12, which encode the aromatic aminotransferase and aspartate aminotransferase respectively. The *tyrB* gene was isolated from a cosmid carrying the nearby *dnaB* gene, identified by its ability to complement a *dnaB* lesion. Deletion and linker insertion analysis located the *tyrB* gene to a 1.7-kilobase *NruI*–*HindIII*-digest fragment. Sequence analysis revealed a gene encoding a 43 000 Da polypeptide. The gene starts with a GTG codon and is closely followed by a structure resembling a ρ independent terminator. The *aspC* gene was cloned by screening gene banks, prepared from a prototrophic *E. coli* K12 strain, for plasmids able to complement the *aspC tyrB* lesions in the aminotransferase-deficient strain HW225. Sub-cloning and deletion analysis located the *aspC* gene on a 1.8-kilobase *HincII*–*StuI*-digest fragment. Sequence analysis revealed the presence of a gene encoding a 43 000 Da protein, the sequence of which is identical with that previously obtained for the aspartate aminotransferase from *E. coli* B. Considerable overproduction of the two enzymes was demonstrated. We compared the deduced protein sequences with those of the pig mitochondrial and cytoplasmic aspartate aminotransferases. From the extensive homology observed we are able to propose that the two *E. coli* enzymes possess subunit structures, subunit interactions and coenzyme-binding and substrate-binding sites that are very similar both to each other and to those of the mammalian enzymes and therefore must also have very similar catalytic mechanisms. Comparison of the *aspC* and *tyrB* gene sequences reveals that they appear to have diverged as much as is possible within the constraints of functionality and codon usage.

INTRODUCTION

Transamination is involved, directly or indirectly, in the biosynthesis of most of the amino acids. In *Escherichia coli*, transamination is the final step in the synthesis of aspartate, phenylalanine, tyrosine, isoleucine, valine, leucine and alanine. In addition, aminotransferases are involved in intermediate stages of lysine, arginine, histidine and serine biosynthesis. Aminotransferases almost without exception require pyridoxal 5'-phosphate as a cofactor, which is covalently bound to the enzyme by the formation of a Schiff base with the ϵ -amino group of a lysine residue. There is only one isolated example of an aminotransferase that does not appear to require pyridoxal 5'-phosphate (Jack & McMahon, 1978). In the case of the mammalian aspartate aminotransferases, the mechanism is relatively well understood and is thought to proceed via a double-displacement process in which the enzyme shuttles between pyridoxal and pyridoxamine forms (Kirsch *et al.*, 1983). Transamination reactions are freely reversible and they play a role in the catabolism of

certain amino acids. In the direction of amino acid biosynthesis the usual amino-group donor is thought to be glutamate.

In *E. coli* there are four aminotransferases that catalyse terminal steps in amino acid biosynthesis. These enzymes have now been characterized both genetically and biochemically.

The so-called aromatic aminotransferase, the product of the *tyrB* gene, is active in the synthesis of phenylalanine, tyrosine, aspartate and leucine. Its identification as the aromatic aminotransferase is based on the kinetic properties of the purified enzyme (Powell & Morrison, 1978), on its apparent regulation by the tyrosine repressor (Silbert *et al.*, 1963; Wallace & Pittard, 1969; Collier & Kohlaw, 1972) and by the study of aminotransferase-deficient mutants (Gelfand & Steinberg, 1977; Powell & Morrison, 1978).

The *aspC*-gene product, the aspartate aminotransferase, has been shown to be active in the synthesis *in vivo* of aspartate, tyrosine and phenylalanine (Gelfand & Steinberg, 1977). Studies on the purified enzyme have

Abbreviations used: kb, kilobases; bp, base-pairs.

* Present address: Nuffield Research Institute, John Radcliffe Hospital, Headly Way, Headington, Oxford OK3 9DU, U.K.

† Present address: Imperial Biotechnology Ltd., 263–269 City Road, London EC1V 1JX, U.K.

‡ To whom correspondence and requests for reprints should be addressed.

Table 1. Summary of the *E. coli* aminotransferase activities

Gene	Enzyme	Product(s)	Reference
<i>aspC</i>	Aspartate aminotransferase	L-Aspartate, L-phenylalanine, L-tyrosine	Gelfand & Steinberg (1977)
<i>tyrB</i>	Aromatic aminotransferase	L-Phenylalanine, L-tyrosine, L-leucine, L-aspartate	Gelfand & Steinberg (1977)
<i>ilvE</i>	Branched-chain aminotransferase	L-Isoleucine, L-valine, L-leucine, L-phenylalanine	Lawther <i>et al.</i> (1979)
<i>avtA</i>	Alanine/valine aminotransferase	L-Alanine, L-valine, α -aminobutyrate	Falkingham (1979)
<i>serC</i>	Phosphoserine aminotransferase	L-Phosphoserine	Umbarger <i>et al.</i> (1963)
<i>hisC</i>	Histidinol phosphate aminotransferase	L-Histidinol phosphate	Goldschmidt & Cater (1970)
<i>dapD</i>	N-Succinyl-L-diaminopimelate aminotransferase	N-Succinyl-L-2,6-diaminopimelate	Peterkofsky & Gilvarg (1961)
<i>argD</i>	Acetylornithine aminotransferase	N-Acetylornithine	Vogel <i>et al.</i> (1963)

Table 2. Bacterial strains

Strain	Genotype	Source	Reference
W3110	F—	Prototroph	M. Yudkin
E107	F—	<i>thr-1 leuB6 thi-1 dnaB107</i> <i>deoC1 lacY1 tonA21 rpsL67</i>	B. Bachmann
ES430	HfrH	<i>thi-1 malB29 relA1 spoT1</i>	B. Bachmann
AB1886	F—	<i>thr-1 leuB6 thi-1 argE3 his-4</i> <i>proA2 uvrA6 lacY1 galK2 mtl-1</i> <i>xyl-5 ara-14 rpsL31 tsx-33</i> <i>supE44</i>	B. Bachmann
DG44	F—	<i>thr-1 leu-6 thi-1 proA2 argE3</i> <i>hisG4 lacY1 galK2 ara-14</i> <i>xyl-5 mtl-1 rpsL31 tsx-33</i> <i>supE44 recB21 recC22 sbcB15</i> <i>hsdS hppT29 tyrB507 aspC13</i>	M. Doel
MC1061	F—	<i>araD139 Δ(ara-leu)7697</i> <i>Δ(lacIPOZY)74 galU</i> <i>galK hsdR rpsL</i>	W. Tacon
HW225		As MC1061 but <i>aspC13 tyrB507</i> <i>recA56 srLa : : Tn10</i>	Present work
HW857	F—	<i>Δ(aspC) KanR tyr B/Tn10</i>	Present work

confirmed its broad substrate specificity (Powell & Morrison, 1978), which is analogous in some respects to the behaviour of some of the mammalian aspartate aminotransferases (Mavrides & Christen, 1978). Physical studies on the aspartate aminotransferase and aromatic aminotransferase have revealed a similar organization of two identical 40000 Da subunits. CNBr fragmentation and amino acid analysis failed to reveal any similarity (Powell & Morrison, 1978).

The branched-chain aminotransferase, the product of the *ilvE* gene, is predominantly involved with the biosynthesis of leucine, isoleucine and valine, although it also has the ability to synthesize phenylalanine *in vivo* (Gelfand & Steinberg, 1977; Kline *et al.*, 1977; Lee-Peng *et al.*, 1979).

The *avtA* gene encodes the alanine/valine aminotransferase. This enzyme is competent in valine biosynthesis but not isoleucine biosynthesis. It is one of the enzymes involved in alanine biosynthesis, though mutations in *avtA* do not confer a requirement for alanine (Whalen & Berg, 1982).

Of the aminotransferases that are involved in the intermediate steps of amino acid biosynthesis only the phosphoserine aminotransferase, the *serC*-gene product, has been extensively characterized (Dempsey, 1969;

Clarke *et al.*, 1973). The products of the *E. coli* aminotransferases *in vivo* are listed in Table 1.

As a first step towards dissecting the molecular biology of the bacterial aminotransferases, we here describe below the cloning and analysis of the *tyrB* and *aspC* genes from *E. coli* K12. This is the first report of the sequence of any aromatic aminotransferase.

MATERIALS AND METHODS

Media, bacterial strains and plasmids

The minimal medium used was that of Vogel & Bonner (1956), supplemented with glucose at 0.2%. Where appropriate, this was supplemented with 0.4% casamino acids (casein hydrolysate) or individual amino acids at 50 μ g/ml. Other media were as described in Miller (1972). Bacterial strains used are listed in Table 2. Plasmids used are shown in Table 3.

Cosmid cloning

Cosmid gene banks were constructed by ligating *E. coli* K12 DNA, isolated from a prototrophic strain (W3110) and partially digested with restriction endonuclease *Sau3AI*, to suitably prepared plasmid pH79 DNA followed by packaging *in vitro* and transduction into the

Table 3. Plasmids used and constructed

Plasmid	Description	Source
pAT153		Twigg & Sherratt (1980)
pHC79		Hohn & Collins (1980)
pME37	<i>Cla</i> I fragment carrying <i>tyrB</i> in pAT153	Present work
pME56	<i>Eco</i> RI deletion of pME37	Present work
pME57	<i>Bam</i> HI deletion of pME37	Present work
pME58	<i>Bam</i> HI deletion of pME56	Present work
pME64	<i>Nru</i> I deletion of pME58	Present work
pME133	<i>Bgl</i> II clone of <i>aspC</i>	Present work
pME101	<i>Hind</i> III- <i>Sal</i> I sub-clone of <i>aspC</i> from pME137	Present work
pIF100	<i>Hinc</i> II- <i>Stu</i> I sub-clone of <i>aspC</i> (carries <i>Bgl</i> II linker at <i>Stu</i> I site)	Present work

dnaB strain E107. The methodology used was essentially as described by Ish-Horowicz & Burke (1981).

Assay of aspartate aminotransferase activity

A commercially available assay kit, purchased from Sigma Chemical Co., was used to determine aminotransferase activity. The assay measures the production of oxaloacetate from aspartate with 2-oxoglutarate as the amino-group acceptor. The formation of oxaloacetate is followed by coupling to NADH oxidation with malate dehydrogenase and measuring the rate of change in absorbance at 340 nm. The rates in crude extracts were corrected for endogenous NADH oxidase activity by measuring the rate of NADH oxidation in the absence of aspartate, 2-oxoglutarate and malate dehydrogenase. Specific activities are expressed as μmol of NADH oxidized/min per mg of protein at 37 °C. Protein concentrations were determined by the method of Bradford (1976), with human serum albumin as standard.

Non-denaturing polyacrylamide-gel electrophoresis

Native gels were run and stained for phenylalanine aminotransferase activity as described previously (Gelfand & Steinberg, 1977).

Construction of the aminotransferase-deficient strain HW225

Detailed analysis of the *tyrB* and *aspC* clones required an *aspC tyrB* mutant that could maintain ColEI-derived vectors such as plasmid pAT153. The *aspC* and *tyrB* lesions were therefore moved from strain DG44 into the readily transformable strain MC1061. This was achieved by isolating *Tn10* insertions linked to *tyrB* and *aspC* and transducing the mutant alleles of these two genes from strain DG44 into strain MC1061 by using bacteriophage PlcI. Initially, an *aspC* Tet^R transductant was identified by screening for lack of the characteristic *aspC* band obtained after staining native polyacrylamide gels for phenylalanine aminotransferase activity. The *aspC* derivative of strain MC1061 was cured of *Tn10* (Bochner *et al.*, 1980) and used as a recipient for transducing in the *tyrB* lesion, again by employing linkage with *Tn10*. The desired *TyrB aspC* recombinants could be easily identified by their requirement for aspartate and tyrosine. This double mutant was again cured of *Tn10* and transduced to *recA* by using bacteriophage P1 grown on the

recA srLA::Tn10 strain HW19 to give strain HW225. Standard *Tn10* methodology was employed throughout this construction (Kleckner *et al.*, 1977). Also used in this work was a strain W3110 derivative carrying the *tyrB* lesion from strain DG44 and a deletion of the entire *aspC* gene (strain HW 857). The *aspC* deletion was constructed *in vitro* and introduced into strain W3110 *tyrB* by forced recombination.

Sequence analysis

Both the chemical method of Maxam & Gilbert (1977) and the chain-termination method of Sanger (Sanger *et al.*, 1977; Biggin *et al.*, 1983) were employed. Extensive use was made of the Messing series of M13 vectors (Messing & Vieira, 1982). In a number of instances the sequence information obtained from a single M13 clone was extended by the use of suitable oligonucleotides synthesized at G. D. Searle and Co. by automated phosphoramidite chemistry (Urdea *et al.*, 1983).

RESULTS AND DISCUSSION

Cloning of the *tyrB* gene

The *tyrB* gene was mapped to 92 min on the *E. coli* genetic map (Gelfand & Steinberg, 1977; Gelfand & Rudo, 1977). As we initially lacked a suitable background for the direct selection of *tyrB* clones we elected to clone the nearby gene *dnaB* (Wechsler & Gross, 1971) on a cosmid. Since cosmid clones carry approx. 1 min of genetic material, i.e. 40 kb, we reasoned that there was a good chance that such clones would also carry *tyrB* as a non-selected marker.

A cosmid bank prepared from a partial restriction endonuclease *Sau*3AI digestion of *E. coli* W3110 DNA ligated into plasmid pHC79 (Hohn & Collins, 1980) was constructed in the strain E107. This strain is temperature-sensitive for growth owing to the presence of a conditional lesion in the *dnaB* gene. Of 1273 cosmid clones isolated in E107, eight were identified as carrying *dnaB* by their ability to grow at the restrictive temperature (42 °C). Plasmid DNA was prepared from these eight isolates and the clones were further characterized.

Restriction analysis indicated that the clones all carried large inserts and that they were indeed related (results not shown). Re-transformation of the *dnaB* strain established that the growth at 42 °C was due to a plasmid-encoded gene and not due to reversion of the original *dnaB* allele. The eight clones were analysed for the presence of genes known to map close to *tyrB* (Bachmann, 1983).

malB, a locus concerned with maltose utilization, maps at 91.5 min on the genetic map, approx. 0.5 min from *tyrB*. The eight *dnaB* clones were introduced into the *malB* strain ES430. None of the transformants acquired the ability to ferment maltose, and we concluded that they did not extend as far as *malB*.

Another nearby gene, *uvrA*, is a locus concerned with the repair of u.v. damage to DNA (Howard-Flanders *et al.*, 1966). Inactivation of this gene results in increased sensitivity to u.v. The presence of the *uvrA* gene on the *dnaB* clones was tested by introducing them into the *uvrA* strain AB1886 and evaluating the transformants for their sensitivity to u.v. Considerable u.v.-resistance was conferred by all but two of the clones. Since the *uvrA* gene is located on the far side of *tyrB* from *dnaB*, it seemed

likely that the *uvrA*-carrying *dnaB* clones must also carry *tyrB*.

That the clones did indeed carry *tyrB* was shown as follows. The aminotransferase-deficient strain DG44 carries point mutations in the *tyrB* and *aspC* genes and as a result requires exogenous aspartate and tyrosine for growth. The presence of the *recB*, *recC* and *sbcB* mutations in this strain prevents the stable inheritance of ColE1-derived plasmids (Vapnek *et al.*, 1976) such as pHC79 and hence the possibility of complementation studies. Strain DG44 is still recombination-proficient, and so it should be possible to isolate *tyrB*⁺ recombinants after transformation with the *dnaB* clones if they indeed carry the adjacent *tyrB* gene.

To enhance the frequency of introduction of the clones, the cosmid DNA was first subject to packaging *in vivo* into bacteriophage particles and then transduced into strain DG44. The cells were allowed to grow for 1 h in L broth to permit recombination and phenotypic expression, washed and plated on to minimal medium lacking tyrosine and aspartate to select for *tyrB*⁺ recombinants. Such recombinants were observed at a frequency of about 0.25% of the expected transduction rate with four out of six of the clones that had been shown to carry *uvrA*. Since no such *tyrB*⁺ colonies were seen in the control transductions with packaged plasmid pHC79 DNA, we concluded that in all probability these *dnaB* clones carry an intact *tyrB* gene.

Sub-cloning of *tyrB*

Restriction analysis of one of the *dnaB*⁺ *tyrB*⁺ *uvrA*⁺ clones revealed an insert of about 40 kb. This isolate was digested with a number of different restriction enzymes with hexanucleotide recognition sites, and all fragments larger than 2 kb were cloned into appropriately cleaved plasmid pAT153 (Twigg & Sherratt, 1980). The isolated sub-clones were tested for the presence of *tyrB* by

transformation into strain DG44 and screening for *tyrB*⁺ recombinants, as described above. Three different sub-clones were positive in this test, the smallest of which carried a 4.5 kb *Cla*I-digest fragment. Native polyacrylamide gels of extracts prepared from cells carrying this *Cla*I sub-clone revealed considerable overproduction of the aromatic aminotransferase (results not shown). The plasmid, carrying the *Cla*I sub-clone pME37, was used for further studies.

Analysis of the location of the *tyrB* gene was greatly facilitated by the construction of the *tyrB* *aspC* *recA* strain HW225 (see the Materials and methods section). Derivatives of plasmid pAT153 can be maintained stably in this genetic background, making it possible to test for the presence of an active *tyrB* gene by complementation.

The position of the *tyrB* gene was narrowed down by deletion analysis and by the introduction of *Bgl*II linkers at the *Rsa*I sites within the insert. The results of this analysis are given in Fig. 1. The gene was finally located on a 1622 bp *Eco*RI-*Nru*I-digest fragment, more than sufficient to encode a 40000 Da protein.

Determination of the *tyrB* DNA sequence

The nucleotide sequence of the 1729 bp *Eco*RI-*Bam*HI-digest fragment was determined by the chemical sequencing method. Use was made of the unique *Bgl*II sites resulting from the linker mutagenesis. All restriction sites used for end-labelling were sequenced through. The *Eco*RI-*Nru*I site region was sequenced on both strands. The sequencing strategy is depicted in Fig. 1. The final sequence of the *Eco*RI-*Nru*I-digest fragment is shown in Fig. 2.

The *tyrB* gene

Statistical analysis of this region (Staden, 1984c) indicated a bias in positional base frequencies and codon usage, indicating that it is indeed a gene. The sequence

Fig. 1. Deletion, insertion and sequence analysis of the *tyrB* gene

Positions at which 8bp *Bgl*II linkers were inserted are indicated by open arrows, and the direction of transcription is shown by a solid arrow. The strategy used in Maxam & Gilbert sequencing of the gene is illustrated by the arrows at the bottom of the Figure. The direction of the arrows indicates whether the fragment was sequenced from its 3' or 5' end. The position of the arrow, above or below the central line, indicates the strand from which the fragment was derived.

```

TCGGATGAAATACGTGGATTAAATGTTCTGTAATTTTGATTGCTGTGCGGATGCGCGGTGAATGCCTTATCGCGCCAATAAAATCTAAAAATCAATAAGTTGATGTTCTTTTCAT
10      20      30      40      50      60      70      80      90      100     110     120

GCTCTTATAAAGGTGCTGCTCTGGGCGATGTACGTTTGTTCATGAGTCTCACTCTGTGTCTAAATGCGGTTCGCTCTGAACATCCACTCGATCTTTCGCTTCTTCGCGTTTATTGTGTT
130     140     150     160     170     180     190     200     210     220     230     240

      rbs
      ****
TTAACCACCTGCCCGTAAACCTGGAGAACCATCGCGTGTTCACAAAAGTTGACGCCTACGCTGGCGACCCGATTCTTACGCTTATGGAGCGTTTAAAGAAGACCCCTCCGAGGACAAAG
250     260     270     280     290     300     310     320     330     340     350     360

      N L S I G L Y Y N E D G I I P Q L Q A V A E A E A R L N A Q P H G A S L Y L P M
TGAATTTAAATATCGGCTCTGTTACTACACGAAGACCGAATTTATTCACAACCTGCAAGCGTGGCGGAGGCGGAAGCGCGCTGAATGCGCACCTCATGCGCGCTTCGCTTTATTATACCGA
370     380     390     400     410     420     430     440     450     460     470     480

      E G L N C Y R H A I A P L L F G A D H P V L K Q Q R V A T I Q T L G G S G A L K
TGGAGGGCTTAACCTGCTATCGCCATGCCATTGCGCGCTGTGTTGGTGGGACCATCCGCTACTGAAACAACAGCGGTAGCAACCATTCAAACCCCTGGCGGCTCCGGGCGATTGA
490     500     510     520     530     540     550     560     570     580     590     600

      V G A D F L K R Y F P E S G V W V S D P T W E N H V A I F A G A G F E V S T Y P
AAGTGGCGCGGATTTCCTGAAACGCTACTTCCCGGAATCAGGCGTCTGGTTCAGCGATCCCTACCTGGGAAAACAGTAGCAATATTGCGCGGGCTGGATTGCAAGTGAAGTACTTACC
610     620     630     640     650     660     670     680     690     700     710     720

      W Y D E A T N G V R F N D L L A T L K T L P A R S I V L L H P C C H N P T G A D
CCTGGTATGACGAAGCGACTTAACGCGTGGCGCTTAAATGACCTGTGCGGACGCTGAAACATTACCTGCGCGCAGTATGTGTGTGCTGCATCCATGTGTCCACACCCACCGGTGCGG
730     740     750     760     770     780     790     800     810     820     830     840

      L T N D Q W D A V I E I L K A R E L I P F L D I A Y Q G F G A G M E E D A Y A I
ATCTCAGTAATGATCAGTGGGATGCGGTGATTGAAATCTCAAAAGCCGCGAGCTTATTCATCTCGATATTGCCTATCAAGGATTGTTGGTCCGCGTATGGAAGAGGATGCCCTACGCTA
850     860     870     880     890     900     910     920     930     940     950     960

      R A I A S A G L P A L V S N S F S K I F S L Y G E R V G G L S V M C E D A E A A
TTCGCGCATTTGCCAGCGCTGGATTACCGGCTCTGGTGAGCAATTCGTTCTCGAAATTTTCTCCCTTACGCGGAGCGCGTCCGCGGACCTTCTGTATGTGTGAAGATGCCGAAGCCG
970     980     990     1000    1010    1020    1030    1040    1050    1060    1070    1080

      G R V L G Q L K A T V R R N Y S S P P N F G A Q V V A A V L N D E A L K A S W L
CTGCGCGCTACTGGGCAATGAAAGCAACAGTTGCGCCCACTACTCCAGCCGCGAATTTTGTGCGCAGGTGTTGGCTGCAGTCTGAATGACGAGGCAATTGAAAGCCAGCTGGC
1090    1100    1110    1120    1130    1140    1150    1160    1170    1180    1190    1200

      A E V E E M R T R I L A M R Q E L V K V L S T E M P E R N F D Y L L N Q R G M F
TGGCGGAAGTAGAAGATGCGTACTCGCATTTCTGGCAATCGCTCAGGAATTTGGTGAAGGTATTGAAGCACAGATGCCAGAACGCAATTTGATTTATCTGCTTAATCAGCGCGCATGT
1210    1220    1230    1240    1250    1260    1270    1280    1290    1300    1310    1320

      S Y T G L S A A Q V D R L R E E F G V Y L I A S G R M C V A G L N T A N V Q R V
TCAGTTATACCGGTTTAAATGCGGCTCAGGTGTACCGACTACGTTGAAGAAATTTGGTGTCTATCTCATTCGCGCAGCGTCCGATGTGTGTGCGCGGTAAATACCGCAATGTACACGCTG
1330    1340    1350    1360    1370    1380    1390    1400    1410    1420    1430    1440

      A K A F A A V M
TGGCAAGCGGTTTGTCTGCGGTGATGTAATGCAGAAAGCAGGCTCGAGCTACCCAGCTCGAGTGAATTAACGTGTCGTGCTTTCACTCTTTCTTTATAGATGATTTTTTTGATGCC
1450    1460    1470    1480    1490    1500    1510    1520    1530    1540    1550    1560

ATCGTTCTACGTGAGAGATAATAAACGTTGTAGTTCTTTTATTGTTAAGCTTTATCCCAATTATCTG
1570    1580    1590    1600    1610    1620

```

Fig. 2. DNA sequence of the *EcoRI*–*NruI*-digest fragment encompassing the *tyrB* gene

The coding sequence of the *tyrB* gene initiates at the GTG codon at position 276 in the sequence, which is preceded by a good-consensus Shine & Dalgarno sequence as indicated by asterisks (rbs, ribosome-binding site). The 1191 bp sequence encoding the 397 amino acid residues of the aromatic aminotransferase is closely followed by a region of perfect dyad symmetry (indicated by overbars), strongly characteristic of structures implicated in termination of transcription in *E. coli*.

reveals one long open reading frame from residue 201 to residue 1466. Neither of the potential methionine (start) codons in this frame at positions 324 and 480 is preceded by a typical consensus ribosome-binding site (Shine & Dalgarno, 1975). The ATG codons at positions downstream of position 480 would result in a polypeptide of less than 36000 Da and are therefore unlikely to be the authentic translation start. In addition to these potential ATG start codons there is a GTG codon at position 276 that is preceded by a Shine & Dalgarno sequence

(GGAG). We favour this codon as the translation start point over the ATG codon at position 324, since, in addition to the presence of the consensus Shine & Dalgarno box, it gives rise to an *N*-terminal peptide sequence closely homologous to that of both the aspartate aminotransferase of *E. coli* and to the mammalian enzymes. Furthermore this region has been recognized as important for the stabilization of the dimeric enzyme structure (see below).

Assuming the GTG start to be correct on functional

Fig. 3. Structure of the putative terminator downstream of the *tyrB* gene

This region of dyad symmetry, closely following the *tyrB* coding sequence, is capable of forming a perfect stem-loop structure comprising a stem of length nine base-pairs, of which seven are G·C, and a loop of six base-pairs. Of the next 12 bases, ten are either A or T. This structure has a free energy of formation of -105.9 kJ/mol (-25.3 kcal/mol) as determined by the method of Tinoco *et al.* (1973).

grounds, the aromatic aminotransferase consists of 397 amino acid residues and has a calculated molecular mass of 43 505 Da (including the *N*-terminal methionine). This is in agreement with the value derived from polyacrylamide-gel analysis of transcription/translation products *in vitro* from plasmid pME64 (results not shown).

Transcriptional signals for the *tyrB* gene

Analysis of the region upstream of the start of the gene reveals a number of potential promoter sequences with the most likely transcription starts at residue 193 or residue 244. Previous reports have suggested that the expression of the *tyrB* gene may be repressed by tyrosine (Silbert *et al.*, 1963). If this were the case then the *tyrR*-gene

product (Wallace & Pittard, 1969) would be a prime candidate. However, a comparison of the upstream sequence with that of the *aroF* and *aroG* genes (Davies & Davidson, 1982; Hudson & Davidson, 1984) reveals no striking similarities. The strongest homology between *tyrB* and *aroF* occurs at residues 148–166, spanning one of the putative promoter -35 regions. This sequence in *aroF* has been suggested as a potential *tyrR* binding site on the basis of a comparison of *aroF* and *aroG* promoter sequences. Precise identification of the *tyrR* binding site must await the isolation and sequence analysis of *tyrB* regulatory mutants.

The *tyrB* translation stop is a single TAA codon at position 1467. The gene is followed by a region of perfect dyad symmetry commencing at position 1479. The dyad symmetry is followed by a run of eight A·T base-pairs. It is likely that this feature represents the *tyrB* transcriptional terminator (see Fig. 3). The putative terminator is followed by an A+T-rich region and another potential promoter.

Cloning of the *aspC* gene

The *aspC* gene was cloned directly by screening a number of gene banks of *E. coli* W3110 DNA cloned into plasmid pAT153 for plasmids capable of complementing the *tyrB* or *aspC* lesions in strain HW225 (*aspC tyrB recA*). The pooled plasmid library was transformed into strain HW225, and Amp^r transformants were selected on L agar plates supplemented with carbenicillin at 200 µg/ml. The transformants were replica-plated on to minimal medium lacking tyrosine and aspartate to screen for those colonies that had acquired an intact *tyrB* or *aspC* gene and could consequently grow on the minimal medium.

Colonies that had acquired an Asp⁺ Tyr⁺ phenotype were purified and plasmid DNA was isolated from them. Restriction analysis of a number of plasmid isolates revealed two classes. The first class of transformant, isolated from *Bgl*III and *Eco*RI libraries, clearly carried

Fig. 4. Deletion analysis, restriction mapping and sequencing of the *aspC* gene

Sub-clones of *aspC* are shown above the restriction map. Δ depicts the small *Pst*I deletion of pIF7. The relative directions of transcription of *aspC* and *ompF* are illustrated by heavy arrows, with the intergenic strong stem-loop structure indicated ($\uparrow$). The sequencing strategy is illustrated below the gene; arrows with filled tails represent Maxam & Gilbert sequencing, and those with open tails represent regions sequenced by the dideoxy chain-termination method.

Fig. 5. DNA sequence of the *aspC* gene

The *aspC* coding sequence comprises 1188 base-pairs, initiating at the ATG at position 138. This is preceded by a four-base Shine & Dalgarno sequence as indicated by asterisks (rbs, ribosome-binding site). The mRNA initiation point was determined by means of run-off transcripts to correspond closely to the best-consensus -10 and -35 sequences as predicted by computer (Staden, 1984c). These are also indicated in the Figure. No regions of dyad symmetry were detected in the DNA sequence immediately after the TAA stop of the aspartate aminotransferase coding sequence. The deduced amino acid sequence agrees completely with that determined by Kondo *et al.* (1984) for the aspartate aminotransferase of *E. coli* B.

the *tyrB* gene. The second class, isolated from a *Bgl*III library, showed no restriction-site homology with the *tyrB* clone. Subsequent native polyacrylamide-gel-electrophoretic analysis confirmed that one plasmid of this class, pME133, indeed carried the *aspC* gene (results not shown). pME133 is a large plasmid, carrying an insert of about 15 kb, and so the location of the *aspC* gene was narrowed down by sub-cloning fragments and testing for complementation of the *tyrB aspC* alleles in strain HW225. The *aspC* gene was first sub-cloned on a 4.5 kb *Hind*III-*Sal*I-digest fragment to give plasmid pME101. Subsequent sub-cloning and deletion analysis established

that the *aspC* gene is encoded on a 1.5 kb fragment bounded by *Hinc*II and *Stu*I sites. The deletion analysis is summarized in Fig. 4.

Determination of the DNA sequence of the *aspC* gene

The sequence of the *aspC* gene along with its flanking regions was ascertained by a combination of the chemical method of Maxam & Gilbert and the dideoxy chain-termination procedure of Sanger (Biggin *et al.*, 1983). Extensive use was made of synthetic oligonucleotides to allow the sequencing of large fragments cloned into M13 vectors and so minimize the amount of sub-cloning

required. The sequencing strategy is outlined in Fig. 4 and the complete sequence is shown in Fig. 5.

The *aspC* gene

The *aspC*-gene sequence contains one long open reading frame from residue 126 to residue 1325. Statistical analysis identical with that used on the *tyrB* gene revealed that the reading frame probably represents a true gene. The only candidate for the initiation codon is the ATG at position 138. It is preceded by a Shine & Dalgarno consensus sequence GGAA at position 127. The next in-phase ATG occurs too far into the reading frame to account for the known size of the aspartate aminotransferase. The predicted amino acid sequence agrees totally with that obtained from the purified aspartate aminotransferase from *E. coli* B (Kondo *et al.*, 1984). Transcription/translation *in vitro* and subsequent analysis of the *aspC*-gene product revealed a size, based on relative mobility on SDS/polyacrylamide-gel electrophoresis, of 43 000 Da (results not shown).

Transcriptional signals for the *aspC* gene

Attempts to extend the *aspC* sequence further upstream were hindered by an extremely strong compression in the sequence, which hindered read through by the polymerase. This finding, along with the A+T-rich nature of this region of the sequence, indicates the presence of a strong terminator. An examination of the genetic map in the *aspC* region of the chromosome (20.75 min) revealed no genes that had been mapped as very close to *aspC*. The nucleotide sequence of the *ompF* gene (Inokuchi *et al.*, 1982), which maps at 20.9 min and encodes one of the major outer-membrane porins, was obtained from the National Biomedical Research Foundation data base and checked for homology with the upstream region of *aspC*. We found that the *ompF* and *aspC* genes are in fact contiguous and transcribed in the same direction. The published *ompF* sequence extends as far as the first few N-terminal residues of *aspC*-gene product. Since *ompF* is a very highly expressed gene, it presumably demands the presence of a particularly effective terminator and explains the problems that we observed in sequencing the intergenic region. The arrangement of the two genes is illustrated in Fig. 4.

That the *aspC* gene has its own promoter and does not just rely on read through from the *ompF* gene is demonstrated by the following observation. Plasmids such as pIF100, containing restriction fragments carrying the *aspC* gene but lacking the *ompF* promoter, still express *aspC* strongly. Transcription run-off experiments and reverse run-off experiments (Hudson & Davidson, 1984) were performed on an isolated *HincII*-*AvaI*-digest fragment carrying the intergenic region. A predominant RNA species of 129 bp was observed, whose size corresponds well to the predicted transcription start point for the promoter-like sequence found by a computer program that scans for potential *E. coli* promoters (Staden, 1984a,b,c). Transcription probably commences at the A or G residue 33 or 34 bp upstream of the *aspC* translation start. There are no dyad symmetries in the promoter region that might be indicative of regulator binding sites. Expression of *aspC* is not known to be subject to any form of regulation. Limited sequence analysis downstream of the gene has not revealed any terminator-like sequences.

Table 4. Overproduction of the *tyrB*-gene and *aspC*-gene products: measurement of aspartate aminotransferase activities

Aspartate aminotransferase activities were measured in crude extracts and corrected for endogenous NADH oxidase activity. The extracts were prepared from cells grown in minimal medium supplemented with 0.2% glucose and 0.4% casamino acids. Cultures were harvested at $A_{600} = 0.8$, washed once in 20 mM-potassium phosphate buffer, pH 7.8, and resuspended in the same buffer containing 100 μ M-pyridoxal 5'-phosphate. The samples were broken by sonication (2×30 s; 15 μ m) and cell debris was removed by centrifugation (16 500 g for 10 min). Protein concentration was determined with BioRad reagent for determination of protein.

Strain/plasmid	Aspartate aminotransferase activity (units/mg of protein)	Relative activity
HW79	0.38	1
HW857	< 0.05	0
HW857/pIF100	42.2	111
HW857/pME64	12.7	33

Overproduction of the *tyrB*-gene and *aspC*-gene products

Overproduction of both the aromatic aminotransferase and the aspartate aminotransferase caused by plasmids pME64 and pIF100 in strain HW857 (*tyrB aspC* del) was demonstrated by staining native polyacrylamide gels, after electrophoresis, for phenylalanine aminotransferase activity. An intensely staining band was observed in the positions expected for the *aspC*-gene product with plasmid pIF100 and the *tyrB*-gene product with plasmid pME64 (results not shown). The extent of the overproduction was measured by assaying crude extracts for aspartate aminotransferase activity (Table 4). Approx. 100-fold overproduction of aspartate aminotransferase and 30-fold overproduction of the aromatic aminotransferase were obtained compared with a wild-type control.

Comparison of the nucleotide sequences of *tyrB* and *aspC* genes

The *tyrB*-gene and *aspC*-gene products are quite clearly related proteins, as the analysis in the next section shows. This homology is reflected in the similarity of the DNA sequence (see Fig. 6). Both the DNA and the protein sequences have been aligned by a single deletion of three bases for the DNA and of one amino acid residue for the protein (see Fig. 6). The overall DNA homology is 51%. The DNA that encodes regions of the protein not highly conserved has obviously undergone considerable evolutionary change. This must not only reflect the differences in substrate specificity of the two proteins but must also represent the effect of neutral mutations on the functionally less-important regions of the gene. It is interesting that at the codons for conserved amino acid residues the choice of synonymous codons has been effectively randomized, with the only significant exceptions to this (codons for arginine, leucine and glycine) being attributable to a strong codon bias. Thus it would appear that the two genes have diverged as much as is compatible with the continued functioning of the enzymes and the constraints of codon preference in *E. coli*.


```

      10      20      30
M F E N I T A A P A D P I L G L A D L F R A D E R P G K I N
ATG TTT GAG AAC ATT ACC GCC GCT CCT GCC GAC CCG ATT CTG GGC CTG GCC GAT CTG TTT COT GCC GAT GAA COT CCC GGC AAA ATT AAC
   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
GTG TTT CAA AAA GTT GAC GCC TAC GCT GGC GAC CCG ATT CTT ACG CTT ATG GAG CGT TTT AAA GAA GAC CCT CGC AGC GAC AAA GTG AAT
M F Q K V D A Y A G D P I L T L M E R F K E D P R S D K V N

      40      50      60
L G I G V Y K D E T G K T P V L T S V K R A E Q Y L L E N E
CTC GCG ATT GGT GTC TAT AAA GAT GAG ACG GGC AAA ACC CCG GTA CTG ACC AGC GTG AAA AAG GCT GAA CAG TAT CTG CTC GAA AAT GAA
   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
TTA AGT ATC GGT CTG TAC TAC AAC GAA GAC GGA ATT ATT CCA CAA CTG CAA GCC GTG GCG GAG GCG GAA GCG CGC CTG AAT GCG CAG CCT
L S I G L Y Y N E D G I I P Q L Q A V E A E A R L N A Q C P

      70      80      90
T T K N - Y L G I D G I P E F G R C T Q E L L F G K G S A L
ACC ACC AAA AAT TA- CTC GGC ATT GAC GGC ATC CCT GAA TTT GGT GCG TGC ACT CAG GAA CTG CTG TTT GGT AAA GGT AGC GCC CTG
   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
CAT GGC GCT TCG CTT TAT TTA CCG ATG GAA GGG CTT AAC TGC TAT CGC CAT GCT ATT GCG CGC CTG CTG TTT GGT GCG GAT GAT CCG GTA
H G A S L Y L P M E G L N C Y R H A I A P L L F G A D H P V

      100     110     120
I N D K R R A R T A Q T P G G T G A L R V A A D F L A K N T S
ATC AAT GAC AAA CGT GCT CGC ACG GCA CAG ACT CCG GGG GGC ACT GGC GCA CTA CGC GTG GCT GAT TTC CTG GCA AAA AAT ACC AGC
   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
CTG AAA CAA CAG CGC GTA GCA ACC ATT CAA ACC CTT GGC GCG TCC GGG GCA TTG AAA GTG GCG GCG GAT TTC CTG AAA CGC TAC TTC CCG
L K Q Q R V A T I Q T L G G S G A L K V G A D F L K R Y F P

      130     140     150
V K R V W V S N P S W P N H K S V F N S A G L E V R E Y A Y
GTT AAG CGT GTG TGG GTG AGC AAC CCA AGC TGG CCG AAC CAT AAG AGC GTC TTT AAC TCT CCA GGT CTG GAA GTT CGT GAA TAC GCT TAT
   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
GAA TCA GGC GTC TGG GTC AGC GAT CCT ACC TGG GAA AAC CAC GTA GCA ATA TTC GCC GGG GCT GGA TTC GAA GTG AGT ACT TAC CCC TGG
E S G V W V S D T W E N H A I F A G F E V S T Y P W

      160     170     180
Y D A E N H T L D F D A L I N S L N E A Q A G D V V L F H G
TAT GAT CCG GAA AAT CAC ACT CTT GAC TTT GAT GCA CTG ATT AAC AGC CTG AAT GAA GCT CAG GCT GGC GAC GTA GTG CTG TTC CAT GGC
   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
TAT GAC GAA CCG ACT AAC GGC GTG GCG TTT AAT GAC CTG TTG GCG ACG AAA ACA CTA CCT GCC CGC AGT ATT GTG TTG CTG CAT CCA
Y D E A T N G V R F N D L L A T L K T T L P A R S I V L L H P

      190     200     210
C C H N P T G I D P T L E Q W Q T L A Q L S V E K G W L P L
TGC TGC CAT AAC CCA ACC GGT ATC GAC CCT ACG CTG GAA CAA TGG CAA ACA CTG GCA CAA CTC TCC GTT GAG AAA GGC TGG TTA CCG CTG
   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
TGT TGC CAC AAC CCA ACG GGT GCC GAT CTC ACT AAT GAT CAG TGG GAT GCG GTG ATT GAA ATT CTC AAA GCC CGC GAG CTT ATT CCA TTC
C C H N P T G A D L T N D Q W D A V I E I L R A R E L I P F

      220     230     240
F D F A Y O G F A R G L E E D A E G L R A F A A H H K E L
TTT GAC TTC GCT TAC CAG GGT TTT GCC CGT GGT CTG GAA GAA GAT GCT GAA GGA CTG CGC GCT TTC GCG GCT ATG CAT AAA GAG CTG ATT
   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
CTC GAT ATT GCC TAT CAA GGA TTT GGT GCC GGT ATG GAA GAG GAT GCC TAC GCT ATT CCG GCC ATT GCC AGC GCT GGA TTA CCC GCT CTG
L D I A Y Q G F G A G H E E D A Y A I R A I A S A G L P A L

      250     260     270
V A S S Y S K N F G L Y N E R V G A C T L V A A D S E T V D
GTT GCC AGT TCC TAC TCT AAA AAC TTT GGC CTG TAC AAC GAG CGT GTT GGC GCT TGT ACT CTG GTT GCT GCC GAC AGT GAA ACC GTT GAT
   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
GTG AGC AAT TCG TTC TCC AAA ATT TTC TCC CTT TAC GGC GAG CGC GTC GGC GGA CTT TCT GTT ATG TGT GAA GAT GCC GAA GCC GCT GGC
V S N S F S K I F S L Y G E R V G G L S V H C E D A E A A G

      280     290     300
R A F S Q M K A A I R A N Y S N P P A H G A S V V A T I L S
CGC GCA TTC AGC CAA ATG AAA GCG GCG ATT CGC GCT AAC TAC TCT AAC CCA CCA GCA CAC GGC GCT TCT GTT GTT GCC ACC ATC CTG AGC
   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
CGC GTA CTG GCG CAA TTG AAA GCA ACA GTT CGC CGC AAC TAC TCC AGC CCG CCG AAT TTT GGT GCG CAG GTG GTG GCT GCA GTG CTG AAT
R V L G Q L K A T V R R N Y S S P P N F G A Q V V A A V L N

      310     320     330
N D A L R A I W E Q E L T D M R Q R I Q R M R Q L F V N T L
AAC GAT CCG TTA CGT GCG ATT TGG GAA CAA GAG CTG ACT GAT ATG CGC CAG CGT ATT CAG CGT ATG CGT CAG TTG TTC GTC AAT ACG CTG
   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
GAC GAG GCA TTG AAA GCC AGC TGG CTG GCG GAA GTA GAA GAG ATG CGT ACT CGC ATT CTG GCA ATG CGT CAG GAA TTG GTG AAG GTA TTA
D E A L K A S W L A E V E E H R T R I L A H R Q E L V K V L

      340     350     360
Q E K C A N R D F S F I I K Q N G M P S F S G L T K E Q V L
CAG GAA AAA GGC CCA AAC CGC GAC TTC AGC TTT ATC ATC AAA CAG AAC GGC ATG TTC TCC TTC AGT GGC CTG ACA AAA GAA CAA GTG CTG
   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
AGC ACA GAG ATG CCA GAA CGC AAT TTC GAT TAT CTG CTT AAT CAG CGC GGC ATG TTC AGT TAT ACC GGT TTA AGT GCC GCT CAG GTT GAC
S T E H P E R N F D Y L L N Q R G R F S Y T G L S A A Q V D

      370     380     390
R L R E E F G V Y A V A S G R V N V A G M T P D N M A P L C
CGT CTG CCG GAA GAG TTT GCG GTA TAT GCG GTT GCT TCT GGT CCG GTA AAT GTG GCC GGG ATG ACA CCA GAT AAC ATG GCT CCG CTG TGC
   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
CGA CTA CGT GAA GAA TTT GGT GTC TAT CTC ATC GCC AGC GGT CCG ATG TGT GTT GCC GGG TTA AAT ACG GCA AAT GTA CAA CGT GTG GCA
R L R E E F G V Y L I A S G R M C V A G L N T A N V Q R V A

      E A I V A V L
      GAA CCG ATT GTG GCA GTG CTG aspC
      AAG CCG TTT GCT GCG GTG ATG tyrB
      K A F A A V H

```

Fig. 6. Comparison of the *aspC* and *tyrB* DNA sequences

The best alignment of the two sequences is achieved by assuming a deletion of three bases at residue 60 in the *aspC* sequence. Matching residues are indicated by asterisks.

Fig. 7. Amino acid sequence comparison of the aspartate aminotransferase (AspC) and aromatic aminotransferase (TyrB) of *E. coli* with the two pig aspartate aminotransferase isoenzymes

Boxed regions represent homology between either of the *E. coli* enzymes with either of the pig enzymes. Asterisks mark those residues that are involved in coenzyme binding or substrate binding or that are in close proximity to the active site. The mitochondrial enzyme sequence was taken from Barra *et al.* (1980) and the cytoplasmic enzyme sequence from Doonan *et al.* (1975).

Comparison of the *aspC*-gene and *tyrB*-gene products with the sequences of the pig aspartate aminotransferases

Many eukaryotic aspartate aminotransferases have been sequenced, and structures have been determined for isoenzymes from both mitochondria and cytoplasm. The same isoenzymes from different species are more alike (approx. 85% homology) than the different isoenzymes from the same species (approx. 50% homology), but all known eukaryotic aminotransferases have similar sequences (Graf-Hausner *et al.*, 1983) and structures (Harutyunyan *et al.*, 1982). The protein sequences of the *E. coli* enzymes are very similar to those of the eukaryotic aminotransferases, and a comparison of the *E. coli* enzyme sequences and those of the two pig enzymes is given in Fig. 7. The protein sequence of *E. coli* B aspartate aminotransferase has been reported by Kondo *et al.* (1984). We have not detected any differences in primary structure between this and the *E. coli* K12 aspartate aminotransferase.

The following discussion was greatly facilitated by a detailed description of the structure of the mitochondrial aspartate aminotransferase (Jansonius *et al.*, 1985). We

have adopted the residue numbering convention based on the cytoplasmic aspartate aminotransferase sequence. Optimal alignment with the pig enzyme sequences was achieved without the need for insertions in the *E. coli* enzyme sequences. However, deletions were required in several positions to maximize the homology: at residues 1-4, 128-132, 153, 232, 405 and 410-412, and a further deletion, tentatively assigned to position 64, for aspartate aminotransferase. Deletions were also required in the mitochondrial isoenzyme at positions 1-2, 64, 127-128, 131-132, 153, 407 and 411-412 to maximize the homology with the cytoplasmic enzyme. In this case, the deletions are allowable, since these positions occur at the protein surface, where local conformational differences may be more readily accommodated. The *E. coli* aminotransferases have deletions in similar positions except at residue 232, which in the eukaryotic enzyme structures occurs in a surface loop, where again a deletion would not have serious structural consequences. The similarity in deletion positions between the bacterial and mitochondrial enzyme sequences is an intriguing aspect of this work.

The homology between the bacterial and eukaryotic

enzymes is summarized in Table 5. The homology is approx. 40% to either of the eukaryotic isoenzymes, which share approx. 50% homology between themselves. The *E. coli* aminotransferases have approx. 54% homology to the eukaryotic enzyme sequences if identities in both pig isoenzyme sequences are scored, and the homology extends to approx. 70% when conservative substitutions are taken into account. This degree of similarity is consistent with evolutionary divergence from a common ancestral gene at the time of emergence of the eukaryotic cell.

The primary sequence homology is sufficient to suggest that the bacterial enzymes have three-dimensional structures very similar to those determined for the mitochondrial and cytoplasmic isoenzymes. The proposed structural homology can be tested by a close examination of the degree of conservation in the tightly packed interior of the molecule, and also by the conservation of glycine residues, which often have a specific structural role. On the former point, extensive homology extending to approx. 80% identity, with the remainder being conservative substitutions, was found, except for an unusual substitution of tyrosine to arginine at position 399 in the aromatic aminotransferase and to proline in the aspartate aminotransferase. With regard to glycine, of the 22 conserved residues found in eukaryotic aspartate aminotransferase sequences, 19 are conserved in the aromatic aminotransferase and 18 in aspartate aminotransferase from *E. coli*. Proline residues have not been conserved to a high degree (approx. 50%), but in the mitochondrial aspartate aminotransferase structure only three proline residues are internal, and of these only one, Pro-299 on the subunit interface, is invariant in the eukaryotic enzyme sequences. This residue is conserved, as are proline residues identified as taking part in turns, in the bacterial sequences.

The active site of aspartate aminotransferase comprises structural elements from both subunits. Tyr-70 and Arg-292 intrude into the active site of the adjacent subunit, and are thought to have catalytic and substrate-binding functions respectively. Both these residues are conserved in the bacterial enzyme sequences. Consequently, in addition to homologous subunit structures, the bacterial enzymes appear to have homologous intersubunit interactions. This homology also extends to an interaction between the hydrophobic residues Met-5 and Phe-6 of the *N*-terminal arm, which locate into a hydrophobic patch on the surface of the adjacent subunit, helping to stabilize the dimeric structure.

The homology extends to other residues having substrate-binding and catalytic significance. Almost all the residues identified as such in the aspartate aminotransferase structures are conserved as indicated by asterisks in Fig. 7. This includes the attachment of Lys-258 to the ring of the pyridoxal 5'-phosphate cofactor and of the ligands to the phosphate group. Charge compensation for the phosphate ions is provided by Arg-266 and by the positive end of the dipole (Ford *et al.*, 1980) formed by a conserved α -helix (residues 108–122). Hydrogen-bonding interactions to pyridoxal 5'-phosphate also appear to be conserved. Interestingly, the two bacterial sequences and the cytoplasmic enzyme do not have a phosphate ligand, as provided by Ser-107 in the mitochondrial enzyme, but instead have an interaction through Ser-257 (Arnone *et al.*, 1982). To judge from the

Table 5. Summary of the homology between the bacterial and mammalian aminotransferases

The top left section of the Table gives the number of identical residues between any given pair of enzymes. The same data are presented in percentage terms in the bottom right of the Table. Numbers along the diagonal represent the lengths (numbers of residues) of the enzymes. Key: mAAT, pig mitochondrial aspartate aminotransferase; cAAT, pig cytoplasmic aspartate aminotransferase; TyrB, *E. coli* aromatic aminotransferase; AspC, *E. coli* aspartate aminotransferase.

Conserved residues					Homology (%)
TyrB	170	157	161	397 100%	
mAAT	161	196	401 100%	40.8 %	
cAAT	160	412 100%	48.9 %	39.5 %	
AspC	396 100%	40.4 %	40.6 %	42.9 %	
	AspC	cAAT	mAAT	TyrB	

conservation of residues involved in coenzyme binding it would appear that the pyridine ring is in a very similar conformation and environment in all the enzymes, except for the substitution of serine for asparagine at position 297 in the aromatic aminotransferase, the significance of which is unknown. Residues 296 and 297 are further examples of intersubunit interactions, as they originate from the adjacent subunit and form a portion of the wall of the coenzyme-binding pocket.

Concerning substrate binding, residues identified with this function in the eukaryotic enzyme crystal structures are completely conserved. The overall degree of conservation in the coenzyme-binding and substrate-binding sites is so great that it is not possible to discern any features that would account for the differing substrate specificities of any of these four enzymes, including the aromatic aminotransferase. It had been suggested (Jansonius *et al.*, 1985) that small structural differences between the mitochondrial and cytoplasmic isoenzymes must account for their differing substrate specificities and that the residues Asp-15 and Ile-17, in Van der Waals contact with the substrate α -carboxylate-binding residue Arg-386, constitute a specific structural feature of the mitochondrial isoenzymes. Interestingly, Asp-15 and Ile-17 are conserved in the *E. coli* enzymes, which are both capable of catalysing the transamination of aliphatic and aromatic oxo acids, although with different K_m and V_{max} values.

Having demonstrated gross conservation of structurally and functionally important residues it appears certain that the *E. coli* enzymes have structures and mechanisms

that are virtually identical with those determined for the eukaryotic enzymes and are attractive candidates for future protein engineering experiments. The residues that are conserved are concerned with structural integrity and with coenzyme and substrate binding. Even though the external portions of the molecule have undergone extensive divergence, it is interesting to note that structural and mechanistic integrity may be maintained even at the level of 40% homology and over 600×10^6 – 1000×10^6 years of evolution.

We are indebted to Mrs F. Webb for technical assistance, Dr. B. Bachmann for provision of bacterial strains, Dr. J. F. Pardon for help with the computer analysis of sequence data and Dr. B. M. Richards for the use of excellent laboratory facilities.

REFERENCES

- Arnone, A., Briley, P. D., Rogers, P. H., Hyde, C. D., Metzler, C. M. & Metzler, D. E. (1982) in *Molecular Structure and Biological activity* (Griffen, J. F. & Duax, W. L., eds.), pp. 57–77, Elsevier/North-Holland, New York.
- Bachmann, B. J. (1983) *Microbiol. Rev.* **47**, 180–230.
- Barra, D., Bossa, F., Doonan, S., Fahmy, H. M. A., Hughes, G. J., Martini, F., Petruzzelli, R. & Wittmann-Liebold, B. (1980) *Eur. J. Biochem.* **108**, 405–414.
- Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 3963–3965.
- Bochner, B. R., Huang, H.-C., Schieuen, G. L. & Ames, B. N. (1980) *J. Bacteriol.* **143**, 926–933.
- Bradford, M. (1976) *Anal. Biochem.* **72**, 248–254.
- Casadaban, M. & Cohen, S. N. (1980) *J. Mol. Biol.* **138**, 179–207.
- Clarke, S. J., Low, B. & Konisberg, W. H. (1973) *J. Bacteriol.* **113**, 1091–1095.
- Collier, R. H. & Kohlaw, G. (1972) *J. Bacteriol.* **112**, 365–371.
- Davies, W. D. & Davidson, B. E. (1982) *Nucleic Acids Res.* **10**, 4045–4058.
- Dempsey, W. B. (1969) *J. Bacteriol.* **100**, 1114–1115.
- Doonan, S., Doonan, H. J., Hanford, R., Vernon, C. A., Walker, J. M., Airoidi, L. P. da S., Bossa, F., Barra, D., Carloni, M., Fasella, P. B. & Riva, F. (1975) *Biochem. J.* **149**, 497–506.
- Falkingham, J. O., III (1979) *Mol. Gen. Genet.* **176**, 147–149.
- Ford, G. C., Eichele, G. & Jansonius, J. N. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 2559–2563.
- Gelfand, D. H. & Rudo, N. (1977) *J. Bacteriol.* **130**, 441–444.
- Gelfand, D. H. & Steinberg, R. A. (1977) *J. Bacteriol.* **130**, 429–440.
- Goldschmidt, E. P. & Cater, M. S. (1970) *Genetics* **66**, 219–229.
- Graf-Hausner, U., Wilson, K. J. & Christen, P. (1983) *J. Biol. Chem.* **258**, 8813–8826.
- Harutyunyan, E. G., Malashkevich, V. N., Tersyan, S. S., Kochkina, V. M., Torchinsky, Yu. M. & Braunstein, A. E. (1982) *FEBS Lett.* **138**, 113–116.
- Hohn, B. & Collins, J. (1980) *Gene* **11**, 291–298.
- Howard-Flanders, P., Boyce, R. P. & Theriot, L. L. (1966) *Genetics* **53**, 1119–1136.
- Hudson, G. S. & Davidson, B. E. (1984) *J. Mol. Biol.* **180**, 1023–1051.
- Inokuchi, K., Mutoh, N., Matsuyama, S.-I. & Kagamiyama, H. (1982) *Nucleic Acids Res.* **10**, 6957–6968.
- Ish-Horowicz, D. & Burke, J. F. (1981) *Nucleic Acids Res.* **9**, 2989–2998.
- Jack, G. W. & McMahon, P. C. (1978) *Biochim. Biophys. Acta* **523**, 344–357.
- Jansonius, J. N., Eichele, G., Ford, G. C., Picot, D., Thaller, C. & Vincent, M. G. (1985) in *Transaminases* (Christen, P. & Metzler, D. E., eds.), pp. 109–138, John Wiley and Sons, New York.
- Kirsch, J. F., Eichele, G., Ford, G. C., Vincent, M. G. & Jansonius, J. N. (1983) *J. Mol. Biol.* **174**, 497–525.
- Kleckner, N., Roth, J. & Botstein, D. (1977) *J. Mol. Biol.* **116**, 125–159.
- Kline, E. L., Manross, D. N., Warwick, M. L. (1977) *J. Bacteriol.* **130**, 951–953.
- Kondo, K., Wakabayashi, S., Yagi, T. & Kagamiyama, H. (1984) *Biochem. Biophys. Res. Commun.* **122**, 62–67.
- Lawther, R. P., Nichols, B., Zurawski, G. & Hatfield, G. W. (1979) *Nucleic Acids Res.* **7**, 2289–2301.
- Lee-Peng, F.-C., Hermodson, M. A. & Kohlaw, G. B. (1979) *J. Bacteriol.* **139**, 339–345.
- Mavrides, C. & Christen, P. (1978) *Biochem. Biophys. Res. Commun.* **85**, 769–773.
- Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. U.S.A.* **74**, 560–564.
- Messing, J. & Vieira, J. (1982) *Gene* **19**, 269–276.
- Miller, J. H. (1972) *Experiments in Molecular Genetics*, Cold Spring Harbor Laboratory, Cold Spring Harbor.
- Peterkofsky, B. & Gilvarg, C. (1961) *J. Biol. Chem.* **236**, 1432–1438.
- Powell, J. T. & Morrison, J. F. (1978) *Eur. J. Biochem.* **87**, 391–400.
- Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463–5467.
- Shine, J. & Dalgarno, L. (1975) *Nature (London)* **254**, 34–38.
- Silbert, D. F., Jorgensen, S. E. & Lin, E. C. C. (1963) *Biochim. Biophys. Acta* **73**, 232–240.
- Staden, R. (1984a) *Nucleic Acids Res.* **12**, 499–503.
- Staden, R. (1984b) *Nucleic Acids Res.* **12**, 505–519.
- Staden, R. (1984c) *Nucleic Acids Res.* **12**, 521–538.
- Tinoco, I., Borer, P., Dengler, B., Levine, M., Uhlenbeck, O., Crothers, D. & Gralla, J. (1973) *Nature (London)* **246**, 40–41.
- Twigg, A. J. & Sherratt, D. (1980) *Nature (London)* **283**, 216–218.
- Umbarger, H. E., Umbarger, M. A. & Siu, P. L. (1963) *J. Bacteriol.* **85**, 1431–1439.
- Urdea, M. S., Merryweather, J. P., Mullenbach, G. T., Coit, D., Heberlein, U., Valenzuela, P. & Barr, J. (1983) *Biochemistry* **80**, 7461–7465.
- Vapnek, D., Alton, N. E., Bassett, C. L. & Kushner, S. R. (1976) *Proc. Natl. Acad. Sci. U.S.A.* **73**, 3492–3496.
- Vogel, H. J. & Bonner, D. M. (1956) *J. Biol. Chem.* **218**, 97–102.
- Vogel, H. J., Bacon, D. F. & Baich, A. (1963) *Proc. Symp. Informational Macromolecules* 293–300.
- Wallace, B. J. & Pittard, J. (1969) *J. Bacteriol.* **97**, 1234–1241.
- Wechsler, J. A. & Gross, J. D. (1971) *Mol. Gen. Genet.* **113**, 273–284.
- Whalen, W. A. & Berg, C. M. (1982) *J. Bacteriol.* **150**, 739–746.