The nucleotide sequence of the *ilvBN* operon of *Escherichia coli*: sequence homologies of the acetohydroxy acid synthase isozymes

Ronald C.Wek, Craig A.Hauser and G.Wesley Hatfield

Department of Microbiology and Molecular Genetics, California College of Medicine, University of California, Irvine, CA 92717, USA

Received 4 March 1985; Revised and Accepted 7 May 1985

ABSTRACT

Three acetohydroxy acid synthase isozymes, AHAS I (*ilvBN*), AHAS II (*ilvGM*) and AHAS III (*ilvIH*) catalyze the first step of the parallel isoleucine-valine biosynthetic pathway in *Escherichia coli*. Previous DNA sequence and protein purification data have shown that AHAS II and AHAS III are composed of large and small subunits encoded in the *ilvG* and *ilvH* operons, respectively. Recent protein purification and characterization data have demonstrated that the AHAS I isozyme is also composed of large and small subunits (L. Eoyang, L. and P. M. Silverman [1984] J. Bacteriol. 157:184-189). Now the complete DNA sequence of the operon encoding the AHAS I isozyme has been determined. These data show that both AHAS I subunits (M_r 60,400 and M_r 11,100) are encoded in this operon. The coordinant regulation of both genes of the *ilvBN* operon has also been demonstrated. Comparisons of the DNA sequences of the genes encoding all three AHAS isozymes have been performed. Conserved homologies were observed between both the large and small subunits of all three isozymes. The closest homology was seen between the AHAS I and AHAS II isozymes. On the basis of these comparisons a rationale for the evolution of the AHAS isozymes in *E. coli* has been proposed.

INTRODUCTION

Three isozymes catalyze the first step of the parallel isoleucine-valine biosynthetic pathway in *E. coli*. These isozymes are acetohydroxy acid synthase (AHAS) I, II and III (1). The AHAS III isozyme is encoded by the *ilvIH* operon (2,3). Analysis of the DNA sequence of this operon shows an *ilvI* open translational reading frame coding for a 61,800 dalton polypeptide and an *ilvH* reading frame coding for a smaller polypeptide of 17,500 daltons (3). Minicells transformed with a plasmid encoding the *ilvIH* operon synthesized products of these approximate sizes (3). *In vitro* complementation experiments using *ilvI* and *ilvH* mutants suggest that the large polypeptide functions as a catalytic subunit while the smaller polypeptide serves as a valine sensitive regulatory subunit (2).

AHAS II is encoded by the *ilvG* gene which resides in the *ilvGEDA* operon. The *ilvG* gene and the 5' portion of the *ilvE* gene have been sequenced (4). The *ilvG* gene codes for a 59,200 dalton polypeptide which is similar in size to the larger subunit in AHAS III. A comparison of the deduced amino acid sequence of both the *ilvI* and *ilvG* gene products show significant homology (3). The distance between the *ilvG* and *ilvE*

genes is 279 nucleotides. This region contains an open translational reading frame which could encode an 87 amino acid polypeptide with a molecular weight of 9,700 daltons. This predicted polypeptide shows significant homology with the ilvH gene product (3). Recently, Kutney and Schloss (Fed. Am. Abstr. 1984, vol. 43, no. 2630) have reported the purification of AHAS II from S. typhimurium LT2. AHAS II was resolved into a large polypeptide of 56,000 daltons and a smaller polypeptide of 10,000 daltons by SDS polyacrylamide gel electrophoresis. Amino terminal sequence of the large AHAS II subunit from S. typhimurium shows identical sequence homology with the deduced E. coli K-12 sequence with the exception of a substitution of a lysyl residue for an asparaginy residue at position 17. The smaller subunit shows similar but weaker amino terminal homology with 12 residues differing out of 35 residues sequenced. The gene coding for the smaller AHAS II subunit has been designated ilvM and the operon is now referred to as ilvGMEDA.

AHAS I isozyme was originally purified to near homogeneity by Grimmer and Umbarger (5). It was demonstrated that a 60,000 dalton polypeptide possessed both AHAS I catalytic activity and valine sensitivity. More recently, AHAS I was purified using a different protocol by Eoyang and Silverman (6). It was confirmed that AHAS I contained a 60,000 dalton polypeptide; however, a second smaller subunit of 9,500 daltons was also present in equimolar amounts. The physical association of these two subunits was confirmed by the following experiments. Antisera directed against the 60,000 dalton subunit coprecipitated both the large and small polypeptides from cell extracts labeled with [³⁵S]methionine. When the extracts were treated with SDS before immunoprecipitation only the 60,000 dalton subunit was recovered. On the basis of this physical association of the two subunits in cell extracts and with purified enzyme it was concluded that they were both AHAS I enzyme subunits.

AHAS I function is expressed by the ilvB operon (7,8). In this report minicells containing this operon are shown to express two polypeptides of 60,000 and 10,000 daltons. Both polypeptides are inducible by the addition of cAMP which is consistent with both products being under the control of the catabolite repressed ilvB promoter (9). The ilvB operon was sequenced and two open translational reading frames representing coding regions of 60,400 and 11,100 daltons were detected. The deduced amino acid sequences of both the large and small polypeptides show extensive homology with the respective large and small subunits of AHAS II and III. Thus, based on 1) the physical association of the large and small subunits in the purified AHAS I enzyme, 2) the cAMP inducible expression of both polypeptides in minicells containing the ilvB operon and 3) the homology of the deduced amino acid sequences of the two open translational reading frames in the ilvB operon with the subunits of AHAS II and III, it is concluded that AHAS I consists of a large and small subunit encoded by the ilvB

operon. The gene encoding the 60,400 dalton polypeptide is named ilvB and the gene encoding the 11,100 dalton small polypeptide is named ilvN. The operon is now referred to as ilvBN.

It is now clear that each AHAS isozyme consists of a large and small polypeptide. A comparison between both the large and small polypeptide sequences indicates close homology. A model for the evolution of the AHAS isozymes is presented based on these isozyme similarities.

MATERIALS AND METHODS

Minicells

Plasmid pCH4 contains a 6.5 kb PstI insert encoding AHAS I in plasmid vector pACYC177 (8). This plasmid was transformed into E. coli K-12 strain DS410 (10) and minicells were purified essentially as described by Roozen *et al.* (11). Minicells obtained from a 350 ml culture were washed and resuspended in 2 ml of 25% (vol/vol) methionine assay medium (Difco), M9 salts (12) and 0.5% glucose. They were labeled with [³⁵S]methionine (sp. act. 600 to 1400 Ci/mmol; Amersham, Arlington Heights, IL) at 20 μ Ci/ml for 45 min, pelleted and frozen. For induction of protein synthesis with cAMP, the minicells were incubated in 3 mM cAMP for 20 minutes prior to labeling. The pellets were resuspended in 1 ml of Laemlli sample buffer (13) and heated at 90°C for 2 min. A sample of approximately 100,000 cpm was then electrophoresed on a 10% SDS-polyacrylamide gel (13).

DNA Sequencing

The ilvBN operon was sequenced by the dideoxy chain termination method of Sanger (15) as illustrated in Figure 1. All regions of both strands of DNA were sequenced from overlapping recombinant phage M13 clones.

RESULTS

Plasmid pCH4, which has a 6.5 kb PstI insert containing the ilvBN operon, was transformed into E. coli K-12 strain DS410. The [³⁵S]methionine labeled proteins synthesized in minicells produced by this strain were electrophoresed on an SDS polyacrylamide gel and an autoradiograph of this gel is shown in Figure 2. The results show that two polypeptides with molecular weights of 60,000 and 10,000 daltons are encoded by the plasmid pCH4 insert. The synthesis of both polypeptides is stimulated by the addition of cAMP to the minicell system. In the absence of cAMP neither product is detected. This is due to the assay medium which would both catabolite repress and attenuate the operon and possibly further repress the operon due to the rich broth effect. The molecular weights of both polypeptides are consistent with the sizes of the polypeptides derived from purified AHAS I (6).

A restriction map of the ilvBN region is shown in Figure 1. Previously, the ilvB attenuator region was sequenced (8,14) and these sequence data suggested the location of the amino terminus of the ilvB product. The Sanger sequence method (15) was used to determine the entire ilvBN coding region. The 2,462 nucleotide DNA sequence of the non-transcribed strand is shown in Figure 3. The numbering starts at the transcriptional initiation site as proposed by Hauser and Hatfield (8). It is noted that Friden et al. (14) proposed an alternative initiation site at an adenine residue four nucleotides further upstream.

Between positions -75 to 192 are sequences previously demonstrated to be involved in the promotion and attenuation of the ilvB operon (8,14,16,17). A putative 32 amino acid leader polypeptide deduced from the DNA sequence is shown (Fig. 3). This leader polypeptide contains codons for Thr, Leu, Ala and Val, which have been shown in vivo to regulate attenuation (16). At positions 236 through 1924 there is a translational open reading frame with a potential coding region for a polypeptide of 562 amino acids. Such a polypeptide would have a molecular weight of 60,400 daltons which is similar to the large polypeptide produced in pCH4-transformed minicells (Fig. 2). An AGGA at position 225 is a potential ribosome binding site (Fig. 3).

Four nucleotides downstream from the large translational open reading frame stop codon is the start of a second smaller open reading frame. This coding region which extends from position 1928 to 2218 could code for a 96 amino acid peptide with a molecular weight of 11,100 daltons (Fig. 3). This size is similar to the 10,000 dalton polypeptide produced in pCH4-transformed minicells (Fig. 2). A potential ribosome binding site is at position 1915 where there are five G residues (Fig. 3).

Figure 2. Autoradiogram of a 10% SDS polyacrylamide gel displaying [^{35}S]methionine labeled polypeptides synthesized in *E. coli* minicells. In lanes 1 and 2 are the labeled products from minicells containing the plasmid vector pACYC177. The major polypeptides synthesized are β -lactamase (M_r 29,000) and the kanamycin resistant determinant (M_r 27,000). In lanes 3 and 4 are the labeled products from minicells containing the plasmid pCH4, which encodes the *ilvBN* operon. The products shown in lanes 2 and 4 were induced with cAMP before labeling. The numbers to the left are molecular weight markers in kilodaltons. The arrows on the right indicate the major polypeptides encoded by the pCH4 insert, also in kilodaltons.

At the 3'-end of the *ilvBN* operon is a probable rho-independent terminator (18). Centered at position 2253 is a G+C rich stem-loop structure which is followed by eight T residues.

DISCUSSION

The results of this work have further defined the *ilvBN* operon which encodes AHAS I. Minicells transformed with pCH4, a plasmid containing the *ilvBN* operon, produced two polypeptides with molecular weights of 60,000 and 10,000 daltons. Both

[illegible]

1924
 GGC GGC AAT ACT GAA ATC GTC GGC GAA TAA GGC ATG CAA AAC ACA ACT CAT GAC AAC GTA ATT CTG GAG CTC ACC GTT CGC AAC CAT CCG
 Ala Ala Asn Thr Glu Met Val Gly Glu * Met Gln Asn Thr Thr His Asp Asn Val Ile Leu Glu Leu Thr Val Arg Asn His Pro

1954
 GGC GGC AAT ACT GAA ATC GTC GGC GAA TAA GGC ATG CAA AAC ACA ACT CAT GAC AAC GTA ATT CTG GAG CTC ACC GTT CGC AAC CAT CCG
 Ala Ala Asn Thr Glu Met Val Gly Glu * Met Gln Asn Thr Thr His Asp Asn Val Ile Leu Glu Leu Thr Val Arg Asn His Pro

2014
 GGC GTA ATG ACC GAC GTT TGT GGC CTT TTT GGC CGC CGC GCT TTT AAC GTT GAA GGC ATT CTT TGT CTG CCG ATT CAG GAC AGC GAC AAA
 Gly Val Met Thr His Val Cys Gly Leu Phe Ala Arg Arg Ala Phe Asn Val Glu Gly Ile Leu Cys Leu Pro Ile Gln Asp Ser Asp Lys

2044
 GGC GTA ATG ACC GAC GTT TGT GGC CTT TTT GGC CGC CGC GCT TTT AAC GTT GAA GGC ATT CTT TGT CTG CCG ATT CAG GAC AGC GAC AAA
 Gly Val Met Thr His Val Cys Gly Leu Phe Ala Arg Arg Ala Phe Asn Val Glu Gly Ile Leu Cys Leu Pro Ile Gln Asp Ser Asp Lys

2104
 AGC CAT ATC TGC CTA CTC GTC AAT GAC GAC CAG CGT CTG GAG CAG ATG ATA AGC CAA ATC GAT AAG CTG GAA GAT GTC GTG AAA CTG CAG
 Ser His Ile Trp Leu Leu Val Asn Asp Asp Gln Arg Leu Glu Gln Met Ile Ser Gln Ile Asp Lys Leu Glu Asp Val Val Lys Val Gln

2134
 AGC CAT ATC TGC CTA CTC GTC AAT GAC GAC CAG CGT CTG GAG CAG ATG ATA AGC CAA ATC GAT AAG CTG GAA GAT GTC GTG AAA CTG CAG
 Ser His Ile Trp Leu Leu Val Asn Asp Asp Gln Arg Leu Glu Gln Met Ile Ser Gln Ile Asp Lys Leu Glu Asp Val Val Lys Val Gln

2164
 AGC CAT ATC TGC CTA CTC GTC AAT GAC GAC CAG CGT CTG GAG CAG ATG ATA AGC CAA ATC GAT AAG CTG GAA GAT GTC GTG AAA CTG CAG
 Ser His Ile Trp Leu Leu Val Asn Asp Asp Gln Arg Leu Glu Gln Met Ile Ser Gln Ile Asp Lys Leu Glu Asp Val Val Lys Val Gln

2194
 CGT AAT CAG TCC GAT CCG ACG ATG TTT AAC AAG ATC GCG GTG TTT TTT CAG TAA CCGCTCAAG GCTTGAAACAA CATCGCGCTT ATCGTTAAGG
 Arg Asn Gln Ser Asp Pro Thr Met Phe Asn Lys Ile Ala Val Phe Phe Gln *

2227
 CGT AAT CAG TCC GAT CCG ACG ATG TTT AAC AAG ATC GCG GTG TTT TTT CAG TAA CCGCTCAAG GCTTGAAACAA CATCGCGCTT ATCGTTAAGG
 Arg Asn Gln Ser Asp Pro Thr Met Phe Asn Lys Ile Ala Val Phe Phe Gln *

2247
 CGT AAT CAG TCC GAT CCG ACG ATG TTT AAC AAG ATC GCG GTG TTT TTT CAG TAA CCGCTCAAG GCTTGAAACAA CATCGCGCTT ATCGTTAAGG
 Arg Asn Gln Ser Asp Pro Thr Met Phe Asn Lys Ile Ala Val Phe Phe Gln *

2257
 CGT AAT CAG TCC GAT CCG ACG ATG TTT AAC AAG ATC GCG GTG TTT TTT CAG TAA CCGCTCAAG GCTTGAAACAA CATCGCGCTT ATCGTTAAGG
 Arg Asn Gln Ser Asp Pro Thr Met Phe Asn Lys Ile Ala Val Phe Phe Gln *

2267
 TAAGGCGCTA TTTTITTTTAC CCGCCAGGAC AAGACCATCA TCACCGTTGC CCTTATAGAC GATCACCTCA TCGTCGCGTC CGGCTTTGCG CAGCTG

Figure 3. Nucleotide sequence of the non-transcribed strand of the *ilvBN* operon of *E. coli* K-12.

Table 1. Amino acid composition of polypeptides encoded by *ilvBN*, *ilvGM*^a and *ilvH*^b as determined from nucleotide sequences.

Amino acid	<i>ilvB</i>	<i>ilvN</i>	Number of Residues ^c		<i>ilvI</i>	<i>ilvH</i>
			<i>ilvG</i>	<i>ilvM</i>		
Alanine	71 (12.6)	3 (3.1)	63 (11.5)	9 (10.3)	51 (9.0)	9 (5.6)
Arginine	24 (4.3)	5 (5.2)	20 (3.6)	7 (8.1)	26 (4.6)	13 (8.1)
Asparagine	19 (3.4)	7 (7.3)	17 (3.1)	6 (6.9)	17 (3.0)	7 (4.4)
Aspartic acid	29 (5.2)	8 (8.3)	33 (6.0)	3 (3.4)	27 (4.8)	3 (1.9)
Cysteine	9 (1.6)	2 (2.1)	9 (1.6)	2 (2.3)	10 (1.8)	0 (0.0)
Glutamic acid	30 (5.3)	4 (4.2)	25 (4.6)	3 (3.4)	22 (3.9)	12 (7.5)
Glutamine	34 (6.0)	8 (8.3)	34 (6.2)	7 (8.0)	31 (5.5)	8 (5.0)
Glycine	45 (8.0)	3 (3.1)	47 (8.6)	1 (1.1)	49 (8.7)	14 (8.8)
Histidine	14 (2.5)	4 (4.2)	18 (3.3)	4 (4.6)	18 (3.2)	2 (1.3)
Isoleucine	47 (8.4)	7 (7.3)	21 (3.8)	4 (4.6)	30 (5.3)	12 (7.5)
Leucine	47 (8.4)	9 (9.4)	60 (10.9)	7 (8.0)	50 (8.8)	18 (11.3)
Lysine	18 (3.2)	4 (4.2)	17 (3.1)	1 (1.1)	23 (4.1)	8 (5.0)
Methionine	21 (3.7)	4 (4.2)	20 (3.6)	4 (4.6)	23 (4.1)	3 (1.9)
Phenylalanine	17 (3.0)	5 (5.2)	18 (3.3)	3 (3.4)	12 (2.1)	3 (1.9)
Proline	37 (6.6)	3 (3.1)	31 (5.7)	2 (2.3)	28 (5.0)	3 (1.9)
Serine	23 (4.1)	4 (4.2)	22 (4.0)	8 (9.2)	32 (5.7)	14 (8.8)
Threonine	29 (5.2)	5 (5.2)	30 (5.5)	5 (5.5)	31 (5.5)	9 (5.6)
Tryptophan	4 (0.7)	1 (1.0)	7 (1.3)	0 (0.0)	12 (2.1)	0 (0.0)
Tyrosine	10 (1.8)	0 (0.0)	13 (2.4)	0 (0.0)	18 (3.2)	3 (1.9)
Valine	34 (6.0)	10 (10.4)	43 (7.8)	11 (12.6)	56 (9.9)	19 (11.9)
	562 (100)	96 (100)	548 (100)	87 (100)	566 (100)	160 (100)
Molecular weight	60,400	11,100	59,200	9,700	62,000	17,500

a. Sequence data from Lawther et al. (4).

b. Sequence data from Squires et al. (3).

c. Numbers in parentheses are percentages of total number of amino acids.

Table 2. Codon frequencies for polypeptides encoded by *ilvBN*, *ilvGM*^a and *ilvIH*^b as compared with codon usage in 25 *E. coli* genes^c.

Residue and Codon	Codon Frequencies ^d								
	<i>ilvBe</i> ^e	<i>ilvB</i>	<i>ilvN</i>	<i>ilvGe</i> ^f	<i>ilvG</i>	<i>ilvM</i>	<i>ilvI</i>	<i>ilvH</i>	<i>E. coli</i>
Phe UUU	0 (0)	10(59)	5(100)	0 (0)	8(44)	1(33)	6(50)	2(67)	104(44)
Phe UUC	0 (0)	7(41)	0 (0)	0 (0)	10(56)	2(67)	6(50)	1(33)	135(56)
Leu UUA	0 (0)	3 (6)	0 (0)	0 (0)	16(27)	4(57)	7(14)	7(39)	36 (6)
Leu UUG	0 (0)	12(25)	1(10)	0 (0)	10(17)	1(14)	14(28)	2(11)	51 (8)
Leu CUU	0 (0)	1 (2)	2(20)	2(50)*	4 (7)	0 (0)	7(14)	5(28)	54 (9)
Leu CUC	1(33)*	4 (9)	1(10)	0 (0)	4 (7)	0 (0)	4 (8)	1 (6)	41 (7)
Leu CUA	2(67)*	0 (0)	1(10)	1(25)*	1 (2)	0 (0)	1 (2)	0 (0)	11 (2)
Leu CUG	0 (0)	27(57)	5(50)	1(25)	25(42)	2(29)	17(34)	3(17)	432(69)
Ile AUU	0 (0)	21(45)	3(43)	4(80)	1 (5)	0 (0)	16(53)	6(50)	151(37)
Ile AUC	0 (0)	24(51)	3(43)	1(20)	19(90)	3(75)	12(40)	4(33)	252(62)
Ile AUA	0 (0)	2 (4)	1(14)	0 (0)	1 (5)	1(25)	2 (7)	2(17)	2 (1)
Met AUG	2 ---	21 ---	4 ---	1 ---	20 ---	4 ---	23 ---	3 ---	189 ---
Val GUU	5(56)	9(26)	3(30)	0 (0)	15(35)	3(27)	11(20)	4(21)	182(38)
Val GUC	4(44)*	6(18)	2(20)	1(17)	8(19)	4(36)	17(30)	5(26)	62(13)
Val GUA	0 (0)	1 (3)	2(20)	0 (0)	4 (9)	1 (9)	11(20)	2(11)	111(23)
Val GUG	0 (0)	18(53)	3(30)	5(83)	16(37)	3(37)	17(30)	8(42)	130(27)
Ser UCU	0 (0)	2 (9)	0 (0)	0 (0)	2 (9)	0 (0)	7(22)	1 (7)	86(27)
Ser UCC	2(100)	3(13)	1(25)	0 (0)	5(23)	0 (0)	7(22)	2(14)	83(26)
Ser UCA	0 (0)	2 (9)	0 (0)	0 (0)	4(18)	2(25)	2 (6)	2(14)	27 (8)
Ser UCG	0 (0)	4(17)	0 (0)	0 (0)	3(14)	2(25)	8(25)	4(29)	37(11)
Ser AGU	0 (0)	3(13)	0 (0)	0 (0)	3(14)	1(13)	4(13)	2(14)	21 (6)
Ser AGC	0 (0)	9(39)	3(75)	2(100)	5(23)	3(38)	4(13)	3(21)	70(22)
Pro CCU	0 (0)	5(14)	0 (0)	0 (0)	5(16)	0 (0)	2 (7)	0 (0)	24 (9)
Pro CCC	0 (0)	5(14)	0 (0)	0 (0)	2 (6)	0 (0)	7(25)	1(33)	16 (6)
Pro CCA	2(67)	5(14)	0 (0)	1(50)	5(16)	2(100)	8(29)	1(33)	53(20)
Pro CCG	1(33)	22(59)	3(100)	1(50)	19(61)	0 (0)	11(39)	1(33)	174(65)
Thr ACU	3(100)*	4(14)	1(20)	0 (0)	5(17)	0 (0)	9(29)	0 (0)	76(24)
Thr ACC	0 (0)	17(59)	2(40)	0 (0)	21(70)	3(60)	11(35)	6(67)	162(51)
Thr ACA	0 (0)	4(14)	1(20)	1(100)*	2 (7)	2(40)	0 (0)	2(22)	19 (6)
Thr ACG	0 (0)	4(14)	1(20)	0 (0)	2 (7)	0 (0)	11(35)	1(11)	63(20)
Ala GCU	0 (0)	6 (8)	1(33)	2(50)	9(14)	1(11)	7(14)	1(11)	202(28)
Ala GCC	1(20)	24(34)	1(33)	1(25)	14(22)	6(27)	10(20)	1(11)	136(19)
Ala GCA	2(40)	9(13)	0 (0)	1(25)	22(35)	2(22)	11(22)	3(33)	166(23)
Ala GCG	2(40)	32(45)	1(33)	0 (0)	18(29)	0 (0)	23(45)	4(44)	221(30)
Tyr UAU	0 (0)	5(50)	0 (0)	0 (0)	6(46)	0 (0)	11(61)	1(33)	69(41)
Tyr UAC	0 (0)	5(50)	0 (0)	0 (0)	7(54)	0 (0)	7(39)	2(67)	101(59)
Ter UAA	0 ---	1 ---	1 ---	0 ---	0 ---	0 ---	1 ---	1 ---	22(88)
Ter UAG	1 ---*	0 ---	0 ---	1 ---*	0 ---	0 ---	0 ---	0 ---	1 (4)
Ter UGA	0 ---	0 ---	0 ---	0 ---	1 ---	1 ---	0 ---	0 ---	2 (8)
His CAU	0 (0)	10(71)	3(75)	0 (0)	10(56)	2(50)	14(78)	1(50)	42(39)
His CAC	0 (0)	4(29)	1(25)	0 (0)	8(44)	2(50)	4(22)	1(50)	66(61)
Gln CAA	0 (0)	9(26)	2(25)	0 (0)	13(38)	4(57)	12(39)	3(38)	75(27)
Gln CAG	0 (0)	25(74)	6(75)	0 (0)	21(62)	3(43)	19(61)	5(63)	207(73)
Asn AAU	1(50)	9(47)	2(29)	0 (0)	6(35)	6(100)	10(59)	2(67)	57(24)
Asn AAC	1(50)	10(53)	5(71)	0 (0)	11(65)	0 (0)	7(41)	1(33)	179(76)

Table 2 cont.

Residue and Codon	<i>ilvBe</i>	<i>ilvB</i>	<i>ilvN</i>	<i>ilvGe</i>	<i>ilvG</i>	<i>ilvM</i>	<i>ilvI</i>	<i>ilvH</i>	<i>E. coli</i>
Lys AAA	1(100)	14(78)	2(50)	0 (0)	14(82)	1(100)	16(70)	6(75)	296(77)
Lys AAG	0 (0)	4(22)	2(50)	1(100)	3(18)	0 (0)	7(30)	2(25)	90(23)
Asp GAU	0 (0)	17(59)	3(38)	0 (0)	18(55)	1(33)	18(67)	5(71)	175(51)
Asp GAC	0 (0)	12(41)	5(63)	0 (0)	15(45)	2(67)	9(33)	2(29)	168(49)
Glu GAA	0 (0)	21(70)	2(50)	0 (0)	17(68)	3(100)	13(59)	8(67)	328(73)
Glu GAG	0 (0)	9(30)	2(50)	0 (0)	8(32)	0 (0)	9(41)	4(33)	119(27)
Cys UGU	0 (0)	7(78)	2(100)	0 (0)	5(56)	0 (0)	2(20)	0 (0)	21(42)
Cys UGC	0 (0)	2(22)	0 (0)	1(100)	4(44)	2(100)	8(80)	0 (0)	29(58)
Trp UGG	0 ---	4 ---	1 ---	0 ---	7 ---	0 ---	12 ---	0 ---	48 ---
Arg CGU	1(100)	8(33)	2(40)	0 (0)	8(40)	3(43)	18(31)	5(38)	201(58)
Arg CGC	0 (0)	12(50)	3(60)	0 (0)	4(20)	3(43)	9(35)	6(46)	121(35)
Arg CGA	0 (0)	0 (0)	0 (0)	2(100)*	4(20)	0 (0)	5(19)	0 (0)	8 (2)
Arg CGG	0 (0)	3(13)	0 (0)	0 (0)	4(20)	1(14)	1 (4)	2(15)	11 (3)
Arg AGA	0 (0)	1 (4)	0 (0)	0 (0)	0 (0)	0 (0)	1 (4)	0 (0)	4 (1)
Arg AGG	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	2 (8)	0 (0)	1 (0.25)
Gly GGU	0 (0)	12(27)	0 (0)	0 (0)	19(40)	1(100)	13(27)	3(21)	231(48)
Gly GGC	1(100)	26(58)	2(100)	1(33)	21(45)	0 (0)	14(29)	7(50)	197(41)
Gly GGA	0 (0)	3 (7)	0 (0)	2(67)*	1 (2)	0 (0)	7(14)	1 (7)	22 (5)
Gly GGG	0 (0)	4 (9)	0 (0)	0 (0)	6(13)	0 (0)	15(31)	3(21)	33 (7)
Total	32	562	96	32	548	87	566	160	6,503

a. Sequence data from Lawther et al. (4).

b. Sequence data from Squires et al. (3).

c. *E. coli* codon usage compilation from Konigsberg and Godson (20).

d. Numbers represent times codon used in respective gene. Numbers in parentheses represent the percentage of codon usage for respective amino acid.

e. *ilvB* leader polypeptide.f. *ilvG* leader polypeptide.

products were inducible upon addition of cAMP to the minicell system which is consistent with both genes being under the catabolite repressed *ilvB* promoter (9). Analysis of the *ilvBN* operon nucleotide sequence suggests two coding regions for polypeptides with molecular weights of 60,400 and 11,100 daltons. We define these regions between position 236 to 1924 as *ilvB* and between position 1928 to 2218 as *ilvN*.

It is now clear that each of the three AHAS isozymes consists of large and small subunits. In addition to AHAS I, which is encoded by the *ilvBN* operon, the *ilvGM* genes express AHAS II and the *ilvIH* operon codes for AHAS III. In table 1 the amino acid composition of each subunit as deduced from their nucleotide sequences is listed. The selection of the *ilvG* and *ilvM* initiation codons is based on their homology to the amino terminal sequence of these gene products in *S. typhimurium* LT2 (Kutney and Schloss, Fed. Am. Abstr. 1984, vol. 43, no. 2630). The *ilvI* initiation codon is based on the amino terminal sequence of its purified product (19).

Figure 4. Alignment of the amino acid sequences of the large and small subunits of the AHAS isozymes in *E. coli* K-12. The sequences were aligned manually with a minimum number of gaps to maximize the sequence homology. Boxes are drawn around amino acid residues in common between at least two isozyme subunits. The *ilvH* encoded sequence is shown up to residue 94, thereafter no homologies were detected.

The codon usage for each AHAS subunit and the *ilvB* and *ilvG* leader polypeptide, labeled *ilvBe* and *ilvGe*, respectively, is listed in table 2. The codon frequencies of the AHAS subunits reflect closely those found in a compilation of 25 *E. coli* K-12 genes (20). However, the *ilvB* and *ilvG* leader polypeptides show a high frequency of rare codon usage. Examples of rare codon usage are indicated by asterisks. The *ilvB* and

Figure 5. Graphic matrix plots of pairwise comparisons of the *ilvBN*, *ilvGM* and *ilvIH* nucleotide sequences. Figure 5A is the *ilvGM* sequence plotted against the *ilvBN* sequence as described by Maizel and Lenk (22). Each dot represents a minimum of 12 nucleotides matched out of 20 nucleotides compared. The position of the dots correspond to the center of the 20 nucleotides compared from each sequence. Figure 5B and C are similar matrix plots of *ilvIH* vs. *ilvBN* and *ilvIH* vs. *ilvGM*, respectively.

ilvG attenuators are regulated by leucine limitation and both leader polypeptides show rare leucine codon usage. This is also true of the *leu* operon leader polypeptide in *S. typhimurium* where four adjacent leucine residues use three CUA and one CUC codons

Figure 6. Hydropathy profile of the large AHAS subunits encoded by the *ilvB*, *ilvG* and *ilvI* genes. A version of the SOAP program by Kyte and Doolittle (23) was used to determine the mean hydropathy of a moving segment of seven amino acid residues. The values are plotted at the midpoint of each segment. Relatively hydrophobic residue segments appear above the midline. Regions 1, 2 and 3 correspond to sequences in the large AHAS subunits which share close homology.

(21). This higher frequency of rare codon usage in leader polypeptides might reflect a more sensitive regulation of an attenuator mechanism to an amino acid limitation.

The three AHAS isozymes share close amino acid sequence homology. Figure 4 shows the large and small subunits aligned with a minimum number of gaps as to maximize their sequence homology. Boxes are drawn around amino acid residues which are identical between at least two of the isozyme sequences. Previously, a comparison of the *ilvG* and *ilvI* gene products discerned three regions of homology (3). These three regions of homology are also identified in the *ilvB* gene product. Gaps in the sequence alignment and weaker sequence homology are seen outside those regions. In addition to these regions there is conserved homology between the carboxyl termini of the *ilvB* and

Figure 7. Hydropathy profile of the small AHAS subunits encoded by the ilvN, ilvM and ilvH genes.

ilvG gene products that is not shared with the ilvI product. The smaller AHAS subunits encoded by ilvN, ilvM and ilvH also share sequence homology.

The three regions of homology in the AHAS large subunits are illustrated more clearly in two-dimensional matrix plots. Figure 5A shows a matrix plot as described by Maizel and Lenk (22) comparing the ilvBN nucleotide sequence with ilvGM. Each dot represents a minimum of 12 nucleotides matched out of 20 nucleotides compared. Three distinct regions of nucleotide homology can be readily discerned. The numbering of the ilvBN sequence corresponds to that in Figure 3. The ilvGM numbering is based on the proposed initiation site of Lawther *et al.* (4). Similar regions of homology are shown in matrix plots of ilvIH versus ilvBN and ilvIH versus ilvGM in Figure 5B and 5C, respectively. These matrix plots were also performed with the stringency explained above. The ilvIH numbering is based on that of Squires *et al.* (3) which starts at a HincII site upstream of the coding region.

Another means of comparing the AHAS subunits was made by evaluating their hydropathy profile. Kyte and Doolittle (23) have proposed an amino acid hydropathy

Figure 8. The percentage of identical amino acid residues between the *ilvB*, *ilvG* and *ilvI* encoded large subunits.

scale with values ranging from -4.5 for the most hydrophilic residue, arginine, to +4.5 for the most hydrophobic residue, isoleucine. A version of their SOAP program was used to determine the mean hydropathy of a moving segment of seven amino acid residues. The values are plotted at the midpoint of each segment, proceeding from the amino terminus on the left to the carboxy terminus on the right. The hydropathy profiles of the *ilvB*, *ilvG* and *ilvI* gene products are illustrated in Figure 6. The three regions of homology show a similar profile as would be expected from their close amino acid sequence homology. Again the *ilvB* and *ilvG* carboxy termini, which are distinct from the *ilvI* terminus, show similar hydropathy profiles. The hydropathy profile of the small AHAS subunits is shown in Figure 7.

To quantitate the relatedness between the AHAS subunits the percentage of amino acid sequence homology was determined. The polypeptide sequences aligned in Figure 4 were compared only at amino acid residues where all three sequences are represented. That is to say, only amino acid residues which exhibit no gap in all three polypeptides were scored. Figure 8 lists the percentage of identical amino acid residues shared between the *ilvB*, *ilvG* and *ilvI* gene products. The products encoded by *ilvB* and *ilvG* share the closest homology, while *ilvB* and *ilvI* products are the most distant. The small AHAS subunits exhibit weaker sequence homology. A comparison of 86 amino acid residues aligned in all three polypeptides as shown in Figure 4 gives the following results: *ilvN* and *ilvM* gene products share identical residues at 21 positions (24%);

Figure 9. The proposed evolution of the *ilvBN*, *ilvGM* and *ilvIH* operons in *E. coli*.

products encoded by ilvM and ilvH are also identical at 21 residues (24%); and the ilvN and ilvH gene products are conserved at 26 residues (30%).

From an evolutionary perspective the differences in the conserved homology between the large AHAS polypeptides can be explained by the model illustrated in Figure 9. An ancestral AHAS consisted of a large polypeptide of approximately 60,000 molecular weight and a smaller molecular weight polypeptide. The polypeptides were encoded by two genes which were co-transcribed with the large polypeptide coding at the 5'-end. The ancestral AHAS genes were duplicated into genes XX' and YY'. The prime designating the smaller gene. Genes XX' evolved into the ilvIH operon. The other genes YY' underwent a second duplication event leading to two copies of the genes. These two copies diverged into ilvBN and ilvGM, respectively.

This model points to a greater homology between ilvBN and ilvGM. The ilvB and ilvG genes share the most amino acid residues and as illustrated in the matrix plots (Fig. 5), a greater nucleotide sequence homology. The ilvN and ilvM encoded products are comparable in size with molecular weights of 11,100 and 9,700, respectively, whereas the ilvH coded polypeptide has a molecular weight of 17,500. AHAS I and AHAS II both require flavin adenine dinucleotide (FAD) as a cofactor while it has not been possible to show such a requirement for AHAS III (24). From a regulation standpoint, both the ilvBN and ilvGMEDA operons are under control by attenuation (8,14,16,25,26,27). The ilvB and ilvG leader polypeptides deduced from DNA sequences are both 32 amino acids long and both attenuators are regulated by leucine and valine limitation. Alternatively, the ilvIH operon is repressed by leucine by a mechanism which appears to be other than attenuation (28). And finally, the transcriptional expression of the ilvBN and ilvGMEDA operons is positively effected by integration host factor encoded by himA and himD, whereas the ilvIH operon shows no such property (29).

Currently, the functional role of the small AHAS subunits is not understood. By genetic analysis the ilvH subunit appears to confer valine sensitivity upon the AHAS III isozyme (2). It has not yet been demonstrated whether the ilvN encoded small subunit plays a similar role in the valine sensitivity of AHAS I. One surprising point concerning the AHAS II small subunit is that this ilvM encoded product in E. coli K-12 and S. typhimurium LT2 share weak homology. Kutney and Schloss (Fed. Am. Abstr. 1984, vol. 43, no. 2630) found that of 35 amino terminal residues sequenced in S. typhimurium, 12 were different in a comparison with the predicted E. coli K-12 ilvM polypeptide sequence. This is in contrast to the amino terminal comparison of the ilvG encoded polypeptide where one substitution was found out of 34 residues sequenced. This weak conservation of ilvM sequence between two closely related enteric bacteria suggests that this polypeptide does not have an important catalytic role.

ACKNOWLEDGEMENTS

We thank Thomas Menees and Elaine Ito for technical support. We are also grateful to Jann Sharp for discussions on codon usage, Toai Nguyen for assistance with computer programming and Denis Heck and Gregg Duester for helpful discussions. This work was supported in part by a grant (GM 24330) from the National Institutes of Health.

REFERENCES

1. Guardioli, J., DeFelice, M., Lamberti, A. and Iaccarino, M. (1977) *Mol. Gen. Genet.* 156,17-25.
2. DeFelice, M., Guardioli, J., Esposito, B. and Iaccarino, M. (1974) *J. Bacteriol.* 120, 1068-1077.
3. Squires, C. H., DeFelice, M., Devereux, J. and Calvo, J. (1983) *Nucleic Acids Res.* 11, 5299-5313.
4. Lawther, R. L., Calhoun, D. H., Adams, C. W., Hauser, C. A., Gray, J. and Hatfield, G. (1981) *Proc. Natl. Acad. Sci. USA* 78, 922-925.
5. Griminger, H. and Umbarger, H. E. (1979) *J. Bacteriol.* 137, 846-853.
6. Eoyang, L. and Silverman, P. M. (1984) *J. Bacteriol.* 157, 184-189.
7. Newman, T., Friden, P., Sutton, A. and Freundlich, M. (1982) *Mol. Gen. Genet.* 186, 378-384.
8. Hauser, C. A. and Hatfield, G. W. (1984) *Nucleic Acids Res.* 11, 127-134.
9. Sutton, A. and Freundlich, M. (1980) *Mol. Gen. Genet.* 178, 179-183.
10. Rothstein, S. J., Jorgensen, R. A., Postle, K. and Reznikoff, W. S. (1980) *Cell*, 19, 795-805.
11. Roozen, D., Fenwick, R. and Curtiss, R. (1971) *J. Bacteriol.* 107, 21-33.
12. Miller, J. H. (1972) *Experiments in molecular genetics.* Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
13. Laemmli, U. K. (1970) *Nature* 227, 680-685.
14. Friden, P., Newman, T. and Freundlich, M. (1982) *Proc. Natl. Acad. Sci. USA* 79, 6156-6160.
15. Sanger, F., Nicklen, S. and Coulson, A. (1977) *Proc. Natl. Acad. Sci. USA* 74, 5463-5467.
16. Hauser, C. A. and Hatfield, G. W. (1984) *Proc. Natl. Acad. Sci. USA* 81, 76-79.
17. Friden, P., Tsui, P., Okamoto, K. and Freundlich, M. (1984) *Nucleic Acids Res.* 12, 8145-8160.
18. Rosenberg, M. and Court, D. (1979) *Ann. Rev. Genet.* 13, 319-353.
19. Lago, C. T., Sannia, G., Marino, G., Squires, C. H., Calvo, J. M. and DeFelice, M. (1985) *Biochem. Biophys. Acta* 824, 74-79.
20. Konigsberg, W. and Godson, G. N. (1983) *Proc. Natl. Acad. Sci. USA* 80, 687-691.
21. Gemmil, R. M., Wessler, S. R., Keller, E. B. and Calvo, J. M. (1979) *Proc. Natl. Acad. Sci. USA* 76, 4941-4945.
22. Maizel, J. V. and Lenk, R. P. (1981) *Proc. Natl. Acad. Sci. USA* 78, 7665-7669.
23. Kyte, J. and Doolittle, R. F. (1982) *J. Mol. Biol.* 157, 105-132.
24. DeFelice, M., Levinthal, M., Iaccarino, M. and Guaroli, J. (1979) *Microbiol. Rev.* 43, 42-58.
25. Lawther, R. P. and Hatfield, G. W. (1980) *Proc. Natl. Acad. Sci. USA* 77, 1862-1866.
26. Nargang, F. E., Subrahmanyam, C. S. and Umbarger, H. E. (1980) *Proc. Natl. Acad. Sci. USA* 77, 1823-1827.
27. Adams, C. W. and Hatfield, G. W. (1985) *J. Biol. Chem.*, in press.
28. DeFelice, M. and Levinthal, M. (1977) *Biochem. Biophys. Res. Commun.* 79, 82-87.
29. Friden, P., Voelkel, K., Sternglanz, R. and Freundlich, M. (1984) *J. Mol. Biol.* 172, 573-579.