Sequence for human argininosuccinate synthetase cDNA

Hans-Georg O.Bock*, Tsung-Sheng Su, William E.O'Brien, and Arthur L.Beaudet

Department of Pediatrics and Cell Biology, Baylor College of Medicine, Houston, TX 77030, USA

Received 10 June 1983; Revised and Accepted 31 August 1983

ABSTRACT

The nucleotide sequence for human argininosuccinate synthetase cDNA was determined by analysis of six clones isolated from a single experiment. The sequence covered 1623 nucleotides including 76 bases of poly(A) and contained a 1236 nucleotide open reading frame encoding a protein of 46,434 daltons. In one cDNA isolate, a cloning artifact or perhaps RNA polymerase error involving addition of an A in a region of six A's within the coding sequence was documented. Single base variations in the 3' untranslated region were examined in detail since detection of DNA polymorphisms in the cDNAs could imply over-expression of both alleles at the active locus in canavanineresistant cells, i.e. a trans-acting mechanism for enzyme overproduction. However, the sequence from five cDNAs suggested some single base artifacts, and DNA polymorphism remains uncertain. The occurrence of three tandem arginine codons in the 5' untranslated region of the cDNA suggested the possibility of an interaction of arginyl-tRNA with mRNA to regulate RNA processing or half-life as a mechanism for arginine-mediated repression.

INTRODUCTION

Argininosuccinate synthetase functions in the urea cycle to convert citrulline and aspartate to argininosuccinate. A deficiency of this enzymic activity in humans is recognized as the disorder citrullinemia, which is inherited in an autosomal recessive manner (1). In some cultured cell lines, argininosuccinate synthetase is subject to metabolite regulation by arginine (2). Variant cultured human cell lines resistant to the arginine analog canavanine have an argininosuccinate synthetase specific activity which is 180-fold higher than that of parental cells. This canavanine-resistant phenotype is stable and the enzymic activity is not subject to an arginine-mediated regulation (3,4). The steady state level of this mRNA correlates closely with the specific activity of the enzyme in the cell extract, indicating regulation at a pretranslational level (4,5). There is no evidence for gene amplification in canavanine-resistant cells (5). There are multiple processed argininosuccinate synthetase pseudogenes (6). However, there is abundant evidence that none of these pseudogenes is expressed, and all of the available data are consistent with the presence of a single active gene (6-8).

We report here the DNA sequence of several cDNA clones for human argininosuccinate synthetase obtained from a single cloning experiment. The sequence provides data relevant to possible mechanisms of canavanine resistance and arginine-mediated regulation.

METHODS

The isolation of the cDNA clones labeled pAS1, pAS2, and pAS3 from a recombinant cDNA library constructed from pBR322 and poly(A)⁺ RNA from canavanine-resistant RPMI-2650 cells was described previously (5). The cDNA clones pAS4, pAS9 and pAS12 were isolated by rescreening that library with nick-translated pAS1. All isolates were from a single cloning experiment using mRNA from one cell harvest.

Sequence Analysis

Restriction maps of all isolates were prepared. Restriction fragments were isolated, labeled at 5' ends, and sequenced by the method of Maxam and Gilbert (9).

RESULTS

Sequencing Strategy and Nucleic Acid Sequence

A total sequence of 1623 nucleotides, including 76 bases of poly(A), was obtained from human argininosuccinate synthetase cDNAs of clones pAS1, pAS2, pAS3, pAS4, pAS9, and pAS12 (Figures 1 and 2). Seventy-five bases 5' to the first ATG sequence were determined. Beginning with the most 5' ATG, there was an open reading frame of 1236 nucleotides, which encoded a protein with a molecular weight of 46,434 daltons. The amino acid content and molecular weight predicted for argininosuccinate synthetase by this nucleotide sequence agreed very well with values determined for the bovine liver enzyme (10).

The 3' untranslated sequence contained 233 nucleotides between the TAG terminator codon and the poly(A) sequence. Potential poly(A) recognition sequences were present at nucleotides 1526-1532 (ATAAAAA) and nucleotides 1537-1544 (AATTAAAA). Each is a variation of the canonical eukaryotic sequence, AATAAA, which typically is located 11-30 nucleotides 5' to the poly(A) addition site (11,12). The proximity of the poly(A) sequence to the AATTAAAA sequence suggested that bases 1526-1532 (ATAAAAA) may represent the actual poly(A) recognition sequence.

The codon usage was nonrandom. Whereas the coding region (bases 76-



Figure 1. The sequencing strategy for human argininosuccinate synthetase cDNAs. The uppermost line provides a restriction map for the entire cDNA sequence. The sequence numbering proceeds in a 5' to 3' direction. Shown below and relative to this restriction map are the six cDNA inserts studied, each presented together with its own sequencing strategy. The solid, heavy lines represent protein coding regions, while the open, heavy lines represent untranslated regions. The cDNA fragments were labeled at the 5' ends, and the direction and extent of each sequence is indicated directly below its respective cDNA map.

1311) had an overall G + C content of 56%, the G + C contents of positions 1, 2, and 3 of these codons were 57%, 37% and 75%, respectively. Relatively few codons contained the dinucleotide CG, similar to other eukaryotic DNA sequences.

Sequence Variations

Our efforts were placed originally on sequencing pAS1, the longest clone available. Upon completion of this sequencing, we were unable to identify a reading frame to encode a protein of the expected size. Analysis of additional cDNAs revealed that pAS1 contained a single base (A) insertion to give seven A's at a site where the sequences of both pAS3 and pAS12 contained only 6 tandem A's (Figures 2 and 3, nucleotides 757-762). Deletion of one of the seven A's at this position in pAS1 provided an open reading frame of the expected size. Additional evidence that only six A's occurred in this region was available from the sequence of the exon containing this region from a genomic clone and from the sequence of three processed pseudogenes (H.G.

			-				MSSKG
CGAGCCCGAGTGGTT 10	CACTGCACT	GTGAAAACAGA 30	TTCCAGACGC 40	CGGGGAACTC 50	ACGCCTCCAAT 60	CCCAGACGCT 70	80 90
S V V T. A	v s c	GLDT	SCI	LVW	LKEQ	GYD	V I A Y L
TCCGTGGTTCTGGCC	ACAGTOCO	GCCTGGACAC	CTCGTGCATC	CTCGTGTGG	CTGAAGGAACA	AGGCTATGAC	GTCATTGCCTATCTG
100	110	120	130	140	150	100	1/0 100
A N I G Q	K E D	F E E A		A L K GCACTGAAG	L G A K	K V F AAAGGTGTTC	I E D V S ATTGAGGATGTCAGC
190	200	210	220	230	240	250	260 270
REFVE	EFI	WPAI	Q S S	ALY	EDRY	LLG	TSLAR
AGGGAGTTTGTGGAG	GAGTTCATC 200	TGGCCGGCCAT	CCAGTCCAGC 310	GCACTGTAT 320	GAGGACCGCTA 330	CCTCCTGGGC 340	ACCTCTCTTGCCAGG 350 360
200	230						
P C I A R CCCTGCATCGCCCGC	K Q V AAACAAGTG	E I A Q GAAATCGCCCA	R E G GCGGGGAGGGG	A K Y GCCAAGTAT	V S H G GTGTCCCACGG	GCGCCACAGGA	AAGGGGAACGATCAG
370	380	390	400	410	420	430	440 450
VRFEL	s c v	SLAP	QIK	V I A	PWRM	PEF	YNRFK
GTCCGGTTTGAGCTC. 460	AGCTGCTAC 470	480	490	GTCATTGCT 500	510	520	530 540
GRNDL	MRY		GTP	TPV	TPEN	PWS	MDENL
GCCCCCAATGACCTG	ATGGAGTAC	GCAAAGCAACA	COCCATTCCC	ATCCCGGTC	ACTCCCAAGAA	CCCGTGGAGC	ATGGATGAGAACCTC
550	560	570	580	590	600	610	620 630
M H I S Y		I L E N			P G L Y		
640	650	660	670	680	690	700	710 720
A P N T P	DIL	EIEF	K K G	V P V	KVTN		сттно
GCCCCCAACACCCCT	GACATTCTC	GAGATCGAGTT	CAAAAAAGGG	GTCCCTGTG.	AAGGTGACCAA	CGTCAAGGAT	GGCACCACCCACCAG
730	740	750	760	770	780	790	800 810
730 ••••••••••••••••••••••••••••••••••••	740 P M V	750	760	770	780	790	800 810
730 T S L E L ACCTCCTTGGAGCTC	740 F M Y FTCATGTAC	750 L N E V CTGAACGAAGT	760 A G K CGCGGGCAAG	770 H G V CATGGCGTG	780 G R I D GGCCGTATTGA	790 IVE CATCGTGGAG	800 810 N R F I G AACCGCTTCATTGGA
730 T S L E L Acctccttggagete: 820	740 FMY TTCATGTAC 830	750 LNEV CTGAACGAAGT(840	760 A G K CGCGGGCAAG 850	770 H G V CATGGCGTG 860	780 G R I D GGCCGTATTGA 870	790 IVE CATCGTGGAG. 880	800 810 N R F I G AACCGCTTCATTGGA 890 900
730 T S L E L ACCTCCTTGGAGGTC 820 M K S R G	740 FMY TTCATGTAC 830 IYE	750 L N E V CTGAACGAAGT(840 T P A G	760 A G K CGCCGCGCAAG 850 T I L	770 H G V CATGGCGTG 860 Y H A	780 G R I D GGCCGTATTGA 870 H L D I	790 IVE CATCGTGGAG 880 EAF	800 810 N R F I G AACCGCTTCATTGGA 890 900 T M D R E
730 T S L E L ACCTCCTTGGAGCTC 820 M K S R G ATGAAGTCCCGAGGT, 910	740 F M Y FTCATGTAC 830 I Y E Atctacgag, 920	750 L N E V CTGAACGAAGT(840 T P A G Acccccagcagg(930	760 A G K CCCCCCCCAAG 850 T I L CACCATCCTT 940	770 H G V CATGGCGTG 860 Y H A TACCATGCT 950	780 G R I D GGCCGTATTGA 870 H L D I CATTTAGACAT 960	790 I V E CATCGTGGAG. 880 E A F CGAGGCCTTC. 970	800 810 N R F I G AACCGCTTCATTGGA 890 900 T M D R E ACCATGGACCGGGAA 980 990
730 T S L E L ACCTCCTTGGAGGCTC 820 M K S R G ATGAAGTCCCGAGGT. 910 V R K I K	740 F M Y TTCATGTAC(830 I Y E ATCTACGAG, 920 O G L	750 LNEV CTGAACGAAGT(840 TPAG ACCCCAGCAGG(930 GLKF	760 A G K CCCCGCGCCAAG 850 T I L CACCATCCTT 940 A E L	770 H G V CATGCCGTG 860 Y H A TACCATGCT 950 V Y T	780 GRID GGCCGTATTGA 870 HLDI CATTTAGACAT 960 GLRP	790 I V E CATCGTGGAG. 880 E A F CGAGGCCTTC. 970	800 810 N R F I G AACCGCTTCATTGGA 890 900 T M D R E ACCATGGACCGGGAA 980 990
730 T S L E L ACCTCCTTGGAGCTC 820 M K S R G ATGAAGTCCCGAGGT, 910 V R K I K GTGCGCGAAAATCAAA	740 F M Y TTCATGTAC(830 I Y E ATCTACGAG. 920 Q G L CAAGGCCTG	750 L N E V CTGAACGAAGT 840 T P A G ACCCCAGCAGGA 930 G L K F GCCTTGAAATT	760 A G K CCCCGCGCAAG 850 T I L CACCATCCTT 940 A E L TCCTGACCTG	770 H G V CATGCCGTG 860 Y H A TACCATGCT 950 V Y T GTGTATACCO	780 G R I D GGCCGTATTGA 870 H L D I CATTTAGACAT 960 G L R P GGTTTACGGCC	790 I V E CATCGTGGAG. 880 E A F CGACGCCTTC. 970 S P E TAGCCCTGAG	800 810 N R F I G AACCGCTTCATTGGA 890 900 T M D R E AACCATGGACCGGGAA 980 990 C E F V R TGGGAATTTGTCCGC
730 T S L E L ACCTCCTTGGAGCTC: 820 M K S R G ATGAAGTCCCGAGGT, 910 V R K I K GTGCGCGAAAATCAAAA 1000	740 F M Y TTCATGTAC: 830 I Y E ATCTACGAG. 920 Q G L CAAGGCCTG4 1010	750 L N E V CTGAACGAAGT 840 T P A G ACCCCAGCAGGA 930 G L K F GGCTTGAAATT 1020	760 A G K CCCCCCCCCAAG 850 T I L CACCATCCTT 940 A E L TGCTGAGCTG 1030	770 H G V CATEGCEGTG 860 Y H A TACCATECT 950 V Y T GTGTATACC 1040	780 G R I D GGCCGTATTGA 870 H L D I CATTTAGACAT 960 G L R P GGTTTACGGCC 1050	790 I V E CATCGTGGGAG. 880 E A F CGACGCCTTC. 970 S P E TACCCCTGAG 1060	800 810 N R F I G AACCGCTTCATTGGA 890 900 T M D R E AACCATGGACCGGGAA 980 990 C E F V R TGTGAATTTGTCCGC 1070 1080
730 T S L E L ACCTCCTTGGAGCTC 820 M K S R G ATGAAGTCCCGAGGT. 910 V R K I K GTGCGCAAAGTCAAAI 1000 H C I A K CACTGCATCGCCAAGG	740 FMY FTCATGTAC 830 IYE ATCTACGAG. 920 QGL CAAGGCCTG4 1010 SQE FCCCCAGGAG4	750 L N E V CTGAACGAAGT 840 T P A G ACCCCAGCAGGA 930 G L K P EGCTTGAAATT 1020 R V E G	760 A G K CGCGGGCAAG 850 T I L CACCATCCTT 940 A E L TGCTGAACCTG 1030 K V Q	770 H G V CATCGCCGTG 860 Y H A TACCATGCT 950 V Y T GTGTATACCO 1040 V S V GTGTCCGTCC	780 G R I D GGCCGTATTGA 870 H L D I CATTTAGACAT 960 G L R P GGTTTAGCGCC 1050 L K G Q L K G Q	790 I V E CATCGTGGGG, 880 E A F CGAGGCCTTC. 970 S P E TAGCCCTGAG 1060 V I I	800 810 N R F I G AACCGCTTCATTGGA 890 900 T M D R E ACCATGGACCGGGGAA 980 990 C E F V R TGTGGAATTTGTCCGG 1070 1080 L G R E S CTCTCCCCCCACCTCAC
730 T S L E L ACCTCCTTGGAGCTC 820 M K S R G ATGAAGTCCCGAGGT. 910 V R K I K GTGCGCAAAATCAAAA 1000 H C I A K CACTGCATCGCCAAGT 1090	740 F M Y TTCATGTAC 830 I Y E ATCTACGAC 920 Q G L CAAGGCCTGG 1010 S Q E TCCCAGGAGG	750 L N E V CTGAACGAAGTO 840 T P A G ACCCCAGCAGCA 930 G L K F GGCTTGAAATT 1020 R V E G CCAGTCGAACCO 1110	760 A G K CCCCCCCCCAAG 850 T I L CACCATCCTT 940 A E L TGCTGAGCTG 1030 K V Q SAAAGTCCAG 1120	770 H G V CATGECGTG 860 Y H A TACCATGCT 950 V Y T GTGTATACCA 1040 V S V GTGTCCGTCC 1130	780 G R I D GGCCGTATTGA 870 H L D I CATTTAGACAT 960 G L R P GGTTTAGGCCC 1050 L K G Q LICAAGGGCCA 1140	790 I V E CATCGTGGAG. 880 E A F CGAGGCCTTC. 970 S P E TAGCCCTGAG 1060 V Y I GGTGTACATCC 1150	800 810 N R F I G AACCGCTTCATTGGA 890 900 T M D R E ACCATGGACCGGGAA 980 990 C E F V R TGTGAATTTGTCCGC 1070 1080 L G R E S CTCCCGCCGGGAGTCC 1160 1170
730 T S L E L ACCTCCTTGGAGCTC 820 M K S R G ATGAAGTCCGGAGGT. 910 V R K I K GTGCGCAAAATCAAA 1000 H C I A K CACTGCATCGCCAAGC 1090 P L S L Y	740 F M T TTCATGTAC: 830 I Y E ATCTACCAG. 920 Q G L CAAGGCCTGG 1010 S Q E TCCCAGGAG: 1100 N E E	750 L N E V CTGAACGAAGT 840 T P A G ACCCCAGCAGCA 930 G L K F G C K F 1020 R V E G CCAGTGGAAGGG 1110 L V S M	760 A G K CCCCCCCCCAAG 850 T I L CACCATCCTT 940 A E L TGCTCACCT 1030 K V Q CAAAGTCCAG 1120 N V Q	770 H C V CATGCCGTG 860 Y H A TACCATGCT 950 V Y T GTGTATACC 1040 V S V GTGTCCCTCC 1130 G D Y	780 G R I D GGCCGTATTGA 870 H L D I CATTTAGACAT 960 G L R P GGTTTACGCC 1050 L K G Q LTCAAGGGCCA 1140 E P T D	790 I V E CATCGTGGAG. 880 E A F CGACGCCTTC. 970 S P E TAGCCCTGAG 1060 V Y I GGTGTACATC: 1150	800 810 N R F I G AACCGCTTCATTGGA 890 900 T M D R E ACCATGGACCGGGAA 980 990 C E F V R TGTGAATTTGTCCGC 1070 1080 L G R E S CTCGGCCGCGGAATTCC 1160 1170 F I N I N
730 T S L E L ACCTCCTTGCAGCTC 820 M K S R G ATGAAGTCCCGAGGT. 910 V R K I K GTGCCCAAAATCAAAA 1000 H C I A K CACTGCATCGCCAAGC 1090 P L S L Y CCACTGTCTCTCTACL 1180	740 F M T TTCATGTAC 830 I Y E ATCTACGAG 920 Q G L CAAGGCCTG 1010 S Q E FCCCAGGAG4 1100 N E E ATCAGCAGC	750 L N E V CTGAACGAAGT 840 T P A G ACCCCAGCAGCAGG 930 G L K F GGCTTGAAATT 1020 R V E G CGAGTGGAAGGA 1110 L V S M TTGGTGAGCATT 1200	760 A G K CCCCGCGCCAAG 850 T I L CACCATCCTT 940 A E L CCTGAGCTG 1030 K V Q GAAAGTGCAG 1120 N V Q GAACGTGCCAG 1210	770 H C V CATGCCGTG 860 Y H A TACCATCCT 950 V Y T GTGTATACC 1040 V S V GTGTCCCGTCC 1130 G D Y GGTATTATT 1220	780 G R I D GGCCGTATTGA 870 H L D I CATTTAGACAT 960 G L R P GGTTTAGGCCC 1050 L K G Q LICAAGGGCCAA 1140 E P T D GAGCCAACTGA 1230	790 I V E CATCGTGGAG. 880 E A F CGAGGCCTTC. 970 S P E TAGCCTGAG 1060 V Y I GGTGTACATC. 1150 A T G TGCCACCGGG	800 810 N R F I G AACCGCTTCATTGGA 890 900 T M D R E ACCATGGACCGGGAA 980
730 T S L E L ACCTCCTTGGAGGTC 820 M K S R G ATGAAGTCCCGAGGT. 910 V R K I K GTGCGCAAAATCAAAA 1000 H C I A K CACTGCATCGCCAAGT 1090 P L S L Y CCACTGTCTCTCTACL 1180	740 F M T TTCATGTAC 830 I Y E ATCTACGAG 920 Q G L CAAGGCCTG 1010 S Q E FCCCAGGAGG 1100 N E E AATGAGGAGG	750 L N E V CTGAACGAAGT 840 T P A G ACCCCAGCAGGA 930 G L K F GCCTTGAAATT 1020 R V E G CGAGTGGAAGGG 1110 L V S M CTGGTGAGCATC 1200	760 A G K CCCCGCGCCAAG 850 T I L CACCATCCTT 940 A E L TCCTGACCTG 1030 K V Q GAAAGTGCAG 1120 N V Q GAACGTCCAG 1210	770 H C V CATGCCGTG 860 Y H A TACCATGCT 950 V Y T GTGTATACC 1040 V S V GTGTCCGTCC 1130 C D Y GGTGATTATT 1220	780 G R I D GCCCTATTGA 870 H L D I CATTTAGACAT 960 G L R P GOTTTACGCCC 1050 L K G Q LTCAAGGGCCAA 1140 E P T D GAGCCAACTGA 1230	790 I V E CATCGTGGAG. 880 E A F CGAGGCCTTC. 970 S P E TAGCCCTGAG 1060 V Y I GGTGTACATC 1150 A T G TGCCACCGGG 1240	800 810 N R F I G AACCGCTTCATTGGA 890 900 T M D R E ACCATGGACCGGGAA 980
730 T S L E L ACCTCCTTGGAGGCTC 820 M K S R G ATGAAGTCCCGAGGT. 910 V R K I K GTGCGCAAAATCAAAA 1000 H C I A K CACTGCATCGCCAAGC 1090 P L S L Y CCACTGTCTCTCTACL 1180 S L R L K TCCCTCAGGCTGAAGC	740 F M Y TTCATGTAC 830 I Y E ATCTACGAG 920 Q G L CAAGGCCTG 1010 S Q E TCCCAGGAGG 1100 N E E AATCACGAGG 1190 E Y H GAATATCATC	750 L N E V CTGAACGAAGT 840 T P A G ACCCCACGCAGGA 930 G L K P CGCTTGAAATT 1020 R V E G CGAGTGGAAGGA 1110 L V S M CTGGTGGAGCACT 1200 R L Q S CGTCTCCCAGAGG	760 A G K CCCCGCGCAAG 850 T I L CACCATCCTT 940 A E L TGCTGACCTG 1030 K V Q GAAAGTGCAG 1120 N V Q GAACGTGCAG 1210 K V T CAAGGTCACT	770 H C V CATGCCGTG 860 Y H A TACCATGCT 950 V Y T GTGTATACCI 1040 V S V GTGTCCGTCC 1130 C D Y GGTGATTAT 1220 A K GCCAAATAG,	780 G R I D GCCCTATTGA 870 H L D I CATTTAGACAT 960 G L R P GCTTTACGCCC 1050 L K G Q LICAACGGCCA 1140 E P T D GAGCCAACTGA' 1230	790 I V E CATCGTGGAG 880 E A F CGACGCCTTC. 970 S P E TACCCCTGAG 1060 V Y I GGTGTACATCC 1150 A T G TGCCACCGGG 1240	800 810 N R F I G AACCGCTTCATTCGA 890 900 T M D R E ACCATGGACCGGGAA 980 90 C E F V R TCTGAATTTGTCCGC 1070 1080 L G R E S CTCCGCCCGGGAGTCC 1160 1170 F I N I N TTCATCAACATCAAT 1250 260
730 T S L E L ACCTCCTTGGAGGCTC: 820 M K S R G ATGAAGTCCCGAGGT. 910 V R K I K GTGCGCAAAATCAAAA 1000 H C I A K CACTGCATCGCCAAGG: 1090 P L S L Y CCACTGCTCTCTTCACA 1180 S L R L K TCCCTCAGGCTGAAGG 1270	740 F M Y TTCATGTAC 830 I Y E ATCTACGAG. 920 Q G L CAAGGCCTG 1010 S Q E TCCCAGGAGG 1100 N E E MATCAGGAGGA 1190 E Y H GAATATCATU 1280	750 L N E V CTGAACGAAGT 840 T P A G ACCCCACCACAGA 930 G L K P EGCTTGAAATT 1020 R V E G CCAGTCGAAGCA 1110 L V S M CTGGTGAGCAGCAT 1200 R L Q S EGTCTCCCAGAGCA	760 A G K CCCCGCCAAG 850 T I L CACCATCCTT 940 A E L TGCTGACCTG 1030 K V Q SAAAGTGCAG 1120 N V Q SAACGTGCAG 1210 K V T CAAGGTCACT 1300	770 H G V CATGCCGTG 860 Y H A TACCATGCT 950 V Y T GTGTATACCI 1040 V S V GTGTCCGTCC 1130 G D Y GGTGATATAT 1220 A K GCCAAATAC, 1310	780 G R I D GCCCTATTGA 870 H L D I CATTTAGACAT 960 G L R P GGTTTACGCCC 1050 L K G Q LTCAACGGCCAA 1140 E P T D GAGCCAACTGAY 1230	790 I V E CATCGTGCAG. 880 E A F CGACGCCTTC. 970 S P E TAGCCCTGAG 1060 V Y I GGTGTACATCC 1150 A T G TGCCACCGGGC 1240 ATGAGGAGCTT 1330	800 810 N R F I G AACCGCTTCATTCGA 890 900 T M D R E AACCATCGACCGGGAA 980 990 C E F V R TGTGAATTTGTCCGC 1070 1080 L G R E S CTCCGCCGGGGAGTCC 1160 1170 F I N I N TTCATCAACATCAAT 1250 1260 EGGGGCCTCCTCAATT 1340
730 T S L E L ACCTCCTTGGAGCTC: 820 M K S R G ATGAAGTCCCGGAGGT. 910 V R K I K GTGCGCAAAATCAAAA 1000 H C I A K CACTGCATCGCCAAGG: 1090 P L S L Y CCACTGTCTCTCTACL 1180 S L R L K TCCCTCAGGCTGAAGG 1270 TGCGAGATCCCCCAAGGT	740 F M Y TTCATGTAC 830 I Y E ATCTACGAG. 920 Q G L CAAGGCCTGG 1010 S Q E TCCCAGGAG 1100 N E E ANTCAGGAGGAG 1190 E Y H SAATATCAT	750 L N E V CTGAACGAAGT 840 T P A G ACCCCACCACAGA 930 G L K P GGCTTGAAATT 1020 R V E G CCAGTGGAAGCA 1110 L V S M L V S M TGGGTGAGCAT 1200 R L Q S CGTCTCCAGAGCA 1290	760 A G K CCCCGCCAAG 850 T I L CACCATCCTT 940 A E L TGCTGAGCTG 1030 K V Q SAAACTCCAG 1120 N V Q SAACCTCCAG 1210 K V T CAACGTCACT 1300	770 H G V CATGCCGTG 860 Y H A TACCATGCT 950 V Y T GTGTATACCI 1040 V S V GTGTCCGTCC 1130 G D Y GGTGATATAT 1220 A K GCCAAATAG, 1310	780 G R I D GGCCGTATTGA 870 H L D I CATTTAGACAT 960 G L R P GGTTTACGCCC 1050 L K G Q LTCAAGGGCCA 1140 E P T D GAGCCAACTGAT 1230 CCCGTGTACA 1230 CGTCTCCCCCG	790 I V E CATCGTGGAG 880 E A F CGACGCCTTC. 970 S P E TAGCCCTGAG 1060 V Y I GGTGTACATC 1150 A T G TGCCACCGGG 1240 ATGAGGAGCT 1330 GCTGCGCAGCG	800 810 N R F I G AACCGCTTCATTEGA 890 900 T M D R E AACCGCGGGAA 980 990 C E F V R TGTGAATTTGTCCGC 1070 1080 L G R E S CTCGGCCGGGGAGTCC 1160 1170 F I N I N TCATCAACATCAATT 1250 1260 GGGGGCCTCCTCAATT 1340 1350 TAGTGGGGCTGCCAG
730 T S L E L ACCTCCTTGGAGCTC: 820 M K S R G ATGAAGTCCCGAGCT. 910 V R K I K GTGCGCAAAATCAAAA 1000 H C I A K CACTGCATCGCCAAGC 1090 P L S L Y CCACTGTCTCTCTACL 1180 S L R L K TCCCTCAGGCTGAGC 1270 TGCAGATCCCCCAAGC 1360	740 F M Y TTCATGTAC 830 I Y E ATCTACGAG. 920 Q G L CAAGCCCTGG 1010 S Q E TCCCAGGAGG 1100 N E E MATCACCAGC 1190 E Y H SAATATCAT 1280 CACAGCCCCC 1370	750 L N E V CTGAACGAAGT 840 T P A G ACCCCACCAGGO 930 G L K P EGCTTGAAATT 1020 R V E G CCACTGGAAGCG 1110 L V S M L V S M L V S M CCCACTGCACCAGGO 1290 R L Q S CCTCTCCACAGGO 1290	760 A G K CCCCGCGCAAG 850 T I L CACCATCCTT 940 A E L TGCTGAGCTG 1030 K V Q SAAGGTCCAG 1210 N V Q SAACGTCCAG 1210 K V T CAAGGTCACT 1300	770 H G V CATGCCGTG 860 Y H A TACCATGCT 950 V Y T GTGTATACCI 1040 V S V GTGTCCCTCC 1130 G D Y GGTGATATAT 1220 A K GCCAAATAG, 1310 ATTGTGACT 1400	780 G R I D GCCCTATTGA 870 H L D I CATTTAGACAT 960 G L R P GGTTTACGCCC 1050 L K G Q LTCAAGGGCCA 1140 E P T D AGCCCATGTACA 1230 CTTCTCCCCCG 1410	790 I V E CATCGTGGAG 880 E A F CGAGGCCTTC. 970 S P E TAGCCCTGAG 1060 V Y I GGTGTACATC: 1150 A T G TGCCACCGGC 1240 ATCAGGAGCTY 1330 GCTGCCACCG	800 810 N R F I G AACCGCTTCATTEGA 890 900 T M D R E AACCATEGACCGEGAA 980 990 C E F V R TEGGAATTTGTCCGC 1070 1080 L G R E S CTCGGCCCGGGAATTCC 1160 1170 F I N I N TCATCAACATCAATT 1250 CGGGCCCTCCTCAATT 1340 1350 TAGTGGGGCTGCCAG 1430
730 T S L E L ACCTCCTTGGAGCTC: 820 M K S R G ATGAAGTCCCGAGGT, 910 V R K I K GTGCGCAAAATCAAAA 1000 H C I A K CACTGCATCGCCAAG 1090 P L S L Y CCACTGTCTCTCTAAC 1180 S L R L K TCCCTCAGGCTGAGC 1270 TGCAGATCCCCCAAGT 1360	740 F M Y TTCATGTAC 830 I Y E ATCTACGAG. 920 Q G L CAAGGCCTGGI 1010 S Q E TCCCAGGAGGI 1100 N E E ATTATCATCATI 1280 CACAGGCCCC 1370 CTCGTCCCCC	750 L N E V CTGAACGAAGT(840 T P A G ACCCCACCAGGA 930 G L K P EGCTTGAAATT1 1020 R V E G CGAGTGGAAGCG 1110 L V S M TTGGTGAGCGATC 1200 R L Q S CGTCTCCCAGAGC 1290 CTCAACCCTCC 1380 CCTGAACCCTCC 1470	760 A G K CCCCGCGCAAG 850 T I L CACCATCCTT 940 A E L TGCTGAGCTG 1030 K V Q GAAGGTCCAG 1210 N V Q GAACGTCCAG 1210 K V T CAAGGTCACT 1300 CAAGGTCACT 1390 CAAGGTTGT 1480	770 H G V CATGCCGTG 860 Y H A TACCATGCT 950 V Y T GTGTATACCC 1040 V S V GTGTCCTCC 1130 G D Y GGTCATTATY 1220 A K GCCAAATAG. 1310 ATTGTGACT 1490	780 G R I D GCCCTATTGA 870 H L D I CATTTAGACAT 960 G L R P GGTTTACGCCC 1050 L K G Q LTCAAGGGCCCA 1140 E P T D SAGCCAACTGACA 1320 CGTCTCCCCCG 1410 SAAGCCTCCCCCGCCCCCCCCCCCCCCCCCCCCCCCCCCC	790 I V E CATCGTGGAG. 880 E A F CGACGCCTTC. 970 S P E TAGCCCTGAG 1060 V Y I GGTGTACATC: 1150 A T G TGCCACCGGG 1240 ATCACGAGCTC 1420 GGTGGCACCGG	800 810 N R F I G AACCGCTTCATTCGA 890 900 T M D R E AACCATCGACCGCAA 980 990 C E F V R TGTGAATTTGTCCGC 1070 1080 L G R E S CTCGCCCGGGAATTCC 1160 1170 F I N I N TTCATCAACATCAATT 1250 CGGCGCCTCCTCAATT 1340 1350 TAGTGGGGCTGCCAG 1440 STCGCGCACCTATAAA 1520 1520
730 T S L E L ACCTCCTTGGAGCTC: 820 M K S R G ATGAAGTCCCGAGGT. 910 V R K I K GTGCGCAAAATCAAAA 1000 H C I A K CACTGCATCGCCAAGC 1090 P L S L Y CCACTGTCTCTCTACL 1180 S L R L K TCCCTCAGGCATCCCAAGC 1360 GCCCCAGCTTTGTTCC 1450	740 F M Y TTCATGTAC 830 I Y E ATCTACGAG. 920 Q G L CAAGGCCTGG 1010 S Q E TCCCAGGAGG 1100 N E E ATTATCATCAT 1280 E Y H SAATATCAT 1370 E Y H SAATATCAT 1370	750 L N E V CTGAACGAAGT 840 T P A G ACCCCACCACGA 930 G L K F GGCTTGAAATT 1020 R V E G CGACTGGAAGCGA 1110 L V S M CTGGTGAACCATC 1200 R L Q S CGTCTCCACAGC 1290 CTATTGTTGTGJ 1380 CCTGAAGCCTCC 1470	760 A G K CCCCGCGCAAG 850 T I L CACCATCCTT 940 A E L TGCTGAGCTG 1030 K V Q GAAGTGCAG 1120 N V Q GAACGTGCAG 1210 K V T CAGGGCCAGT 1210 K V T CAACGTCAGTCAG 1390 CAACGTTGT 1480	770 H G V CATGCCGTG 860 Y H A TACCATGCT 950 V Y T GTGTATACCA 1040 V S V GTGTCCCTCC 1130 G D Y GGTCATTATT 1220 A K GCCAAATAG. 1310 ATTGTGACT. 1400 CATCGAGGG 1490	780 G R I D GGCCGTATTGA 870 H L D I CATTTAGACAT 960 G L R P GGTTTACGCCC 1050 L K G Q LICAAGGGCCA 1140 E P T D SAGCCAACTGA 1230 CGTCTCCCCCG 1410 CGCCGCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC	790 I V E CATCGTGGAG. 880 E A F CGACGCCTTC. 970 S P E TAGCCCTGAG 1060 V Y I GGTGTACATC: 1150 A T G TGCCACCGGC: 1240 ATCACGAGCTT: 1420 GGTGTGCACCGGC: 1420	800 810 N R F I G AACCGCTTCATTCGA 890 900 T M D R E AACCATCGACCGCAA 980 990 C E F V R TGTGAATTTGTCCGC 1070 1080 L G R E S CTCGCCCGGGAGTCC 1160 1170 F I N I N TTCATCAACATCAATT 1250 1260 CGGGGCCTCCTCAATT 1340 1350 TAGTGGGGCTGCCAG 1440 GTGCGGAGCTATAAA 1520 1530

Figure 2. The nucleotide and amino acid sequences of human argininosuccinate synthetase cDNA. Three tandem arginine codons 5' to the initiator ATG are overlined. Arrows indicate the locations of the observed nucleotide variations.



Figure 3. Summary of the Nucleotide Variations observed within Multiple Human Argininosuccinate Synthetase cDNA Clones (positions 757-762, 1320, 1431, and 1555).

Bock and S.O. Freytag, unpublished observations).

Nucleotide variations also were detected at positions 1320, 1431, and 1555 in the 3' untranslated region (Figures 2 and 3). At position 1320, pAS2 had a C while pAS1, pAS3, pAS9, and pAS12 had T's. At position 1431, pAS1 had an A, while pAS2 and pAS9 had G's. At position 1555, pAS1 and pAS9 both had an A, while pAS2 had a T.

Arginine Codons in the 5' Untranslated Region

Three tandem arginine codons occurred in the 5' untranslated region. They were in the same reading frame as the coding sequence and were separated from the initiator ATG by 27 nucleotides. No ATG initiator codon was present upstream from the arginine codons although two GTG codons (an initiator codon in some instances) did occur upstream in this reading frame. No terminator codons were between the arginine codons and the initiator ATG. This sequence was from canavanine-resistant cells and could conceivably be different in wild type cells. The 5' untranslated region was 59% G + C (44/75 bases), and this could favor the occurrence of the CGX series of arginine codons.

DISCUSSION

There are multiple argininosuccinate synthetase-like sequences in the human genome. Our approach to interpreting and integrating the data from six cDNA clones assumed the existence of a single active diploid locus for argininosuccinate synthetase, including canavanine-resistant cells. The strongest argument for this interpretation was the analysis of cultured skin fibroblasts with mutations in the structural gene. These data indicated that all detectable mRNA was derived from a single locus in those cells (8). In addition, extensive analysis of genomic clones indicated a large expressed gene with at least eleven introns but no evidence for gene duplication (6).

The general features of the cDNA sequence were typical for a eukaryotic gene. The assigned coding region predicted an amino acid content in close agreement with that published for bovine liver enzyme (10). The content of basic amino acids (21 arginines and 33 lysines in 412 amino acids) was unusually high and was consistent with the pI of 9.0 reported for the human enzyme (13).

The occurrence of sequence variations both in the coding region and in the 3' untranslated region led to the sequencing of additional cDNA clones from a single cloning experiment. Assuming that there is only one active diploid locus, we interpreted the extra A at position 757-762 in pAS1 to represent an RNA polymerase error or cloning artifact. The occurrence of six A's in two cDNAs, in three processed pseudogenes and in the genomic exon supported this interpretation. We could not eliminate the possibility that the tissue culture cell line was heterozygous for a frameshift mutation. Sequence variations in the 3' untranslated region occurred at three sites, but the most 3' variation could have involved slight differences in the site of polyadenylation and variations at bases 1320 and 1431 were less ambiguous. Initially this region was sequenced in multiple clones in an attempt to demonstrate genetic heterogeneity in the 3' untranslated portion of the mRNA. Additional attempts to determine if DNA polymorphisms were present were unsuccessful. Although the sequence at position 1320 would involve the presence or absence of the sequence CGCG, a restriction enzyme recognition site for FnuDII or ThaI, efforts to analyze for such a variation were unsuccessful due in part to the presence of a very large number of pseudogenes which complicate Southern blotting analysis of genomic DNA (5). We have been unsuccessful, as have most other laboratories, in using Sl nuclease analysis to detect single base variations in the mRNA sequence (8). Assuming a single active locus, the isolation of two sequences in a small sample of clones would suggest that enzyme overproduction in canavanine-resistant cells involves increased expression of both alleles at the locus, i.e. a transacting mechanism. Three different combinations were identified at the 1320/1431 bases; T/A in pAS1, C/G in pAS2, and T/G in pAS9 (Figure 3). Again assuming a single active locus, these three combinations indicated at

least one artifactual result although polymorphism still could have occurred at either position. Single base artifacts at position 1320 in pAS2 and at position 1431 in pAS1 could explain the data. In concert with this possibility, the sequence in two processed pseudogenes agreed with that in pAS9 (S.O. Freytag and H.G. Bock, manuscript in preparation). Other approaches to distinguish whether the canavanine-resistant phenotype is trans-acting or cis-acting are being pursued.

The occurrence of three tandem arginine codons, nucleotides 40-48, in the 5' untranslated region may be relevant to the repression of this enzymic activity by arginine. The steady state level of mRNA for argininosuccinate synthetase is regulated by the arginine concentration of tissue culture medium (2,5). The sequence determined may include virtually all of the 5' untranslated region since two processed pseudogenes diverged from each other within four bases of the 5' end of pAS4 (S.O. Freytag and H.G. Bock, manuscript in preparation). A limited survey of the 5' untranslated region from over thirty other eukaryotic cDNAs indicated a need for cautious interpretation. Arginine codons occurred in clusters in a number of cDNAs, particularly if the G + C content was high, but no other instance of three tandem arginine codons was observed. The presence of tandem codons for a regulatory amino acid is similar to findings in bacterial amino acid operons which are regulated by attenuation (14-17). A similar mechanism may occur in yeast (18) and in SV40 (19). In our system, the absence of any evidence for translation of this region and the absence of a sequence resembling a termination site for transcription suggested that attenuation as described in bacteria would be unlikely. However, interaction of arginine codons with arginyl-tRNA could still serve as a component of a regulatory mechanism. A recent report (20) that nuclear RNA processing or stability was influenced by the occurrence of a nonsense codon in the β -globin gene, provides a precedent for interaction of mRNA with tRNA to affect RNA processing or stability in higher eukaryotic cells.

ACKNOWLEDGEMENTS

This work was supported by National Institutes of Health Research Fellowship Award GM07466-02 and grant GM 27593. We thank Drs. Svend O. Freytag and Jeffrey Rosen for constructive criticism of this manuscript and Lynn Loewenstein for help in the preparation of this manuscript.

*To whom correspondence should be addressed

REFERENCES

- Walser, M. (1983) in Metabolic Basis of Inherited Disease, Stanbury, J.B., Wyngaarden, J.B., Fredrickson, D.S., Goldstein, J.L., and Brown, M.S. Eds., pp.402-438, McGraw-Hill, New York.
- 2. Schimke, R.T. (1964) J. Biol. Chem. 239, 136-145.
- 3. Jacoby, L.B. (1978) Somatic Cell Genetics 4, 221-231.
- 4. Su, T.-S., Beaudet, A.L. and O'Brien, W.E. (1981) Biochemistry 20, 2956-2960.
- Su, T.-S., Bock, H.-G.O., O'Brien, W.E. and Beaudet, A.L. (1981) J. Biol. Chem. 256, 11826-11831.
- Beaudet, A.L., Su, T.-S., Bock, H.-G.O., Freytag, S.O. and O'Brien, W.E. (1983 in Recombinant DNA Applications to Human Disease, Caskey, C.T., White, R., ed., pp. 97-103, Cold Spring Harbor Laboratory, New York.
- Su, T.-S., Bock, H.-G.O., Beaudet, A.L. and O'Brien, W.E. (1982) J. Clin. Invest. 70, 1334-1339.
- 8. Su, T.-S., Beaudet, A.L. and O'Brien, W.E. (1983) Nature 301, 533-534.
- 9. Maxam, A.M. and Gilbert, W. (1980) Methods Enzymol. 65, 499-560.
- 10. Ratner, S. (1982) Proc. Natl. Acad. Sci. U.S.A. 79, 5197-5199.
- 11. Fitzgerald, M. and Shenk, T. (1981) Cell 24, 251-260.
- 12. Proudfoot, N.J. and Brownlee, G.G. (1974) Nature 252, 359-362.
- Hudson, L.D., Erbe, R.W. and Jacoby, L.B. (1980) Proc. Natl. Acad. Sci. U.S.A. 77, 4234-4238.
- 14. Keller, E.B. and Calvo, J.M. (1979) Proc. Natl. Acad. Sci. U.S.A. 76, 6186-6190.
- 15. Yanofsky, C. (1981) Nature 289, 751-758.
- 16. Platt, T. (1981) Cell 24, 10-23.
- 17. Kolter, R. and Yanofsky, C. (1982) Ann. Rev. Genet. 16, 113-134.
- Andreadis, A., Hsu, Y.-P., Kohlhaw, G.B. and Schimmel, P. (1982) Cell 31, 319-325.
- 19. Hay, N., Skolnik-David, H. and Aloni, Y. (1982) Cell 29, 183-193.
- Takeshita, K., Scarpa, A., Chan, L. and Benz, E.J., Jr. (1983) Clin. Res. 31, 479A.