Volume 9 Number 22 1981 Nucleic Acids Research

The sequence of human serum albumin cDNA and its expression in E. coli

Richard M.Lawn, John Adelman, Susan C.Bock, Arthur E.Franke, Catherine M.Houck, Richard C. Najarian, Peter H.Seeburg and Karen L.Wion

Department of Molecular Biology, Genentech Inc., 460, Point San Bruno Boulevard, South San Francisco, CA 94080, USA

Received 31 August 1981

ABSTRACT

A recombinant plasmid has been constructed which contains the mature protein coding region of the human serum albumin (HSA) gene. Bacteria containing this plasmid synthesize HSA protein under control of the E. coli trp promoter-operator. The DNA sequence and predicted protein sequence of HSA were determined from the cDNA plasmid and are compared to existing data obtained from direct protein sequencing. The DNA sequence predicts a mature protein of 585 amino acids preceded by a 24 amino acid "prepro" peptide.

INTRODUCTION

Human serum albumin (HSA) is the major protein component in adult plasma. The protein is produced in the liver and is largely responsible for maintaining normal osmolarity in the bloodstream and functions as a carrier for numerous small molecules (1, 2). The apparent fetal counterpart of HSA is α-fetoprotein (3). The two proteins have similar physicochemical properties and have reciprocal levels of expression in fetal and adult liver (4). Rat serum albumin and α-fetoprotein share 34 percent amino acid and 50 percent nucleotide sequence homology (5). Only partial protein sequence data exists for human α-fetoprotein (6-8). The complete protein sequence of HSA has recently been published (9-12). The published protein sequences disagree in about 20 residues as well as in the total number of amino acids in the mature protein [584 amino acids (9); 585 (12)]. Some evidence suggests that HSA is initially synthesized as a precursor molecule (13, 14). The precursor forms of bovine (15) and rat (16) serum albumin are known to contain 25 and 24 amino acid "prepro" peptide sequences respectively.

Knowledge of the DNA sequence organization of human serum albumin and α-fetoprotein genes will help elucidate evolutionary, regulatory and functional properties of these related proteins. The availability of

purified HSA produced through recombinant DNA technologies more abundantly than is now possible from blood fractionation may have widespread research and clinical impact. In this paper we report the isolation of cDNA clones spanning the entire sequence of the protein coding and 3' untranslated portions of HSA mRNA. These cDNA clones were used to construct a recombinant plasmid which directs the expression in E. coli of the mature HSA protein from the trp promoter. We also present the complete nucleotide and predicted amino acid sequence of HSA.

MATERIALS AND METHODS

Synthesis and Cloning of cDNA. Poly (A)+ RNA was prepared from quickly frozen human liver samples obtained from biopsy or from cadaver donors by either ribonucleoside-vanadyl complex (17) or guanidinium thiocyanate (18) procedures. cDNA reactions were performed essentially as described in (19) employing as primers either oligo-deoxynucleotides prepared by the phosphotriester method (20) or oligo $(dT)_{12-18}$ (Collaborative Research). For typical cDNA reactions 25-35 μg of poly (A)+ RNA and 40-80 pmol of oligonucleotide primer were heated at 90° for 5 minutes in 50 mM NaCl. The reaction mixture was brought to final concentrations of 20 mM Tris HCl pH 8.3, 20 mM KCl, 8 mM $MgCl_2$, 30 mM dithiothreitol, 1 mM dATP, dCTP, dGTP, dTTP (plus ^{32}P-dCTP (Amersham) to follow recovery of product) and allowed to anneal at 42°C for 5'. 100 units of AMV reverse transcriptase (BRL) were added and incubation continued at 42° for 45 minutes. Second strand DNA synthesis, SI treatment, size selection on polyacrylamide gels, deoxy (C) tailing and annealing to pBR322 which was cleaved with PstI and deoxy (G) tailed, were performed as previously described (21, 22). The annealed mixture was used to transform E. coli K-12 strain 294 (23) by a published procedure (24).

Screening of Recombinant Plasmids with ^{32}P-labelled Probes. E. coli transformants were grown on LB-agar plates containing 5 μg/ml tetracycline, transferred to nitrocellulose filter paper (Schleicher and Schuell, BA85) and tested by hybridization using a modification of the in situ colony screening procedure (25). ^{32}P-end labelled (26) oligodeoxynucleotide fragments of from 12 to 16 nucleotides in length were used as direct hybridization probes, or ^{32}P-cDNA probes were synthesized from RNA using oligo(dT) or oligodeoxynucleotide primers (19). Filters were hybridized overnight in 5X Denhardt's solution (27), 5xSSC (1xSSC = 0.15 M NaCl 0.015 M Na Citrate), 50 mM Na phosphate pH 6.8, 20 μg/ml salmon

sperm DNA at temperatures ranging from 4° to 42° and washed in salt concentrations varying from 1 to 0.2xSSC plus 0.1 percent SDS at temperatures ranging from 4° to 42° depending on the length of the ^{32}P-labelled probe (28). Dried filters were exposed to Kodak XR-2 X-ray film using DuPont Lightning-Plus intensifying screens at -80°.

DNA Preparation and Restriction Enzyme Analysis. Plasmid DNA was prepared in either large scale (29) or small scale ("miniprep"; 30) quantities and cleaved by restriction endonucleases (New England Biolabs, BRL) following manufacturers conditions. Slab gel electrophoresis conditions and electroelution of DNA fragments from gels have been described (31).

DNA Sequence Determination. DNA sequences were established by both the method of Maxam and Gilbert (26) utilizing end-labelled DNA fragments and by dideoxy chain termination (32) on single stranded DNA from phage M13 mP7 subclones (33) utilizing synthetic oligonucleotide (20) primers. Each region was independently sequenced several times.

Construction of 5' End of Albumin Gene for Direct Expression of HSA. 10 µg (~16 pmol) of the ~1200 bp PstI insert of plasmid F-47 was boiled in H_2O for 5 minutes and combined with 100 pmol of ^{32}P-end labelled 5' primer (dATGGATGCACACAAG). The mixture was quenched on ice and brought to a final volume of 120 µl of 6 mM Tris HCl pH 7.5, 6 mM $MgCl_2$, 60 mM NaCl, 0.5 mM dATP, dCTP, dGTP, dTTP at 0°. 10 units of DNA polymerase I Klenow fragment (Boehringer-Mannheim) were added and the mixture incubated at 24° for 5 hr. Following phenol/chloroform extraction, the product was digested with HpaII, electrophoresed in a 5 percent polyacrylamide gel, and the desired 450 bp fragment electroeluted. The single stranded overhang produced by XbaI digestion of the vector plasmid pLeIF A25 (21) was filled in to produce blunt DNA ends by adding deoxynucleoside triphosphates to 10 µM and 10 units DNA polymerase I Klenow fragment to the restriction endonuclease reaction mix and incubating at 12° for 10 minutes. Restriction endonuclease fragments (0.1 - 1 µg in approximate molar equality) were annealed and ligated overnight at 12° in 20 µl of 50 mM Tris HCl pH 7.6, 10 mM $MgCl_2$, 0.1 mM EDTA, 5 mM dithiothreitol, 1 mM rATP with 50 units T4 ligase (N.E. Biolabs). Further details of plasmid construction are discussed below.

Protein Analysis. Two ml cultures of recombinant E. coli strains were grown in either LB or M9 media plus 5 µg/ml tetracycline to densities of A_{550} = 1.0, pelleted, washed, repelleted, and suspended in 2 ml of LB or

supplemented M9 (M9 + 0.2 percent glucose, 1 μg/ml thiamine, 20 μg/ml standard amino acids except methionine which was 2 μg/ml; tryptophan was excluded). Each growth media also contained 5 μg/ml tetracycline and 100 μCi ^{35}S-methionine (NEN; 1200 Ci/mmol). After 1 hr incubation at 37°, bacteria were pelleted, freeze-thawed and resuspended in 200 μl 50 mM Tris HCl pH 7.5, 0.12 mM NaEDTA then placed on ice for 10 minutes following subsequent additions of lysozyme to 1 mg/ml, NP40 to 0.2 percent, and NaCl to 0.35 M. The lysate was adjusted to 10 mM $MgCl_2$ and incubated with 50 μg/ml DNase I (Worthington) on ice for 30 min. Insoluble material was removed by mild centrifugation. Samples were immunoprecipitated with rabbit anti-HSA (Cappel Labs) and staphylococcal absorbent (Pansorbin; Cal Biochem) as described (34), and subjected to SDS polyacrylamide gel electrophoresis (35).

RESULTS AND DISCUSSION

cDNA Cloning. Initial cDNA clones were obtained by priming human liver mRNA with oligo (dT). They were screened by colony hybridization with both total liver cDNA (to identify clones containing abundant RNA species) and with two ^{32}P-labelled cDNA preparations obtained by priming liver mRNA with two sets of four 11-base oligodeoxynucleotides synthesized to represent the possible coding variations for amino acids 546-549 and 294-297 of HSA. Positive colonies never contained more than about the 3' half of the protein coding region of the expected HSA mRNA sequence. (The longest of these recombinants was designated B-44.) Since existing procedures were unable to directly copy an mRNA of the expected size (~2000 bp), synthetic oligodeoxynucleotides were prepared to correspond to the antimessage strand at regions near the 5' extreme of B-44. From the nucleotide sequence of B-44, we constructed a 12 base oligodeoxynucleotide corresponding to amino acids 369-373. This was used to prime cDNA synthesis of liver mRNA and produce cDNA clones in pBR322 containing the 5' portion of the HSA message while overlapping the existing B-44 recombinant. Approximately 400 resulting clones were screened by colony hybridization with a 16 base oligodeoxynucleotide fragment located slightly upstream of the 12 base priming region in the mRNA sequence we had thus far determined. Approximately 40 percent of the colonies hybridized to both probes. Many of those colonies which failed to contain hybridizing plasmids presumably resulted from RNA self-priming or priming with contaminating oligo (dT) during reverse transcription, or lost the 3'

region containing the sequence used for screening. "Miniprep" amounts of plasmid DNA from hybridizing colonies were digested with PstI. Three recombinant plasmids contained sufficiently large inserts to code for the remaining 5' portion of the HSA message. Two of these (F-15 and F-47) contained the extreme 5' coding portion of the mature protein message but failed to extend back to a PstI site necessary for joining with B-44 to reform the complete albumin gene. Recombinant F-61 possessed this PstI site but lacked the entire 5' end. A three part reconstruction of the entire message sequence was possible employing restriction endonuclease sites in common with the part length clones F-47, F-61 and B-44 (Fig. 1).

An additional cDNA clone extending even farther 5' was obtained by similar oligodeoxynucleotide primed cDNA synthesis (from a primer corresponding to amino acid codons no. 175-179). Although not employed in the construction of the mature HSA expression plasmid, this cDNA clone (P-14) allowed determination of the DNA sequence of the "prepro" peptide coding and 5' non-coding regions of HSA mRNA.

The mature HSA mRNA sequence was joined to a vector plasmid for direct expression of the mature protein in E. coli via the trp promoter-operator. The plasmid pLeIF A25 directs the expression of human leukocyte interferon A (IFNα2) (21). It was digested with XbaI and the cleavage site "filled in" to produce blunt DNA ends with DNA polymerase I Klenow fragment and deoxynucleoside triphosphates. After subsequent digestion with PstI, a "vector" fragment was gel purified that contained pBR322 sequences and a 300 bp fragment of the E. coli trp promoter, operator, and ribosome binding site of the trp leader peptide terminating in the artificially blunt ended XbaI cleavage site. A 15-base oligodeoxynucleotide was designed to contain the initiation codon ATG followed by the 12 nucleotides coding for the first four amino acids of mature HSA as determined by DNA sequence analysis of clone F-47. In a process referred to as "primer repair", the gene containing PstI fragment of F-47 was denatured, annealed with excess 15-mer and reacted with DNA polymerase I Klenow fragment and deoxynucleoside triphosphates. This reaction extends a new second strand downstream from the annealed oligonucleotide, degrades the single stranded DNA upstream of codon number one and then polymerizes upstream three nucleotides complementary to ATG. In addition, when this product is blunt-end ligated to the prepared vector fragment, its initial adenosine residue recreates an XbaI restriction site. Following the primer repair reaction, the DNA was digested with HpaII and a 450 bp fragment containing the 5' portion of the

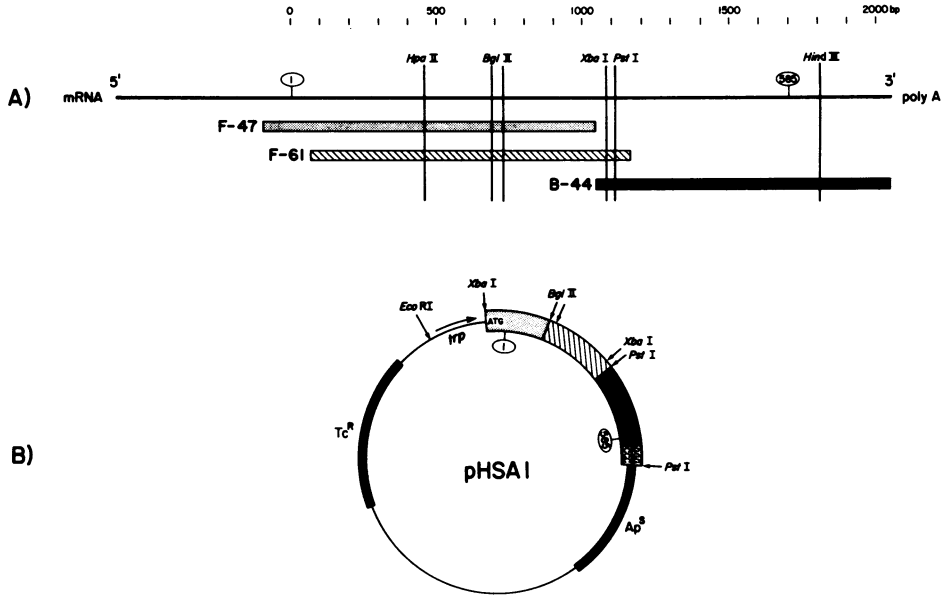

Fig. 1 Construction of pHSA1

A) The top line represents the mRNA coding for the human serum albumin protein and below it are the regions contained in the cDNA clones F-47, F-61 and B-44 described in the text. The initial and final amino acid codons of the mature HSA mRNA are indicated by circled 1 and 585 respectively. Restriction endonuclease sites involved in the construction of pHSA1 are shown by vertical lines. An approximate size scale in nucleotides is included.

B) The completed plasmid pHSA1 is shown with HSA coding regions derived from cDNA clones shaded as in A). Selected restriction sites and terminal codons number 1 and 585 are indicated as above. The E. coli trp promoter-operator region is shown with an arrow representing the direction of transcription. G:C denotes an oligo (dG:dC) tail. The leftmost XbaI site and the initiation codon ATG were added synthetically. The tetracycline (Tc) and ampicillin (Ap) resistance genes in the pBR322 portion of pHSA1 are indicated by a heavy line. See the text for further details.

mature albumin gene was gel purified (see Fig. 1). This fragment was annealed and ligated to the vector fragment and to the gel isolated HpaII to PstI portion of F-47 and used to transform E. coli cells. Diagnostic restriction endonuclease digests of plasmid minipreps identified the recombinant A-26 which contained the 5' portion of the mature albumin coding region ligated properly to the trp promoter-operator. For the final steps in assembly, the A-26 plasmid was digested with BglII plus PstI and

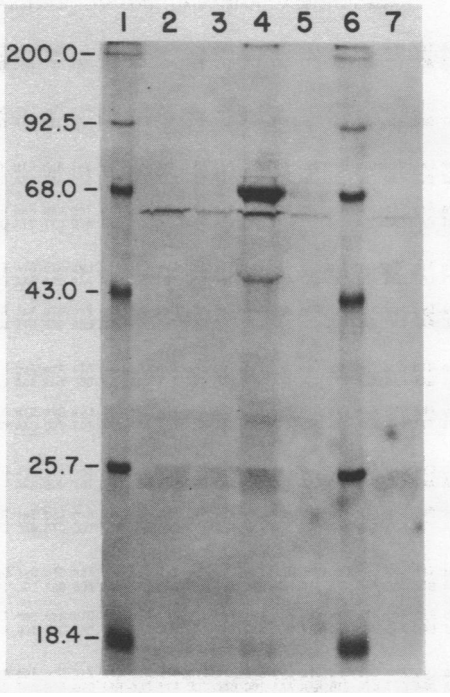

Fig. 2. Immunoprecipitation of Bacterially Synthesized HSA.
E. coli cells transformed with albumin expression plasmid pHSA1 (lanes 4 and 5) or control plasmid pLeIFA25 (containing an interferon α gene in the identical expression vehicle; lanes 2, 3 and 7) were grown in ^{35}S-methionine-supplemented media. Samples in lanes 2, 4 and 7 were induced for expression from the trp promoter in M9 media lacking tryptophan; samples in lanes 3 and 5 were grown in tryptophan-containing LB broth to repress the trp promoter. Each sample lane of the autoradiograph of the SDS-polyacrylamide gel presented here contains labeled protein immunoprecipitated from 0.75 ml of cells at a density of A_{550} = 1 (see Methods). Lanes 1 and 6 contain radioactive protein standards (BRL) whose molecular weight in kilodaltons is indicated at the left. Bacterially synthesized HSA is seen in lane 4 comigrating with the 68,000 dalton (d) ^{14}C-labeled bovine serum albumin standard. Increased production of serum albumin in the induced versus repressed culture of pHSA1 represents higher levels of synthesis of plasmid encoded protein rather than a difference in ^{35}S-methionine pool specific activities for minimal versus rich media (data not shown). The sharp band at 60,000 d is an apparent artifact; this band is seen in both induced and repressed pHSA1 and control transformants, and binds to preimmune (lane 7) as well as anti-HSA IgGs (lanes 2-5). The minor 47,000 d band in lane 4 is apparently plasmid encoded and may represent a prematurely terminated form of bacterially synthesized HSA.

```
                                                                                                                    AGGATGTCTTCTGGCAATTTCATATAAGTATTTTTTCAAAAATGTCTCTTCTGTCAACCCCACGCCTTTGGC

    (Prepro)
    Met Lys Trp Val Thr Phe Ile Ser Leu Leu Phe Leu Phe Ser Ser Ala Tyr Ser Arg Gly Val Phe Arg Arg
ACA ATG AAG TGG GTA ACC TTT ATT TCC CTT CTT TTT CTC TTT AGC TCG GCT TAT TCC AGG GGT GTG TTT CGT CGA

1(Mature)
Asp Ala His Lys Ser Glu Val Ala His Arg Phe Lys Asp Leu Gly Glu Glu Asn Phe Lys Ala Leu Val Leu Ile
GAT GCA CAC AAG AGT GAG GTT GCT CAT CGG TTT AAA GAT TTG GGA GAA GAA AAT TTC AAA GCC TTG GTG TTG ATT

                                                                                                                        50
Ala Phe Ala Gln Tyr Leu Gln Gln Cys Pro Phe Glu Asp His Val Lys Leu Val Asn Glu Val Thr Glu Phe Ala
GCC TTT GCT CAG TAT CTT CAG CAG TGT CCA TTT GAA GAT CAT GTA AAA TTA GTG AAT GAA GTA ACT GAA TTT GCA

Lys Thr Cys Val Ala Asp Glu Ser Ala Glu Asn Cys Asp Lys Ser Leu His Thr Leu Phe Gly Asp Lys Leu Cys
AAA ACA TGT GTA GCT GAT GAG TCA GCT GAA AAT TGT GAC AAA TCA CTT CAT ACC CTT TTT GGA GAC AAA TTA TGC

                                                                                                                        100
Thr Val Ala Thr Leu Arg Glu Thr Tyr Gly Glu Met Ala Asp Cys Cys Ala Lys Gln Glu Pro Glu Arg Asn Glu
ACA GTT GCA ACT CTT CGT GAA ACC TAT GGT GAA ATG GCT GAC TGC TGT GCA AAA CAA GAA CCT GAG AGA AAT GAA

Cys Phe Leu Gln His Lys Asp Asp Asn Pro Asn Leu Pro Arg Leu Val Arg Pro Glu Val Asp Val Met Cys Thr
TGC TTC TTG CAA CAC AAA GAT GAC AAC CCA AAC CTC CCC CGA TTG GTG AGA CCA GAG GTT GAT GTG ATG TGC ACT

                                                                                                                        150
Ala Phe His Asp Asn Glu Glu Thr Phe Leu Lys Lys Tyr Leu Tyr Glu Ile Ala Arg Arg His Pro Tyr Phe Tyr
GCT TTT CAT GAC AAT GAA GAG ACA TTT TTG AAA AAA TAC TTA TAT GAA ATT GCC AGA AGA CAT CCT TAC TTT TAT

Ala Pro Glu Leu Leu Phe Phe Ala Lys Arg Tyr Lys Ala Ala Phe Thr Glu Cys Cys Gln Ala Ala Asp Lys Ala
GCC CCG GAA CTC CTT TTC TTT GCT AAA AGG TAT AAA GCT GCT TTT ACA GAA TGT TGC CAA GCT GCT GAT AAA GCT

                                                                                                                        200
Ala Cys Leu Leu Pro Lys Leu Asp Glu Leu Arg Asp Glu Gly Lys Ala Ser Ser Ala Lys Gln Arg Leu Lys Cys
GCC TGC CTG TTG CCA AAG CTC GAT GAA CTT CGG GAT GAA GGG AAG GCT TCG TCT GCC AAA CAG AGA CTC AAA TGT

Ala Ser Leu Gln Lys Phe Gly Glu Arg Ala Phe Lys Ala Trp Ala Val Ala Arg Leu Ser Gln Arg Phe Pro Lys
GCC AGT CTC CAA AAA TTT GGA GAA AGA GCT TTC AAA GCA TGG GCA GTG GCT CGC CTG AGC CAG AGA TTT CCC AAA

                                                                                                                        250
Ala Glu Phe Ala Glu Val Ser Lys Leu Val Thr Asp Leu Thr Lys Val His Thr Glu Cys Cys His Gly Asp Leu
GCT GAG TTT GCA GAA GTT TCC AAG TTA GTG ACA GAT CTT ACC AAA GTC CAC ACG GAA TGC TGC CAT GGA GAT CTG

Leu Glu Cys Ala Asp Asp Arg Ala Asp Leu Ala Lys Tyr Ile Cys Glu Asn Gln Asp Ser Ile Ser Ser Lys Leu
CTT GAA TGT GCT GAT GAC AGG GCG GAC CTT GCC AAG TAT ATC TGT GAA AAT CAG GAT TCG ATC TCC AGT AAA CTG

                                                                                                                        300
Lys Glu Cys Cys Glu Lys Pro Leu Leu Glu Lys Ser His Cys Ile Ala Glu Val Glu Asn Asp Glu Met Pro Ala
AAG GAA TGC TGT GAA AAA CCT CTG TTG GAA AAA TCC CAC TGC ATT GCC GAA GTG GAA AAT GAT GAG ATG CCT GCT

Asp Leu Pro Ser Leu Ala Ala Asp Phe Val Glu Ser Lys Asp Val Cys Lys Asn Tyr Ala Glu Ala Lys Asp Val
GAC TTG CCT TCA TTA GCT GCT GAT TTT GTT GAA AGT AAG GAT GTT TGC AAA AAC TAT GCT GAG GCA AAG GAT GTC

                                                                                                                        350
Phe Leu Gly Met Phe Leu Tyr Glu Tyr Ala Arg Arg His Pro Asp Tyr Ser Val Val Leu Leu Leu Arg Leu Ala
TTC CTG GGC ATG TTT TTG TAT GAA TAT GCA AGA AGG CAT CCT GAT TAC TCT GTC GTG CTG CTG CTG AGA CTT GCC

Lys Thr Tyr Glu Thr Thr Leu Glu Lys Cys Cys Ala Ala Ala Asp Pro His Glu Cys Tyr Ala Lys Val Phe Asp
AAG ACA TAT GAA ACC ACT CTA GAG AAG TGC TGT GCC GCT GCA GAT CCT CAT GAA TGC TAT GCC AAA GTG TTC GAT

                                                                                                                        400
Glu Phe Lys Pro Leu Val Glu Glu Pro Gln Asn Leu Ile Lys Gln Asn Cys Glu Leu Phe Lys Gln Leu Gly Glu
GAA TTT AAA CCT CTT GTG GAA GAG CCT CAG AAT TTA ATC AAA CAA AAC TGT GAG CTT TTT AAG CAG CTT GGA GAG

Tyr Lys Phe Gln Asn Ala Leu Leu Val Arg Tyr Thr Lys Lys Val Pro Gln Val Ser Thr Pro Thr Leu Val Glu
TAC AAA TTC CAG AAT GCG CTA TTA GTT CGT TAC ACC AAG AAA GTA CCC CAA GTG TCA ACT CCA ACT CTT GTA GAG

                                                                                                                        450
Val Ser Arg Asn Leu Gly Lys Val Gly Ser Lys Cys Cys Lys His Pro Glu Ala Lys Arg Met Pro Cys Ala Glu
GTC TCA AGA AAC CTA GGA AAA GTG GGC AGC AAA TGT TGT AAA CAT CCT GAA GCA AAA AGA ATG CCC TGT GCA GAA

Asp Tyr Leu Ser Val Val Leu Asn Gln Leu Cys Val Leu His Glu Lys Thr Pro Val Ser Asp Arg Val Thr Lys
GAC TAT CTA TCC GTG GTC CTG AAC CAG TTA TGT GTG TTG CAT GAG AAA ACG CCA GTA AGT GAC AGA GTC ACA AAA

                                                                                                                        500
Cys Cys Thr Glu Ser Leu Val Asn Arg Arg Pro Cys Phe Ser Ala Leu Glu Val Asp Glu Thr Tyr Val Pro Lys
TGC TGC ACA GAG TCC TTG GTG AAC AGG CGA CCA TGC TTT TCA GCT CTG GAA GTC GAT GAA ACA TAC GTT CCC AAA

Glu Phe Asn Ala Glu Thr Phe Thr Phe His Ala Asp Ile Cys Thr Leu Ser Glu Lys Glu Arg Gln Ile Lys Lys
GAG TTT AAT GCT GAA ACA TTC ACC TTC CAT GCA GAT ATA TGC ACA CTT TCT GAG AAG GAG AGA CAA ATC AAG AAA

                                                                                                                        550
Gln Thr Ala Leu Val Glu Leu Val Lys His Lys Pro Lys Ala Thr Lys Glu Gln Leu Lys Ala Val Met Asp Asp
CAA ACT GCA CTT GTT GAG CTT GTG AAA CAC AAG CCC AAG GCA ACA AAA GAG CAA CTG AAA GCT GTT ATG GAT GAT

Phe Ala Ala Phe Val Glu Lys Cys Cys Lys Ala Asp Asp Lys Glu Thr Cys Phe Ala Glu Glu Gly Lys Lys Leu
TTC GCA GCT TTT GTA GAG AAG TGC TGC AAG GCT GAC GAT AAG GAG ACC TGC TTT GCC GAG GAG GGT AAA AAA CTT

Val Ala Ala Ser Gln Ala Ala Leu Gly Leu End
GTT GCT GCA AGT CAA GCT GCC TTA GGC TTA TAA CATCTACATTTAAAAGCATCTCAGCCTACCATGAGAATAAGAGAAAGAAAATGAA

GATCAAAAGCTTATTCATCTGTTTTCTTTTTCGTTGGTGTAAAGCCAACACCCTGTCTAAAAAACATAAATTTCTTTAATCATTTTGCCTCTTTTCTCT

GTGCTTCAATTAATAAAAAATGGAAAGAATCTAATAGAGTGGTACAGCACTGTTATTTTTCAAAGATGTGTTGCTATCCTGAAAATTCTGTAGGTTCTG

TGGAAGTTCCAGTGTTCTCTCTTATTCCACTTCGGTAGAGGATTTCTAGTTTCTGTGGGCTAATTAAATAAATCACTAATACTCTTCTAAGTT Poly(A)
```

the approximately 4 kb fragment was gel purified. This was annealed and ligated to a 390 bp PstI, BglII partial digestion fragment purified from F-61 and a 1000 bp PstI fragment of B-44. Restriction endonuclease analysis of resulting transformants identified plasmids containing the entire HSA coding sequence properly aligned for direct expression of the mature protein. One such recombinant plasmid was designated pHSA1. When E. coli containing pHSA1 is grown in minimal media lacking tryptophan, the cells produce a protein which specifically reacts with HSA antibodies and comigrates with HSA in SDS polyacrylamide electrophoresis (Fig. 2). No such protein is produced by identical recombinants grown in rich tryptophan containing broth, implying that production in E. coli of the putative HSA protein is under control of the trp promoter-operator as designed. Efforts are now underway to increase the level of protein production in this and other expression systems above the present modest levels. To insure the integrity of the HSA structural gene in the recombinant plasmid, pHSA1 was subject to DNA sequence analysis.

DNA Sequence Analysis

The sequence of the albumin cDNA portion and the adjoining regions of pHSA1 were determined by both the chemical degradation method of Maxam and Gilbert (26) and the dideoxy chain termination procedure employing templates derived from single stranded M13 mP7 phage derivatives (32, 33). The DNA sequence is shown in Fig. 3 along with the predicted amino acid sequence of the HSA protein. The DNA sequence farther 5' to the mature HSA coding region was also determined from the cDNA clone P-14 and is included in Fig. 3.

DNA sequence analysis confirmed that the artifical initiation codon

Fig. 3. Nucleotide and Amino Acid Sequence of Human Serum Albumin.
The DNA sequence of the mature protein coding and 3' untranslated regions of HSA mRNA were determined from the recombinant plasmid pHSA1 and the DNA sequence of the prepro peptide coding and 5' untranslated regions were determined from the plasmid P-14 (see text). Predicted amino acids are included above the DNA sequence and are numbered from the first residue of the mature protein. The preceding 24 amino acids comprise the prepro peptide. The five amino acid residues which disagree with the protein sequence of HSA reported by both Dayhoff (9) and Moulon et al. (12) are underlined. The above nucleotide sequence probably does not extend to the true 5' terminus of HSA mRNA. In the albumin direct expression plasmid pHSA1, the mature protein coding region is immediately preceded by the E. coli trp promoter-operator-leader peptide ribosome binding site (36, 37), an artificial XbaI site, and an artificial initiation codon ATG; the prepro region has been excised. The nucleotides preceding HSA codon no. 1 in pHSA1 read 5'-TCACGTAAAAAGGGTATCTAGATG.

and the complete mature HSA coding sequence directly follows the E. coli trp promoter-operator as desired. The ATG initiator follows the putative E. coli ribosome binding sequence (36) of the trp leader peptide (37) by 9 nucleotides.

Translation of the DNA sequence of pHSA1 predicts a mature HSA protein of 585 amino acids. Various published protein sequences of HSA disagree at about 20 amino acids. Our sequence differs by eleven residues from Moulon et al. (12), and by 28 residues from that reported in the Atlas of Protein Sequence and Structure (9). The Atlas (9) sequence is credited as arising primarily from Behrens et al. (10) with some residues deriving from Moulon et al. (12). Most of these differences represent inversions of pairs of adjacent residues or glutamine-glutamic acid disagreements. Only at five of the 585 residues does our sequence differ from the residue reported in both the Atlas (9) and in Moulon et al. (12), and three of these five differences represent glutamine-glutamic acid interchanges. The five common differences are underlined in Figure 3. At all discrepant positions the nucleotide sequencing has been carefully rechecked and it is unlikely that DNA sequencing errors are the cause of these reported differences. The possibility of artifacts introduced by cDNA cloning cannot be ruled out. However, other likely explanations exist for the amino acid sequence differences among various reports. These include changes in amidation (affecting glutamine-glutamic acid discrimination) occurring either in vivo or during protein sequencing (38). Polymorphism in HSA proteins may also account for some differences; over twenty genetic variants of HSA have been detected by protein electrophoresis (39) but have not yet been analyzed at the amino acid sequence level. It is also worth noting that our predicted HSA protein sequence is 585 amino acids long, in agreement with reference (12) but not reference (9). The difference is accounted for by the deletion (in ref. 9) of one phenylalanine (Phe) residue in a Phe-Phe pair at amino acids 156-157.

As sequence information becomes available, it will be possible to compare albumin genes between different species and with the related α-fetoprotein. When compared to the DNA sequence of a rat serum albumin cDNA clone (16) the mature HSA sequence we report shares 74 percent homology at the nucleotide and 73 percent homology at the amino acid level. (The rat albumin protein is one amino acid shorter than HSA; the carboxy terminal residue of HSA is absent in the rat protein.) All 35 cysteine residues are located in identical positions in both proteins. The

predicted "prepro" peptide region of HSA shares 76 percent nucleotide and 75 percent amino acid homology with that reported from the rat cDNA clone (16). Interspecies sequence homology is reduced in the portion of the 3' untranslated region which can be compared (the published rat cDNA clone ends before the 3' mRNA terminus). The HSA cDNA contains the hexanucleotide AATAAA 28 nucleotides before the site of poly(A) addition. This is a common feature of eukaryotic mRNAs first noted by Proudfoot and Brownlee (40).

It has been postulated that albumin and α-fetoprotein genes arose from duplication of a common ancestral gene (5, 41, 42). The two structural gene sequences are 50 percent homolgous in rats (5). The complete human α-fetoprotein sequence is not yet available, so this comparison cannot yet be made for the human proteins. Further studies will focus on inter- and intra-species relationships of structure, function, and expression of this related gene pair.

ACKNOWLEDGEMENTS

We wish to thank Daniel Palermo for valuable participation in early phases of this project, Stephen Rowe for procuring tissue samples, Roberto Crea, Mark Vasser and coworkers for oligonucleotide synthesis, and Jeanne Arch and Alane Gray for assistance in manuscript and illustration preparation, respectively. This research was supported by Genentech, Inc.

REFERENCES

1. Rosenoer, V.M., Oratz, M., Rothschild, M.A. Eds. (1977) Albumin Structure, Function and Uses, Pergamon Press, Oxford.
2. Peters, T. (1977) Clin. Chem. (Winston-Salem, N.C.) 23, 5-12.
3. Ruoslahti, E. and Terry, W.D. (1976) Nature 260, 804-805.
4. Sala-Trepat, J.M., Dever, J., Sargent, T.D., Thomas, K., Sell, S. and Bonner, J. (1979) Biochemistry 18, 2167-2178.
5. Jagodzinski, L.L., Sargent, T.D., Yang, M., Glackin, C. and Bonner, J. (1981) Proc. Natl. Acad. Sci. USA 78, 3521-3525.
6. Ruoslahti, E. and Terry, W.D. (1976) Nature 260, 804-805.
7. Yachnin, S., Hsu, R., Heinrikson, R.L. and Miller, J.B. (1977) Biochim. Biophys. Acta 493, 418-428.
8. Aoyagi, Y., Ikenaka, T. and Ichida, F. (1977) Cancer Research 37, 3663-3667.
9. Dayhoff, M. (1978) Atlas of Protein Sequence and Structure, Vol. 5, Suppl. 3, p. 266, National Biomedical Research Foundation, Washington.
10. Behrens, P.Q., Spiekerman, A.M. and Brown, J.R. (1975) Fed. Proc. 34, 591.
11. Brown, J.R. (1977) in Rosenoer et al. (Ref. no. 1), pp. 27-52.
12. Meloun, B., Moravek, L. and Kostka, V. (1975) Febs Letters 58, 134-137.
13. Judah, J.O., Gamble, M., and Steadman, J.H. (1973) Biochem. J. 134, 1083-1091.

14. Russell, J.H. and Geller, D.M. (1973) Biochem. Biophys. Res. Commun. 55, 239-245.
15. MacGillivray, R.T., Chung, D.W. and Davie, E.W. (1979) Eur. J. Biochem. 98, 477-485.
16. Sargent, T.D., Yang, M. and Bonner, J. (1981) Proc. Natl. Acad. Sci. USA 78, 243-246.
17. Berger, S.L. and Birkenmeier, C.S. (1979) Biochemistry 18, 5143-5149.
18. Ullrich, A., Shine, J., Chirgwin, R., Pictet, R., Tischer, E., Rutter, W.J. and Goodman, H.M. (1977) Science 196, 1313-1315.
19. Goeddel, D.V., Yelverton, E., Ullrich, A., Heyneker, H.L., Miozzari, G., Holmes, W., Seeburg, P.H., Dull, T., May, L., Stebbing, N., Crea, R., Maeda, S., McCandliss, R., Sloma, A., Tabor, J.M., Gross, M., Familletti, P.C. and Pestka, S. (1980) Nature 287, 411-416.
20. Crea, R. and Horn, T. (1980) Nucleic Acids Res. 8, 2331-2348.
21. Goeddel, D.V., Heyneker, H.L., Hozumi, T., Arentzen, R., Itakura, K., Yansura, D.G., Ross, M.J., Miozzari, G., Crea, R. and Seeburg, P.H. (1979) Nature 281, 544-548.
22. Goeddel, D.V., Shepard, H.M., Yelverton, E., Leung, D. and Crea, R. (1980) Nucleic Acids Res. 8, 4057-4074.
23. Backman, K., Ptashne, M. and Gilbert, W. (1976) Proc. Natl. Acad. Sci. USA 73, 4174-4178.
24. Hershfield, V., Boyer, H.W., Yanofsky, C., Lovett, M.A. and Helinski, D.R. (1974) Proc. Natl. Acad. Sci. USA 71, 3455-3459.
25. Grunstein, M., and Hogness, D.S. (1975) Proc. Natl. Acad. Sci. USA 3961-3965.
26. Maxam, A.M. and Gilbert, W. (1980) Methods Enzymol. 65, 499-560.
27. Denhardt, D.T. (1966) Biochem. Biophys. Res. Commun. 23, 461-467.
28. Wallace, R.B., Johnson, M.J., Hirose, T., Miyake, T., Kawashima, E.H. and Itakura, K. (1981) Nucleic Acids Research 9, 879-893.
29. Blin, N. and Stafford, D.W. (1976) Nucleic Acids Res. 3, 2303-2308.
30. Birnboim, H.C. and Doly, J. (1979) Nucleic Acids Research 7, 1513-1523.
31. Lawn, R.M., Adelman, J., Franke, A.E., Houck, C.M., Gross, M., Najarian, R. and Goeddel, D.V. (1981) Nucleic Acids Research 9, 1045-1052.
32. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
33. Messing, J., Crea, R. and Seeburg, P.H. (1981) Nucleic Acids Res. 9, 309-321.
34. Kessler, S.W. (1976) J. Immunology 117, 1482-1490.
35. Laemmli, U.K. (1970) Nature 277, 680-685.
36. Shine, J. and Dalgarno, L. (1974) Proc. Natl. Acad. Sci. USA 71, 1342-1346.
37. Platt, T., Squires, C. and Yanofsky, C. (1976) J. Mol. Biol. 103, 411-420.
38. Robinson, A.B. and Rudd, C.J. (1974) in Current Topics in Cellular Regulation, Horecker, B.L. and Stadtman, E.R., Eds., Vol. 8, 247-295, Academic Press, New York.
39. Weitkamp, L.R., Salzano, F.M., Neel, J.V., Porta, F., Geerdink, R.A. and Tarnoky, A.L. (1973) Ann. Hum. Genet., Lond. 36, 381-391.
40. Proudfoot, N.J. and Brownlee, G.G. (1976) Nature 263, 211-214.
41. Gorin, M.B., Cooper, D.L., Eiferman, F., Van de Rijn, P. and Tilghman, S.M. (1981) J. Biol. Chem. 256, 1954-1959.
42. Kiousois, D., Eiferman, F., Van de Rijn, P., Gorin, M.B., Ingram, R.S. and Tilghman, S.M. (1981) J. Biol. Chem. 256, 1960-1967.