Sequence homology between RNAs encoding rat α -fetoprotein and rat serum albumin

(molecular evolution/amino acid sequence/cDNA/gene duplication)

LINDA L. JAGODZINSKI, THOMAS D. SARGENT*, MARIA YANG, CARLOTTA GLACKIN, AND JAMES BONNER

Division of Biology, California Institute of Technology, Pasadena, California 91125

Contributed by James F. Bonner, March 17, 1981

ABSTRACT We have determined the sequences of the recombinant DNA inserts of three bacterial plasmid cDNA clones containing most of the rat α -fetoprotein mRNA. The resultant nucleotide sequence of α -fetoprotein was exhaustively compared to the nucleotide sequence of the mRNA encoding rat serum albumin. These two mRNAs have extensive homology (50%) throughout and the same intron locations. The amino acid sequence of rat α -fetoprotein has been deduced from the nucleotide sequence, and its comparison to rat serum albumin's amino acid sequence reveals a 34% homology. The regularly spaced positions of the cysteines found in serum albumin are conserved in rat α -fetoprotein, indicating that these two proteins may have a similar secondary folding structure. These homologies indicate that α -fetoprotein and serum albumin were derived by duplication of a common ancestral gene and constitute a gene family.

 α -Fetoprotein (AFP) and serum albumin are major plasma proteins both synthesized in mammalian liver parenchymal cells. AFP is also produced in the embryonic yolk sac (1) and is a single chain glycoprotein with a molecular weight of 70,000 containing 4% carbohydrate (2–6). AFP synthesis is associated with developmental and neoplastic processes (7–9), and it is the dominant plasma protein of the mammalian fetus. After birth the concentration of AFP decreases to a trace level in the serum of healthy adult mammals (10, 11). On the other hand, production of serum albumin, a single chain polypeptide with a molecular weight of 66,000, increases severalfold after birth to become the predominant protein in the adult serum (7, 12, 13). This inverse relationship in expression of AFP and serum albumin and the physical similarities between these two proteins suggest that AFP may be the fetal analog of serum albumin.

The developmental alterations in the expression of AFP and serum albumin are of interest as an example of eukaryotic gene regulation (14). Analysis of the sequence and the structural organization of these two genes should elucidate the evolutionary relationship between AFP and serum albumin, and may also aid in understanding their regulation. The mRNA nucleotide sequence and the structural organization of the rat serum albumin gene have been established by sequence analysis (15-17). The amino acid sequence data available for AFP have been compared to those for serum albumin and a degree of internal homology has been shown to exist between these two proteins (18-20). The structural organizations of the mouse AFP and serum albumin genes have been estimated from electron micrographs of R-loops, and there is some similarity in this regard between the rat albumin gene and the mouse AFP and albumin genes (21, 22). The mouse AFP amino acid sequence has been deduced from the nucleotide sequence of its mRNA. A comparison of this sequence to that of human and bovine albumin revealed a 32% homology and regularly spaced cysteines (23). These homologies suggest that AFP and serum albumin are related, possibly derived by the duplication of a common ancestor. A comparison of rat AFP (RAFP) and rat serum albumin amino acid and nucleotide sequences and the evolutionary significance of their homologies are discussed.

METHODS

Cloning Procedures. Rat AFP mRNA was purified from Morris hepatoma 7777 (13). cDNA was synthesized from this template as described (13). The resultant cDNA had a numberaverage size of approximately 1000 nucleotides and contained a small amount of much longer material. The cDNA was rendered double stranded by sequential treatment with *Escherichia coli* DNA polymerase I and S1 nuclease (24) and was prepared for insertion into the plasmid pBR322 at the *Pst* I site (25). The mixture of DNA fragments was ligated and used to transform *E. coli* strain χ 1776 (26). The AFP clones were selected by use of the filter colony hybridization method of Grunstein and Hogness with ³²P-labeled AFP cDNA as a probe (27). cDNA clones with an insert size of 600 nucleotides or longer were selected for characterization. All clones were subsequently transferred to *E. coli* strain HB101 for growth (28).

Restriction Enzyme Mapping. Restriction endonucleases were obtained from Bethesda Research Laboratories (Rockville, MD) and New England BioLabs and used according to the manufacturers' instructions, with minor modifications. The digested DNA was analyzed by electrophoresis on 6% polyacrylamide or 1-1.5% agarose gels in 50 mM Tris-borate/1 mM EDTA, pH 8.3 (29).

Sequence Analysis. Plasmid cDNAs were freed of low molecular weight nucleic acid contaminants by exclusion from a column of Sepharose CL-2B, cleaved with an appropriate restriction endonuclease, dephosphorylated with bacterial alkaline phosphatase (Bethesda Research Laboratory), and labeled at the 5' ends with phage T4 kinase (Boehringer Mannheim) and $[\gamma^{-32}P]$ ATP. After digestion with a second restriction endonuclease, labeled DNA fragments were isolated from gels by a modified crush and soak method (30) and purified by chromatography on benzoylated DEAE-cellulose (Boehringer Mannheim) (16, 17). DNA sequence determination was done according to the procedures of Maxam and Gilbert with minor modifications (30). The products of the "G>A," "A>C," "C," and "C+T" reactions were electrophoresed on 0.4-mm-thick 8% acrylamide gels as described by Sanger and Coulson (31).

Statistical Analysis. The RAFP mRNA sequence data were divided into 80-nucleotide segments with overlaps of 10 nucleotides. Each segment was compared to the entire nucleotide

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

Abbreviations: AFP, α -fetoprotein; RAFP, rat AFP.

^{*} Present address: National Cancer Institute, National Institutes of Health, Bethesda, MD 20205.



FIG. 1. Restriction map of RAFP cDNA clones and the sequencing strategy. The restriction map of RAFP is shown below the scale in hundreds of nucleotides as measured from the presumed 5' terminus of the mRNA. The RAFP mRNA inserts of the three clones pRAF87, pRAF65, and pRAF6 are shown. The jagged line indicates the poly(A) tail region. The closed circles represent the restriction enzymes sites used in the sequence determinations: HI, *Hinf*I; HII, *Hpa*II; HIII, *Hind*III; B, *Bam*HI; T, *Taq*I; S, *Sal*I; PI, *Pst*I; PII, *Pvu*II. The direction and extent of the sequencing reactions are indicated by the arrows.

sequence of rat serum albumin by use of a computer program, SEQCMP, written by R. F. Murphy and J. W. Posakony (15). AFP segments of 40% or higher homology to rat serum albumin were identified in this manner. A rotary program, ROTCMP, was also used to compare these two DNA sequences (17). A single unambiguous alignment of these two sequences was thus obtained and is shown in Fig. 2.

RESULTS

Seven cDNA plasmid clones containing RAFP mRNA sequences were subjected to restriction endonuclease mapping analysis. Three overlapping RAFP clones (pRAF 87, pRAF 65, and pRAF 6) containing 85% of the total mRNA sequence were selected for further restriction enzyme mapping and DNA sequence determination. Fig. 1 illustrates the relationship of these three cDNA clones to RAFP mRNA. The sequencing strategy is represented by arrows in Fig. 1 along with the complete site map of each enzyme used in these experiments. pRAF 6 contains a stretch of 60 adenylate residues that presumably represent the 3' poly(A) tail of the mRNA. As indicated, much of the mRNA sequence has been read from both strands of the cloned DNA, and most of the determinations were repeated at least once. The sequence data presented in this paper are considered highly reliable. Certain sequences did not appear normally on the autoradiograms: the EcoRII site, C-C-A-G-G appeared as C-A-G-G, and the sequence C-C-G-T-T-C-G-A-A appeared as C-G-T-T-C-G-A-A. These artifacts were detected by sequencing both strands of the DNA.

The nucleotide sequence of cloned RAFP is presented in Fig. 2, along with the cDNA sequence of rat serum albumin (15). The two DNA sequences were aligned by introducing one 3-nucleotide gap into each to maximize their homology. A 50% overall homology is obtained by aligning rat serum albumin and RAFP as shown in Fig. 2, and no other alignment approaches this in significance. In no way could these two sequences have randomly obtained this level of homology. There is no evidence

for regions of nonhomology, rearrangements, or deletions in either mRNA other than the two triplets already mentioned.

The sequences for both mRNAs have been grouped into codon triplets and translated into the amino acid sequence shown in Fig. 2. To choose the correct reading frames for the rat serum albumin and RAFP mRNA sequences, it was assumed that no frameshifts or premature termination codons (TAA, TGA, TAG) occurred in either mRNA. The presence of multiple termination codons throughout two of the three possible reading frames made it possible to unambiguously infer the amino acid sequence for both of these proteins. The resultant amino acid homology between rat serum albumin and RAFP was 34% overall. Both genes have the same termination codon, TAA. Near the 3' end of each gene is the putative polyadenylylation signal sequence A-A-T-A-A which is located 145 nucleotides from the rat serum albumin termination codon and 110 nucleotides from RAFP's terminator (32). Benoist et al. (33) have identified another characteristic sequence located near the polyadenylylation site, T-T-T-C-A-C-T-G-C. A similar sequence, T-T-T-C-A-A-C-T-G-T, is found immediately to the left of the polyadenylylation site of the RAFP gene. In general, the untranslated portions of these two genes are more divergent than their translated portions, having only a 25% nucleotide sequence homology.

In the comparison of RAFP and rat serum albumin genes, two regions of amino acid nonhomology are found at positions 120–137 and 362–374. The DNA sequence homology of region 362–374 is the same as the overall homology (49%), whereas region 120–137 has only a 33% DNA homology.

DISCUSSION

The amino acid and nucleotide sequence homologies reported here strongly indicate that serum albumin and AFP genes are the product of the duplication of a common ancestral gene. In addition to the overall amino acid homology, almost all of the

FIG. 2 (on following page). Nucleotide sequence comparison. Except for approximately 300 uncloned nucleotides at the 5' end of the mRNA, the nucleotide sequence of the RAFP mRNA is shown. The inferred amino acid sequence of RAFP is also indicated, as are the nucleotide and amino acid sequences of rat serum albumin (ALB). Matching nucleotides between the two sequences are indicated by the dots. The matching amino acids are underlined. All cysteines are boxed. The vertical lines through each sequence represent the location of the introns in the respective genes. The upper-case letters designate the exons in the rat serum albumin gene (16, 17). Amino acids are numbered from the 5' end of the cloned RAFP mRNA. The dashes in the DNA represent the 3-nucleotide inserts. \$, End of rat serum albumin mRNA and start of poly(A) tail.

.

AI AI	LB J A FP ¥	95 AG ** AA	ser TCC asn AAC	ilu ATI ⊴ln CAG	hi CA 1e CT	s ti C Al U Se A Ti	nr 1 CT C er v CT G	eu TC al TG	Phe TTC Phe TTT	sly GGA leu CTG	<u>as</u> GAC <u>as</u> GA1	199 AAU SIGA	10 5 leo 5 TTo 1 ilo A AT		s al C GC S hi C CA	a il 5 AT 5 AT 5 AL 6 GA	u pr T CC u th G AC	o ly A AA Ar si G GA	is <u>le</u> 16 Cl 14 <u>le</u> 14 Cl		s as T GA T as T AA	ap as AC AA An 1s AC AA	20 In ty IC TA	r <u>91</u> C GG r <u>91</u> T GG	y 91 T GA Y	u le A CT - Ph - TT •	u al G GC e se C TC	a as T GA r gl A GG	PCY CTG YCY TG		ala GCA sasr CAAC	a lys AAA s sin CAA ••
AI AI	LB s C FP s A	ln AA er GC	slu GAG sly GGG • •	₽ro CCC val GTG	명 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1		is a GA A GA C	SD : AC (is : AC (slu GAG sln CAG ••	CYS TGT CYS TGT	Phe TTC lec CTC	2 <u>1ec</u> CTC <u>1ec</u> CTC	40 91 6 CAU 1 a1 6 GC1	n hi G CA B ar: CG	s <u>ly</u> C AA S <u>ly</u> C AA	s as G GA S 19 G AA	P as T GA s th G AC	r as C AA r al C GC		C AA	n C P se C TC •	- le - CT r va T GT		0 PT 0 CC 0 PT A CC		e sl C CA e hi C CA	n ar G AG s Ph C TT	s <u>Pr</u> G CC e <u>Pr</u> T CC	G GA G GA G GA G GA	u ala 6 GC u thi A AC	3 slu F GAG r ala F GCC r •	i ala GCC slu GAG •
AL AF	_B m A FPs A ●	et TG er GT	CYS TGC CYS TGC	thr ACC Pro CCA	se TC al GC •	r ph C TI a ty A TA •	ie si ic Ci ir si ic Gi	In g AG (Lu g AA (slu GAG slu GAA	asn AAC asn AAC	FTO CCT arg AGG	thr ACC ala GCG	AGC AGC Met ATC	Phi TT Set TC	e le T CT T il C AT +	u sl G GG u as A AA	y hi A CA n th C AC	s ty C TA r Ph G TT	r le T TT e il C AT	u hi G CA U ty C TA	s gl T GA r as C GA ••	u va A GT P va T GT	1 al T GC 1 se G TC	a ar C AG r 19 G AA •	ars AGA ars AGA AGC	a hi CA asi AA	5 <u>PT</u> CC 1 <u>PT</u> CC CC	o ty T TA D Phi C TTI	r Phi T TTI 2 lei C CTI	tyi TAI	ala GCC ala GCT	CCA CCA PTO CCC
AL AF	.Bs G Pt A	lu : AA (hr : CC (leu CTC ilu ATT	<u>1eu</u> CTT <u>1eu</u> CTT	ty TA TA	r ty C TA r le C TT	it <u>a</u> it <u>a</u> iu <u>a</u> iA <u>6</u>		slu GAG sla GCT	lys AAA gln CAG	tyr TAC tyr TAT	asn AAT asp GAC	510 GAG 195 AAG	GTI GTI ala GCA	l le r CT a va GT	u th G AC 1 pr T CC •	r sl C CA o al A GC	n <u>cy</u> G TG a <u>cy</u> A TG	C TG	s th C AC s ly C AA	r sil A GA s al A GC •	u se G TC a as T GA	r asi T GAI P asi C AAI • 140	⊳ 199 C AAA h met C AT(• •	ala GCA selu GAC •	8 814 6 GC(9 910 6 GA(•		s le CTC Phe TTC	u thi G ACA S Slr C CAC	FFC CCE thr GCCE	195 AAG 195 AAG	leu CTT ars AGA
AL AF	.Ba Gi Pa Gi	SP a AT (la s CA l	ala GCC Ser ICA	val GTG met ATG	ly AA al GC	5 91 A GA B 19 A AA	u 19 6 A4 5 9] 6 64	AA O Lu 1 AA T	ala GCA Leu TA	leu CTG ars AGA	val GTC slu GAA •	ala GCA sly GGA • •	ala GCT Ser AGC	val GTC met ATC	l ars CGI Leu CTC •	ILE E CA Jasi CAA C	n ar: G AGU n slu F GAU	s me G AT U hi G CA	t ly G AA s va T GT	5 <u>C4</u> 6 T6 1 <u>C4</u> 6 T6	s se TC al GC	r se C AG a va A GTi	r me T AT(1 il) A AT(t sir 3 CAC J ars A AGA	ars AGA 195 AAG •	<u>Phe</u> TTI <u>Phe</u> TTI	9 <u>919</u> 660 9 <u>919</u> 660	a slu A GAC a set A TCC	ars AGA ars CGC	ala GCC asn AAT	Phe TTC Leu CTC	lys AAA ⊴ln CAG ∙
AL AF	B a Gl F <u>a</u> Gl	La t CC T La V CT C	GG al TA	ala GCA leu CTC	val GTA ilt ATC	l al A GC J il C AT	a ar f CG u ly T AA •	IT A	et TG eu TA	ser AGC Ser AGT	<u>sln</u> CAG <u>sln</u> CAA	ars AGA 1ys AAG •	Phe TTC Phe TTC Phe TTT	013 013 013 013 013	asr AA1 159 AAA	ala GC ala GC ••	GAU GAU Basi AA1	J Pho G TN H ilo T AT	e al GC L th F AC	a <u>sl</u> A GA r <u>sl</u> T GA ••	110 ATC 110 ATC	u thi C ACC u ars T CGC	170 199 AAA 199 AAC	i <u>leu</u> TTG <u>leu</u> CTG	<u>ala</u> GCA <u>ala</u> GCC ••	thr ACA lec CTC		val GTT val GTC	thr ACC ala GCC	lys AAA his CAC	<u>ilu</u> ATC <u>ilu</u> ATC	asn AAC his CAC ••
al. Af	B 1 Af F sl Gf	46 6 46 6 46 0	1 u AG 1 n AG	CYS TGC CYS TGC		CA CA CA	5 <u>81</u> C GG 5 <u>81</u> C GG	y a G y C y C •	AC AC	leu CTG ala GCG	leu TTG met ATG	<u>slu</u> GAA <u>slu</u> GAG	CYS TGC CYS TGC	ala GCG leu CTA	asp GAT sln CAG	GAC GAC GAT	ars AGC Sly GGC	3 a1 6 GC 9 10 6 GA	a sl A GAN J Set A AGO	u lec A CTI n val C GTC •	ala GCC met CATC	a lya C AAC t thi G ACA •	200 s tyr 5 TAC • his • CAT	ATG ATG ATG	суз Тбт суз Тбт •••	slu GAG ser TCT	asr AAC slr CAA	sin CAG sin CAA	ala GCC ⊴lu GAA	thr ACT ilu ATT	ilu ATC leu CTG	ser TCC ser TCG
AL	B <u>se</u> AC AC		ys AA ys AA	leu CTG thr ACA	slr CAG ala GCA	al GC sl GA	B C Y T T G C Y G T G • •		ys GT ys GC •	BSP GAT LYS AAA I	lys AAG leu CTA	019 CCA 019 019 000	220 val GTG thr ACC	leu CTG ilu ATC	sln CAG slu GAG	lys AAA leu CTC	ser TCC sly GGC	slr CAC tyr TAT		ile CTC il ATA	ala GCT ilu ATT	s slu GAC his CAT	230 J thr ACA ala GCC	<u>910</u> GAA <u>910</u> GAA	his CAT asn AAT	asp GAC sly GGC	asn AAC asp GAC	ilu ATT lys AAA •		ala GCC slu GAA	asp GAT gly GGT •	leu CTG leu TTA
ALI AFI	B PT CC b th AC	0 5 C T T 1 T C	er : CA / eu a TG /	ilu ATA BSN AAC	ala GCT Pro CCA	ala GCI set AGC	3 85 GA 91 GA 04	P E C T C E F G T	he v TT (he 1 TT 1	val s GTT (Leu s TG (slu GAG sly GGA	asp GAT asp GAC	250 lys AAG arg AGA •	slu GAA asn AAT	val GTG Phe TTT	cys TGT ala GCC	lys AAG gln CAG	asr AAC Phe TTT	tyr TAT Ser TCT	ala GCT Ser TCA	<u>410</u> GAG <u>410</u> GAG	ala GCC slu GAA	260 <u>195</u> AAG <u>195</u> AAA	asp GAT leu CTC	val GTC leu CTG	Phe TTC Phe TTT	leu CTG met ATG	⊴ly GGC ala GCA ∙	thr ACE SET ACE	<u>Phe</u> TTT <u>Phe</u> TTT	<u>leu</u> TTG <u>leu</u> CTT	tyr TAT his CAT
ALI AFF	9 <u>91</u> GA 9 <u>91</u> GA			CA CA	ars AGA ars AGA	ars AGG asn AAT	CAU CAU CAU		ro CC G CC A	ISP 1 IAT 1 ISN 1 IAC C	AC AC EU CTT	ser TCC Pro CCT	80 <u>val</u> GTG <u>val</u> GTC	Ser TCC Ser TCA	leu CTG val GTC	leu CTG ilu ATT	leu CTG leu CTA	arg AGA 1ys AAA	leu CTT thr ACT	<u>ala</u> GCT <u>ala</u> GCT	195 AAG 195 AAG	lys AAA ser TCA	290 <u>tyr</u> TAT <u>tyr</u> TAC ••	slu GAA sln CAG	ala GCC slu GAA •	thr ACA ilu ATT •	<u>leu</u> CTG <u>leu</u> TTG	<u>slu</u> GAG <u>slu</u> GAG	<u>195</u> AAG <u>195</u> AAG	CYS TGC CYS TGC	cys 1GT Ser TCC	ala GCT sln CAG
ALB AFP	sl GA se TC	u si A GG r si T GA •	iy a GC G Lu t AA A	AT C AT C CG C	010 TO: 010 AO:	Pro CCT Ser TCG •	ala GCC 199 AAA		45 t 50 T 45 S 50 C	yr 9 AC G In a AG G •	GC SP AC	3 thr ACA asn AAT •	10 val GTG met ATG	leu CTT slu GAA	ala GCA slu GAA • •	<u>slu</u> GAA <u>slu</u> GAA	Phe TTT leu TTG	<u>sln</u> CAG <u>sln</u> CAA	PTO CCT 195 AAA	leu CTT his CAC ∙	val GTA ilu ATC	slu GAA sln CAG	320 <u><u><u><u></u></u><u><u></u><u></u><u></u><u><u></u><u></u><u></u><u></u><u></u><u></u><u></u><u></u><u></u><u></u><u></u><u></u></u></u></u></u>	۶ro CCT Ser AGC	lys AAG gln CAG	asn AAC ala GCA	<u>1eu</u> TTG <u>1eu</u> CTG	val GTC ala GCC	195 AAA 195 AAA	thr ACT Sln CAA	asn AAC ser AGT	CSS TGT CSS TGT
AL B	si GA asi AA	u <u>1</u> e 5 C1 5 <u>1e</u> 7 C1			31u GAG 31n CAG	<u>195</u> AAG <u>195</u> AAG	lec CTI <u>lec</u> TTA	00 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	A G A C	lu t AG T ro t CC T		3 919 GGA tyr TAC	40 Phe TTC leu TTA ••	aln CAA aln CAA	asn AAC asn AAT	ala GOC leu CTG	val GTT Phe TTC	<u>1eu</u> CTG <u>1eu</u> CTC	val GTT ilu ATT	arg CGA gly GGC	tyr TAC tyr TAC	thr ACC thr ACG	350 sln CAG ars CGG	<u>195</u> AAA 195 AAG	ala GCA ala GCT	019 019 019 019	<u>sln</u> CAG <u>sln</u> CAG	vəl GTG 1eu CTG ••	ser TCG thr ACC	thr ACC ser TCA	Pro CCA ala GCA	thr ACT slu GAG
ALB AFP	le CTC Le CTC	L Va C GT L 11 S AT	1 9 6 6 9 8 4 6	lu a AG (SP) AC (ela GCA Leu CTG	ala GCA thr ACA	arg AGA gly GGG	85 AA 19 AA	in 1 IC C IS MI IG A	eu s TG G et v TG G •••	ly GA al TG	3 ars AGA ser AGC	70 val GTG ilu ATT	aly GGC ala GCC •	thr ACC ser TCA	lys AAG thr ACG • •	<u>C95</u> TGT C95 TGC	<u>C95</u> TGT C95 TGC ••	thr ACC sln CAG	<u>leu</u> CTT <u>leu</u> CTC	Pro CCT Ser AGT	slu GAA slu GAG ••	380 ala GCT slu GAG •	sin CAG Ivs AAA	AGA AGA ars CGG	leu CTG ser TCC	₽ro CCC ala GCC ●●	суз ТСТ суз ТСС ••	və1 616 sly 66T ∙	<u>slu</u> GAA <u>slu</u> GAG	asr GAC ⊴1y 1 GGA (J TAT Leu CTG
ALB AFP	let CTC ala GCC	J Se G TC B as C GA	ra IG Fi CA	la i CC A lu t TT T	TC TC AC	leu CTG ilu ATT	asn AAC sly GGA	ar CG hi CA •	s <u>1</u> T C s <u>1</u> C T		ys GT ys GT	4 Val GTG (Leu a ITA (00 leu CTG ars AGA	LAT CAT CAT	<u>slu</u> GAG <u>slu</u> GAG	lys AAG ala GCA	thr ACC asn AAC		val GTG val GTG	ser AGC asn AAC • •	slu GAG ser TCC	lys AAG gly GGT	10 Val GTC ilu ATC	thr ACC asn AAC • •	lys AAG his CAC	CYS TGC CYS TGC	CYS TGT CVS TGC	ser AGT Ser AGT	dly GGG ser TCC	ser TCC ser TCG	leu v ITG (tyr s TAT 1	val STG Ser ICC
ALB AFP	slu GAA asr AAC	AG AG AG	a a Ci a Ai Ai Ai Ai	rs P 56 C rs 1 56 C	TO CA EU TC	сяг 161 160 160	Phe TTC ilu ATC	se TC th AC	ra] TG(rse CA(la 1 CT C Pr Pi GC T	eu f TG A he I TT C	4: thr v ACA (Leu a CTG 4	30 Val GTT BTS AGG	JAC JAC JAC	<u>slu</u> GAG slu GAA	thr ACA thr ACC	tyr TAT tyr TAC	val GTC val GTC	014 013 014 100	195 AAA PTO CCA •	slu GAG PTO CCA	Phe TTT Phe TTC ••	140 195 AAA Ser TCT	ala GCT slu GAA	slu GAG as⊧ GAC ●●	thr ACC 195 AAA •	Phe TTC Phe TTC	thr ACC ilu ATC • •	Phe TTC Phe TTC	his CAC his CAC	ser <u>a</u> FCT 0 lys <u>a</u> AAG 0	SAT SAT
ALB AFP	ilu ATC leu CTG	су ТС ТС ТС	s t.) C A(s s) C C/	nr 1 CA C In a NA G	eu TC 1a CT	۴۲۵ CCA sln CAG	asp GAC sly GGC	ly: AAC ars CGA	s sl G GA s al A GC	lu 19 AG A4 Ja 16 CC C1	45 5 46 0 90 5 1A 0	44 11n i 2AG 4 11n t AG 4	SO LLU I ATA A Shr M ACC A	AG AG TG	195 AAG 195 AAA	<u>eln</u> CAA eln CAA	KiL ACG slu GAG	ala GCT leu CTT	leu CTC leu CTC	ala GCT ilu ATT	⊴lu GAG asn AAC •	4 <u>1eu</u> CTG <u>1eu</u> CTG	70 <u>val</u> GTG val GTG	<u>195</u> AAA 195 AAA	his CAC MIN CAA	192 AAG 195 AAG	019 000 019 019	lys AAG glu GAA	ala GCC met ATG	thr ACA thr ACA	<u>110</u> a 3AA G 110 9 3AG G	iar iar iar
ALB AFP	sln CAG sln CAG	le CTC his CAC	u 19 3 AA 3 al 3 GC	/s t 16 A .a a 26 G	hr CG 1a C1	va1 GTG va1 GTC	met AIG thr ACT	s1: 661 a1: 601	I GA			49 CAC ers CTG	20 21n F 2AA 1 219 1 36C C	TC (TC) TC :	val GTG Leu TTG ••	asp GAC slu GAG	195 AAG 195 AAG	<u>C95</u> TGT C95 TGC	<u>Cys</u> TGC Cys TGC	<u>195</u> AAG <u>195</u> AAA	ala GCT asr GAC •	31a GCC sin CAG	00 <u>asp</u> GAC asp GAT	lys AAG (sln : CAG (BSP A GAT A Blu A GAA L	35N AC 313 3CC •	CYS 160 CYS 161	Fhe TTC Fhe TTC	ala GCC ala GCA	thr s ACT (Lys s AAA (N 66 19 61
AL B AF P	<u>FT0</u> CCA FT0 CCA	asr AA(199 AA() <u>le</u> ; CT ; <u>le</u> ; TT		al a fr (lu 9 ff -	ala GCT Ser FCC	arg AGA lys AAA • •	set AGC thr ACT	r ly C AA r ar C G	1 64 a al a 21 a 24 a 24 a 24 a 1 a 1 a 1 a 1 a 1 a 1 a 1 a 1 a 1 a 1		52 1a 1 CC T 1a 1 CT T	eu a TA G eu a TG G	1a 1 000 1 19 0 66 0	TAA TAA A Val T GTT T	ACAC TAA TAA	ATCA ACATI	CAAC CTCC	CATC AUGA	M TCAG GGAA	N GCTAI GAAAI	CCCT GGAC	GAGAI AAAAI	90000 9001	AGAC BIGTE	CATGI CGACI	AAGA GCTT	CTCA IGGT(GACT GTGAC	ICATO SCITT	тстт тсбб	ст тт

FIG. 2. (Legend appears at the bottom of the preceding page.)

cysteines present in rat serum albumin are also present at precisely the same location in RAFP, with the exception of cysteine 254 and 299 (Fig. 2). In serum albumin these two cysteines form a disulfide bond that results in a loop in the protein structure. Assuming that the other cysteine residues are crosslinked in the same pattern with both proteins, AFP would have a larger loop at this location than serum albumin. The near perfect conservation of the cysteines in RAFP and rat serum albumin is also observed in mouse AFP (23) and suggests that this feature is of some selective importance.

Amino acid sequence homologies between AFP and serum albumin have been observed by other investigators (18-20), although there appears to be no homology between their aminoterminal regions (34). A more thorough interspecies comparison of mouse AFP amino acid sequence with that of human and bovine serum albumin indicates that AFP and serum albumin are homologous throughout their entire lengths, except for the first 52 amino acid residues (23). A thorough comparison (intraspecies) of RAFP and rat serum albumin amino acid sequence (Fig. 2) reveals that there is in fact extensive homology at all regions of these two proteins, although there is somewhat less amino acid homology (30% vs. 39%) at the amino terminus than at the carboxyl terminus. In spite of the decreasing amino acid homology as one approaches the amino terminus of the proteins, the nucleotide sequence homology of the mRNAs remains constant throughout. Comparisons of rat and mouse AFP amino acid and nucleotide sequences (23) indicate that they are 85-87% homologous.

The structural organization of the mouse AFP and serum albumin genes as obtained by measurements of R-loops indicated that these two genes have a similar distribution of introns and exons (22). Assuming that serum albumin and AFP are indeed related and duplicated genes, a prediction can be made concerning the locations of the introns in the RAFP sequence. Exon sizes and intron locations have been established by sequence analysis of rat serum albumin genomic clones (16, 17) and are indicated by the vertical lines drawn through the rat serum albumin sequence (Fig. 2). The uppercase letters represent the exons as labeled by Sargent et al. (16, 17). We postulate that the locations of the introns in these two genes are similar. Six intron locations have been established in RAFP by sequence analysis of genomic clones and are indicated by vertical lines in Fig. 2 (unpublished results). Five of these splice sites align exactly with those of rat serum albumin. The sixth is slightly off due to its location in the untranslated region. This stability of intron locations is observed in other eukaryotic gene families (35–40). The greater conservation of the exons between AFP and serum albumin genes than that of their nucleotide sequence is further evidence of their duplication.

Serum albumin and AFP appear to have similar secondary protein folding structures and genomic organization of coding segments (22). The serum albumin protein has internal homologies from which Brown (41) proposed that this protein consists of three domains. This same triplet periodicity is reflected in the nucleotide sequence of the rat serum albumin mRNA and the sizes of the exons (16, 17). An internal comparison of the exon sequences of the rat serum albumin gene has led to a proposed model for the evolution of this gene (16, 17). The model predicts that a 5-exon ancestral gene evolved by a series of at least three intragenic duplication events into the 15-exon/14intron/three-domain ancestor of rat serum albumin (and RAFP). Subsequently, an intergenic duplication resulted in the appearance of serum albumin and AFP as distinct genes and proteins. The sequence and structural homologies reported here support this evolutionary model. Thus, AFP and serum albumin represent a gene family.

We thank Barbara Hough-Evans for helpful suggestions in the preparation of this manuscript. This work was supported in part by Grants 5 T32 GM 07616, and GM 13762, awarded by the National Institute of General Medical Sciences.

- Gitlin, D., Pernicelli, A. & Gitlin, G. M. (1971) Cancer Res. 32, 979–982.
- 2. Ruoslahti, E. & Seppala, M. (1971) Int. J. Cancer 7, 218-225.
- 3. Smith, C. J. & Kelleher, P. C. (1973) Biochim. Biophys. Acta 317, 231-242.
- Watanabe, A., Taketa, K. & Kosaka, K. (1975) Ann. N.Y. Acad. Sci. 259, 95-108.
- Sell, S., Jalowayski, I., Bellone, C. & Wepsic, H. T. (1972) Cancer Res. 32, 1184–1189.
- 6. Watabe, H. (1974) Int. J. Cancer 13, 377-389.
- 7. Abelev, G. I. (1974) Transplant. Rev. 20, 1-37
- Ruoslahti, E., Pihko, H. & Seppälä, M. (1974) Transplant. Rev. 20, 38–60.
- 9. Sell, S. & Becker, F. F. (1978) J. Natl. Cancer Inst. 60, 19-26.
- 10. Sell, S. & Gord, D. R. (1973) Immunochemistry 10, 439-442.
- Masseyeff, R., Gilli, J., Krebs, B., Bonet, C. & Zrihen, H. (1974) Biomedicine (Paris) 21, 353–357.
- 12. Van Furth, R. & Adinolfi, M. (1969) Nature (London) 222, 1296–1299.
- Sala-Trepat, J. M., Dever, J., Sargent, T. D., Thomas, K., Sell, S. & Bonner, J. (1979) *Biochemistry* 18, 2167–2178.
- Liao, W. S. L., Conn, A. R. & Taylor, J. M. (1980) J. Biol. Chem. 255, 10036–10039.
- Sargent, T. D., Yang, M. & Bonner, J. (1981) Proc. Natl. Acad. Sci. USA 78, 243–246.
- Sargent, T. D., Jagodzinski, L. J., Yang, M. & Bonner, J. (1981) Mol. Cell. Biol., in press.
- 17. Sargent, T. D. (1981) Dissertation (Calif. Inst. Tech., Pasadena, CA).
- 18. Ruoslahti, E. & Terry, W. D. (1976) Nature (London) 260, 804–805.
- Liao, W. S. L., Hamilton, R. W. & Taylor, J. M. (1980) J. Biol. Chem. 255, 8046-8049.
- 20. Innis, M. A. & Miller, D. L. (1980) J. Biol. Chem. 255, 8994-8996.
- Gorin, M. B. & Tilghman, S. M. (1980) Proc. Natl. Acad. Sci. USA 77, 1351–1355.
- Kioussis, D., Eiferman, F., Van de Rijn, P., Gorin, M., Ingram, R. S. & Tilghman, S. M. (1981) J. Biol. Chem. 256, 1960–1967.
- Gorin, M. B., Cooper, D. L., Eiferman, F., Van de Rijn, P. & Tilghman, S. M. (1981) J. Biol. Chem. 256, 1954–1959.
- Higuchi, R., Paddock, G. V., Wall, R. & Salser, W. (1976) Proc. Natl. Acad. Sci. USA 73, 3146–3150.
- 25. Roychoudhury, R., Jay, E. & Wu, R. (1976) Nucleic Acids Res. 3, 101-116.
- Curtiss, R., III, Pereira, D. A., Hsu, J. C., Hull, S. C., Clarke, J. E., Maturin, L. J., Sr., Goldsmith, R., Moody, R., Inoue, M. & Alexander, L. (1977) in *Proceedings of the 10 Miles International Symposium*, eds. Beers, R. F., Jr. & Bassett, E. G. (Raven, New York), pp. 45-56.
- Grunstein, M. & Hogness, D. S. (1975) Proc. Natl. Acad. Sci. USA 72, 3961–3965.
- Kushner, S. R. (1978) in Proceedings of the International Symposium on Genetic Engineering, eds. Boyer, H. W. & Nicosia, S. (Elsevier/North-Holland, New York), pp. 17-23.
- 29. Maniatis, T., Jeffrey, A. & Van de Sande, H. (1975) Biochemistry 14, 3787-3794.
- Maxam, A. M. & Gilbert, W. (1980) Methods Enzymol. 65, 499-560.
- 31. Sanger, F. & Coulson, R. (1978) FEBS Lett. 87, 107-110.
- 32. Proudfoot, N. J. & Brownlee, G. G. (1976) Nature (London) 263, 211-214.
- Benoist, C., O'Hare, Breathnach, R. & Chambon, P. (1980) Nucleic Acids Res. 8, 127–142.
- Peters, E. H., Nishi, S. & Tamaoki, T. (1978) Biochem. Biophys. Res. Commun. 83, 75–82.
- Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., DeRiel, J. K., Forget, B. G., Weissman, S. M., Slighton, J. L., Blechi, A. E., Smithies, O., Baralle, F. E., Shoulder, C. C. & Proudfoot, N. J. (1980) Cell 21, 653-668.

Biochemistry: Jagodzinski et al.

.

- Nishioka, Y. & Leder, P. (1979) Cell 18, 875–882.
 Leder, A., Miller, H. I., Hamer, D. H., Seidman, J. G., Norman, B., Sullivan, M. & Leder, P. (1978) Proc. Natl. Acad. Sci. USA 75, 6187–6191.
- Tiemeier, D. C., Tilghman, S. M., Polsky, F. I., Seidman, J. G., Leder, A., Edgell, M. H. & Leder, P. (1978) Cell 14, 237–245.
- 39. Konkel, D. A., Maizel, J. V., Jr. & Leder, P. (1979) Cell 18, 865-873.
- Perler, F., Efstradiatis, A., Lomedico, P., Gilbert, W., Kolodner, R. & Dodgson, J. (1980) Cell 20, 555-565.
 Brown, J. R. (1976) Fed. Proc. Fed. Am. Soc. Exp. Biol. 35,
- 2141-2144.