MOLECULAR AND CELLULAR BIOLOGY, Oct. 1981, p. 871-883 Vol. 1, No. 10
0270-7306/81/100871-13$02.00/0

Fine Structure and Evolution of the Rat Serum Albumin Gene

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

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

Received 3 March 1981/Accepted 30 June 1981

The exons, their boundaries, and approximately half of the intronic deoxyribonucleic acid of the rat serum albumin gene were sequenced. In addition to the 14 exons identified earlier by R-loop analysis, a small exon was detected between the "leader" exon (Z) and exon B. The leader exon encoded the 5′-untranslated portion of albumin messenger ribonucleic acid and the "pre-pro" oligopeptide present on the nascent protein. The sites of initiation and termination of transcription were tentatively identified by comparison of the 5′ and 3′ gene-flanking sequences with those of other eucaryotic genes. All 28 intron/exon junctions conformed to the "GT-AG rule" (Breathnach et al., Proc. Natl. Acad. Sci. **75:** 4853-4857, 1978). The three homologous domains of albumin were encoded by three subgenes that consisted of four exons each and evolved by intragenic duplication of a common ancestor. The second and forth exons of each subgene appeared to be the result of an even earlier duplication event. We propose a model for the evolution of this gene that accounts for the observed patterns of exon size and homology.

The typical eucaryotic genome contains on the order of 10,000 structural genes. Although usually considered to be single copy, many of these genes, possibly all of them, are organized into families of sequences related by homology and sometimes also by function. These gene families have arisen by a long process of sequence duplication and mutational divergence of a relatively small number of ancestral precursors.

When the boundaries of a duplication event fall outside of the transcriptional initiation and termination signals (intergenic duplication), the result is the creation of a new gene. When the duplication boundaries fall within a transcription unit, an intragenic duplication has occurred, and a single gene has been expanded into a larger, internally redundant one. Analysis of protein sequence data has revealed many examples of this latter phenomenon (1). It has also been demonstrated at the nucleic acid level for the mouse immunoglobulin heavy-chain constant region (41), the chicken ovomucoid gene (38), and rat serum albumin (35). These proteins have periodic homology that constitutes portions of their amino acid sequence, the basic repeating unit of which represents the vestige of duplicated ancestral sequence. It is instructive to consider such genes as a special variety of gene family whose members are fused into a single transcription unit and encode a single protein rather than a family of smaller polypeptides.

As mutations become fixed in a gene family, its members grow less homologous. When their messenger ribonucleic acid (mRNA) homology falls sufficiently low, the genes will become operationally single copy, as they will no longer be capable of cross-reaction in a molecular hybridization. Beyond this point, protein sequence analysis is needed to demonstrate relatedness. Direct analysis of the actual genetic deoxyribonucleic acid (DNA) sequence is an ideal measure of homology because related genes can retain statistically significant similarity even after their proteins appear to be unrelated. Furthermore, certain features of eucaryotic genes, particularly the pattern of interruptions by introns, are often very rigidly conserved, and these can be helpful in identifying extremely divergent gene families.

We used nucleotide sequence analysis of the rat serum albumin gene and its mRNA to elucidate the structural details and evolutionary history of this protein. We identified a "leader" exon at the 5′ end of the albumin gene that encodes both the "signal" oligopeptide present on nascent albumin and most or all of the 5′ untranslated portion of the mRNA. The sequences immediately flanking the albumin gene were compared with the equivalent regions of other genes, and possible sites for the initiation and termination of albumin gene transcription were thereby identified. The gene duplication events that are evident from the internal ho-

† Present address: Laboratory of Biochemistry, National Cancer Institute, National Institutes of Health, Bethesda, MD 20205.

mology of the mRNA and protein have been quite clearly preserved in the intron/exon structure.

MATERIALS AND METHODS

Materials. Restriction endonucleases were purchased from New England Biolabs, Bethesda Research Laboratories, or Boehringer Mannheim Corp. Bacterial alkaline phosphatase was purchased from New England Biolabs. T4 kinase was purchased from Boehringer Mannheim. All enzymes were used according to the manufacturer's instructions. [γ-^{32}P]adenosine triphosphate was purchased as a crude, carrier-free aqueous solution from ICN, and used within 24 h of receipt.

Clones. The rat serum albumin genomic clones were isolated from a library of rat liver genomic DNA as described previously (34). Most of the nucleotide sequence data and restriction endonuclease site map data were derived from "subclones" of these genomic clones. The *Eco*RI fragments designated A, B, C, D, E, and F (Fig. 1) were isolated from agarose gels and ligated to vector DNA, pBR325 (5) cleaved with *Eco*RI. The J:I ratio (13) was approximately 1.0. *Eco*RI fragment J was cleaved with *Hin*dIII, which generates three fragments. The two *Hin*dIII-*Eco*RI fragments were ligated with an equimolar amount of the vector, the larger *Hin*dIII-*Eco*RI fragment of pBR325. *Escherichia coli* strain HB101 was transformed according to the method of Kushner (22). All operations were performed in accordance with National Institutes of Health guidelines for experiments involving recombinant DNA.

Sequencing. Subclone plasmid DNA (or recombinant lambda phage DNA in the case of exons C and I) was freed from low-molecular-weight nucleic acid contaminants by exclusion from Sepharose CL2B, cleaved with an appropriate restriction endonuclease, dephosphorylated with bacterial alkaline phosphatase, and labeled at the 5′ ends with T4 kinase and [γ-^{32}P]-adenosine triphosphate. After digestion with a second restriction endonuclease, labeled DNA fragments were isolated from agarose or acrylamide gels and purified of soluble contaminants as follows. The DNA, eluted into 0.1 M NaCl-10 mM tris(hydroxymethyl)aminomethane (Tris) (pH 7.4)-1 mM ethylenediaminetetraacetic acid-0.1% sodium dodecyl sulfate, was bound to a 0.1-ml column of benzoylated diethylaminoethylcellulose (Boehringer Mannheim). This column was then washed with 3 to 5 ml of 0.2 M NaCl-10 mM Tris (pH 7.4), which removed essentially all non-nucleic acid contaminants. The pure DNA was eluted with 200 to 300 μl of 0.6 M NaCl-10 mM Tris (pH 7.4)-20% (vol/vol) ethanol and precipitated by addition of an equal volume of isopropyl alcohol and freezing in crushed dry ice for at least 10 min. DNA was sequenced according to minor modifications of the methods of Maxam and Gilbert (24). The products of the G>A, A>C, C, and C+T reactions were electrophoresed on 0.4-mm 8% acrylamide gels (33). Up to 450 nucleotides could be read from a single labeled site. Figure 1 shows the regions of the albumin gene that were sequenced.

RESULTS

Identification of intron/exon junctions and putative capping and polyadenylation sites by DNA sequence analysis. Figure 1 shows a revised map of the rat serum albumin gene. The sizes of the introns and exons are presented in Table 1. Exons were identified and their sizes were inferred by comparison of genomic clone sequence data with the nucleotide sequence of the albumin complementary (cDNA) clones (35). Most intron sizes were estimated by electrophoretic mobility on agarose or acrylamide gels and were accurate to approximately 5%. Introns CD, JK, and LM have been completely sequenced, so their precise sizes are known. This map is very similar to one presented previously (34), which was based on electron microscopic analysis of R-loops. There are two significant differences. The exon presently designated "A" was not detected by the R-loop experiments, presumably due to its small size and proximity to the end of the restriction fragment used in the hybridization reaction. We renamed the first exon "Z" after establishing the correct structure of the 5′ end of the albumin gene by sequence analysis. Also, the width of exon C was erroneously measured as 95 base pairs, which is less than half its actual size. This latter discrepancy is difficult to explain, and was

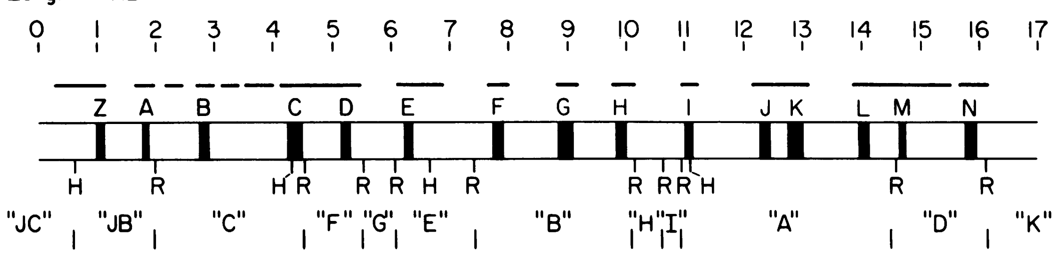

FIG. 1. *Map of the rat serum albumin gene. Black vertical bars denote exons. The horizontal bars indicate the regions of the cloned gene that have been sequenced to date. H, HindIII; R, EcoRI. The letters in quotation marks are the names of the restriction fragments that were subcloned and sequenced.*

TABLE 1. *Exon and intron sizes and locations in protein*

Exon		Intron			
Designation	Size[a]	Designation	Size[b]	Location in protein	Location in codon[c]
Z	(105)	ZA	697	*his-27*	1
A	58	AB	890	*leu-46*	2
B	133	BC	1364	*ile-90/his-91*	3
C	212	CD	809	*his-161*	2
D	133	DE	938	*lys-205/leu-206*	3
E	98	EF	1422	*trp-238*	2
F	130	FG	978	*arg-281/ala-282*	3
G	215	GH	779	*thr-353*	2
H	133	HI	1042	*val-397/leu-398*	3
I	98	IJ	1182	*ala-430*	2
J	139	JK	327	*tyr-476/leu-477*	3
K	224	KL	997	*thr-451*	2
L	133	LM	556	*glu-595/gly-596*	3
M	62	MN	1048	Untranslated	
N	(140)				

[a] The sizes of exons Z and N depend upon the location of the capping and polyadenylation sites, respectively. The approximate values given in parentheses correspond to the assignments for these shown in Fig. 2. The sizes of other exons are ambiguous by as many as six nucleotides due to terminal redundancy of introns. The values given result from the assumption that splicing conforms to the GT-AG rule.

[b] Intron sizes (except CD, JK, and LM) are estimated from the electrophoretic mobility of restriction fragments and are accurate to approximately 5%. Introns CD, JK, and LM have been completely sequenced.

[c] 1, Intron falls between the first and second nucleotides of the codon corresponding to the amino acid given in column 5; 2, it falls between the second and third; 3, it falls between codons.

unfortunate since it temporarily obscured the threefold substructure of the albumin gene. Of the 15,000 base pairs of this gene, a total of approximately 8,000 were sequenced, including all of the exons and their boundaries and over 6,000 nucleotides of intronic DNA. Although it has been possible to identify the approximate beginning and end of the albumin gene by R-loop measurements and blot hybridizations (34), the exact location of the capping and polyadenylation sites are not known due to the failure of the 5′ and 3′ extremities of the albumin mRNA to appear in any of the cDNA clones. Comparison of the 5′- and 3′-flanking sequences to those of other well-characterized eucaryotic genes reveals certain homologies that suggest locations for the termini of the albumin transcription unit.

At the 5′ end of exon Z, the sequences CCAAT and TATATT were found −120 and −65, respectively, from the ATG translation initiation codon. These are probably variants of similar sequences found about 80 and 30 nucleotides, respectively, upstream from the capping sites of most eucaryotic genes for which data are available (14). On this basis, the most likely capping site of the albumin gene is one of the A residues in the region indicated in Fig. 2. This assignment would predict that the distance from the cap to the second *Hin*dIII site of the albumin mRNA is about 650 nucleotides. We estimated this distance to be 600+/−20 nucleotides by alkaline electrophoresis of a cDNA preparation extended from a restriction fragment primer hybridized to the mRNA (T. Sargent, unpublished data). At the 3′ end of the albumin gene, the putative polyadenylation signal sequence, AATAAA, was located 145 nucleotides from the termination codon (121 nucleotides from the beginning of exon N). The 3′ end of the albumin gene is probably approximately 17 nucleotides downstream from this hexanucleotide. Benoist et al. (2) have identified another characteristic sequence located near the polyadenylation site; the consensus from several mRNA's is TTTTCACTGC. A similar sequence, TTTTCTCTGT, was located 19 nucleotides upstream from the AATAAA in albumin mRNA. Figure 2 also indicates the approximate 3′ end of exon N, as determined by R-loop measurements, and the 3′ end of the cDNA clone sequence. If the gene termini are in fact at the proposed position, this would give a total length of approximately 2,030 nucleotides for the non-polyadenylated portion of rat serum albumin mRNA, which is consistent with our earlier measurements (32).

The intron/exon junction sequences of a large number of genes have been determined (23, 36). The consensus sequences for 5′ and 3′ intron boundaries are (A/C)AG-GTAAGT and TYTYYYTCAG-G, respectively (Y = T or C). It is

```
Exon Z
                  -120
ATTTTGTAATGGGGTAGGAACCAATGAAATGAAAGGTTAGTGTGGTTAATGACCTACAGT

              -65                        -38
TATTGGTTAGAGAAGTATATTAGAGCGAGTTTCTCTGCACACAGACCACCTTTCCTGTCA
                                          cap site?
                  -1+1
ACCCCACTGCCTCTGGCACAATGAAGTGGGTAACCTTTCTCCTCCTCCTCTTCATCTCCG
                    Met

GTTCTGCCTTTTCTAGGGGTGTGTTTCGCCGAGAAGCACGTAAGCTAGGTA
                                         intron ZA


Exon N

TTTCAAGGCTACCCTGAGAAAAAAAGACATGAAGACTCAGGACTCATCTCTTCTGTTGGT
intron MN
                                                           *
GTAAAACCAACACCCTAAGGAACACAAATTTCTTTGAACATTTGACTTCTTTTCTCTGTG
                                                  TTTTCACTGC
       †                  ⁑
CCGCAATTAATAAAAAATGGAAGGAATATACTCTGTGGTTCGGAGGTCTGTCTTCCAACG
                             polyA site ?

GCGCGTCTCACCCTGGCGGGCTCTAGGGCTGGGGGAAACCCTCGGTTTCCTCCCTTCATC
```

FIG. 2. *Terminal sequences of the rat serum albumin gene. Exon Z includes the hydrophobic signal peptide sequence (nucleotides 16 to 36), the methionine initiation codon, and the CCAAT and TATATTA sequences associated with the capping regions of other eucaryotic genes (see text). The location of the putative capping site is indicated. Exon N consists of most of the 3' noncoding region of albumin mRNA and the putative polyadenylation or termination site (⁑), which is usually situated approximately 17 nucleotides downstream from the sequence AATAAA (2). Another typical sequence, similar to TTTTCACTGC, is often found downstream from AATAAA. It is found upstream from this in exon N. (*) the end of the cloned cDNA sequence; (†) the approximate 3' end of exon N according to R-loop measurements.*

almost always possible to define the splice junctions of an intron so that the first two nucleotides are GT and the last two AG [the GT-AG rule; 7]. All of the albumin gene intron/exon junctions were similar to these models. Of the 14 introns, 6 had unambiguous splice junctions. The remaining eight had up to four potential splice sites due to a small amount of terminal redundancy in the intronic sequence. All six defined introns conformed to the GT-AG rule, and all of the eight ambiguous introns could be construed to do so (Fig. 3).

Albumin introns contain simple sequences that are repeated elsewhere in the genome. Intron CD contained an interrupted palindrome, located 200 nucleotides from exon C: $(GT)_{55}$ followed by 83 nucleotides of complex sequence and then the complementary polydinucleotide, $(AC)_{85}$. Both polydinucleotide stretches had a few variant positions. The sequence $(CT)_{19}(GT)_{17}$ occurred 43 nucleotides to the left of exon E, in intron DE. These regions of the cloned albumin gene result in smears of intense radioactivity when hybridized to Southern blots of rat genomic DNA (T. D. Sargent, Ph.D thesis, California Institute of Technology, Pasadena, 1981), and thus presumably contain reiterated sequences. Intron EF gives similar, although much weaker, results. The polyadenylate-cytidylate common to introns CD and DE also appears to be reiterated in the mouse genome (26), and it is present in an intron in the rat prolactin gene (20) and in repeated elements flanking the rat serum albumin gene (Sargent, unpublished data). Oligo cytidylate-thymidylate has been found in sea urchin histone gene spacer regions (40). Intron EF contains only one tract of simple sequence, 28 T residues located 400 nucleotides from exon E. In addition, there are five tandem repeats of AAAAC plus several slightly mutated copies located 126 nucleotides into intron FG. As determined by hybridization of labeled genomic subclone DNA to rat genomic DNA blots, the rat serum albumin exons and introns are single copy except for the repetitive regions mentioned above (Sargent, Ph.D thesis, 1981). The guanine-plus-cytosine content of albumin introns (43%) is significantly lower than the exonic level of 50%. The value for the whole rat genome is 42%, calculated from the melting temperature of rat DNA (8, 42). There is a two- to threefold underrepresentation of the dinucleotide CG in albumin introns and exons, as has been observed in other eucaryotic DNA sequences (10).

Albumin gene consists of three homologous segments. The internal periodicity of the albumin gene has been demonstrated by statistical analysis of its mRNA sequence (35). The divisions in the mRNA sequence proposed on this basis corresponded to the boundaries between exons D and E and between exons H and I. Thus, the albumin gene was divided into three homologous subgenes (Fig. 4A). Each contained four exons and corresponded to a structural domain of the albumin polypeptide (9). In addition, there was a leader exon, Z, and a 3′ untranslated exon, N, which did not fit into the threefold pattern of homology. The small, partially translated exon M may be the remnant of a very early duplication event in albumin evolution (see Discussion and Fig. 6b). Except for exon A, it appears that the corresponding exons of these subgenes have remained remarkably similar in size. There were a total of 19 positions where the same amino acid was present in all three subgenes. Of these, 10 were cysteine residues. The remaining nine were highly conserved in rat, human, and bovine serum albumins (24 of 27 possible triple matches). In general, the amino acid sequences encoded by the three rat albumin

```
                                ↓
exon Z(3')        GCCGAGAAGCACGTAAGCTAGGTA
exon A(5')        CCATTCCCACAGACAAGAGTGAGA
     cDNA         GCCGAGAAGCACACAAGAGTGAGA
                                ↓
exon A(3')        TTTCAAAGGCCTGTAAGTTAAGAG
exon B(5')        CCTGTCTTTCAGAGTCCTGATTGC
     cDNA         TTTCAAAGGCCTAGTCCTGATTGC
                                ↓
exon B(3')        GACAAGTCCATTGTGAGTACATTC
exon C(5')        TCTTCCACTTAGCACACTCTCTTC
     cDNA         GACAAGTCCATTCACACTCTCTTC
                                ↓
exon C(3')        CTTTCTGGGACAGTGAGTACCCAG
exon D(5')        CCCATAATTCAGCTATTTGCATGA
     cDNA         CTTTCTGGGACACTATTTGCATGA
                                ↓
exon D(3')        CTGACACCGAAGGTAATCCCTGGA
exon E(5')        TTCTTTTGGTAGCTTGATGCCCTG
     cDNA         CTGACACCGAAGCTTGATGCCCTG
                                ↓
exon E(3')        CTTCAAAGCCTGGTATATGAATTT
exon F(5')        TTCCTTTTTCAGGGCAGTAGCTCG
     cDNA         CTTCAAAGCCTGGGCAGTAGCTCG
                                ↓
exon F(3')        GCGGATGACAGGGTAAAGAGGGGG
exon G(5')        CCATTCTCACAGGCAGAACTTGCC
     cDNA         GCGGATGACAGGGCAGAACTTGCC
                                ↓
exon G(3')        CTTCCTGGGCACGTGAGTAGATGC
exon H(5')        CGCCTCAATTAGGTTTTTGTATGA
     cDNA         CTTCCTGGGCACGTTTTTGTATGA
                                ↓
exon H(3')        TACGGCACAGTGGTAGGTTTCCGC
exon I(5')        TTTATCTTGCAGCTTGCAGAATTT
     cDNA         TACGGCACAGTGCTTGCAGAATTT
                                ↓
exon I(3')        ATTCCAAAACGCGTGAGAGTTTTT
exon J(5')        TTTGTTACACAGCGTTCTGGTTCG
     cDNA         ATTCCAAAACGCCGTTCTGGTTCG
                                ↓
exon J(3')        GTGGAAGACTATGTGAGTCTTTTA
exon K(5')        TCTCTTCTTTAGCTGTCTGCCATC
     cDNA         GTGGAAGACTATCTGTCTGCCATC
                                ↓
exon K(3')        AAAGAAGCAAACGTGAGGATATAT
exon L(5')        GTCCTGCTGCAGGGCTCTCGCTGA
     cDNA         AAAGAAGCAAACGGCTCTCGCTGA
                                ↓
exon L(3')        TTCGCCACTGAGGTAACAAATGTC
exon M(5')        TTTCCTGTTCAGGGGCCAAACCTT
     cDNA         TTCGCCACTGAGGGGCCAAACCTT
                                ↓
exon M(3')        CAACCATCTCAGGTAACTATACTC
exon N(5')        TGTGTTTTCAAGGCTACCCTGAGA
     cDNA         CAACCATCTCAGGCTACCCTGAGA


Rat serum albumin consensus:

                           A     A
exon 3'           *********C A*GTGAGTA*****
                             G
exon 5'           **TYTYYYTCAGC*T*********


"Universal" consensus:

                           A
exon 3'           *********CAGGTAAGT******
exon 5'           **TYTYYYTCAGG***********
```

FIG. 3. *Splice junctions. The terminal nucleotide sequences of exons and the relevant cDNA sequences are listed. Genomic DNA homology to the cDNA is underlined, and the overlaps represent potential splice sites. The arrows indicate sites that conform to the GT-AG rule. Consensus sequences are presented for albumin splicing configurations and for other eucaryotic genes (23).*

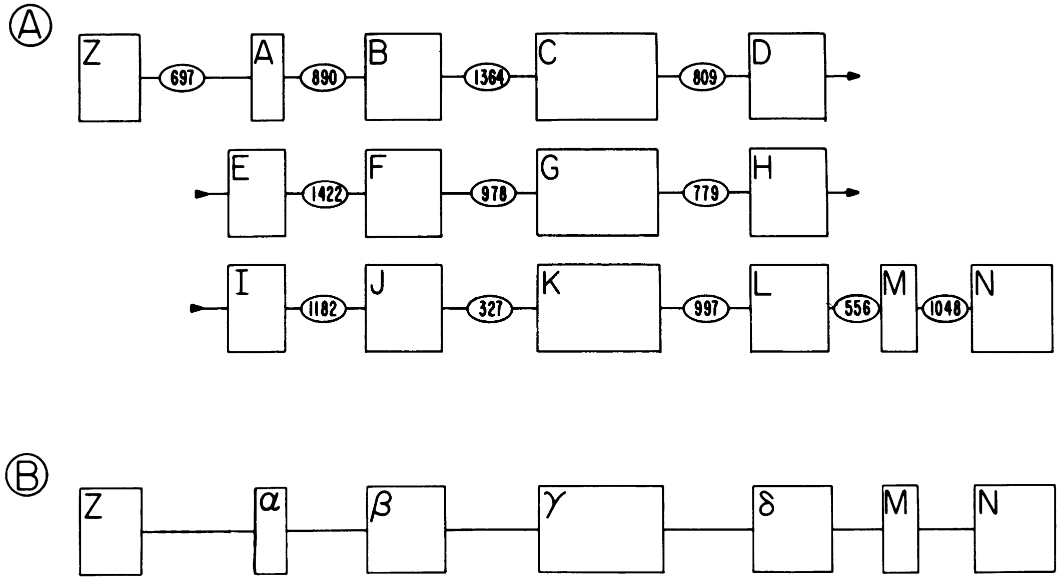

FIG. 4. *(A) Divisions in the albumin gene. The three albumin subgenes are illustrated. The boundaries corresponding to introns DE and HI were predicted from the internal homology of the cDNA sequence (Sargent, Ph.D thesis, 1981). The exons are drawn to scale, and intron sizes are specified in the ellipses. (B) The proto-albumin gene. Our model of the immediate evolutionary precursor to the albumin gene consists of the leader exon Z and 3′ exons M and N plus four exons, α, β, γ, and δ, that are equivalent to exons A, B/F/J, C/G/K, and D/H/L, respectively. Intron sizes are arbitrary.*

subgenes have diverged greatly. The average interdomain amino acid homology was only 20%. Figure 5 shows the best alignment of the polypeptides encoded by the various exons. Table 2 presents a summary of the internal homology.

We found no evidence for conservation of intron sequences between albumin subgenes. The variations in the intron/exon junctions showed no particular pattern, nor was there any noticeable homology in the intronic DNA that was sequenced. The corresponding introns in each subgene were quite different in size (Fig. 4A), suggesting that, as with globin genes (14, 26), albumin introns tend to diverge relatively rapidly, both in sequence and in length. The repeated elements present in introns CD, DE, and EF were absent from the corresponding introns in other subgenes.

DISCUSSION

The structure of the rat serum albumin gene is not unlike that of other eucaryotic protein-encoding sequences that have been analyzed at the nucleotide level. The short blocks of sequences that are characteristic of DNA flanking the 5′ and 3′ ends of many eucaryotic genes are present at the positions predicted from the R-loop data, allowing us to infer possible sites of transcriptional initiation and termination on this basis (Fig. 2). The AT-rich "Pribnow box" is evidently involved in the interaction between procaryotic genes and RNA polymerases (29), and the similar sequences found upstream from eucaryotic capping sites may serve an equivalent purpose. At the 3′ end of eucaryotic genes, the sequence AATAAA is usually found approximately 17 nucleotides upstream from the polyadenylation site, and thus may be involved in the specification of transcriptional termination (14, 30). Strictly speaking, the functions of flanking sequences are not known. The conserved sequences upstream from the 5′ end of 5S ribosomal RNA genes appear not to be involved in initiating transcription (31). On the other hand, similar experiments involving in vitro transcription of cloned conalbumin and ovalbumin genes (43) suggest that the 40 nucleotides preceding the capping site of this gene may be necessary for correct initiation. At the present time, the primary significance of these short sequences is their frequent appearance at particular positions relative to structural genes.

Properties of albumin gene introns. The albumin mRNA-encoding sequence is interrupted, as are other eucaryotic genes, by introns. Of the total gene length of 14,900 nucleotides, 12,900 represent intronic DNA. The intron/exon and exon/intron junctions are similar to the

```
                                   **
          *                 *     ***  *
Exon A    -------------HKSEIAHRFKDLGEQHFKGL
     E    LDAVKEKALVAAVRQRMKCSSMQRFGERAFKAW
     I    LAEFQPLVEEPKNLVKTNCELYEKLGEYGFQNA
                   .         .         .

                *                       *          *
          ** * **  *     *  ***  **   * **  **     *  *
     B    VLIAFSQYLQKCPYEEHIKLVQEVTDFAKTC-VADENAEN-CDKSI
     F    AVARMSQRFPNAEFAEITKLATDVTKINKECCHGDLLE---CADDR
     J    VLVRYTQKAPQVSTPTLVEAARNLGRVGTKCCTLPEAQRLPCVEDY
                   .         .         .         .

                  *              **         *      *                   *
                 **     *    **  ** *   **  **     ** *   *  *  * *    *    * * ****
     C    HTLFGDKLCAIPKLRDNYGELADCCAKQEPERNECFLQHKDDN-PNLPPFQ---RPEAEAMCTSFQENPTSFLGH
     G    -AELAKYMCE-NQATIS-SKLQACCDKPVLQKSQCLAETEHDNIPADLPSIAADFVEDKEVCKNYAEAKDVFLGT
     K    LSAILNRLCVLHEKTPVSEKVTKCCSGSLVERRPCFSALTVDETYVPKEFKAETFTFHSDICTLPDKEKQIKKQT
                   .         .         .         .         .         .         .

           * *     *      *             **   *   *
           * *  ****      **  * **   * *** * ** **  *
     D    YLHEVARRHPYFYAPELLYYAEKYNEVLTQCCTESDKAACLTPK
     H    FLYEYSRRHPDYSVSLLLRLAKKYEATLEKCCAEGDPPACYGTV
     L    ALAELVKHKPKATEDQLKTVMGDFAQFVDKCCKAADKDNCFATE
                   .         .         .         .
```

FIG. 5. *Internal amino acid homology. Peptides encoded by the four sets of equivalent exons were aligned for maximal homology by introducing gaps in the shorter sequences. Two of three matches are denoted by an asterisk, and three of three are denoted by a double asterisk. The first 13 amino acids of exons E and I are absent from exon A. When a codon is split by an intron, it is awarded to the exon which includes two of the three nucleotides. The numerical amino acid and nucleotide homologies are summarized in Table 2.*

TABLE 2. *Summary of internal homology*[a]

Comparison	Amino acids (%)	Nucleotides (%)
Exons		
A and E	3/20 (15)	26/58 (45)
A and I	4/20 (20)	22/58 (38)
E and I	5/33 (15)	37/98 (38)
B and F	11/43 (26)	51/130 (39)
B and J	6/44 (14)	46/133 (35)
F and J	8/43 (19)	55/130 (42)
C and G	16/71 (23)	93/212 (44)
C and K	10/71 (14)	72/212 (34)
G and K	11/72 (15)	86/215 (40)
D and H	18/44 (41)	68/133 (51)
D and L	9/44 (20)	51/133 (38)
H and L	10/44 (23)	57/133 (43)
Subgenes		
1 and 2	48/179 (27)	238/536 (44)
1 and 3	29/179 (16)	191/536 (36)
2 and 3	34/192 (18)	235/576 (41)

[a] The exons were aligned as shown in Fig. 5. In each comparison, the total was taken to be the length of the shorter sequence; i.e., gaps were ignored in the tabulation.

consensus sequences that have been adduced by analysis of a large number of genes.

An unusual property of albumin introns is the presence in introns CD and DE of the polydinucleotides AC and TC, which are repeated elsewhere in the rat and other genomes. Repeated sequences in introns might facilitate recombinations at these sites which could destroy the albumin gene. It would be interesting to know how long these simple intronic repeated sequences have been present. Human (J. W. Hawkins and A. Dugaiczyk, J. Cell Biol. **87:** 111a, 1980), mouse (19), and chicken (18) albumin genes are under investigation in other laboratories, so it should be possible to answer this question soon.

Of course, it is possible that sequences other than the simple tracts are repeated in the rat genome. All hybridization data obtained to date are consistent with the interpretation that the repeated nature of introns CD, DE, and EF is due to these elements, but definition of the limits of a repeated sequence is difficult without extensive nucleotide sequence data from many diverse copies of the repeated element.

Relationship between exons and protein domains. Serum albumin consists of three structural domains of approximately 190 amino acids each (9). These polypeptide segments are similar in secondary and tertiary structure and exhibit weak but significant overall amino acid homology. The nascent protein also has a short "signal" peptide attached to the amino terminus. The periodic nature of the albumin gene is even more pronounced in the nucleic acid sequence of the mRNA, and is in turn reflected in an obvious way in the pattern of exon sizes. The three structural domains correspond to exons ABCD, EFGH, and IJKL, which we have named subgenes 1, 2, and 3. The divergence at the protein and nucleic acid levels is extensive and nonuniform (Table 2). Since the subgenes presumably result from saltatory duplication and the divergence time for all exons within a given subgene is therefore identical, the relatively extensive conservation of some exons means either that there has been greater selection on these regions of the gene or that the basic mutation rate is different for each exon. Similar disparities in divergence have been observed in the exons of globin genes (14).

A great deal of diversity has evolved in the albumin protein by intragenic duplication followed by fixation of nucleotide substitutions. The polypeptides encoded by homologous exons are quite different from one another and may have correspondingly different functions. The possible correlation between exons and protein functional domains has been discussed at length (12, 16). Except for the leader exon (see below), this proposal may not be particularly appropriate to albumin. A number of substances have been found to bind to serum albumin; the active sites for each ligand are usually confined to one structural domain (reviewed by Peters and Reed [28]), and in many cases are probably encoded by individual exons. However, there is no reason to presume that these activities existed before the assembly of the ancestral albumin gene, since there has been such extensive amino acid sequence divergence since the subgene duplication events. Furthermore, the binding site for copper(II) ions consists of the first three amino acids of secreted bovine serum albumin (6), so this functional domain is disrupted by intron ZA (Table 1). Fatty acids apparently bind to the hydrophobic clefts between albumin structural domains, which are separated in the gene by introns DE and HI. The interpretation of these results is further complicated by the observation that in humans (reviewed by Gitlin and Gitlin [17]) and rats (15), the complete absence of serum albumin, analbuminemia, is almost asymptomatic. This suggests that albumin may have no vital function at all, which casts some doubt on the significance of those functional assignments that have been made.

In addition to the 12 subgene exons that encode most of the albumin protein, there are 3 that have special significance. At the 5′ end of the gene is a leader exon, Z. This exon encodes the 5′ untranslated region of albumin mRNA, the 18-amino-acid "pre" peptide (39), which includes the signal sequence (4) present on the amino terminus of nascent albumin, and the 6-amino-acid "pro" peptide. Exon Z also encodes the first 2⅓ amino acid residues of the secreted protein. Leader exons are found attached to other genes that encode secreted proteins. Mouse immunoglobulin light-chain genes are organized in this manner (3), as are the chicken ovomucoid (38) and conalbumin (11) genes and the bovine preproopiocortin gene (25). Although the structures of the leader exons differ considerably, they do seem to encode an equivalent protein functional domain, the signal peptide. At the 3′ end are two exons, M and N, that, like exon Z, seem not to be part of the subgenes. Exon N consists entirely of untranslated mRNA sequences, including the putative 3′ terminus of the albumin gene. The significance of such an exon is not clear. Obviously, it has nothing to do with protein function, and 3′ untranslated sequences tend to diverge more rapidly than the coding region. However, these elements are conserved to some extent (30), so they presumably have some significant, if unknown, role (37). Exon N does have a function in the sense that its sequence includes the polyadenylation site. Exon M is partly translated—the termination codon, TAA, is included in its sequence, and it encodes the COOH-terminal 13 amino acids. It does show slight homology to the 5′ third of exon I (15 of 40 matching nucleotides), which may be the result of one of the duplication events (Fig. 6).

In summary, exon Z can be regarded as encoding a functional domain, i.e., the signal peptide, and the three subgene clusters clearly encode the three structural domains of albumin. Other correlations between the exons of the albumin gene and domains in the protein are probably not justified at present.

Evolution of the serum albumin gene. Perhaps the most striking aspect of the serum albumin gene is the clarity with which its evolutionary history is preserved in its sequence. Brown (9) inferred a triplex structure for this gene from the pattern of internal amino acid homology and concluded that the three structural domains evolved by duplication of a common ancestor gene. This hypothesis is strongly supported by our observations.

Brown estimated the two duplications to have taken place 700 million years ago, an extrapolation from the amino acid sequence homology between domains and between human and bovine albumins and from the fossil record of mammalian radiation. This presumes that the selective pressures on albumin over the past 80 million years are indicative of the preceding billion or so, which may be incorrect and misleading. However, chicken serum albumin is approximately the same size as rat serum albumin (18), so it may have a similar three-domain structure.

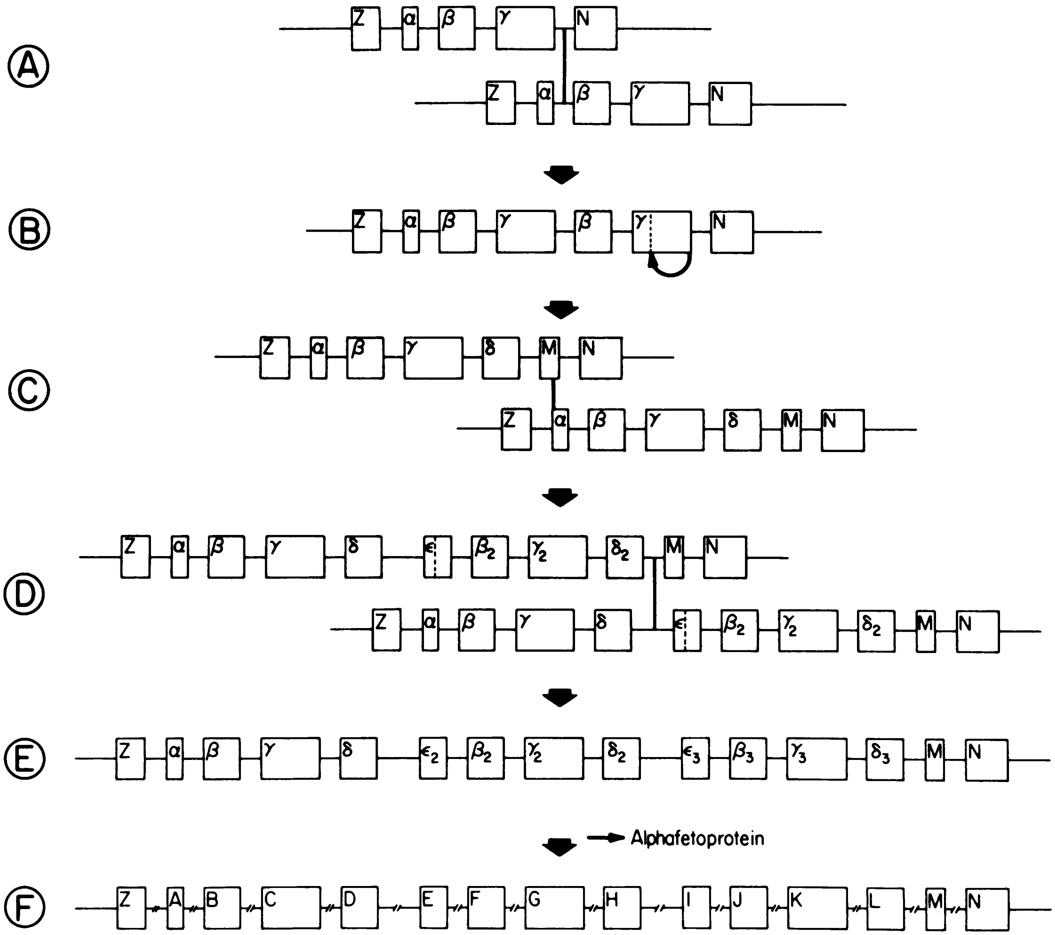

FIG. 6. *Model for the evolution of the rat serum albumin gene. (A) Unequal crossover between two copies of a five-exon gene duplicates the third and fourth exons. (B) The superfluous second γ exon shrinks to 62 nucleotides by a deletion or by evolution of a new splicing signal, creating the proto-albumin gene. (C) Unequal crossover between two alleles of the proto-albumin gene. The recombination sites are 40 nucleotides into exon M and the first nucleotide of exon α. This achieves a simultaneous duplication of most of the protein-coding sequence and a frameshift to compensate for the duplication of 3n + 2 nucleotides of the first subgene. The second exon β also diverges slightly into exon δ. (D) Second major intragenic duplication, with boundaries within introns, results in three approximately equivalent subgenes encoding three similar protein domains. (E) Divergence of exons to approximately 40% DNA homology and approximately 20% protein homology and extensive divergence of intron sequence and sizes. The α-fetoprotein–albumin gene duplication probably occurred during this period. (F) The rat serum albumin gene.*

If this is so, the domain duplications probably predate the bird-mammal divergence that occurred about 300 million years ago (44).

Because the albumin gene is so complex, there are many plausible models that could explain its evolution. We have applied two principal criteria to advance one of these alternatives: (i) the number of recombination and rearrangement steps should be minimized; and (ii) the translational reading frame should be continuously maintained throughout albumin evolution. Any model will have to account for the observed patterns of internal homology, which we interpret as indications of sequence duplication, and also the 40-nucleotide difference in the sizes of exons A versus E and I.

Before the two duplication events, the "proto-albumin" gene may have had a structure similar to that shown in Fig. 4B; a leader exon, four protein-encoding exons, exon M, and an untranslated exon equivalent to N. Because of the greater homology between protein domains I and II, Brown concluded that domain III is the "oldest" and would represent the ancestral albumin gene. However, we believe that this is incorrect and that the proto-albumin gene was equivalent to the first subgene attached to the 5'- and 3'-terminal exons Z, M, and N. The reason for this conclusion is that if an exon equivalent to I or E were spiced to the leader exon Z, a frameshift would result, since exon Z terminates after the first nucleotide in a codon and exons I and E begin between codons, assuming that the GT-AG rule is and has always been followed (Table 1). The leader exon could have been one nucleotide shorter before the duplications, but this would have to have changed when it became associated with what is now exon A. It is extremely unlikely that a frameshifted gene would survive long enough to become fixed in the population. Nor is there any obvious mechanism that could have simultaneously deleted 40 nucleotides from the 5' end of the ancestor to exon A and added a single nucleotide to the ancestor to exon Z. Therefore, we favor the hypothesis that exons Z and A have been associated all along; i.e., subgene 1 is the evolutionary precursor to the other two.

Subgene 1 has $3n + 2$ nucleotides, and this was presumably the case when it represented the proto-albumin gene. As such, a simple intragenic duplication of exons α through δ would result in a new second subgene that would be translated out of phase. Figure 6C illustrates a mechanism for the first subgene duplication that circumvents this difficulty. A recombination event between exon α and exon M of two different copies of the proto-albumin gene would duplicate the first subgene and transfer 40 nucleotides ($3n + 1$) to the 5' end of the new exon ϵ. This accomplishes a simultaneous duplication and compensating frameshift, so the enlarged gene would encode a translatable protein. This model also explains the larger size of exon E versus exon A. The extra 40 nucleotides of exon ϵ would not have been in the correct reading frame, and the 13 amino acids encoded by this DNA may serve merely as a linker polypeptide whose particular sequence is not important. This region of serum albumin is the most divergent segment of the protein when rat, human, and bovine albumins are compared, which is consistent with this interpretation.

Since subgene 2 has $3n$ nucleotides (currently 576), its duplication does not present a reading frame problem. This could be accomplished by recombination within introns δ_2M and $\delta\epsilon$ (Fig. 6D). The result of this (Fig. 6E) would be a 15-exon albumin gene with three homologous subgenes. The greater amino acid homology between domains I and II is not explained by our model, but this disparity is not overwhelming (27% amino acid homology versus 16 and 18%; Table 2), and the DNA homologies are even more similar (44% versus 36 and 41%). Furthermore, any argument based upon such minor differences in homology is considerably weakened by the observation (Table 2; 14) that exons can accumulate mutations at peculiar rates.

The period represented by the space between Fig. 6E and 6F probably lasted at least 300 million years and involved the fixation of a large number of nucleotide substitutions as well as several 3-, 6-, 9-, and 12-nucleotide deletions and insertions in the albumin exons (Table 1). The intronic homology also disappeared during this period. Another important evolutionary event that evidently took place was an intergenic duplication that resulted in the creation of two genes that evolved into what are now recognized as albumin and α-fetoprotein. The evidence for this is a very significant level of homology (50%) between the rat serum albumin and rat α-fetoprotein mRNA sequences (21, 21a, 35).

Possible duplication events preceeding the triplication of subgenes. It is also possible to infer the nature of even earlier evolutionary events that gave rise to the proto-albumin gene. There is a remarkable similarity between exons β and δ and their duplication products. The codon interruption patterns are identical, and the exon sizes and the positions of the cysteine residues are nearly so. The DNA sequence homology between these exons is 47/133, 41/130, and 47/133 matches for the B-D, F-H, and J-L pairs, respectively, when they are aligned with their 5' borders in register (Fig. 7). Assuming that DNA sequence homology values are Poisson distributed and that the basic probability of an accidental match is 25%, the probabilities that these or higher levels of homology are accidents are 0.014, 0.084, and 0.014, respectively. Although none of these values is small enough to convincingly exclude accidental homology, the probability that all three pairs of exons independently acquired homology in this way is miniscule (the product of the three pairwise accident probabilities is 0.000017. The accident probability for 135/396 matches is 0.00034). Nor is it likely that exons β and δ were initially different and converged to a high enough level of homology to account for the present similarity of their descendants, considering the extent of overall divergence that has taken place since the subgene duplications. Thus, we conclude that a duplication event was responsible for the generation of exons β and δ from a common ancestor. Since these two exons are separated by exon

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
		val	leu	ilu	ala	phe	ser	gln	tyr	leu	gln	lys	cys	pro	tyr	glu	glu	his	ilu	lys	leu	val	gln
exon B	A	GTC	CTG	ATT	GCC	TTT	TCC	CAG	TAT	CTC	CAG	AAA	TGC	CCA	TAT	GAA	GAG	CAT	ATC	AAA	TTG	GTG	CAG
		tyr	leu	his	glu	val	ala	arg	arg	his	pro	tyr	phe	tyr	ala	pro	glu	leu	leu	tyr	tyr	ala	glu
exon D	C	TAT	TTG	CAT	GAA	GTT	GCC	AGG	AGA	CAT	CCT	TAT	TTC	TAT	GCC	CCA	GAA	CTC	CTT	TAC	TAT	GCT	GAG
		ala	val	ala	arg	met	ser	gln	arg	phe	pro	asn	ala	glu	phe	ala	glu	ilu	thr	lys	leu	ala	thr
exon F	G	GCA	GTA	GCT	CGT	ATG	AGC	CAG	AGA	TTC	CCC	AAT	GCT	GAG	TTC	GCA	GAA	ATC	ACC	AAA	TTG	GCA	ACA
		phe	leu	tyr	glu	tyr	ser	arg	arg	his	pro	asp	tyr	ser	val	ser	leu	leu	leu	arg	leu	ala	lys
exon H	G	TTT	TTG	TAT	GAA	TAT	TCA	AGA	AGG	CAC	CCC	GAT	TAC	TCC	GTG	TCC	CTG	CTG	CTG	AGA	CTT	GCT	AAG
		val	leu	val	arg	tyr	thr	gln	lys	ala	pro	gln	val	ser	thr	pro	thr	leu	val	glu	ala	ala	arg
exon J	C	GTT	CTG	GTT	CGA	TAC	ACC	CAG	AAA	GCA	CCT	CAG	GTG	TCG	ACC	CCA	ACT	CTC	GTG	GAG	GCA	GCA	AGA
		ala	leu	ala	glu	leu	val	lys	his	lys	pro	lys	ala	thr	glu	asp	gln	leu	lys	thr	val	met	gly
exon L	G	GCT	CTC	GCT	GAG	CTG	GTG	AAA	CAC	AAG	CCC	AAG	GCC	ACA	GAA	GAT	CAG	CTG	AAG	ACG	GTG	ATG	GGT

	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44		
	glu	val	thr	asp	phe	ala	lys	thr	cys	val	ala	asp	glu	asn	ala	glu	asn	cys	asp	lys	ser	ilu		
exon B	GAA	GTA	ACA	GAC	TTT	GCA	AAA	ACA	TGT	GTC	GCT	GAT	GAG	AAT	GCC	GAA	AAC	TGT	GAC	AAG	TCC	ATT		
	lys	tyr	asn	glu	val	leu	thr	gln	cys	cys	thr	glu	ser	asp	lys	ala	ala	cys	leu	thr	pro	lys		
exon D	AAA	TAC	AAT	GAG	GTT	CTG	ACC	CAG	TGC	TGC	ACA	GAG	TCT	GAC	AAA	GCA	GCC	TGC	CTG	ACA	CCG	AAG		
	asp	val	thr	lys	ilu	asn	lys	glu	cys	cys	his	gly	asp	leu	leu	glu	cys	ala	asp	asp	arg			
exon F	GAC	GTT	ACC	AAA	ATC	AAC	AAG	GAG	TGC	TGT	CAC	GGC	GAC	CTG	TTG	GAA	TGC	GCG	GAT	GAC	AGG			
	lys	tyr	glu	ala	thr	leu	glu	lys	cys	cys	ala	glu	gly	asp	pro	pro	ala	cys	tyr	gly	thr	val		
exon H	AAA	TAT	GAA	GCC	ACA	CTG	GAG	AAG	TGC	TGT	GCT	GAA	GGC	GAT	CCT	CCT	GCC	TGC	TAC	GGC	ACA	GTG		
	asn	leu	gly	arg	val	gly	thr	lys	cys	cys	thr	leu	pro	glu	ala	gln	arg	leu	pro	cys	val	glu	asp	tyr
exon J	AAC	CTG	GGA	AGA	GTG	GGC	ACC	AAG	TGT	TGT	ACC	CTT	CCT	GAA	GCT	CAG	AGA	CTG	CCC	TGT	GTG	GAA	GAC	TAT
	asp	phe	ala	gln	phe	val	asp	lys	cys	cys	lys	ala	ala	asp	lys	asp	asn	cys	phe	ala	thr	glu		
exon L	GAC	TTC	GCA	CAA	TTC	GTG	GAC	AAG	TGT	TGC	AAG	GCT	GCC	GAC	AAG	GAT	AAC	TGC	TTC	GCC	ACT	GAG		

FIG. 7. *β versus δ exon homology. The nucleotide and amino acid sequences of exons B and D, F and H, and J and L are aligned with the 5′ ends in register. Matching nucleotides are indicated, and the cysteine residues are underlined. There are five sites of extensive homology:amino acid residue numbers 2, 10, 30, 31, and 32.*

γ, the most straightforward assumption is that this intervening exon was also duplicated at the same time as the precursor to exons β and δ. Figure 6A illustrates a hypothetical recombination event between two alleles of a five-exon precursor gene that would result in a duplication of the third and fourth exons. The translational reading frame would be preserved, but the second γ exon would have to be eliminated before the next duplication event. This could occur via either a deletion or the evolution of a new 3′ splicing junction, which we have depicted in Fig. 6B as resulting in the appearance of exon M. Although there is some weak homology supporting this relationship between exon M and the descendents of exon γ, it is not statistically significant by our criterion. The primary justification for the proposed mechanism is that it is simple; i.e., it accounts for both the genesis of exon M and the elimination of the superfluous exon γ in a single evolutionary event. Other interpretations are of course not excluded. There are 18 positions of amino acid homology between second and fourth subgene exons, out of a total of 131 sites compared. Six of these are cysteines, and the rest are clustered primarily at three points; the second amino acid is usually leucine (5/6), the tenth is usually proline (5/6), and half of the positions immediately preceeding the double cysteines are lysine residues. The double cysteines are in the same alignment as the 5′ exonic boundaries and the single cysteines with the 3′ boundaries, so the different exon lengths are probably due to insertions or deletions between the second and third cysteine residues. The homology score can be improved by allowing for this, but we have not attempted to calculate the statistical significance of such comparisons.

Nucleotide sequence determination and analysis has been an effective approach to the study of this complex gene. The intricate pattern of introns and exons originally elucidated by R-loop analysis has been completely resolved, the putative termini of the transcription unit have been located, and the triplication model of albumin evolution has been confirmed and elaborated. We have been able to infer the probable nature of a series of ancient duplications of multiexon sequences by inspection of the nucleotide sequence of existing genes. A number of points have emerged that suggest future lines of investigation. There is weak homology between exons descended from α, β, and γ, and it may be possible to adduce earlier events in albumin evolution by a more sophisticated analysis of these data. Since this would represent the origin of a primitive multiexon gene, it might serve as a paradigm for the evolution of all introns. It would also be interesting to know if the albumin-α-fetoprotein gene family exists as cluster of sequences, if it has additional members, and if mechanisms other than intragenic and intergenic duplications were involved in its evolution.

Intragenic duplication has evidently been the principal evolutionary mechanism for the accretion of exons and introns by the ancestral precursor to the albumin (and α-fetoprotein) gene. According to our model, at least 10 of 14 introns and 10 of 15 exons were generated in this fashion. In view of the large number of proteins with internal periodicity (1), it is apparent that this has been an important evolutionary source of the complexity of eucaryotic genes and genomes.

ACKNOWLEDGMENTS

We thank Jim Posakony for helpful discussions and Margaret Chamberlin for technical advice and for the gift of pBR325 subcloning vector DNA.

This project was supported by Public Health Service grants 5 T32 GM 07616 and GM 13762, awarded by the National Institute of General Medical Sciences.

LITERATURE CITED

1. **Barker, W. C., L. K. Ketcham, and M. O. Dayhoff.** 1978. Composition of proteins, p. 359-362. *In* M. O. Dayhoff (ed.), Atlas of protein sequence and structure, vol. 5, suppl. 3. National Biomedical Research Foundation, Washington, D.C.
2. **Benoist, C., K. O'Hare, R. Breathnach, and P. Chambon.** 1980. The ovalbumin gene—sequence of putative control regions. Nucleic Acids Res. **8**:127-142.
3. **Bernard, O., N. Hozumi, and S. Tonegawa.** 1978. Sequences of mouse immunoglobulin light chain genes before and after somatic changes. Cell **15**:1133-1144.
4. **Blobel, G., and D. Dobberstein.** 1975. Transfer of proteins across membranes. II. Reconstitution of functional rough microsomes from heterologous components. J. Cell Biol. **67**:852-862.
5. **Bolivar, F.** 1978. Construction and characterization of new cloning vehicles. III. Derivatives of plasmid pBR322 carrying unique EcoRI sites for selection of EcoRI generated recombinant DNA molecules. Gene **4**: 121-136.
6. **Bradshaw, R. A., W. T. Shearer, and F. R. N. Gurd.** 1968. Sites of binding of copper (II) ion by peptide (1-24) of bovine serum albumin. Biol. Chem. **243**:3817-3825.
7. **Breathnach, R., C. Benoist, K. O'Hare, and P. Chambon.** 1978. Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. Proc. Natl. Acad. Sci. U.S.A. **75**: 4853-4857.
8. **Britten, R. J., D. E. Graham, and B. R. Neufeld.** 1974. Analysis of repeating DNA sequences by reassociation. Methods Enzymol. **29E**:363-418.
9. **Brown, J. R.** 1976. Structural origins of mammalian albumin, Fed. Proc. **35**:2141-2144.
10. **Catterall, J. F., J. P. Stein, P. Kristo, A. R. Means, and B. W. O'Malley.** 1980. Primary sequence of ovomucoid messenger RNA as determined from cloned complementary DNA. J. Cell Biol. **87**:480-487.
11. **Cochet, M., F. Gannon, R. Hen, L. Maroteaux, F. Perrin, and P. Chambon.** 1979. Organization and sequence studies of the 17-piece chicken conalbumin gene. Nature **282**:567-574.
12. **Crick, F.** 1979. Split genes and RNA splicing. Science **204**:264-271.
13. **Dugaiczyk, A., H. W. Boyer, and H. M. Goodman.** 1975. Ligation of EcoRI endonuclease-generated DNA fragments into linear and circular structures. J. Mol. Biol. **96**:171-184.
14. **Efstratiadis, A., J. W. Posakony, T. Maniatis, R. M. Lawn, C. O'Connell, R. A. Spritz, J. K. DeRiel, B. G. Forget, S. M. Weissman, J. J. Slightom, A. E. Blechl, O. Smithies, F. E. Baralle, C. C. Shoulders, and N. J. Proudfoot.** 1980. The structure and evolution of the human α-globin gene family. Cell **21**:653-668.
15. **Esumi, H., M. Okui, S. Sato, T. Sugimura, and S. Nagase.** 1980. Absence of albumin mRNA in the liver of analbuminemic rats. Proc. Natl. Acad. Sci. U.S.A. **77**:3215-3219.
16. **Gilbert, W.** 1978. Why genes in pieces? Nature **271**:501.
17. **Gitlin, D., and J. D. Gitlin.** 1975. Genetic alterations in the plasma proteins of man, p. 321-374. *In* F. W. Putnam (ed.), The plasma proteins, vol. 2. Academic Press, Inc., New York.
18. **Gordon, J. I., A. T. H. Burns, J. L. Christmann, and R. G. Deeley.** 1978. Cloning of a double-stranded cDNA that codes for a portion of chicken preproalbumin. J. Biol. Chem. **253**:8629-8639.
19. **Gorin, M. B., and S. M. Tilghman.** 1980. The structure of the alpha-fetoprotein gene in the mouse. Proc. Natl. Acad. Sci. U.S.A. **77**:1351-1355.
20. **Gubbins, E. J., R. A. Mauer, M. Lagrimini, C. R. Erwin, and J. E. Donelson.** 1980. Structure of the rat prolactin gene. J. Biol. Chem. **255**:8655-8662.
21. **Innis, M. A., and D. L. Miller.** 1980. α-Fetoprotein gene expression. Partial DNA sequence and COOH terminal homology to albumin. J. Biol. Chem. **255**:8994-8996.

21a. **Jagodzinski, L. L., T. D. Sargent, M. Yang, C. Glackin, and J. Bonner.** 1981. Sequence homology between RNAs encoding rat α-fetoprotein and rat serum albumin. Proc. Natl. Acad. Sci. U.S.A. **78**:3521-3525.

22. **Kushner, S. R.** 1978. Improved method for transformation of E. coli with Col E1 derived plasmids, p. 17-23. *In* A. W. Boyer and S. Nicosia (ed.), Proceedings of the International Symposium on Genetic Engineering. Elsevier/North-Holland Biomedical Press, New York.
23. **Lerner, M. R., J. A. Boyle, S. M. Mount, S. L. Wolin, and J. A. Steitz.** 1980. Are sn RNPs involved in splicing? Nature (London) **283**:220-224.
24. **Maxam, A. M., and W. Gilbert.** 1980. Sequencing end-labeled DNA with base-specific chemical cleavages. Methods Enzymol. **65**:499-560.
25. **Nakanishi, S., Y. Teranishi, M. Noda, M. Notake, Y. Watanabe, H. Kakidani, H. Jingami, and S. Numa.** 1980. The protein-coding sequence of the bovine ACTH-β-LPH precursor gene is split near the signal peptide region. Nature (London) **287**:752-755.
26. **Nishioka, Y., and P. Leder.** 1979. The complete sequence of a chromosomal mouse α-globin gene reveals elements conserved throughout vertebrate evolution. Cell **18**:875-882.
27. **Nishioka, Y., and P. Leder.** 1980. Organization and complete sequence of identical embryonic and plasmacytoma K V-region genes. J. Biol. Chem. **255**:3691-3694.
28. **Peters, T., Jr., and R. G. Reed.** 1977. Serum albumin: conformation and active sites, p. 11-20. *In* T. Peters and I. Sjoholm (ed.), Albumin: structure, biosynthesis, function. Pergamon Press, New York.
29. **Pribnow, D.** 1975. Nucleotide sequence of an RNA polymerase binding site at an early T7 promotor. Proc. Natl. Acad. Sci. U.S.A. **72**:784-788.

30. **Proudfoot, N. J., and G. G. Brownlee.** 1976. 3′ Noncoding region sequences in eukaryotic messenger RNA. Nature (London) **263**:211-214.

31. **Sakonju, S., D. F. Bogenhagen, and D. D. Brown.** 1980. A control region in the center of the 5S RNA gene directs specific initiation of transcription: 1. The 5′ border of the region. Cell **19**:13-25.

32. **Sela-Trepat, J. M., T. D. Sargent, S. Sell, and J. Bonner.** 1979. α-Fetoprotein and albumin genes of rats: no evidence for amplification-deletion or rearrangement in rat liver carcinogenesis. Proc. Natl. Acad. Sci. U.S.A. **76**:695-699.

33. **Sanger, F., and R. Coulson.** 1978. The use of thin acrylamide gels for DNA sequencing. FEBS Lett. **87**: 107-110.

34. **Sargent, T. D., J.-R. Wu, J. M. Sala-Trepat, R. B. Wallace, A. A. Reyes, and J. Bonner.** 1979. The rat serum albumin genes: analysis of cloned sequences. Proc. Natl. Acad. Sci. U.S.A. **76**:3256-3260.

35. **Sargent, T. D., M. Yang, and J. Bonner.** 1981. Nucleotide sequence of cloned rat serum albumin messenger RNA. Proc. Natl. Acad. Sci. U.S.A. **78**:243-246.

36. **Seif, I., G. Khoury, and R. Dhar.** 1979. BKV splice sequences based on analysis of preferred donor and acceptor sites. Nucleic Acids Res. **6**:3387-3389.

37. **Setzer, D. R., M. McGrogan, J. H. Nunberg, and R. T. Schimke.** 1980. Size heterogeneity in the 3′ end of dihydrofolate reductase messenger RNAs in mouse cells. Cell **22**:361-370.

38. **Stein, J. P., J. F. Catterall, P. Kristo, A. R. Means, and B. W. O'Malley.** 1980. Ovomucoid intervening sequences specify functional domains and generate protein polymorphism. Cell **21**:681-687.

39. **Strauss, A. W., C. D. Bennett, A. M. Donohue, J. A. Rodkey, and A. W. Alberts.** 1977. Rat liver preproalbumin: complete amino acid sequence of the prepiece. J. Biol. Chem. **252**:6846-6855.

40. **Sures, I., J. Lowry, and L. H. Kedes.** 1978. The DNA sequence of sea urchin (S. purpuratus) H2A, H2B, H3 histone coding and spacer regions. Cell **15**:1033-1044.

41. **Tucker, P. W., K. B. Marcu, N. Newell, J. Richards, and F. R. Blattner.** 1979. Sequence of the cloned gene for the constant region of murine of 2b immunoglobulin heavy chain. Science **206**:1303-1306.

42. **Wallace, R. B., T. D. Sargent, R. F. Murphy, and J. Bonner.** 1977. Physical properties of chemically acetylated rat liver chromatin. Proc. Natl. Acad. Sci. U.S.A. **74**:3244-3248.

43. **Wasylyk, B., C. Kedinger, J. Corden, O. Brison, and P. Chambon.** 1980. Specific in vitro initiation of transcription on conalbumin and ovalbumin genes and comparison with adenovirus-2 early and late genes. Nature (London) **285**:367-373.

44. **Wilson, A. C., S. S. Carlson, and T. J. White.** 1977. Biochemical evolution. Annu. Rev. Biochem. **46**:573-639.