The Rat Serum Albumin Gene

Thesis by

Thomas D. Sargent

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1981

(Submitted February 11, 1981)

ACKNOWLEDGEMENTS

I thank the National Science Foundation for supporting me during my first three years of graduate education. I am grateful to Mark, Jim, Eric, Bruce and the two Bills for teaching me so much, to James and Inselore for their patience, to Linda, Carlotta and Maria for carrying more than their share of the burden, to Lody for protecting me from evil, to my parents and family for their unflagging support and encouragement and to all my friends and colleagues at Caltech for setting standards that I have tried so hard to meet. Most of all, I am grateful to Maggie for being there.

ABSTRACT

The messenger RNA's that encode the proteins rat serum albumin (RSA) and rat alphafetoprotein (RAFP) have been purified to virtual homogeneity by a combination of immunoprecipitation of polysomes and other physical isolation methods. Radioactive cDNA copies of these mRNA's have been prepared and used to monitor the changes in abundance of RSA- and RAFP-synthesizing polysomes in neonatal rat liver and in Morris hepatoma 7777, a rat liver tumor. These cDNA probes have also been used to determine the stability and reiteration frequency of their respective genes in various rat tissues.

The RSA mRNA sequence has been converted into a series of bacterial plasmids by recombinant DNA methodology, and the RSA gene has been isolated from a library of recombinant bacteriophage. Restriction endonuclease site mapping, R-loop mapping, "Southern" blot and extensive nucleotide sequence determination have been employed to elucidate the organization of the cloned sequences.

The rat serum albumin gene has been found to be interrupted at fourteen locations by introns. The fifteen exons are spread over approximately 15,000 nucleotides of contiguous chromosomal DNA. The evolutionary history of albumin has been deduced by analysis of the patterns of

internal periodic homology in this gene. Albumin apparently evolved by a series of at least three intragenic duplications followed by accumulation of many point mutations and small deletions. These events probably occurred over 300 million years ago. The rat alphafetoprotein gene has been shown to be related to the rat serum albumin gene by a common ancestor gene, and thus two (or possibly more) highly complex genes with many exons and many protein domains have evolved by duplication mechanisms. That this might be an important general source of the complexity of eukaryotic genomes is discussed.

Table of Contents

INTRODUCTION

The subject of this dissertation is the analysis of the gene that encodes rat serum albumin. Before going into the scientific details, a word about the historical significance of this type of research is appropriate. The concept of "heredity" must have been extant, in some primitive form, in the minds of the first practitioners of agriculture, perhaps ten thousand years ago. That animals and plants "bred true" had to be appreciated in order to cultivate successfully. However, a meaningful, systematic approach to the science of heredity did not really commence until Mendel did his famous experiments with Pisum sativum in the 1860's, or rather with the "rediscovery" of his work some 35 years later by Hugo de Vries and others.

The concept of the "gene" was refined by Morgan and his students in the early part of the present century, but the chemical nature of the genetic material remained a mystery until the second great heuristic advance, the elucidation of the role and structure of DNA by Avery, MacLeod and McCarty, Franklin, Watson and Crick, and many others, of course. This is usually regarded as the beginning of "molecular biology", and in the quarter century since the publication of Watson and Crick's paper, the field progressed rapidly,

particularly in the area of bacterial molecular genetics. Progress was not so rapid, however, with eukaryotes. By 1975 the enormous complexity of the eukaryotic genome, especially that of vertebrates, had become a limiting obstacle to research. Methods such as kinetic analysis of nucleic acid renaturation, X-ray diffraction and neutron scattering by chromatin, primitive sequence determination and others available during the preceding period were simply incapable of resolving the basic features of individual eukaryotic genes. For this reason, the invention of "recombinant DNA technology", rapid DNA sequence determination and "Southern blots" have precipitated what can be fairly considered to be the third revolution in the biology of heredity. The research described in this dissertation began during this revolution and was part of it.

Perhaps the most remarkable discovery of the past few years was the presence of interruptions, or "introns", with the mRNA-coding sequences ("exons") of most eukaryotic genes. The ubiquity of these elements suggests that they have been around for a long time; introns are probably as ancient as are eukaryotes. It is likely that the discontinuous nature of eukaryotic genes has been important in the evolution of eukaryotic genomes (Crick, 1979). This appears to be the case with serum albumin. An argument is

advanced in Chapter 4 that unequal recombinations between introns of different copies of a primitive albumin gene were responsible for the internal periodic homology seen in this protein. The same argument can be applied to chicken ovomucoid (Stein et al., 1980) and the multiple constant region domains of immunoglobulins (Tucker et al., 1979). There are many other examples of proteins with internal periodic homology (Barker, Ketcham and Dayhoff, 1978), and most or all of these are probably also examples of the result of intragenic duplication of exons by recombination within introns. Gilbert (1978) and others have speculated that similar recombinations between the introns of unrelated genes might have been an important evolutionary mechanism for the generation of diversity in eukaryotic genomes. According to this theory, exons tend to encode discrete protein functional "domains", and by "shuffling exons" via recombination between different genes, new composites of multiple protein functions could be rapidly created, and possibly become fixed in the population, if they proved useful. The extent to which this model applies to eukaryotic genes remains to be seen. The boundaries of protein functional domains are not always well-defined, and many protein functions are probably the net result of polypeptides encoded by more than one exon. Furthermore, any novel association of exons requires that the

translational reading frame is maintained. As discussed in Chapter 4, many (at least 2/3) potential exon reassortments will not survive because they lead to translational frameshifts in the resultant mRNA.

Another consequence of the presence of introns in the albumin and other genes is that they must be removed somehow before the gene sequence can be translated into protein. It is evident that this takes place by enzymatic excision of the intronic RNA from the primary transcript, with concomitant ligation of consecutive exons (for example, Nordstrom et al., 1979). Nucleotide sequence determination has been instrumental in beginning to analyze this process. There is little homology between introns, even between introns of closely related genes (Efstratiadis et al, 1980). However, surveys of sequences immediately surrounding intron/exon boundaries has revealed that there exist "canonical" junction sequences that are found at the ends of all introns, including those of different phyla (Seif, Khoury and Dhar, 1979). The "GT-AG rule", first enunciated by Breathnach et al. (1978) has thus been expanded as follows: The exon-intron boundary is similar to "(A/C)AG-GTAAGT, and the intron-exon boundary is similar to "TYTYYYTCAG-G" (Y=T or C). This rule can be applied to all 28 of the intron/exon junctions in the rat serum albumin gene. This consensus sequence assumes additional

significance when compared to the nucleotide sequence of a small nuclear RNA, "U1" (Reddy et al., 1974, Lerner et al., 1980, Rogers and Wall, 1980). It is possible to construe the formation of a duplex structure including the 3' end of U1 RNA and the 5' and 3' extremities of any intron. This structure would bring adjacent exons into physical proximity, and could conceivably be the specific recognition substrate of splicing enzymes. Avvedimento et al. (1980) have been able to predict the location of intermediate splicing sites within an intron of the chicken collagen gene on this basis, which strongly supports the validity of the hypothesis that "U1" is involved in intron removal. The possibility that extensive gene regulation is exercised at the level of "transcript processing" (Wold et al., 1978, Kamalay and Goldberg, 1980) makes this area even more exciting.

Perhaps the major motivation to examine eukaryotic genes is the desire to understand how they are regulated. Serum albumin, and the related protein alphafetoprotein (AFP), represent promising subjects for research in this area. As discussed in Chapter 1, these proteins are characteristic synthetic products of liver hepatocytes, and as such are examples of how the process of "terminal differentiation" applies to individual genes. It is apparent from results from our laboratory and others' (Liao,

Conn and Taylor, 1980) that albumin production (and also that of AFP) is controlled primarily at the level of transcription, i.e. the albumin gene is actively transcribed into mRNA in the liver but not in other tissues. Unfortunately, we cannot even begin to explain why this is so. There was a naive expectation, at least initially, that examination of the chromosomal DNA that flanks the 5′ and 3′ termini of the albumin gene might provide some insight into how this gene is regulated. We have sequenced several thousand nucleotides at both ends of the albumin gene, and these data are tabulated in the Appendix. The significance of most of this information, if any, is not yet apparent. Immediately "upstream" from the putative capping site of albumin mRNA we have found the oligonucleotides CCAAT and TATATT, which have been found in similar form and positions near many other eukaryotic genes (Efstratiadis et al., 1980). These may be involved in the positioning of RNA polymerase prior to the initiation of transcription (Wasylyk et al., 1980), but they give no clue as to how it is decided to transcribe or not to transcribe the ensuing DNA. One hopes that this problem will be solved soon, and the "flanking" DNA sequences of the rat serum albumin gene may prove helpful.

It should be kept in mind that we have not cloned the albumin gene, but merely its DNA component. In the nuclei

of rat cells, this DNA is associated with histones, in the form of nucleosomes and probably with other proteins as well. The "anatomy" of this or any other eukaryotic gene must be considered to include these proteins and the manner in which they interact with the DNA and with one another. In other words, analysis of genomic DNA is only the first step in the analysis of chromatin

For several years a regulatory role for "middle repetitive" sequences has been contemplated (Britten and Davidson, 1969). We have concluded that the rat serum albumin gene resides within a 20,000 nucleotide region of chromosomal DNA that is probably devoid of such elements. This was unexpected, since "middle repetitive" sequences are spread throughout approximately 70% of the rat genome as a whole (Pearson, Wu and Bonner, 1978). There are sequences in three introns that are repeated many times elswhere in the genome, and there is a highly repeated sequence element located 600 nucleotides from the 3′ end of the rat serum albumin gene. However, none of these elements seem to be "middle repetitive", as described by Britten and Davidson. Furthermore, there are 2000-3000 nucleotides of single-copy DNA immediately preceding the 5′ terminus of this gene. The arguments concerning the probable existence of "gene batteries" under coordinated control are still valid, but the role, if any, of reiterated sequences in the regulation

of genes like albumin is as obscure as ever.

The primary contributions of the work leading to this thesis are the creation of a rat genomic library, the isolation of the rat serum albumin gene and the elucidation of its fine structure, (Chapters 2 and 4), the determination of the nucleotide sequence of rat serum albumin messenger RNA and the inferred amino acid sequence of rat serum albumin (Chapter 3), and the elucidation of the role played by intragenic and intergenic sequence duplication in the evolution of what can be regarded as the albumin gene family (Chapter 4).

REFERENCES

Avvedimento, V.E., Voseli, G., Yamada, Y., Maizel, J.V. Jr., Pastan, I. and de Crombrugghe, B. (1980) Cell 21, 689-696.

Barker, W.C., Ketcham, L.K. and Dayhoff, M.O. (1978) in Atlas of Protein Sequence and Structure, ed. Dayhoff, M.O. (National Biomedical Research Foundation, Washington, D.C.) vol. 5, suppl. 3, pp. 359-362.

Breathnach, R., Benoist, C., O'Hare, K. and Chambon, P. (1978) Proc. Natl. Acad. Sci. USA 75, 4853-4857.

Britten, R.J. and Davidson, E.H. (1969) Science 165, 349-357.

Crick, F. (1979) Science 204, 264-271.

Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, C., Spritz, R.A., DeRiel, J.K., Forset, B.G., Weissman, S.M., Slightom, J.L., Blechl, A.E., Smithies, O., Baralle, F.E., Shoulders, C.C. and Proudfoot, N.J. (1980) Cell 21, 653-668.

Gilbert, W. (1978) Nature 271, 501.

Kamalay, J.C. and Goldberg, R.B. (1980) Cell 19, 935-946.

Lerner, M.R., Boyle, J.A., Mount, S.M., Wolin, S.L. and Steitz, J.A. (1980) Nature 283, 220-224.

Liao, W.S.L., Conn, A.R. and Taylor, J.M. (1980) J. Biol. Chem. 255, 10036-10039.

Nordstrom, J.L., Roop, D.R., Tsai, M.J. and O'Malley, B.W. (1979) Nature 278, 328-331.

Pearson, W.R., Wu, J.-R. and Bonner, J. (1978) Biochemistry 17, 51-59.

Reddy, R., Ro-Choi, T.S., Henning, D. and Busch, H. (1974) J. Biol. Chem. 249, 6486-6494.

Rogers, J. and Wall, R. (1980) Proc. Natl. Acad. Sci. USA 77, 1877-1879.

Seif, I., Khoury, G. and Dhar, R. (1979) Nucleic Acids Res. 6, 3387-3398.

Stein, J.P., Catterall, J.F., Kristo, P., Means, A.R. and O'Malley, B.W. (1980) Cell 21, 681-687.

Tucker, P.W., Marcu, K.B., Newell, N., Richards, J. and Blattner, F.R. (1979) Science 206, 1303-1306.

Wasylyk, B., Kedinger, C., Corden, J., Brison, O. and Chambon, P. (1980) Nature 285, 367-373.

Wold, B.J., Klein, W.H., Housh-Evans, B.R., Britten, R.J. and Davidson, E.H. (1978) Cell 14, 941-950.

CHAPTER ONE

Preliminary Characterization of the Rat Serum Albumin and Rat Alphafetoprotein mRNA's and Genes.

SUMMARY

The messenger RNA's encoding the rat serum proteins albumin (RSA) and alphafetoprotein (RAFP) have been purified and used as templates for the synthesis of radioactive cDNA. The sizes of the full-length copies of these two mRNA's were very similar, and coincide with the estimated length of their non-poly(A) sequences. Hybridization of these probes to various preparations of rat liver polysomal RNA reveals that following birth, the level of RSA mRNA in liver remains relatively constant at approximately 68,000 (+/-50%) copies per liver cell. RAFP concentrations in the same cells fall precipitously after birth, from 18,000 at parturition to 10 or fewer copies per cell by age 6 weeks. The rate and extent of hybridization of RSA and RAFP cDNA's to rat genomic DNA suggest that these genes occur a few times per haploid genome. There is no detectable amplification or deletion of either RSA or RAFP genes associated with neoplastic transformation or liver differentiation. The unchanged patterns of hybridization of RSA and RAFP cDNA

probes to "Southern blots" of restriction endonuclease-disested nuclear DNA from liver and tumors indicate that there are no substantial rearransements of these senes durins tumor formation.

INTRODUCTION

Albumin and alphafetoprotein are the major proteins of adult and fetal serum, respectively. Both are synthesized at high levels in the liver parenchymal cells which constitute most of the mass of that organ. As such these proteins represent convenient examples of "terminal differentiation" at the level of individual genes. Alphafetoprotein synthesis in liver cells declines abruptly following parturition in mammals and the protein is virtually absent from adult serum. Partial hepatectomy, chemically induced hepatocyte necrosis or chronic exposure to chemical carcinogens can induce re-expression of the alphafetoprotein gene in the adult animal (Sell and Becker, 1978). Elevated serum concentration of this protein is in fact a widely recognized symptom of liver damage or neoplasia (Sell et al., 1976). In most cases, renewed synthesis of alphafetoprotein is accompanied by diminished albumin synthesis (Schreiber et al., 1971). However, there is not much alteration, perhaps a twofold increase, in albumin gene activity during the early postnatal extinction of alphafetoprotein synthesis. Nevertheless, there does seem to be some inverse relationship in the expression of these two structural genes, and a comparative analysis of their regulatory mechanisms promises to be a constructive

approach.

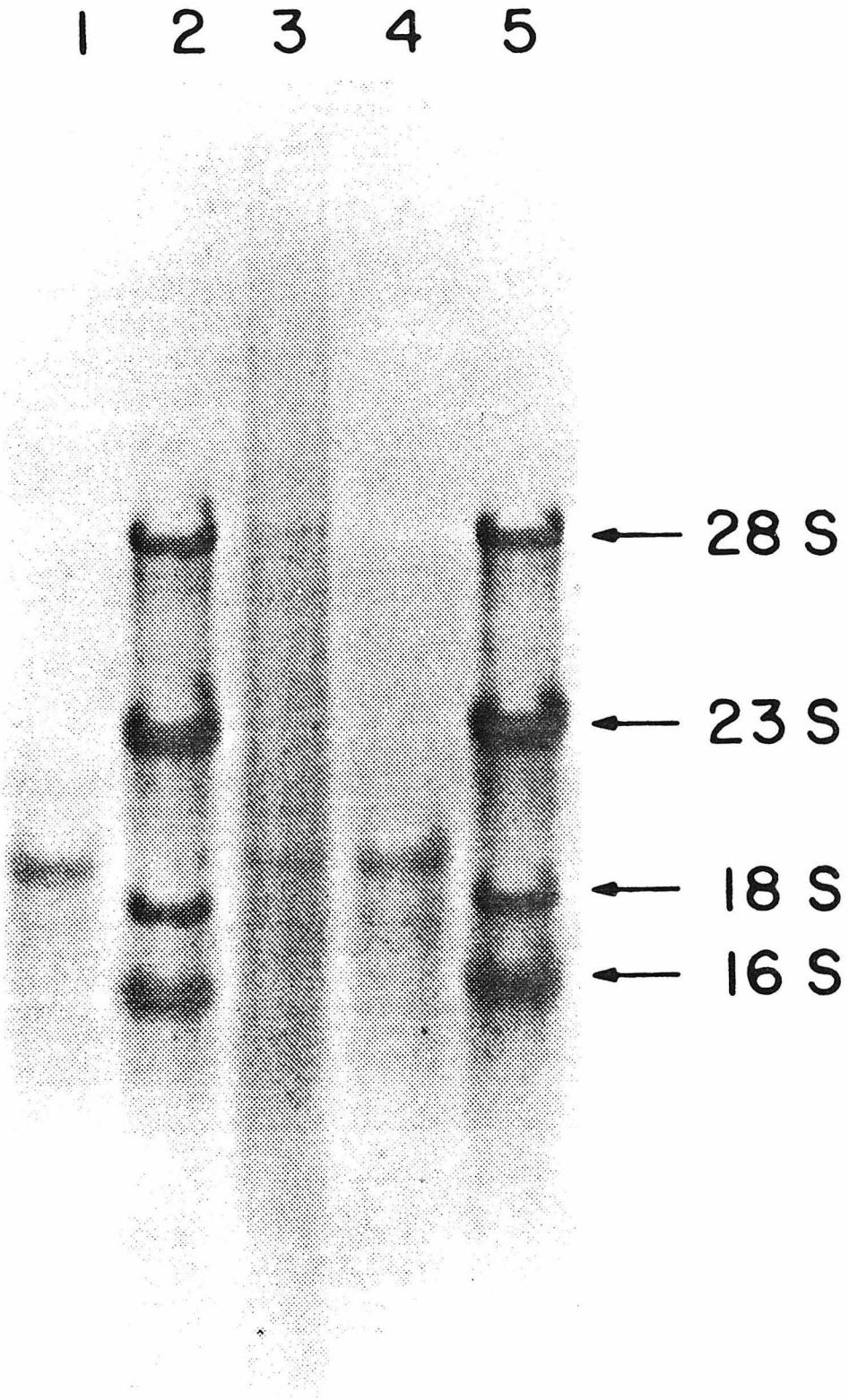
Another important feature of albumin and alphafetoprotein is that these proteins share many physical properties. Both proteins are approximately the same size (Peters et al., 1979), and there is some evidence of amino acid sequence homology, particularly at the carboxyl terminus (Ruoslahti and Terry, 1976). There is also a significant level of immunogenic cross-reactability between the denatured proteins (Ruoslahti and Engvall, 1976). It is therefore quite possible that these proteins arose by gene duplication of a common ancestor. For this reason is would be interesting to compare the structure of the albumin and alphafetoprotein mRNA's and genes.

We have purified the messenger RNA's encoding rat serum albumin (RSA) and rat alphafetoprotein (RAFP) and have used these as templates for the synthesis of radioactive cDNA by reverse transcription. Hybridization of RSA and RAFP cDNA probes to different preparations of polysomal RNA has been used to quantitate levels of these mRNA sequences in neonatal rat liver and in a hepatic tumor, Morris hepatoma 7777. The cDNA probes have also been hybridized to genomic DNA, both in solution and bound to filters in the form of "Southern blots" to characterize the RSA and RAFP genes in these cells.

RESULTS AND DISCUSSION

CHARACTERIZATION OF ALBUMIN AND ALPHAFETOPROTEIN MESSENGER RNA'S. RSA and RAFP mRNA's were purified by immunoprecipitation (Taylor and Schimke, 1973) of liver or hepatoma polysomes, respectively, followed by poly(U)-Sepharose 4B affinity chromatography and size fractionation on denaturing sucrose gradients, as described in Materials and Methods. These procedures yielded RNA's that were approximately 90-95% pure as determined by analysis of in vitro translation products (data not shown). Figure 1 illustrates the pattern obtained when the RSA (lane 4) and RAFP mRNA (lane 1) preparations were fractionated by electrophoresis on agarose gels in 10mM methylmercury hydroxide, which completely denatures RNA secondary structure (Bailey and Davidson, 1976). Both RNA species are represented by a single predominant band migrating slightly slower than the 18S rat ribosomal RNA marker. From this mobility we estimate that RSA and RAFP mRNA's have a molecular weight of approximately 765,000 daltons, or 2180 nucleotides. This includes a stretch of polyadenylic acid residues, the length of which was estimated by hybridizing an excess of radioactive polyuridylic acid to the purified mRNA and measuring the saturation values by ribonuclease treatment (data not shown). Such measurements indicated

Legend to Fisure 1. Electrophoresis of purified RAFP and RSA mRNA's. Various amounts of different RNA samples were electrophoresed on a 15 cm asarose slab sel containins 10 mM methylmercury hydroxide. Lane 1; 0.75 us RAFP mRNA. Lanes 2 and 5; 4 us of a mixture of 16S and 23S ribosomal RNA from E. coli and 18S and 28S ribosomal RNA from rat liver. Lane 3; 4 us of poly(A)-containins RNA from rat liver polysomes. Lane 4; 0.75 us of RSA mRNA.

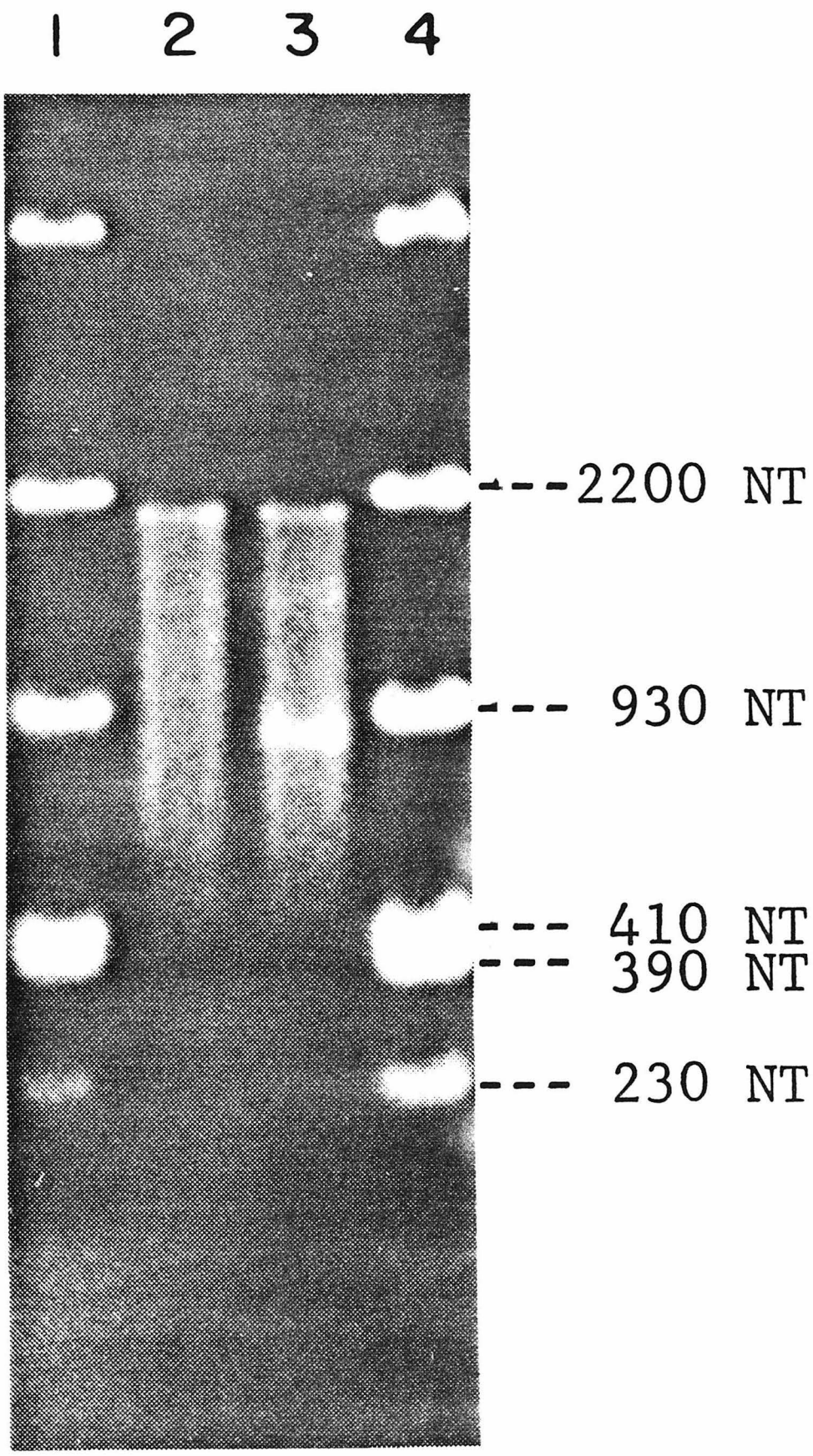

1 2 3 4 5
28 S
23 S
18 S
16 S

that these mRNA molecules have approximately 100-200 adenylate residues at their 3′ ends. Thus the non-poly(A) segments of these mRNA's are both approximately 2030 nucleotides in length.

Figure 2 (lanes 2 and 3) is an autoradiogram of [32-P]-labeled RAFP and RSA cDNA preparations that were fractionated by alkaline agarose gel electrophoresis. Since a large molar excess of oligo-dT was used to prime the reverse-transcription reaction, cDNA synthesis should have been initiated at the 5′ end of the poly(A) tract, and thus the size of "full length" cDNA should correspond the the non-poly(A) portions of the mRNA's. As expected, both cDNA preparations contain a distribution of different sized molecules that stops abruptly at approximately 2030 nucleotides. In addition, it is interesting to note the presence of a 900-nucleotide band in the albumin cDNA lane. This presumably results from premature termination of cDNA polymerization at a site located 900 nucleotides from the poly(A) tract. The nature of this termination site is unknown, but it does not seem to be present in RAFP mRNA.

Human and bovine serum albumins are 584 and 582 amino acid residues in length (Dayhoff, 1976), and rat albumin is of similar size (Chapter 3). There is also a 24-amino acid "pre-pro" peptide present on newly synthesized rat serum

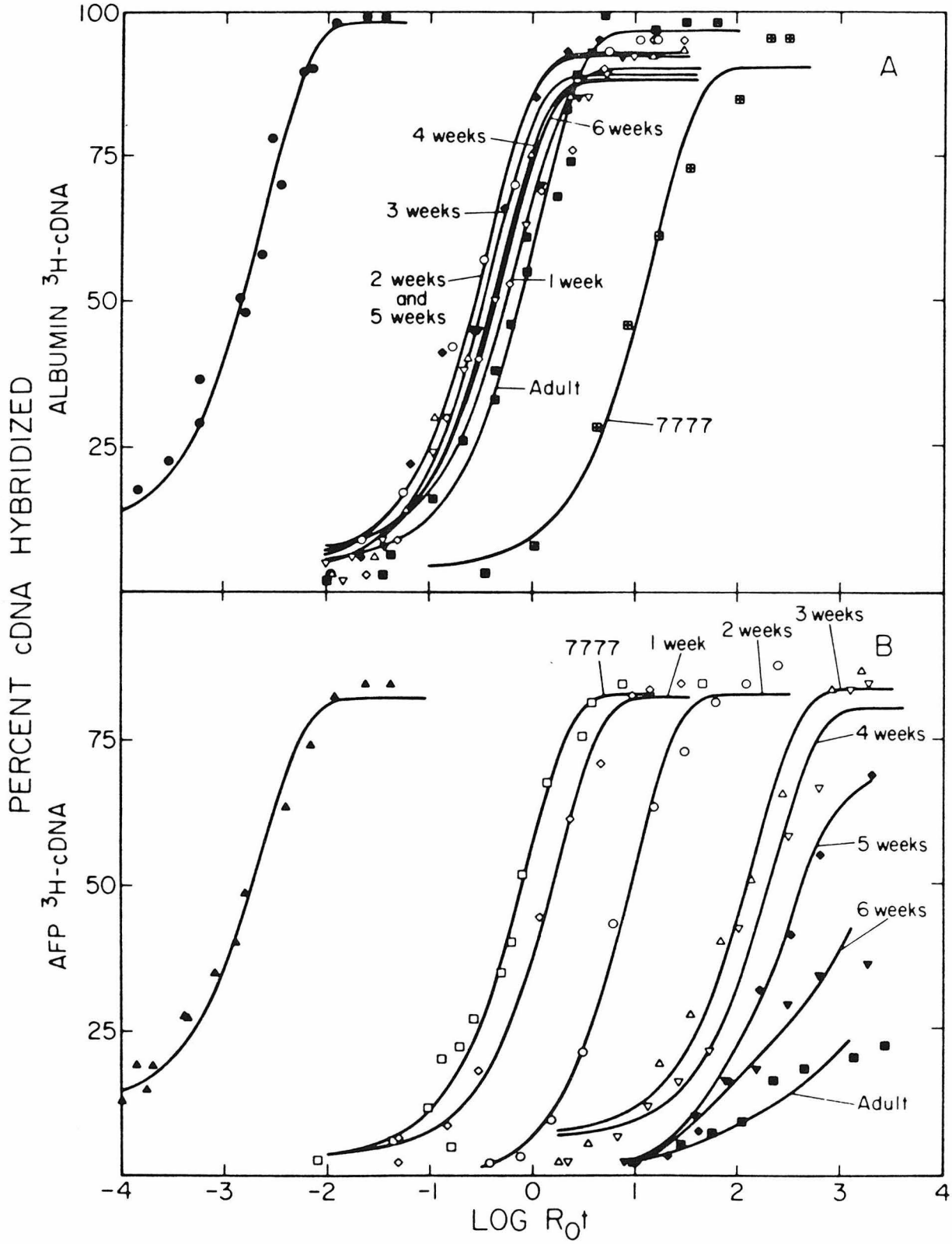
Legend to Figure 2. Electrophoresis of RAFP and RSA cDNA. [32-P]-labeled RAFP and RSA cDNA were prepared from the purified mRNA templates, as described in Materials and Methods. Aliquots were electrophoresed on a 1% agarose slab gel containing 30 mM NaOH, 2 mM EDTA as described (McConnel, Simon and Studier, 1977). Following electrophoresis, the gel was neutralized, dried and exposed to X-ray film. The resulting autoradiogram is shown in reverse contrast. Lanes 1 and 4; [32-P]-labeled Hind III fragments of PM2 DNA used as markers. The molecular weights of these fragments were estimated by comparison to various restriction endonuclease fragments of ØX174 DNA, and are listed in nucleotides at the right. Lane 2; RAFP cDNA. Lane 3; RSA cDNA.


1
2
3
4
2200 NT
930 NT
410 NT
390 NT
230 NT

albumin (Strauss et al., 1977) that is cleaved away during the secretion process. This total of 609 amino acids would be encoded by 1827 nucleotides, so RSA mRNA probably contains approximately 200 nucleotides of untranslated sequence, in addition to the poly(A). Human AFP is also approximately 585 amino acids in length, and possesses a "pre-pro" peptide (Peters et al., 1979), so a similar amount of untranslated sequence should be found on its mRNA.

ABUNDANCE LEVELS OF ALBUMIN AND ALPHAFETOPROTEIN mRNA IN DEVELOPING RAT LIVER AND IN MORRIS HEPATOMA 7777. If reasonably homogeneous mRNA is available, it is possible to quantitate the concentration of homologous sequences in a sample of RNA by measuring the rate at which a radioactive cDNA probe hybridizes to a vast molar excess of that RNA. Polysomal RNA was purified from the livers of neonatal rats aged 1, 2, 3, 4, 5, and 6 weeks, from adult rat livers and from a tumor, Morris hepatoma 7777. These RNA's were used to "drive" RAFP and RSA cDNA probes. The extent of reaction was measured by digestion of unreacted cDNA with S1 nuclease. The kinetic parameters were evaluated according to Equation 1 by a computerized least squares approach (Materials and Methods). The hybridization data along with the best fitting solution to the pseudo first order rate equation are presented in Figure 3. The rate constants corresponding to the curves shown in this figure are given

Legend to Fisure 3. Polysomal RNA-driven hybridization of RSA (Panel A) and RAFP (Panel B) cDNA's. RNA excess hybridizations were carried out as described in Materials and Methods and assayed by S1 nuclease disestion. Total polysomal RNA was isolated from the livers of neonatal rats ased 1 week (◇), 2 weeks (○), 3 weeks (△), 4 weeks (▽), 5 weeks (◆) and 6 weeks (▼), and from adult rat liver (■) and from Morris hepatoma 7777 (⊞) as described. Hybridization of the cDNA to purified RSA (●) and RAFP (▲) mRNA's are also included for comparison. The curves represent the best fittins solutions to Equation 1.


PERCENT cDNA HYBRIDIZED
ALBUMIN ^{3}H-cDNA
AFP ^{3}H-cDNA
LOG R_0t
A
B
4 weeks
6 weeks
3 weeks
2 weeks and 5 weeks
1 week
Adult
7777
100
75
50
25
-4
-3
-2
-1
0
1
2
3
4
5 weeks

in Table 1.

Both purified mRNA's hybridize to their cDNA copies at a rate of approximately 400 $M^{-1}sec^{-1}$. Taking the hybridization of single-stranded ØX174 DNA to a molar excess of complementary RNA as a standard (k=170 $M^{-1}sec^{-1}$, complexity=5385 nucleotides, Galau, Britten and Davidson, 1977), this rate corresponds to a complexity of 2300 nucleotides, which is reasonably consistent with our other measurements. It was not possible to detect any cross-reaction of RSA mRNA to RAFP cDNA (or vice versa), even at reduced criteria (data not shown). This means that these cDNA probes are sequence-specific, and that the level of homology between the two mRNA's must be less than 70-80% overall. (The actual extent of RSA-RAFP homology is 40-50% at the DNA level; Innis and Miller, 1980, Chapter 3 and our unpublished data).

The measurements of RSA and RAFP mRNA levels in liver polysomes (Table 1) reveals that following birth, there is little change in the concentration of albumin mRNA whereas RAFP mRNA has virtually disappeared by 5-6 weeks. These mRNA levels coincide with the concentration of albumin and alphafetoprotein in the serum. the latter declines from 2-3 mg/ml to less than 100 ng/ml by age 6 weeks (Sell and Becker, 1978). Morris hepatoma 7777 synthesizes and

Lesend to Table 1.

(a) Total ribosomal RNA isolated from the source indicated, or purified mRNA. (b) Pseudo first-order rate constant corresponding to best fitting solution of rate equation 1. (c) Calculated by dividing the rate of the polysomal RNA-driven reaction by the rate of the appropriate mRNA-driven reaction. Values in parentheses are crude estimates corresponding to unterminated reactions.

Table 1. Polysomal RNA-cDNA Hybridization Parameters.

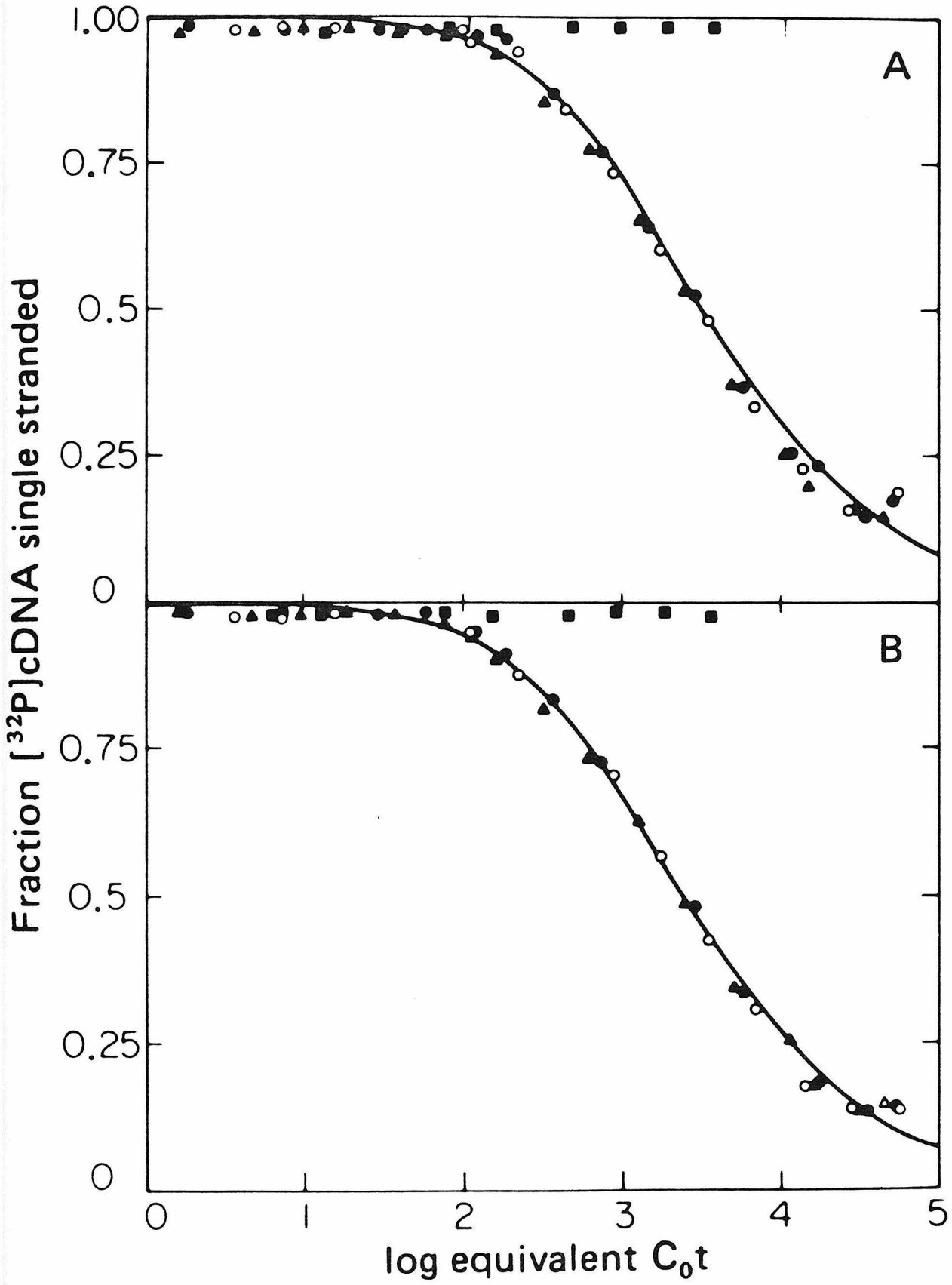
Driver Source(a)	Tracer	Rate ($M^{-1}sec^{-1}$) (b)	Fraction Drivins(c)
RSA mRNA	RSA cDNA	4.10×10^{2}	1.00
1 wk. liver	"	1.33×10^{0}	3.24×10^{-3}
2 wk. liver	"	2.77×10^{0}	6.76×10^{-3}
3 wk. liver	"	2.24×10^{0}	5.46×10^{-3}
4 wk. liver	"	1.73×10^{0}	4.22×10^{-3}
5 wk. liver	"	2.77×10^{0}	6.76×10^{-3}
6 wk. liver	"	1.73×10^{0}	4.22×10^{-3}
Adult liver	"	1.02×10^{0}	2.49×10^{-3}
Hepatoma	"	7.07×10^{-2}	1.72×10^{-4}
RAFP mRNA	RAFP cDNA	3.98×10^{2}	1.00
1 wk. liver	"	6.08×10^{-1}	1.53×10^{-3}
2 wk. liver	"	1.02×10^{-1}	2.56×10^{-4}
3 wk. liver	"	8.42×10^{-3}	2.12×10^{-5}
4 wk. liver	"	5.01×10^{-3}	1.26×10^{-5}
5 wk. liver	"	(2.0×10^{-3})	(5.0×10^{-6})
6 wk. liver	"	$(< 10^{-3})$	$(< 10^{-6})$
Adult liver	"	$(< 10^{-4})$	$(< 10^{-7})$
Hepatoma	"	1.26×10^{0}	3.17×10^{-3}

secretes large quantities of AFP and a somewhat lesser amount of albumin. As expected, these mRNA species are present at high levels in the polysomal fraction from this tumor. The RAFP mRNA is approximately 18 times more prevalent than RSA mRNA. Tse, Morris and Taylor (1978) have made similar measurments. These authors found a 3-4 fold difference in mRNA concentrations, which agrees more closely with the different RAFP and RSA sythesis rates in this tumor.

PRELIMINARY CHARACTERIZATION OF THE ALBUMIN AND ALPHAFETOPROTEIN GENES. Before undertaking to elucidate the fine structure and regulatory mechanisms of the RSA and RAFP genes, it is desirable to ascertain the number of copies of these sequences in the rat genome, i.e. the reiteration frequency. It is also important to address the possibility of rearrangements or amplifications of genetic DNA associated with alterations in expression, especially in tumor cells.

The rate at which sheared genomic DNA "drives" tracer quantities of a cDNA probe is directly proportional to the genomic reiteration frequency of the relevant gene (Britten, Graham and Neufeld, 1974). Figure 4 shows the results of such an experiment with RSA(A) and RAFP(B) cDNA's. Milligram quantities of rat liver, rat kidney and Morris hepatoma nuclear DNA sheared to 300-400 nucleotides were

Legend to Figure 4. Association reactions of RSA cDNA (Panel A) and RAFP cDNA (Panel B) with nuclear DNA from rat liver (●), rat kidney (○) and from Morris hepatoma 7777 (▲). Reaction mixtures of approximately 1.5 mg of sheared nuclear DNA and 5-10 ng of [32-P]-labeled cDNA were prepared, denatured and incubated to various values of equivalent Cot as described in Materials and Methods. (■); "mock" association reactions with Micrococcus lysodeikticus DNA. Duplex formation was assayed by digestion with S1 nuclease. The curves represent the best fitting solutions to Equation 2.


1.00
0.75
0.5
0.25
0
A
B
Fraction [32P]cDNA single stranded
0
1
2
3
4
5
log equivalent C0t

denatured and allowed to reassociate in the presence of radioactive cDNA. It is important that the nonradioactive "driver" DNA is present in large excess relative to the cDNA probe. A sufficient mass ratio of "driver" to "tracer" was employed to assure that the molar ratio of genes to homologous cDNA sequences was at least 40. The second order rate constants were evaluated by a least squares approach according to Equation 2 (Materials and Methods) and are listed in Table 2. Two conclusions can be drawn from these data. First, there is no discernible difference in the concentration of either RSA or RAFP sequences in any of the driver DNA's utilized, and therefore the enormous differences in the level of expression of these genes in liver, kidney and tumor cells are not due to gene amplification or deletion. Second, the observed reassociation rate constants for RSA and RAFP cDNA tracers (Table 2) are approximately three to four times higher than the observed rate constant of the "single-copy" component of rat DNA (Pearson, Wu and Bonner, 1978), which suggests that these genes are present at approximately 3-4 copies each per haploid genome.

Another method for estimation of gene reiteration frequency is to hybridize increasing amounts of pure, radioactive cDNA to a constant mass of genomic DNA. The saturation plateau of such a reaction should correspond to

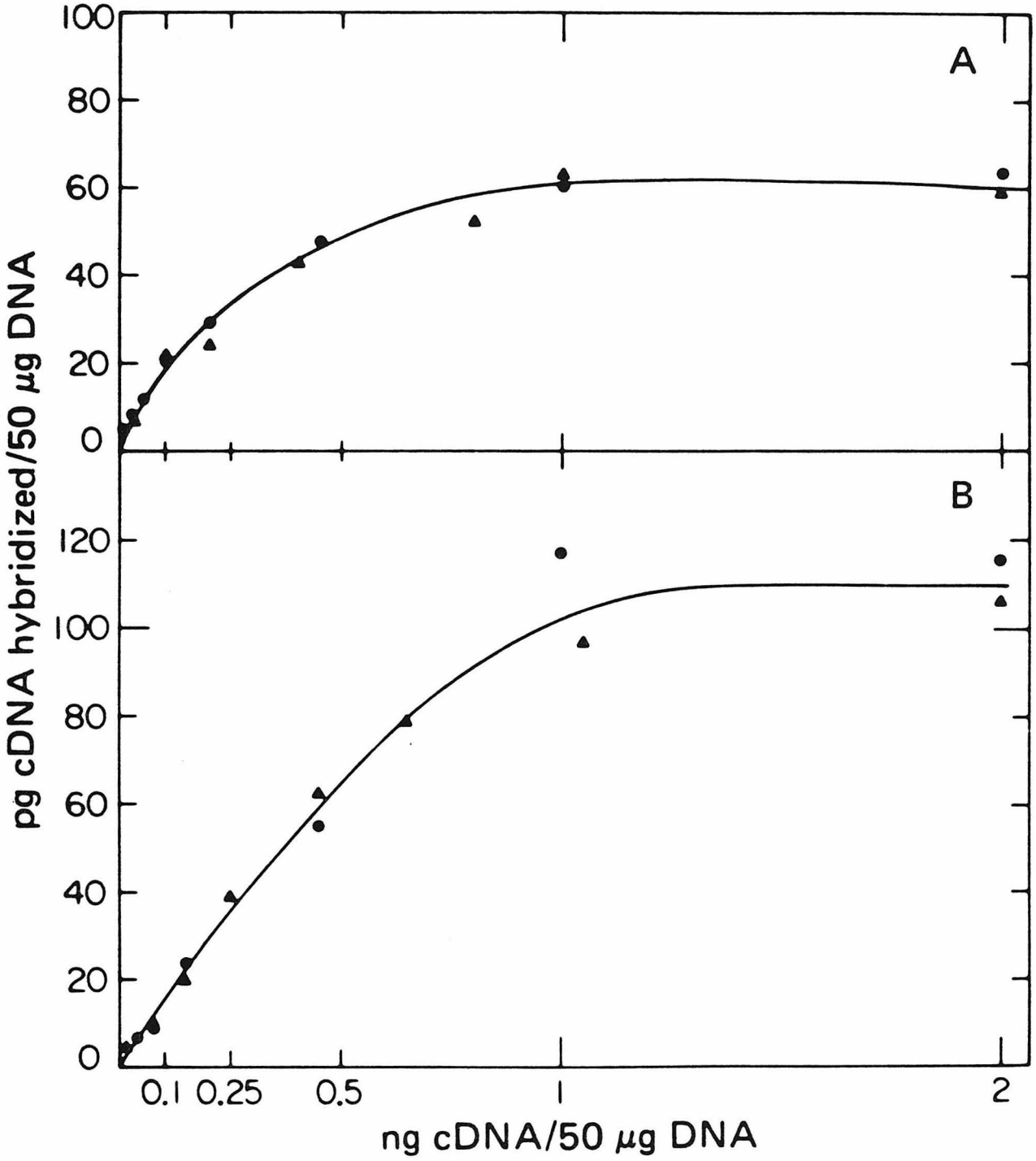
Table 2. Rates of Reassociation of cDNA with Nuclear DNA

Driver(a)	Tracer	Rate ($M^{-1}sec^{-1}$)(b)	Copies/genome(c)
Kidney DNA	RSA cDNA	1.17×10^{-3}	3.5
Liver DNA	"	1.11×10^{-3}	3.3
Hepatoma DNA	"	1.12×10^{-3}	3.3
Kidney DNA	RAFP cDNA	1.32×10^{-3}	3.9
Liver DNA	"	1.34×10^{-3}	4.0
Hepatoma DNA	"	1.42×10^{-3}	4.2

(a) Nuclear DNA isolated, sheared and hybridized to cDNA as described in Materials and Methods. (b) Best fitting solution to rate equation 2. (c)Number of copies per haploid rat genome calculated by dividing the observed rates by the published rate for the "single-copy" component of rat liver DNA; 3.9×10^{-4} (Pearson, Wu and Bonner, 1978).

the fraction of genomic sequences that is complementary to the cDNA probe. Figure 5 displays the results of such saturation hybridization experiments. RSA and RAFP cDNA's were used to titrate sheared, denatured nuclear DNA from rat liver and Morris hepatoma 7777, and hybrids were assayed by their resistance to S1 nuclease (Monahan, Harris and O'Malley, 1976). Conversion from the plateau values in radioactivity to mass of hybridized cDNA was done by assuming that the specific activity of the cDNA was equal to that of the radioactive precursors as specified by the manufacturer. Thus it can be calculated that there are 60 pg of RSA-encoding sequences per 50 ug of either rat liver or hepatoma DNA, a fractional value of 1.2×10^{-6}. Since only one strand of the albumin gene is capable of hybridizing to the cDNA, this fraction should be multiplied by two. The haploid rat genome consists of approximately 2.9×10^{9} base pairs, so there are $1.2 \times 10^{-6} \times 2 \times 2.9 \times 10^{9}$ = 6720 base pairs of RSA-encoding sequence per haploid genome. This is equivalent to about three copies of the 2130 nucleotide mRNA sequence, which is in good agreement with the hybridization kinetic data. Similarly, there are approximately 5 copies of the RAFP gene according to the saturation plateau level. This is somewhat higher than the reiteration frequency as inferred from the kinetic data. This may indicate that there are more RAFP than RSA

Legend to Fisure 5. Saturation hybridization of RSA cDNA (Panel A) and RAFP cDNA (Panel B) to nuclear DNA from rat liver (●) and from Morris hepatoma 7777 (▲). Increasins amounts of [32-P]-labeled cDNA were hybridized to a constant mass (50 us) of sheared, denatured nuclear DNA as described in Materials and Methods. Duplex formation was assayed by disestion with S1 nuclease.


A
B
100
80
60
40
20
0
120
100
80
60
40
20
0
pg cDNA hybridized/50 μg DNA
0.1
0.25
0.5
1
2
ng cDNA/50 μg DNA

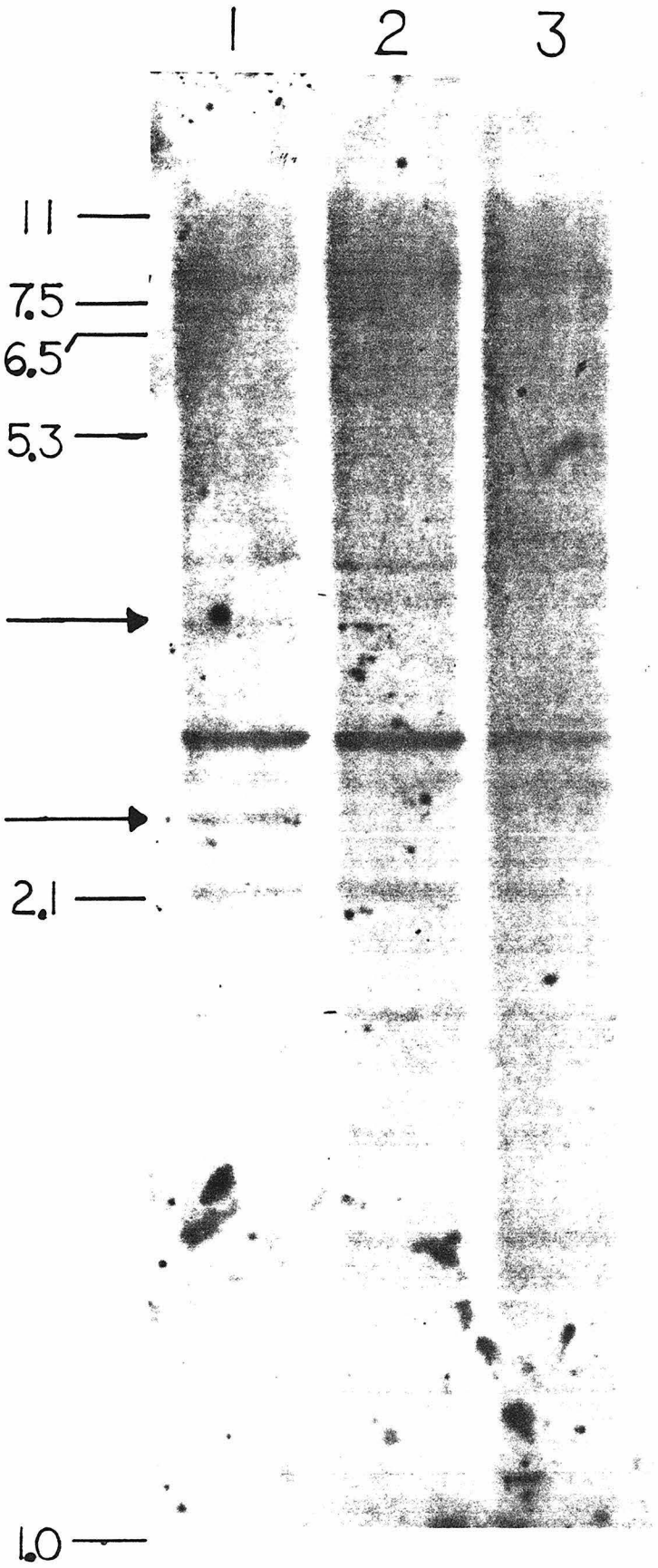
sequences in the rat genome, or that the RAFP cDNA is more contaminated with other mRNA sequences that the RSA cDNA. Again, there are no observed differences between tumor and liver DNA.

These values for the absolute reiteration frequency of the RSA and RAFP genes should not be taken too literally. While there may in fact be multiple versions of the RSA and RAFP genes, there are uncertainties inherent in both the kinetic and saturation measurements that complicate the interpretation of these data. The reassociation reactions were assayed by S1 nuclease digestion, which introduces considerable deviation from ideal second order kinetics (Smith, Britten and Davidson, 1975). The "single copy" renaturation reaction was assayed by the hydroxylapatite chromatography method, which is sensitive primarily to the nucleation of double-stranded structures, and therefore the form of the observed single-copy renaturation reaction is much closer to second order. Furthermore, the RSA (see Chapters 2 and 4) and the mouse AFP (Tilghman et al., 1979) genomic mRNA sequences are interrupted by intervening sequences, and this could have an effect on the kinetics of the nuclear DNA/cDNA driver-tracer reaction. The saturation hybridization method is rather sensitive to contaminating sequences in the cDNA driver. Interpretation of the results of such experiments also depends on accurate knowledge of

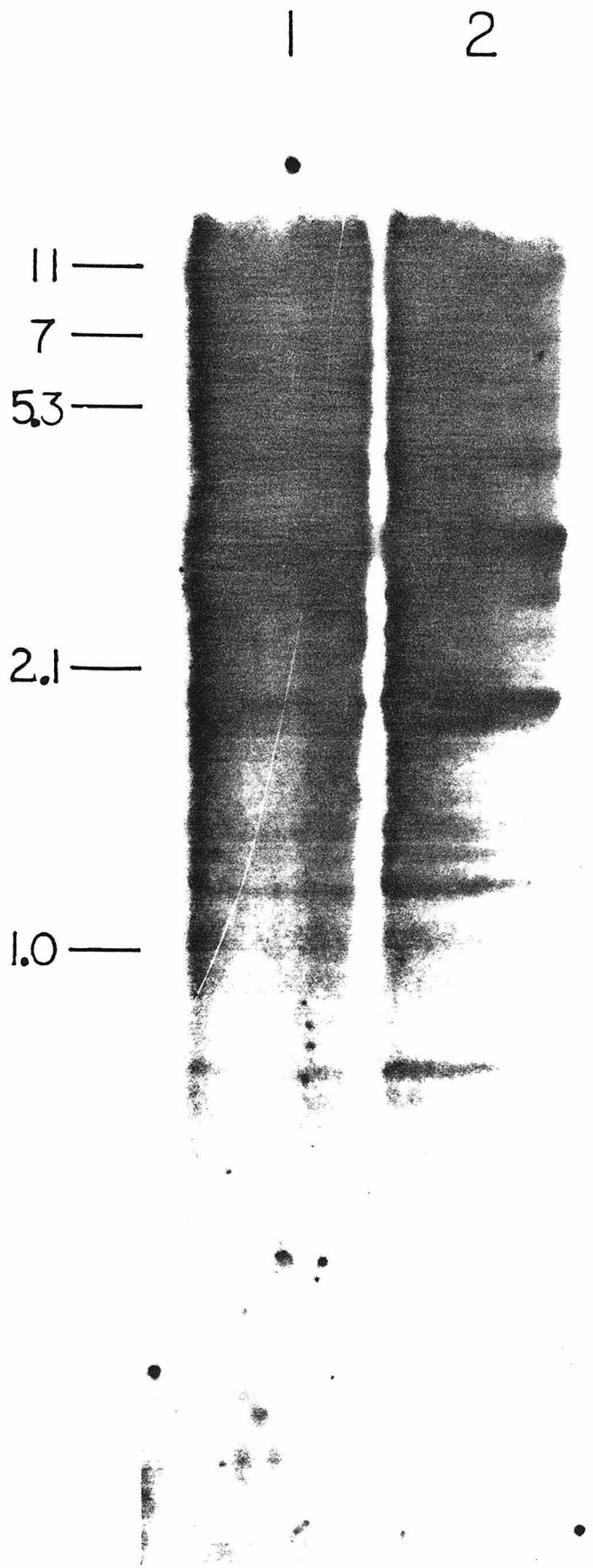
the cDNA specific activity, which we have not measured directly. These uncertainties might lead to significant overestimates or underestimates in gene copy number. Better controls, including internal standards, would undoubtedly improve the accuracy of these determinations, but we feel that the problem of detailed characterization of these and other eukaryotic genes is much more effectively approached by means of recombinant DNA methodology. This is the subject of the ensuing chapters. For the moment the conclusion that the RSA and RAFP genes are "approximately single-copy" is sufficiently precise.

Although gene amplification is clearly not responsible for the stimulation of RSA and RAFP production in liver and tumor cells, it is possible that some form of gene rearrangement might be involved, akin to that observed during differentiation of antibody-producing cells (e.g., Brack et al., 1978). Substantial rearrangements of genomic DNA should be reflected in the pattern of restriction endonuclease sites in the vicinity of the effected gene. This pattern can be detected by the "Southern blot" technique (Southern, 1975), and the results of such experiments are shown in Figures 6 and 7. Lanes 1, 2 and 3 of Figure 6 are blots of liver DNA from Sprague-Dawley rats, liver DNA from Buffalo rats and DNA from Morris hepatoma 7777, respectively. All were digested with Eco R1,

Legend to Figure 6. Filter hybridization of RSA cDNA to Eco R1 fragments of Sprague-Dawley rat liver DNA (lane 1), Buffalo rat liver DNA (lane 2) and Morris hepatoma 7777 DNA (lane 3). Restriction endonuclease digestion, "Southern" transfer, and hybridization to [32-P]-cDNA were carried out as described in Materials and Methods. The numbers at the left correspond to the positions and molecular weight (in KB) of marker restriction fragments that were co-electrophoresed with the genomic DNA and visualized by ethidium bromide staining. The arrows point to the positions of the bands at 2.7 and 3.9 KB that were present in the Sprague-Dawley rat liver DNA and missing in the other two samples.


1
2
3
11
7.5
6.5
5.3
2.1
1.0

Legend to Figure 7. Filter hybridization of RAFP cDNA to Eco R1 fragments of Buffalo rat liver DNA (lane 1) and Morris hepatoma 7777 DNA (lane 2), as in Figure 6.


1
2
11
7
5.3
2.1
1.0

electrophoresed, transferred to nitrocellulose filters and hybridized with [32-P]-labeled RSA cDNA as described in Materials and Methods. Many bands appear in the autoradiograms, most of which are identical in all three lanes. No differences occur between the Buffalo rat liver and Morris hepatoma DNA's. A few bands are unique to the Sprague-Dawley rat DNA, and this may be due to interstrain heterogeneity at the genomic level. Figure 7 shows a similar experiment comparing Buffalo rat liver DNA and Morris hepatoma 7777 DNA blots hybridized to [32-P]-labeled RAFP cDNA. Again, no differences are detected, and we conclude from this that major rearrangements of the RSA or the RAFP genes do not take place during hepatocarcinogenesis, at least not in this particular tumor. Many of the large number of bands visualized are apparently due to the small amount of contaminating sequences in our cDNA preparations. These contaminants are probably hybridizing to a variety of completely different genes that are expressed in liver or tumor cells. This conclusion is based upon the observations, described in Chapter 2, that cloned RSA cDNA sequenes hybridize only to six different Eco R1 fragments of rat genomic DNA. A similar observation has been made concerning RAFP sequences (unpublished data).

In summary, the concentration of the mRNA encoding RAFP has been found to decline rapidly in the livers of neonatal

rats. This decline coincides with the extinction of circulating AFP in this animal. RSA mRNA concentrations do not change substantially during this period.

The genes that encode RSA and RAFP are included within the "single-copy" fraction of the rat genome, although there is some indication that each may be repeated three or four times. There are no differences in these reiteration frequencies when DNA from rat liver, rat kidney and from Morris hepatoma 7777 are compared. Nor is there any detectable difference in the gross organization of the RSA or RAFP genes when DNA from normal liver and tumor are compared by the Southern blot technique.

Thus we conclude that the dramatic alterations in RSA and RAFP expression that take place during development and neoplastic transformation are due to regulation of the level of transcription of these genes into stable mRNA.

MATERIALS AND METHODS

MATERIALS. Poly(U) Sepharose 4B and cyanogen bromide-activated Sepharose 4B were obtained from Pharmacia. Oligo(dT) was purchased from Collaborative Research. Actinomycin D and nonradioactive deoxynucleoside triphosphates were obtained from P/L Biochemicals. Avian Myeloblastosis Virus Reverse Transcriptase was a gift from Dr. J. Beard, Life Sciences Inc., St. Petersburg, FL. Methylmercury hydroxide was from Alfa Chemicals, Ventron Corp., Beverly, MA. PM2 DNA was a gift from Dr. H.B. Gray, Jr., University of Houston, TX. Eco R1 was purchased from Boehringer-Mannheim and used according to the manufacturer's instructions. Nitrocellulose membrane filter paper was from Millipore. [32-P]-nucleoside triphosphates were purchased from either New England Nuclear or Amersham. [3-H]-labeled nucleoside triphosphates were purchased from New England Nuclear.

ANIMALS AND HEPATOMAS. Male rats of the Sprague-Dawley strain were purchased from Simonsen Labs, Inc. (Gilroy CA), and were used for isolation of albumin mRNA and for preparation of liver RNA fractions from different stages of development. Morris hepatoma 7777 was maintained and transplanted into both gastrocnemius muscles of mature

Buffalo strain rats as described elsewhere (Sell et al., 1974). Only exponentially growing tumors were used and special care was taken to remove any necrotic or infected tissue.

PREPARATION AND PURIFICATION OF ANTIGENS AND ANTIBODIES. Rat serum albumin (RSA) was purified from "Fraction V" (Sigma) according to published procedures (Taylor and Schimke, 1973). The final RSA preparation was shown to be homogeneous by SDS gel electrophoresis and by double-diffusion Ouchterlony analysis. Analytical amounts of immunochemically pure rat alphafetoprotein (RAFP) were originally obtained from rat amniotic fluid by repeated fractionation on an isoelectric focusing column (Sell et al., 1972). This preparation was used to raise monospecific antibodies in rabbits. The Immunoglobulin G (IgG) fraction from the resulting antisera were coupled to cyanogen bromide activated Sepharose 4B (Cuatrecasas, 1970) and used as an immunoadsorbent for large scale purification of RAFP.

The purified antigens were also used to raise antibodies in goats. The IgG fractions from the resulting antisera was further purified by adsorption to RSA or RAFP bound to Sepharose 4B. The adsorbed antibody was eluted with sterile 0.2 M glycine, pH 2.7. The eluant was neutralized, precipitated with ammonium sulfate and dialyzed

against 50 mM tris, pH 7.6, 150 mM NaCl, 5 mM $MgCl_2$.

A nonimmune goat IgG fraction was injected subcutaneously into a burro. The burro (anti-goat IgG) IgG was then purified by ammonium sulfate fractionation and made ribonuclease-free by chromatography on CM cellulose and DEAE cellulose (Schimke et al., 1974).

PREPARATION OF POLYSOMES. Polysomes were prepared from normal rat liver or from Morris hepatoma 7777 essentially as described by Schimke et al. (1974). Seventeen mL of the postmitochondrial supernatant was layered onto 1 mL of 1 M sucrose in 25 mM tris, pH 7.5, 25 mM NaCl, 5 mM $MgCl_2$, 0.5 mg/ml heparin sulfate over 15 mL of 2.5 M sucrose in the same buffer. The polysomes were pelleted onto the 2.5 M sucrose cushion by centrifugation at 25,000 rpm for 4 hr at 4°C. and removed from this interface by puncturing the side of the tube with a sterile syringe and needle. Polysomes were used immediately either for immunoprecipitation or for extraction of total polysomal RNA.

ISOLATION OF SPECIFIC POLYSOMES BY IMMUNOPRECIPITATION. The polysomes removed from the 2.5 M sucrose interface were dialyzed overnight at 4°C. against a large volume of 50 mM tris, pH 7.6, 150 mM NaCl, 5 mM $MgCl_2$, 100 ug/ml heparin. The dialysate was centrifuged in a Beckman SW 27 rotor for

10 min at 11,000 rpm to remove aggregates. The clarified polysome supernatant usually had an optical density at 260 nm of 10-15. 74 ug/mL goat anti-RSA or 50 ug/mL goat anti-RAFP were added and incubated for 45 min at 0°C. The resulting polysome-antibody complexes were precipitated by adding a 48 fold (RSA) or 40 fold (RAFP) excess of burro anti-goat IgG and incubation for an additional 90 min with occasional gentle mixing. The mixture was then layered onto 7.5 mL of 0.5 M sucrose over 14 mL of 1.0 M sucrose, both buffered with 50 mM tris, pH 7.6, 150 mM NaCl, 5 mM $MgCl_2$ plus 1% each Triton X100 and sodium deoxycholate and centrifuged for 16 min at 11,000 rpm in a SW 27 rotor at 2°C. The pellet was suspended in the same buffer and re-sedimented through another discontinuous sucrose gradient, as before.

PURIFICATION OF RSA AND RAFP mRNA. RNA was extracted from immunoprecipitated polysomes as previously described (Sala-Trepat et al., 1978). Ethanol-precipitated RNA was collected by centrifugation, washed with 66% ethanol, dried by lyophilization and dissolved in 20 mM tris, pH 7.5, 2 mM EDTA and 1% SDS. The RNA was heated at 65°C. for 10 min and then chilled in an ice bath. The poly(A) RNA was bound in this buffer to a poly(U)-Sepharose 4B column and eluted with 90% formamide containing 10 mM tris, pH 7.5, 1 mM EDTA

and 0.2% SDS. The eluate was adjusted to 0.2 M NaCl and precipitated with three volumes of ethanol overnight at -20°C. Rechromatography of the poly(A) RNA over a second poly(U) Sepharose column was generally performed to remove residual ribosomal RNA contamination.

RSA and RAFP mRNA's were further purified from the poly(A)-containing RNA preparations by sedimentation through linear 5-20% sucrose gradients containing 50% dimethyl sulfate. The RNA was dissolved in 600 uL of 40 mM tris, pH 7.5, 50% dimethyl sulfate, 0.1 M LiCl, 5 mM EDTA and 0.5% SDS and heated at 40°C. for 10 min. The samples were then layered onto 12 mL gradients and centrifuged for 20 hr at 41,000 rpm in a Beckman SW41 rotor. The peak of RNA sedimenting at 17S was collected and precipitated with ethanol.

DETERMINATION OF THE POLY(A) CONTENT OF RNA PREPARATIONS. The content of poly(A) in the different RNA preparations was assayed by hybridization to [3-H]-poly(U) essentially as described by Gillespie, Marshall and Gallo (1972).

ELECTROPHORESIS OF RNA AND DNA ON DENATURING AGAROSE GELS. Fractionation of RNA on 1.2% agarose slab gels containing 10 mM methylmercury hydroxide was as described by Bailey and Davidson (1976). After electrophoresis for approximately 9

hr at 4 volts/cm, the gels were rinsed in 1 M ammonium acetate and stained for 45 min at room temperature with a solution of 1 M ammonium acetate, 1 ug/mL ethidium bromide. Electrophoresis of DNA on 1% agarose gels containing 30 mM NaOH, 2 mM EDTA was done as described by McConnel et al. (1977). Following electrophoresis, gels were neutralized by soaking in 1 M Na acetate, pH 5.0 for 30 min, and then dried onto a piece of Whatman 3MM filter paper on a heated suction platform.

SYNTHESIS OF COMPLEMENTARY DNA (cDNA) FROM RSA AND RAFP mRNA TEMPLATES. cDNA's were synthesized in 25-100 uL reaction volumes containing the following: 50 mM tris, pH 8.4, 10 mM $MgCl_2$, 60 mM NaCl, 20 mM dithiothreitol, 100 ug/mL Actinomycin D, 50 ug/mL oligo-dT, 100 ug/mL bovine serum albumin, 20 ug/mL mRNA, and 300-600 units/mL reverse transcriptase. The reactions also contained all four deoxynucleoside triphosphates at a concentration of 0.5 mM for nonradioactive and 0.025 mM for radioactive dNTP's. The reactions were incubated at 46°C. for 1 hr. RNA was then hydrolyzed by treatment with 0.5 N NaOH at 70°C. for 20 min. Unincorporated dNTP's, degraded RNA and other low molecular weight materials were removed by chromatography over Sephadex G200. The excluded fractions were precipitated with ethanol.

RNA EXCESS HYBRIDIZATIONS TO RADIOACTIVE RSA AND RAFP cDNA. RNA excess hybidizations to [3-H]-labeled cDNA was carried out in 10-50 uL reaction volumes containing 10 mM tris, pH 7.4, 0.18 M NaCl, 1 mM EDTA, 0.1% SDS, 3000 cpm of the appropriate cDNA (sp. act. 1.4×10^6 cpm/ug) and varying amounts of RNA. The reaction mixtures were sealed in silanized capillary tubes and incubated at 68°C. for varying times, up to 72 hr. Hybrid formation was detected by treatment of samples with S1 nuclease (Monahan, Harris and O'Malley, 1976). The data are plotted as percent of the cDNA probe in hybrid versus log Rot, where "Rot" is the product of the RNA concentration in moles of nucleotides per liter and the incubation time in seconds. These data were analyzed by use of a computer program that identifies the best fitting solution to Equation 1:

$$c/Co = \exp(-kRot)$$

by a least-squares algorithm (Pearson, Davidson and Britten, 1977). "c/Co" is the fraction of probe remaining unreacted at time "t" (in sec), k is the first order rate constant (in $M^{-1}sec^{-1}$) and Ro is the molar concentration of RNA.

PREPARATION OF NUCLEAR DNA. Nuclear DNA to be used in solution hybridizations was prepared from rat liver, rat

kidney and Morris hepatoma 7777 by a modification of the citric acid procedure of Hisashi et al. (1966). The nuclear pellets were dissolved in 100 mM NaCl, 1 mM EDTA, 1% SDS, 10 mM tris, pH 8.0. Proteinase K was then added to a concentration of 100 us/mL and the samples were incubated at 37°C. for 2 hr. The buffer was then adjusted to 0.1 M tris, pH 8.0, and the solutions were extracted 2-3 times with 1 vol of phenol/chloroform (1:1) and the DNA was spooled from the aqueous phase after addition of 2 vol of cold ethanol. This DNA was dissolved in 50 mM tris, pH 7.5, 100 mM NaCl, 2 mM EDTA and treated with 50 us/mL of pancreatic ribonuclease A for 90 min at 37°C. SDS and proteinase K were then added to final concentrations of 0.5% and 50 us/mL, respectively, and incubation was continued for another 60 min at 37°C. The solution was then extracted with phenol/chloroform and precipitated with ethanol. Purified DNA was dissolved in 10 mM Na acetate, pH 6.8 and sheared in a VirTis model 60 homogenizer to a weight average length of approximately 300-400 nucleotides. Fragment sizes of the final preparations were measured on 5-20% alkaline sucrose gradients.

NUCLEAR DNA/cDNA REASSOCIATION REACTIONS UNDER CONDITIONS OF DNA EXCESS. DNA/cDNA reactions were performed in 10-50 uL volumes in silanized capillary tubes containing 0.5 M NaCl,

1 mM EDTA, 20 mM tris, pH 7.5, 0.1% SDS, 1000-2000 cpm of [32-P]-labeled cDNA (sp. act. 2.2×10^8 cpm/ug) and varying amounts of sheared nuclear DNA. The reaction mixtures were denatured by heating at 105°C. for 4 min and were incubated at 69°C. for various times up to 72 hr. The extent of reaction was measured by S1 nuclease digestion (Monahan, Harris and O'Malley, 1976). Cot values were corrected to compensate for the high salt concentration (Britten, Graham and Neufeld, 1974). The data were evaluated according to Equation 2:

$$c/Co = P(1+kCot)^{-0.44}$$

by a computerized least squares approach, as above. P is the fraction of the cDNA in duplex form at termination, c/C0 is the fraction unreacted at time "t", Cot is the product of DNA concentration in M nucleotides and incubation time in sec., corrected for the salt concentration, as above, and k is the second order rate constant. The exponent -0.44 is used to correct for the non-second order kinetics of reactions assayed by the S1 nuclease method (Smith, Britten and Davidson, 1975, Morrow, 1974).

SATURATION HYBRIDIZATION EXPERIMENTS. Fifty micrograms of nuclear DNA was mixed with increasing amounts of either RSA

or RAFP cDNA labeled with [32-P]-dCTP to a specific activity of 5×10^7 cpm/ug in 10 uL of 0.5 M NaCl, 1 mM EDTA, 0.1% SDS, 20 mM tris, pH 7.5. The reaction mixtures were sealed in silanized capillary tubes, denatured by heating at 105 C. for 4 min and incubated at 69 C for 65 hr. Duplex formation was assayed by digestion of unreacted cDNA with S1 nuclease (Monahan, Harris and O'Malley, 1976), and background resistance to S1 was measured by digestion of "mock" reactions containing 50 ug of Micrococcal DNA rather than rat DNA.

SOUTHERN BLOTS OF RAT DNA. High molecular weight cellular DNA was extracted from rat livers and from Morris hepatoma 7777 by minor modifications of the method described by Blin and Stafford (1976). The DNA preparations obtained were digested three times by a three-fold excess of the restriction endonuclease Eco R1. The resulting fragments were separated by electrophoresis on 0.7% agarose gels containing 0.1 ug/mL of ethidium bromide, and the DNA was visualized under ultraviolet light.

FILTER HYBRIDIZATIONS. DNA fragments were transferred to nitrocellulose filters by minor modifications of the technique of Southern (1975). The filter strips were hybridized to 20 ng/mL [32-P]-labeled RSA or RAFP cDNA (sp.

act. 2×10^8 cpm/ug) at 65°C. for 20 hr in several mL of hybridization solution containing 1 M NaCl, 10-fold concentrated Denhardt's solution (Denhardt, 1966), 6 mM EDTA, 0.1% SDS, 25 ug/mL denatured, sheared E. coli DNA and 90 mM tris, pH 7.2. After hybridization, the filters were washed at 63°C. in 75 mM NaCl, 15 mM tris, pH 7.2, 1 mM EDTA, 0.1% SDS, 0.1% Na pyrophosphate and exposed to preflashed Kodak XR-5 X-ray film with a Cronex Lightning Plus intensifier screen for 24-48 hr at -80°C.

ACKNOWLEDGEMENTS

The research described in the preceding Chapter has been published as: Sala-Trepat, J.M., Sarsent, T.D., Sell, S. and Bonner, J. (1979) Proc. Natl. Acad. Sci. USA 76, 695-699, and Sala-Trepat, J.M., Dever, J., Sarsent, T.D., Thomas, K., Sell, S. and Bonner, J. (1979) Biochemistry 18, 2167-2178. T.D.S. was supported during this period by a Predoctoral Fellowship awarded by the National Science Foundation.

REFERENCES

Bailey, J.M. and Davidson, N. (1976) Anal. Biochem. 70, 75-85.

Blin, N. and Stafford, D.W. (1976) Nucleic Acids Res. 3, 2303-2308.

Brack, C., Hirama, M., Lenhard-Schuller, R. and Tonegawa, S.(1978) Cell 15, 1-14.

Britten, R.J., Graham, D.E. and Neufeld, B.R. (1974) in Methods in Enzymology 29, 363-418.

Cuatrecasas, P. (1970) J. Biol. Chem. 245, 3059-3065.

Dayhoff, M.O. (1976) Atlas of Protein Sequence and Structure, ed. Dayhoff, M.O. (National Biomedical Research Foundation, Washington, D.C.) vol. 5 suppl. 2, 266-267.

Denhardt, D.T. (1966) Biochem. Biophys. Res. Commun. 23, 641-646.

Galau, G.A., Britten, R.J. and Davidson, E.H. (1977) Proc. Natl. Acad. Sci. USA 74, 1020-1023.

Gillespie, D., Marshall, S. and Gallo, R.C. (1972) Nature New Biol. 236, 227-231.

Hisashi, K., Adams, H.R. and Busch, H. (1966) Cancer Res. 26, 2196-2201.

Innis, M.A. and Miller, D.L. (1980) J. Biol. Chem. 255, 8994-8996.

McConnel, M.W., Simon, M.N. and Studier, F.W. (1977) J. Mol. Biol. 110, 119-146.

Monahan, J.J., Harris, S.E. and O'Malley, B.W. (1976) J. Biol. Chem. 251, 3738-3748.

Morrow, J. (1974) Ph.D. Dissertation, Stanford University.

Pearson, W.R., Davidson, E.H. and Britten, R.J. (1977) Nucleic Acids Res. 4, 1727-1735.

Pearson, W.R., Wu, J.-R. and Bonner, J. (1978) Biochemistry 17, 51-59.

Peters, E.H., Nishi, S., Miura, K., Lorscheider, F.L., Dixon, G.H.and Tamaoki, T. (1979) Cancer Res. 39, 3702-3706.

Ruoslahti, E. and Ensvall, E. (1976) Proc. Natl. Acad. Sci. USA 73, 4641-4644.

Ruoslahti, E. and Terry, W.D. (1976) Nature (London) 260, 804-805.

Sala-Trepat, J.M., Sewage, M. and Bonner, J. (1978) Biochim. Biophys. Acta 519, 173-193.

Schimke, R.T., Palacios, R., Sullivan, O., Kiely, M.L., Gonzalez, C. and Taylor, J.M. (1974) in Methods in Enzymol. 30,631-648.

Schreiber, G., Urban, J., Zahringer, J., Reutles, W. and Frosch, W.(1971) J. Biol. Chem. 246, 4531-4538.

Sell, S., Becker, F., Leffert, H. and Watabe, H. (1976) Cancer Res. 36, 4239-4249.

Sell, S. and Becker, F.F. (1978) J. Natl. Cancer Inst. 60, 19-26.

Sell, S., Jalowayski, I., Bellone, C. and Wepsic, H.T. (1972) Cancer Res. 32, 1181-1189.

Sell, S., Wepsic, H.T., Nickel, R. and Nichols, M. (1974) J. Natl. Cancer Inst. 52, 133-137.

Smith, M.J., Britten, R.J. and Davidson, E.H. (1975) Proc. Natl. Acad. Sci. USA 72, 4805-4809.

Southern, E. (1975) J. Mol. Biol. 98, 503-517.

Strauss, A.W., Bennett, C.D., Donohue, A.M., Rodkey, J.A. and Alberts, A.W. (1977) J. Biol. Chem. 252, 6846-6855.

Taylor, J.M. and Schimke, R.T. (1973) J. Biol. Chem. 248, 7664-7668.

Tilshman, S.M., Kioussis, D., Gorin, M.B., Ruiz, J. and Ingram, R.S. (1979) J. Biol. Chem. 254, 7393-7399.

Tse, T.P.H., Morris, H.P. and Taylor, J.M. (1978) Biochemistry 17, 2121-2128.

CHAPTER TWO

The Rat Serum Albumin Gene: Analysis of Cloned Sequences

SUMMARY

The rat serum albumin gene has been isolated from a recombinant library containing the entire rat genome cloned in the Lambda phage Charon 4A. R-loop and restriction analysis has revealed that this gene is split into at least 14 fragments (exons) by thirteen intervening sequences (introns), and that it occupies a minimum of 14.5 KB of genomic DNA.

INTRODUCTION

Recent advances in recombinant DNA technology have made it possible to obtain virtually any desired genomic sequence in cloned form, provided an appropriate hybridization probe is available. We have used these techniques to isolate the rat serum albumin gene.

Serum albumin synthesis and secretion is one of the major phenotypic characteristics of vertebrate liver. Observation of the activity and state of this gene during development and in adult tissues should be informative as to the process of terminal differentiation. Albumin synthesis is essentially constitutive in adult liver cells, but does respond significantly to a variety of stimuli (Peters, 1975). It is also expressed to variable extents in different hepatoma cell lines (Tse, Morris and Taylor, 1978). The availability of cloned albumin genomic DNA will greatly facilitate the study of this variable expression, particularly at the level of transcript processing.

Determination of the sequence organization of the albumin gene is also of interest, especially with regard to the disposition of repetitive elements and intervening sequences. Although regulatory (Britten and Davidson, 1969) and evolutionary (Gilbert, 1978) significance has been postulated, the functional role, if any, of these features

of eukaryotic genomes remains unknown. Bovine and human serum albumins exhibit a threefold pattern of internal homology, leading to the conclusion that the albumin gene evolved by intragenic duplication of a smaller (200 amino acid) protein (Brown, 1976). If this hypothesis is correct, it should be reflected in the structure of the gene.

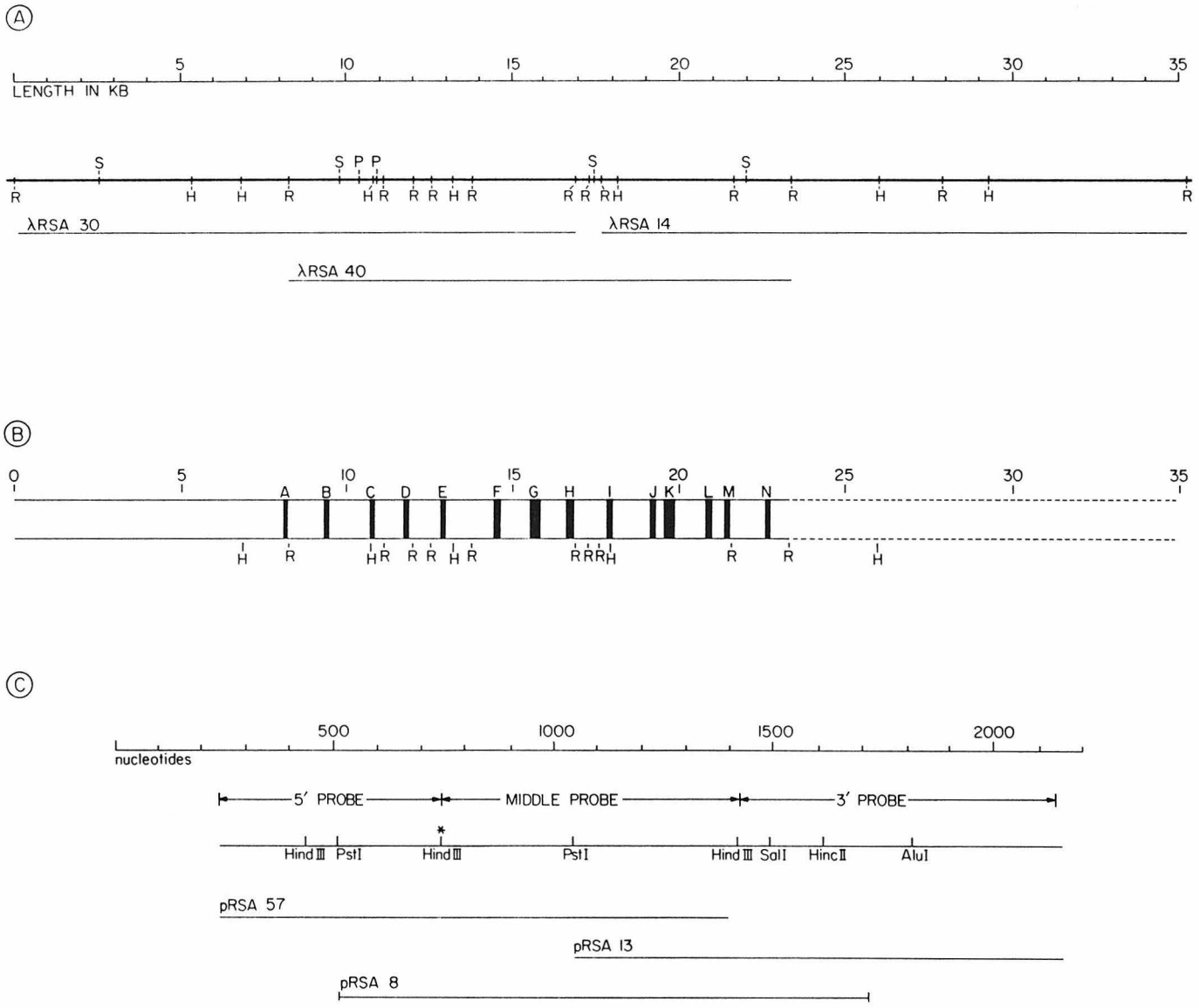
The comparative studies that will be possible as other genes are extracted from the rat and other organisms can be expected to provide considerable insight into the regulation and evolution of the eukaryotic genome.

RESULTS

The recombinant DNA methodology that makes it possible to extract specific sequences from a large genome involves considerable manipulation of DNA, including ligation of a mixture of restriction fragments and several rounds of replication in the bacterial host. This creates the potential for two particularly serious artifactual modifications of genomic sequences; (1), ligation of noncontiguous restriction fragments and (2), genetic rearrangement during propagation. The partial restriction library approach used in this study provides an effective mechanism for detecting the former artifact. Independently generated clones can be used to confirm the legitimacy of the restriction map of DNA common to two or more clones, since the probability of any given spurious ligation event occurring more than once in the production of a recombinant genome library is negligible. Figure 1A shows the map of restriction sites for Eco R1, Hind III and Sac I in the cloned rat serum albumin (RSA) gene. The region included in clone λRSA40 is confirmed by the maps of λRSA14 and λRSA30.

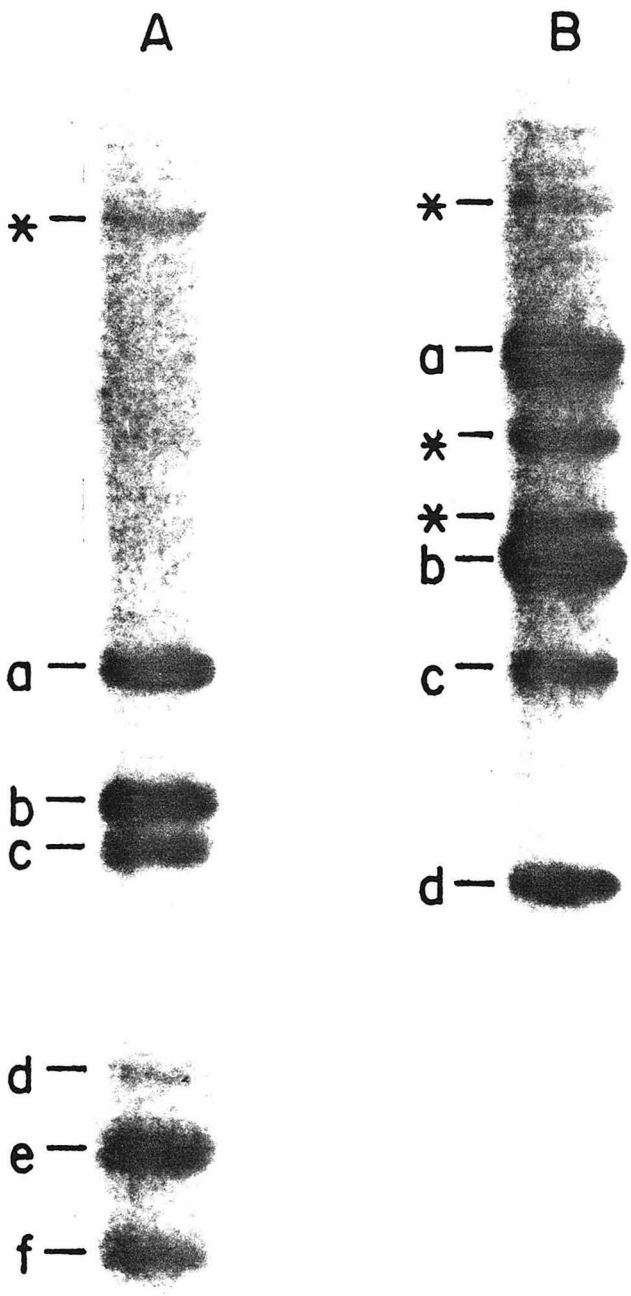
Verification of non-overlapping regions and exclusion of gross genetic artifacts such as deletions or rearrangements depend upon comparisons made between the cloned sequences and the rat genome, which is accomplished

Legend to Fisure 1. (A) Partial restriction site map for three albumin genome clones, λRSA30, λRSA40 and λRSA14. R=Eco RI, H=Hind III, S=Sac I, P=Pst I. Molecular weights were estimated by comparison to phase PM2 DNA digested with Hind III. (B) Intron/Exon map of the albumin gene deduced from R-loops. Black bars=exons, white bars=introns. The dashed line indicates region not tested for R-loop formation. Scale same as in (A). (C) Restriction site map for three albumin cDNA clones, pRSA57, pRSA8 and pRSA13. Asterisk indicates Hind III site not found in genomic DNA (see text).


A
5
10
15
20
25
30
35
LENGTH IN KB
S
S P P
S
S
R
H
H
R
H R
R R H R
R R R H
R
R
H
R
H
R
λRSA 30
λRSA 14
λRSA 40
B
0
5
10
15
20
25
30
35
A B C D E F G H I J K L M N
H
R
H R
R R H R
R R R H
R
R
H
C
500
1000
1500
2000
nucleotides
5′ PROBE
MIDDLE PROBE
3′ PROBE
*
Hind III
PstI
Hind III
PstI
Hind III
SalI
HincII
AluI
pRSA 57
pRSA 13
pRSA 8

by use of "Southern blots" of genomic and cloned rat DNA. Figure 2A shows the pattern obtained when rat DNA is digested with Eco R1 (2A) or Hind III (2B), fractionated by electrophoresis, transferred to nitrocellulose and driven by cloned albumin cDNA labeled with [32-P] by nick translation (Maniatis, Jeffrey and Kleid, 1975). The Eco R1 digestion results in seven bands complementary to the albumin probe. Band "d" is quite faint but is clearly visible in the original autoradiogram and in blots probed with albumin cDNA (Chapter 1). A mixture of two different plasmid clones, pRSA13 and pRSA57, which includes approximately 85% of the albumin mRNA sequence complexity (Chapter 3), was used as probe in these experiments. The restriction site map for these cDNA clones is shown in Figure 1C. Since there are no Eco R1 sites present in the probe sequence, the presence of several bands in the genome blot suggests that either the gene exists in multiple divergent copies or is interrupted by sequences not present in the mRNA, or conceivably both. Evidence will be presented that shows that the albumin gene is in fact interrupted. Figure 2B shows the result of a similar experiment with rat DNA digested with Hind III. A total of seven bands can be visualized. Four of these, labeled "a", "b", "c", "d" are consistent with the restriction map obtained from the genome clones, as are all but the largest band in Figure 2A. The remaining bands,

Legend to Figure 2. Rat genome blots. (A) Eco RI-digested and (B) Hind III-digested rat liver DNA analyzed as described in text. The molecular weights (in KB) of the lettered bands are as follows: Lane A; a=3.9, b=3.1, c=2.7, d=1.6, e=1.3, f=1.0 KB. Lane B; a=7.9, b=4.9, c=3.9, d=2.5


A
B
*
a
b
c
d
e
f
*
a
*
*
b
c
d

indicated by asterisks, are unexpected. These anomalous bands appear with variable intensity in different experiments and are probably due to partial digestion. However, their appearance also is consistent with the hypothesis that there are multiple, slightly divergent albumin genes in the rat. The patterns obtained with cloned RSA cDNA probes are much less complex than those shown in Chapter 1 which resulted from hybridization of [32-P]-cDNA to similar blots. It is likely that most of the fainter bands visualized by cDNA hybridization (Chapter 1, Figure 6) are due to contaminants in the mRNA preparation.

Figure 3 illustrates analogous experiments performed with DNA from the various genomic clones. In these sets of blots the albumin cDNA probe has been cleaved with Hind III and electrophoretically fractionated into "5′", "middle", and "3′" probes (Figure 1B). The patterns generated are entirely consistent with that seen with whole rat liver DNA. The molecular weights are given in the legends to Figures 2 and 3. The similarity of the clone and genomic DNA blot patterns suggests that there has been little or no disruption of the RSA gene during the cloning procedure, although the resolution limit of the genome blots is probably about 100 nucleotides, and small modifications or point mutations would most likely be undetected. A curious aspect of these results is revealed in Figure 3F and 3G.

Legend to Figure 3. Albumin genome clone blots. Panel A; Photograph of gel of Eco RI-digested λRSA40 DNA. Panels B, C and D; triplicate blots of DNA shown in Panel A probed with 5′, middle and 3′ fragments of albumin cDNA clones, respectively (see Figure 1C). Panel E; Photograph of gel of Hind III-digested λRSA30 (left lane), λRSA40 (middle lane) and RSA14 (right lane). Panels F, G and H; triplicate blots of DNA shown in panel E hybridized with 5′(F), middle (G) and 3′ (H) probes as above. The molecular weights of the lettered bands are the same as that given for the corresponding bands in Figure 2. The order of the fragments in the map (Figure 1A) is as follows: Eco RI; c,f,e,b,a,d. Hind III; c,d,b,a.

A
B
C
D
E
F
G
H
a
b
c
d
e
f
a
b
c
d
c+λ
d
b+λ
b
d
a+λ
a

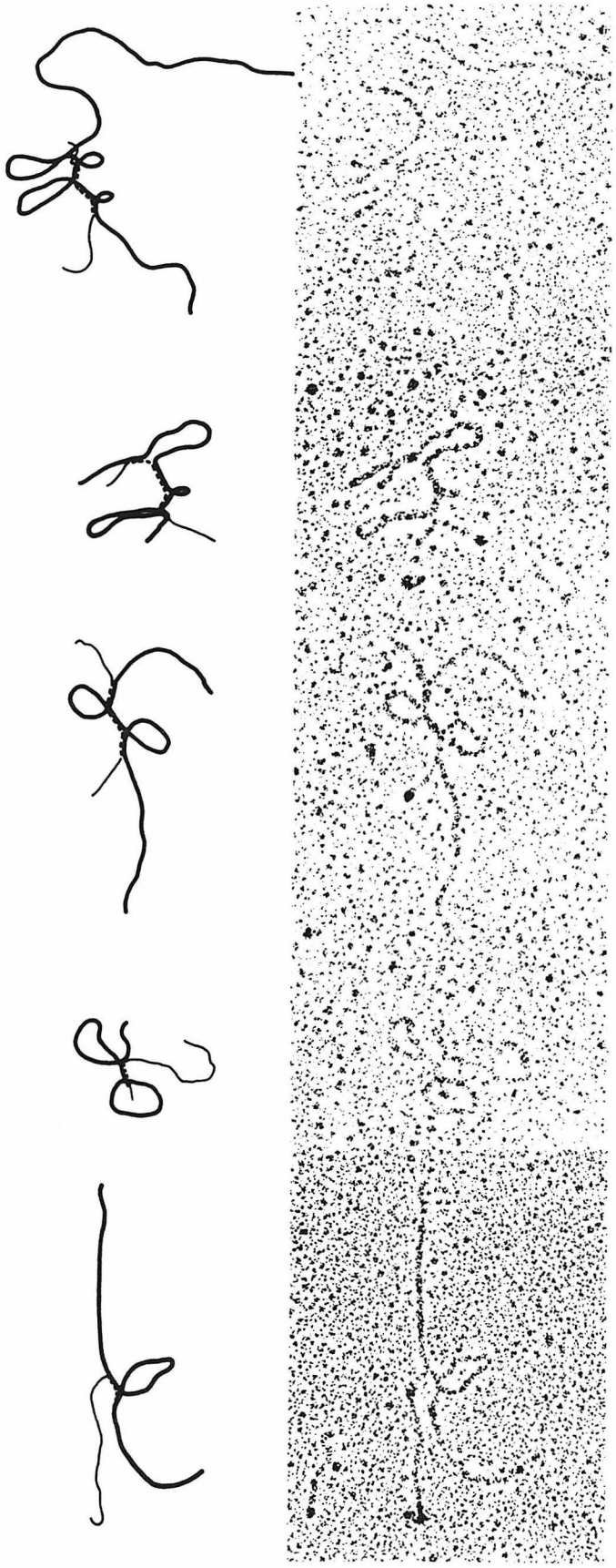
Both the 5′ and middle albumin mRNA probes react with the same Hind III fragment (band "d") which implies that the Hind III site separating these two probes (indicated by an asterisk in Figure 1B) does not exist in the genome. The trivial explanation of probe cross-contamination is unlikely since there is no analogous reaction when middle and 3′ probes are compared. Sequence rearrangement is not responsible since the phenomenon is observed with two different genome clones, λRSA30 and λRSA40, and two different cDNA clones, pRSA13 and pRSA8 (data not shown). Furthermore, this unexpected pattern is seen when 5′ and middle albumin probes are used to drive Hind III genome blots (unpublished data). It is conceivable that this restriction site is created in the mRNA by the fusion of two adjacent exons. As demonstrated in Chapter 4, the latter explanation is correct. Splicing of exons D and E generates a Hind III site in the mRNA sequence that is not present in the genomic DNA.

The pattern of hybridizations with the various probes shown in Figure 3 also establishes the colinearity of the genomic and mRNA sequences. The polarity of the cDNA clones was established by nucleotide sequence determination (Chapter 3) and thus the direction of transcription, from left to right in Figure 1, can be inferred. These blot hybridization data, in conjunction with the restriction maps

suggest that within the cloned locus there exists only one albumin mRNA sequence which is interrupted at least five times and is dispersed over a minimum of 14 KB.

To verify these observations and improve the resolution beyond the limits of restriction analysis, DNA from the genomic clones was digested with either Eco R1 or Hind III, mixed with purified albumin mRNA and incubated under conditions suitable for "R-loop" formation (White and Hogness, 1977). The hybrid structures that formed were then visualized by electron microscopy. The result of these experiments was the identification of a total of thirteen interruptions in the albumin mRNA sequence. Selected electron micrographs along with interpretive drawings of the R-loops are shown in Figure 4. The various hybrids that formed could be identified by their contour lengths, determined by comparison to SV40 DNA spread on the same grids. This permits ordering of R-loop structures by consulting the genome clone restriction maps but does not specify the correct orientation of the fragments. In most cases the polarity can be deduced from the Hind III fragment R-loops by identification of the molecules containing the left or partial right arms of the Charon 4A vector. In one case, shown in Figure 4 (second panel from left), it was necessary to assume that the end lacking a visible DNA "branch" corresponds to the Hind III site located in the

Legend to Figure 4. Selected electron micrographs along with interpretive drawings of R-loops. The restriction fragments involved in the R-loops are, from left to right, Hind III fragments c, d and b, Eco RI fragment a and Hind III fragment at (from λRSA40). In the tracings at the top of the figure, heavy solid lines indicate DNA and the thin lines represent albumin mRNA. The bar at bottom right represents 500 BP.



mRNA sequence, that is, within exon C. This assumption has since been found to be correct (Chapter 4). The results obtained from Eco R1-digested clones are more difficult to interpret due to the larger number of fragments and their similar sizes. However, the structures visualized are consistent with the Hind III fragment R-loops. Certain hybrids do not form efficiently, apparently due to close proximity to a restriction cleavage, and these tend to appear with one but not the other digestion. Also, the temperature optima for hybrid formation with the various fragments differed over a few degrees and had to be adjusted accordingly. The data from which the exon/intron map was deduced are summarized in Table 1. These values are in some cases the sum of two measurements which terminate on opposite sides of the same restriction site. All of the structures represented by the schematic diagram shown in Figure 1C are supported by a minimum of 6 measurements made from different unambiguous examples. However, it is important to bear in mind the uncertainty inherent in data of this type. Aside from possible interpretive errors, small exons and introns might not be detectable by electron microscopy. Furthermore, very small hybrids might not be stable under the conditions used for hybridization or spreading the DNA. DNA displacement loops alongside exons were not visible in most cases. Presumably this is due to

Legend to Table 1.

Mean sizes in nucleotide pairs of exons and introns in the RSA genomic clones. Introns are designated by two letters, indicating the adjacent exons. Lengths were determined by comparison with simian virus 40 DNA contour lengths, as measured from the electron micrographs. These data were used to construct the schematic diagram shown in Figure 1B. n; number of samples measured for each exon or intron. SD; standard deviation of measurements.

Table 1. Intron and exon lengths from R-loops formed with RSA mRNA and RSA genomic clone restriction fragments.

Exon	Length(SD)	n	Intron	Length(SD)	n
A	102 (29)	12	AB	927 (64)	6
B	111 (31)	11	BC	1370 (37)	21
C	95 (35)	10	CD	946 (145)	10
D	148 (34)	10	DE	978 (42)	10
E	108 (32)	12	EF	1458 (88)	19
F	163 (34)	13	FG	920 (65)	13
G	245 (39)	15	GH	778 (66)	14
H	189 (53)	13	HI	908 (62)	10
I	133 (26)	10	IJ	1161 (90)	11
J	131 (27)	16	JK	298 (88)	11
K	239 (31)	14	KL	1011 (74)	16
L	146 (28)	13	LM	434 (94)	13
M	108 (29)	14	MN	1054 (163)	13
N	125 (58)	10			

the collapse of single-stranded DNA, which may be a property of smaller R-loops prepared by the methods we have utilized (Dusaiczyk et al., 1978)

Close examination of the RNA involved in the R-loop structures reveals a continuous translocation of duplex regions from the 5′ end to the 3′ end of the mRNA, supporting the conclusion drawn from the blot experiments that there is only one albumin gene in the cloned complex and that it is colinear with the mRNA sequence, although interrupted, of course. Another interesting observation is that exons A and N seem very near the termini of the mRNA. Very little, if any RNA extends beyond the 5′ side of exon A. This implies that unless a tiny "leader" exon is separated from the rest of the gene by a huge intron or there is more than 8 KB of RNA cleaved from the 5′ end of a primary transcript, the initiation of transcription should be located somewhere on the 5′ end of clone λRSA30 (See Chapter 4, Figure 2). Similarly, the small "whisker" of RNA visible on the 3′ side of exon N is probably mostly poly(A) (Chapter 1), suggesting that the albumin mRNA coding sequence terminates at this point in the genome. Neither of the Eco R1 fragments located 3′ to this terminal exon react with albumin cDNA (unpublished data). Since these fragments represent over 10 KB of genomic DNA it is unlikely that there are any albumin exons not present on these clones.

DISCUSSION

The isolation of the albumin genomic sequences in cloned form makes it possible to determine the structure of the albumin gene - or more properly its DNA component - to the nucleotide level of resolution if desired. However, it is important to recall that the cloned gene is no longer functional rat DNA but nonfunctional bacteriophage DNA. The possibility that rearrangements might occur is not trivial and could be rather misleading. We have argued that the results shown are not artifactual since they were derived from independently generated clones and since the patterns generated by two different restriction enzymes are very similar when either clone or genomic DNA is digested and analyzed by the Southern blot technique.

Based upon the R-loop map and the various blots shown it can be tentatively concluded that all of the albumin coding sequence is located within the 35 KB of contiguous cloned rat DNA and that it is divided into at least fourteen mRNA-encoding sequences or "exons" by thirteen intervening sequences or "introns". The possibility remains that additional exons or introns exist that are too small to be easily detectable by electron microscopy. As shown by extensive DNA sequence analysis (Chapter 4), the conclusions drawn from the R-loop data are essentially correct, except

for the width of exon C (actually 212 BP) and the existence of a small (58 BP) exon between A and B. The possibility of multiple variants of the albumin gene within the individual rat used to generate the library has not been excluded. However, we have isolated a total of nine albumin clones from the library and with the exception of one obvious ligation artifact, all contain linear permutations of the same rat DNA restriction fragments. It seems unlikely that other, different albumin genes would have escaped inclusion in the rat library or detection by plaque hybridization. On the other hand, evidence is presented in Chapter 4 (Figure 4) suggesting that other versions of the albumin gene exist in other individual rats of the Sprague-Dawley strain. The faintness of hybridization band d in the Eco R1 genomic blot experiment is unexplained. Fragment "d" contains 174 base pairs of homology to the cDNA clone (pRSA 13) that was hybridized to the blot. This is almost twice as much as the 98 base pairs (Chapter 4) of homology between the cDNA and R1 fragment "e", which hybridizes much more intensely than does fragment "d".

There seems to be no simple pattern to the size or position of the exons in this gene. The interruptions occur along the entire length of the mRNA sequence. The middle Hind III site in the mRNA has been aligned with amino acids 205/206 by nucleotide sequence determination (Chapter 3,

Figure 2). This would mean that approximately 200-300 nucleotides of untranslated RNA reside at the 3′ end of the albumin mRNA (Fig. 1C). If exon N does in fact contain the 3′ end of the mRNA, then one and possibly two introns may be located within the untranslated sequence (Table 1). Introns located in antibody genes have been found to correspond to junctions between functional domains in the immunoglobulin protein (Sakano et al., 1979), and it has been suggested that this relationship may be a general property of eukaryotic genes and proteins (Gilbert, 1978). Serum albumin can be visualized as consisting of three domains further subdivided into a total of nine subdomains (Brown, 1976). It is not possible to conclude from the present R-loop data that there exists any direct relationship between albumin protein structural domains and the exons in the albumin gene. This is due to the uncertainties associated with the measurement of such small exons by electron microscopy. The elegant relationship of RSA gene exons to the pattern of internal homology in the protein and mRNA is demonstrated in Chapters 3 and 4 by nucleotide sequence analysis.

Assuming that we have not inadvertently cloned an inactive variant of the albumin gene, the arrangement of introns and exons has obvious ramifications regarding the processing of the albumin mRNA precursors. First, the

transcription unit is at least 14.5 KB in length, and by analogy to other systems (Tilghman et al., 1978, Catterall et al., 1979) the primary transcript can be expected to be of the same size or longer. This is three times larger than the value reported by Shafritz and coworkers (Strair et al., 1977) suggesting that the 26S species they detected represents accumulated processing intermediates. Second, there should be a minimum of thirteen different nuclear intermediates which have been processed to some extent. If multiple pathways exist then there could be many more species.

By performing hybridization experiments with nuclear RNA from rat liver and hepatoma cells and DNA from the albumin clones it will be possible to test the gene structure we have adduced and to determine the mechanism and kinetics of albumin mRNA biosynthesis.

MATERIALS AND METHODS

RAT GENOME LIBRARY. High molecular weight liver DNA was extracted from an adult male Sprague-Dawley rat (Simonsen Labs, Inc. Gilroy CA) by the method of Blin and Stafford (1976) and aliquots were digested with Eco R1 (Boehringer-Mannheim) under conditions adjusted to cleave either 1/3 or 1/5 of the Eco R1 sites in an equivalent amount of bacteriophage Lambda DNA. The fragments resulting from this partial digestion were sedimented through a 10-30% sucrose gradient and the material between 10 and 20 KB was recovered by ethanol precipitation. 2.5 micrograms of this rat DNA was ligated with 8.5 micrograms of a preparation of Charon 4A "cloning fragments" (Sternberg, Tiemeier and Enquist, 1977, Blattner et al, 1977). This recombinant DNA was packaged in vitro using extracts prepared from defective Lambda lysogens (NS 428 and NS 433) provided by N. Sternberg (cited above). The method used was that of Hohn and Murray (1977). Approximately 2,000,000 independent clones were obtained. The library was amplified 100,000-fold by subconfluent plating on DP50SupF (Maniatis et al., 1978). The library is stored in 10 mM tris, pH 7.2, 100 mM NaCl, 5 mM $MgCl_2$, 100 ug/mL gelatin, 30% (v/v) glycerol at 5°C. The titer decays under these storage

conditions with a half-life of approximately one year.

cDNA CLONES. cDNA was synthesized from purifed albumin mRNA as described (Sala-Trepat et al., 1979, Chapter 1), except that all four dNTP's were present in the reaction mixture at a concentration of 0.5 mM. The precursor specific activity was approximately 1 Ci/mmol [32-P]-dCTP. This cDNA contained a small amount of full-length material and had a number average molecular weight of approximately 1000 NT. It was rendered double stranded by sequential treatment with E. coli DNA polymerase I and S1 nuclease (Higuchi et al., 1976). The resulting DNA had a number average molecular weight of 600 NTP. An average of 10 dCMP residues were polymerized per 3'-end by terminal transferase (Roychoudhury, Jay and Wu, 1976, W. Rowekamp, personal communication). 40 nanograms of the tailed albumin was mixed with 200 nanograms of pBR322 DNA that had been cleaved with Pst 1 and similarly tailed with dGMP residues (a gift of W. Rowekamp), and induced to co-circularize by incubation at 42°C. for 4 hr followed by 16 hr of slow cooling to 4°C. E. coli strain X-1776 was transformed with this mixture (Curtiss et al., 1977), by a calcium-manganese procedure as described by Villa-Komaroff et al. (1978). Several hundred colonies were obtained on supplemented L-agar plates containing 15 ug/mL tetracycline. Clones that

were resistant to tetracycline and sensitive to ampicillin (20 ug/mL) were subsequently screened by the filter colony hybridization method of Grunstein and Hogness (1975) using [32-P]-labeled albumin cDNA as a probe. The most intensely reacting clones were selected and plasmid DNA was prepared. Proof of their identity was obtained by comparison of partial nucleotide sequences to the existing amino acid sequence of rat serum albumin (Isemura and Ikenaka, 1978, Chapter 3). All manipulations were carried out according to the N.I.H. Guidelines for Recombinant DNA Research.

SCREENING. Approximately 1,000,000 plaques from the rat library were screened by a modification of the method of Benton and Davis (1977), using as a probe nick translated albumin cDNA clones. Nine different genome clones were obtained, three of which, λRSA14, λRSA30 and λRSA40 are the subject of the present report.

R-LOOPS. 50 nanograms of recombinant phage DNA was digested with either Eco RI or Hind III, mixed with 40 nanograms of purified albumin mRNA in a total volume of 20 microliters of 70% recrystallized formamide, 0.4M NaCl, 5mM EDTA, 80 mM Pipes pH 7.4 (White and Hogness, 1977) and incubated at 50-53°C. for 24 hr. The hybrids were spread for electron

microscopy by the modified Kleinschmidt method (Davis, Simon and Davidson, 1971). The grids were rotary shadowed with platinum and palladium (80/20) and viewed in a Phillips 300 electron microscope. DNA contour lengths were measured using a Hewlett-Packard digitizer.

BLOTS. Rat liver DNA from the same preparation used to make the genomic library was digested three times with a sixfold excess of either Eco R1 or Hind III, extracted with phenol and precipitated with ethanol. 10 micrograms of the digested DNA was fractionated on an 8 mm-thick 0.8% agarose slab gel buffered with 50 mM tris, 18 mM NaCl, 2 mM EDTA, 20 mM sodium acetate, pH 7.4. Electrophoresis was at 2.5 volts per cm for 16 hr. Clone DNA was digested once with a tenfold excess of restriction endonuclease and 0.2 micrograms was electrophoresed on a 2mm-thick gel. Transfer to nitrocellulose (Millipore) was essentially as described by Southern (1975). Washing after hybridization was done at 62-63°C. with a descending series of salt concentrations from 1.0 to 0.1 M NaCl in Denhardt's solution (Denhardt, 1966), 0.1% SDS, 0.1% sodium pyrophosphate, 2 mM EDTA, 30 mM tris pH 7.4.

ACKNOWLEDGEMENTS

The preceding Chapter was published essentially as it appears in this dissertation as: Sargent, T.D., Wu, J.-R., Sala-Trepat, J.M., Wallace, R.B., Reyes, A.A. and Bonner, J. (1979) Proc. Natl. Acad. Sci. USA 76, 3256-3260. I am grateful to Walter Rowekamp for his advice and help in cloning the rat serum albumin messenger RNA. T.D.S. was supported during this period by a Predoctoral Fellowship awarded by the National Science Foundation, and by Grant No. 5 T32 GM 07616, awarded by the National Institute of General Medical Sciences, N.I.H.

REFERENCES

Benton, W.D. and Davis, R.W. (1977) Science 196, 180-182.

Blattner, F.R., Williams, B.G., Blechl, A.E., Denniston-Thompson, K., Faber, H.E., Furlong, L.-A., Grunwald, D.J., Kiefer, D.O., Moore, D.D.,Schumm, J.W., Sheldon, E.L. and Smithies, O. (1977) Science, 196, 161-169.

Blin, N. and Stafford, D.W. (1976) Nucleic Acids Res. 3, 2303-2308.

Britten, R.J. and Davidson, E.H. (1969) Science 165, 349-357.

Brown, J.R. (1976) Federation Proc. 35, 2141-2144.

Catterall, J.F., Stein, J.P., Lai, E.C., Woo, S.L.C., Dusaiczyk, A., Mace, M.L., Means, A.R. and O'Malley, B.W. (1979) Nature 278, 323-327.

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. and Alexander, L.(1977) in Proceedings of the 10th Miles International Symposium, eds. Beers, R.F., Jr. and Bassett, E.G. (Raven, New York), pp. 45-56.

Davis, R.W., Simon, M. and Davidson, N. (1971) Methods Enzymol. 21D, 413-428.

Denhardt, D.T. (1966) Biochem. Biophys. Res. Comm. 23, 641-646.

Dusaiczyk, A., Woo, S.L.C., Lai, E.C., Mace, M.L. Jr., McReynolds, L. and O'Malley, B.W. (1978) Nature 274, 328-333.

Gilbert, W. (1978) Nature 271, 501.

Grunstein, M. and Hogness, D.S. (1975) Proc. Natl. Acad. Sci. USA 72, 3961-3965.

Higuchi, R., Paddock, G.V., Wall, R. and Salser, W. (1976) Proc. Natl. Acad. Sci USA 73, 3146-3150.

Hohn, B. and Murray, K. (1977) Proc. Natl. Acad. Sci. USA 74, 3259-3263.

Isemura, S. and Ikenaka, T. (1978) J. Biochem. 83, 35-48.

Maniatis, T., Hardison, R.C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G.K. and Efstradiatis, A. (1978) Cell 15, 687-701.

Maniatis, T., Jeffrey, A. and Kleid, D.G. (1975) Proc. Natl. Acad. Sci. USA 72, 1184-1188.

Peters, T., Jr. (1975) in The Plasma Protein, vol. 1, Putnam, F.W., Ed., 2nd ed., New York, N.Y., Academic Press, pp. 133-181.

Roychoudhury, R., Jay, E. and Wu, R. (1976) Nucleic Acids Res. 3, 101-116.

Sakano, H., Rogers, J.H., Huppi, K., Brack, C., Traunecker, A., Maki, R., Wall, R. and Tonegawa, S. (1979) Nature 277, 627-633.

Sala-Trepat, J.M., Dever, J., Sargent, T.D., Thomas, K., Sell, S. and Bonner, J. (1979) Biochemistry 18, 2167-2178.

Southern, E. (1975) J. Mol. Biol. 98, 503-517.

Sternberg, N., Tiemeier, D. and Enquist, L. (1977) Gene 1, 255-280.

Strair, R.K., Yap, S.H., Nadal-Ginard and Shafritz, D.A. (1977) J. Biol. Chem 253, 1328-1331.

Tilghman, S.M., Curtiss, P.J., Tiemeier, D.C., Leder, P. and Weissman, C. (1978) Proc. Natl. Acad. Sci. USA 75, 1309-1313.

Tse, T.P.H., Morris, H.P., and Taylor, J.M. (1978) Biochemistry 17, 3121-3128.

Villa-Komaroff, L., Efstradiatis, A., Broome, S., Lomedico, P., Tizard, R., Naber, S.P., Chick, W.L. and Gilbert, W. (1978) Proc. Natl. Acad. Sci. USA 75, 3727-3731.

White, R.L. and Hosness, D.S. (1977) Cell 10, 177-192.

CHAPTER THREE

Nucleotide Sequence of Cloned Rat Serum Albumin Messenger RNA

SUMMARY

The recombinant DNA inserts of three bacterial plasmid clones containing nearly all of the rat serum albumin mRNA complexity have been sequenced. Statistical analysis of the nucleotide sequence reveals a pattern of repeated internal homology that divides the mRNA into three approximately equal segments. Each of these homologous segments corresponds to one protein "structural domain", and the extensiveness of the homology strongly supports the "intragenic triplication" model of albumin evolution. In addition, the complete amino acid sequence of rat pre-pro-albumin has been inferred from the nucleotide sequence data.

INTRODUCTION

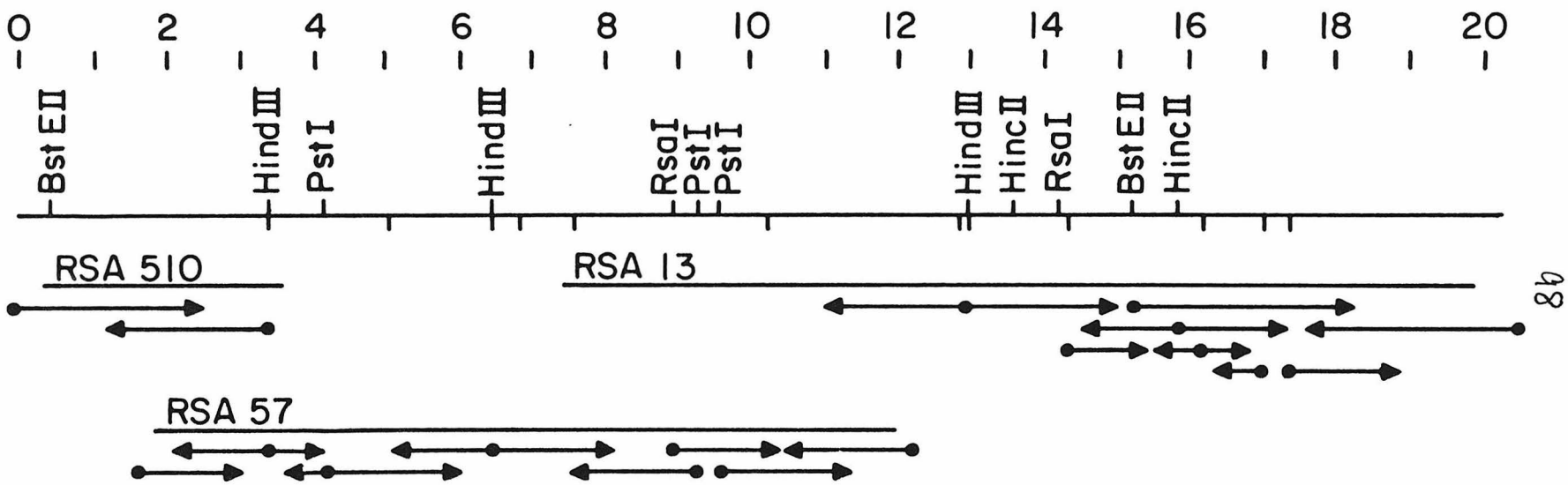
The protein serum albumin has several attributes that make it an attractive subject for experimental investigation. It is the predominant and characteristic synthetic product of adult vertebrate liver and therefore a convenient example of controlled gene expression in terminally differentiated cells. In mammalian embryos, there is a reciprocal relationship between the expression of albumin and its fetal counterpart, alpha-fetoprotein, which is an interesting problem of developmental biology (Chapter 1). Perhaps the most striking property of serum albumin is the remnants in its amino acid sequence of the evolutionary history of this protein. Disulfide crosslinks generate a pattern of loops that is repeated threefold, defining the three structural domains of serum albumin. These domains exhibit significant amino acid homology in addition to the cysteine residues, and it has been suggested by Brown (1976) that albumin evolved by intragenic triplication of a smaller protein corresponding to one domain, which may have in turn evolved from a much smaller sequence by an earlier series of duplications and partial deletions. However, this evolutionary hypothesis is based upon amino acid sequence homology between domains that is not overwhelming, and could conceivably be due to convergent evolution of originally nonhomologous sequences.

To address this question and as a basis for further research, we have cloned the rat serum albumin messenger RNA as a series of recombinant DNA plasmids and have determined the nucleotide sequence of these clones, which include all of the complexity of albumin mRNA from the amino-terminal codon to within approximately 30 nucleotides of the polyadenylation site. A statistical analysis of these data reveals extensive internal homology in the albumin mRNA that essentially verifies the intragenic triplication hypothesis of albumin evolution.

RESULTS

The strategy used in sequencing the RSA mRNA is shown in Figure 1. Most of the sequencing operations were repeated at least twice, and the resulting data are probably free of errors, although mutations associated with the cloning procedure cannot be ruled out. Approximately 35 nucleotides at the extremities of the albumin mRNA failed to appear in any of the cDNA clones. The balance extends 1956 nucleotides from the middle of the ATG/Met, corresponding to the amino terminus of pre-pro-albumin, to 130 nucleotides into the 3′ untranslated region. The sequence data are shown in Figure 2. Assuming code universality, the amino acid sequence of rat pre-pro-albumin is readily inferred, and is listed above the DNA sequence. This amino acid sequence generally concurs with published sequences that have been determined by conventional methods (Strauss et al., 1977, Isemura and Ikenaka, 1978 and T. Ikenaka, personal communication). There are a few discrepancies, however. Amino acid positions 353(Thr), 357(Glu), 402(Gln), 453(Asn), 454(Leu) and 456(Arg) are specified in Ikenaka's unpublished sequence data as Lys, Asp, Ala, Leu, Gly and Glx, respectively. There are several possible explanations for these differences, but they are probably not due to erroneous DNA sequence determinations. When the amino acid sequences of rat, human and bovine serum albumins are

Legend to Figure one. Sequencing strategy and restriction map of albumin cDNA clones. The albumin mRNA "inserts" of the recombinant DNA plasmids pRSA510, pRSA57 and pRSA13 are shown. The nucleotide sequence was determined by labelling restriction endonuclease-digested DNA at the sites indicated by the closed circles and subjecting this "end-labelled" DNA to the chemical sequencing procedure of Maxam and Gilbert (1980). The direction and extent of the sequencing determinations are indicated by the arrows. Most determinations were performed at least twice. The scale is in hundreds of nucleotides. The downward "tics" correspond to Alu I sites.


0
2
4
6
8
10
12
14
16
18
20
BstEII
HindIII
PstI
HindIII
RsaI
PstI
PstI
HindIII
HincII
RsaI
BstEII
HincII
RSA 510
RSA 13
RSA 57

Legend to Figure two. Nucleotide sequence. Except for approximately 35 nucleotides from either end of the mRNA that failed to be cloned, this is the complete sequence of the albumin messenger RNA. The inferred amino acid sequence of rat pre-pro-albumin is also indicated. The "pre" piece is amino acid residues 1-18 and the "pro" piece is residues 19-24.

```
met lys trp val thr phe leu leu leu leu phe ile ser gly ser ala phe ser arg gly val phe arg arg glu ala his lys ser glu ile ala his
ATG AAG TGG GTA ACC TTT CTC CTC CTC CTC TTC ATC TCC GGT TCT GCC TTT TCT AGG GGT GTG TTT CGC CGA GAA GCA CAC AAG AGT GAG ATC GCC CAT      99

arg phe lys asp leu gly glu gln his phe lys gly leu val leu ile ala phe ser gln tyr leu gln lys cys pro tyr glu glu his ile lys leu
CGG TTT AAG GAC TTA GGA GAA CAG CAT TTC AAA GGC CTA GTC CTG ATT GCC TTT TCC CAG TAT CTC CAG AAA TGC CCA TAT GAA GAG CAT ATC AAA TTG     198

val gln glu val thr asp phe ala lys thr cys val ala asp glu asn ala glu asn cys asp lys ser ile his thr leu phe gly asp lys leu cys
GTG CAG GAA GTA ACA GAC TTT GCA AAA ACA TGT GTC GCT GAT GAG AAT GCC GAA AAC TGT GAC AAG TCC ATT CAC ACT CTC TTC GGA GAC AAG TTA TGC     297

ala ile pro lys leu arg asp asn tyr gly glu leu ala asp cys cys ala lys gln glu pro glu arg asn glu cys phe leu gln his lys asp asp
GCC ATT CCA AAG CTT CGT GAC AAC TAC GGT GAA CTG GCT GAC TGC TGT GCA AAA CAA GAG CCC GAA AGA AAC GAG TGT TTC CTG CAG CAC AAG GAT GAC     396

asn pro asn leu pro pro phe gln arg pro glu ala glu ala met cys thr ser phe gln glu asn pro thr ser phe leu gly his tyr leu his glu
AAC CCC AAC CTG CCA CCC TTC CAG AGG CCG GAG GCT GAG GCC ATG TGC ACC TCC TTC CAG GAG AAC CCT ACC AGC TTT CTG GGA CAC TAT TTG CAT GAA     495

val ala arg arg his pro tyr phe tyr ala pro glu leu leu tyr tyr ala glu lys tyr asn glu val leu thr gln cys cys thr glu ser asp lys
GTT GCC AGG AGA CAT CCT TAT TTC TAT GCC CCA GAA CTC CTT TAC TAT GCT GAG AAA TAC AAT GAG GTT CTG ACC CAG TGC TGC ACA GAG TCT GAC AAA     594

ala ala cys leu thr pro lys leu asp ala val lys glu lys ala leu val ala ala val arg gln arg met lys cys ser ser met gln arg phe gly
GCA GCC TGC CTG ACA CCG AAG CTT GAT GCC CTG AAA GAG AAA GCA CTG GTC GCA GCT GTC CGT CAG AGG ATG AAG TGC TCC AGT ATG CAG AGA TTT GGA     693

glu arg ala phe lys ala trp ala val ala arg met ser gln arg phe pro asn ala glu phe ala glu ile thr lys leu ala thr asp val thr lys
GAG AGA GCC TTC AAA GCC TGG GCA GTA GCT CGT ATG AGC CAG AGA TTC CCC AAT GCT GAG TTC GCA GAA ATC ACC AAA TTG GCA ACA GAC GTT ACC AAA     792

ile asn lys glu cys cys his gly asp leu leu glu cys ala asp asp arg ala glu leu ala lys tyr met cys glu asn gln ala thr ile ser ser
ATC AAC AAG GAG TGC TGT CAC GGC GAC CTG TTG GAA TGC GCG GAT GAC AGG GCA GAA CTT GCC AAG TAC ATG TGT GAG AAC CAG GCC ACT ATC TCC AGC     891

lys leu gln ala cys cys asp lys pro val leu gln lys ser gln cys leu ala glu thr glu his asp asn ile pro ala asp leu pro ser ile ala
AAA CTG CAG GCT TGC TGT GAT AAG CCA GTG CTG CAG AAA TCC CAG TGT CTC GCT GAG ACA GAA CAT GAC AAC ATT CCT GCC GAT CTG CCC TCA ATA GCT     990

ala asp phe val glu asp lys glu val cys lys asn tyr ala glu ala lys asp val phe leu gly thr phe leu tyr glu tyr ser arg arg his pro
GCT GAC TTT GTT GAG GAT AAG GAA GTG TGT AAG AAC TAT GCT GAG GCC AAG GAT GTC TTC CTG GGC ACG TTT TTG TAT GAA TAT TCA AGA AGG CAC CCC    1089

asp tyr ser val ser leu leu leu arg leu ala lys lys tyr glu ala thr leu glu lys cys cys ala glu gly asp pro pro ala cys tyr gly thr
GAT TAC TCC GTG TCC CTG CTG CTG AGA CTT GCT AAG AAA TAT GAA GCC ACA CTG GAG AAG TGC TGT GCT GAA GGC GAT CCT CCT GCC TGC TAC GGC ACA    1188

val leu ala glu phe gln pro leu val glu glu pro lys asn leu val lys thr asn cys glu leu tyr glu lys leu gly glu tyr gly phe gln asn
GTG CTT GCA GAA TTT CAG CCT CTT GTA GAA GAA CCT AAG AAC TTG GTC AAA ACT AAC TGT GAG CTT TAC GAG AAG CTT GGA GAG TAT GGA TTC CAA AAC    1287

ala val leu val arg tyr thr gln lys ala pro gln val ser thr pro thr leu val glu ala ala arg asn leu gly arg val gly thr lys cys cys
GCC GTT CTG GTT CGA TAC ACC CAG AAA GCA CCT CAG GTG TCG ACC CCA ACT CTC GTG GAG GCA GCA AGA AAC CTG GGA AGA GTG GGC ACC AAG TGT TGT    1386

thr leu pro glu ala gln arg leu pro cys val glu asp tyr leu ser ala ile leu asn arg leu cys val leu his glu lys thr pro val ser glu
ACC CTT CCT GAA GCT CAG AGA CTG CCC TGT GTG GAA GAC TAT CTG TCT GCC ATC CTG AAC CGT CTG TGT GTG CTG CAT GAG AAG ACC CCA GTG AGC GAG    1485

lys val thr lys cys cys ser gly ser leu val glu arg arg pro cys phe ser ala leu thr val asp glu thr tyr val pro lys glu phe lys ala
AAG GTC ACC AAG TGC TGT AGT GGG TCC TTG GTG GAA AGA CGG CCA TGT TTC TCT GCT CTG ACA GTT GAC GAG ACA TAT GTC CCC AAA GAG TTT AAA GCT    1584

glu thr phe thr phe his ser asp ile cys thr leu pro asp lys glu lys gln ile lys lys gln thr ala leu ala glu leu val lys his lys pro
GAG ACC TTC ACC TTC CAC TCT GAT ATC TGC ACA CTC CCA GAC AAG GAG AAG CAG ATA AAG AAG CAA ACG GCT CTC GCT GAG CTG GTG AAA CAC AAG CCC    1683

lys ala thr glu asp gln leu lys thr val met gly asp phe ala gln phe val asp lys cys cys lys ala ala asp lys asp asn cys phe ala thr
AAG GCC ACA GAA GAT CAG CTG AAG ACG GTG ATG GGT GAC TTC GCA CAA TTC GTG GAC AAG TGT TGC AAG GCT GCC GAC AAG GAT AAC TGC TTC GCC ACT    1782

glu gly pro asn leu val ala arg ser lys glu ala leu ala ter
GAG GGG CCA AAC CTT GTT GCT AGA AGC AAA GAA GCC TTA GCC TAA ACACATCACAACCATCTCAGGCTACCCTGAGAAAAAAAGACATGAAGACTCAGGACTCATCTCTTCTGTTG    1881

GTGTAAAACCAACACCCTAAGGAACACAAATTTCTTTGAACATTTGACTTCTTTTCTC
```

compared (Dayhoff, 1976), 61% of the positions are found to be identical in all three proteins. This homology is rather evenly distributed over the length of the protein. However, it is interesting to note that of the 35 cysteine residues, 34 are exactly conserved in all three albumins. The one difference is due to absence in rat albumin of the residues between the 17th and 18th cysteines. This near-perfect conservation implies that changes in the size of the loops generated by cystine crosslinks are highly deleterious, whereas the primary sequence of the protein can diverge rather freely, at a rate typical of eukaryotic proteins (Wilson, Carlson and White, 1977). There is nothing in the 3′ untranslated portion similar to the "AATAAA" sequence found near the polyadenylation sites of most mRNA's (Benoist et al., 1980)). This is due to a cloning failure rather than a unique feature of albumin mRNA. The DNA sequence of the 3′ end of the rat serum albumin gene has been determined (Chapter 4, Figure 2) and the sequence GCAATTAATAAAAAATGG is found 134 nucleotides downstream from the termination codon, TAA.

The DNA sequence has been subjected to a statistical analysis designed to identify regions of extensive internal homology. The results of this analysis are summarized in Table 1. Depending upon how one defines "extensive homology", a large variety of alignments can be found. When

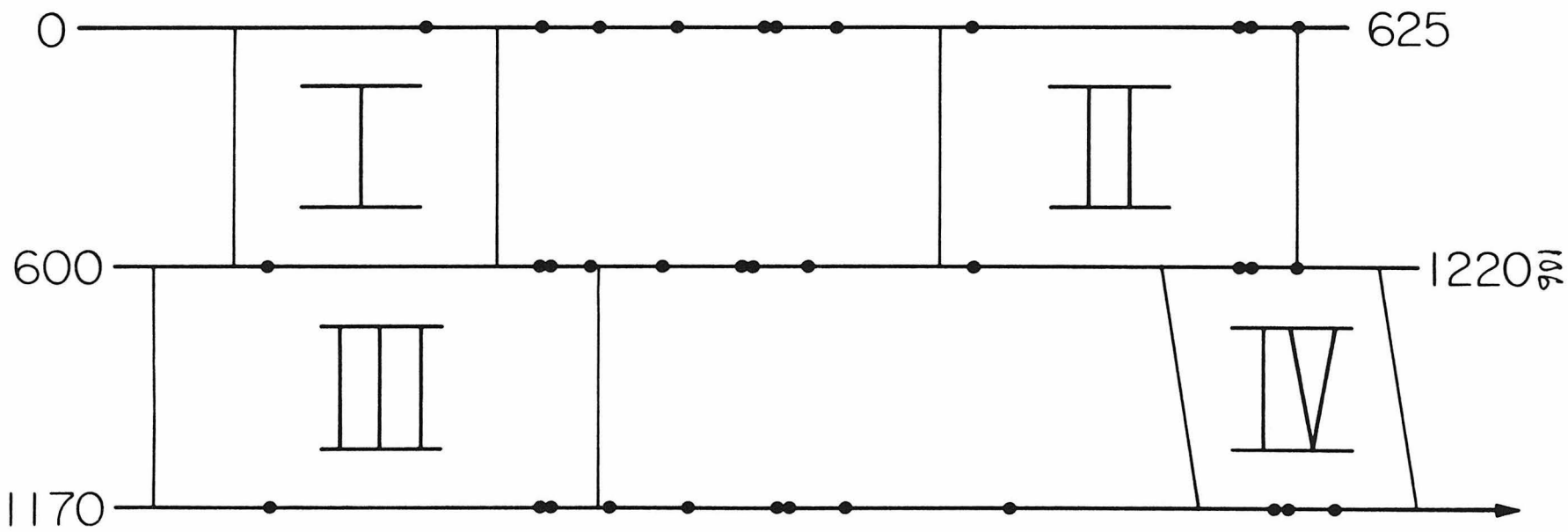
Legend to Table 1. Pa(.25) and Pa(.28) are accident probabilities calculated for random match probabilities, "p", of 0.25 and 0.28, respectively (see Methods). Block V is not aligned in phase with the codon reading frame, so an amino acid comparison is meaningless.

Table 1. Summary of homology

Homology Block	Location	DNA Homology (%)	Pa(.25)	Pa(.28)	Amino Acid Homology (%)
I	80-209 vs. 656-785	57/130 (44)	6.3×10^{-5}	9.5×10^{-4}	10/43 (23)
II	427-603 vs. 1003-1179	97/177 (55)	2.0×10^{-10}	2.4×10^{-8}	24/59 (41)
III	616-835 vs. 1192-1411	84/220 (38)	1.7×10^{-4}	3.8×10^{-3}	11/73 (15)
IV	1707-1815 vs. 1113-1221	49/111 (44)	1.7×10^{-4}	1.8×10^{-3}	7/36 (19)
V	1038-1180 vs. 1402-1544	57/143 (40)	6.4×10^{-4}	6.7×10^{-3}	----

the specification is made that the homologous regions be at least 100 base pairs long and have an Accident Probability, Pa, of 7.0×10^{-4} or less (Materials and Methods), only five regions of homology qualify. Four of these, I-IV are in approximate phase with one another. The fifth homology block is distinct from the others, being offset 212 nucleotides. It is also not aligned in phase with the codon reading frame, as are blocks I-IV, and thus seems unlikely to represent "legitimate" homology, although it is certainly statistically significant ($Pa=6.4 \times 10^{-4}$). Block V has been disregarded in establishing the domain boundaries. Its relationship, if any, to albumin evolution remains to be elucidated. There are no other homologous alignments with Pa values within a factor of 10 of those listed in Table 1. The four pairs of homologous sequence unambiguously define an internal alignment, shown in Figure 3. The protein is thus divided into three blocks that correspond almost exactly to the three "domains" described in bovine and human albumins by Brown (1976). This alignment places most of the cysteine residues in direct phase. Inspection of the positions of these homology blocks and cysteine residues leads to the inference that there have been several small deletions and insertions of 3, 6, 9 or 12 nucleotides at various positions in the albumin mRNA.

Legend to Figure three. The horizontal lines symbolize the mRNA sequence, divided into three overlapping segments, the boundaries of which are indicated at the ends of each line. The vertical lines denote the four "homology blocks" that define the three domains of albumin. The closed circles correspond to cysteine residues.


0
625
600
1220
1170
I
II
II
IV

When the amino acid sequences within each homology block are compared, the results are in accord with the DNA homology, except that the percentage of matches is significantly lower for amino acids than for nucleotides, especially for blocks I, III and IV (Table 1).

DISCUSSION

Did albumin evolve by intragenic triplication? There is no doubt that the internal homology we have found in albumin mRNA is statistically significant. The protein is in fact composed of three homologous "domains". There are only two ways to explain partial sequence homology; 1, initial identity followed by mutational divergence and 2, convergent evolution of two initially distinct sequences. There are several arguments against the latter alternative, the most convincing of which is based on the fact that the internal homology of albumin is much greater at the level of the DNA than at the level of the amino acid sequence (Table 1). This cannot be easily explained by convergent evolution driven by selective pressure on albumin protein structures. Intragenic triplication followed by partial divergence is the only reasonable explanation for the observed structure of rat serum albumin. It is possible that other duplication events preceded and followed this triplication. We have not yet been able to conclusively identify vestiges of intra-domain homology that would indicate an earlier series of intragenic duplications, but there is a high background of relatively weak internal homology that is not in phase with the four main blocks that define the domains. A more sophisticated analysis of the rat serum albumin mRNA sequence might reveal periodicity in this background and

thereby identify a more primitive "proto-albumin" sequence, if it exists. This question is more effectively addressed by analysis of the albumin gene rather than its mRNA. The homology between exons B and D, F and H, and J and L (Chapter 4) suggests that at least one duplication event may have preceded the triplication of domains. There is also convincing evidence that serum albumin and alpha-fetoprotein are related proteins, viz. there is approximately 40-50% homology between the mRNA sequences that encode these two proteins (Innis and Miller, 1980, and unpublished data of our laboratory). This suggests quite clearly that an intergenic duplication of a common ancestor gene occurred subsequent to the domain triplications.

A fundamental problem of biology is to explain the complexity of the eukaryotic genome. Duplication and divergence of genomic DNA may account for much of this diversity. Surveys of the primary sequences of many different proteins reveal a number of clear examples of internal homology (Barker, Ketcham and Dayhoff, 1978), most of which are probably due to intragenic duplications. Furthermore, many, conceivably most, genes are members of families that arose by intergenic duplications (Hood, Campbell and Elgin, 1975, Long and Dawid, 1980). After a structural sequence has been duplicated, presumably one copy can accumulate various mutations at a rather unrestrained

rate. This evolutionary mechanism could convert a simple protein with only one function into a family of complex proteins with many different functions.

METHODS

CLONING PROCEDURES. The production of two of the plasmid clones used in the present experiments (pRSA57 and pRSA13) has been described in Chapter 2. The plasmid clone pRSA510 was produced by "extending" a primer fragment of pRSA57 toward the 5′ end of the albumin mRNA (Figure 1). The Hind III fragment nearest the 5′ end of pRSA57 was labeled by "filling in" the Hind III sites with [32-P]-dCTP plus the other nonradioactive dNTP's and E. coli DNA polymerase I. The appropriate radioactive restriction fragment was isolated by electrophoresis and hybridized to rat liver mRNA, which is approximately 7% albumin-encoding sequences by mass (Chapter 1). The hybridization was carried out in 0.18 M NaCl, 10 mM tris pH 7.0, 0.1 mM EDTA at 60°C. for 30 minutes. Equimolar amounts of primer DNA and albumin mRNA sequences were included in the reaction. The "Cot" (see Chapter 1) was sufficient after 30 minutes for 90% of the primer to renature. The specific heteroduplex thus formed was treated without further purification with AMV reverse transcriptase and then with sodium hydroxide to generate a cDNA that extended from the second Hind III site to within a few nucleotides of the cap of the albumin mRNA. This cDNA had an approximate size of 660 nucleotides, as determined by electrophoresis on an alkaline agarose gel, followed by

exposure to X-ray film. This material was converted to double-stranded cDNA by treatment with E. coli DNA polymerase I (Higuchi et al., 1976, Chapter 2). The "full-length" double stranded cDNA was isolated by agarose gel electrophoresis, purified by chromatography on benoylated DEAE cellulose (see below) and "tailed" with oligo-dG according to published procedures (Roychoudhury, Jay and Wu, 1976). The tailed albumin DNA was mixed with an equimolar amount of vector DNA, which was prepared as follows: Superhelical plasmid pBR322 DNA was digested with the restriction endonuclease Cla I, which cleaves pBR322 once, between the Eco R1 and Hind III sites, within an essential region of the tetracycline resistance gene. Multiple digestions were carried out to assure complete digestion, but care was taken to avoid nicking or otherwise degrading the DNA. The Cla I sites were "filled in" by treatment with E. coli DNA polymerase I and low specific activity [32-P]-dCTP and nonradioactive dGTP. This material was fractionated from residual undigested plasmid DNA by agarose gel electrophoresis (there was no visible superhelical DNA on the gel after staining with ethidium bromide). The linear DNA was extracted from the agarose by a modified crush and soak method (Maxam and Gilbert, 1980) and purified of soluble agarose contaminants by chromatography on benzoylated DEAE cellulose, as described

below. This material was "tailed" with oligo dC. The duplex cDNA and vector DNA were induced to co-circularize as described in Chapter 2. This recombinant DNA was used to transform the E. coli strain MC1061 (Casadaban and Cohen, 1980) according to the method of Kushner (1978). Clones that were sensitive to tetracycline and resistant to ampicillin were further screened by hybridization to a restriction fragment of the rat serum albumin gene containing the "leader" exon (i.e. subclone "JB"; Chapter 4).

SEQUENCING. Labeled DNA for sequence determination was prepared by digesting pRSA510, pRSA57 or pRSA13 with an appropriate restriction endonuclease, indicated in Figure 1 by a closed circle. This DNA was then dephosphorylated with bacterial alkaline phosphatase (New England Biolabs), denatured and labeled with [32-P]-ATP and T4 kinase as described by Maxam and Gilbert (1978), with slight modifications. The labeled DNA was digested with an appropriate second restriction endonuclease, and the desired fragment separated from others by agarose gel electrophoresis. DNA was isolated from the gel by a "crush and soak" method. This DNA was often found to contain soluble gel contaminants that interfered with the sequencing reactions. This problem was alleviated by binding the eluted DNA to a small (0.1 mL) column of benzoylated DEAE

cellulose (Boehringer/Mannheim) in 0.1 M NaCl, 10 mM tris, pH 7.2. The column was then washed with 2-3 mL of 0.3 M NaCl, and finally the pure DNA was eluted with a few hundred uL of 0.6 M NaCl, 20% ethanol, 10 mM tris pH 7.2 and precipitated by addition of 0.7 volumes of isopropyl alcohol and freezing at -80°C. The partial chemical modification and hydrolysis reactions were performed exactly as described by Maxam and Gilbert. The "G>A", "A>C", "C" and "C+T" reactions were used. Electrophoresis was on either 8% or 20% acrylamide gels as described by Sanger and Coulson (1978).

STATISTICAL ANALYSIS. The validity of a given homology between two sequences was evaluated by calculation of an "accident probability",Pa, which is the probability that a homology equal to or greater than that being considered might arise accidentally. The equation is a summation of the Poisson distribution;

$$Pa = \sum_{i=n}^{N} [\exp(-Np)](Np)^{i}/(i!)$$

where N is the length in nucleotides over which the homology is measured, n is the number matches in this interval and p is the probabililty that any given position will be a match, which will be equal to 0.25 if there is no preference for any of the four nucleotides at any given position. Although

in some cases minor deviation from ideal randomness was observed, the result of over one thousand random comparisons of 200 nucleotide fragments of these sequences was an average homology of 51.2 matches per comparison (p=0.256), with a standard deviation of 5.4 (data not shown). Unless specified otherwise (i.e. Table 1), Pa values were calculated assuming p=0.25. The mRNA sequence was divided into several segments, each of which was compared with all the others. A computer program, written by R.F. Murphy and J.W. Posakony, was used to search for stretches of sequence that met or exceeded arbitrary criteria of homology. No allowance was made for gaps, and thus the "homology blocks" tend to have rather sharp boundaries near sites of deletions. Homology extending over a minimum of 100 nucleotides with a maximum Pa of 7×10^{-4} was considered "legitimate" and used to establish the internal alignments of the albumin mRNA sequence.

ACKNOWLEDGEMENTS

The preceding Chapter has been published essentially as it appears in this dissertation as: Sargent, T.D., Yang, M. and Bonner, J. (1981) Proc. Natl. Acad. Sci. USA, In press. I thank Martha Bond for the gift of purified E. coli DNA polymerase I and Bruce Wallace for his advice and help with the transformation procedures. This project was supported by Grant No. 5 T32 GM 07616, awarded by the National Institute of General Medical Sciences, N.I.H.

REFERENCES

Barker, W.C., Ketcham, L.K. and Dayhoff, M.O. (1978) Atlas of Protein Sequence and Structure, ed. Dayhoff, M.O. (National Biomedical Research Foundation, Washington, D.C.) vol. 5, suppl. 3, pp. 359-362.

Benoist, C., O'Hare K., Breathnach, R. and Chambon, P. (1980) Nuc. Acids Res. 8, 127-142.

Brown, J.R. (1976) Federation Proc. 35, 2141-2144.

Casadaban, M.J. and Cohen, S.N. (1980) J. Mol. Biol. 138, 179-207.

Dayhoff, M.O. (1976) Atlas of Protein Sequence and Structure, ed. Dayhoff, M.O. (National Biomedical Research Foundation, Washington, D.C.) vol 5, suppl. 2, pp. 266-277.

Hisuchi, R., Paddock, G.V., Wall, R. and Salser, W. (1976) Proc. Natl. Acad. Sci. USA 73, 3146-3150.

Hood, L., Campbell, J.H. and Elgin, S.C.R. (1975) Annu. Rev. Genetics 9, 305-353.

Innis, M.A. and Miller, D.L. (1980) J. Biol. Chem. 255, 8994-8996.

Isemura, S. and Ikenaka, T. (1978) J. Biochem. 83, 35-48.

Kushner, S.R. (1978) in Proceedings of the International Symposium on Genetic Engineering, eds. Boyer, H.W. and Nicosia, S. (Elsevier/North-Holland Biomedical Press, New York) pp. 17-23.

Long, E.O. and Dawid, I.B. (1980) Annu. Rev. Biochem. 49, 727-764.

Maxam, A.M. and Gilbert, W. (1980) in Methods in Enzymology, eds. Grossman, L. and Moldave, K. (Academic Press, New York) vol. 65: Nucleic Acids, Part I pp. 499-560.

Roychoudhury, R., Jay, E. and Wu, R. (1976) Nuc. Acids Res. 3, 101-116.

Sala-Trepat, J.M., Dever, J., Sargent, T.D., Thomas, K., Sell, S. and Bonner, J. (1979) Biochemistry 18, 2167-2178.

Sanger, F. and Coulson, R. (1978) FEBS Letters 87, 107-110.

Sargent, T.D., Wu, J.R., Sala-Trepat, J.M., Wallace, R.B., Reyes, A.A. and Bonner, J. (1979) Proc. Natl. Acad. Sci. USA 76, 3256-3260.

Strauss, A.W., Bennet, C.D., Donohue, A.M., Rodkey, J.A. and Alberts, A.W. (1977) J. Biol. Chem. 252, 6846-6855.

Wilson, A.C., Carlson, S.S. and White, T.J. (1977) Annu. Rev. Biochem. 46, 573-639.

CHAPTER FOUR

The Fine Structure and Evolution of the Rat Serum Albumin Gene

SUMMARY

The exons, their boundaries and approximately half of the intronic DNA of the rat serum albumin gene have been sequenced. In addition to the 14 exons identified earlier by R-loop analysis (Sargent et al., 1979), a small exon has been detected between the "leader" exon (Z) and exon B. The leader exon encodes the 5'-untranslated portion of albumin mRNA and the "pre-pro" oligopeptide present on the nascent protein. The sites of initiation and termination of transcription have been tentatively identified by comparison of the 5' and 3' gene-flanking sequences to those of other eukaryotic genes. All 28 intron/exon junctions conform to the "GT-AG rule" (Breathnach et al., 1978). The three homologous "domains" of albumin are encoded by three "subgenes" that consist of four exons each and evolved by intragenic duplication of a common ancestor. The second and fourth exons of each subgene appear 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.

INTRODUCTION

The typical eukaryotic 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 (Barker, Ketcham and Dayhoff, 1978). It has also been demonstrated at the nucleic acid level for the mouse immunoglobulin heavy chain constant region (Tucker et al., 1979), the chicken ovomucoid gene (Stein et al., 1980) and rat serum albumin (Sargent, Yang and Bonner, 1981). These proteins have periodic homology that constitutes portions of their amino acid sequence, the basic repeating unit of which

represents the vestige of the 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 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 in order to demonstrate relatedness. Direct analysis of the actual genetic 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 eukaryotic genes, particularly the pattern of interruptions by introns, are often very rigidly conserved, and these can be helpful in the identification of extremely diversent gene families.

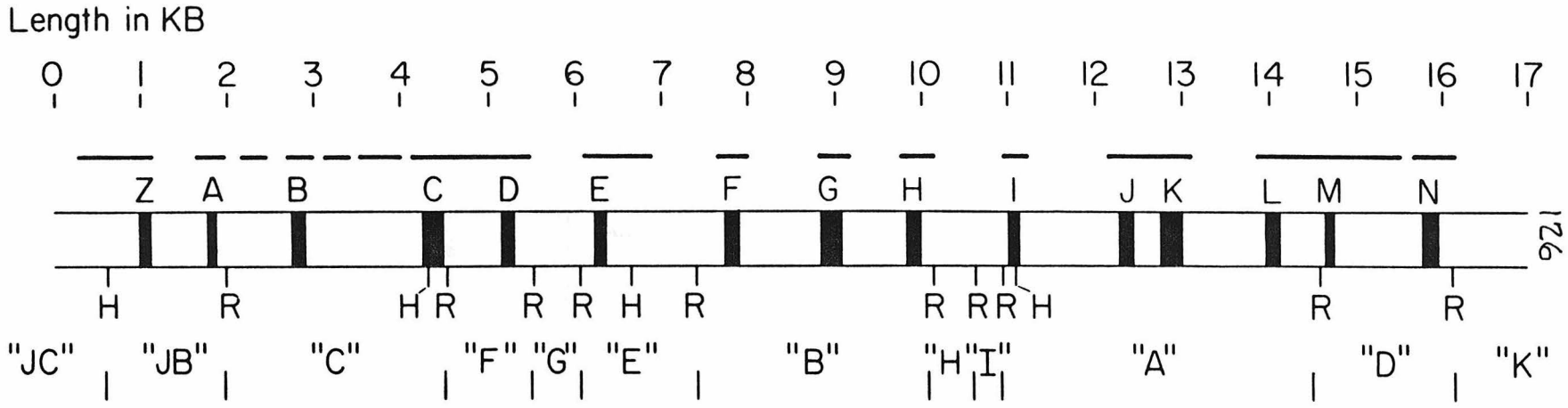
We have used nucleotide sequence analysis of the rat serum albumin gene and its mRNA to elucidate the structural details and evolutionary history of this protein. We have identified a "leader" exon at the 5′ end of the albumin gene that encodes both the "signal" oligopeptide present on

nascent albumin and most or all of the 5'-untranslated portion of the mRNA. The sequences immediately flanking the albumin gene have been compared to the equivalent regions of other genes, and possible sites for the initiation and termination of albumin gene transcription thereby identified. The gene duplication events that are evident from the internal homology of the mRNA and protein have been quite clearly preserved in the intron/exon structure.

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 inferred by comparison of genomic clone sequence data to the nucleotide sequence of the albumin cDNA clones (Chapter 3). Most intron sizes were estimated by electrophoretic mobility on agarose or acrylamide gels and are 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 in Chapter 2, that was based upon 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 of its actual size. This latter discrepancy is difficult to explain, and was unfortunate as it temporarily obscured the threefold

Legend to Figure 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=Hind III, R=Eco R1. The letters in quotation marks are the names of the restriction fragments that were subcloned and sequenced.


Length in KB
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
Z A B C D E F G H I J K L M N
126
H R H R R R H R R R R H H
"JC" "JB" "C" "F" "G" "E" "B" "H" "I" "A" "D" "K"

Legend to Table 1. Exon and intron sizes and locations. (*)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 Figure 2. The sizes of other exons are ambiguous by as many as 6 nucleotides due to terminal redundancy of introns. The values given result from the assumption that splicing conforms to the "GT-AG rule". ($)Intron sizes (except CD, JK and LM) were estimated from the electrophoretic mobility of restriction fragments and are accurate to approximately 5%. Introns CD, JK and LM have been completely sequenced. (**)"1" means that the intron falls between the first and second nucleotides of the codon corresponding to the amino acid given in column 5; "2", between the second and third; "3", between codons. in all cases it was assumed that the GT-AG rule is followed.

Table 1

Exon and intron sizes and locations in protein

exon	size*	intron	size$	location in protein	location in codon**
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 551	2
L	133	LM	556	Glu 595/Gly 596	3
M	62	MN	1048	untranslated	
N	(140)				

substructure of the albumin gene. Of the 15,000 base pairs of this gene, a total of approximately 8000 were sequenced, including all of the exons and their boundaries and over 6000 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 (Sargent et al., 1979), 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 eukaryotic 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 are 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 eukaryotic genes for which data are available (Efstratiadis et al., 1980). On this basis, the most likely capping site of the albumin gene is one of the A residues in the region indicated in Figure 2. This assignment would predict that the distance from the cap to the second Hind III site of the albumin mRNA would be about 650 nucleotides. We estimated

Legend to Figure 2. Terminal sequences of the rat serum albumin gene. Exon Z includes the hydrophobic "signal" peptide sequence (nucleotides 16-36), the methionine initiation codon and the "CCAAT and "TATATTA" sequences associated with the capping regions of other eukaryotic 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 (Benoist et al., 1980). Another typical sequence, similar to TTTTCACTGC, is often found downstream from the 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.

Exon Z

-120

ATTTTGTAATGGGGTAGGAACCAATGAAATGAAAGGTTAGTGTGGTTAATGACCTACAGT

-65 -38

TATTGGTTAGAGAAGTATATTAGAGCGAGTTTCTCTGCACACAGACCACCTTTCCTGTCA

cap site?

-1 +1

ACCCCACTGCCTCTGGCACAATGAAGTGGGTAACCTTTCTCCTCCTCCTCTTCATCTCCG

Met

GTTCTGCCTTTTCTAGGGGTGTGTTTCGCCGAGAAGCACGTAAGCTAGGTA

intron ZA

Exon N

TTTCAAGGCTACCCTGAGAAAAAAAGACATGAAGACTCAGGACTCATCTCTTCTGTTGGT

intron MN

*

GTAAAACCAACACCCTAAGGAACACAAATTTCTTTGAACATTTGACTTCTTTTCTCTGTG

TTTTCACTGC

† * *

CCGCAATTAATAAAAAATGGAAGGAATATACTCTGTGGTTCGGAGGTCTGTCTTCCAACG

poly A site ?

GCGCGTCTCACCCTGGCGGGCTCTAGGGCTGGGGGAAACCCTCGGTTTCCTCCCTTCATC

this distance to be 660+/-20 nucleotides by alkaline electrophoresis of a cDNA preparation "extended" from a restriction fragment primer hybridized to the mRNA (Chapter 3, Materials and Methods section). At the 3' end of the albumin gene, the putative polyadenylation signal sequence, AATAAA, is 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. (1980) have identified another characteristic sequence located near the polyadenylation site; the consensus from several mRNA's is TTTTCACTGC. A similar sequence, TTTTCTCTGT, is 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 positions, this would give a total length of approximately 2030 nucleotides for the non-poly(A) portion of rat serum albumin mRNA, which is consistent with the measurements presented in Chapter 1.

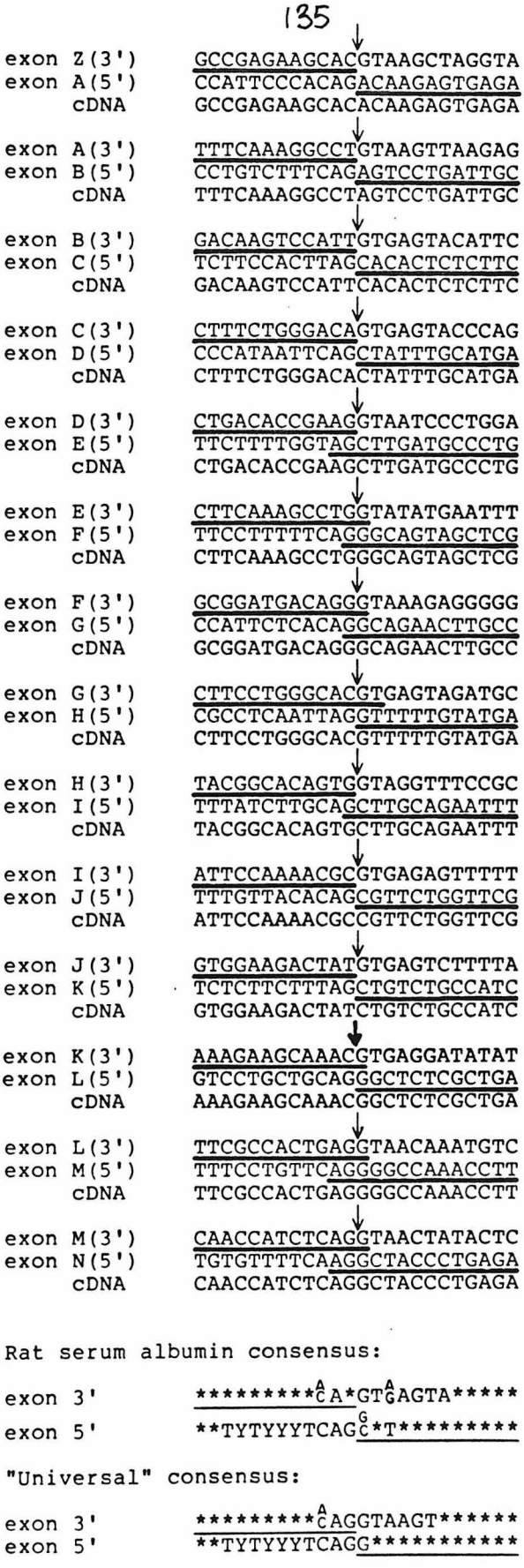
The intron/exon junction sequences of a large number of genes have been determined (Seif, Khoury and Dhar, 1979, Lerner et al., 1980). The consensus sequences for 5' and 3' intron boundaries are (A/C)AG-GTAAGT and TYTYYYTCAG-G, respectively (Y=T or C). It is almost always possible to

define the splice junctions of an intron so that the first two nucleotides are GT and the last two AG (the "GT-AG rule", Breathnach et al., 1978). All of the albumin gene intron/exon junctions are similar to these models. Of the fourteen introns, six have unambiguous splice junctions. The remaining eight have up to four potential splice sites due to a small amount of terminal redundancy in the intronic sequence. All six defined introns conform to the "GT-AG rule", and all of the eight ambiguous introns can be construed to do so (Figure 3).

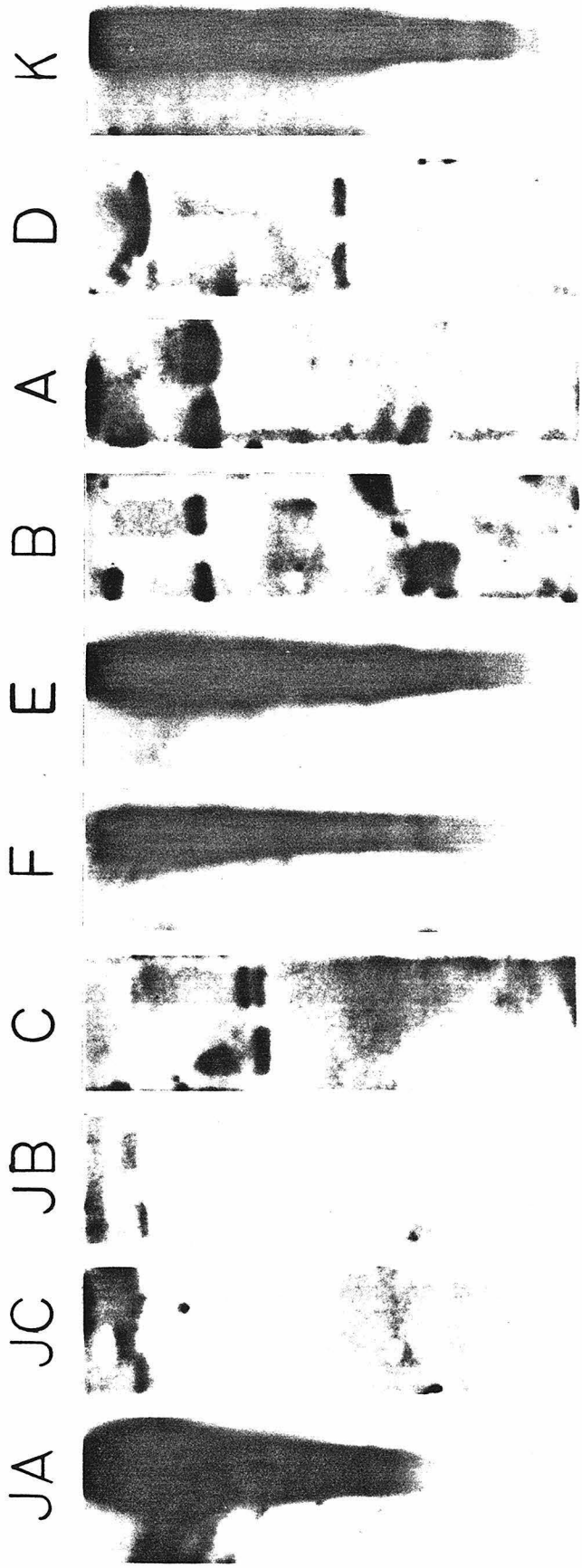
ALBUMIN INTRONS CONTAIN "SIMPLE" SEQUENCES THAT ARE REPEATED ELSEWHERE IN THE GENOME. Figure 4 shows the results obtained when [32-P]-labeled genomic subclones were hybridized to "Southern blots" of Eco R1-digested rat DNA. Each blot consists of two lanes. On the right is 5 ug of Eco R1-digested rat DNA, and on the left is 100 pg of either λRSA30 (Panels JA, JB and JC), λRSA40 (Panels A, B, C, D, E, and F) or λRSA14 (Panel K), digested with Eco R1. These clone blot lanes correspond to the equivalent of one copy per haploid of the 5′, central 15KB or 3′ regions, respectively of the RSA gene (see Chapter 2, Figure 1A). Each panel is named according to the subclone used as the hybridization probe. The locations of the various subclones of the RSA gene are indicated in Figure 1, except for "JA",

Lesend to Fisure 3. Splice Junctions.

The terminal nucleotide sequences of exons and the relevant cDNA sequences are listed. Genomic DNA homolosy 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 splicins confisurations and for other eukaryotic senes (Lerner et al., 1980).


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:
exon 3' *********A/CA*GTA/GAGTA*****
exon 5' **TYTYYYTCAGG/C*T*********
"Universal" consensus:
exon 3' *********ACAGGTAAGT******
exon 5' **TYTYYYTCAGG**********

Legend to Figure 4. Genome blots were performed as described in Materials and Methods. Each panel consists of two lanes run side-by-side on the same gel; on the right is a lane of 5 ug of Eco R1-digested rat liver DNA, and on the left is 100 pg of an Eco-R1-digested genomic clone; RSA30 in Panels JA, JC and JB, RSA40 in Panels C, F, E, B, A and D, and RSA14 in Panel K. 100 pg is equivalent to one copy per haploid genome of the RSA gene when 5 ug of genomic DNA is used. The panels are labeled accordins to the subclone used as the radioactive hybridization probe. The panels are presented in the same order as the relevant restriction fragments appear in the rat serum albumin gene. Eco R1 fragments G, H and I were not used as hybridization probes. Exposure to Kodak XR5 X-ray film was done for either 7 hours without intensification (Panels JA, F, E and K) or for 24 hours with an intensifier screen (all other panels).


K
D
A
B
E
F
C
JB
JC
JA

which is a 5 KB Eco R1-Hind III fragment located to the left (5′) of "JC". Filters were hybridized, washed at 63°C, in 100 mM NaCl and exposed for varying times as described in Materials and Methods.

Subclones JA, E, F, and K generate smears of intense radioactivity on the rat DNA lanes. These blots were exposed overnight without intensifier screens, and the radioactivity in their "clone" lanes does not appear under these conditions. Subclones A and C appear as doublets in the genomic DNA lanes. This was an unexpected finding, as previous experiments indicated only single A and C bands in rat DNA (Chapter 2 and unpublished data). This appears to be due to heterogeneity in the Sprague-Dawley rat population. Different combinations of one or two bands homologous to either subclone A or C have been visualized in the DNA of three different individual Sprague-Dawley rats, even though all specimens were purchased from the same source. This heterogeneity was observed whether genomic or cDNA clones were utilized as hybridization probes. The intensity of these doublets in the A and C genomic lanes is comparable, or slightly less, than that of the "single-copy" genomic clone lanes, and we conclude that there are two slightly different versions of the serum albumin gene in this rat, probably one copy per haploid genome. Subclones JC, JB, D and B are also present at "single-copy" levels in

the rat genome. D and B appear as single bands in genomic DNA, as does JB. Subclones JB and JC are Hind III-Eco R1 and Hind III-Hind III fragments, respectively, of a 9 KB Eco R1 fragment ("J"), and hybridization to such a genomic band at single-copy intensity can be seen at the top of the JB lane. The equivalent region of the JC lane is partially obscured by filter background, but it is clear that neither does subclone JC contain any significantly reiterated sequences. The repeated sequence nearest to the putative 5' terminus of the RSA gene is located approximately 4000 nucleotides away, in fragment JA.

In an attempt to characterize the repeated sequences located within RSA introns (those located in subclones E and F), we determined the complete nucleotide sequence of the Eco R1 restriction fragments E and F (see the Appendix to this dissertation). Three introns were found to contain "simple" sequences that are evidently responsible for at least part of the repeated nature of subclones E and F. Intron CD contains an interrupted palindrome, 200 nucleotides from exon C; (GT)55 followed by 83 nucleotides of "complex" sequence then the complementary polydinucleotide, (AC)85. Both polydinucleotide stretches have a few variant positions. The sequence (CT)19(GT)17 occurs 40 nucleotides to the left of exon E, in intron DE. The poly (AC) common to these two elements appears to be

reiterated in the mouse genome (Nishioka and Leder, 1980), and it is present in an intron in the rat prolactin gene (Gubbins et al., 1980) and in repeated elements flanking the rat serum albumin gene (unpublished data, see the Appendix to this dissertation). Oligo (CT) has been found in the spacer region between sea urchin histone H2A and H1 genes (Sures, Lowry and Kedes, 1978). Intron EF also appears to include some reiterated sequences. The Hind III fragment to the right of exon E hybridizes weakly with reiterated rat DNA (data not shown). The only "simple" sequence found within this intron is a (T)28 tract 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.

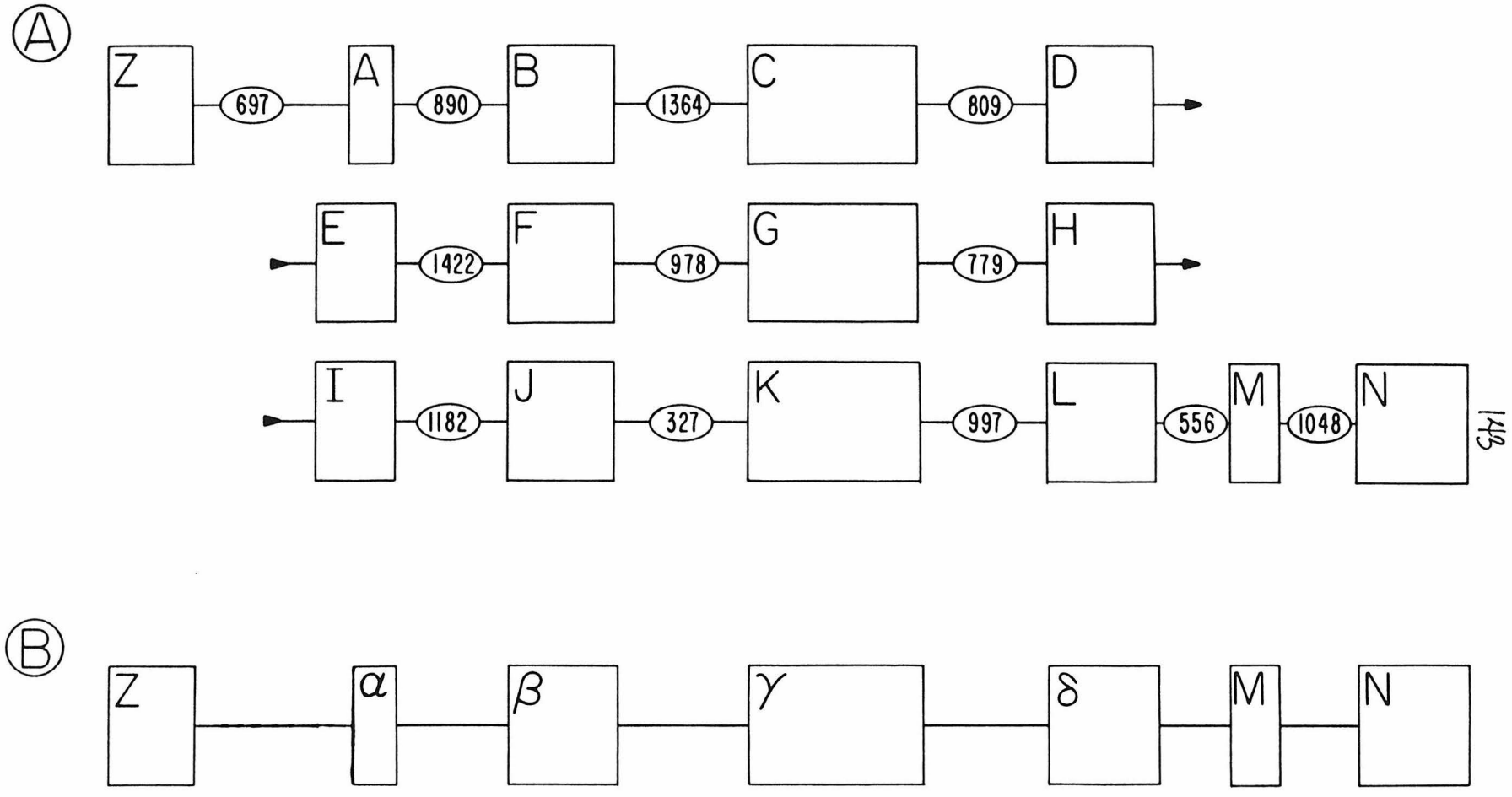
The G+C content of albumin introns (43%) is significantly lower than the exonic level of 50%. The value for the whole rat genome is 42%, calculated from the melting temperature of rat DNA (Wallace et al., 1977, Britten, Graham and Neufeld, 1974). There is a two to three-fold underrepresentation of the dinucleotide CG in albumin introns and exons, as has been observed in other eukaryotic DNA sequences (Catterall et al., 1980).

THE ALBUMIN GENE CONSISTS OF THREE HOMOLOGOUS SEGMENTS. The internal periodicity of the albumin gene has been

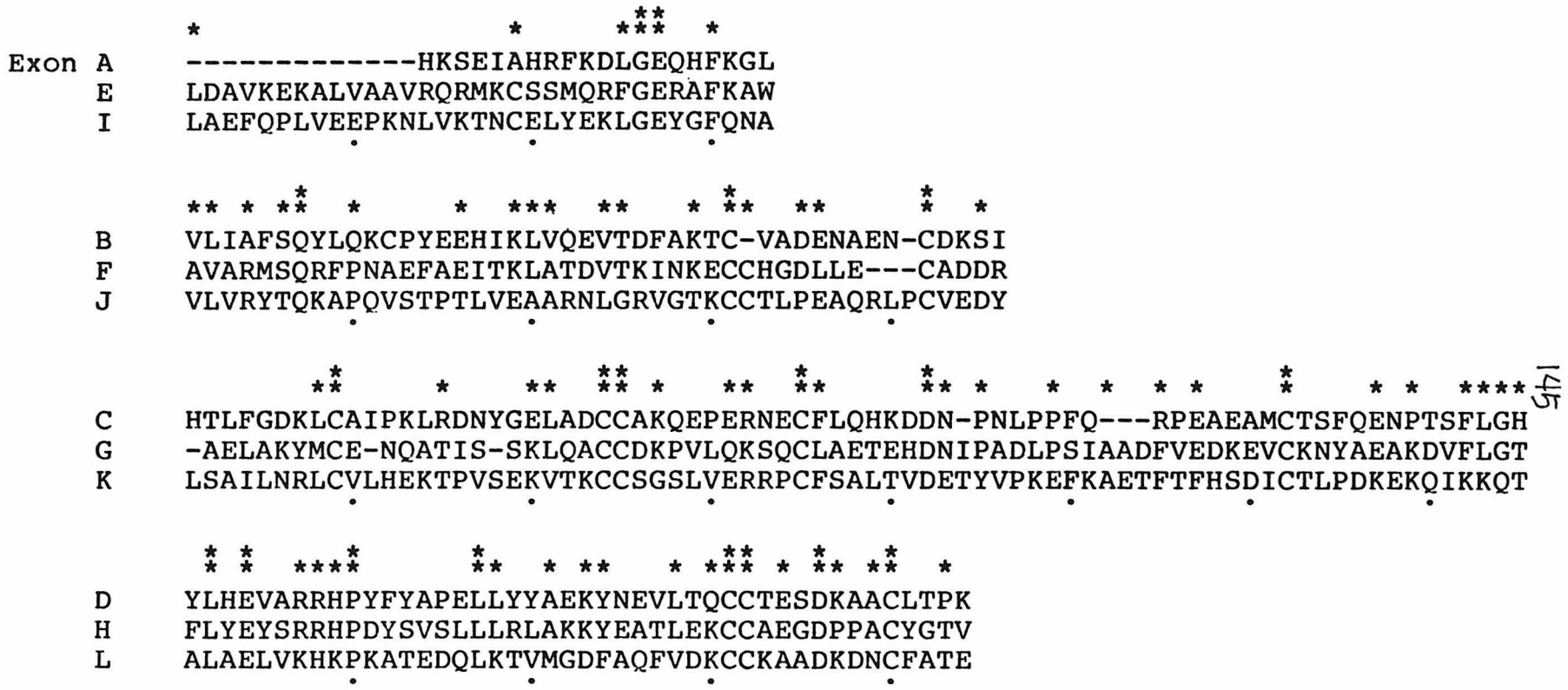
demonstrated by statistical analysis of its mRNA sequence (Sargent, Yang and Bonner, 1981). The divisions in the mRNA sequence proposed on this basis correspond to the boundaries between exons D and E and between exons H and I. Thus the albumin gene is divided into three homologous "subgenes", illustrated in Figure 5a. Each contains four exons, and corresponds to a "structural domain" of the albumin polypeptide (Brown, 1976). In addition there is a leader exon, Z, and a 3' untranslated exon, N, which do 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, Figure 7b). Except for exon A, the corresponding exons of these subgenes have remained remarkably similar in size. There is a total of 19 positions where the same amino acid is present in all three subgenes. Of these, ten are cysteine residues. The remaining nine are highly conserved in rat, human and bovine serum albumins (24 out of 27 possible triple matches). In general, the amino acid sequences encoded by the three rat albumin subgenes have diverged greatly. The average interdomain amino acid homology is only 20%. Figure 6 shows the best alignment of the polypeptides encoded by the various exons. A summary of the internal homology is presented in Table 2.

Legend to Figure 5. (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 et al., 1980). 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.


A
Z
697
A
890
B
1364
C
809
D
E
1422
F
978
G
779
H
I
1182
J
327
K
997
L
556
M
1048
N
1485
B
Z
α
β
γ
δ
M
N

Legend to Figure 6. Internal amino acid homology. Peptides encoded by the fours sets of equivalent exons were aligned for maximum homology by introducing gaps in the shorter sequences. Two out of three matches are denoted by an asterisk, and three out of three 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 2 of the 3 nucleotides. The numerical amino acid and nucleotide homologies are summarized in Table 2.


Exon A
E
I
B
F
J
C
G
K
D
H
L

Legend to Table 2. Summary of internal homology.

The exons were aligned as shown in Figure 6. In each comparison, the total was taken to be the length of the shorter sequence, i.e., gaps were ignored in the tabulation.

Table 2

Summary of internal homology

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

We find no evidence for conservation of intron sequences between albumin subgenes. The variations in the intron/exon junctions show no particular pattern, nor is there any noticeable homology in the intronic DNA that has been sequenced. As shown in Figure 5a, the corresponding introns in each subgene are quite different in size, suggesting that as with globin genes (Nishioka and Leder, 1979, Efstratiadis et al., 1980), albumin introns tend to diverge relatively rapidly, both in sequence and in length. The repeated elements present in introns CD, DE and EF are absent from the corresponding introns in other subgenes.

DISCUSSION

The structure of the rat serum albumin gene is not unlike that of other eukaryotic 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 eukaryotic 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 (Figure 2). The AT-rich "Pribnow box" is evidently involved in the interaction between prokaryotic genes and RNA polymerases (Pribnow, 1975) and the similar sequences found upstream from eukaryotic capping sites may serve an equivalent purpose. At the 3′ end of eukaryotic 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 (Proudfoot and Brownlee, 1976, Efstratiadis et al., 1980). 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 (Sakonju, Bogenhagen and Brown, 1980). On the other hand, similar experiments involving in vitro transcription of cloned conalbumin and ovalbumin genes (Wasylyk et al., 1980) 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 eukaryotic 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 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 (Hawkins and Dugaiczyk, 1980), mouse (Gorin and Tilghman, 1980) and chicken (Gordon et al., 1978) albumin genes are under investigation in other laboratories, so it should be possible to answer this soon.

It is apparent that aside from the repeated elements in the introns, the rat serum albumin gene is "single copy". There seem to be no "middle repetitive" sequences near or within this gene. This stretch of essentially single copy DNA spans approximately 20,000 nucleotides, which is nearly 10 times longer than the average stretch of single copy DNA in the rat genome, as determined by measuring the renaturation rates of rat DNA sheared to various lengths (Pearson, Wu and Bonner, 1978). This may prove to be a peculiar property of the rat serum albumin gene, but general conclusions regarding the disposition of repeated sequences relative to structural genes will have to await the analysis of several more genes, preferably representing different categories of mRNA prevalence (i.e. genes encoding mRNA's present at low levels in the cytoplasm).

THE RELATIONSHIP BETWEEN EXONS AND PROTEIN DOMAINS. Serum albumin consists of three "structural domains" of approximately 190 amino acids each (Brown, 1976). These polypeptide segments are similar in secondary and tertiary structure and exhibit a small 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 non-uniform, as shown in 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 that either 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 (Efstratiadis et al., 1980).

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 (Gilbert, 1978; Crick, 1979). 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, and the active sites for each ligand are usually confined to one structural domain (reviewed by Peters and Reed, 1977), and

in many cases are probably encoded by individual exons. However, there is no reason to presume that these activities existed prior to the assembly of the ancestral albumin gene, since there has been such extensive amino acid sequence diversence 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 (Bradshaw, Shearer and Gurd, 1968), 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, 1975) and rats (Esumi et al., 1980) 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 twelve subgene exons that encode most of the albumin protein, there are three 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 (Strauss et al., 1977) which includes the "signal" sequence

(Blobel and Dobberstein, 1975) present on the amino terminus of nascent albumin, and the 6 amino acid "pro" peptide. Exon Z also encodes the first 2 1/3 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 (Bernard, Hozumi and Tonegawa, 1978), as are the chicken ovomucoid (Stein et al., 1980) and conalbumin (Cochet et al., 1979) genes and the bovine preproopiocortin gene (Nakanishi et al., 1980). 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 (Proudfoot and Brownlee, 1976), so they presumably have some significant, if unknown, role (Setzer et al., 1980). 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 (15/40 matching nucleotides), which may be the result of one of the duplication events (Figure 7).

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.

THE 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 (1976) 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 MY ago, an extrapolation from the amino acid sequence homology between domains and between human and bovine albumins, and the fossil record of mammalian radiation. This presumes that the selective pressures on

albumin over the past 80 MY 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 (Gordon et al., 1978), so it may have a similar three-domain structure. If this is so, the domain duplications probably predate the bird-mammal diversence that occurred about 300 MY ago (Wilson, Carlson and White, 1977).

Because the albumin gene is so complex, there are many plausible models that could explain its evolution. We have applied two principal criteria in order to select one of these alternatives: One, the number of recombination and rearrangement steps should be minimized, and two, 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 the size of exons A versus E and I.

Prior to the two duplication events, the "proto-albumin" gene may have had a structure similar to that shown in Figure 5b; 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 spliced 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 prior to the duplications, but this would have to change 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 7c 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 diversent 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. As shown in Figure 7d, this could be accomplished by recombination within introns δ2M and δε . The result of this (7e) 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 as Table 2 illustrates, this disparity is not overwhelming (27% amino acid homology versus 16% and 18%), and the DNA

Legend to Figure 7. Model for the evolution of the rat serum albumin gene.

(A) Unequal crossover between two copies of a 5-exon gene duplicates the third and fourth exons.

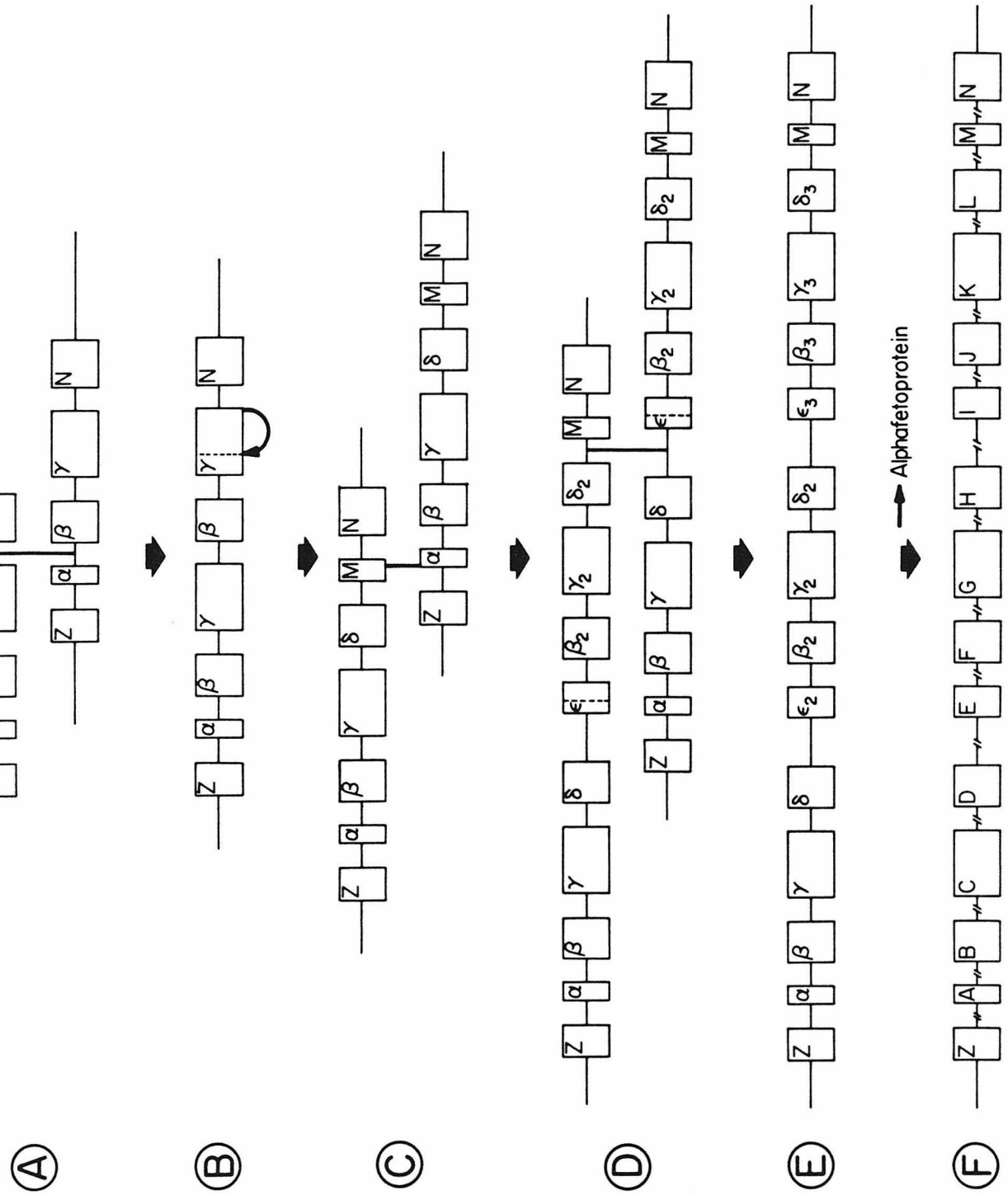
(B) The new second γ exon shrinks to 62 nucleotides by a deletion or by evolution of a new splicing signal, creating the "proto-albumin gene".

(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 alphafetoprotein-albumin gene duplication may have occurred during this period.

(F) The rat serum albumin gene.


A
B
C
D
E
F
Alphafetoprotein

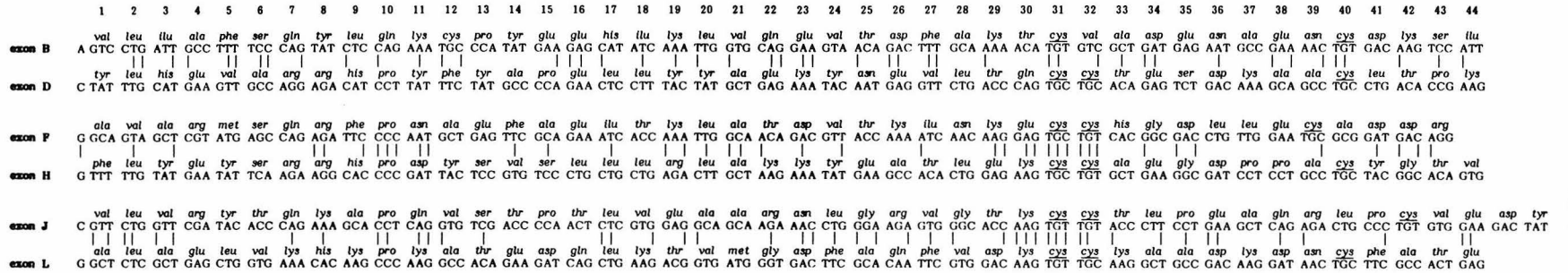
homologies are even more similar (44% versus 36% and 41%). Furthermore, any argument based upon such minor differences in homology is considerably weakened by the observation (Table 2, Efstratiadis et al., 1980) that exons can accumulate mutations at peculiar rates.

The period represented by the space between Fisure 7e and 7f 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 alphafetoprotein. The evidence for this is a very significant level of homology (40-50%) between the rat serum albumin and rat alphafetoprotein mRNA sequences (Sargent, Yang and Bonner, 1981, Innis and Miller, 1980, Jagodzinski, Sargent and Bonner, unp.).

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 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 (Figure 8). Assuming that DNA sequence homology values are Poisson-distributed and that the basic probability of an accidental match is 1/4, the probabilities that these or higher levels of homology are accidents are 0.014, 0.084 and 0.014, respectively. While none of these values are 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 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. Figure 7a 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 δ exon would have to be

Legend to Figure 8. 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 5 sites of extensive homology; amino acid residue numbers 2, 10, 30, 31 and 32.


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44
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 GAA GTA ACA GAC TTT GCA AAA ACA TGT GTC GCT GAT GAG AAT GCC GAA AAC TGT GAC AAG TCC ATT
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 AAA TAC AAT GAG GTT CTG ACC CAG TGC TGC ACA GAG TCT GAC AAA GCA GCC TGC CTG ACA CCG AAG
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 GAC GTT ACC AAA ATC AAC AAG GAG TGC TGT CAC GGC GAC CTG TTG GAA TGC GCG GAT GAC AGG
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 AAA TAT GAA GCC ACA CTG GAG AAG TGC TGT GCT GAA GGC GAT CCT CCT GCC TGC TAC GGC ACA GTG
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 AAC CTG GGA AGA GTG GGC ACC AAG TGT TGT ACC CTT CCT GAA GCT CAG AGA CTG CCC TGT GTG GAA GAC TAT
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 GAC TTC GCA CAA TTC GTG GAC AAG TGT TGC AAG GCT GCC GAC AAG GAT AAC TGC TTC GCC ACT GAG

eliminated prior to the next duplication event. This could occur either via a deletion or the evolution of a new 3' splicing junction (Figure 7b). 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 preceding 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 multi-exon sequences by inspection of the nucleotide

sequence of existing genes. A number of points have emerged that suggest future lines of investigation. In particular, it would be interesting to know if the albumin-alphafetoprotein 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. In view of the complexity of the albumin gene, it would also be interesting to study the processes that convert what should be a 15,000 nucleotide-long primary transcript into the 2030 nucleotide mRNA.

Intragenic duplication has been shown to be the principal evolutionary mechanism for the accretion of exons and introns by the ancestral precursor to the albumin (and alphafetoprotein) 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 (Barker, Ketcham and Dayhoff, 1978), it is apparent that this has been an important source of diversity in the evolution of the eukaryotic genome.

MATERIALS AND METHODS

MATERIALS. Restriction endonucleases were purchased from New England Biolabs, Bethesda Research Laboratories, or Boehringer-Mannheim. 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]-ATP was purchased as a crude, carrier-free aqueous solution from I.C.N., and used within 24 hours of delivery.

CLONES. The rat serum albumin genomic clones were isolated from a library of rat liver genomic DNA as described previously (Sargent et al. 1979). Most of the nucleotide sequence data and restriction endonuclease site map data were derived from "subclones" of these genomic clones. The Eco R1 fragments designated A, B, C, D, E, and F (Figure 1) were isolated from agarose gels and ligated to vector DNA, pBR325 (Bolivar, 1978) cleaved with Eco R1 The "J:I ratio" (Dugaiczyk, Boyer and Goodman, 1975) was approximately 1.0. Eco R1 fragment J was cleaved with Hind III, which generates three fragments. The two Hind III-Eco R1 fragments were ligated with an equimolar amount of the vector, the larger Hind III-Eco R1 fragment of pBR325. E. coli strain HB101 was transformed according to the method of Kushner (1978). All operations were performed in accordance

with NIH guidelines for experiments involving recombinant DNA.

SEQUENCING. Subclone plasmid DNA (or recombinant Lambda phage DNA in the case of exons C and I) was purified of low molecular weight nucleic acid contaminants by exclusion from Sepharose CL2B, cleaved with an appropriate restriction endonuclease, dephosphorylated with bacterial alkaline phosphatase and labelled at the 5′ ends with T4 Kinase and γ-[32-P]-ATP. Following digestion with a second restriction endonuclease, labelled DNA fragments were isolated from agarose or acrylamide gels, purified of soluble contaminants by chromatography on benzoylated DEAE cellulose and sequenced according to minor modifications of the methods of Maxam and Gilbert (1980). The products of the "G>A", "A>C", "C" and "C+T" reactions were electrophoresed on 0.4 mm 8% acrylamide gels (Sanger and Coulson, 1978). Up to 450 nucleotides could be read from a single labelled site. Occasionally, a nucleotide was not visible on the sequencing gel exposure. Whenever a conflict occurred between the exonic and the cDNA sequence data, the complementary strand from the appropriate restriction fragment was sequenced. All discrepancies between the cDNA and exon sequences were eliminated in this manner. The regions of the albumin gene that were sequenced are denoted in Figure 1.

BLOTS AND FILTER HYBRIDIZATION. High molecular weight rat liver DNA was isolated from different adult male Sprague-Dawley rats by the method of Blin and Stafford (1976) with the modification that ethanol precipitation was substituted for dialysis. Aliquots of this DNA were digested three or four times with a four-fold excess of Eco R1 (Boehringer/Mannheim), extracted with phenol and chloroform (1:1 mixture) and precipitated. Electrophoresis on 0.7% agarose gels was performed by standard techniques. Before transferring the electrophoretically fractionated DNA to nitrocellulose filters, the gel was immersed for 100 sec in 0.25 M HCl at room temperature to partially depurinate the DNA. During the 15 minute denaturation in 0.5 N NaOH the DNA molecular weight is reduced to approximately 1000 nucleotides (data not shown), and this appears to improve the transfer of DNA, especially that above 5000 nucleotides in length, from the agarose to the filter. Otherwise, the transfer was performed essentially as described by Southern (1975). Hybridization was carried out as described in Chapter 2, except that 20 mL glass scintillation vials were used instead of plastic bags as hybridization chambers. After hybridization, the blots were washed for 2-3 hrs in high salt buffer (Chapter 2) followed by an overnight wash at 63°C. in 0.5X SET (1X SET is 0.15 M NaCl, 30 mM tris pH 7.2, 2 mM EDTA), 0.1% SDS. Blots that appeared highly

radioactive following this wash procedure (i.e. those hybridized to subclones JA, E, F and K) were exposed for a few hours without intensification. Other lanes were exposed for 12-72 hours with an intensifier screen.

ACKNOWLEDGEMENTS

The experiments described in the preceding Chapter were conducted in collaboration with Dr. Linda Jasodzinski and Ms. Maria Yang. This project was supported by Grant No. 5 T32 GM 07616, awarded by the National Institute of General Medical Sciences, N.I.H.

REFERENCES

Baralle, F.E. and Brownlee, G.G. (1978) Nature 274, 84-87.

Barker, W.C., Ketcham, L.K. and Dayhoff, M.O. (1978) in Atlas of Protein Sequence and Structure, ed. Dayhoff, M.O. (National Biomedical Research Foundation, Washington, D.C.) vol. 5, suppl. 3, pp. 359-362.

Benoist, C., O'Hare, K., Breathnach, R. and Chambon, P. (1980) Nucleic Acids Research 8, 127-142.

Bernard, O., Hozumi, N. and Tonesawa, S. (1978) Cell 15, 1133-1144.

Blobel, G. and Dobberstein, D. (1975) J. Cell Biol. 67, 852-862.

Bolivar, F. (1978) Gene 4, 121-136.

Bradshaw, R.A., Shearer, W.T. and Gurd, F.R.N. (1968) J. Biol. Chem. 243, 3817-3825.

Breathnach, R., Benoist, C., O'Hare, K. and Chambon, P. (1978) Proc. Natl. Acad. Sci. USA 75, 4853-4857.

Britten, R.J., Graham, D.E. and Neufeld, B.R. (1974) in Methods in Enzymology, eds. Grossman, L. and Moldave, K. (Academic Press, New York) vol. 29E, pp. 363-418.

Brown, J.R. (1976) Federation Proceedings 35, 2141-2144.

Catterall, J.F., Stein, J.P., Kristo, P., Means, A.R. and O'Malley, B.W. (1980) J. Cell Biol. 87, 480-487.

Cochet, M., Gannon, F., Hen, R., Maroteaux, L., Perrin, F. and Chambon, P. (1979) Nature 282, 567-574.

Crick, F. (1979) Science 204, 264-271.

Dugaiczyk, A., Boyer, H.W. and Goodman, H.M. (1975) J. Mol. Biol. 96, 171-184.

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., Slightom, J.L., Blechl, A.E., Smithies, O., Baralle, F.E., Shoulders, C.C. and Proudfoot, N.J. (1980) Cell 21, 653-668.

Esumi, H., Okui, M., Sato, S., Sugimura, T. and Nagase, S. (1980) Proc. Natl. Acad. Sci. USA 77, 3215-3219.

Gilbert, W. (1978) Nature 271, 501.

Gitlin, D. and Gitlin, J.D. (1975) in The Plasma Proteins, ed. Putnam, F.W. (Academic Press, New York) vol. II pp. 321-374.

Gordon, J.I., Burns, A.T.H., Christmann, J.L. and Deeley, R.G. (1978) J. Biol. Chem. 253, 8629-8639.

Gorin, M.B. and Tilshman, S.M. (1980) Proc. Natl. Acad. Sci. USA 77, 1351-1355.

Gubbins, E.J., Maurer, R.A., Lasrimini, M., Erwin, C.R. and Donelson, J.E. (1980) J. Biol. Chem. 255, 8655-8662.

Hawkins, J.W. and Dusaiczyk, A. (1980) J. Cell Biol. 87, 111a (abstract).

Innis, M.A. and Miller, D.L. (1980) J. Biol. Chem. 255, 8994-8996.

Kushner, S.R. (1978) in Proceedings of the International Symposium on Genetic Engineering, eds. Boyer, H.W. and Nicosia, S. (Elsevier/ North-Holland Biomedical Press, New York) pp. 17-23.

Lerner, M.R., Boyle, J.A., Mount, S.M., Wolin, S.L. and Steitz, J.A. (1980) Nature 283, 220-224.

Maxam, A.M. and Gilbert, W. (1980) in Methods in Enzymology, eds. Grossman, L. and Moldave, K. (Academic Press, New York) vol. 65: Nucleic Acids, Part I, pp. 499-560.

Nakanishi, S., Teranishi, Y., Noda, M., Notake, M., Watanabe, Y., Kakidani, H., Jingami, H. and Numa, S. (1980) Nature 287, 752-755.

Nishioka, Y. and Leder, P. (1979) Cell 18, 875-882.

Nishioka, Y. and Leder, P. (1980) J. Biol. Chem. 255, 3691-3694.

Pearson, W.R., Wu, J.-R. and Bonner, J. (1978) Biochemistry 17, 51-59.

Peters, T. Jr. (1975) in The Plasma Proteins, ed. Putnam, F.W. (Academic Press, New York) vol. I, pp. 133-181.

Peters, T. Jr. and Reed, R.G. (1977) in Albumin: Structure, Biosynthesis, Function, eds.Peters, T. and SJoholm, I. (Persamon Press, New York) pp. 11-20.

Pribnow, D. (1975) Proc. Natl. Acad. Sci. USA 72, 784-788.

Proudfoot, N.J. and Brownlee, G.G. (1976) Nature 263, 211-214.

SakonJu, S., Bosenhasen, D.F. and Brown, D.D. (1980) Cell 19, 13-25.

Sala-Trepat, J.M., Sarsent, T.D., Sell, S. and Bonner, J. (1979) Proc. Natl. Acad. Sci. USA 76, 695-699.

Sanser, F. and Coulson, R. (1978) FEBS Letters 87, 107-110.

Sargent, T.D., Wu, J.-R., Sala-Trepat, J.M., Wallace, R.B., Reyes, A.A. and Bonner, J. (1979) Proc. Natl. Acad. Sci. USA 76, 3256-3260.

Sargent, T.D., Yang, M. and Bonner, J. (1981) Proc. Natl. Acad. Sci. USA, in press.

Seif, I., Khoury, G. and Dhar, R. (1979) Nucleic Acids Res. 6, 3387-3398.

Setzer, D.R., McGrogan, M., Nunberg, J.H. and Schimke, R.T. (1980) Cell 22, 361-370.

Stein, J.P., Catterall, J.F., Kristo, P., Means, A.R. and O'Malley, B.W. (1980) Cell 21, 681-687.

Strauss, A.W., Bennett, C.D., Donohue, A.M., Rodkey, J.A. and Alberts, A.W. (1977) J. Biol. Chem. 252, 6846-6855.

Sures, I., Lowry, J. and Kedes, L.H. (1978) Cell 15, 1033-1044.

Tucker, P.W., Marcu, K.B., Newell, N., Richards, J. and Blattner, F.R. (1979) Science 206, 1303-1306.

Wallace, R.B., Sargent, T.D., Murphy, R.F. and Bonner, J. (1977) Proc. Natl. Acad. Sci. USA 74, 3244-3248.

Wasylyk, B., Kedinger, C., Corden, J., Brison, O. and Chambon, P. (1980) Nature 285, 367-373.

Wilson, A.C., Carlson, S.S. and White, T.J. (1977) Ann. Rev. Biochem. 46, 537-639.

APPENDIX

The accumulated (13,600 NT as of 1-20-81) nucleotide sequence of DNA associated with the rat serum albumin gene is tabulated. Unsequenced regions are denoted as "gaps", and the approximate sizes of these are given. Intronic DNA is typed in lower case, and exonic DNA in upper case.

The order of the restriction fragments of the cloned rat serum albumin gene is as follows: JA, JC, JB, C, F, G, E, B, H, I, A, D and K, from the 5′ to the 3′ direction. JA is an Eco R1-Hind III fragment, JC is a Hind III-Hind III fragment and JB is a Hind III-Eco R1 fragment. All others are Eco R1-Eco R1 fragments. Fragments G, H and I were not sequenced at all. The nucleotide sequence begins approximately 5000 nucleotides from the 5′ end of JA.

gggattttgtgtgtgtgtgtgtgtgtgtatgtgtgtgtgtgtgtatgtgtgtgtgcgtg
tggtgtgtgtgtgtgtggtgtgtatgtgtgtgtgtgtgtgtggtgtgtatgtgtgtgtgtgt
tgttgttgttgttcccttctagatgatgaaaataggattgtgtttccaaacaacaaatg
attctcagctactctctagttttaaagccccccctacaagcacaagaaatgaccatcccag
ctattttctattcggaaactaaactcctagtttaacttttgcgtttatgctcctgccaac
ttcttctgaaggctaactaaccatctagtttaaaagcttacagtgtctcaagtgtgaact
gggaggaagccacagcgataaaaacaagtgtggaggactctcagaagtcttgtgtcctca
ccccatactcaggctatgtgtgaaacgggtctgaggtgttttagaggagacactagtaga
gcaggcacagtaatgatggtcctcacctttgcaacctatttcctctgacaccacaggcts

ataattastascasaccasttatctcacttaacctaaaassatsastcstctstcaasac

ctcatcccaatcaccccatsstccaatatataccasss

(sar of 120 NT)

sa*tcttaatastccasaaatsaaataacaacasactastcccctsaasacacasstsatc

asaacstaaaaaastaaaaatstctacctctasaattacttccactsssssscasaastss

assacasaattaascccaaaacctcatastssaaatssaaaaaaaaaatactaaacssaaa

asatssttctttcttssataaattcctccssaaacttcctasaaasastsstcastattt

stttsaasacastascccatscctastcctstccttststssatctssataaacscaaaca

astasscaatscttaastccttsaasstattsta

(sar of 440 NT)

sactsttsaasctccaataatttaasacataattcc*scaactsasccctctcctcssca

stcctaattcaatstcttacacssscatattctaattaatcttcascascccaasaattaa

astscttaactcttsacacaatctccatscttacasttcatctcaattsaaaasaaatsc

ttaaacasttccatttsatattsassaatttatactctsctsaaactcctsscattcatc

ctatacstatattctatcctsasasssaatsaascccacattattsttasttacttctaasc

atcctstttttcaascttctasaasatcctaactttaaaaasatsasssaasacttssaa

scccsastcaactcccccttcctcacacsaastcttcacssccctctacatasaaacatat

ttsssaasaacaaaccttscaactcascttttcaaatstsccctsctteccctscaaacts

aaatscccaaatsssasacasasasattaaactcttatstaaaatttsatstttacata

actttaatsaattsacaaascctttsssssstttsssssascttcasatsscaaacatacs

caasssatttasttaaacaacttttttttttctttttsscaassatsstatsatttsta

atsssstassaaccaatsaaatsaaassttastststsattaatsatctacasttattsstt

asasaastatattasascsastttctctscacaCAGACCACCTTTCCTGTCAACCCCACT

GCCTCTGGCACAATGAAGTGGGTAACCTTTCTCCTCCTCCTCTTCATCTCCGGTTCTGCC

TTTTCTAGGGGTGTGTTTCGCCGAGAAGCACstaascattctatstttctcatctctac

exon2

ttttatttstcstastaacssaasctassta

(sap of 450 NT)

ssaassctsssaacttcasssaatastttaastsastsaasasaassstsaaaascascst

sttstaatttsatstcttcaatatttaaatatstttstsstttttctctcccccattc

ccacasACAAGAGTGAGATCGCCCATCGGTTTAAGGACTTAGGAGAACAGCATTTCAAAG

GCCTstaasttaasasscstsaaastccacttaatstttssatastattasttcassttc

tasaaaccctttcccactttsaststscaaattstasascascactcassaatasaaat

(sap of 100 NT)

stsstttttscatscactaaaacaaattaascasaaaaaaaatsttaatsaaatsttact

tttttcttactcctscasacataccctttsatstataasttacttttaaastcatsaas

caaaassaaatcataasaaatctcccttcscscttscctttctttsacttscctatttct

scstaactttcttsasastcaascaccscastattststststsstssataaaacasasaas

caaasascstaacsssctttsassaassttscctttccststttttctatctssscactaaatct

assascaaastcactataattaaaassasaaaaaaaaasccatssssascsacatsttcac

acaactcscsssctccsctcctctsssasaascctastttactcastcststaaaactt

cssaataacaccatctactttastastcsttaaaaatscasatctcctctctattctas

aaatatccccctcaassacttcatttttaaaatatttctttccttcctcactaattttat

tstasaaatcattcaacstcatasaaatt

(sap of 150 NT)

aastatacacatscacaatatsctstttassaasaascccacctscctcctactsassttt

sctttctscctstctttcasAGTCCTGATTGCCTTTTCCCAGTATCTCCAGAAATGCCCA

TATGAAGAGCATATCAAATTGGTGCAGGAAGTAACAGACTTTGCAAAAACATGTGTCGCT

GATGAGAATGCCGAAAACTGTGACAAGTCCATTstsastacattctsatttctsassatc

tacttttctscctttstaattccsastttttcaaassaaascaaatcacstctctcaca

cttaass

exon A

exon B

(gap of 80 NT)

gactcctttcctagtaattgcagaagctacactcacaaagtctcactaatatgactaaa

tgtgaccttgagagtgacagtgacgatacacatgccaaggagagatgaggggaaagaccaca

agaacaaagctatagggcaaccaaagaatgctgagagtaggagaaattgccttctttgagg

aagagtgcaccaactgtgtacccaataccaatgggcttcaaaactgcaggtctgagctct

tctgggcacataatctgtgggcggcac

(gap of 140 NT)

aatttctatgacgttgaatgatttctacaataaaattagtgaggaaggaaagaaatattt

taaaaatgaagtccttgaggggagatatttctagaatagagaggagatctgcatttttaac

gactactaaaagtagatggtgttattccgaagttttacacgactgagtaaaactaggctt

ctcccagaggagcggagcccgcgagttgtgtgaacatgtcgctccccatggctttttttt

ctcctttaattatagtgactttgctcctagatttagtgccagatagaaaacacggaagg

caaccttcctcaacgccgttagctctttgcttctctgttttatccaccacacacaatact

gcggtgcttgattctcaagaaagttacgcagaaatagggcaagtcaaaggaaggcaagcgc

gaaggggagatttcttatgatttcctttgcttcatgactttaaaaagtaacttatacatc

aaagggtatgtctgcaggagtaagaaaaaagtaacatttcattaacatttttttctgct

taatttgttttagtgcatgcaaaaaccac

(gap of 90 NT)

ggtgcatatcctctatgtacgtacatgcaggtacatgttttcttaaccgagctttatcag

cacatttggtttcgacgctgaccttttggctatagaggaagtcgggcaccggcaggggatct

gaccctttctgtcttccacttagCACACTCTCTTCGGAGACAAGTTATGCGCCATTCCAA

AGCTTCGTGACAACTACGGTGAACTGGCTGACTGCTGTGCAAAACAAGAGCCCGAAAGAA

ACGAGTGTTTCCTGCAGCACAAGGATGACAACCCCAACCTGCCACCCTTCCAGAGGCCGG

AGGCTGAGGCCATGTGCACCTCCTTCCAGGAGAACCCTACCAGCTTTCTGGGACAgtgag

tacccagacatctatgcttgaagataaaaatggggcaatagcttagccggtagcacttag

exon C

sctssсastсatssaaaссссaacactsascсttsssatscattssaatsaattctaaaa

acataatctsataccataaascтctacacsscsacaatcaaacaatsacstttsaacts

saaaacasatasssacctststsaaatattctsasactsasaaattscaaaatattctsaa

tactctsaasststttststsaatscststscststststscststststststscatstst

scstscstscatststststststststststststststststststststttsstcats

ttctctssatcaatttcastttastcascaacattacatcacctaaatttaactacttac

caaaaatscacacatsttaacattcacasacacacacatatccacacacasacacacaca

cacacasacacacacasatacasacacacacasacacasacacacacacasacacacca

casacacacacacasacacacacasatacasacacacacasacacasatacacacacaca

cacacacaaacacacacctctctttctcacttaacaasasacaaasaccatcactcccs

sstascсacacactsastsaaactsscсaattacaaatscсacatatatsssaasstats

ssttsactttaataatttccstaststtctstsctstsssaaatsactscctatsttctt

cttcttcctcttcccataattcasCTATTTGCATGAAGTTGCCAGGAGACATCCTTATTT

CTATGCCCCAGAACTCCTTTACTATGCTGAGAAATACAATGAGGTTCTGACCCAGTGCTG

CACAGAGTCTGACAAAGCAGCCTGCCTGACACCGAAGstaatcccctssaaasacssaasa

aatatctscctcstastccaacctccaatcttatsscstasattcassasactcsattttc

tsaatsaaacstctcttcccctcttaasttaccastssatttttsttctcatacсttccaa

aacscascastsstctsasasasssaattc

(sap of 550 NT; Eco R1 frasment G)

saattcttacctascatscacassatttcaactcctccccacccaaaccaattstttsss

catasssctstssssaactttsaaassssstctctctctctctctctctctctctctct

ctctctctctststststststststststststststststststaaaaataacctttccaa

ttttttcaactssatttctttsstasCTTGATGCCGTGAAAGAGAAAGCACTGGTCGC

AGCTGTCCGTCAGAGGATGAAGTGCTCCAGTATGCAGAGATTTGGAGAGAGAGCCTTCAA

AGCCTGstatatsaattttctttaatatcacastccacatctttctsatsatascststcc

aatccaaaaggctgtgcctttataaggcactttatgaaaagacgatagaggtgaaatcgg

attaatagtagtaagttgacatttttgtctcatgctttattttaaattatcaaagaattt

agaagacttttgtttttttcctaaaagaagacatgtatttaaaatctcatttgcttataa

gactgggactcgttttcaaagcttaaggaacgcagtcaaggtgattaatttacttaagtt

gtgttctcctaggctaaaagtgactgttttatgtcctgataattgaatttagccatcaaa

gaagctggagttattactttgaattttattacaactccaactgtgtgtttttttttttaa

ctttttttttttctctgagctgggactgaacctagggccttgcgcttgctgggcaagtgc

tctaccactgagctaaatccccaaccccccaactgtgttcttaatctcatgtttctgttt

gatacctcctccttt

(gap of 70 NT)

ttaacgtcgtcgtatgggttctctggaatccgtcagcacaattatcctcacgaagcattc

ctactatcctactatccgaattatcacagccaaccaaatcaacacacgaggcaacttcag

tatcaaacaatacactgttgtgcaggccgtgactggctcttgcaaaggagtctctttatt

ggagttaaggaactgaaaatgatcttggaatgagagtcttggcttgtgccatagaactta

gaggctgtctgccggaacagtttcttggcgggaggatgcatggttagaactctctaagct

ctctggaagttaagaattttgtttccccatcaggaactttagaattc

(gap of 120 NT)

ttttactgctttcctccaaatacactatgataataattccttgatgggtttcttttttc

exonF

ctttttcagGGCAGTAGCTCGTATGAGCCAGAGATTCCCCAATGCTGAGTTCGCAGAAAT

CACCAAATTGGCAACAGACGTTACCAAAATCAACAAGGAGTGCTGTCACGGCGACCTGTT

GGAATGCGCGGATGACAGGgtaaagagggggattgcaggctcagtcagtgaatgcaagtg

tttggtcttcacccgagatcacttgatggtctgtctggtgt

(gap of 850 NT)

exonG

ttgccttccattctcacagGCAGAACTTGCCAAGTACATGTGTGAGAACCAGGCCACTAT

CTCCAGCAAACTGCAGGCTTGCTGTGATAAGCCAGTGCTGCAGAAATCCCAGTGTCTCGC

TGAGACAGAACATGACAACATTCCTGCCGATCTGCCCTCAATAGCTGCTGACTTTGTTGA

GGATAAGGAAGTGTGTAAGAACTATGCTGAGGCCAAGGATGTCTTCCTGGGCACgtgagt

agatgcctttctcttcgtggctttggcgaagcgtgtggtgtaggaggctacactaggctt

cct

(gap of 600 NT)

exon H
tcagctcagcctatgttcgcctcaattagGTTTTTGTATGAATATTCAAGAAGGCACCCC

GATTACTCCGTGTCCCTGCTGCTGAGACTTGCTAAGAAATATGAAGCCACACTGGAGAAG

TGCTGTGCTGAAGGCGATCCTCCTGCCTGCTACGGCACAGTggtaggtttccgcgagcca

agaacactcacggtgctagctgggctttctgtcaggagaaagacgagctgaattc

(gap of 850 NT; Eco R1 fragments H and I)

gtgtcaagtgacctctttgggctctcggaagggcctgtaaacacaattcttttatcttg

exon I
cagCTTGCAGAATTTCAGCCTCTTGTAGAAGAACCTAAGAACTTGGTCAAAACTAACTGT

GAGCTTTACGAGAAGCTTGGAGAGTATGGATTCCAAAACGCgtgagagtttttttcctt

gatcaacttgtaattatattaaacattatataggccacccagcatatg

(gap of 900 NT)

tatgctatcactccagaagtaaagatactttgattagtgggggaaaaaaatcttaattgg

tttgggcgaatcttcttgaaatatattacatcaggccagttttctgtctggattaaacca

exon J
tgtaagttattgtaagtaataactttttgttacacagCGTTCTGGTTCGATACACCCAGA

AAGCACCTCAGGTGTCGACCCCAACTCTCGTGGAGGCAGCAAGAAACCTGGGAAGAGTGG

GCACCAAGTGTTGTACCCTTCCTGAAGCTCAGAGACTGCCCTGTGTGGAAGACTATgtga

gtcttttaaacaacatcaaagttaacaggggacaggctgttgtcctcagaccttgataaa

tctaactttcaggagcaagggggttctaatatgtgcttacgtgtgtatagtatccgggtc

atacccctctctcccatatctaactaagatattgaaacagtttttaataagtctcacatgt

gcatcacatgtgcagtttcaaagacacacaaattaccaaactgggcaaggctcgcttcat

agtggctagtccaacatgacagtttttagtgcaccggagagcggatgtttgtggaatgct

exon N

tstsaccttccctctccsctttctcttctttasCTGTCTGCCATCCTGAACCGTCTGTG

TGTGCTGCATGAGAAGACCCCAGTGAGCGAGAAGGTCACCAAGTGCTGTAGTGGGTCCTT

GGTGGAAAGACGGCCATGTTTCTCTGCTCTGACAGTTGACGAGACATATGTCCCCAAAGA

GTTTAAAGCTGAGACCTTCACCTTCCACTCTGATATCTGCACACTCCCAGACAAGGAGAA

GCAGATAAAGAAGCAAACstsassatatattctttcscatctctctsttttcatattsca

tatstattsacastsasaaccstcaactssacaa

(sap of 700 NT)

tttssctcccssatttassaccttaaactstssectcsccatccttcccstcttatttte

exon L

ctscaccssastsaatstctccttsctstccttsctttctstcctsctscasGGCTCTCG

CTGAGCTGGTGAAACACAAGCCCAAGGCCACAGAAGATCAGCTGAAGACGGTGATGGGTG

ACTTCGCACAATTCGTGGACAAGTGTTGCAAGGCTGCCGACAAGGATAACTGCTTCGCCA

CTGAGstaacaaatstcttctccatttsatttststasccttccatttctstscactg

stcassttasasctccssaactcacatactssttaatstsatcaatccasattttsttt

sctacacaactsttastasaaaccsacttacstasctcttaatttttatcttctaccaca

ctsctscatattacatssttattatcactatttsstttcaaatttstsacscatacsaac

atssttssaaatacttstaaasccccasaatcsctaactcatttaascctttscccctsaaat

sctacttttsaactsasttattttatacactaatssasaacatttaccattcaatstct

saatcatttcacattcctccsstscccctaacaacastttatcctttttattttssa*****

*****saattcaccaatssaactstaattaascaassccassscacatscatttaccatst

atcatcaccsatssctatssaastscaaaccctaatastcctsctaatatttttctaaca

exon M

tccaccatttttttcctsttcasGGGCCAAACCTTGTTGCTAGAAGCAAAGAAGCCTTAGC

CTAAACACATCACAACCATCTCAGstaactatactcsssaatttaaaacacaatcataa

tcatttttcctaaaacsatcaasatcsataascatttsccaasascccacatssataassc

cascccsscattcttasstcctctcstatsstcstcaatttsssttccatttstasatats

aaactsaacactattsstctassttstacaacacasassassscaaccaaactasssasas

aasacaccaaaattaaaaattccaaasasccacasasttsatcstsstttts ctctssct

tacascaaccactatstasastcscccatsttascccaacctasastatcctctsaatsa

asctsaaaccasatsaatsaacatttttsssts caaaacatttststatcaatsattccat

stassttts acccatttsttctstsaactcatttcaccasaastcastsssttctctttc

ataactsatctsctttcaaatttctsatstsccacatsctaaacttt aacasaatttatc

ttcttatsacaaascascctcctttsaacatatastssctacatssctatst**acaatt

ctsssttts sasccttttssaasasaattttcascattssstsataatssaasacttts

staaaaaattstssctccatatsaaaacaaaaacttacttttt atttaascaaasaacca

atassctataassctatttcaaaatcacttcastttta saaasaattsssastttts cat

cattttsasaasacasctttcctttstaataaattaastaataascasstscccasaaat

sstsasacctctsactttsacacttsasssaccastsatsaasasssssststcacataa

atatcstststcttctscccacc

(sar of 100 NT)

aaaaacatattcaaattttcacaaassacacattcastttctctttststttcaasGCT

exon N

ACCCTGAGAAAAAAAGACATGAAGACTCAGGACTCATCTCTTCTGTTGGTGTAAAACCAA

CACCCTAAGGAACACAAATTTCTTTGAACATTTGACTTCTTTTCTCTGTGCCGCAATTAA

TAAAAAATGGAAGGAATATActctstsstt cssasstctstcttccaacsscscstctca

ccctsscsssctctasssctsssssaaaccctcsstttcctcccttcatctcsscsssssa

tacsssctccatatsscaacacaatssasstaaatsttaatattctcccasstt atstsc

sstsssc acssaaaccaatsttccaacatcacsssatsaattcsaacsasascttscctt

tscsstsssaatsasttsssstt sssssacaaassasatassaatsscttt aaasstsact

asttsaactsstaaastaasatsasttsassaaaaaaactaattcaasatsasassasaa

asscttsasaascccts ctssaatcasaaactssstsatctssasssstactsaascaca

ctssatttsacsaaccctaasastcacacastcaaasattcststtaaaatctasstscc

actttcactsctassctctttststtsttccasscatcasttsctcatatasactatsaa

ttttatagacttatccaacactgggtatgagtttacatacatagtcatttatagccatgt

gaaatttttactgtacgtggctttagctgtgcatattattctacaactggctctttttta

tttatttatttatttttttattatcttgagtatttcttatatacatttcaagtgttatt

ccctttcccggtttccggacagacatcaccctccccccctccccttccttgtgggtgttcc

cctcccaaacctcccccccattgccacccctcccccgcatagtctagttcactgggggttcag

tcttagcaggacccagggcttcccccttccactggtgctcttactaggatattcattgcta

cctatggggacagagtcaggggtcagtccatgtatagtctttaggtagtggcttagtcccct

ggaagctctggttgcttgacattgttgtacttttgggggtctcgagcaccttcaaactctt

ccagttctttctgattccttcaacaggggacccattctcagttcagtggtttgccgcccgg

cattcgccccctgtatttgccgtattccggccgtgtctctcaggagcgatctacatccggc

tcctgtcggtctgcacttctttgcttcatccatcttgtctaattgggtggctgtatatgt

atgggccacctgtggggcaggctctgaatgggtgttccttcagtctctgttttaatcttt

gcctctcccttcccctgccaagggtattcttttcctcatttaaagaaggagtgaagcatt

cacattttgaccatccgtcttgagtttcctttgttctagggacctagggtaatgcaagca

tttgggctaatagccacttatcaatgagtgcataccatgtatgtctttctgtgagtgggt

tagctcactcaggatgatattttccagttccaaccatttgcctacgaagtcataaactcg

ttgttttgatagctgagtaatattccattgtgtagatgtaccacattttctgtatccat

tcctttgttgaaggcatctgggttctttccagtttctggctattataaataaggctgcg

atgaacatagtggagcacgtgtctctttatatgttgaggcatcttttgggtatagcccca

agagaggtatggctggatcctcaggcagttcaatgtccaattttctgaggaacctccaga

ctgatttccagaatggttttaccagtctgcaatcccaccaacaatggaggagtgttcctc

tttctccgcatcctcgccagcatctactgtcacctgagttttgatcttagccattctca

ctggtgtgaggtgaaatctcaggctacaactggctcttaatcctcctcccttacagtttc

actttaagcatagtacttgaagcggtcatgagaccatgtgtaaagagaggtccttttcaa

tcctcttaagcctgcatcctcctttctctcaacctgagctttggcctccctggtaaaagt

atcaaasasasaaaaasactsssaaaasstssaatssasattsssccctsacctsatasa

aacctsctaatssasctaatttascasataatttsaataasststtsctattcaactss

tccaataatssaccasscatssaascactatctssasssscaccsscteasacccactcts

acataccaaacsssaacttaaaactttsasasaattatsectcccastscactatttcct

sascccasttctsctttttstssastcaacacacsaactaaasasacccastscactctst

ttttsasasssaatccatsssaaaataatttctatctstctstcatctstctsctsstct

ctstattttccctcatsatccasaassctsssctactasaacscascasscttactssss

tststststststststststststststststststststststststsctasasct

sstctcatasaacactsssaacatssssaacatsttaasctstsascatsatttaataca

sssttttasstsatttttttatttcsttaaassscaststttsccaasttttcttcaccsts

aasttattactttccccctttscatasscttttctsssssaascaaaccattaaatacasct

ctccccaatstsasacatsaaaaccattaasctttctssaasssssaacsstcctssttcc

ttttattasasaatssattscaaaasatcaststactsactctscccctctctastsassa

ctcactstttctataactttscasaatasaactasssaaaccaastatstasacatacaca

cacatcctctatctscstatatacccatctactcatctacccatctacccatctcccctt

tttcccatttacccatctatccatctacccatctaaccatctacctatttacccatctac

acatctacccttctacccatctacccctctacccctctacccatctacccatctactcts

ctccccttttctaccctsctcccc