Organization and sequence of the human α -lactalbumin gene

Len HALL,*† David C. EMERY,* Michael S. DAVIES,† David PARKER†§ and Roger K. CRAIG†

*Department of Biochemistry and Unit of Molecular Genetics, University of Bristol, Bristol BS8 1TD, U.K., and †Cancer Research Campaign Endocrine Tumour Molecular Biology Group, Medical Molecular Biology Unit, Courtauld Institute of Biochemistry, Middlesex Hospital Medical School, London W1P 7PN, U.K.

A recombinant bacteriophage containing the entire α -lactalbumin gene was isolated from a human genomic library constructed in bacteriophage λ L47. Within this recombinant the 2.5 kb α -lactalbumin gene is flanked by about 5 kb of sequence on either side. The complete nucleotide sequence of the gene and its immediate flanking sequences were determined and compared with those of the rat α -lactalbumin gene. These studies showed that the size, organization and sequence of the exons have been highly conserved, whereas the introns have diverged considerably. In particular, the first intron of the human gene was found to contain an *Alu* repetitive sequence not present in the rat. A high degree of homology (67%) was also observed in the 5' flanking regions, extending as far as 655 nucleotide residues upstream of the transcriptional initiation site. Comparison of the 5' flanking sequences of these two α -lactalbumin genes with those of five casein genes has revealed the presence of a highly conserved region [consensus sequence: RGAAGRAAA(N)TGGACA-GAAATCAA(CG)TTTCTA], extending from position –140 to –110 in all seven sequences examined, suggesting a possible regulatory role in the hormonal control or tissue-specific expression of milk protein genes in the mammary gland.

INTRODUCTION

Peptide and steroid hormones play a major role in the regulation of tissue differentiation during development. One of the ways in which this is achieved is by modulating the expression of a number of gene products that are characteristic of the target cell. Although recent research has established the importance of specific hormone-receptor binding sites in the expression of a number of steroid hormone regulated genes, little is known about the way in which peptide hormones modulate specific gene expression.

Milk-protein genes (encoding a number of caseins and the whey protein α -lactalbumin) provide an attractive system for studying hormonal control since, during lactation, their expression is modulated in the epithelial cells of the mammary gland by an intricate combination of peptide and steroid hormones (Banerjee, 1976). Of particular interest is the differential regulation of the α -lactalbumin and casein genes during the onset of lactation (Ono & Oka, 1980a,b; Burditt *et al.*, 1981). It is generally accepted that insulin and prolactin are required for maximum α -lactalbumin expression, whereas in contrast insulin, prolactin and glucocorticoids are needed for maximum casein gene expression (Guyette *et al.*, 1979; Ono & Oka, 1980a,b; Matusik & Rosen, 1978), though some studies have also suggested a role for glucocorticoids in α -lactalbumin expression (Ono & Oka, 1980b). Evidence points to a predominant role for prolactin in milk-protein mRNA accumulation, through post-transcriptional mechanisms, in addition to the modulation of transcription (Houdebine *et al.*, 1978; Guyette *et al.*, 1979; Bathurst *et al.*, 1980).

Our current understanding of milk-protein gene regulation is based on a combination of studies on the

whole animal, mammary explants, organ cultures and isolated epithelial cells cultured on collagen substrates (see Hall & Campbell, 1986). Such studies, though valuable, do not permit direct manipulation of milk-protein genes as a means of determining the sequences that are involved in their hormonal regulation and tissue-specific expression. Isolation and subsequent expression of transfected or micro-injected milk-protein genes, either in hormonally responsive mammary-gland cell lines or transgenic animals, provide alternative experimental avenues.

As an essential step towards an understanding of the mechanisms and sequences involved in the hormonal regulation of human α -lactalbumin, we have isolated a 12 kb DNA fragment containing the complete gene and its flanking sequences from a genomic library. DNA sequence analysis shows that the gene is interrupted by three introns, the first of which contains an *Alu* repetitive sequence. Comparison of the complete human α -lactalbumin gene and its 5' and 3' flanking sequences with the rat α -lactalbumin gene (Qasba & Safaya, 1984) shows identical structural organization and identifies extensive homology within the 5' flanking regions of the two genes.

In an attempt to identify possible common regulatory elements within the promoter regions of the structurally unrelated α -lactalbumin and casein genes, we have carried out detailed comparisons of the 5' flanking regions of two α -lactalbumin and five casein sequences. These analyses have revealed the presence of a region with the consensus sequence RGAAGRAAA(N)TGG-ACAGAAATCAA(CG)TTTCTA, which is highly conserved in all seven milk-protein sequences examined, both in position as well as sequence, extending from 110 to 140 nucleotide residues upstream of the transcrip-

† To whom correspondence should be addressed.

§ Present address: Wellcome Research Laboratories, Langley Court, Beckenham, Kent BR3 3BS, U.K.

tional initiation site. This is the first report of a large common structural element within the promoter regions of both the α -lactalbumin and casein genes, a candidate control element for the hormone-mediated transcription or mammary-gland specific expression of milk-protein genes.

MATERIALS AND METHODS

Materials

Restriction endonucleases, 'nick-translation' kits, the Klenow fragment of DNA polymerase I and [α - 32 P]dATP (> 800 Ci/mmol) were purchased from Amersham International. NEN GeneScreen *Plus* hybridization transfer membrane was obtained from Du Pont (U.K.) Ltd., Southampton, U.K. All other chemicals were of AnalaR quality or the highest grade available.

Recombinant plasmids

The construction, characterization and sequence analysis of pHB-35, a recombinant plasmid containing a cDNA copy of human α -lactalbumin mRNA (including the complete protein-coding region) inserted into the *Eco*RI site of pAT153 by using homopolymeric dA-dT tails, have been described previously (Hall *et al.*, 1981, 1982).

Preparation of human placental DNA

Frozen human placental tissue (2 g) was pulverized with a hammer, then ground to a powder in a mortar and pestle cooled with liquid N₂. It was then rapidly transferred to a loose-fitting Teflon-in-glass Potter-Elvehjem homogenizer containing 20 ml of solution containing 2% (w/v) SDS, 0.5 mg of proteinase K/ml, 1 mM-EDTA and 20 mM-Tris/HCl buffer, pH 7.6, and homogenized briefly at room temperature. After incubation at 45 °C for 1 h the homogenate was made 0.1 M with respect to NaCl and extracted with an equal volume of phenol/chloroform (1:1, v/v). Total nucleic acid was then precipitated from the aqueous phase with 2.5 vol. of ethanol, and reprecipitated several times to remove traces of phenol. The nucleic acid was redissolved in 150 mM-NaCl/15 mM-sodium citrate buffer, pH 7, and incubated at 37 °C for 1 h in the presence of 40 units of ribonuclease A/ml and 200 units of ribonuclease T₁/ml. After the addition of SDS to a final concentration of 0.1% (w/v) and proteinase K to 0.5 mg/ml, incubation at 37 °C was continued for a further 1 h. The solution was then extracted with phenol/chloroform (1:1, v/v) and repeatedly ethanol-precipitated as described above. Finally, the purified DNA was redissolved in water and stored at 4 °C in the presence of a drop of chloroform.

Construction and screening of a genomic library

High-molecular-mass human placental DNA was partially digested with restriction endonuclease *Sau*3A, then centrifuged through a linear 10–40% (w/v) sucrose gradient in 1 M-NaCl/10 mM-Tris/HCl buffer, pH 8, at 120000 g for 18 h at 20 °C. Fractions containing DNA within the size range 11–20 kb were pooled and concentrated by ethanol precipitation.

Bacteriophage λ L47.1 DNA (Loenen & Brammar, 1980) was digested with restriction endonucleases *Bam*HI and *Sal*I (the latter to reduce the size of the middle fragment) and arms were purified by density-gradient centrifugation under conditions identical with

those described above for partially digested genomic DNA.

Purified bacteriophage λ DNA arms and size-selected genomic DNA (in a 4:1 weight ratio) were incubated at 42 °C for 1 h in 10 μ l of 66 mM-Tris/HCl buffer, pH 7.5, containing 6.6 mM-MgCl₂, 10 mM-dithiothreitol and 0.4 mM-ATP, at a total DNA concentration of 150–200 μ g/ml. Bacteriophage T4 DNA ligase was then added to a final enzyme concentration of 100 units/ml and the mixture was incubated at 15 °C overnight. Chimeric bacteriophage DNA was packaged *in vitro* (Jeffreys *et al.*, 1980) with extracts from mutant lysogen strains BHB2690 and BHB2688, and the resulting bacteriophage particle titre was determined on *Escherichia coli* Q359, a P2 lysogen that will support infection only by recombinant and not by wild-type bacteriophage (see Lederberg, 1957).

Recombinant bacteriophage containing sequences complementary to human α -lactalbumin mRNA were identified by screening approximately 500000 independent plaques (Benton & Davis, 1977) with a 525 bp *Pvu*II–*Sac*I fragment containing most of the coding and 3' non-coding regions of human α -lactalbumin cDNA, obtained from recombinant plasmid pHB-35 (Hall *et al.*, 1981, 1982), and 32 P-labelled to high specific radioactivity by 'nick translation' (Rigby *et al.*, 1977).

Southern-blot analysis

Human lymphocyte genomic DNA was restricted with appropriate restriction endonucleases, and the products were separated by electrophoresis on a 0.8% (w/v) agarose gel. DNA was then blotted by capillary transfer (Southern, 1975) on to an NEN GeneScreen *Plus* membrane, hybridized at 42 °C in the presence of 50% (v/v) formamide and subsequently washed exactly as recommended by the membrane manufacturer except that depurination was omitted. 32 P-labelled pHB-35 plasmid DNA (> 10⁸ c.p.m./ μ g) prepared by 'nick translation' (Rigby *et al.*, 1977) was used as the probe. Bands were detected by autoradiography at –70 °C with Kodak X-Omat S X-ray film and Kodak Lanex intensifying screens.

DNA sequence analysis

Suitable restriction fragments (either single gel-purified species or *A*luI or *Hae*III digests of larger purified fragments) were cloned in M13 mp8 or mp9 vectors. DNA sequence analysis was carried out by the dideoxy chain termination method of Sanger *et al.* (1977, 1980). Autoradiographs were read manually and sequence overlaps were aligned by eye.

RESULTS

Isolation and characterization of human α -lactalbumin genomic clones

A human genomic library in bacteriophage λ L47 (500000 independent plaques) was screened with a high-specific-radioactivity 32 P-labelled nick-translated human α -lactalbumin cDNA probe (derived from recombinant plasmid pHB-35; Hall *et al.*, 1982). Four strong signals were obtained on the primary screen, and each was plaque-purified through either four or five successive rounds of screening. Subsequent restriction analysis of these four independently isolated recombinants revealed them all to have indistinguishable restric-

Fig. 1. Genomic organization of the human α -lactalbumin gene and flanking regions

(a) Cleavage sites for a number of hexanucleotide-specific restriction enzymes within recombinant bacteriophage λ h α LA-1 are indicated. Those fragments whose size has been confirmed by Southern-blot analysis of total human genomic DNA (see b) are indicated by heavy lines. (b) Southern-blot analysis of total human genomic DNA with an α -lactalbumin cDNA probe. Human lymphocyte genomic DNA was restricted, blotted and probed with a human α -lactalbumin cDNA probe as described in the Materials and methods section. Lane A, *Bam*HI; lane B, *Eco*RI; lane C, *Hind*III; lane D, *Pvu*II. Arrows indicate the positions of *Hind*III-restricted bacteriophage λ DNA size markers.

tion maps (results not shown), suggesting that they were all derived from a single initial recombinant DNA molecule. All subsequent work was carried out on one of these clones, designated λ h α LA-1.

Detailed restriction analysis of λ h α LA-1 DNA with a series of hexanucleotide-specific endonucleases revealed that this recombinant bacteriophage contained about 12 kb of human genomic sequence (Fig. 1a). When this was mapped by Southern-blot analysis with cDNA probes, the α -lactalbumin exons were found to be located in the centre of the insert, almost exclusively within a 3.05 kb *Sac*I DNA fragment. Southern-blot analysis of total human genomic DNA (Fig. 1b) restricted with *Bam*HI, *Eco*RI, *Hind*III and *Pvu*II, and probed with α -lactalbumin cDNA, established that the

sizes of the exon-containing restriction fragments were identical with those mapped in recombinant λ h α LA-1, at least as far as the positions of the *Eco*RI and *Pvu*II sites about 2 kb upstream and 2 kb downstream of the α -lactalbumin gene. The absence of the large terminal *Hind*III fragment and large terminal *Bam*HI fragment, both of as yet undetermined size, may reflect the absence of a depurination step in our blotting protocol, since depurination gave rise to the selective loss of the small *Hind*III fragments. Southern blotting, in addition to substantiating the gross organization of the cloned α -lactalbumin gene, also demonstrates that it is unlikely that human α -lactalbumin is a member of a closely related family of genes.

For nucleotide sequence analysis by the dideoxy chain

Fig. 2. Sequencing strategy for the human α -lactalbumin gene

Only sites for restriction enzymes that were utilized for DNA sequencing are shown. Arrows indicate the direction and extent of individual sequencing runs. Owing to the nature of the shotgun cloning, many fragments were sequenced several times, although for clarity only a single arrow is shown for each different restriction fragment.

-700 -650
 GAGCTCCTGGGCTCAAGTGATCCACCAGACTCGGCTCCCAAAATGCCGGATTACAGGTGTGAGCCACTGTGCTGGCCTAGATGCTTTTCATACAGGCTTTTCAATTAT
 -500 -550
 GCATTTTCCTTAAGTAGGAAGTCTTAAGATCCAAGTTATATCGGATTGTTGTAGTCTACGTTCCCATATTCTATTCTTCTGAGCCTTCAGTCATGAGCTACCATAT
 -500 -450
 TAAAGAACTAATTCTGGGCTTGTTACATGGCTGGATTGGTTGGACAAGTGCCAGCTCTGATCCTGGACTGTGGCATGTGATGACATACACCCCTCTCCACATTCTGC
 -400 -350 -300
 ATGTCTCTAGGGGGGAAGGGGAAGCTCGGTATAGAACC1TTATTGTATTTTCTGATTGCCTCACTTCTTATATTGCCCCATGCCCTTCTTTGTCTCTCAAGTAACCAG
 -250 -200
 AGACAGTGCTTCCCAGAACCAACCTTACAAGAAACAAAGGGCTAAACAAAGCCAAATGGGAAGCAGGATCATGGTTTGAACCTTTTCTGGCCAGAGAACAATACCTGCTA
 -150 -100
 TGGACTAGATACTGGGAGAGGAAAGGAAAGTAGGGTGATTATGGAAGGAAGCTGGCAGGCTCAGCGTTTCTGTCTGGCATGACCAGTCTCTCTTCACTTCTCTCTCT
 -50 1
 AGATGTAGGGCTTGGTACCAGAGCCCTGAGGCTTTCTGCATGAATATAAATAAATGAACTGAGTGATGCTTCCATTTCAAGTTCTTGGGGTAGCCAAA ATG AGG
 Met Arg
 -19
 50 100
 TTC TTT GTC CCT CTG TTC CTG GTG GGC ATC CTG TTC CCT GCC ATC CTG GCC AAG CAA TTC ACA AAA TGT GAG CTG TCC CAG CT
 Phe Phe Val Pro Leu Phe Leu Val Gly Ile Leu Phe Pro Ala Ile Leu Ala Lys Gln Phe Thr Lys Cys Glu Leu Ser Gln Le
 -10 150 -1 +1 200 10
 G CTG AAA GAC ATA GAT GGT TAT GGA GGC ATC GCT TTG CCT GAA T GTGAGTCCCTGCCTCTGTGTTTCATCCATCTCTCATACGTTCTCTCTCT
 u Leu Lys Asp Ile Asp Gly Tyr Gly Ile Ala Leu Pro Glu L
 20 250 300
 CCATCCCTCTTTCTTCCACTTCGCCCCCTCCACTTTTACTTAATTATCTAATCATCTCTTTTCTGCTCATTTCGATACCTTTTATTTCATGTATGTATATGTATGT
 350 400
 ATTTATTTATTTTGGAGTGGAGTTTCGCTCTTGTTCGCCAGACTGGAGTGCAATGGTGAATCTCGGCTCACTGCAACCTCCGCCCTCCGTTCAAGTGATTCTCCTG
 450 500
 CCTCAGCCTCCCAAGTAGCTGGAATTACAGGCACCCACCACCATGCCTGGCTAATTTTGTATTTTGTAGAGACAGGGTTTCAACCATGTTGGCCAGGCTGGTCTCAAAC
 550 600 649
 TTCTGACCTCAGGTGATCCGCCCTCCTCAGCCTCCCAAGTGTGGGATTACAAGCGTGAGCCATCATGCCTGGCCCCATTTATTTTCTATCCTTTCTTCTCTTATTG
 700 750
 TCTGATTTTTTTTGAATTCTCCATCTCATCAAGAACTCTGAGCTTTGCCATCTTTGGAGATTGGCTGGAAAGCATTTTGTCTGAGAATTACAGTTCTCTCTTTATG
 800 850
 CAGATCCTGTACATCTCTGTGTATCTCTTTCTCATCTTTCCCTCAG TG ATC TGT ACC ATG TTT CAC ACC AGT GGT TAT GAC ACA CAA GCC ATA
 eu Ile Cys Thr Met Phe His Thr Ser Gly Tyr Asp Thr Gln Ala Ile
 30 40
 900
 GTT GAA AAC AAT GAA AGC ACG GAA TAT GGA CTC TTC CAG ATC AGT AAT AAG CTT TGG TGC AAG AGC AGC CAG GTC CCT CAG TC
 Val Glu Asn Asn Glu Ser Thr Glu Tyr Gly Leu Phe Gln Ile Ser Asn Lys Leu Trp Cys Lys Ser Ser Gln Val Pro Gln Se
 50 60 1000
 A AGG AAC ATC TGT GAC ATC TCC TGT GAC A GTGAGTAGCCCTATAACCTCTTTCTCTGTTTCTGAGGCTGCCCTTGGGATAATCTCCTTTTATG
 r Arg Asn Ile Cys Asp Ile Ser Cys Asp L
 70 1050 1100
 GCCAAGCAGACCTCAGGCTTCATTGCCTTGGCTGGGCTCTATAAAAATTGTGGGACTTGAATTGGCAGTACTGAGTAAGAAGCTGTTGGATTTTTCATGGTCATCAAAT
 1150 1200 1250
 CCCAGACAGTTCCTTGAGGTTTCAGTGGTAGACAATCGAGCTGTCTGAGAGTCTGGAATCTGATTGTCTGCATTTTCAGGGTAAGTCAGTTGATGAAGCTGATGATT
 1300 1350
 CTCCAGAGATATCCCAGGGAATGAAGGAAGTCCCTACCCAGGGTTAGACATTACCACATTTGGTCTTTTCATAGAAAGACAACAGGCACAGCCTTGAGTTTAGAGAA
 1400 1450
 CCCACTGGATCCAGGGGTTAGGGGAAGTCTGCTGGGTAATACTTGTGAGTGTCTCAATCCTTTCCCTGTAACCTCTGCCAG AG TTC CTG GAT GAT G
 ys Phe Leu Asp Asp A
 80 1500 1550
 AC ATT ACT GAT GAC ATA ATG TGT GCC AAG AAG ATC CTG GAT ATT AAA GGA ATT GAC TAC TG GTGAATCCTTATTCTATTCTTATTCTC
 sp Ile Thr Asp Asp Ile Met Cys Ala Lys Lys Ile Leu Asp Ile Lys Gly Ile Asp Tyr Tr
 90 100
 1600 1650
 CCATCCTCTCTCTTACCCATTAGCCAGCACCCCTTCTCTTACCTATCTCTTGGTCATTAACTAGAATACAGTGTCTGAAACAAAGCTTACCTAGAGACTC
 1700 1750
 AGGTTTCTGTATTATTAAGCTCTCTCGCTCCGCTCCTTGGTAGCAATTTTCTAATAAGGGGTTGCCTAATGGAGGGCTCAGACCCAGGCTCTCTTCACTTAGACTTGGA
 1800 1850
 CATCTAATCCACTTGTAGTCTATGCCCCAAGCAAGCTGTTGGTAACATTGCATCTCTTTTTAACCTACAATTTTCTTGATATTTTTTATGGACTGTATTCCA
 1900 1950 1998
 CTTGATGGCTTGTGCTGCTTGACATCAGGCCAGGAATGTCTTCTGTAATTTCTGTCACGCTCTTCCACTTCAGCCCTCCTGGGAATGAATGTAAGATTACGTACAGCT
 2050
 AACTCACCTTGTCCCTTCTCCATTATCAG G TTG GCC CAT AAA GCC CTC TGC ACT GAG AAG CTG GAA CAG TGG CTT TGT GAG AAG TTG
 p Leu Ala His Lys Ala Leu Cys Thr Glu Lys Leu Glu Gln Trp Leu Cys Glu Lys Leu
 110 120 123
 2100 2150
 TGA GTGCTGCTGTCTTGGCACCCTGCCACTCCCACTCCTGGAATACCTCTTCCCTAATGCCACCTCAGTTTGTCTTCTTCTGTTCCTCCCAAGCTTATCTGTCT
 Stop
 2200 2250 2300
 CTGAGCCTTGGGCCCTGTAGTGACATACCGAATTTCTGAAGACTATTTTCCAGGGATGCCTGAGTGGTGCAGTCTGAGCCTTAGACCTTACTCAGTGCCTTCGATGGCAC
 2350 2400
 TTTCACTACAGCACAGATTTACCTCTGTCTTGAATAAGGTCCCCTTTGAACTCAGTGGCTGTAATTTTTTCCCTCCCTGGAGGGAAGGGGAAGAAATAGGATGAGTAG
 ...GAAGTC(A)_n (cDNA)
 2452 2500
 GTGGACACTGAAGCCATAGGTATAGCCACCTTCCATCTCTACTGAAGAAGAAGTAGGCTGAATTTACAATAGAAAGGTGAAGTTACTGTCTGTACCAACTCAATGCAA
 2550 2575
 CAAACTTTTATTGATCACTAATCTATTCAAGGAAGCTAGACGGATCC

Table 1. Exon and intron sizes for the human and rat α -lactalbumin genes

	No. of nucleotide residues						
	Exon I	Intron I	Exon II	Intron II	Exon III	Intron III	Exon IV
Human α -lactalbumin	159	648	159	489	76	499	333
Rat α -lactalbumin	165	341	159	429	76	1016	328

termination method, the 3.05 kb *SacI* fragment of λ h α LA-I was subcloned into the plasmid pUC13 (to produce recombinant plasmid ph α LA-1), and convenient restriction fragments were further subcloned into M13. The nucleotide sequence of the extreme 3' end of the gene and the 3' flanking sequence were obtained from a subcloned 1.9 kb *EcoRI* fragment (ph α LA-13). Sequencing was performed on both strands and across all restriction sites by using the strategy depicted in Fig. 2, and the resulting nucleotide sequence (3309 bases) is shown in Fig. 3. Comparison of the sequence with that of human α -lactalbumin cDNA (Hall *et al.*, 1982) revealed the presence of four exons and showed no discrepancies in those regions of sequence common to genomic and cloned cDNA sequences.

Comparison of the position of the exon-intron boundaries in the human and rat (Qasba & Safaya, 1984) α -lactalbumin genes reveals that all three introns occur at identical positions in the two genes, namely within the codons for amino acid residues Leu-26 (intron I), Lys-79 (intron II) and Trp-104 (intron III) in the case of the human gene. The latter two residues are conserved in both species, whereas Leu-26 is replaced by Trp-26 in the rat protein. This strict homology in the sites of insertion of introns suggests that the exon-intron organization of these two genes was established before the divergence of the two species. Whereas the position of the exon-intron boundaries is conserved, as is the size of the exons (Table 1), the sizes of the introns differ markedly between these two species. This is particularly apparent for intron I, which is much larger in the human gene, and intron III, which is much larger in the rat, both a reflection of inserted repetitive sequences (see below).

Examination of the nucleotide sequence for conserved regions thought to be involved in gene transcription and RNA processing reveals a number of features in common with other mammalian single-copy genes. The position of nucleotide 1, the putative transcriptional start site (CAP), an adenosine residue in the centre of a pyrimidine-rich region, has been provisionally designated on the basis of homology in this region with the rat (Qasba & Safaya, 1984) and guinea-pig (Laird, 1985) α -lactalbumin genes. The scarcity of lactating human mammary-gland RNA precluded the necessary S1 mapping or primer-extension experiments required for a more precise determination of the transcriptional start site. Preceding the initiation site is an A-rich region

containing the sequence TATAAA(T), identical with the consensus sequence known as the Hogness box, found within the promoter region (centred around position -25) of most eukaryotic genes. However, the other common feature of many eukaryotic promoters, the so-called CAAT box, which normally lies between about -80 and -70, is not present in a recognizable form in the human α -lactalbumin promoter region.

In common with other eukaryotic genes, the dinucleotides GT and AG occur at the 5' and 3' ends respectively of the introns (Mount, 1982). The sequences CTCAT (intron I), CTCAA, CTCAG and GTCAG (intron II), and CTCAC (intron III), all show reasonable homology to the weakly conserved lariat branch site (Keller & Noon, 1984) and are found close to the corresponding splice acceptor site. At the 3' end, cleavage and polyadenylation occurs at a site C(A) 17 nucleotide residues downstream from the 'polyadenylation signal' (AATAAA), immediately preceding a T+G-rich region containing an oligo(dT)-rich sequence, similar to sequences found in several other eukaryotic genes and implicated in precursor cleavage and polyadenylation (Birnstiel *et al.*, 1985).

Identification of an *Alu* repetitive sequence within the first intron of the human α -lactalbumin gene

Southern-blot analysis of total human genomic DNA, with various subclones of λ h α LA-1 as probes, revealed the presence of a highly repetitive sequence within the first intron of the α -lactalbumin gene. This was subsequently shown after nucleotide sequence analysis to be a member of the *Alu* repetitive family (Houck *et al.*, 1979). Comparison of the α -lactalbumin-associated repetitive sequence (which was found to be inserted in a reversed orientation relative to the α -lactalbumin gene) with the consensus *Alu* sequence (Deininger *et al.*, 1981) gave an overall nucleotide homology of 88% (Fig. 4). In addition, it contained the various conserved internal elements shared by most *Alu* and *Alu*-like sequences. These include (see Fig. 4) the postulated RNA polymerase III promoter boxes 'A' and 'B' (Fowlkes & Shenk, 1980), a sequence (GAGGCNGAGGC) that corresponds to the T-antigen-binding sequence of the SV40 replication origin (Jelinek *et al.*, 1980), a short conserved symmetrical sequence (CCAGCCTGG) of unknown function, and a region (GCCTGGCC) overlapping the above symmetrical sequence that can base-pair with the 5' end of the *Alu* sequence (GGCCAGGC) (see Rogers, 1985). Similar *Alu*-like repetitive sequences have previously been found in a number of mammalian genes, both within introns in either orientation, as well as within the 3' untranslated regions of a few mature mRNAs (Rogers, 1985).

A computer search for other repeated sequences

Fig. 3. Complete DNA sequence of the human α -lactalbumin gene and immediate flanking sequences

The sequence is numbered from the putative transcriptional start site (see the text). The site of polyadenylation in the transcript is indicated.

Fig. 4. *Alu* sequence homology

The *Alu* repetitive sequence found within the first intron of the human α -lactalbumin gene (nucleotides 334–616 in Fig. 3) is shown aligned with the *Alu* consensus sequence. Dots indicate positions at which the two sequences differ. The postulated RNA polymerase III promoter boxes 'A' and 'B' are shown. Underlined regions are features common to most *Alu* and *Alu*-like sequences (see the text).

identified a second *Alu* repetitive sequence in the 5' flanking region of the human α -lactalbumin gene, extending from nucleotide position –656 (the precise point at which divergence occurs between the human and

rat 5' flanking sequences; see below) to a point beyond the limit of sequence determination.

Characterization of sequence homology between the human and rat α -lactalbumin genes and their flanking sequences

Comparison of the degree of homology between the human and rat α -lactalbumin genes and their immediate flanking sequences by using DIAGON, an interactive computer graphics program (Staden, 1982), demonstrated a high degree of homology within the exons and in addition in the 5' flanking region of the two genes (Fig. 5). Evidence for some homology was also evident in intron II. In contrast, introns I and III have diverged considerably, both in length and in sequence, such that at this level of match no significant homology was apparent. Part of the difference in length found in intron I can be accounted for by the presence of the *Alu* repetitive sequence in the human gene, and the corresponding difference in length in intron III reflects the presence of a B2 repetitive sequence in the rat α -lactalbumin gene (see Fig. 5 and Rogers, 1985). The di- and tri-nucleotide repeats (TG)₂₄, (TG)₂₁, (TCC)₂₃ and (TAT)₁₈ present in introns I and III of the rat α -lactalbumin gene (see Fig. 5) are not present in the human gene. Alignment of the rat and human α -lactalbumin genes, allowing gaps for best fit (results not shown), showed homology between exons 1, 2, 3 and 4 to be 73%, 81%, 84% and 72% respectively. Overall, in terms of the mature processed mRNA, there was 83% homology in the 5' non-coding region, 78% in the protein-coding region and 70% in the 3' non-coding region. In the 5' flanking regions of the genes there was 68% overall homology from nucleotide positions –1 to –655, after which no alignment could be found, owing to the presence of the *Alu* repeat in the human sequence. Within the limited rat and human sequence available for the 3' flanking regions, homology was 63% between the two species.

Fig. 5. DIAGON analysis of the human and rat α -lactalbumin genes

Areas of homology between the human and rat α -lactalbumin genes are illustrated by means of a DIAGON plot (Staden, 1982) (parameters: span length = 7, score = 7). Exons are indicated as boxes. 'Alu' and 'B2' represent the repetitive sequences found associated with the human and rat genes respectively. a–d represent repeating di- or tri-nucleotides found within the rat gene: a, (TCC)₂₃; b, (TG)₂₄; c, (TAT)₁₈; d, (TG)₂₁.

Overall consensus:	R G A A G R A A A(N) T G G A C A G A A A - T C A A(C G) T T T C T A	
Rat α -casein:	 T - T C T	86.2%
Rat β -casein:	. A - . A . . A - C . . T -	82.7%
Rat γ -casein:	 C - . C C . G . . . G . - . . . -	79.3%
Bovine α_{s1} -casein:	. . . G . . . C T - . A C A . T G . - C T	69.0%
Guinea-pig α_{s2} -casein:	 C - . G . - . . . - - -	79.3%
Human α -lactalbumin:	 G C . . . - . . . G C - . . . G C G G	75.8%
Rat α -lactalbumin:	. . . G G T . . . - . . . G C - . . . G G C G	72.4%

Fig. 6. Conserved region within the 5' flanking sequences of seven milk-protein genes

The regions from about -140 to -110 of two α -lactalbumin and five casein genes were aligned with the introduction of a small number of gaps (-) to maximize homology. Dots represent homology with the derived consensus sequence. Percentage homologies for each sequence are expressed relative to the consensus sequence.

DISCUSSION

Our primary interest in the comparative analysis of human and rat α -lactalbumin gene sequences relates to the possible identification of conserved sequences of potential regulatory significance. One of the most interesting features of this analysis is the considerable conservation of sequence homology (67%) in the 5' flanking regions of the human and rat α -lactalbumin genes, extending as far as the *Alu* repeat at position -656 in the human sequence. Although such comparisons do highlight a number of highly conserved regions, the availability of only two α -lactalbumin gene sequences does not provide the basis of lengthy speculation as to their potential biological significance at this stage. In view of the potential regulatory role of glucocorticoids (Ono & Oka, 1980b) it is, however, worth noting that the hexanucleotide CTTCCT, its inverse complement AGGAAG, or a very close derivative of these, are found seven times within the human and rat 5' flanking regions, and that these are all in conserved regions. In two cases this hexanucleotide forms part of the larger sequence, CTTCCTAGA. These sequences bear some resemblance to the hexanucleotide TGT(T/C)CT, shown to be part of a number of glucocorticoid receptor binding sites (Scheidereit *et al.*, 1983; Karin *et al.*, 1984; Renkawitz *et al.*, 1984; von der Ahe *et al.*, 1985), but little homology with the more extended consensus sequence for glucocorticoid receptor binding, (T/C)GGTN(A/T)CA(A/C)-(A/T)NTGT(T/C)CT (Scheidereit *et al.*, 1983; Cato *et al.*, 1984; Karin *et al.*, 1984; Renkawitz *et al.*, 1984; von der Ahe *et al.*, 1985). No sequences showing homology with oestrogen receptor and progesterone receptor binding sites (see von der Ahe *et al.*, 1985) were identified.

Although the α -lactalbumin and casein genes are structurally and evolutionarily unrelated, they are likely to share common regulatory features, since both are expressed in a tissue-specific manner in the lactating mammary gland and both require the peptide hormone prolactin for maximal expression. In an attempt to identify regions of sequence similarity within potential regulatory regions of the casein and α -lactalbumin genes, we have carried out comparisons between their 5' flanking regions. This analysis included the human and rat (Qasba & Safaya, 1984) α -lactalbumin sequences, the recently reported rat α -, β -, γ - and bovine α_{s1} -casein genomic sequences (Yu-Lee *et al.*, 1986) and our unpublished guinea-pig α_{s2} -casein sequence. A significant finding was a run of about 30 nucleotide residues, highly conserved in casein and α -lactalbumin genes, both in terms of

sequence and position (from about -110 to -140), in all seven genes that we examined (Fig. 6). This corresponded to one of the regions that had already been found to be conserved within the three different rat casein genes (Yu-Lee *et al.*, 1986). Homology between the human and rat α -lactalbumin sequences in this region was 87%. After the introduction of a small number of gaps to maximize alignment an overall consensus sequence RGAAGRAAA(N)TGGACAGAAATCAA(CG)TT-TCTA was derived for both casein and α -lactalbumin genes. Comparison of each individual milk-protein sequence with this consensus sequence produced levels of homology ranging from 69% to 86%. The average homology relative to the consensus sequence was 74% for the two α -lactalbumin genes and 79% for the five casein genes, suggesting that the analysis was not unduly biased by the greater number of casein sequences. In view of the evolutionary relationship between the α -lactalbumin and egg-white lysozyme genes (they exhibit significant nucleotide sequence homology within the coding regions and share an identical exon-intron organization; see Hall *et al.*, 1982; Qasba & Safaya, 1984), and the involvement of glucocorticoids in lysozyme expression, we have included the lysozyme gene (Grez *et al.*, 1981) in our analysis of 5' flanking sequences. However, no significant homology was observed between the α -lactalbumin and lysozyme flanking sequences, and, in particular, the highly conserved region found within the α -lactalbumin and casein genes was not present.

This is the first report of a region of homology common to the 5' flanking regions of both the casein and the α -lactalbumin genes. What makes this finding even more significant is the fact that this sequence is located at exactly the same place relative to the transcriptional start site in all milk-protein genes so far examined. At the moment we can only speculate as to its function, the most obvious being a role in prolactin-mediated transcriptional control or mammary-gland-specific gene expression. However, definitive evidence for a regulatory role in milk-protein gene expression for this consensus must await a detailed study of the properties of manipulated milk-protein genes, after their re-introduction into hormone-responsive mammary-gland cell lines or transgenic animals.

We acknowledge the assistance of Professor W. J. Brammar in the construction of the genomic library. This work was supported by a grant from the Medical Research Council and in part by the Cancer Research Campaign.

REFERENCES

- Banerjee, M. R. (1976) *Int. Rev. Cytol.* **47**, 1-97
- Bathurst, I. C., Craig, R. K., Herries, D. G. & Campbell, P. N. (1980) *Eur. J. Biochem.* **109**, 183-191
- Benton, W. D. & Davis, R. W. (1977) *Science* **196**, 180-182
- Birnsteil, M. L., Busslinger, M. & Strub, K. (1985) *Cell* (Cambridge, Mass.) **41**, 349-359
- Burditt, L. J., Parker, D., Craig, R. K., Getova, T. & Campbell, P. N. (1981) *Biochem. J.* **194**, 999-1006
- Cato, A. C. B., Geisse, S., Wenz, M., Westphal, H. M. & Beato, M. (1984) *EMBO J.* **3**, 2771-2778
- Deininger, P. L., Jolly, D. J., Rubin, C. M., Friedmann, T. & Schmid, C. W. (1981) *J. Mol. Biol.* **151**, 17-33
- Fowlkes, D. M. & Shenk, T. (1980) *Cell* (Cambridge, Mass.) **22**, 405-413
- Greß, M., Land, H., Giesecke, K. & Schutz, G. (1981) *Cell* (Cambridge, Mass.) **25**, 743-752
- Guyette, W. A., Matusik, R. J. & Rosen, J. M. (1979) *Cell* (Cambridge, Mass.) **17**, 1013-1023
- Hall, L. & Campbell, P. N. (1986) *Essays Biochem.* **22**, 1-26
- Hall, L., Davies, M. S. & Craig, R. K. (1981) *Nucleic Acids Res.* **9**, 65-84
- Hall, L., Craig, R. K., Edbrooke, M. R. & Campbell, P. N. (1982) *Nucleic Acids Res.* **10**, 3503-3515
- Houck, C. M., Rinehart, F. P. & Schmid, C. W. (1979) *J. Mol. Biol.* **132**, 289-306
- Houdebine, L.-M., Devinoy, E. & Delouis, C. (1978) *Biochimie* **60**, 57-63
- Jeffreys, A. J., Wilson, V., Wood, D., Simons, J. P., Kay, R. M. & Williams, J. G. (1980) *Cell* (Cambridge, Mass.) **21**, 555-564
- Jelinek, W. R., Toomey, T. P., Leinwand, L., Duncan, C. H., Biro, P. A., Choudary, P. V., Weissman, S. M., Rubin, C. M., Houck, C. M., Deininger, P. L. & Schmidt, C. W. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 1398-1402
- Karin, M., Haslinger, A., Holtgreve, H., Richards, R. I., Krauter, P., Westphal, H. M. & Beato, M. (1984) *Nature* (London) **308**, 513-519
- Keller, E. B. & Noon, W. A. (1984) *Proc. Natl. Acad. Sci. U.S.A.* **81**, 7417-7420
- Laird, J. E. (1985) Ph.D. Thesis, University of London
- Lederberg, S. (1957) *Virology* **3**, 496-513
- Loenen, W. A. M. & Brammar, W. J. (1980) *Gene* **10**, 249-259
- Matusik, R. J. & Rosen, J. M. (1978) *J. Biol. Chem.* **253**, 2343-2347
- Mount, S. M. (1982) *Nucleic Acids Res.* **10**, 459-472
- Ono, M. & Oka, T. (1980a) *Cell* (Cambridge, Mass.) **19**, 473-480
- Ono, M. & Oka, T. (1980b) *Science* **207**, 1367-1369
- Qasba, P. K. & Safaya, S. K. (1984) *Nature* (London) **308**, 377-380
- Renkawitz, R., Schutz, G., von der Ahe, D. & Beato, M. (1984) *Cell* (Cambridge, Mass.) **37**, 503-510
- Rigby, P. W. J., Dieckmann, M., Rhodes, C. & Berg, P. (1977) *J. Mol. Biol.* **113**, 237-251
- Rogers, J. H. (1985) *Int. Rev. Cytol.* **93**, 187-279
- Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463-5467
- Sanger, F., Coulson, A. R., Barrell, B. G., Smith, A. J. H. & Roe, B. A. (1980) *J. Mol. Biol.* **143**, 161-178
- Scheidereit, C., Geisse, S., Westphal, H. M. & Beato, M. (1983) *Nature* (London) **304**, 749-752
- Southern, E. M. (1975) *J. Mol. Biol.* **98**, 503-517
- Staden, R. (1982) *Nucleic Acids Res.* **10**, 2951-2961
- von der Ahe, D., Janich, S., Scheidereit, C., Renkawitz, R., Schutz, G. & Beato, M. (1985) *Nature* (London) **313**, 706-709
- Yu-Lee, L.-Y., Richter-Mann, L., Couch, C. H., Stewart, A. F., Mackinlay, A. G. & Rosen, J. M. (1986) *Nucleic Acids Res.* **14**, 1883-1902

Received 25 June 1986/22 September 1986; accepted 14 November 1986