Structure and Function of Milk Protein Genes

JEAN-CLAude MERCIER and JEAN-LUC VILOTTE
Laboratoire de Génétique Biochimique
Institut National de la Recherche Agronomique
Centre de Recherches de Jouy-en-Josas
78352 Jouy-en-Josas Cedex, France

ABSTRACT
Interspecies comparisons of cDNA and mosaic milk protein genes have confirmed their high rate of evolution, but the overall gene organization has been conserved. The three Ca-sensitive casein genes, which share common motifs in the promoter region and contain similar sequences that encode signal peptide and multiple phosphorylation sites, probably derived from a common ancestor. a1- and a2-casein genes, divided into many small exons, undergo complex splicing, and the deleted caseins arise from exon skipping. The four bovine casein genes are clustered on 200 kb of chromosome 6. a-Lactalbumin and $\beta$-lactoglobulin pseudogenes occur in ruminants.

Study of the expression of native and modified milk protein genes in mammary cell lines and transgenic animals and DNA footprinting have shown the occurrence of important regulatory motifs in the proximal 5' flanking region, including one recognized by a specific mammary nuclear factor. Good stage- and tissue-specific expression has been obtained in transgenic animals with milk protein genes having less than a 3-kb 5' flanking region.

Better knowledge of both the structure and function of milk protein genes, which has already allowed the use of powerful techniques for the rapid identification of alleles, offers the potential for the genetic modification of milk composition.

(Key words: milk protein, messenger ribonucleic acid, gene, structure)

INTRODUCTION

Previous structural and genetic studies of the major milk proteins from cattle and from a few other species have contributed greatly to the knowledge of the relevant genes. Elucidation of the primary structure of casein components in the 1970s led to the identification of only four types of casein ($\alpha_{s1}$-, $\alpha_{s2}$-, $\beta$-, and $\kappa$-) and the characterization of most genetic variants. It was clearly established that the heterogeneity of whole casein arose from incomplete $\alpha$-phosphorylation of caseins, $\alpha$-glycosylation of $\kappa$-casein, partial proteolysis by plasmin, and genetic polymorphism. These genetic variants were used as markers for Mendelian segregation analyses, which showed the transmission of solely the parental casein haplotypes to the progeny. The tight linkage of $\alpha_{s1}$- and $\beta$-casein genes was first demonstrated in 1964 (55), and the relative order of the three-casein locus, $\alpha_{s1}$-$\beta$-$\kappa$, was postulated in 1973 (58).

Some interesting evolutionary features were inferred from amino acid sequence analysis of milk proteins: 1) the evolution of $\alpha_{s1}$- and $\alpha_{s2}$-casein genes by intragenic duplication as deduced from the internal similarity observed in $\alpha_{s1}$- and especially $\alpha_{s2}$-casein polypeptide chains (18); 2) the probable common origin, proposed in 1977 (47), of the three Ca-sensitive casein genes ($\alpha_{s1}$-, $\alpha_{s2}$-, and $\beta$-), which share similar multiple phosphorylation sites (104) and signal peptides (47); 3) the high evolutionary rate of caseins (106); and 4) the evolutionary relationship between $\kappa$-casein and fibrinogen (76), between $\alpha$-lactalbumin and lysozyme (16), and between $\beta$-lactoglobulin, retinol-binding protein (51, 113), and human placental protein 14 (73).

Much concurrent progress was made in the study of the complex endocrine balances that
differ among species and are responsible for the changes in structure and activity of mammary tissue. In the 1970s, the role of prolactin and steroid hormones in the induction and modulation of milk protein synthesis, and especially the striking correlation between the level of specific mRNA and the rate of protein synthesis, were clearly recognized [reviewed in (103, 125, 146, 154)].

The rapid development of molecular biology methodology in the 1980s gave tremendous stimulus to current research in the dairy field. Indirect knowledge of milk protein genes, which was limited to the coding frame, has been confirmed by direct analysis of mRNA (cDNA) and genes. This analysis has also provided new insight and perspectives.

The ease and automation of DNA sequencing have greatly facilitated the characterization of mammary cDNA and genes and, consequently, of milk proteins in various species. Some 60 cDNA and 20 genes from 12 species have already been completely sequenced. These cDNA or genes can now be modified in vitro by site-directed mutagenesis and then expressed in various systems, such as bacteria, yeast, baculovirus-infected insect cells, and COS cells (SV40-transformed African Green monkey kidney cells), after insertion into adequate vectors. Comparison of the physicochemical properties of mutated proteins should provide interesting information on the relationship between structure and function.

Also, study of genetic polymorphism at the nucleotide level has led to the discovery of new alleles, has provided information about the mechanism responsible for the occurrence of deleted caseins, and has enabled animals to be genotyped at birth, a major advance for selection.

Finally, expression analysis of native or modified genes using in vitro transcription systems, mammary cell lines, and transgenic animals has greatly improved the knowledge of the functioning of milk protein genes and has offered the potential for genetic modification of milk composition.

In the present review, we attempt to summarize the main features of the structure and function of milk protein mRNA and genes and give an overview of the current practical application of this knowledge.

MILK PROTEIN-ENCODING mRNA

The mRNA content steadily increases in mammary epithelial cells from midpregnancy to lactation. At that stage, those encoding major milk proteins can account for up to 60 to 80% of total mRNA. An earlier study (62) carried out on rat mammary gland explants showed that casein mRNA accumulation, with a steady-state level of about 90,000 mRNA per cell, was due to an increase of the transcription rate (two- to fourfold) and an efficient stabilization of those mRNA (half-life x 17 to 25). Furthermore, differential rates of accumulation were found for each type of mRNA (126). Subsequent studies of milk protein gene expression in the mouse have confirmed the differential stage specificity. For example, β-casein and α-lactalbumin mRNA begin to accumulate at mid (66, 84, 112) and late pregnancy (137), respectively.

The abundance of specific mRNA in the lactating mammary gland greatly facilitated the screening of mammary cDNA libraries. The 60 or so relevant mRNA sequenced to date from a dozen species share the general organization described for mRNA encoding secretory proteins. As illustrated in Figure 1, the mRNA contain 1) a 7-methylguanosine cap, which seems to play a dual role by protecting the 5' untranslated region against degradation enzymes and by facilitating the binding of the ribosomal 40S subunit-Met-tRNAmet initiation factors complex, 2) a coding frame delimited by the initiation and stop codons, and 3) an untranslated 3' region with the recognition signal for polyadenylation located 13 to 20 nucleotides upstream from the poly(A) tail. At least 20% of translatable milk protein mRNA have a very short, if any, poly(A) tail (107). Sizes of mRNA encoding caseins, β-lactoglobulin, α-lactalbumin, and whey acidic protein (WAP) range from 549 nucleotides for rabbit WAP to 1349 nucleotides for rat α-casein, similar to αs1-casein, excluding the poly(A) tail, and the coding frame represents on average 60 to 70% of the mRNA. The human lactoferrin mRNA (124), a major whey protein, comprises 712 codons, of which 19 code for the signal peptide. The coding frame is flanked by 5' and 3' untranslated regions with sizes of 30 to 40 and 181 nucleotides, respectively.

Interspecies comparisons have confirmed the high rate of evolution of milk proteins,
particularly caseins. The lower evolutionary rate of the 5' and 3' untranslated regions compared with the coding frame of casein mRNA (103) suggests an evolutionary constraint for maintaining local structures facilitating translation or involved in the stability of mRNA. The striking similarity and conservation of the multiple phosphorylation sites (104) and signal peptides (47) of α₁-, α₂- and β-caseins have been confirmed at the nucleotide level. As illustrated in Figure 2, most nucleotide substitutions observed in the signal peptide-encoding region did not specify any amino acid replacement. In contrast, numerous substitutions occur at all three positions of many codons in the region encoding the mature protein. This high selection pressure strongly suggests that the efficient transfer of the most abundant milk proteins across the endoplasmic reticulum membrane requires the structural integrity of the transient signal peptide, the conformation of which presumably being optimal (107). A covalent linkage may improve the interaction between the signal peptide and the signal recognition particle, or the signal sequence receptor, or both, as suggested by the constant occurrence of the cysteine residue at position -7.

The aforementioned evolutionary relationship between milk proteins and other proteins was confirmed by cDNA comparison. For example, similarity was 62% between ovine β-lactoglobulin and human placental protein 14 cDNA (78). Moreover, the amino acid se-

---

Figure 1. Schematic representation of the transcriptional units and cognate mRNA encoding the major specific milk proteins. These genes and mRNA are those encoding the four caseins, β-lactoglobulin, α-lactalbumin, and whey acidic protein (WAP). Numbers indicate the extreme sizes of transcriptional units, mRNA and relevant proteins, and of constitutive regions in the approximately 12 species investigated so far. The first and the last exon may only comprise untranslated nucleotide stretches (hatched boxes), depending on the type of gene. The black box refers to the signal peptide-encoding region. The consensus sequences for splicing indicate the 5' splice, branch point, and acceptor sites, respectively. A lower case letter indicates the nucleotide most frequently found at a given position. E = Exon; nt = nucleotide; N = any nucleotide; R and Y = purine and pyrimidine nucleotides, respectively; aa = amino acid.
quence inferred from cDNA sequences (32, 36, 70) clearly indicated that the WAP belongs to the "four-disulfide core" family, which includes neurophysins, wheat germ agglutinin, protease inhibitors, and venom toxins.

STRUCTURE OF MILK PROTEIN-ENCODING GENES

Structure of Tissue-Specific Genes

Tissue-specific eukaryotic genes have the canonical structure described in Figures 1 and 3. The transcription unit contains from 5' to 3', the consensus sequences corresponding to the cap site; the numerous donor, branch point, and acceptor sites required for splicing; and the AATAAAA polyadenylation signal followed by a GT-rich nucleotide stretch, signaling the end of transcription. The proximal 5' flanking region contains the ATA box and possibly other motifs that signal the transcription site to RNA polymerase II and ubiquitous transcription factor TFII D. Gene activation requires the recognition by various effectors of consensus sequences or local conformations involved in induction and modulation of expression as well as stage and tissue specificity. Many of these cis-acting motifs are 5' proximal to the transcription unit, but some can be 5' distal or even occur within the transcription unit or the 3' flanking region. Some important 5' distal elements, called the locus control region, responsible for the tissue-specific accessibility and sequential amplification of clustered genes such as those of the globin family, have been well studied (29, 110). Other distal elements, matrix attachment regions, are involved in the partition of the genome into topologically distinct functional domains [see (115) for recent review] and might also be important for gene expression. Recent studies of constructs stably integrated in the cell genome showed that the chicken lysozyme 5' matrix attachment regions mediated elevated and position-independent gene transcription (139), even with heterologous promoters and cell lines (116).

General Features of Milk Protein-Encoding Genes

The score of milk protein-encoding genes sequenced so far are mosaic genes. The transcriptional units, with sizes ranging from 1.7 to 18.5 kb, comprise between 3 and 18 introns made up of 81 to 5800 bp. Introns often contain repetitive sequences that can represent, for example, 14% of the bovine αs2-casein gene. Most exons are quite short, and their sizes range from 21 to 525 bp. In contrast to the whey protein genes, no codon is split off by intron in the casein genes, and the ribonucleotide stretches encoding multiple phosphorylation sites of caseins (-Ser-Ser-Ser-Glu-Glu-) are generated by splicing.

Localization and identification of regulatory sequences have been carried out in different ways. A computer search for known cis-acting motifs recognized by hormone-receptor complexes or nuclear factors (44) allowed the localization of several potential recognition sites for glucocorticoid and progesterone receptors and for other effectors located essentially in the 5' flanking region. Similarly, some structural motifs shared by several milk protein genes or conserved during the evolution of a given gene have been identified by sequence comparison just upstream from the transcription unit. More recently, expression study of modified genes in mammary cell epithelial lines and transgenic animals gave evidence of the occurrence of important cis-acting motifs in the proximal 5' flanking region and, to a lesser extent, in the 3' flanking region. However, only a few of these ligand-binding elements have been identified using footprinting, interference methylation, gel retardation, and oligonucleotide competition techniques.

Structure of Casein Genes

Since the first reports from 1983 to 1985 (77, 168) on the partial organization of rat casein genes, other nucleotide sequences have been published, including those relevant to the four bovine casein genes, αs1- (81), αs2- (54), β- (13), and κ-casein (4) (Figure 4). The striking similarity between two regions of bovine αs2-casein (18) and between nucleotide stretches of the relevant mRNA (138) is clearly apparent on the gene where they correspond to both groups of 5 exons, VII through XI and XII through XVI, which obviously arose from a duplication. Casein genes seem to be structurally quite different because sizes of the transcription units range from 8.5 to 18.5 kb and the number of introns ranges from 4 to 18. However, the
hypothesis of a common ancestor for the α₁s-, α₂-, and β-casein genes (47) was substantiated by finding common sequence motifs in the proximal 5' flanking regions (Figure 3) and the similar organization pattern of the first four exons (Figure 4) first observed by Rosen's group (77). In particular, the second exon comprises the remaining part of the 5' untranslated region and the coding frame for the signal peptide and the first two amino acids of the mature polypeptide chain. Therefore, Jones et al. (77) proposed that the present casein genes derived from a primitive gene made up of a few exons, one corresponding to the 5' untranslated region, and others encoding the signal peptide, a simple phosphorylation site, and a hydrophobic peptide. This gene might have grown through intragenic duplication and then undergone intergenic duplications with divergent evolution of the new genes. This model

Figure 2. Nucleotide sequences of casein-encoding cDNA, in the region specifying the signal peptide. The previously published figure (108) has been updated with sequences of cDNA encoding goat (87), pig (2), mouse (59), and kangaroo (26) α₁s-casein (CN), goat (14), and pig (3) α₂-CN, mouse γ-CN (T. Sasaki, 1992, unpublished; European Molecular Biology Laboratory (EMBL) bank: D10215, Germany), goat (M. A. Persuy, 1992, unpublished results), pig (1), human (95), and kangaroo (27) β-CN, goat (A. Coll et al., 1991, unpublished; EMBL bank: X60763, Germany), pig (90), guinea pig (63), human (R. S. Menon et al., 1991, unpublished; EMBL bank: M73628, Germany; S. Bergstroem et al., 1992, unpublished; EMBL bank: X64417, Germany), and rabbit (E. Deviny, 1992, personal communication) α-CN. Bracketed letters in the left margin refer to the original name of the casein. G.PIG = guinea pig; KANG. = kangaroo; AA = amino acid; CONS. refers to the consensus sequence of the signal peptides of the three bovine α₁s-caseins: α₂-caseins, and β-CN. Italicized numbers refer to the codons of the signal peptide (-) and the mature casein (+). Dashes represent nucleotides identical to those of each bovine cDNA taken as reference. Lower case letters refer to substituted nucleotides that did not specify any change of codons. The one-letter symbols for amino acids are written in italics. Boldfaced amino acids indicate the most frequent amino acids at a given position.

### SIGNAL PEPTIDE MATURE PROTEIN

<table>
<thead>
<tr>
<th>α₁s-CN</th>
<th>COM</th>
<th>-15</th>
<th>-10</th>
<th>-5</th>
<th>-1</th>
<th>+1</th>
<th>+5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHEEP</td>
<td>ATG ATT ACT ACC ACC ACT ACC ACC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GOAT</td>
<td>ATG ATT ACT ACC ACC ACT ACC ACC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAT</td>
<td>-C- C- T- T- T- T- T- T- T-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOUSE</td>
<td>-C- C- T- T- T- T- T- T- T-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HABBIT</td>
<td>-C- C- T- T- T- T- T- T- T-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RABBIT</td>
<td>-C- C- T- T- T- T- T- T- T-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AA</td>
<td>M K L I L F P T S C C L V A A T P A S L P V A N L M K L V F S Y K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>α₂-CN</th>
<th>COM</th>
<th>-15</th>
<th>-10</th>
<th>-5</th>
<th>-1</th>
<th>+1</th>
<th>+5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHEEP</td>
<td>ATG ATT ACT ACC ACC ACT ACC ACC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GOAT</td>
<td>ATG ATT ACT ACC ACC ACT ACC ACC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAT</td>
<td>-C- C- T- T- T- T- T- T- T-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOUSE</td>
<td>-C- C- T- T- T- T- T- T- T-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HABBIT</td>
<td>-C- C- T- T- T- T- T- T- T-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RABBIT</td>
<td>-C- C- T- T- T- T- T- T- T-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AA</td>
<td>M K L I L F P T S C C L V A A T P A S L P V A N L M K L V F S Y K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>β-CN</th>
<th>COM</th>
<th>-15</th>
<th>-10</th>
<th>-5</th>
<th>-1</th>
<th>+1</th>
<th>+5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHEEP</td>
<td>ATG ATT ACT ACC ACC ACT ACC ACC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GOAT</td>
<td>ATG ATT ACT ACC ACC ACT ACC ACC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAT</td>
<td>-C- C- T- T- T- T- T- T- T-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOUSE</td>
<td>-C- C- T- T- T- T- T- T- T-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HABBIT</td>
<td>-C- C- T- T- T- T- T- T- T-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RABBIT</td>
<td>-C- C- T- T- T- T- T- T- T-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RABBIT</td>
<td>-C- C- T- T- T- T- T- T- T-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AA</td>
<td>M K L I L F P T S C C L V A A T P A S L P V A N L M K L V F S Y K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>κ-CN</th>
<th>COM</th>
<th>-15</th>
<th>-10</th>
<th>-5</th>
<th>-1</th>
<th>+1</th>
<th>+5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHEEP</td>
<td>ATG ATT ACT ACC ACC ACT ACC ACC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GOAT</td>
<td>ATG ATT ACT ACC ACC ACT ACC ACC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAT</td>
<td>-C- C- T- T- T- T- T- T- T-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOUSE</td>
<td>-C- C- T- T- T- T- T- T- T-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HABBIT</td>
<td>-C- C- T- T- T- T- T- T- T-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RABBIT</td>
<td>-C- C- T- T- T- T- T- T- T-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AA</td>
<td>M K L I L F P T S C C L V A A T P A S L P V A N L M K L V F S Y K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
was recently refined by Groenen et al. (54), who gave evidence of a closer evolutionary relationship between the \( \alpha_\text{S2} \) and \( \beta \)-casein genes on the basis of nucleotide sequence and exon size similarities.

In contrast, the \( \kappa \)-casein gene does not share any common pattern with other casein genes. It was postulated (76) to be evolutionarily related to the fibrinogen gene family, which encodes proteins that are functionally similar to \( \kappa \)-casein in that their limited proteolytic cleavage triggers the clotting of blood. If so, these genes have much diverged because the most striking homology observed between the cDNA involves a nucleotide stretch corresponding to the 5' end of \( \kappa \)-casein exon IV and the 3' end of \( \gamma \)-fibrinogen exon II (4).

Interspecies comparison of casein genes showed, as expected, a greater divergence between homologous introns, which show greater differences in nucleotide sequence, and often in size, than do exons. This divergence is mainly due to the frequent occurrence of repetitive sequences of different types, which also occur in the flanking regions.

Many repetitive DNA elements belong to the A family of artiodactyl retroposons. For more details, the reader is referred to the original papers on casein genes already mentioned. Nevertheless, the overall organization of each type of gene has been conserved in present mammals as illustrated by the structural comparison of \( \beta \)-casein genes from five species (Figure 4).

**Chromosomal Location of Casein Genes and Organization of the Clustered Bovine Casein Genes**

As previously mentioned, earlier genetic studies showed a tight linkage between the four bovine casein genes with the possible relative order: \( \alpha_\text{S1} \)-, \( \beta \)-, and \( \kappa \)-casein loci (58). Similarly, Mendelian segregation analysis of casein DNA fragments by restriction fragment length polymorphism (RFLP) showed the same linkage in the ovine species (37, 89).

This linkage was confirmed by probing panels of somatic cell hybrids, which gave evidence of casein synteny in mouse (61) and rabbit (31). Casein loci have been assigned to chromosome 5 in mouse (49), 12 in rabbit (50), 4 in human (102) and sheep (69), and 6 in cattle (145). Chromosomes 4 and 6 are difficult to discriminate in domestic ruminants, hence, the discrepancy between the chromosomal assignments of casein loci in cattle and sheep (Figure 5). Recently, restriction mapping of bovine inserts of yeast artificial chromosome clones gave the order \( \alpha_\text{S1} \)-\( \beta \)-\( \alpha_\text{S2} \)-\( \kappa \) within the 200-kb casein locus cluster (45, 145), which might be dependent on a locus control region. Casein genes appear to occur as a single copy per haploid chromosome set because no related sequence has ever been reported.

**Structure of Whey Protein Genes**

Sequences of genes encoding whey proteins were first reported in 1984 for rat \( \alpha \)-lactalbumin (121) and for mouse and rat WAP (22); at present, a dozen sequences are known. Structures of \( \alpha \)-lactalbumin and WAP genes are very simple: a 2-kb transcriptional unit divided into 4 exons (Figure 6). The \( \beta \)-lactoglobulin gene has a 4.7-kb transcriptional unit comprising 6 introns. The partially sequenced murine lactotransferrin gene comprises at least 16 exons (134).

As illustrated in Figure 6, the organization of whey protein genes has been conserved during evolution, and the striking similarity of organization between genes encoding proteins thought to be evolutionarily related strongly supports the proposed common origin.

A consensus sequence coined “milk box” (64, 82) might be shared by proximal 5' flanking regions of \( \alpha \)-lactalbumin and “calcium-sensitive” casein genes (Figure 3). Part of this nucleotide stretch might also be common to \( \beta \)-lactoglobulin and WAP genes (152).

**Chromosomal Location of Whey Protein Genes and Occurrence of Pseudogenes**

The \( \alpha \)-lactalbumin locus is localized on chromosome 5, 3, and 12 in the bovine (145), ovine (74), and human (34) species, respectively, and the murine WAP locus might occur on chromosome 11. The \( \beta \)-lactoglobulin locus was assigned to chromosome 3 in sheep and 11 in goat and cow (68) (Figure 5).

One pseudogene has been reported for \( \beta \)-lactoglobulin in the ovine and caprine species (A. J. Clark, 1991, personal communication, and A. Sanchez, 1992, personal communication), and complex RFLP genomic patterns indicate the occurrence of at least five \( \alpha \)-
lactalbumin-related sequences in domestic ruminants (136). Analysis of two of them showed 80% similarity with the α-lactalbumin transcriptional unit downstream from exon II (136, 150, 152). Evolutionarily related lysozyme and α-lactalbumin genes underwent several duplications, which occurred most likely for the α-lactalbumin gene before the divergence of goat, sheep, and cattle.

**FUNCTION OF MILK PROTEIN-ENCODING GENES**

**Expression of Casein Genes**

Expression of endogenous casein genes was first investigated in mammary gland explants and primary cultures. Later, mammary epithelial cell lines such as murine COMMA-1D (33) and HC11 (8) provided a model system best adapted for the study of regulated expression of both the endogenous β-casein gene and transfected native or shortened β-casein genes from various species, as well as β-casein promoter-driven hybrid genes, as further discussed. The synergistic action of lactogenic hormones (8, 35, 39, 40, 52, 119, 127, 128, 166, 167) and extracellular matrix (43, 127 and citations therein) on β-casein gene expression has been well established. In cell cultures, glucocorticoid in the presence of insulin is required for rapid and strong induction of the β-casein promoter by prolactin. The former hormone may act indirectly (40) 1) in regulating glucocorticoid-sensitive genes producing effectors trans-acting on the β-casein gene (128, 167), or stabilizing β-casein transcripts (119), or both and 2) in disrupting nucleosomes.

Figure 3. Some of the important structural motifs identified in the 5' flanking regions of the major specific milk protein-encoding genes from various species. The ATA box and possibly other signals indicate the start of transcription to the complex polymerase II-transcription factor IID (TF II D). Other boxes refer to structural motifs shared by the genes encoding the three "calcium-sensitive" αs1-1, αs2-2, and β-caseins (CN). The box with rounded edges indicates that the motif is also common to the genes encoding β-lactoglobulin (β-LG), α-lactalbumin (α-LA), and whey acidic protein (WAP). This motif is recognized by a specific mammary nuclear factor (158). The "milk box" consensus sequence, common to the genes encoding α-LA, WAP, and "calcium-sensitive" caseins, partially overlap two motives. Lower case letters indicate nucleotides occurring much less frequently at a given position. R and Y refer to purine and pyrimidine nucleotides, respectively. Numbers indicate the position upstream from the transcriptional unit (TU).

(123) at the β-casein locus, which contains consensus sequences for glucocorticoid-receptor complexes (39).

Protein-binding sites (128), some spanning the motifs shared with αs1- and αs2-casein genes, have been identified in the 5' flanking region of the β-casein gene. A strong and a weak site were recognized by a mammary gland-specific factor, but other complexes were down-regulated during induction, suggesting a transcriptional derepression (128). Furthermore, mutation of the mammary gland-specific factor-binding site of β-casein gene abolished lactogenic hormone induction of this gene in HC11 cells. The mammary gland-specific factor protein was developmentally and environmentally regulated, probably through its phosphorylation state (128). Comparative expression analysis between native and altered rat (39), murine (40, 166), and bovine (127) β-casein genes transfected into mammary cell lines suggested that most, if not all, prevalent cis regulatory elements mediating the hormonal and extracellular matrix effects might occur within 2.6 kb of the 5' flanking region. However, some discrepancies occurred with the very low expression in transgenic mice for a rat β-casein transgene comprising a 3.5-kb 5' flanking region and a 3-kb 3' flanking region (85, 86) and in transgenic rabbits for a chimeric interleukin-2 gene driven by a 2-kb β-casein 5' flanking region (20).

Figure 4. The organization of the four bovine casein (CN)-encoding genes and the similarity of organization between the genes encoding β-CN from five species. The drawings are based on the nucleotide sequences reported for the bovine genes encoding β- (13), αs1- (81), αs2- (54), and κ-CN (4). Only the exons, which are represented as high boxes, are not at scale. Their base pair sizes are indicated below each drawing. Two numbers are indicated whenever an exon comprises both a noncoding and a coding nucleotide stretch (black boxes). Underlined exons are those that can be skipped during processing of pre-mRNA in the case of either a particular bovine or caprine allele (bracketed letter) or a given species (bracketed name). Exon VI of αs2-CN pre-mRNA is partly skipped in sheep. Distribution and respective base pair sizes of exons (splicing of boldfaced exons IV and V generates a nucleotide stretch encoding the multiple phosphorylation site) and introns (italicized numbers) of rat (77), mouse (165), cow (13), sheep (Provot, 1989, unpublished results), and rabbit (143) β-CN genes are indicated in the large box. TU = Transcriptional unit.

Contrast, expression was high, stage-specific, and mammary tissue-specific in transgenic mice carrying a caprine \( \alpha \)-casein transgene with a 3-kb 5' flanking region and a 6-kb 3' flanking region (112). Recently, secretion of the human cystic fibrosis transmembrane conductance regulator, associated with milk fat globule membrane, was achieved by expressing a goat \( \beta \)-casein transgene substituted between exons 2 and 7 with a cystic fibrosis transmembrane conductance regulator cDNA (38). Despite some successes, further experiments are clearly needed to identify important cis-acting elements essential for \( \beta \)-casein gene expression.

Data on regulatory elements controlling other casein genes are scarcer, mainly because complete sequencing of these longest genes and establishment of better adapted mammary epithelial cell lines have just been achieved. Recently, Groenen et al. (54) reported the strong binding of a mammary gland-specific nuclear factor and of octamer-binding factor 1 to the conserved sequences at positions −90 and −50 in the bovine \( \alpha_2 \)-casein gene, respectively. The octamer-binding factor also bound to 3 weak sites at positions −210, −260, and −480. Furthermore, a hybrid bovine \( \alpha_2 \)-casein-human urokinase gene, comprising a 21-kb 5' flanking region and a 2-kb 3' flanking region of \( \alpha_1 \)-casein gene as well as exonic sequences relevant to the 5' and 3' untranslated regions and the C-terminal half of the casein, promoted the specific secretion of 1 to 2 mg of urokinase/ml milk in transgenic mice (98).

**Expression of Whey Protein Genes**

Expression of modified and unmodified \( \beta \)-lactoglobulin, \( \alpha \)-lactalbumin, and WAP genes has been studied in transfected mammary cell lines and transgenic animals.

**WAP Gene.** The WAP is abundantly expressed in mammary epithelial cells, and the relevant mRNA accounts for 10 to 15% of mammary polyadenylated RNA in lactating rat, mouse, and rabbit (70, 71, 72). Low concentrations of WAP mRNA are detectable in mammary tissue of virgin and early pregnant animals (67, 72, 118). The relative several 1000-fold increases occurring at midlactation result from the proliferation and differentiation of epithelial cells (about a 10-fold change in the proportion of rat alveolar tissue (72)) and the steady-state accumulation of WAP mRNA at late pregnancy (up to 50-fold around d 16 in mouse) and, to a lesser extent, after parturition. Study of the hormonal regulation and important cis-acting elements of the WAP gene has been made difficult by its poor expression in mammary cell cultures unless proper three-dimensional alveoli-like structure and interacting matrix are maintained or mimicked (23, 41, 118, 130). Induction and maintenance of endogenous WAP gene expression depend on the synergistic action of lactogenic hormones (prolactin, glucocorticoid, and insulin) and cell-cell and cell-extracellular matrix interactions. When mammary epithelial cells lack the correct spatial structure and cellular environment, secretion of at least one inhibitor might be responsible for the specific posttranscriptional suppression of WAP expression (23). According to the analysis of hybrid genes WAP-myc (129) and WAP-CAT (chloramphenicol acetyltransferase) (41) in cell cultures, several (but not all) regulatory motifs of the WAP gene, including elements responsive to lactogenic hormones, are scattered along the 2.5-kb 5' flanking region, because induction and expression of the constructs correlated with the length of the 5' flanking region (41).
These features were confirmed by analysis of transgenic animals carrying native and modified WAP genes. The WAP hybrid genes comprising about 2.5 kb of the WAP 5' flanking region fused to human oncogenes (6, 129), human tissue plasminogen activator (42, 53, 117, 118), or human growth hormone (60, 122) genes were predominantly expressed in the mammary gland of transgenic mice and goat (42), but, overall, at variable and lower levels than the endogenous WAP gene and more precociously. A hybrid CAT gene driven by .5 kb of the WAP 5' flanking region did not express significant activity (30). In contrast, a hybrid α1-antitrypsin gene driven by 17.6 kb of the rabbit WAP 5' flanking region expressed 6 mg of human α1-antitrypsin/ml of milk (11). Mouse (19) and rat (9) WAP transgenes comprising the 2.6- and 95-kb 5' flanking sequence and the 1.6- and 1.4-kb 3' flanking sequence, respectively, were also expressed earlier, at midpregnancy, and at levels between 3 and 54% and between 1 and 95% of the endogenous WAP gene during lactation of transgenic mice. Further study of the rat WAP transgene (30) with shortened 3' ends showed that the transgene with only 70 bp of the 3' flanking region was expressed at uniformly high levels and that deletion of the 3' end of the transcriptional unit reduced WAP mRNA

![Figure 6](image-url)

Figure 6. Organization of the specific whey protein-encoding genes from various species and of some evolutionary related genes. The coding frame is in black. Sizes of exons (standard and boldfaced numbers refer to untranslated (hatched boxes) and coding regions, respectively) and introns (italicized numbers) are expressed as base pairs. The first exon always comprises the 5' untranslated region followed by a nucleotide stretch encoding the signal peptide and the beginning of the mature protein. The genes encoding ovine (65) and bovine (L. J. Alexander et al., 1992, unpublished; EMBL bank: XI4710, Germany) β-lactoglobulin (β-LG) and the evolutionarily related human placental protein 14 (147) and murine urinary protein share a very similar organization. The 5' untranslated region of the gene encoding the retinol-binding protein is divided into two exons, and the last two exons are equivalent to the pairs IV plus V and VI plus VII of the β-LG gene. The exon sizes given for the genes encoding murine urinary proteins (MUP) and retinol-binding protein (RBP) were taken from Ali and Clark (5). The genes encoding rat (121), mouse (149), guinea pig (82), bovine (151), goat (152), and human (64) α-lactalbumin (α-LA), and chicken (79) and human (114) lysozyme are structurally quite similar. The organization of the genes encoding rat (22), mouse (22), and rabbit (144) whey acidic protein (WAP) is also well conserved.
20-fold during lactation. The murine WAP gene was also successfully expressed in transgenic swine at 0.5 to 1.5 g/L of milk (133, 156). An impaired mammary development that was probably due to the precocious expression of WAP was observed in some lines of transgenic mice (21) and pigs (132), suggesting the involvement of WAP in mammary cell development and differentiation. Recent experiments with a murine WAP transgene linked to a matrix attachment region indicate that the ratio of expressing lines and the developmental regulation of the WAP transgene in transgenic mice can be improved (155).

Gel retardation and nuclease protection assays allowed the identification of 4 mammary nuclear protein-binding sites in the -175 to -88 region, upstream from the murine WAP transcriptional unit (91), which shares common motifs with proximate 5' flanking regions of other milk protein genes. Accordingly, in vitro transcription of the mouse WAP promoter and competition assays with fragments from this region showed that the nucleotide stretch -175 to -25 could stimulate transcription (93).

β-Lactoglobulin Gene. In the ovine species, the β-lactoglobulin gene is already expressed at midpregnancy, and the level of the relevant mRNA, which is much higher than those of casein mRNA during gestation, increases slowly until parturition and more rapidly thereafter (48, 67). At d 20 of lactation, β-lactoglobulin mRNA accounts for about 5% of total mammary poly(A) RNA, a 20-fold increase from levels at midpregnancy (48). In cultured ovine mammary explants, expression of the β-lactoglobulin gene appears to be less dependent on lactogenic hormones than the casein genes are, and glucocorticoid and insulin have only a slightly synergistic effect on prolactin induction (120). The β-lactoglobulin gene seems to behave similarly in the porcine species (131). In transgenic mice, patterns of expression of the β-lactoglobulin transgene and the endogenous β-casein gene appear to be similar (67). A small, gradual accumulation of the relevant mRNA occurs until midpregnancy, followed by a rapid increase: 5 and 65 to 80% of the midlactation mRNA level at midpregnancy and at parturition, respectively. In marsupials, concentration of β-lactoglobulin mRNA increases in late lactation, and induction may depend on prolactin alone (26). The β-lactoglobulin promoter is sensitive to prolactin-induced signals, as demonstrated by prolactin-induction of CAT activity in Chinese hamster ovary cells cotransfected with a prolactin-receptor expression plasmid and a hybrid gene comprising the 4-kb β-lactoglobulin 5' flanking region fused to the CAT reporter gene (88).

The native ovine β-lactoglobulin gene comprising a 4.3-kb 5' flanking sequence and a 7.3- or 1.6-kb 3' flanking sequence was efficiently and specifically expressed in the mammary gland of transgenic mice (135). Study of 5' shortened constructs (66) indicated that 8 kb upstream from the transcription unit was sufficient for high, tissue-specific expression. The region -406 to -149 appears to be essential, but not necessarily sufficient, for high, tissue-specific expression (162). Multiple-binding sites recognized by various nuclear effectors were identified in this region by in vitro binding assays (158) including at least five binding sites for nuclear factor I and three sites recognized by a specific mammary gland factor termed milk protein-binding factor. The recognition site for milk protein-binding factor might be a 13-bp palindromic nucleotide stretch, GATTCCNGGAACC, that is structurally similar to structural motifs shared by proximal 5' flanking regions of the genes encoding other main milk proteins (149).

Study of various β-lactoglobulin minigenes and derived hybrid genes indicated that intronic sequences are involved in the efficiency of expression, probably through interaction with the upstream 5' flanking sequences (161). A hybrid human α1-antitrypsin minigene driven by a 4.3-kb β-lactoglobulin promoter was highly expressed in the lactating mammary glands of mice (7) and ewe (164); yields were up to 7 and 35 mg of active human α1-antitrypsin/ml of milk, respectively. In contrast, expression of another hybrid gene, made from a human antithemophilic factor IX cDNA inserted into the 5' untranslated region of the ovine β-lactoglobulin gene, was less successful; yield was 25 ng of factor IX/ml of milk in a transgenic ewe (25).

α-Lactalbumin Gene. In mice, induction of α-lactalbumin gene expression requires the synergistic action of insulin and prolactin (or placental lactogen) and is maximal in the presence of $<3 \times 10^8 M$ glucocorticoid (109).
contrast, high concentration of this hormone inhibits \( \alpha \)-lactalbumin gene induction at both RNA and protein levels (46). Thyroid hormone and prostaglandins can reverse this inhibitory effect (10, 142). \( \alpha \)-Lactalbumin gene expression is inhibited by progesterone in mammals, and cyclic AMP might also be a negative regulator (111). In marsupials, \( \alpha \)-lactalbumin gene expression depends only on prolactin and is not inhibited by progesterone (28).

Efficient mammary tissue-specific expression of bovine (153) and caprine (137) \( \alpha \)-lactalbumin genes in transgenic mice was obtained; yields were up to 4 and 3.7 mg of exogenous \( \alpha \)-lactalbumin/ml of milk. The relevant transgenes comprised .75- and 8.5-kb 5' flanking regions and .34- and 9.5-kb 3' flanking regions, respectively. Analysis of \( \alpha \)-lactalbumin mRNA and protein in transgenic mice carrying a guinea pig \( \alpha \)-lactalbumin transgene (96), with about 1.2-kb 5' flanking region and .4-kb 3' flanking region, also showed an efficient mammary expression of the transgene. High concentrations of exogenous \( \alpha \)-lactalbumin and endogenous \( \beta \)-casein mRNA in sebaceous glands were also reported (96) but not confirmed by other authors (112, 137, 149). Expression analysis of the aforementioned bovine transgene shortened at the 5' end (137) or substituted with a trophoblast interferon cDNA in the coding frame (140) indicated that .4-kb 5' flanking region and .34-kb 3' flanking region might be sufficient for mammary targeting and correct developmental expression but not for high expression. Accordingly, several sites binding nuclear mammary effectors in vitro, −1062 to −1040, −1004 to −970, and −125 to −85, were identified by footprinting analysis (92). The latter binding site might be recognized by nuclear factor 1 (92).

Occurrence of Allelic and Nonsallelic Deleted Caseins and Their Evolutionary Importance

The translatable mature mRNA in the cytoplasm derive from primary transcripts through complex processing: addition of a methylated nucleotide (capping) to the 5'end of nascent pre-mRNA; polyadenylation of the shortened 3'end of the primary transcript; and fixation of ribonucleic proteins and methylation of internal adenosine residues, excision of introns and splicing of exons by the spliceo-
some machinery, and ultimately partial deadenylation.

Splicing of the 19 exons of \( \alpha_{s1} \)- and 18 exons of \( \alpha_{s2} \)-casein pre-mRNA must be quite complex, and, in retrospect, the finding of deleted \( \alpha_{s1} \)- and \( \alpha_{s2} \)-caseins is not surprising. Bovine \( \alpha_{s1} \)-casein A and \( \alpha_{s2} \)-casein D and caprine \( \alpha_{s1} \)-casein D and F lack an internal stretch of 13 (57), 9 (56), 11, and 37 (17) amino acid residues, respectively, each one, except F, corresponding to a single exon. Ovine milk always contains two types of \( \alpha_{s2} \)-casein differing by an internal deletion of 9 amino acid residues, and they are the translation products of four types of mRNA (12) arising from partial skipping of 2 exons relevant to the 5' untranslated region and the coding frame. Similarly, study of caprine \( \alpha_{s1} \)-casein F transcripts showed the occurrence of 10 or so different types of mRNA arising from exon skippings (87). A single nucleotide deletion in exon IX and both insertions of 11 and 3 nucleotide stretches in downstream intron might be responsible for skipping of exons IX, X, and XI. In particular, a presumptive stem loop formation between the 11 nucleotide stretch and the intron 5' splice site might impair the recognition of that site by small nuclear ribonucleoprotein particle U1 (U1 snRNP).

Sequence study of genomic DNA from two cows with genotype \( \alpha_{s2} \)-casein DD showed the occurrence of the nucleotide stretch encoding the missing peptide (15). The mutation likely responsible for exon VIII skipping in pre-mRNA D might be the substitution T/G affecting the last nucleotide of that exon compared with other alleles, i.e., the consensus sequence R/GTRRTG of the 5' splice site (Figure 1). The deleted cDNA clone, encoding an \( \alpha_{s1} \)-casein of type A (97) and isolated from a mammary cDNA library from a homozygous \( \alpha_{s1} \)-casein B cow, indicated that the processing mechanism is not fully accurate. Moreover, the deleted clone provided indirect evidence that the rare allele A differs from its counterparts by at least one mutation inducing skipping of one exon.

Interspecies sequence comparisons of known homologous caseins and of the relevant cDNA often showed marked differences in size, and many deleted regions are obviously encoded by a distinct exon. Exon skipping
might be one of the mechanisms involved in the rapid evolution of caseins. Any apparently minor mutation, including a single nucleotide substitution, could induce a major alteration of the polypeptide chain, provided that it affects any local nucleotide sequence or conformation required for correct processing of the pre-mRNA. Consequently, structural differences between casein genes from various species might be less important than expected from protein and cDNA comparisons. Recent studies of the human \( \beta \)-casein gene (95, 101) showed that the missing peptide, compared with other \( \beta \)-caseins, was actually encoded by a cryptic exon III. Subsequently, in vitro transcription of a mutated human \( \beta \)-casein gene demonstrated that the four purines interrupting the polypyrimeidic end of human intron 2 were responsible for the human differential processing (100). One can predict that exon skipping is probably responsible, at least in part, for the shorter mRNA, such as those encoding mouse and rat \( \alpha_\text{S}_2 \)-casein and ovine \( \alpha_\text{S}_1 \)-casein.

**PRACTICAL APPLICATIONS**

The present knowledge of the structure and function of the major milk protein genes has already been applied for selecting animals with interesting dairy genotypes and for creating transgenic animals used as models for studying mammary carcinoma, for example, or as producers of exogenous proteins of high value, such as pharmaceuticals.

Milk protein-encoding cDNA have been used for probing genomic DNA of domestic dairy breeds, either to identify, by RFLP, the animals with a known interesting genotype or to search for new alleles. The DNA phenotyping analysis is already in use for selecting cattle carrying the \( k \)-casein B allele, which is associated with a faster coagulation rate and a firmer curd, and male goats carrying \( \alpha_\text{S}_1 \)-casein alleles A, B, or C, which are associated with higher yields of \( \alpha_\text{S}_1 \) and whole caseins.

The tedious and time-consuming RFLP technique is now replaced by simpler and quicker automated techniques whenever the nucleotide sequence surrounding the mutation of interest is known: analysis of DNA amplified in vitro by polymerase chain reaction; i.e. (99), using restriction enzymes or allelic specific oligonucleotide probes or allelic specific primers. The ligase chain reaction technique (159) may also be applied for this purpose.

The overall localization of most essential regulatory elements responsible for induction, modulation, and stage and tissue specificity of expression of milk protein genes, together with the improvement of transgenesis techniques, has opened the door to genetic modification of milk composition with two main purposes: 1) production of a milk better adapted to nutritional or technological needs, or both; [i.e., production of "maternized milk" (24, 75, 105, 163)]; and 2) utilization of the remarkable capacity of synthesis and secretion of the mammary gland for making large amounts of valuable exogenous proteins (83, 160) using the aforementioned regulatory elements for targeting a high expression of the chimeric gene. Successful transfer and expression of \( \beta \)-lactoglobulin, \( \alpha \)-lactalbumin, WAP, \( \beta \)- and \( \alpha_\text{S}_1 \)-casein genes, or derived constructs, mainly in transgenic mice and, to a lesser extent, in rat, rabbit, swine, and domestic ruminants, have been reported in many papers (38, 112). The reported yields of exogenous functional human urokinase, growth hormone, \( \alpha_1 \)-antitrypsin, and tissue plasminogen activator reached 2, 10, 37, and 3 mg/ml of milk in mouse (98), rabbit (E. Devinoy, 1992, personal communication), sheep (164), and goat (42), respectively. The production of recombinant proteins into the milk of transgenic animals seems to be a valuable alternative to production in cell culture.

**CONCLUSIONS**

Since the advent of molecular biology, remarkable progress has been achieved in the structural and functional analysis of milk protein-encoding mRNA and genes and possibly the inferred amino acid sequences. In particular, the complete nucleotide sequences of the genes encoding the four bovine caseins, \( \alpha \)-lactalbumin, \( \beta \)-lactoglobulin, and the rodent and rabbit WAP are now available. This knowledge has allowed prediction of which fragments of these genes might be difficult to detect with a cDNA probe or should not be used as probes for RFLP studies. Moreover, the synthesis of appropriate allelic-specific oligonucleotide probes and primers and the
development of simple genetic screening techniques have been made possible. These techniques have been thoroughly tried and tested for identifying the few alleles known to be associated with interesting dairy traits. Their usefulness in phenotyping analysis of DNA from neonates and from sperm is obvious, and their standardization depends on the identification of new alleles of economic interest. The prospects are bright with the present mapping of the genome, which should lead to the discovery of genes controlling quantitative lactation traits.

These cDNA and genes can be modified, inserted in various vectors, and expressed in miscellaneous systems to produce mutated milk proteins for analyzing structure-function relationship and to elaborate new polypeptides with novel technological properties. This developing field has already gained some success with recombinant lysozyme (94) and α-lactalbumin (141, 148) secreted from yeast at a concentration of several milligrams per liter of culture and the production of κ-casein (80) and α-lactalbumin (157) by Escherichia coli. However, the production of modified milk proteins by large animals is not for immediate use. The technology exists today for using homologous recombination to modify endogenous genes of embryonic stem cells available in the mouse species and to obtain chimeric animals that can be inbred to produce animals homozygous for the modified gene. But presently only a few specialized laboratories fully control the entire process in mice, and the availability of true embryonic stem cells in domestic species must be confirmed.

The specific expression of chimeric genes driven by milk protein promoters in the mammary gland of transgenic animals appears to be promising for production of large amounts of scarce and costly foreign proteins. Some successful experiments on expression level have been published, but "we cannot see the wood for the trees"; overall, the milk yields were too low for commercial production. Moreover, many failures were obviously not reported. Major progress in the field requires a better knowledge of regulation of milk protein genes. The identification of essential elements involved in gene expression, pre-mRNA processing, and mRNA stability is a prerequisite for making efficient constructs without time-consuming and costly adjustment. Other technical impediments, such as the copy-independent and site-dependent expression of the transgene, might be obviated by introduction of an appropriate locus control region or matrix attachment region element in the constructs. In any case, the mouse is still the experimental animal of choice for testing the feasibility of any project.

ACKNOWLEDGMENTS

We are particularly grateful to M.A.M. Groenen, R.J.M. Dijkhof, A.J.M. Verstege, and J. J. Van der Poel, who have provided the manuscript on the bovine α₂-casein gene prior to publication, and to E. Devinoy and A. Sanchez for communicating unpublished results. The bibliography is not exhaustive because of restrictions on the number of references, and we apologize to authors whose pertinent work was not included for those reasons.

NOTE ADDED IN PROOF

The casein gene group has been recently mapped to cattle chromosome 4 (band q32), numbered according to the ISCNDA 89 standard (International System for Cytogenetic Nomenclature of Domestic Animals) (H. Hayes, E. Petit, C. Bouniol, and P. Popescu. 1993. Localization of the α₂-casein gene (CASAS2) to the homoeologous cattle, sheep and goat chromosomes 4 by in situ hybridization. Cytogenet. Cell Genet. (in press).)

REFERENCES

6 Andres, A. C., C. A. Schonesberger, B. Groner, L.
Symposium: Genetic Perspectives on Milk Proteins


56 Grosclaude, F., P. Joudrier, and M. F. Mahé. 1978. Polymorphisme de la caséine $\alpha_2$ bovine: étroite liaison du locus $\alpha_2$-Cas avec les loci $\alpha_1$-Cas, $\beta$-Cas et $\kappa$-Cas: mise en évidence d'une délétion dans le variant $\alpha_2$-Cas-D. Ann. Génét. Sér. Anim. 10:313.


Symposium: Genetic Perspectives on Milk Proteins

3095


SYMPOSIUM: GENETIC PERSPECTIVES ON MILK PROTEINS


lactalbumin obtained by limited proteolysis of a fusion protein expressed at high levels in *Escherichia coli*. J. Biol. Chem. 264:21116.


