The nucleotide sequences corresponding to bovine α_{S2}- and β-casein mRNAs have been determined by cDNA analysis. Both sequences appear to be complete at their 5' ends. The nucleotide sequence of α_{S2}-casein, when compared with the corresponding canine A sequence, helps to define the boundaries of a large amino acid repeat (~80 residues) whereas comparisons with the nucleotide sequences of rat α- and mouse κ-casein mRNAs also reveal extensive sequence similarities. An alignment of these four sequences shows that the divergence of their translated regions has been characterized by the duplication and deletion of discrete segments of sequence that probably correspond to exons. A high degree of nucleotide substitution is also found when the four sequences are compared, except for well-conserved leader-peptide and phosphorylation-site sequences and, to a lesser extent, the 5'-untranslated regions. Similar comparison of the bovine and rat β-caseins shows that their divergence has involved a high rate of nucleotide substitution but that no major insertions or deletions of sequence have occurred. The several splice sites that have been defined in the rat β-casein gene are likely to have been conserved in the bovine. The contrasting evolutionary histories of the α- and β-casein coding sequences correlate with the distinctive functions of these proteins in the casein micelle system in milk.

Introduction

The caseins comprise the major protein fraction of the milk from most species. Their function is to transport calcium phosphate in milk—and hence to provide the suckling infant with a source of calcium and phosphorus for bone formation—as well as to contribute to the requirement for amino acids.

Bovine milk contains four caseins whose genes are linked in the order α_{S1*}, α_{S2*}, β-, and κ-casein (Grosclaude et al. 1973, 1979), and these are expressed in a coordinate manner during lactation. The caseins have been extensively studied at the protein
level, and their amino acid sequences are known (Mercier et al. 1971, 1973; Ribadeau-Dumas et al. 1972; Brignon et al. 1977).

The two α-caseins and β-casein are described as "calcium-sensitive" because they precipitate in the presence of low concentrations of this cation (Waugh 1971). In milk, the caseins occur as very large aggregates termed casein micelles, in which the calcium-sensitive caseins are maintained in stable suspension as a result of their interaction with κ-casein (Mackinlay and Wake 1971). Approximately 5% of the weight of casein micelles consists of colloidal calcium phosphate, which is sequestered by the micelles through interaction of the inorganic component with clustered serine phosphate residues of the calcium-sensitive caseins (Sleigh et al. 1979). In this way calcium phosphate is prevented from precipitating out of milk and is maintained in a form that can be readily assimilated by the young animal.

Phosphorylation of the serine clusters occurs posttranslationally as the result of the activity of one or more specific casein kinases located in the Golgi complex (Mackinlay et al. 1977). The recognition sequence for the casein kinase(s) is Ser/Thr-X-Y, where X may be any amino acid and Y is a glutamate, aspartate, or serine phosphate residue (Mercier 1981). The phosphorylation sites typically occur in the sequence Ser-Ser-Glu-Glu, sometimes preceded by Ser-X or extended in the other direction by Ser-X-Glu. It has been shown for the three rat calcium-sensitive casein genes that a splice junction occurs between the pair of glutamate codons so that sequences from two exons are involved in creating the clustered phosphorylation sites (Jones et al. 1985; Yu-Lee et al. 1986).

The caseins, being secreted proteins, possess N-terminal signal peptides that direct the passage of the newly synthesized polypeptides into the lumen of the endoplasmic reticulum, whereupon the signal peptides are removed from the primary translation products to yield the mature caseins. The leader peptide sequences are very similar among all calcium-sensitive caseins examined. This was first observed at the amino acid level for the ovine caseins (Mercier et al. 1978) and has led to the proposition that the calcium-sensitive caseins are evolutionarily related. The sequences of cDNAs determined subsequently for the caseins of the rat (Rattus rattus) (Blackburn et al. 1982; Hobbs and Rosen 1982), mouse (Mus musculus) (Hennighausen et al. 1982), guinea pig (Cavia cobaya) (Hall et al. 1984a, 1984b), and cow (Bos taurus) (Stewart et al. 1984) support this notion. The calcium-sensitive caseins are therefore thought to constitute a gene family that has arisen from an ancestral sequence by an extensive series of duplications and modifications while retaining similar 5'-untranslated, leader-peptide, and phosphorylation-site sequences. The remainder of the sequences appear to be subject to few functional constraints and have accumulated many nucleotide substitutions, as revealed by comparative sequence analysis.

In the present paper we present the complete cDNA sequences for bovine αS2- and β-caseins, thereby completing our characterization of the bovine caseins at the mRNA level. The relationship of bovine αS2-casein to cavine casein A, rat γ-, and murine ε-caseins is described herein and shows that these αS2-like caseins have diverged through extensive sequence rearrangements and nucleotide substitutions. The divergence of the bovine and rat β-caseins, by contrast, appears to have only involved extensive nucleotide substitution.

Material and Methods

The construction and identification of pBβC14 has been described (Willis et al. 1982). Plasmids pBβC16 and pBαS2C23 were isolated from the same cDNA library.
Because these three clones were incomplete copies of their corresponding mRNAs, another cDNA library was constructed using the method of Land et al. (1981). This cDNA library was screened by colony hybridization (Grunstein and Wallis 1979), using [α-32P]dATP-labeled HinfI restriction fragments from pBβC16 and pBaS2C23. Plasmids pBβC468, pBaS2C170, and pBaS2C411 were identified.

All sequencing was performed by using the method of Maxam and Gilbert (1980). Free-energy values for RNA secondary structure were calculated using the method of Ninio (1979) as detailed in de Wachter et al. (1982).

Computer alignments (figs. 2C, 4) were obtained by using the method of Taylor (1984) as implemented in the ALIGN program in the package described by Harr et al. (1986). The parameters were set to count 1 for each mismatch plus 1 for each gap introduced plus 1 for each residue that has a gap opposite it (i.e., gap penalty = 1 + length of gap).

Results and Discussion

cDNA Sequences

Restriction maps and sequencing strategies for the αs2- and β-casein cDNA clones are shown in figure 1. Plasmid pBaS2C170 contained an insert of 669 bp, excluding the GC tails. Plasmid pBaS2C411 contained an insert of 515 bp, including a poly A tail of 13 residues but excluding the GC tails. The overlapping sequences imply an mRNA of 1,024 nucleotides, excluding the poly A tail with 5'-untranslated, translated, and 3'-untranslated regions of 56, 666, and 302 nucleotides, respectively (fig. 2). It is likely that the sequence is complete at the 5' end, since it commences with the dinucleotide AT (see below) and is directly comparable to the 5'-untranslated region determined from the analysis.

![Fig. 1. Restriction maps of αs2- and β-casein cDNAs and strategies for their sequence determination. Coding sequences are indicated by thickened lines; jagged lines represent GC tails; and the numbers define the orientations, where known, of the cDNA inserts with respect to the pBR322 vector (all cDNAs cloned into the PstI site at position 3612). Arrows represent the extent and direction of sequence obtained from the restriction sites indicated by short vertical lines. Key: A = AaiI; D = DdeI; F = HinfI; H = HaelIII; K = KpnI; 9 = Sau96 I; and Y = HpaII.](image-url)
FIG. 2.—Alignment of the αS2-like casein cDNAs. The sequences for bovine αS2, cavine A (Hall et al. 1984b), rat γ (Hobbs and Rosen 1982), and mouse e-casein (Hennighausen et al. 1982) are compared, with similar sequences aligned vertically. 2A: 5'-Untranslated regions. Numbers refer to nucleotides with the last nucleotide of the untranslated region being set at −1. 2B: Translated regions. Numbers refer to codons, with the first codon of the mature proteins set at +1. U and D refer to the upstream and downstream portions, respectively, of the large duplication. Except for their C-terminal extremities, the mature protein-coding regions were aligned as follows: the closely related bovine and cavine sequences were aligned on the basis of dot-matrix comparisons (performed as described by Reisner and Bucholtz [1983]), gaps in these sequences being positioned to correspond to putative splice junctions. The more divergent rat and mouse sequences
were fitted to this alignment in a similar manner, viz., by identifying segments of these sequences that correspond to putative individual exons (see text) and then visually aligning these to corresponding segments of the bovine and canine sequences. The alignment downstream of bovine codon 116 was arrived at by considering amino acid sequences and aligning amino acids that are either identical or similar. This alignment indicates that the last two codons were not part of the duplication, a conclusion consistent with the presence of a splice junction at this position in the p-casein gene (see fig. 4). 2C: 3'-untranslated regions. Numbers refer to nucleotides. The polyadenylation signals are underlined. The bovine and canine 3' noncoding regions were aligned using the ALIGN program (see Material and Methods).
of the rat γ-casein cDNA and genomic sequences (Yu-Lee and Rosen 1983). The predicted leader-peptide sequence is identical to the published ovine αs2-casein leader-peptide sequence (Mercier et al. 1978). The predicted mature protein sequence varies from the published bovine αs2-casein A sequence (Brignon et al. 1977) only at amino acid 87, which is predicted to be glutamine rather than glutamate.

Plasmid pBβC468 contained an insert of 945 bp, excluding the GC tails. Plasmid pBβC16 contained an insert of 800 bp, excluding the GC tails and including a poly A tail of 19 residues. The overlapping sequences imply a mRNA of 1,089 nucleotides, excluding the poly A tail, with 5'-untranslated, translated, and 3'-untranslated regions of 56, 672, and 361 nucleotides, respectively (fig. 4). The first base of the 5'-most HinfI site shown on the restriction map is continuous with the G tail. Another β-casein cDNA isolated, pBβC319, also displays this feature. On this basis and from a comparison with rat β-casein cDNA and genomic sequences (Jones et al. 1985), we assume that the sequence presented is complete at the 5' end. The predicted leader-peptide sequence is identical to the published ovine β-casein leader-peptide sequence (Mercier et al. 1978), and the predicted mature protein sequence agrees with the published bovine β-casein A2 sequence (Ribadeau-Dumas et al. 1972), with the following exceptions: (1) amino acid 117, which is predicted to be glutamate rather than glutamine; (2) amino acids 137 and 138, which are predicted to be leucine-proline rather than proline-leucine; (3) amino acid 175, which is predicted to be glutamate rather than glutamine; and (4) amino acid 195, which is predicted to be glutamate rather than glutamine.

In both mRNAs the protein-coding region is predicted to commence with the first AUG encountered, and both conform to the observations of Kozak (1981). Both were found to contain the conventional polyadenylation signal (Fitzgerald and Shenk 1981).

The αs2-Like Caseins

The 5'-Untranslated Regions

A comparison of the 5'-untranslated regions of the αs2-like casein cDNAs is presented in figure 2A. The cavine and murine sequences are incomplete at their 5' ends. Comparisons of these sequences—as well as those of rat β- and the three αs1-like caseins for which sequences are also available—reveal a number of common features. All begin with the dinucleotide AT, which may be related to the efficiency of initiation of their translation, as in other families of mRNAs in which the first two bases are conserved (Breathnach and Chambon 1981). Their translation-initiation sites are also conserved, with a purine -3 and a C -1 to the initiating AUG.

Bases -14 to -11 consist of the sequence AGGA in the αs2-like and β-caseins and of AGAT in the αs1-like caseins. These have been shown to be the junctions between exons I and II in the three rat caseins, in bovine αs1-casein (Yu-Lee et al. 1986), and in bovine β-casein (J. Bonsing, unpublished results).

In the αs2-like caseins, the sequences from -36 to -13 contain a string rich in pyrimidines followed by a string rich in purines. These are able to form stable hairpin loops. For bovine αs2-casein, the calculated ΔG value in kilocalories per mole for this structure is -7.4 (when the thermodynamic binding model is used) or -9.6 (when the empirical binding model, as detailed in de Wachter et al. [1982], is used). This structure comprises nucleotides -36 to -27 base pairing with nucleotides -21 to -13 to form a 9-bp stem including a single G:U bp, a single looped-out nucleotide (-31), and an unpaired loop containing nucleotides -26 to -22. Similar structures are possible
for casein A and rat γ-casein with calculated values of −7.1 or −9.6 and −5.7 or −6.6 kcal/mol, respectively. Comparable secondary structures cannot be drawn for other published casein sequences, although they possess similar purine and pyrimidine strings. Occurring within the stem portions of these loops is the sequence GGAA, also found in the region of 18S rRNA proposed to interact with mRNAs during initiation of translation (Baralle 1983). These secondary structures may therefore modulate the rate of initiation of translation of the αS2-like casein mRNAs, the products of which are all minor components in the milks in which they occur.

The Translated Regions

Figure 2B presents an alignment of the translated regions of the four αS2-like casein cDNAs. A prominent feature of these sequences is a repeated structure, first detected in the amino acid sequence of bovine αS2-casein (Brignon et al. 1977) and extended here to give a tandem repeat consisting of upstream (U) codons 33–125 and downstream (D) codons 126–205 in the bovine sequence and codons 33–113 and 114–206 in the cavine sequence. Rat codons 34–104 and 105–161 also correspond to portions of this tandem repeat. In murine ε-casein the repeat is absent, but this sequence has a unique repetition of codons 12–33 in codons 53–75.

Predominantly hydrophilic amino acid sequences are encoded by (1) the bovine sequence up to approximately codon 68, (2) the sequences in other species aligned with this section, and (3) the corresponding downstream repeated sequences. These are shown divided into a number of segments in figure 3. The boundaries of these segments, most of which are likely to correspond to existing or ancestral splice junctions, have been defined according to two criteria: (1) identity with established splice junction sequences, particularly those adjacent to phosphate clusters (Jones et al. 1985; Yu-Lee et al. 1986), with such boundaries including those at codons 11–12, 41–42, 59–60, and 67–68, and (2) coincidence with gaps in the alignment, e.g., following codon 33, where gaps occur at this position in the mouse and rat sequences and following codons 67–68, where a gap occurs at this position in the cavine sequence. A boundary following codon 50 may be inferred since a gap occurs after codon 141, the corresponding position in the upstream repeat. Further evidence for boundaries at codons 50–51 and 59–60 is provided by the deletion that characterizes the αS2-casein D variant.

FIG. 3.—Summary of the sequence alignments of the αS2-like caseins. Numbering refers to the first codon of each segment, except for termination codons and cavine 67, murine 33 and 75, and rat 33 and 104, which are the last codons in the sections of sequence that precede them. 5'ut = 5'-Untranslated; L = leader peptide-coding region; P = major phosphorylation site found within this segment; 3'ut = 3'-untranslated region; and A, = poly (A) tail. Minor gaps in the sequence alignment of fig. 2 have been omitted. Also shown are regions of 3'-untranslated sequences that show detectable similarity on the basis of dot-matrix analysis.
(Grosclaude et al. 1979). Although the presence of splice sites at these positions remains to be experimentally demonstrated, the situation clearly parallels that in the rat β-casein gene, where the hydrophilic phosphate cluster-containing N-terminus of the molecule has been shown to be encoded by a series of short exons postulated to have arisen as the result of a series of duplications of an ancestral sequence (Jones et al. 1985).

The bovine sequences from codons 68–125, the C-terminal codons 150–207, and the corresponding sequences in the other αs2-like casein encode predominantly hydrophobic protein domains. These sequences have been subject to a high rate of amino acid replacement. For example, of the 42 nucleotide differences found between bovine codons 83–125 and cavine codons 68–113, 30.5 (73%) are replacement differences. We believe that this implies selective neutrality toward hydrophobic amino acid replacements in these domains—and that they do not code for structurally important amino acid sequences, in contrast to the well-conserved phosphorylation-site sequences. However, constraints have acted to maintain a relatively high degree of hydrophobicity in these domains as a whole, a characteristic important for the protein-protein interactions involved in casein micelle formation (Waugh 1971). Figure 3 summarizes the relationships between the four αs2-like caseins.

The β-Caseins

An alignment of the bovine and rat β-casein cDNAs is presented in figure 4. Many nucleotide substitutions have occurred during the divergence of these caseins, but no major insertions, deletions, or other sequence rearrangements—such as are found in the αs2- and αs1-like caseins (Stewart et al. 1984)—are evident. From the amino acid sequence of human β-casein (Greenberg et al. 1984), we infer that this is also the case for that species. In the rat, four short exons encode the N-terminal hydrophilic sequence that contains a single phosphate cluster, whereas the remainder of the molecule, of markedly hydrophobic character, is encoded by a single exon. The conservation of rat and bovine sequences that span the splice sites of the rat β-casein gene (indicated by arrowheads in fig. 4) suggests that the arrangement of exons and introns has been conserved in the rat and bovine β-casein genes.

The absence of major sequence rearrangements in the β-caseins and the maintenance of hydrophilic N-terminal and hydrophobic C-terminal domains, despite many amino acid replacements, points to the operation of functional constraints acting to conserve the overall architecture of the molecule. Consistent with this conclusion is recently obtained evidence that shows β-casein to be important in determining the surface properties of casein micelles and essential for curd formation when milk is clotted by the proteolytic enzyme chymosin (Pearse et al. 1986). Curd formation is physiologically important since it ensures the retention of milk protein in the stomach of the infant and allows further digestion to occur.

A similar conservation of structure is found for the κ-caseins, which are also important in determining properties of casein micelles, including the maintenance of micelle stability and the initiation of curd formation when this protein is cleaved by chymosin at a specific Phe-Met bond. The importance of the β- and κ-caseins is indicated by their occurrence in all eutherian milks so far examined (Jenness 1979).

The α-caseins, by contrast, occur in varying amounts or may be absent altogether, as is the case for human milk. These caseins appear to be much less important for curd formation than the β- and κ-caseins (Pearse et al. 1986), and their role is probably restricted to determining the capacity of casein micelles for calcium phosphate transport.
FIG. 4.—The nucleotide sequence of bovine β-casein cDNA shown aligned with the corresponding rat sequence (Blackburn et al. 1982). Alignment was carried out using the ALIGN program (see Material and Methods). Translation-initiation and -termination codons and the first codons of each of the mature proteins are underlined. Arrowheads indicate the position of splice junctions in the rat sequence (Jones et al. 1985).

Thus, the structure-function relationships of the different caseins are reflected in the manner in which each has been modified during the evolution of mammalian species.
Acknowledgments

We wish to thank Dr. A. Rcisner and G. Kornfeld for assistance with computer analyses. This work was supported by the Australian Research Grants Scheme, the Australian Dairy Research Committee, and by Commonwealth Postgraduate Awards to A.F.S. and J.B.

LITERATURE CITED


WALTER M. FITCH, reviewing editor

Received July 7, 1986; revision received November 14, 1986.