

Complete Nucleotide Sequences of Bovine α_{S2} - and β -Casein cDNAs: Comparisons with Related Sequences in Other Species¹

A. Francis Stewart,² John Bonsing, Craig W. Beattie,³ Farida Shah,⁴ Ian M. Willis,⁵ and Antony G. Mackinlay

School of Biochemistry, University of New South Wales

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 γ - and mouse ϵ -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

1. Key words: bovine α_{S2} -casein mRNA, bovine β -casein mRNA, interspecific sequence comparisons, intragenic duplication, casein evolution.

2. Current address: Institute for Cell and Tumour Biology, German Cancer Research Centre.

3. Current address: Division of Surgical Oncology, University of Illinois College of Medicine.

4. Current address: Department of Genetics, University of Malaysia.

5. Current address: Department of Molecular Biophysics and Biochemistry, Yale University.

Address for correspondence and reprints: Dr. A. G. Mackinlay, School of Biochemistry, University of New South Wales, Kensington, N.S.W. 2033, Australia.

Mol. Biol. Evol. 4(3):231–241, 1987.

© 1987 by The University of Chicago. All rights reserved.

0737-4038/87/0403-0003\$02.00

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-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 now 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 α_{s2} - and β -caseins, thereby completing our characterization of the bovine caseins at the mRNA level. The relationship of bovine α_{s2} -casein to bovine casein A, rat γ -, and murine ϵ -caseins is described herein and shows that these α_{s2} -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 pB β C14 has been described (Willis et al. 1982). Plasmids pB β C16 and pB α_{s2} C23 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 - ^{32} P]dATP-labeled *Hinf*I restriction fragments from pB β C16 and pB α_{s2} C23. Plasmids pB β C468, pB α_{s2} C170, and pB α_{s2} C411 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 α_{s2} - and β -casein cDNA clones are shown in figure 1.

Plasmid pB α_{s2} C170 contained an insert of 669 bp, excluding the GC tails. Plasmid pB α_{s2} C411 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

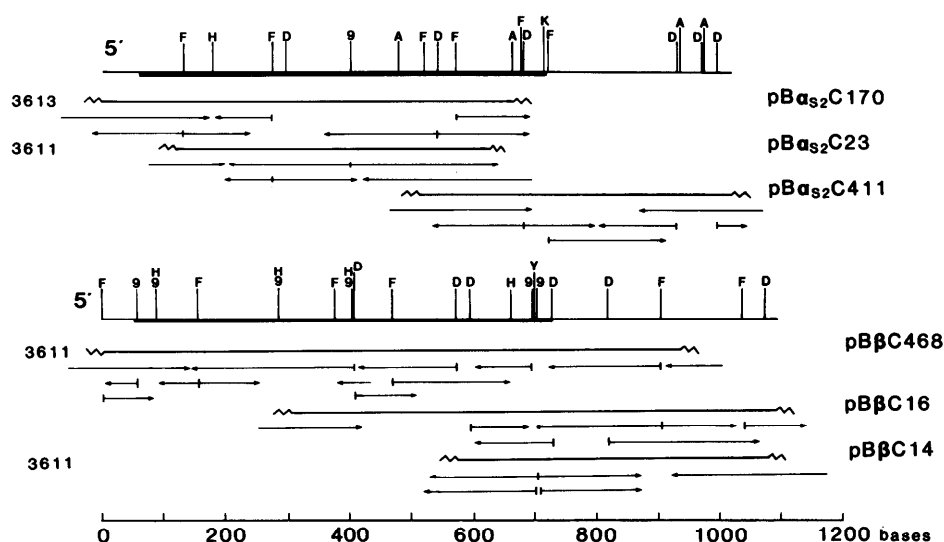


FIG. 1.—Restriction maps of α_{s2} - and β -casein cDNAs and strategies for their sequence determination. Coding sequences are indicated by thickened lines; jagged lines represent GC tails; and the numbers define the orientations, where known, of the cDNA inserts with respect to the pBR322 vector (all cDNAs cloned into the *Pst*I site at position 3612). Arrows represent the extent and direction of sequence obtained from the restriction sites indicated by short vertical lines. Key: A = *Alu*I; D = *Dde*I; F = *Hinf*I; H = *Hae*III; K = *Kpn*I; 9 = *Sau*96 I; and Y = *Hpa*II.

2A: 5' UNTRANSLATED REGIONS

BOVINE α_{s2}	-56	ATATTCCATTGCCTGGACTACTTGTCTTCCTTTTAGGAAACGAGGACAAAGCAAAC
CAVINE A	-46	ACCTGGCCTTCTTGTCTTCCACTTGGGAAATAAGGACTCAGTAAAC
RAT γ	-56	ATCATCATCTACCTATTCTCTCTCTTCCACTTGGGAAACGAAGGATCAAGTAAAC
MURINE ϵ	-23	TAGGAAGCAAGGACTCATTAATC

2B: TRANSLATED REGIONS

SIGNAL PEPTIDE CODING REGIONS

BOVINE	-15	ATG AAG TTC TTC ATC TTT ACC TGC CTT TTG GCT GTT GCC CTT GCA	-1
CAVINE	-15	ATG AAG CTC TTC ATC TTC ACC TGC CTC TTG GCT GTG GCT CTC GCA	-1
RAT	-15	ATG AAG TTC TTC ATC TTC ACC TGC CTG GTG GCT GCT GCT CTG GCT	-1
MURINE	-15	ATG AAG TTC ATC ATT CTG ACT TGC CTT TTG GCC GTT GCT CTT GCA	-1

MATURE PROTEIN CODING REGIONS

BOVINE	1	AAG AAT ACG ATG GAA CAT GTC TCC TCC AGT GAG GAA TCT ATC	14
CAVINE	1	AAG CAC AAG TCA GAG CAA CAG TCC TCC AGT GAG GAA TCT GTC AGC	15
RAT	1	AAG CAT GCG GTA AAG GAT AAA CCC TCC AGT GAG GAA TCT GCC AGT	15
MURINE	1	AAG CAG AGG ATG GAG CAA TAC ATC TCC AGT GAG GAG TCC ATG GAT	15
MURINE	53		56
		GTA TCT GAC ATC	
BOVINE	15	ATC TCC CAG GAA ACA TAT AAG CAG GAA AAG AAT ATG GCC ATT AAT	22
CAVINE	16	ATC TCC CAG GAA AAA TTT AAG GAC AAG AAT ATG GAC ACT ATT	22
RAT	16	GTC TAC TTG GGA AAG TAT AAG CAG GGC AAT AGT GTG TTC TTC CAA	30
MURINE	16	AAC TCT CAA GAA AAC TTT AAG CAG AAT ATG GAT GTG GCC TTT TTT	30
MURINE	57	ATT TCT CAG CAA CAA TAC AAC CAG AAA ATG ATG GAC ATG AGT	70
BOVINE	30	CCC AGC AAG	32
CAVINE	30	TCT AGT GAG	32
RAT	31	ACA CCT CAG	33
MURINE	31	CCC AGT CAG	33
MURINE	71	GTG AGT GCC AGA GAG	75
BOVINE U	33	GAG AAC CTT TGC TCC ACA TTC TGC AAG GAA GTT GTA AGG AAC GCA	43
CAVINE U	33	GAG ACC ATT TGT GCC TCC CTC TGC AAG GAA GCT ACA AAG AAC ACA	43
BOVINE D	126	GAG CAG CTC TCC ACC AGT GAG GAA AAT TCA AAG AAG ACC	139
CAVINE D	114	GAA CAG ATC TCC ACC ATT GAG GAC ATC CTA AAG AAG ACT	126
MURINE	34	GAA ACT GTA GAG AAC ATT	39
BOVINE U	48	AAT GAA GAG GAA TAT TCT ATC GGC TCA TCT AGT GAG GAA	60
CAVINE U	48	CCT AAA ATG GCA TTC TTT AGC AGA TCT TCT AGC GAG GAA	60
BOVINE D	139	GTG GAC ATG	142
CAVINE D	127	ACT GCT GTG GAA TCC TCT AGC AGT TCA TCT ACG GAG AAA	139
RAT	34	GAC TCT GCT AGC AGT TCA TCC AGC GAG GAA	43
MURINE	40	TAC ATT CCC CAA ATG	45
BOVINE U	61	TCT GCT GAA GTT GCC ACT GAG GAA GTT AAG ATT ACT GTG GAC GAT	75
CAVINE U	61	TTT GCT GAC ATT CAC AGA GAG	75
RAT U	44	TCA TCT GAG GAA ATC AGC GAA AAA ATT GAA CAG TCT GAG GAA CAA	58
BOVINE D	143	TCA ACA GAA GTA TTC ACT AAG AAA ACT AAA CTG ACT GAA GAA GAA	157
CAVINE D	140	TCC ACT GAC GTA TTT ATT AAG AAA ACT AAG ATG GAT GAA GTA CAA	154
RAT D	105	CAG GAG TCC ACC AGC ACC TCA GTG GAG GAA ATC CTG	116
MURINE	46	TCT GTT GAA GCT CCC ATG AAG	52
MURINE	76	AAA ACT GTG ATG ACT GAA GAA AGT	83

FIG. 2.—Alignment of the α_{s2} -like casein cDNAs. The sequences for bovine α_{s2} , cavine A (Hall et al. 1984b), rat γ - (Hobbs and Rosen 1982), and mouse ϵ -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

2B: TRANSLATED REGIONS (CONTINUED)

BOVINE	U	76	AAG	CAC	TAC	CAG	AAA	GCA	CTG	AAT	GAA	ATC	AAT	CAG	TTT	TAT	CAG	90
CAVINE	U	68								AAC	AAA	AAA	GAT	CAA	CTT	TAT	CAG	75
RAT	U	59	AAG	GTC	AAC	CTA	AAT	CAG	CAG	AAG	AAA	AGC	AAG	CAA	TTT	TCC	CAG	73
BOVINE	D	158	AAG	AAT	CGC	CTA	AAT	TTT	CTG	AAA	AAA	ATC	AGC	CAG	CGT	TAC	CAG	172
CAVINE	D	155	AAG	CTT	ATC	CAA	AGT	CTC	CTG	AAC	ATA	ATT	CAT	GAA	TAT	TCC	CAG	169
RAT	D	117	AAG	AAG	ATC	ATT	GAT	ATC	GTC	AAG	TAC	TTC		CAA		TAT	CAA	129
MURINE		84	AAG	AAC	ATC	CAA	GAC	TAC	ATG	AAC	AAA	ATG	AAA	CGA		TAC	AGC	97
BOVINE	U	91	AAG	TTC			CCC	CAG	TAT	CTC	CAG	TAT	CTG	TAT	CAA	GGT	CCA	103
CAVINE	U	76	AAG	TGG	ATG	GTA	CCC	CAG	TAT	AAC	CCG	GAT	TTT	TAC	CAA	CGT	CCC	90
RAT	U	74	GAC	TCT	TCC	TTC	CCC	CAA	ATT	TGC		ACA	CCT	TAT	CAA	CAG	CAG	87
BOVINE	D	173	AAA	TTC	GCC	TTG	CCC	CAG	TAT	CTC	AAG	ACT	GTT	TAT	CAG	CAT	CAG	187
CAVINE	D	170	AAG	GCC	TTC	TGG	TCC	CAG	ACC	CTA	GAA	GAT	GTT	GAC	CAG	TAT	CTG	184
RAT	D	130	CAG	CTC	ACC	AAC	CCC	CAC	TTT	CCC	CAG	GCT	GTT	CAT	CCA		CAG	143
MURINE		98	AAG	ATC	ACC	TGG	CCA	CAG	TTT	GTA	AAG	CTT	CTC	CAT	CAA	TAC	CAG	112
BOVINE	U	104	ATT	GTT	TTG	AAC	CCA	TGG	GAT	CAG	GTT	AAG	AGG	AAT	GCT	GTT	CCC	118
CAVINE	U	91	GTT	GTC	ATG	AGC	CCG	TGG	AAT	CAG	ATT	TAC	ACA	AGA	CCC	TAC	CCC	105
RAT	U	88	AGT	AGT	GTG	AAC	CAA	CGG	CCC	CAG	CCT	AAT			GCC	ATC		99
BOVINE	D	188	AAA	GCT	ATG	AAG	CCA	TGG	ATT	CAA	CCT	AAG	ACA	AAG	GTT	ATT	CCC	202
CAVINE	D	185	AAA	TTT	GTA	ATG	CCG	TGG	AAT	CAC	TAT	AAT	ACA	AAT	GCT	GAC	CAA	199
RAT	D	144	ATA	CGT	GTG	AGC	TCT	TGG	GCC		CCC	AGT		AAG	GAC	TAC	ACC	156
MURINE		113	AAA	ACC	ATG	ACC	CCG	TGG	AGC						TAT	TAC	CCA	122
BOVINE	U	119	ATT		ACT	CCC	ACT	CTG	AAC	AGA								125
CAVINE	U	106	ATT	GTG	CTT	CCC	ACT	TTG	GGC	AAA								113
RAT	U	100	TAT	GAC	GTC	CCC	AGC											104
BOVINE	D	203	TAT	GTG						AGG	TAC	CTT	TAA					208
CAVINE	D	200	GTC	GAT		GCC	TCT	CAG	GAG	AGA	CAA	GCT	TGA					209
RAT	D	157	TTC			CCA	ACT	GCT	CGC	TAC	ATG	GCT	TGA					165
MURINE		123	TCT		ACC	CCC	AGC				CAG	GTT	TGA					129

2C: 3' UNTRANSLATED REGIONS

	725	740	760	780	
BOVINE	TAAGATTCTGGAATTAACTGCTTCTACCTGGTTATGGTTGGACTGGAAAAATCTATCTTCTACAGTTT				
CAVINE	T GATTCTTGGATTAACTGTTTTTACCTTGCTTGGTTTAACTGGAAAAATCTACCTCTACAGTCT				
	718	740	760	780	
	800	820			
BOVINE	CATA TCTACCACCTTTA CTTCATACAACCAGCATGTTTGG			GTGG	
CAVINE	CTTTCTCTGCCC TTTTAACTTCACGCTCCAGCATATTTAGACAGACTTACAGACAAGACAGAAATCC				
	800	820			
	840	860	880	900	
BOVINE	TGCATCAATAAGA CAAATCGCAGAGTATTTCTGTAATTATTTATACCTCTCTGTTAGTTTATATG				
CAVINE	TGAGCCA AAGAAGACCCCCACAGGGCGCCTTCGGRAATT CCAGTCTTTCTGCTGGTCCGTTGCA				
	860	880	900		
	920	940	960		
BOVINE	AATATACTGGGTTTATATGTTGGTTTCATACAAACTTAGCTGATGATTATTTAAAAATCCTTTCCCTAC				
CAVINE	AATCTGTTGGATCCGTA TTGTTGCATAGGAGCTGCGTTGATAACTAATTTAACGTTTCTTCGAA				
	920	940	960		
	980	1000	1020		
BOVINE	CCTTTCTGAGCTATAGAAATATTTTTCTTTCCCTCTGAGAAATAAAATTTCAAGTGC(A)				
CAVINE	CTTTTTTTTCTGTTAAACCACTTACTTTTCTCGCAAAATAAACTCTAAGTGCATCCCTT(A _n)				

FIG. 2 (Continued)

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 cavine 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 β -casein gene (see fig. 4). 2C: 3'-Untranslated regions. Numbers refer to nucleotides. The polyadenylation signals are underlined. The bovine and cavine 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 α_{s2} -casein leader-peptide sequence (Mercier et al. 1978). The predicted mature protein sequence varies from the published bovine α_{s2} -casein A sequence (Brignon et al. 1977) only at amino acid 87, which is predicted to be glutamine rather than glutamate.

Plasmid pB β C468 contained an insert of 945 bp, excluding the GC tails. Plasmid pB β C16 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 *Hinf*I site shown on the restriction map is continuous with the G tail. Another β -casein cDNA isolated, pB β C319, 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 glutamine rather than glutamate; 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 α_{s2} -Like Caseins

The 5'-Untranslated Regions

A comparison of the 5'-untranslated regions of the α_{s2} -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 α_{s1} -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 α_{s2} -like and β -caseins and of AGAT in the α_{s1} -like caseins. These have been shown to be the junctions between exons I and II in the three rat caseins, in bovine α_{s1} -casein (Yu-Lee et al. 1986), and in bovine β -casein (J. Bonsing, unpublished results).

In the α_{s2} -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 α_{s2} -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 cavine 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 α_{s2} -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 α_{s2} -like casein cDNAs. A prominent feature of these sequences is a repeated structure, first detected in the amino acid sequence of bovine α_{s2} -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 α_{s2} -casein D variant

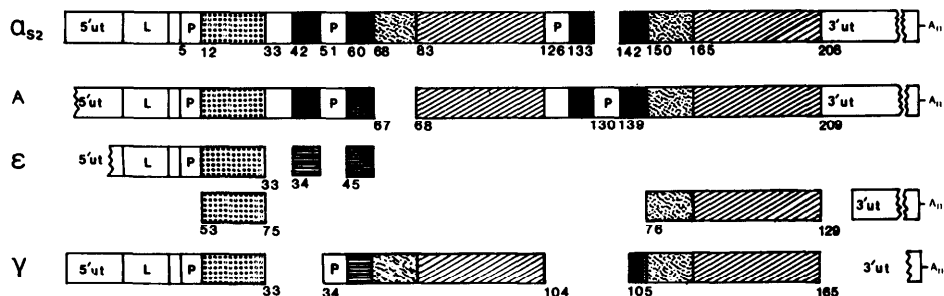


FIG. 3.—Summary of the sequence alignments of the α_{s2} -like caseins. Numbering refers to the first codon of each segment, except for termination codons and cavine 67, murine 33 and 75, and rat 33 and 104, which are the last codons in the sections of sequence that precede them. 5'ut = 5'-Untranslated; L = leader peptide-coding region; P = major phosphorylation site found within this segment; 3'ut = 3'-untranslated region; and A_n = 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 α_{s2} -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 α_{s2} -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 α_{s2} - and α_{s1} -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.

BOVINE	1	ATTC	AGCTCCTCCTTCACTTCTGTGCTCTACTTTGAAAAAGGAATTGAGAGCC	
RAT	1	ATCCTCTGAGCTTCACTCTCTCT	CTTGTCTCCGCT	AAAGACTTGCAGCC
BOVINE	57	<u>ATGAAGGTCTCATCCTTGCTGCCTGGTGGCTCTGGCCCTTGCAAGAGAGCTGGAAGAAC</u>		
RAT	53	<u>ATGAAGGTCTTCATCCTTGCTGCCTTGTGGCACCTTGTCTTGCAGGGGAGAAGGATGCAT</u>		
BOVINE	118	TCAATGTACTCGGTGAGATTGTGGAAAGCCTTTCAAGCAGCGAGGAATCAATTACACGCAT		
RAT	114	TCACTGTGTCTCTGAGACT	GGTAGTATTTC	CAGTGAGGAATCCGTGAACATAT
BOVINE	179	CAATAAGAAAAATTGAGAAAGTTTCAGAGTGAGGAACAGCAGCAACAGAGGATGAATCCAG		
RAT	169	CAATGAGAAACTTCAGAAGGTGAAACTCATGGACAGGTGCAGTCCGAGGATGTGCTCCAG		
BOVINE	240	GATAAAATCCACCCCTTTTGCCCGAGACACAGCTCTCTAGTCTATCCCTTCCCTGGGCCCATCC		
RAT	230	AATAAATTTCACTCCGGCAITTCAGTCAGAACCCAAAGGCCATTCCATATGCCAGACCATCT		
BOVINE	301	CTAACAG	CCTCCACAAAAACATCCCTCCTCTTACTCAAACCCCTGTGGTG	GT
RAT	291	CTTGCACTCCCAATTCACAAAAACATCCAGCCTATTGTCTCAACCCCTGTGGTGCCAACTGA		
BOVINE	353	GCCGCCTTTCTTCAGCCTGAAGTAATGGGAGTCTCCAAAGTGAAGGAGGCTATGGCCCT		
RAT	352	TGGGCCTATCATTTCTCTCTGAAGTGAATCCCTTCTTAAAGCTAAAGCCACTGTCTTCCC		
BOVINE	414	AAGCACAAAGAAATGCCCTTCCCTAAATATCCAGTTGAGC	CCTTTACTGAAA	GCCAGA
RAT	413	AAGCACAAACAGATGCCCTTCCCTTAACTCTGAAACCGTGCTCCGCCTCTTTAACTCTCAAA		
BOVINE	472	GCCTGACTCTCACTGATGTTGAAAATCTGCACCTTCTCTGCTCTGCTCCAGTCTTGGAT		
RAT	474	TCCCCAGTCT	TGATCTTGCTAATCTGCACCTTCTCTAGTCTCCAGCCAG	CTCCAG
BOVINE	533	GCACCAGCCTCACCAGCCTCTTCTCCAACT	GTCATGTTTCTCCTCAGTCCGTGCTG	
RAT	530	GCA	CAAATTGTGCAGGCTTTTCTCTCAGACTCCCGCGGTGGTTTCTCTCAGCCCCAGCTG	
BOVINE	591	TCCCTTTCTCAGTCCAAAGTCTGCTGTTCCTCCAGAAAGCAGTGCCCTATCCCGAGAG		
RAT	590	TCTCATCTCAGTCCAAAGTCACTGCTGTGTCAGCAACTAGCCCCCTCTTCCAACAAG		
BOVINE	652	ATATGCCCATTCAGGCCTTTCTGCTGTACCAGGAGC	CTG TA	CTCGTCTG T
RAT	651	GTATGCTGTTCAGACCTTCTGCACTACCTAGATCTTCTGCTTAACCCCAACCTCCAGTT		
BOVINE	704	CCGGGGAC	CCTTC	CCTATTATGTCTAAGAGGATTTCAAAGTGAATGC CC
RAT	712	CC	TGCCACTCAACAACTTCATTCACCTTCTCTAAGAGGATTTCCGGGTATTCTCT	
BOVINE	754	CCTCCTCA	CTTTTGAATTGACTGCGACTGGAAATATGGCAACTTTTCAATCCTTGC	
RAT	772	CCTCATCATTTTTTTTTTGAATTGACTGAAACTGGAAATGTACAATCATTTTCATCTGGG		
BOVINE	810	ATCATGTTACTAA	GATA	ATTTTTAAATGAGTATACATGGAACAAAAAATGAAACTTT
RAT	833	ATCATGCTACAAAAGTGATATATTTCTGAATGAACTACATGG	AAAAAAGAAATTTT	
BOVINE	867	ATTCCTTTATTATTATTTATG	CTTTTTCATCTTAAGTTGAATTTGAATCAT	AAACTAT
RAT	891	ATTCCTTTATTATTCTATGTATTATATGATATTCAATTTGAATTTGACCCATGGACTCTAT		
BOVINE	924	ATATTTCAAATTTTAATTCA	ACATTAGCATAAAAGTTCATTTTAACTTGAAAATATCA	
RAT	952	ATGTGTCTCAATATATTTTCATATAAATACTACAAATGTTCAATATTGAGTTTAA	A	
BOVINE	984	TGAACATATCAAAATATGTATAAAAAATAATTTCTG	AAATTG	TGATTATTATTTCCT
RAT	1007	TGA	AAGTCAATA	ATGTA CAAAATAGCTCATGAAAAATGGATTACTATTATTCTT
BOVINE	1040	TAAGAATCTATTTCCTAACCAAGTCATTTCATAAAATTAATCC	TTAGGCAT(A ⁿ)	
RAT	1064	TGGGAACCAATTCCTTTCCAGTCATTTCATAAAATAATCCTTTAGGCAT(A ⁿ)		

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. Reisner 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

- BARALLE, F. E. 1983. The functional significance of leader and trailer sequences in eukaryotic mRNAs. *Int. Rev. Cytol.* **81**:71-106.
- BLACKBURN, D. E., A. A. HOBBS, and J. M. ROSEN. 1982. Rat β casein cDNA: sequence analysis and evolutionary comparisons. *Nucleic Acids Res.* **10**:2295-2307.
- BREATHNACH, R., and P. CHAMBON. 1981. Organisation and expression of eukaryotic split genes coding for proteins. *Annu. Rev. Biochem.* **50**:349-383.
- BRIGNON, G., B. RIBADEAU-DUMAS, J.-C. MERCIER, J.-P. PELLISIER, and B. C. DAS. 1977. Complete amino acid sequence of bovine α_{s2} -casein. *FEBS Lett.* **76**:274-279.
- DE WACHTER, R., M.-W. CHEN, and A. VANDENBERGHE. 1982. Conservation of secondary structure in 5S ribosomal RNA: a uniform model for eukaryotic eubacterial, archaebacterial and organelle sequences is energetically favourable. *Biochimie* **64**:311-329.
- FITZGERALD, M., and T. SHENK. 1981. The sequence 5'-AAUAAA-3' forms part of the recognition site for polyadenylation of late SV40 mRNAs. *Cell* **24**:251-260.
- GREENBERG, R., M. L. GROVES, and H. J. DOWER. 1984. Human β -casein: amino acid sequence and identification of phosphorylation sites. *J. Biol. Chem.* **259**:5132-5138.
- GROSCLAUDE, F., P. JOUDRIER, and M.-F. MAHE. 1979. A genetic and biochemical analysis of a polymorphism of bovine α_{s2} -casein. *J. Dairy Res.* **46**:211-213.
- GROSCLAUDE, F., J.-C. MERCIER, and B. RIBADEAU-DUMAS. 1973. Genetic aspects of cattle casein research. *Neth. Milk Dairy J.* **27**:328-340.
- GRUNSTEIN, M., and J. WALLIS. 1979. Colony hybridisation. *Methods Enzymol.* **68**:379-389.
- HALL, L., J. E. LAIRD, and R. K. CRAIG. 1984a. Nucleotide sequence determination of guinea-pig casein B mRNA reveals homology with bovine and rat α_{s1} -caseins and conservation of the non-coding regions of the mRNA. *Biochem. J.* **222**:561-570.
- HALL, L., J. E. LAIRD, J. C. PASCALL, and R. K. CRAIG. 1984b. Guinea-pig casein A cDNA: nucleotide sequence analysis and comparison of the deduced protein sequence with that of bovine α_{s2} -casein. *Eur. J. Biochem.* **138**:585-589.
- HARR, R., P. FALLMAN, M. HAGGSTROM, L. WAHLSTROM, and P. GUSTAFSSON. 1986. GENIEUS: a computer system for DNA and protein sequence analysis containing an information retrieval system for the EMBL data library. *Nucleic Acids Res.* **14**:273-284.
- HENNIGHAUSEN, L. G., A. STEUDLE, and A. E. SIPPEL. 1982. Nucleotide sequence of cloned cDNA coding for mouse ϵ -casein. *Eur. J. Biochem.* **126**:569-572.
- HOBBS, A. A., and J. M. ROSEN. 1982. Sequence of rat α - and γ -casein mRNAs: evolutionary comparison of the calcium-dependent rat casein multigene family. *Nucleic Acids Res.* **10**:8079-8098.
- JENNESS, R. 1979. Comparative aspects of milk proteins. *J. Dairy Res.* **46**:197-210.
- JONES, W. K., L.-Y. YU-LEE, S. M. CLIFT, T. L. BROWN, and J. M. ROSEN. 1985. The rat casein multigene family: fine structure and evolution of the β -casein gene. *J. Biol. Chem.* **260**:7042-7050.
- KOZAK, M. 1981. Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes. *Nucleic Acids Res.* **9**:5233-5252.
- LAND, H., M. GREZ, H. HAUSER, W. LINDENMAIER, and G. SCHUTZ. 1981. 5' Terminal sequences of eucaryotic mRNAs can be cloned with high efficiency. *Nucleic Acids Res.* **9**:2251-2266.
- MACKINLAY, A. G., and R. G. WAKE. 1971. κ -Casein and its attack by rennin (chymosin). Pp. 175-215 in H. A. MCKENZIE, ed. *Milk proteins: chemistry and molecular biology*. Vol. 2. Academic Press, New York.

- MACKINLAY, A. G., D. W. WEST, and W. MANSON. 1977. Specific casein phosphorylation by a casein kinase from lactating bovine mammary gland. *Eur. J. Biochem.* **76**:233–243.
- MAXAM, A. M., and W. GILBERT. 1980. Sequencing end-labelled DNA with base-specific chemical cleavages. *Methods Enzymol.* **65**:499–560.
- MERCIER, J.-C. 1981. Phosphorylation of caseins: present evidence for an amino acid triplet code posttranslationally recognized by specific kinases. *Biochimie* **63**:1–17.
- MERCIER, J.-C., G. BRIGNON, and B. RIBADEAU-DUMAS. 1973. Structure primaire de la casein κ B bovine: s  quence compl  te. *Eur. J. Biochem.* **35**:222–235.
- MERCIER, J.-C., F. GROSCLAUDE, and B. RIBADEAU-DUMAS. 1971. Structure primaire de la caseine α _{S1}-bovine: s  quence compl  te. *Eur. J. Biochem.* **23**:41–51.
- MERCIER, J.-C., G. HAZE, P. GAYE, and D. HUE. 1978. Amino terminal sequence of the precursor of ovine β -lactoglobulin. *Biochem. Biophys. Res. Commun.* **82**:1236–1245.
- NINIO, J. 1979. Prediction of pairing schemes in RNA molecules—loop contributions and energy of wobble and non-wobble pairs. *Biochimie* **61**:1133–1150.
- PEARSE, M. J., P. M. LINKLATER, R. J. HALL, and A. G. MACKINLAY. 1986. The effect of casein composition and casein dephosphorylation on the coagulation and syneresis of artificial micelle milk. *J. Dairy Res.* **53**:381–390.
- REISNER, A. H., and C. A. BUCHOLTZ. 1983. Apparent relatedness of the main component of ovine 1.714 satellite DNA to bovine 1.715 satellite DNA. *EMBO J.* **2**:1145–1149.
- RIBADEAU-DUMAS, B., G. BRIGNON, F. GROSCLAUDE, and J.-C. MERCIER. 1972. Structure primaire de la caseine β bovine. *Eur. J. Biochem.* **25**:505–514.
- SLEIGH, R. W., T. B. SCULLEY, and A. G. MACKINLAY. 1979. The binding of β -casein to hydroxyapatite: the effect of phosphate content and location. *J. Dairy Res.* **46**:337–342.
- STEWART, A. F., I. M. WILLIS, and A. G. MACKINLAY. 1984. Nucleotide sequences of bovine α _{S1}- and κ -casein cDNAs. *Nucleic Acids Res.* **12**:3895–3907.
- TAYLOR, P. 1984. A fast homology program for aligning biological sequences. *Nucleic Acids Res.* **12**:447–455.
- WAUGH, D. F. 1971. Formation and structure of casein micelles. Pp. 3–85 in H. A. MCKENZIE, ed. *Milk proteins: chemistry and molecular biology*. Vol. 2. Academic Press, New York.
- WILLIS, I. M., A. F. STEWART, A. CAPUTO, A. R. THOMPSON, and A. G. MACKINLAY. 1982. Construction and identification by partial nucleotide sequence analysis of bovine casein and β -lactoglobulin cDNA clones. *DNA* **1**:375–386.
- YU-LEE, L.-Y., L. RICHTER-MANN, C. H. COUCH, A. F. STEWART, A. G. MACKINLAY, and J. M. ROSEN. 1986. Evolution of the casein multigene family: conserved sequences in the 5' flanking and exon regions. *Nucleic Acids Res.* **14**:1883–1902.
- YU-LEE, L.-Y., and J. M. ROSEN. 1983. The rat casein multigene family. I. Fine structure of the γ -casein gene. *J. Biol. Chem.* **258**:10794–10804.

WALTER M. FITCH, reviewing editor

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