Complete Nucleotide Sequence of the Bovine β-casein Gene

John Bonsing, Jennifer M. Ring[A]*, A. Francis Stewart*[B] *and Antony G. Mackinlay*

School of Biochemistry, University of New South Wales, P.O. Box 1, Kensington, N.S.W. 2033.
[A] Present address: School of Zoology, University of Cambridge, U.K.
[B] Present address: Institute for Cell and Tumour Biology, German Cancer Research Centre.

Abstract

The β-casein gene is a member of a small gene family encoding the calcium-sensitive caseins, which are specifically synthesized and secreted by the mammary gland during lactation in response to both peptide and steroid hormones. The caseins are involved in the transport of calcium phosphate in milk, which is important for bone development in the infant mammal. We report here the organization and complete DNA sequence of the 8·5 kb long bovine β-casein gene. Comparison with the rat β-casein gene reveals that the exons of both genes correspond exactly. The 5′ flanking sequences of all Ca-sensitive casein genes are conserved within the proximal 200 bp and contain several elements that probably function as *cis*-acting regulatory elements, including an octamer-like motif, an SV40-type core enhancer and a sequence that appears to be common to all lactoprotein genes. The latter sequence is flanked on either side by 12 bp direct repeats. These direct repeats are themselves each part of sequences that display two-fold symmetry. The first 30 nucleotides of the 3′ flanking regions in the bovine and rat β-caseins are well conserved, indicating that they are likely to be involved in the mechanism of 3′ end processing of the primary transcript.

Introduction

The β-casein gene encodes one of a number of proteins found in milk that are synthesized and secreted by the mammary gland during lactation. The bovine milk-specific proteins include the three calcium(Ca)-sensitive caseins (α_{s1}-, α_{s2}- and β-casein), κ-casein and the whey proteins α-lactalbumin and β-lactoglobulin.

In milk, the four caseins form loosely organized aggregates called casein micelles that are able to sequester and transport calcium phosphate at concentrations above its solubility product (Waugh 1971). This specific transport of calcium phosphate in milk by the caseins is essential to the newly born infant for bone development.

The cDNAs corresponding to the various lactoproteins have been characterized for a number of species (for a review see Bonsing and Mackinlay 1987) and from this work it is clear that the Ca-sensitive caseins are evolutionarily related and comprise a small gene family. In addition, the unrelated κ-casein gene is genetically linked to those of the Ca-sensitive caseins (Grosclaude *et al.* 1973, 1979). The expression of this gene cluster is coordinately induced in response to peptide and steroid hormones.

Currently, the genes encoding the lactoproteins in a number of species are being studied in order to better understand their tissue specific and developmentally regulated expression. The sequence of the rat β-casein gene plus its flanking regions (Jones *et al.* 1985) and the 5′ flanking sequences of the rat α-, γ- and bovine α_{s1}-casein genes have been published (Yu-Lee *et al.* 1986). Of the whey proteins, the α-lactalbumin gene has been studied in the rat (Qasba and Safaya 1984), human (Hall *et al.* 1987) and cow (Vilotte *et al.* 1987), the

0004-9417/88/040527$03.00

β-lactoglobulin gene has been studied in the sheep (Ali and Clark 1988) and the whey acidic protein (WAP) gene has been studied in the rat (Yu-Lee and Rosen 1983) and mouse (Campbell *et al.* 1984).

In this paper, we report the characterization and complete DNA sequencing of the bovine *β*-casein gene, 1722 bp of the 5′ flanking region and 117 bp of the 3′ flanking region.

Materials and Methods

The isolation and identification from a λEMBL3 library of clones with bovine casein genomic inserts has previously been described (Yu-Lee *et al.* 1986). Large scale preparation, by plate lysis, and purification of λ-clone DNA were essentially as described by Maniatis *et al.* (1982).

Fragments of the λ-clones were isolated from low melting agarose and ligated into pUC 8 according to the method of Crouse *et al.* (1983) and pUC sub-clone DNA was prepared using the method of Ish-Horowicz and Burke (1981), scaled up to employ 100 ml cultures.

Restriction analysis, hybridization studies and the subcloning of DNA fragments into M13 vectors used the standard techniques (Maniatis *et al.* 1982). Dideoxy sequencing of M13 clones with both the Klenow fragment of *E. coli* DNA polymerase (Promega Biotec) and T7 DNA polymerase (Pharmacia) utilized both universal and specific oligonucleotide primers based upon the *β*-casein DNA sequence, synthesized on an Applied Biosystems DNA Synthesizer 380B.

Sequence analysis was performed with the aid of the COMPARE and ALIGN programs, contained in the GENEUS package described by Harr *et al.* (1986) and NAAS programs (Genesearch, Broadbeach, Queensland).

Results

Restriction and Sequence Analysis

Screening of the bovine genomic library as previously described (Yu-Lee *et al.* 1986) led to the isolation of two clones that hybridized specifically to the bovine *β*-casein cDNA pB*β*C468 (Stewart *et al.* 1987). Upon restriction analysis, these two clones, λB*β*C8 and λB*β*C12, were found to be independent but overlapping clones and both clones were found to hybridize to end-labelled restriction fragments representative of the entire mRNA. The clone λB*β*C8 contained 1·2 kb more sequence at the 3′ end of the gene than did λB*β*C12, whereas the latter clone contained 2·5 kb more sequence at the 5′ end of the gene than did λB*β*C8.

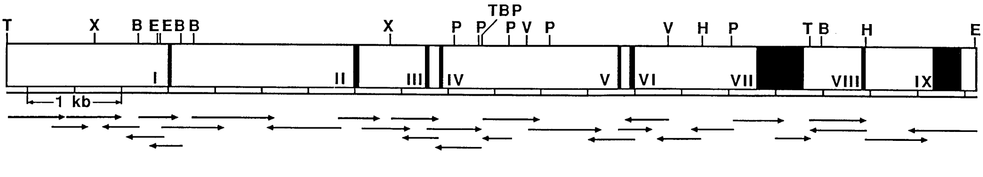

Fig. 1. Restriction map of the bovine *β*-casein gene and sequencing strategy. Exons are shown as black bars and are numbered. Arrows represent the extent and direction of sequence obtained from M13 subclones. Restriction sites are indicated by vertical lines. Key: B = Bgl II, E = Eco RI, H = Hind III, P = Pst I, T = Taq I, V = Pvu II, X = Xba I. A scale in 0·5 kb divisions is located below.

Restriction of λB*β*C8 with Taq I produced three insert-containing fragments, which were ligated into pUC 8 for further studies. Hybridization studies showed that the entire *β*-casein gene was contained within an 8·6 kb region defined by Eco RI restriction sites. The restriction map obtained for the bovine *β*-casein gene including the positions of the exons and the sequencing strategy employed is shown in Fig. 1. The complete DNA sequence of the bovine *β*-casein gene, 1722 bp of the 5′ flanking region and 117 bp of 3′ flanking region is presented in Fig. 2. The start of exon I is based on two independent *β*-casein cDNAs (Stewart *et al.* 1987).

```
                  -1700                                             -1650
TCGAATCCATCTCTATCAATTAATGTAATTCAAAATTGGTGAGAGACAGTCATTAGGAAATTCTCTGTTTATTGCACAAT
                                      -1600
ATGTAAAGCATCTTCCTGAGAAAAGGGAAATGTTGAATGGGAAGGACATGCTTTCTTTTGTATTCCTTTTCTCAGAAATC
        -1550                                             -1500
ACACTTTTTTGCCTGTGGCCTTGGCAACCAAAAGCTAACACATAAAGAAAGGCATATGAAGTAGCCAAGGCCTTTTCTAG
                            -1450
TTATATCTATGACACTGAGTTCATTTCATCATTTATTTTCCTGACTTCCTCCTGGGCCATATGAGCAGTCTTAGAATGAA
  !                                             -1350
TATTAGCTGAATAATCCAAATGCATAGTAGATGTTGATTTGGGTTTTCTAAGCAATACAAGACTTCTATGACAGTGAGAT
                  -1300                                             -1250
GTATTACCATCCAACACACATCTCAGCATGATATAAATGTAAGGTATATTGTGAAGAAAAATTATCAATTATGTCAAAGT
                                      -1200
GCTTACTTTAGAAGATCATCTATCTGTCCCAAAGCTGTGAATATATATATTGAACATAATTAATAGACGAAACAAACCTT
        -1150                                             -1100
GTAAAAATGAGTAGTGTAAAATACAACTACATTTATGAACATCTATCACTAAAGAGGCAAAGAAAGTTGAGGACTGCTTT
                            -1050
TGTAAATGGGCTCTTATTAATGAAAAGTACTTTTGAGGTCTGGCTTAGACTCTATTGTAGTACTTATGGTAAGACCCTCC
  !                                              -950
TCTTGTCTGGGCTTTCATTTTCTTTCTTCCTTCCCTCATTTGCCCTTCCATGAATACTAGCTGATAAACATTGACTCACT
                   -900                                              -850
ATAAAAGATATGAGGCCAAACTTGAGCTGTCCATTTTAATAAATCTGTATAAATAATATTTGTTCTACAGAAGTATCTCT
                                       -800
AAATAAATGTACTTTCTCTCTTAAAATCCCTCAACAAATCCCCACTATCTAGAGAATAAGATTGACATTCCCTGGAGTCA
         -750                                              -700
CAGCATGCTTTGTCTGCCATTATCTGACCCCTTTCTCTTTCTCTCTTCTCACCTCCATCTACTCCTTTTTCCTTGCAATA
                             -650
CATGACCCAGATTCACTGTTTGATTTGGCTTGCATGTGTGTGTGCTGAGTTGTGTCTCACTCTTGTCAACCCCATGAATG
  !                                              -550
ACAGTCCACCAGGCTCCACTATTTCCAGTTAAGAATACTGGAGTGGATTGTGTTTCCTACTTCATTTGATTAATTTAGTG
                   -500                                              -450
ACTTTTTAAATTTTTTTCCATATTCAGGAGGCTATTCTTTCCTTTTAGTCTATACTGTCTTCGCTCTTCAGGTCTAAGCT
                                       -400
ATCATCATGTGCTTGTTAGCTTGTTTCTTTCTCCATTATAGCATAAACACTAACAACTATTCAGGTTAGCATGAGATTGT
         -350                                              -300
GTTCTTTGTGTGGCCTGTGTATTTCTGGTGTGTATTAGAATTTACCCCAAGATCTCAAAGACCCACCGAATACTAAAGAG
                             -250
ACCTCATTGTAGTTACAATAATTTGGGGACTGGGCCAAAACTTCCGTGTGTCCCAGCCAAGGTCTGTAGCTACTGGACAA
  !                                              -150
TTTAATTTCCTTTATCAGATTGTGAATTATTCCCTTTAAAATGCTCCCCAGAATTTTTGGGGACAGAAAAATAGGAAGAA
                   -100                                               -50
TTCATTTTCTAATCATGCAGATTTCTA GGAATTCAAATCCACTATTGGTTTTATTTCAAACCACAAAATTAGCATGCCAT
                                        -1 EXON I
TAAATACTATATATAAACAACCACAAAATCAGATCATTATCC ATTCAGCTCCTCCTTCACTTCTTGTCCTCTACTTTGG
        INTRON I                                             100
AAAAAAG GTAAGAATCTCAGATATAATTTCATTGTATCTGCTACTCATCTTTATTTCAGACTAGGTTAAAATGTAGAAA
                               150
GAACATAATTGCTTAAAATAGATCTTAAAAATAAGGATGTTTAAGATAAAGTTTACAGTATTTTCAGCAAATTTGTTAAA
 200                                                250
AAATAGAAGCAACTATAAAGATTTGTAACAGTGGTTGCTATTTTCTTTACCACGAGACTAGTTAACAGGCTGTATTAAAA
                     300                                                350
GATCTTTTCTTGAATTAAATATTTTCAATTTGATTAAACATACCTCAGCCATAAAGGCAAGCACATTTAATTTATACTAT
                                         400
GGGAATTTGAATAATTGTTACTGAAGAAGCTCTACCAACAAAAAGTTTATAGAGCTAGCATATTTAGTCAAGAGATAAAG
           450                                                  500
AGGGTTGTTAGGATACATGTGCTATTTGAAAGGTATTTATAAAAGAAGAGTATATTTATTAAAATTGCTCAGAACATCCA
                               550
AATTTCAAGTTTATCATTTATCTTACAATATTTCAAAAATATTAAAATAGATACATGAAATACAGAAGTAAATTAAAGAG
 600                                                650
AAAGTATTTTATTTTGTAAAAAAAAATTCTAGGTTGGACAGGGAGTACCAGGAAACAAAAAACAATGAAAAATGTGATCT
                     700                                                750
GACAGAAATTATAGCTCAAAGTATAGTAGTCAGTAATGAAATGGCTTAAAAATTGGCATATAAAATGCTAATTATAAAAT
                                         800
AAACAAAATGTAATAATACCCTCCCTACATGTAATGAACTCTGAGTATTATACTCTTTTTTGAAGTCTTGACAATGAAAA

           850                                                  900
TTTATTTAGACTTTTATAGACATCTTGGATAAAGTAAAACAAATTACGAATTAGCATCCATGAGAAAAATATAGAAAAAT
                               950
TTCTTAATGTAGTTTGCAAATCTGGGGATTGAAGATGTGTGTCAAGAGATTGTGATGGCAGACATTTTTTTTCAGACTAT
1000                                               1050
AAAATGCACAAACAACCATTTAATACATTTTGGTCAAAAATAGTATGTATTTTATTTTATGCTACAGGAGAGTAGTCTAA
                    1100                                                   1150
AGTAGGACTGGGCAGAGATCTGACACCCTGGTAATCACCGAGAGATAGTACACAGTCTCTGTAGAGAAAATAAGCATAGT
```

```
                                            1200
GTATGATCTCTAAAATTATGTGGACAAAGGGGAGATAACATTAGGCATGTGGGGATGAAGACTGAGTACAGAAGAACAAT
           1250                                                      1300
CTAGTCAGTCCAAGAAAACATGTGGATCAATGGAACAAATAGAAGAAATGCTAAAATGAAACAGAAGTCTTACTGGAAAT
                                 1350
AAAAGATATGAGGAAGACAAACATTCATGAAAATCACTTAGTTTAGTAGAGAAAAGATAAAAATAAAGTATTACCTTCTT
1400                                                   1450
CTTCATATACATTGTTTGATCAGATGCCCCTCAATAAAACTGAGTCTCCAACAGAACTGAAACTTTAATATTTTGTTCAC
                      1500                                                     1550
TGCTCTAATCCCAGAATCTAAGACATATCTGGCAATAAAAATTAATAAATAAATATTTTTAATAAGTAAATCAATCACTT
                                           1600
AATTTTTCTGTAAGTATCTGTAACTTCTCTTCTGTCTTTCCAAAAAACACTCATAAGTACTGTGAATAAGATGAAAAGAG
           1650                                                      1700
TGAAATAAGATATAGGCTGTTAGCTGAAAACATCTGGATGGCTGGCAGTGAAACATTAACTTGAAATGTAAGATTAATGA
                                 1750
GTAATAGTAAATTTTAACCTTGGCCGTATGATAAAATGTCTATTAATATTTTTCTAAAATACAGGGCTTTTTGTTTTTGC
1800                                                   1850
CATGAGGTTTGCAGGATCTTGGTTCCCTGATGAGGGATCAAACCTGGGCTCCCCTGGAAGCACGGAGTCTTAGATATTTG
                      1900                                                    1950
TATTATACACTATCTTTGGTTTCTTTTAAAGGGAAGTAATTCTACTTAAATAAGAAAATAGATTGACAAGTAATACACTA
                        EXON II                  2000
TTTCCTCATCTTCCCATTCCCAG GAATTGAGAGCC ATG AAG GTC CTC ATC CTT GCC TGC CTG GTG GCT
                                     Met Lys Val Leu Ile Leu Ala Cys Leu Val Ala
                        INTRON II
CTG GCC CTT GCA AGA GAG GTAAATACAGAAAAAATGTTGAAATAAATAAGACTAGTACTATCTGCTATGTGTAG
Leu Ala Leu Ala Arg Glu
 !                                                  2150
AAAATTCATTACCAACATTGTAAATGTATAAATAATGCACAATCTCAGATTTTTTTGAATGCTAAGAAAGTCATTTACG
                   2200                                                       2250
TTCATCCACTATCTCAGTAGTATCCTATGGGACCACAAGTCTGAGTCTAGTGCTTTCTATAGTATTGTACCATCTGTACC
                                         2300
ATCAATCCCTAAAGAAAAAAGAAAATAAACCAATAAGCAACAGACTAACAAGAAGGAACACAGATAAGAACAAAAAGTGA
        2350                                                   2400
GTAATATTGCATAAATACAATTGCATGCATATACAATCTAGATAAATATATCTTATTCCAGTGATGAAATATTTGTATCC
                              2450
CTTACTGTAGAGTGCTAGGTTTAGCTGTGTCTATTCAACACAGGATGATACTCCAGAGGATGGTATATCAGACAACAATA
 !                                                  2550
ATAAATATGTTCATAATTATAATAAAAAGTGTTCAGTAAAAATTAAAATAACTCCTTTTCTGTTACCCATAAAAACTCTT
                   2600                                                       2650
CATTAAAGTAAAACAAAAATATACTAATGAAAGTTACTAAATTTAAAAGACTCTCAAAAGACATATAACATTTTTATTTT
                                         2700
TCAGATTTGTGAAATAGATAGCTCTGAATAAAGCAAGTAAAAATTAGGTAGGAAAATATTTAATAATGAGTTGACTGTGG
        2750                 EXON III                            INTRON III
GAACTAAAGTGTTTTTTTTTCTCTTTAG CTG GAA GAA CTC AAT GTA CCT GGT GAG GTAAGATATTTTTAT
                             Leu Glu Glu Leu Asn Val Pro Gly Glu
                                         2850
ACAAAGAAAAAAATTAATTTAACTGTAAAATAGTAACAGTCTCTAATGATCTGGCAGAAGACTCAGCTAATTGTCAATTT
        2900      EXON IV                             INTRON IV   2950
TTATTTTTCCTTTATAG ATT GTG GAA AGC CTT TCA AGC AGT GAG GTAAGATAGTGTTCATTCAGAGGCAA
                  Ile Val Glu Ser Leu Ser Ser Ser Glu
                                         3000
TTTCCCAAATTTAGAGCAATAAAATGCTGTATTATCTTTTTGTGTTACATTAATGGCAACCCACTCCAGTATTCTTGCCT
        3050                                                   3100
GGAAAATCCCATAGAGGAGGAGCCTGGTAGGCTGCAGTCCACGGGGTCGCTAAGAGTCGGACAGGACTGAGCGACTTCCC
                              3150
TTTCACTTTTCACTTTCATGCCTTGGAGAAGGAAATGGCAACCCACTCCAGTTTTCTTGCCGGGAGAATCCCAGGGACGG

 !                                                 3250
GGGAGCCTGGTGGGCTGCCGTCTATGGGGGTCGCACAGAGTTGGACACGACTGAAGCGACTTAGCAGTAGCAGCAGCAGG
                   3300       ⇒⇒⇒⇒⇒⇒⇒⇒⇒⇒⇒⇒⇒⇒⇒⇒⇒⇒⇒                       3350
ACAGTTAAGGTTTCTCTAATAGCTCAGTTGGTAAAGAATCTGCCTGCAGTGCAGGAGGACCCTGGGTTCGATTCCTCAAG
                                         3400
ATCTGCAGGAGAAGGGATAGGCTACCCATTCCAGTGTTCTTTAAGCCAATGTGGCTATGTACTGACGGGTAACTATTGTC
        3450                                                   3500
AATTTCCACTCTGTATATTTAAAGGAATAAATGTGTAGAAGGTTTAATATTCTAGTAATTTCTAAATGGGTTTGTTATTT
                              3550
GAAATTGTGTCATTGTGCCCTGCTTTTTTTCCTTAATGAACTGTACAGTCCTCTTTTCTGTCTTGAACTTTCTATGTTAA
 !                                                 3650
CCTCCTTCCATGCTTTGGCATTTATTGAGCACTTTCTGTTTCAGACTTTGACTAGGAACTGCAGTACAAACTAGAAAGAG
                   3700                                                       3750
GGATGCCCTTTGTAAAGTGTGAGCAGACATGCAAATGGACATATTTTATTATTATACAAGCAATCCAGTACACAAGAGGC
                                         3800
AGTGAGAATGAGTGTAGTCCTAAATCTGCCTGGTGGAATGAGGTAGATAATAACCCTATGCCACTCTTTTTGGCTCTGTC
        3850                                                   3900
ATCAGCTGTTGGTTAAATAGTGCATCAATATACTTTGTCTCTTCACAAAGGTCAAAAGATGCTGCTGCTGCTGCTAAGTC
```

```
                         3950
GCTTCAGTCATGTCCGACTCTGTGCGACCTCATAGATGGCAGCCTGCCAGGCTCCCCCATCCCTGGGATTCTCCAGGCAA
 !                                         4050
GAATACTGGAGTGGGTTGCCATTTCCTATTCCAATGCATAAAAGTGAAAAGTGAAAGTGAAGTCGCTCAGTCGTGTCCGA
              4100                                 4150
CTCCTAGTGACCCCGTGGACTGCAGCCTACCAGGCTCCTCCATCCATGGGATTTTCCAGGCAAGAGTACTGGAGTGGGTT
                                 4200
GCCATTGCCATCTCCAAAGATCCTTATAGGAGGGTATGTATTTCTTATAATTCATTAGAAGCCTAAACATAACCAGGGAA
       4250                                       4300
CTAAGAATGATTAACAAGTTCATGCGTGGTTATTATTATATATTTTCAATGACTATTATTCTTTTAGAACCAGATGAAAA
                         4350
TATTAGAGATCATTTGTTGTCCTAAGGAGAGAACAGGATGATTGAGAGACATGTATGCATGCAAAGTTACTTCAGCCCTT
 !                                         4450
TCCAACTCTGTATGACCCTATGGACTGTAACCTGCCAGGTTCCTCTGTCTATGGGATTCTCCAGGCAAGAATACAGGAGT
                4500                       ⇐⇐⇐⇐⇐⇐⇐⇐⇐⇐⇐⇐⇐⇐⇐⇐⇐⇐⇐⇐
GGGTTGTCATCTCCTCCAGGGCCTAGGAATTGACCTGCATCTCTTACGTCTCCTGCATTGGCAGGCAGGTTCTTTACCAC
                                 4600
TAGCACCACCTGGAAGCCCGATTAGTATCTGTTAAATGCCTCTTTGAGTACTATGCTCTCTAATCCTCTTTTTCTTGATT
       4650                                       4700
GCATCATCTTTCTTTTTATACAACAGCCTTATTCAGAAGAGTGGAACATAAACTTTCAGCCATAAAATAATGATATTATC
                         4750
AAATGAGCTGTCCATATTAATCTATTAAATTTCTTCATTTTCTGATTTATGTTGACAAATAAGAATTTTTTTAAAGCTA
 !                            EXON V                  4850    INTRON V
GACCTGATTTTATTTTTATTTTTCCAAAG GAA TCT ATT ACA CGC ATC AAT AAG GTAAAACCCCTCATATTT
                              Glu Ser Ile Thr Arg Ile Asn Lys
                        4900                                                     EXON
AAATGTACATTTTTTTAAATTTCATGTTTGATTTTTATAAACAGCATTTCTTTATGTATTTTTTTTTAACCAG AAA A
VI !                                                                      Lys I
                                                   INTRON VI  5000
TT GAG AAG TTT CAG AGT GAG GAA CAG CAG CAA ACA GAG GTAATTTGTTCACTATGAGTATATTTTGA
le Glu Lys Phe Gln Ser Glu Glu Gln Gln Gln Thr Glu
                            5050
GAAGTATTATGAAACATAACACATAAAAAGATTTATAATAATTATGTTCAGTCTAAGAATGGTAATATAAATGTCAGTGC
  5100                                        5150
AAGAAATAAAAACTTTGACAAAATGAAAATATTTTAAAAATATAAACGCATTTTAAAACACATAATCAAATTTCACAGTA
                   5200                                           5250
TAGAATAAATAGCTAAGAATAATTATTGATGTATTCATTTTACTAATGGTATACCTGGTTTTAATAACTGCATATTAGTA
                                     5300
GGAACATTTCCAGACTAGGGACTGTGATCCCCTTATTCTAATGATGGATATGCTGATGAAAGACAGTAGGGTGACAGTGT
           5350                                       5400
GGCACTAATCCTCATCTGATCATTTATCAGCTGTATAACCTTGGCTCCATGTTTCTCTGTACATCATTTTCTTCACCTGT
                            5450
AAATTGAGAATATTCATAATTACCCAGAGTTGATGAACTGACACACAATGAATATTCAGTGGTTTTATATTATATTTGAT
  5500                                        5550
AGCTTTTATACTCACATTTATGGATGTGTGGAGTTCTAAAAAGTATTTCCATTGCCCAGATGAGAGAAGTGAGGTACAGG
                   5600                                           5650
ACAATTGAGTATGCAAATGTGTGACCATACCACATAGTTATTAAATAGCAGAACTTGCTTAAAAACAAGGATTTGCGGAC

                                     5700
AATGTAAAATTCTTTCATTATATTACTCTTGTGGTAACATATTTATCTAATTATGATATTTAAGCTTTCCTCTTTTATAA
           5750                                       5800
TTGAAGTTTGATTGTTTGGCACTTAGGCCAAATTCTAAATCAAAATGAATTTACAACTTGATGCCTTTGAAGACTCAAGA
                            5850
TTACCACCTTCTACCAAGAGAAGTAGTGCTAGAAGTTGGCCATTGTTAAGGAACTCCTTGAATTAAAAAAACACATATTA
  5900                                        5950
AGACTTAGTTTTCATTAAAACAAACAAAAATAAACCTCAGAGTAACTTTTAAAGTCTTTTTAAAATGGATCTTTCTTTGT
                   6000                                           6050
TATATGAAACCAGTTTGGACTATTATCCAAAGTATGTAGCTACCACTCTGCAGGAACTCAGGAAGAGGTGGAATAAGTGT
                                     6100
TGAAATCTCCAAACCCTGATTTCACTTGACTCTCTGATTTCACCTGTGAAGAAAGTGGGTTAATGAGAAATCCTTCAGTG
           6150                                       6200
AGCATTTTACTCATTAGTCTTCATATGACCCCAATTTCTTAACCAAACCAAATGGAAGATTTTCTTTCTCTCTCTTCACT
                            6250
GAATTATGTTTTAAAAAGAGGAGGATAATTCATCATGAATAACAATTATAACTGGATTATGGACTCAAAGATTTGTTTTC
  6300      EXON VII                                               6350
CTTCTTTCCAG GAT GAA CTC CAG GAT AAA ATC CAC CCC TTT GCC CAG ACA CAG TCT CTA GTC
            Asp Glu Leu Gln Asp Lys Ile His Pro Phe Ala Gln Thr Gln Ser Leu Val
                       6400  *
TAT CCC TTC CCT GGA CCC ATC CAT AAC AGC CTC CCA CAA AAC ATC CCT CCT CTT ACT CAA
Tyr Pro Phe Pro Gly Pro Ile His Asn Ser Leu Pro Gln Asn Ile Pro Pro Leu Thr Gln
                                        6450
ACC CCT GTG GTG GTG CCG CCT TTC CTT CAG CCT GAA GTA ATG GGA GTC TCC AAA GTG AAG
Thr Pro Val Val Val Pro Pro Phe Leu Gln Pro Glu Val Met Gly Val Ser Lys Val Lys
                           6500
GAG GCT ATG GCT CCT AAG CAC AAA GAA ATG CCC TTC CCT AAA TAT CCA GTT GAG CCC TTT
Glu Ala Met Ala Pro Lys His Lys Glu Met Pro Phe Pro Lys Tyr Pro Val Glu Pro Phe
```

```
               6550
ACT GAA AGC CAG AGC CTG ACT CTC ACT GAT GTT GAA AAT CTG CAC CTT CCT CTG CCT CTG
Thr Glu Ser Gln Ser Leu Thr Leu Thr Asp Val Glu Asn Leu His Leu Pro Leu Pro Leu
6600
                                                                      6650
CTC CAG TCT TGG ATG CAC CAG CCT CAC CAG CCT CTT CCT CCA ACT GTC ATG TTT CCT CCT
Leu Gln Ser Trp Met His Gln Pro His Gln Pro Leu Pro Pro Thr Val Met Phe Pro Pro
                                                         6700
CAG TCC GTG CTG TCC CTT TCT CAG TCC AAA GTC CTG CCT GTT CCC CAG AAA GCA GTG CCC
Gln Ser Val Leu Ser Leu Ser Gln Ser Lys Val Leu Pro Val Pro Gln Lys Ala Val Pro
                                            6750
TAT CCC CAG AGA GAT ATG CCC ATT CAG GCC TTT CTG CTG TAC CAG GAG CCT GTA CTC GGT
Tyr Pro Gln Arg Asp Met Pro Ile Gln Ala Pro Gln Leu Tyr Gln Glu Pro Val Leu Gly
                           6800     INTRON VII
CCT GTC CGG GGA CCC TTC CCT ATT ATT GTAAGTCTAAATTTACTAACTGTGCCTGTTTAACTTCTGATGTT
Pro Val Arg Gly Pro Phe Pro Ile Ile
  !                                                50
TGTATGATATTCGAGTAATTAAGAGTCCTATAAAAAAATGAATAATGAATGGTTCCAAAATAAGCATAGCTGAGATTAAT
                   6950                                                  7000
GATTGTCAGCATTAGTTATAAATAGAATAAGCTGGAGAACCTTCACCTCCCCTCCACCACCAGATCTCAATGTCTAGGCT
                                         7050
TACCCGTGGAGATTCTGATGTAATTGTTCTTTCTATGTAGAAGAAACTTATTGGGAAGAAATAATATAATGGACTATGAT
           7100                                                7150
TTAATTGGTCTGTTGAGAACCAATTAAATTAGATGAAAGCGATTAAGTACAATAAAGCCAAAATTGAATTTGATAATCTC
                               7200
ATTTGGCTAAGAATAACAAACCTAAGAAGGTTTGCTATTTTCTACAATTTTGAAGTTCTCCTTATGCACAATTATTTCAC
  !                                                7300
CACATGACTCATTTCACATCGTGTTTTTGATATATGAGCATATGAGGGAAAAATACTGAGATGCTTATTTCAATACTCAG
                   7350                                                    7400    EX
GGAAAATTTATTGCCAAAAGGCAAGAAATGTATAATTCATTCACTTATTTTATTTTATTATTTTTTTTATTTTTAAG GT
                                                                                Va
ON VIII                                    INTRON VIII
C TAA GAGGATTTCAAAGTGAATGCCCCCTCCTCACTTTTG GTAAGCTTTAGGATATTGGAGGCAGACTGATCATTTT
l Stop
             7500                                              7550
TATAGTTAATATCTTTTACATTTCATTTTCCTGGATAAGCCCCAATAGTAGCAATTTCCATCAGTGTACCAGCTTAAAGA

                                  7600
TTAATTATAAATTTATTTTCAATGATTGACTGTTATTTACTGGCCTGAAATTATGTATCTGTTATATTTCAAATAATGCA
   7650                                                7700
AAACTGTATATATATGGTGTTTACAGATTTGATTGGTTTTCTTTCAATAGCCTATATCCTTATTATTGATTGTCATCATT
                       7750                                                  7800
TATAGAAAAAACTGAAAATAATTTCTTATACTTTTATGTAAACCTGTTAGAGCTTATTTTAAAGATCAACTGCATTCACA
                                          7850
TTTCTAATCTAGTCATTATGAGCTTCAATAGTTTTATCTCACTTAAAATATATATATTGTCTTTTAATTCATGAGTCAAA
             7900                                                7950
ATACAATCTCACAGTCCAGATATGGGACTTAAAAGGGGGATAGAATATAGTTTTGATATTCTTAACAATACACATCCTTT
                                  8000
TGTGATCATGATTCAGCAGACATTTAATAAAATGATTCCAAGTAAGCCGATGTTTGGTCCTAGAGGAATTTTTATAACCT
   8050                                              8100
TTAAGAGAAGGCATAGCATGGTGTTTTTGTAATAAGATTTCTTTTATGAAAAAGTCACACCAAAATTGCAAATGGGGGTG
                       8150                          EXON IX             8200
AGATGAAGAGTTATAACATATAACTAAATCTATGTTTGTTCTCTATTCCACAG AATTGACTGCGACTGGAAATATGGCA
                                           8250
ACTTTTCAATCCTTGCATCATGTTACTAAGATAATTTTTAAATGAGTATACATGGAACAAAAAATGAAACTTTATTCCTT
              8300                                               8350
TATTTATTTTATGCTTTTTCATCTTAATTTGAATTTGAGTCATAAACTATATATTTCAAAATTTTAATTCAACATTAGCA
                                  8400
TAAAAGTTCAATTTTAACTTGGAAATATCATGAACATATCAAAATATGTATAAAAATAATTTCTGGAATTGTGATTATTA
    8450                       poly (A) signal          3' FLANK
TTTCTTTAAGAATCTATTTCCTAACCAGTCATTTCAATAAATTAATCCTTAGGCAT ATTTAAGTTTTCTTGTCTTTATT
                         8550                                                8600
ATATTTTTTTTAATGAAATTGGTCTCTTTATTGTTAACTTAAATTTATCTTTGATGTTAAAAAGAGCTGTGGAAAATTAA

AATTGGATAGAATTC
```

Fig. 2. Complete DNA sequence of the bovine β-casein gene and flanking regions. Numbering is relative to the start point of transcription. The encoded amino acids are shown below the sequence. The retroposon elements are underlined and the inverted repeats identified by arrows above the sequence. The 67th codon of the mature protein coding region (denoted by an asterisk) specifies histidine, indicating that this sequence codes for the A_1 variant of bovine β-casein.

Organization of the Bovine β-casein Gene

The bovine β-casein gene is 8·5 kb in length and contains nine exons. The sequences associated with the splice sites conform to the consensus sequence (Mount 1982). Within the coding region, all splice junctions occur between codons.

The exons generally encode distinct portions of the mRNA and protein, as follows: the first exon contains the first 44 bp of the 5′ untranslated region. Exon II encodes the remaining 12 bp of the untranslated region, the entire signal peptide and the first two codons of the mature protein. The N-terminal hydrophilic region of the mature protein is encoded by four short exons (III to VI), arranged in pairs with short introns of about 100 bp between the exons of each pair, but with the pairs separated from each other by 1·9 kb.

The last four codons of exon IV and the first codon of exon V encode the amino acids Ser Ser Ser Glu Glu. The serine residues are post-translationally phosphorylated by a specific casein kinase that recognises the substrate Ser-X-Y, where X is any amino acid and Y is either Glu or Ser-P (Mercier 1981). Phosphorylation of these serines is crucial to the transport of calcium phosphate by β-casein and therefore the ability to correctly process the two exons involved must be essential for maintaining the gene in a functional state.

Exon VII, which is the longest exon in the gene (498 bp), encodes all but the last residue of the comparatively hydrophobic remainder of the mature protein. Exon VIII contains the last codon of the mature protein, the stop codon and the first 36 bases of the 3′ untranslated region. Exon IX comprises the remaining 322 bp of the 3′ untranslated region.

There are several repetitive DNA elements in the β-casein gene and its 5′ flank. All are examples of the A-family of artiodactyl retroposon (Rogers 1985) and contain at least one copy of the 117 bp consensus sequence defined by Watanabe *et al.* (1982). These elements are indicated in Fig. 2. One is found in the 5′ flank in the region between −650 and −539 and consists of the complement of the Watanabe consensus and with a $(CA)_5$ tail.

Intron IV contains three retroposon elements, the first two consisting of a pair of sequences conforming to the Watanabe consensus, linked by a sequence containing $(CACTTT)_n$ and tailed with $(AGC)_n$, features that are characteristic of these elements (Rogers 1985). The first is located 84 bp downstream of exon IV and the second, which is in the opposite orientation to the first and the Watanabe consensus, is located 966 bp downstream of exon IV. A third copy, consisting of a single Watanabe consensus sequence, is loacated 200 bp after the 3′ end of the second copy.

Another feature of intron IV indicated in Fig. 2 is the occurrence of two repeated sequences, 30 bp long and inverted with respect to each other, located about 30 bp after the first and third retroposon elements in the intron.

Discussion

Comparison of the Bovine and Rat β-casein Genes

When the bovine β-casein gene is compared to its rat counterpart, it becomes clear that very little change has occurred in this gene since the divergence of the two species from their common ancestor. The organization of the two genes is the same, with the exons of the two genes directly corresponding to each other. The overall sizes of the genes are similar, with the 7·2 kb long rat gene being 1·3 kb smaller. The sizes of the introns are generally similar, the main discrepancy being intron IV, which is about 0·9 kb longer in the cow. This largely accounts for the overall size difference between the two genes, and much of the increased length of intron IV in the bovine gene is due to the retroposon elements described above, which have a combined length of about 660 bp. A comparison of the intron sizes in the two genes is shown below:

	Bovine	Rat
Intron	Length of intron (bp)	
I	1935	1700
II	724	680
III	112	128
IV	1895	970
V	92	90
VI	1320	1107
VII	601	280
VIII	730	845

The sequences of the introns of the two genes display little or no similarity apart from the splice junction donor and acceptor sequences and two sections of conserved sequence, one about 50 bp long, found 556 and 334 bp upstream of exon II in the bovine and rat genes, respectively. The other comprises the first 75 bp following exon VIII.

```
                                                                ⇐⇐⇐⇐⇐
                                                                5'DIR
BOV αs1  -195         ATTT CCTGTATAA  TGAGTCACTTCTTTGTTGTAAACT CTCCTT
RAT α    -199         ATTTGCCATAATAACATGAATCACTCCTTTGTTGGAGACT TTACTC
BOV β    -199         ATTT CCTTTATCAGATTGTGAATTATTCCCTTTAA AATGCTCCCC
RAT β    -192         AGTT CCTTCACCAGCTTCTGAATTGCTGCCTTGTTTAATGTCCCCC
RAT γ    -188         ATTA TCTT ATCA  TGGCCTCAATCAAACGGTTTAAGAACTCCCT

              ⇐⇐⇐⇒⇒⇒⇒⇒⇒⇒⇒⇒ SIM. TO WAP                         ⇐
              REPEAT          ⇒⇒⇒⇒⇒              ⇐⇐⇐⇐⇐
BOV αs1  -152 AGAATTTCTTGGGAG AGGAACTGAACAGAACATTGATTTCCTA TGTGAGAGAA
RAT α    -153 AGAATTTCCCAGAAGAAGGAATTGGACAGAA ATTAATTTCCTA TTTGCAACAA
BOV β    -153 AGAATTTTT GGGGACAGAAAAATAGGAAGA ATTCATTTTCTAATCATGCAGAT
RAT β    -146 AGAATTTCTTGGGAA AGAAAATAGAAAGAA ACC ATTTTCTAATCATGGGAAC
RAT γ    -145 AGAATCT   GTGGA ACAAAATCCAGAGAG AC  AATTTCTAATGATATTGCT

              ⇐⇐⇐⇒⇒⇒⇒⇒
              3'DIR REPEAT                           CORE     OCTAMER
BOV αs1   -99 TTCTTAGAATTTAAATAAA CCTGGTTGGTTAAACTGAAACCACAAAATTAGCAT
RAT α    -100 TTCTTAGAATTTATGTAAAACCTTT  GTCTGAAACAAAACCACAAAATTAGCAT
BOV β    -100 TTCTAGGAATTCAAATCCA CTATT GGTTTTATTTCAAACCACAAAATTAGCAT
RAT β     -94 TTCTTGGAATT AAGGAA  CTTTT GAATATCTTACGAACCACAA ATTAGCAT
RAT γ     -97 TTCTTAGAATTCGAATGT  CTTTTTAGGTATTT  GAAACCACAGAATTAGCAT

                                TATA BOX                        EXON1
BOV αs1   -45 TTTACTAATCAG TAGGTTTAAATAGCTTGGAAGCAAAAGTCT     GCC ATC
RAT α     -47 TTCACTGCTCAG CAAGTTTAAATAGCTGTGGAGCAAACTTCT   CAGCC ATC
BOV β     -47 GCCATTAAATAC   TATATATAAACAACCACAAAATCAGATCATTATCC ATT
RAT β     -44 GTCATTAAGTAT   GGTATATATACAGTCACAGAGTCTGATAG     ACC ATC
RAT γ     -46 ATGATGCTAGAACCTGGTTTAAATAGTGCGGGAGCTACCCACT     GCT ATC
```

Fig. 3. Alignment of the 5′ flanking sequences of five Ca-sensitive casein genes. Sequences are identified at left. Numbering is relative to the start point of transcription and refers to the left-most nucleotide in each line. Sequence elements referred to in the text are underlined and identified above the alignment. Arrows indicate the extent and relative orientation of inverted repeats. 'Sim. to WAP' denotes the region that bears sequence similarity to the mouse whey acidic protein gene; the 3′ seven bp of this region are repeated at the 5′ extremity of these sequences (underlined; see text).

The 5′ Flanking Sequences

The three Ca-sensitive casein genes display a great deal of similarity in their 5′ flanking regions. This is consistent with their presumed evolutionary origins, i.e. gene duplication, and also may explain to some extent their co-ordinate expression at lactation. Previously, Yu-Lee *et al.* (1986), reported that the proximal 200 bp of flanking sequences are conserved both between species and between the different caseins. In this region, the bovine and rat β-caseins diverge by only 26%. Beyond this point, the sequences are more divergent; the −200 to −400 region of the bovine and rat sequences is 46% divergent and several gaps are required to be introduced in order to optimize the alignment.

Fig. 3 presents a comparison of the first 200 bp of the 5′ flanking sequences of the bovine and rat β-, bovine α_{s1}-, rat α- and rat γ-casein genes. Analysis of the five casein sequences within this region reveals a number of sequence elements that are likely to be important for the transcriptional regulation of these genes.

The sequence TATATATAAA is found in the bovine β-casein 5′ flank in the region between −35 to −26, part of which would function as the TATA motif for this gene. A feature of the β-casein genes is the difference in their TATA motifs as compared to the α-caseins ('tata boxes' in Fig. 3). Rat β-casein also has the sequence TATATATA in this region, but the sequence for the three α-caseins is TTTAAATA (Yu-Lee *et al.* 1986). The TATA motifs of the different lactoprotein genes vary a great deal. The rodent whey acidic protein gene motif is TTTAAAT (Yu-Lee and Rosen 1983; Campbell *et al.* 1984), that of the α-lactalbumin gene is TAAATAAAA (Qasba and Safaya 1984; Hall *et al.* 1987; Vilotte *et al.* 1987) and that of the ovine β-lactoglobulin gene is TATAA (Ali and Clark 1988).

The bovine β-casein gene, like the other casein genes studied, does not possess within its 5′ flank a recognisable CCAAT box (Breathnach and Chambon 1981).

The −65 to −45 region contains a sequence that is almost completely conserved between the five caseins for a length of 18 bp. The proximal eight bp of this sequence (ATTAGCAT, 'octamer' in Fig. 3) bear a remarkable resemblance to the octamer sequence (consensus ATTTGCAT), a factor binding sequence that functions in either orientation in many genes (see, e.g. Bohmann *et al.* 1987). The distal seven bp of the conserved sequence plus an additional unconserved residue ('core' in Fig. 3) are similar to the SV40-type core enhancer (consensus GTGG A/T A/T A/T G, Weiher *et al.* 1983), also found in many genes in either orientation. The sequence from −64 to −54 in bovine β-casein is also found in the region between the TATA box and exon I (−24 to −14), but this is not conserved in the other caseins.

The sequences within the region from −160 to −80 comprise inverted and direct repeats. Well-conserved 12 bp direct repeats are found centred at about −150 and −95 ('5′ dir repeat', '3′ dir repeat' in Fig. 3). From the five casein sequences analysed, the 5′ repeat is TC*CYYAGAAT*T*T*, the 3′ repeat is *TTCTT*R*GAATT*Y and the overall consensus is TY*C*TTA*GAAT*TT, the italics indicating complete conservation in each comparison. These two 12 bp direct repeats each also display dyad symmetry that in some cases extends beyond the limits of the repeats themselves. The extents of these inverted repeats are indicated in Fig. 3 with arrows.

Between the two direct repeats, in the region from −140 to −110 is found a sequence in which the outer 7 to 10 bp display complementarity (indicated by arrows in Fig. 3), and containing the tetramer AGAA in the non-complementary centre. The 3′ seven bp of this sequence are repeated further upstream, near −200. The 5′ portion of the sequence is AG-rich, and displays similarity to an element within the mouse WAP gene 5′ flank with the sequence AGAAGGAAGT. This element is protected from DNase I in footprint experiments using nuclear protein extracts derived from rat mammary cells (Lubon and Hennighausen 1987). Similar sequences within the α-lactalbumin and β-lactoglobulin 5′ flanks have also been reported (Lubon and Hennighausen 1987; Vilotte *et al.* 1987). This sequence may therefore be important in directing lactoprotein gene transcription at lactation.

Beyond the conserved 200 bp in the bovine β-casein 5′ flank is found a recognition sequence for the transcriptional factor AP-I (Lee *et al.* 1987), located at −931. AP-I sequences are also found within the gene, at 7252 (intron VII), 7875 (intron VIII) and 8318 (exon IX). The octamer sequence found at −55 is also found at 886 (intron I) and 8356 (exon IX).

The 3′ Flanking Sequences

In the region adjacent to the site of 3′ processing, the sequences of the two β-casein genes are very similar, from within exon IX to a point 30 nucleotides downstream of the processing site (13% divergence in the 3′ flank). Runs of T-residues, the trinucleotide TGT and the sequence TTTATT, located 17 nucleotides downstream of the processing site, are found in this region, as in many genes whose mRNAs are polyadenylated (Birnstiel *et al.* 1985). The proximal sequences of the β-casein 3′ flanking region are therefore likely to be important for 3′ processing in the production of the mRNA.

Conclusions

The structure of the β-casein gene is the least complex of the three members of the gene family. The α_{s1}- and α_{s2}-casein genes, which evidently share a common ancestor with β-casein, appear on the basis of cDNA comparisons (Stewart *et al.* 1984, 1987) to have undergone multiple major structural rearrangements, such as the duplication of exons and groups of exons, exon deletion and the recruitment of new exons. Thus, β-casein probably closely resembles the ancestral casein and during the duplications that led to the gene family, the additional members were able to acquire structural changes, while β-casein remained relatively unchanged in structure and function.

The conservation of one member of this gene family is probably important for micelle formation or structure; β-casein has been shown to be important in determining the surface properties of micelles (Pearse *et al.* 1986), and only the β- and κ-caseins are found in all milks studied (Jenness 1979).

Therefore it is expected that in any mammalian species, the β-casein gene would be found to be very similar to the two documented cases. Consistent with this is the amino acid sequence of human β-casein (Greenberg *et al.* 1984), from which it can be inferred that the protein coding region of its gene has not been subject to structural rearrangements.

Despite the dissimilar arrangements of the protein coding regions of the three members of the Ca-sensitive casein gene family, all possess very similar sequences within the proximal 200 bp of their 5′ flanking regions. Elements within their 5′ flanks are recognizably similar to *cis*-acting elements of other genes. These sequences, by virtue of their similarity and structural complexity, indicate the occurrence of multiple and concurrent interactions in those regions for all three genes that direct their co-ordinate, tissue-specific and developmentally regulated expression. In order to establish precisely where such interactions are occurring, we are currently using gel retardation and DNase I footprinting experiments with nuclear protein extracts obtained from lactating ewe udders.

Acknowledgments

We thank Dr L. J. Alexander for preparation and purification of oligonucleotide primers. This work was supported by the Australian Research Grants Scheme and the Australian Dairy Research Committee. J.B. is a recipient of a Commonwealth Postgraduate Award.

References

Ali, S., and Clark, A. J. (1988). Characterization of the gene encoding ovine beta-lactoglobulin. Similarity to the genes for retinol binding protein and other secretory proteins. *J. Mol. Biol.* **199**, 415–26.

Birnstiel, M. L., Busslinger, M., and Strub, K. (1985). Transcription termination and 3′ processing: the end is in site! *Cell* **41**, 349–59.

Bohmann, D., Keller, W., Dale, T., Scholer, H. R., Tebb, G., and Mattaj, I. W. (1987). A transcription factor which binds to the enhancers of SV40, immunoglobulin heavy chain and U2 snRNA genes. *Nature (London)* **325**, 268–72.

Bonsing, J., and Mackinlay, A. G. (1987). Recent studies on nucleotide sequences encoding the caseins. *J. Dairy Res.* **54**, 447–61.

Breathnach, R., and Chambon, P. (1981). Organization and expression of eukaryotic split genes coding for proteins. *Annu. Rev. Biochem.* **50**, 349–83.

Campbell, S. M., Rosen, J. M., Hennighausen, L., Strech, J. U., and Sippel, A. E. (1984). Comparison of the whey acidic protein genes of the rat and mouse. *Nucleic Acids Res.* **12**, 8685–97.

Crouse, G. F., Frischauf, A., and Lehrach, H. (1983). An integrated and simplified approach to cloning into plasmids and single-stranded phages. *Methods Enzymol.* **101**, 78–89.

Greenberg, R. M., Groves, M. L., and Dower, H. J. (1984). Human β-casein: amino acid sequence and identification of phosphorylation sites. *J. Biol. Chem.* **259**, 5132–8.

Grosclaude, F., Joudrier, P., and Mahe, M.-F. (1979). A genetic and biochemical analysis of a polymorphism of bovine α_{s2}-casein. *J. Dairy Res.* **46**, 211–13.

Grosclaude, F., Mercier, J.-C., and Ribadeau-Dumas, B. (1973). Genetic aspects of cattle research. *Neth. Milk Dairy J.* **27**, 328–40.

Hall, L., Emery, D. C., Davies, M. S., Parker, D., and Craig, R. K. (1987). Organization and sequence of the human α-lactalbumin gene. *Biochem. J.* **242**, 735–42.

Harr, R., Fallman, P., Haggstrom, M., Wahlstrom, L., and Gustafsson, P. (1986). GENEUS, a computer system for DNA and protein sequence analysis containing an information retrieval system for the EMBL data library. *Nucleic Acids Res.* **14**, 273–84.

Ish-Horowicz, D., and Burke, J. F. (1981). Rapid and efficient cosmid cloning. *Nucleic Acids Res.* **9**, 2989–98.

Jenness, R. (1979). Comparative aspects of milk proteins. *J. Dairy Res.* **46**, 197–210.

Jones, W. K., Yu-Lee, L.-Y., Clift, S. M., Brown, T. L., and Rosen, J. M. (1985). The rat casein multigene family. Fine structure and evolution of the β-casein gene. *J. Biol. Chem.* **260**, 7042–50.

Lee, W., Mitchell, P., and Tjian, R. (1987). Purified transcription factor AP-1 interacts with TPA-inducible enhancer elements. *Cell* **49**, 741–52.

Lubon, H., and Hennighausen, L. (1987). Nuclear proteins from lactating mammary glands bind to the promoter of a milk protein gene. *Nucleic Acids Res.* **15**, 2103–21.

Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982). Molecular Cloning: A Laboratory Manual. Cold Spring Harbor, N.Y.

Mercier, J.-C. (1981). Phosphorylation of caseins: present evidence for an amino acid triplet code posttranslationally recognized by specific kinases. *Biochimie (Paris)* **65**, 499–560.

Mount, S. M. (1982). A catalogue of splice junction sequences. *Nucleic Acids Res.* **10**, 459–72.

Pearse, M. J., Linklater, P. M., Hall, R. J., and Mackinlay, A. G. (1986). The effect of casein composition and casein dephosphorylation on the coagulation and syneresis of artificial micelle milk. *J. Dairy Res.* **53**, 381–90.

Qasba, P. K., and Safaya, S. K. (1984). Similarity of the nucleotide sequences of rat α-lactalbumin and chicken lysozyme genes. *Nature (London)* **308**, 377–80.

Rogers, J. H. (1985). The origin and evolution of retroposons. *Int. Rev. Cytol.* **93**, 187–279.

Stewart, A. F., Bonsing, J., Beattie, C. W., Shah, F., Willis, I. M., and Mackinlay, A. G. (1987). Complete nucleotide sequences of bovine α_{s2}- and β-casein cDNAs: Comparisons with related sequences in other species. *Mol. Biol. Evol.* **4**, 231–41.

Stewart, A. F., Willis, I. M., and Mackinlay, A. G. (1984). Nucleotide sequences of bovine α_{s1}- and κ-casein cDNAs. *Nucleic Acids Res.* **12**, 3895–907.

Vilotte, J.-L., Soulier, S., Mercier, J.-C., Gaye, P., Hue-Delahaie, D., and Furet, J.-P. (1987). Complete nucleotide sequence of bovine α-lactalbumin gene: comparison with its rat counterpart. *Biochimie (Paris)* **69**, 609–20.

Watanabe, Y., Tsukada, T., Notaki, M., Nakanishi, S., and Numa, S. (1982). Structural analysis of repetitive DNA sequences in the bovine corticotropin-β-lipotropin precursor gene region. *Nucleic Acids Res.* **10**, 1459–69.

Waugh, D. F. (1971). Formation and structure of casein micelles. In 'Milk Proteins: chemistry and molecular biology.' Vol. 2. (Ed. H. A. McKenzie.) pp. 3–85. (Academic Press: New York.)

Weiher, H., Konig, M., and Gruss, P. (1983). Multiple point mutations affecting the simian virus 40 enhancer. *Science (Wash. D.C.)* **219**, 626–31.

Yu-Lee, L.-Y., and Rosen, J. M. (1983). The rat casein multigene family I. Fine structure of the γ-casein gene. *J. Biol. Chem.* **258**, 10794–804.

Yu-Lee, L.-Y., Richter-Mann, L., Couch, C. H., Stewart, A. F., Mackinlay, A. G., and Rosen, J. M. (1986). Evolution of the casein multigene family: conserved sequences in the 5′ flanking and exon regions. *Nucleic Acids Res.* **14**, 1883–902.

Manuscript received 16 May 1988, accepted 19 October 1988