

Isolation and characterization of the bovine κ -casein gene

Leeson J. ALEXANDER¹, A. Francis STEWART¹, Antony G. MACKINLAY¹, Tatiana V. KAPELINSKAYA²,
Tatiana M. TKACH² and Stanislav I. GORODETSKY²

¹ School of Biochemistry, University of New South Wales, Sydney

² N.I. Vavilov Institute of General Genetics, USSR Academy of Sciences, Moscow

(Received June 3/August 26, 1988) — EJB 88 0653

1. The bovine κ -casein gene has been isolated as a series of overlapping λ clones and shown to consist of five exons distributed over a total length of approximately 13 kb. Most of the mature protein-coding sequence is contained in a single large exon.

2. Approximately 65% of the gene has been sequenced together with portions of the 5'- and 3'-flanking sequences. The immediate 5'-flanking sequence contains several motifs which are characteristic of upstream regions including a TATA box, a CAAT box, a sequence similar to that recognized by transcription factor AP-1 and a purine-rich sequence resembling that found upstream in all other lactoprotein genes. Other possible regulatory sequences are found upstream of exon 4.

3. The organization of the κ -casein gene, together with its upstream sequence, confirms previous conclusions that it is unrelated to the calcium-sensitive-casein gene family to which it is linked. Evidence is presented which supports a previous suggestion that κ -casein and the fibrinogens are evolutionarily related.

4. Intron sequences contain several examples of the A family of the artiodactyl Alu-like repeated sequences, together with a single example of a C-family sequence. The remainders of the introns of the κ -casein gene, compared with the repeat elements and exons, are A + T-rich.

5. Among the λ clones isolated, representatives were found of the A and B genetic variants which can be distinguished by restriction-enzyme analysis. Several other examples of polymorphisms in the non-coding region were found.

The caseins are the predominant proteins in the milk of most species. As well as being the main source of amino acids for the suckling infant, the caseins serve to raise the calcium and phosphate concentrations in milk to levels well in excess of the solubility product of calcium phosphate by forming loosely ordered aggregates, termed micelles, which sequester calcium phosphate. Bone formation in the young animal thus depends on the ability of casein to transport appropriate quantities of calcium phosphate in milk.

In bovine milk more than 20 individual components can be resolved when whole casein is analyzed electrophoretically under dissociating conditions [1]. These result from post-translational modification and genetic variation of four primary translation products which correspond to α_{s1} -, α_{s2} -, β - and κ -caseins. The α_{s1} -, α_{s2} - and β -caseins are insoluble in the presence of calcium ion at the concentrations at which it occurs in milk and these caseins are referred to as the calcium-sensitive caseins. In contrast, κ -casein is insensitive to the presence of calcium and is referred to as the micelle stabilizer because it is essential for the formation of stable casein micelles [2]. Following ingestion, the caseins are immobilized in the stomach as a result of clot formation. This occurs when chymosin (rennin) or pepsin specifically cleaves a single Phe-Met bond in κ -casein to form insoluble para- κ -casein (105 residues) and a soluble macropeptide containing the C-terminal 64 amino acid residues of κ -casein [3].

Two common genetic variants of κ -casein, A and B, are known, along with a recently described third variant [4]. The A variant has a Thr at position 136 and an Asp at position 148 of the mature protein, while the B variant has Ile and Ala residues at these positions respectively [5]. Analysis of the patterns of inheritance of these allelic variants, together with those of the other caseins, indicates that the genes for all four caseins occur as a tightly linked cluster [6, 7]. Analysis of the bovine casein genes as the cDNA level, as well as of their homologues in other species, reviewed in [8], has confirmed that the genes for the three calcium-sensitive caseins are evolutionarily related and constitute a small gene family. It appears that the κ -casein gene, although linked to the other casein genes, is unrelated to them. It has been suggested on the basis of amino acid sequence similarities that κ -casein may be related to the fibrinogens [9].

In this paper we present the results of analyses of cloned genomic DNA sequences that span the bovine κ -casein gene. This is the first κ -casein gene to be reported. As a result of DNA-sequence analysis the sizes and arrangement of the exons are defined. Several nucleotide substitutions and restriction-site polymorphisms are identified which distinguish the A and B alleles.

MATERIALS AND METHODS

Restriction enzymes, Klenow fragment of *Escherichia coli* DNA polymerase and T4 DNA ligase were purchased from commercial sources and used in accordance with manufac-

Correspondence to A. G. Mackinlay, School of Biochemistry, University of New South Wales, P.O. Box 1, Kensington, Sydney, New South Wales, Australia 2033

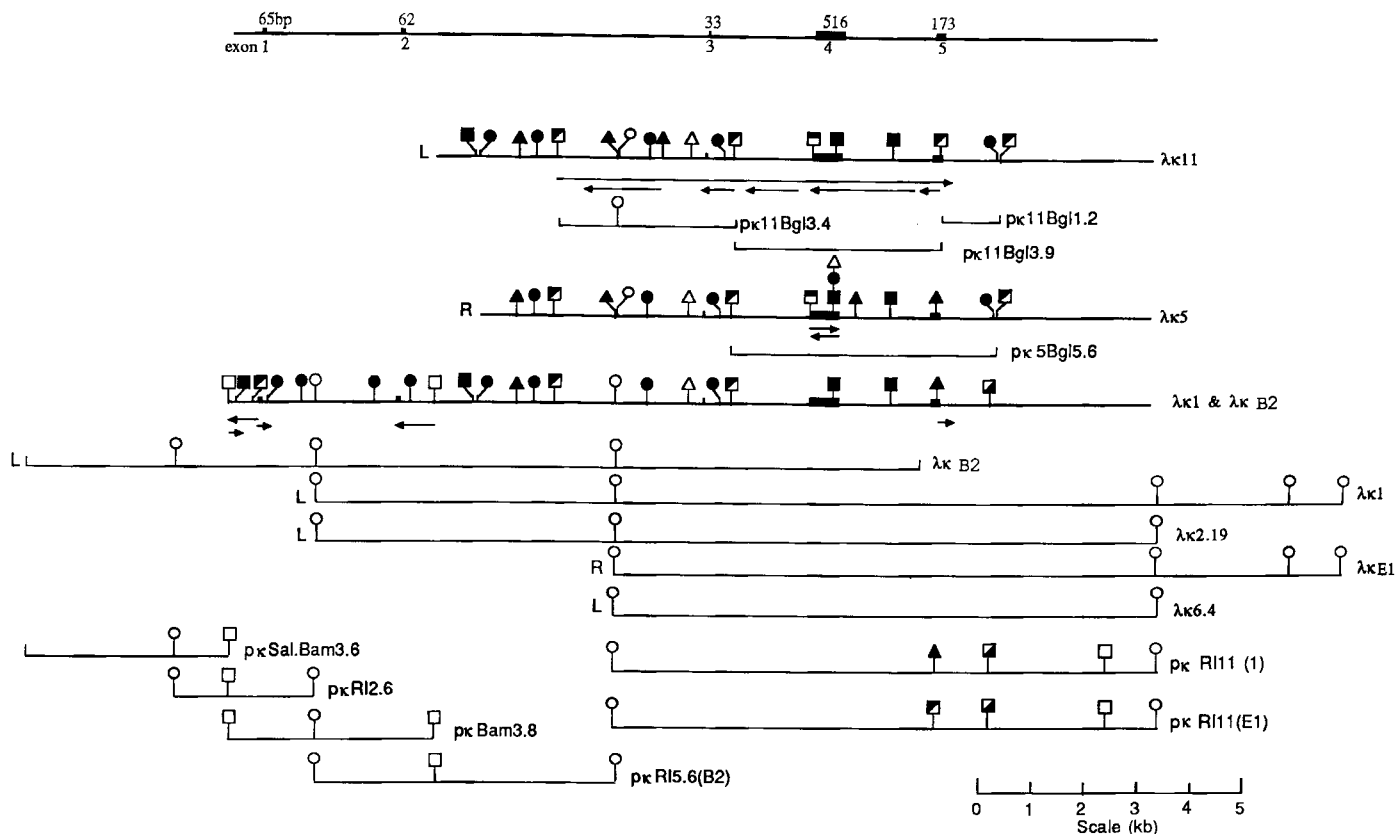


Fig. 1. Restriction maps of the cloned sequences spanning the bovine κ -casein gene. Arrows beneath the maps indicate those regions which were sequenced and the sequencing strategy employed. L and R refer to the left and right arms of the vector respectively. The names and positions within the gene map of various pUC plasmid subclones are also shown. The sizes and positions of the exons are shown at the top of the figure. Restriction sites for the following enzymes are shown: $\circ$ *EcoRI*, $\square$ *BamHI*, $\blacksquare$ *BglII*, $\boxplus$ *CfoI*, $\bullet$ *HindIII*, $\boxminus$ *KpnI*, $\blacktriangle$ *MspI*, $\triangle$ *TaqI* and $\blacksquare$ *XbaI*.

turer's specifications. Nitrocellulose was purchased from Schleicher & Schüll (FRG) and from Amersham. [α - ^{32} P]-dATP was purchased from Bresa, Adelaide, Australia and the Institute of Nuclear Physics, Tashkent, USSR. All chemicals were of analytical grade or the highest grade available. The construction, characterisation and sequences of the κ -casein cDNAs have been described previously [10–12]. Various oligonucleotides used for screening and sequencing were synthesized on an Applied Biosystems DNA synthesizer model 380B.

Standard recombinant DNA techniques were used, as described previously by Maniatis et al. [13]. Suitable restriction fragments were cloned into pUC8 and/or M13mp8. Ordered deletions were prepared by the rapid deletion method of Dale et al. [14] using a Cyclone kit purchased from I.B.I. (New Haven, CT, USA). DNA sequence determinations were performed by the dideoxy-chain-termination method of Sanger et al. [15, 16] and the chemical method of Maxam and Gilbert [17]. Double-stranded sequencing was performed following the protocol of Promega Biotec (Madison, USA). Autoradiographs were read manually and the sequence data analysed by the NAAS program (Genesearch, Broadbeach, Qld, Australia) on an Apple IIe personal computer and the Geneus program [18] on a Vax 11/785 mainframe computer. Of the sequence presented here, the majority was determined in both directions (see Fig. 1).

RESULTS

Isolation and characterization of λ clones $\kappa 1$, $\kappa 2.19$, $\kappa E1$, $\kappa 6.4$ and $\kappa B2$

Two bovine genomic DNA libraries in λ bacteriophage were used, one obtained by cloning DNA partially digested with *EcoRI* into λ EMBL 4, the second consisted of a partial *Sau3A* digest cloned into the *BamHI* site of λ EMBL 3. Four clones, $\lambda\kappa 1$, $\lambda\kappa 2.19$, $\lambda\kappa E1$ and $\lambda\kappa 6.4$ were obtained when the *EcoRI* library was screened as previously described [19]. These clones were restriction mapped using *EcoRI*, *BamHI*, *KpnI* and *MspI* in order to establish overlaps. The orientation of each clone was determined by hybridization, using labelled fragments from the right arm of λ and from the 3'-end of κ -casein cDNA. All four clones contained a common 11-kb *EcoRI* fragment which hybridised with a *PstI* fragment containing nucleotides 329–840 of the cDNA which includes the 3'-untranslated region. Two of the recombinants, $\lambda\kappa 1$ and $\lambda\kappa 2.19$, contained an additional 5.6-kb *EcoRI* fragment which hybridized with the 5' *PstI* fragment (residues 1–328 of the cDNA). The 5.6-kb fragment did not hybridise with a cDNA fragment corresponding to the 5'-untranslated end of the mRNA.

The second, *Sau3A*, library was then screened and among the clones which hybridized to the total cDNA, one clone, $\lambda\kappa B2$, was found to hybridise with a 180-bp fragment from

the 5'-end of the cDNA. It was shown to contain an additional 5.5-kb sequence at the 5' end of the gene compared to the cloned sequences obtained from the first library. In addition to the 5.6-kb *EcoRI* fragment, $\lambda\kappa B2$ contained a 2.6-kb *EcoRI* fragment which hybridized to the 5'-untranslated region of the cDNA. Restriction mapping and Southern blot analysis of DNA from $\lambda\kappa 1$ and $\lambda\kappa B2$ indicated that exon 1 is located in the 1.8-kb *HindIII* fragment.

More precise mapping was possible using subclones obtained in pUC19. Three fragments surrounding the 5' end of the gene were subcloned from $\lambda\kappa B2$: 3.6-kb *Sall*–*BamHI*, 2.6-kb *EcoRI* and 3.8-kb *BamHI* fragments. Investigation of these subclones has shown that exon 1 is located within a 289-bp *BglII*–*HindIII* fragment of plasmid p $\kappa R12.6$. Exon 2 is within a *HindIII* fragment (approx. 750 bp) of p $\kappa Bam3.8$ and is located 2.5 kb downstream from exon 1. Intron 2 is 5.8 kb long.

Isolation and characterization of clones, $\lambda\kappa 5$ and $\lambda\kappa 11$

The partial *Sau3A* bovine genomic library in EMBL 3A of Ruppert et al. [20] was screened as described previously [21] and two κ -casein genomic clones obtained, $\lambda\kappa 5$ and $\lambda\kappa 11$. Clone $\lambda\kappa 11$ is 12.7 kb in length and the entire insert can be excised intact with *BamHI*. Restriction mapping indicates that its 5' end may coincide with the *BamHI* site which is located 2.7 kb downstream from the start point of transcription (Fig. 1). Clone $\lambda\kappa 5$ contains an insert about 1 kb shorter than $\lambda\kappa 11$ and it is also excised intact by *BamHI*. Restriction mapping shows that $\lambda\kappa 5$ and $\lambda\kappa 11$ terminate at their 3' ends either at the same site or within 1–200 bp of each other. These were analysed by Southern blot analysis of restriction digests using the following end-labelled fragments derived from κ -casein cDNA [10] as probes: *HaeIII*–*HinfI* (nucleotides 19–180), *HaeIII*–*EcoRII* (19–105), *HinfI*–*HinfI* (180–249), *HinfI*–*HinfI* (250–575), *HinfI*–*HinfI* (576–820). These analyses indicated that most of the cDNA coding sequence was contained in both clones (Fig. 1). Representative of these results, three *BglII* fragments of size 3.9 kb, 3.4 kb and 1.2 kb from $\lambda\kappa 11$ were found to hybridize to one or more of the above probes. However, neither clone hybridized with an oligonucleotide complementary to nucleotides 21–40 of the cDNA sequence. Hence it was concluded that the 5' end of the gene was absent from both these clones.

Sequence analysis of κ -casein genomic clones

The nucleotide sequence of exon 1 and the 5'-flanking sequence was determined by Maxam and Gilbert sequencing, using an end-labelled *BglII*–*HindIII* fragment of p $\kappa R12.6$, as well as by dideoxy sequencing using a primer complementary to bases 21–40 of the cDNA. Another specific primer was used to obtain a sequence downstream of exon 1. Exon 2 was sequenced using the universal primer and an M13 subclone of p $\kappa Bam3.8$. Maxam and Gilbert sequencing of the pUC8 subclones p $\kappa 11 Bgl1.2$ and p $\kappa 11 Bgl3.9$ yielded sequence corresponding to the final 89 bases of the last exon, together with 127 nucleotides of a downstream sequence (nucleotides 13067–13193 in Fig. 2) as well as defining the 5' end of the last exon. Similarly, chemical sequencing of a 900-bp *MspI*–*KpnI* fragment from p $\kappa 6.4$ yielded a sequence from the same region. There were five differences found between the sequence data obtained from p $\kappa 11 Bgl1.2$ and that from p $\kappa 6.4$; these were T [13093 in p $\kappa 11 Bgl1.2$ (Fig. 2)] to C, T (13095)

to C, T (13103) to C, G (13153) to A and an insertion of a C residue after A (13169).

More extensive sequence analysis was carried out on $\lambda\kappa 11$ by subcloning the 3.9-kb and 3.4-kb *BglII* fragments into M13mp8 and subjecting these to stepwise deletion by the method of Dale et al. [14] and then performing dideoxy sequence analysis on the resulting series of overlapping clones. Both *BglII* fragments were completely sequenced by this method. These two *BglII* fragments were shown to be contiguous using double-strand sequencing on $\lambda\kappa 11$ and an oligonucleotide primer corresponding to residues 8948–8967. The 5' end of the 3.4-kb fragment is numbered +5600 in Fig. 2. Thus, the κ -casein gene was found to consist of five exons extending over a sequence of close to 13 kb, from the start point of transcription.

Exon 1 codes for 65 nucleotides of the 5'-untranslated region, exon 2 contains the remaining 5 nucleotides of the 5'-untranslated region and codes for 19 of the 21 amino acids of the leader peptide. Exon 3 contains only 33 nucleotides while exon 4 codes for the remainder of the mature protein as well as 37 nucleotides of the 3'-untranslated region, the remaining 173 nucleotides of which are contained in exon 5. Introns 1–4 are 2.5 kb, 5.8 kb, 2.0 kb and 1.8 kb in length respectively.

Sequence variation between κ -casein clones

Previous determinations of the κ -casein cDNA sequence have shown that the two common κ -casein variants, A and B, which are characterized by amino acid substitutions at positions 136 and 148, can be distinguished at the DNA level by two single nucleotide substitutions which give rise in the B variant to a *TaqI* and a *HindIII* site, respectively. Restriction mapping of the $\lambda\kappa 5$ and $\lambda\kappa 11$ clones, in conjunction with the sequence data for $\lambda\kappa 11$, showed that $\lambda\kappa 5$ and $\lambda\kappa 11$ correspond to the expected B and A variants, respectively. Double-strand sequencing using specific sequencing primers flanking the nucleotide substitutions at codons 136 and 148 has confirmed that $\lambda\kappa 5$ corresponds to the cDNA sequence determined for the B variant by Gorodetsky et al. [12]. Other substitutions in $\lambda\kappa 5$ include a G in place of an A at position 11006, resulting in loss of a *PstI* site, and also the presence of a T in place of A at position 11013. These substitutions were also found in the B variant cDNA [12].

Another polymorphism results from the substitutions at position 12981 of a T in $\lambda\kappa 11$ for a C in $\lambda\kappa 5$ which destroys the *BglII* site in $\lambda\kappa 11$ and creates an *MspI* site in $\lambda\kappa 5$. This polymorphism, previously detected in the course of cDNA sequence studies [10, 12], also distinguishes clones $\lambda\kappa 1$ and $\lambda\kappa E1$.

Repeated sequence elements in the κ -casein gene

A number of previously recognized repeat elements are found within the sequenced portion of the κ -casein gene. Several examples occur of the 117-bp consensus sequence of Watanabe et al. [22], i.e. the Alu-like A family of Rogers [23], as well as a single example of the tRNA-like C family. Nucleotides 5600–5623 at the 5' end of the 3.4-kb *BglII* fragment correspond to the 3' end of an A repeat (repeat 1). Repeat 2 (6114–6400) is a dimer of the repeated element and is flanked by a direct 11-bp repeat: CAATACAGCAAT and has the short repeating unit (AGC)₅ at its 5'-end. The repeats in this dimer are separated by a 50-bp spacer in which nucleotides 6255–6281 are also found in the spacer region of the dimeric repeat 4 (8161–8187). Repeat 3 (6757–6949)

[illegible][illegible]

Fig. 2

12550
 GTAAATGAAACCAAGTAATGAGTTGCACTATCTATTTTATTGATATGAAATACATTATTTACATGATTGTT
 12600
 TTAATAAATAATCATTATCTCATAGTCACATTTGTTCAAACTCAAGAACTAGATTCTAAGTTAAACACACACT
 12700
 ACTTATAAAAAAGAAATGATAGGACTGTGCTTCTGTGAATATTAGATGAAGACAATTGTATAAGATCATCAGTTC
 12750
 ACTATGTGGTGAATTTAAATATTAAATTTATATTTAGTTAGATAAACTACTTGGATACAGACATGAAGCAATCA
 12800
 CAGAAGCTAAATTAATACATATCATGAGAACTGTTAGCAATGTGAACTAAATCTCTAACTGTCTTTTATAG

Exon V
 12950
 GTCTAGCTGAAACCAATGACTACTTCAAACTTCTCTTGGCCAGTTGCTCTCGTGAACAGAGAAATGATTTT
 13000
 CACAGATCTGGCTCTCTCTCTCTCTTACATTTTACTTTTATGCCAGATTTAAATTTTTTATTGATTCGATAATA
 polyadenylation
 13100
 AAGCCAATCAATGC AATGGATCTCTTTCTTTTAAATTTATTTTAAATTTGGAGGCTAATTACTTTTCAATAT
 13150
 TGTATGGTCTACACACATCAACGTGAATCTGCCACGGGATACAGTGTCCCATCTCTGAAC

Fig. 2. Partial nucleotide sequence of the bovine κ -casein gene. The first nucleotide of exon 1 is designated +1 while the upstream sequence is numbered from -1. The start of exon 2 is set at +2600, while the first nucleotide of the Bg/II site at the 5'-end of the 3.4-kb Bg/II fragment is set at +5600. The amino acid sequence, for the A variant, is shown beneath the nucleotide sequence. Repeat elements are underlined. The polyadenylation site is shown at position 13066

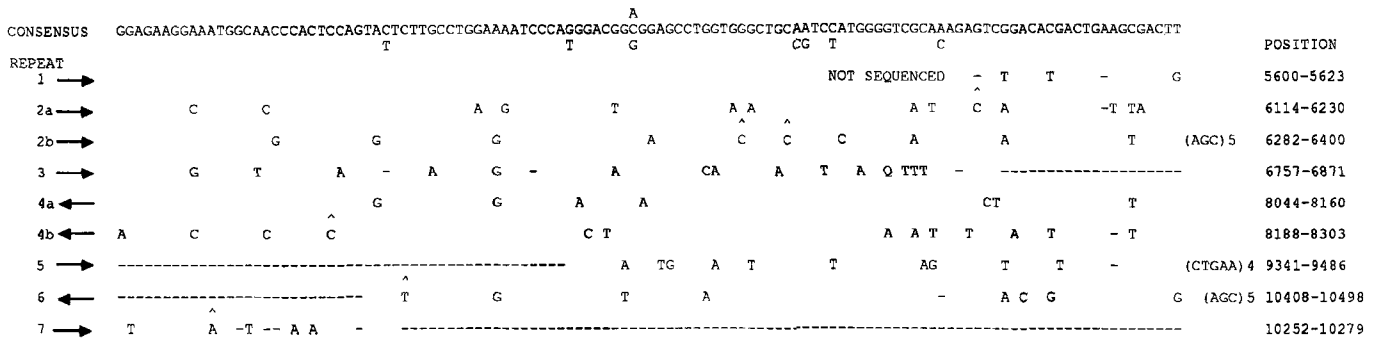


Fig. 3. Repeated sequence elements in the bovine κ -casein gene. The repeats belonging to the A family of Alu-like repeats are shown aligned with the consensus sequence of Watanabe et al. [22]. Numbers indicate position and arrows show the orientation in the sequence in Fig. 2. Key: - indicates a deletion; ^ indicates an insertion of the base below it; Q indicates the insertion of TCA

occurs adjacent to and downstream from a C-family-repeated element (6676-6756). Repeat 3 is flanked by a 10-bp direct repeat, AAAACTATTT. This element does not exhibit short repeats at its 3' end. Repeated element 4 (8044-8303) is present as a dimer in the opposite orientation, contains the conserved spacer region referred to above, and is flanked by an 11-bp direct repeat: TCACAGACTAT. No short repeats are present at the 5'-end. Repeated element 5, which is followed by the sequence (CTGAA)₄, contains a portion of an Alu-like monomer (positions 9427-9486) together with a copy of a repeated sequence element, termed the 73-bp repeat by Skowronski et al. [24]. The overall repeat shows 81% sequence identity with positions 3511-3690 in the 1.709-g/cm³ bovine satellite DNA [24]. No direct repeats are found adjacent to this repeated element. Repeat 6, 10408-10498 is a partial single element in the opposite orientation, situated 30 bp upstream of exon IV. It is similar to bases 26-117 of the consensus sequence and has the short tandem repeat (GCT)₅ at its 5' end. No direct repeats flank this element. Finally, the sequence 10252-10279 shows some similarity to the first 32 bases of the consensus sequence. The A-family-repeated sequences are shown aligned with the consensus sequence of Watanabe et al. [22] in Fig. 3.

DISCUSSION

This paper describes the structure of the first κ -casein gene to be isolated and characterized. These results show that although most of the sequence coding for the mature κ -casein molecule is contained within a single large exon, the overall length of the gene is close to 13 kb. The gene is divided into five exons which correspond approximately to the 5'-untranslated region, the leader peptide, a very short exon encoding the N-terminus of the mature protein, the large fourth exons

which extends 37 nucleotides into the 3'-untranslated region, while the remainder of the 3'-untranslated region is contained in the final exon. Splice junctions within coding sequences occur between codons and the dinucleotides GT and AG occur at the beginning and end of each intron. The sequence downstream of the polyadenylation site contains runs of T residues which are often observed in this position but not the G/T cluster with which the oligo(T) motifs are often associated [25]. Also present is the sequence TTTATT which is the complement of the AATAAA polyadenylation signal. This has also been observed in some other genes, including the gene for the γ chain of human fibrinogen [26].

Nucleotide 1 in Fig. 2 corresponds to the 5' end of the cDNA. Although it has not been shown directly that this is the start point of transcription, several lines of evidence indicate that it must be very close to this position. Previous measurements of the size of the mRNA are consistent with this conclusion [27]. Two independent cDNA clones beginning at this position were isolated from the cDNA library which had been constructed to facilitate generation of full-length clones and which has yielded other full-length cDNAs [28]. cDNAs for rat and mouse κ -casein have also been isolated and these are both shorter than the bovine κ -casein at the 5' end [29, 30]. Furthermore, the A + T-rich sequence TTAAATTA is found at positions -33 to -26, close to where the TATA motif is found in most other genes. In both the bovine α_{s1} - and rat α -casein genes the similar sequence, TTAAATA, is found at positions -29 to -22 in the one case and -31 to -24 in the other [21]. Also a CAAT element is found at positions -78 to -75, as observed in many other genes, although not in those for the calcium-sensitive caseins.

Because there is evidence for the linkage of all four casein genes in cattle [6, 7] and also because the calcium-sensitive caseins α_{s1} -, α_{s2} - and β -caseins, show a clear evolutionary relationship to each other, it is of interest to know whether

κ -casein may also be a member of this gene family or share common elements. Previous comparisons of cDNA sequences have failed to indicate any such relationship and the present data on the structure of the κ -casein gene reinforces this conclusion. A striking feature of the calcium-sensitive casein genes is that, in spite of the extreme divergence of their coding sequences and of the number and position of the exons which encode the mature proteins, there has been complete conservation of the positions within the genes of the first two and the last two exons. This arrangement is exemplified by that of rat β -casein [31], has been confirmed for bovine α_{s1} - and β -casein [21] (and Alexander, Bonsing, Clarke and Mackinlay, unpublished results) and inferred for bovine α_{s2} -casein [28]. According to this criterion, there is no discernible relationship between κ -casein and the other caseins.

Further support for this conclusion comes from the comparison of sequences upstream of the start point of transcription. A variety of conserved sequence motifs are found when the first 200 base pairs of the upstream sequences of the calcium-sensitive caseins are compared [21]. The corresponding portion of the κ -casein gene contains none of these, with the possible exception of a purine-rich sequence which appears to be common to most of the lactoproteins [32]. Two examples of this sequence are found in the κ -casein sequence, at positions -109 to -98 and -170 to -159, but both of these are in the opposite orientation to those seen in the other lactoprotein genes. It has to be concluded therefore that while the three calcium-sensitive casein genes arose as a result of the duplication of an ancestral sequence, the association of the κ -casein gene with these other genes is the result of the recruitment of an unrelated sequence into this linkage group. This may have been in response to a requirement for coordinate regulation of expression of all four genes.

Another sequence which could be involved in regulation of expression of the gene is the sequence TGACGCA which begins at -54. This is a 6/7 match with the sequence TGACTCA which has been shown to be recognized by the transcription factor AP-1 [33]. It is noteworthy that a 7/8 match to the complement of the immunoglobulin enhancer [34] occurs at almost this exact position in the upstream region of the calcium-sensitive casein genes [21].

Jollès et al. [9] have pointed out that κ -casein and the fibrinogens, especially γ -fibrinogen, show extensive sequence similarities at the amino acid level. κ -Casein and the fibrinogens display a number of functional parallels, e.g. their susceptibility to specific limited proteolysis with subsequent formation of a clot. Also, under certain conditions, e.g. in the presence of 50% alcohol, κ -casein exhibits fibrous characteristics [35]. Because of their similar functions, and also because κ -casein is evolving relatively rapidly [36], these sequence similarities could be the result of convergent evolution. One might however expect to see some conservation of splice sites if in fact the two proteins are derived from common ancestral sequences. The evidence on this point is persuasive, as shown in Fig. 4 which illustrates a region which is very similar in κ -casein and γ -fibrinogen, both at the nucleotide (17/24 identity) as well as the amino acid level (5/8 in the cow, 6/8 in the ewe and goat). Little or no similarity occurs on either side of this sequence [36]. The significant feature here is that the region of sequence similarity begins at the first residue of the κ -casein fourth exon and ends at one of only two splice junctions which are conserved in all three fibrinogen genes [37]. In order to account for the overall lack of coincidence of splice junctions in the members of the fibrinogen gene family it has been suggested [37] that the ancestral gene contained

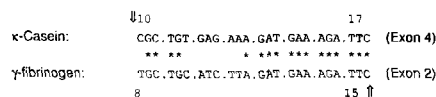


Fig. 4. Comparison of κ -casein and fibrinogen sequences. Codons specifying amino acid residues 10–17 of mature κ -casein are aligned with codons for residues 8–15 of human γ -fibrinogen. Arrows indicate splice sites. Fibrinogen data is from Rixon et al. [26]

many introns and that these have been lost randomly following duplication of the ancestral gene. If one accepts this suggestion, then it becomes likely that the 24-bp sequences in Fig. 4 correspond to one of the ancestral exons, the sequence of which has been conserved, both during the evolution of the fibrinogens as well as since the, presumably, more recent emergence of κ -casein. There are two other splice sites which occur in γ -fibrinogen, at amino acid positions 107/108 and 152/153, which coincide with the beginning of stretches of significant amino acid sequence similarity between κ -casein and γ -fibrinogen [36–38]. This evidence supports the view that κ -casein and γ -fibrinogen are evolutionarily related and that their divergence has been accompanied by the loss of introns and rearrangement of exons.

The overall base composition of the sequenced portion of the κ -casein gene is very A+T-rich. This is due mainly to the high A+T composition of the intron sequences (excluding repeated elements) which ranges over 69–75% A+T while the percentage A+T composition of the exons and the repeated elements ranges over 44–66%. Rogers [23] has noted that the Alu-type repeats tend to occur within A+T-rich sequences and this correlation appears to extend to the other bovine casein genes as well. Intron sequences in these genes are also A+T-rich and they have been shown to contain numerous copies of the A-family Alu-like repeat by probing restriction digests of genomic clones of these genes using a hybridization probe containing the sequences present in repeat 4 of the κ -casein gene (Alexander and Mackinlay, unpublished observations). A particularly A+T-rich sequence is found at positions 10920–11120. The sequence in this region appears to be a candidate for a matrix-attachment region [39, 40]. Downstream from this region, but upstream from exon 4, there is an extensive run of alternating purine and pyrimidine residues (11387–11441): AC(AT)₁₀(AC)₆(AT)₂(AC)₃-TC(AT)₄G. Such a sequence is known to facilitate the formation of the Z-DNA conformation [41] and its occurrence at this position is interesting in view of reports that control elements are often found close to matrix-attachment regions [39].

Clones $\lambda\kappa 5$ and $\lambda\kappa 11$ correspond in sequence and restriction sites to the expected B and A alleles, respectively. Southern blot analysis of *Hind*III and *Taq*I digests of genomic DNA should serve to genotype cattle for these variants. Rando et al. [42] have recently utilized the *Pst*I polymorphism at position 11006 to determine genotypes for κ -casein by this means. The size of the genomic *Pst*I fragments detected by Rando et al. [42] conform with the sequence data presented here. Accurate genotyping of κ -casein variants is likely to be of economic significance since a number of reports indicate that milk containing the B variant has improved processing characteristics, including better retention of fat in cheese and better cheese yield [43, 44]. The results presented here should assist future studies on the allelic variation of the κ -casein gene and may lead to the discovery of additional alleles not detectable by electrophoresis of the gene product.

The authors would like to thank Dr S. Ruppert for supplying the EMBL 3A library. This work was supported by grants to AGM from the Australian Dairy Research Committee and Australian Research Grants Scheme. LJA was supported by the Australian National Research Fellowship Scheme and AFS was a recipient of a Commonwealth Postgraduate Award.

REFERENCES

- Wake, R. G. & Baldwin, R. L. (1961) *Biochim. Biophys. Acta* **47**, 225–239.
- Waugh, D. F. & von Hippel, P. H. (1956) *J. Am. Chem. Soc.* **78**, 4576–4582.
- Mercier, J.-C., Brignon, G. & Ribadeau-Dumas, B. (1973) *Eur. J. Biochem.* **35**, 222–235.
- Seibert, B., Erhardt, G. & Senft, B. (1987) *Anim. Genet.* **18**, 269–272.
- Grosclaude, F., Mahe, M.-F., Mercier, J.-C. & Ribadeau-Dumas, B. (1972) *Eur. J. Biochem.* **26**, 328–337.
- Grosclaude, F., Mercier, J.-C. & Ribadeau-Dumas, B. (1973) *Neth. Milk Dairy J.* **27**, 328–340.
- Grosclaude, F., Joudrier, P. & Mahe, M.-F. (1979) *J. Dairy Res.* **46**, 211–213.
- Bonsing, J. & Mackinlay, A. G. (1987) *J. Dairy Res.* **54**, 447–461.
- Jollès, P., Loucheux-Lefebvre, M.-H. & Henschen, A. (1978) *J. Mol. Evol.* **11**, 271–277.
- Stewart, A. F., Willis, I. M. & Mackinlay, A. G. (1984) *Nucleic Acids Res.* **12**, 3895–3907.
- Gorodetsky, S. I., Kerhulite, D. R., Ivanov, V. N., Kapelinskaya, T. V. & Kaledin, A. S. (1985) *Genetika* **21**, 191–200.
- Gorodetsky, S. I. & Kaledin, A. S. (1987) *Genetika* **23**, 596–604.
- Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular cloning; a laboratory manual*, Cold Spring Harbor Laboratory, NY.
- Dale, R. M. K., McClure, B. A. & Houchins, J. P. (1985) *Plasmid* **13**, 31–40.
- Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl Acad. Sci. USA* **74**, 5463–5467.
- Sanger, F., Coulson, A. R., Barrel, B. G., Smith, A. J. H. & Roe, B. A. (1980) *J. Mol. Biol.* **143**, 161–178.
- Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
- Harr, R., Hallman, P., Haggstromm, M., Wahlstrom, L. & Gustafsson, P. (1986) *Nucleic Acids Res.* **14**, 273–284.
- Gorodetsky, S. I., Tkach, G. M. & Kapelinskaya, T. V. (1988) *Gene* **66**, 87–96.
- Ruppert, S., Scherer, G. & Schutz, G. (1984) *Nature (Lond.)* **308**, 554–557.
- Yu-Lee, L.-Y., Richter-Mann, L., Couch, C. H., Stewart, A. F., Mackinlay, A. G. & Rosen, J. M. (1986) *Nucleic Acids Res.* **14**, 1883–1902.
- Watanabe, Y., Tsukada, T., Notake, M., Nakanishi, S. & Numa, S. (1982) *Nucleic Acids Res.* **10**, 1459–1469.
- Rogers, J. H. (1985) *Int. Rev. Cytol.* **93**, 187–279.
- Skowronski, J., Plucienniczak, A., Bednarek, A. & Jaworski, J. (1984) *J. Mol. Biol.* **177**, 399–416.
- Birnsteil, M. L., Busslinger, M. & Strub, K. (1985) *Cell* **41**, 349–359.
- Rixon, M. W., Chung, D. W. & Davic, E. W. (1985) *Biochemistry* **24**, 2077–2086.
- Willis, I. M., Stewart, A. F., Caputo, A., Thompson, A. R. & Mackinlay, A. G. (1982) *DNA (N.Y.)* **1**, 375–386.
- Stewart, A. F., Bonsing, J., Beattie, C. W., Shah, F., Willis, I. M. & Mackinlay, A. G. (1987) *Mol. Biol. Evol.* **4**, 231–241.
- Thompson, M. D., Dave, J. R. & Nakhasi, H. L. (1985) *DNA (N.Y.)* **4**, 263–271.
- Nakhasi, H. L., Grantham, F. H. & Gullino, P. M. (1984) *J. Biol. Chem.* **259**, 14894–14898.
- Jones, W. K., Yu-Lee, L.-Y., Clift, S. M., Brown, T. L. & Rosen, J. M. (1985) *J. Biol. Chem.* **260**, 7042–7050.
- Lubon, H. & Hennighausen, L. (1987) *Nucleic Acids Res.* **15**, 2103–2121.
- Lee, W., Mitchell, P. & Tjian, R. (1987) *Cell* **49**, 741–752.
- Wirth, T., Staudt, L. & Baltimore, D. (1987) *Nature (Lond.)* **329**, 174–178.
- McKenzie, H. A. & Wake, R. G. (1961) *Biochim. Biophys. Acta* **47**, 240–242.
- Brignon, G., Chtourou, A. & Ribadeau-Dumas, B. (1985) *FEBS Lett.* **188**, 48–54.
- Crabtree, G. R., Comeau, C. M., Fowlkes, D. M., Fornace, A. J., Malley, J. D. & Kant, J. A. (1985) *J. Mol. Biol.* **185**, 1–19.
- Jollès, P. & Henschen, A. (1982) *Trends Biochem. Sci.* **7**, 325–328.
- Cockcrill, P. N. & Garrard, W. T. (1986) *Cell* **44**, 273–282.
- Gasser, S. M. & Laemmli, U. K. (1986) *Cell* **46**, 521–530.
- Nordheim, A. & Rich, A. (1983) *Nature (Lond.)* **303**, 674–679.
- Rando, A., Di Gregorio, P. & Masina, P. (1988) *Anim. Genet.* **19**, 51–54.
- McLean, D. M., Graham, E. R. B., Ponzoni, R. W. & McKenzie, H. A. (1984) *J. Dairy Res.* **51**, 531–546.
- Marziali, A. S. & Ng-Kwai-Hang, K. F. (1986) *J. Dairy Sci.* **69**, 1793–1795.

Note added in proof (received November 15, 1988). Repeat 5 is a truncated version of a highly repeated family of sequences which contain portion of the Alu-type repeat and which are described by Duncan, C. H. [*Nucleic Acids Res.* **15**, 1340 (1987)]. This repeated element begins at residue 9268 (Fig. 2).