

GENE 07695

Structure of the human β -casein encoding gene

(Milk protein gene; recombinant DNA; sequence homology; nucleotide sequence; exon-skipping)

Lennart Hansson^a, Anders Edlund^a, Thore Johansson^a, Olle Hernell^b, Mats Strömquist^a,
Susanne Lindquist^c, Bo Lönnerdal^d and Sven Bergström^c

^aSymbicom AB, S-90124 Umeå, Sweden; ^bDepartment of Pediatrics, University of Umeå, S-90187 Umeå, Sweden. Tel. (46-90) 102-121; ^cDepartment of Microbiology, University of Umeå, S-90187 Umeå, Sweden. Tel. (46-90) 101-963; and ^dDepartment of Nutrition, University of California, Davis, CA 95616, USA. Tel. (1-916) 752-8347

Received by G. Bernardi: 21 June 1993; Revised/Accepted: 23 September/27 September 1993; Received at publishers: 8 November 1993

SUMMARY

The entire human β -casein-encoding gene, *Bca*, was cloned and sequenced. The gene consists of eight exons ranging from 21 to 531 nucleotides (nt) in length and extending over 10 466 nt. Exon-2 contains the translational start, the entire signal sequence and the codons for the two first amino acids of the mature protein. This corresponds to the organization found in other species. The translational stop is localized to exon-7. Exon/intron boundaries are in accordance with the AG/GT rule and conform to suggested consensus sequences. Splice junctions are located between coding triplets. In all other species analyzed, *Bca* has been found to consist of nine exons; however, within intron-2 of the human gene, a sequence omitted from human mRNA, but corresponding to exon-3 of other known *Bca* genes, was revealed.

INTRODUCTION

Casein genes which expression is temporally and spatially regulated by various hormones (Devinoy et al., 1988; Jahn et al., 1989) encode the major milk proteins in most mammals. Human milk casein is composed of two species, β - and κ -casein, of which β -casein is the most abundant. Human β -casein is a 25-kDa protein which is present in several isoforms representing various degrees of phosphorylation (Greenberg et al., 1984; Hansson et al., 1993). Diverse biological effects have been ascribed to β -casein, but aside from enhancement of calcium absorption the experimental evidence for these effects is relatively weak (for review, see Miller et al., 1990).

Complete *Bca* cDNAs have been cloned and sequenced from eight different mammalian species; human (Lönnerdal et al., 1990), mouse (Yoshimura et al., 1986), rat (Blackburn et al., 1982), cow (Stewart et al., 1987), rabbit (Schaerer et al., 1988), sheep (Provot et al., 1989),

pig (Alexander and Beattie, 1992) and kangaroo (Collet et al., 1992). The *Bca* gene structure and sequence have been described for mouse (Yoshimura and Oka, 1989), cow (Bonsing et al., 1988), rat (Jones et al., 1986) rabbit (Thépot et al., 1991) and goat (Roberts et al., 1992). Partial characterization of the human *Bca* gene structure, focusing on the sequence corresponding to the third exon in the gene of other mammals which is located in the second intron, have indicated that exon-skipping is responsible for the deletion of a 9-aa sequence present in β -casein from other species (Martin and Leroux, 1992; Menon et al., 1992).

RESULTS AND DISCUSSION

(a) Isolation of the human *Bca* gene

The human *Bca* gene was cloned from a human genomic library, constructed from human placenta DNA in

Correspondence to: Dr. L. Hansson, Symbicom AB, Box 1451, S-90124 Umeå, Sweden. Tel. (46-90) 190-120; Fax (46-90) 192-332.

Abbreviations: aa, amino acid(s); *Bca*, gene (DNA) encoding

β -casein; bp, base pair(s); cDNA, DNA complementary to mRNA; kb, kilobase(s) or 1000 bp; nt, nucleotide(s); oligo, oligodeoxyribonucleotide; ORF, open reading frame; PCR, polymerase chain reaction; *tsp*, transcription start point(s).

λ EMBL-3 (Clontech, Palo Alto, CA, USA). A genomic clone containing the entire transcribed sequence of the human *Bca* gene was isolated using the previously cloned human *Bca* cDNA as a probe (Lönnerdal et al., 1990). Approx. 6×10^5 individual plaques were analyzed. One plaque hybridizing to the 5'- and 3'-sequences of the cDNA was detected and isolated. After several re-screening experiments, the recombinant phage DNA was purified (Sambrook et al., 1989). The purified DNA was digested with *Sal*I and a fragment of about 14 kb of genomic DNA was isolated by agarose electrophoresis, and cloned into *Sal*I linearized pUC19 plasmid. The resulting plasmid, called pS358, was characterized by restriction enzyme mapping (Fig. 1a).

(b) Structure and sequence of the human *Bca* gene

The exons were mapped by comparing the genomic sequence with the sequence of the human cDNA. The restriction map and the location of the exons and introns are presented in Fig. 1a. The approximate positions of the exons, and the approximate size of the introns were analyzed by PCR amplification using primers complementary to various sequences of the cDNA. The results from the analysis of plasmid pS358 were compared to results obtained with the same PCR primers using human genomic DNA as template. The results obtained with the two different templates were identical (data not shown).

In order to facilitate nt sequence analysis, 21 restriction fragments derived from pS358 were isolated and subcloned into pUC19, resulting in pS361–pS381 (Fig. 1b).

The orientation of the subcloned fragments was determined by PCR analysis. The inserts in the 21 plasmids pS361–381 were subjected to nt sequence analysis. The complete sequence of all subclones was determined using T7 sequencing kit (Pharmacia, Uppsala, Sweden; US Biochemical, Cleveland, OH, USA) on double-stranded templates. Synthetic oligos complementary to pUC19 or *Bca* sequences were used as primers for sequencing reactions. The analyzed human *Bca* genomic nt sequence consists of 10 600 bp (Fig. 2), and contains the complete exon and intron structure of the gene. The human *Bca* gene consists of 8 exons and 7 introns (Fig. 3), and the transcribed region extends over 10 466 nt (Fig. 2). The translational start is localized in exon-2, and the translational stop in exon-7. The exons are relatively small, size range is between 21–531 bp, and the introns range in size from 95–4670 bp (Fig. 3).

As can be noted from Table I, exon/intron boundaries are in accordance with the AG/GT rule and all but one splice junction conform well to the consensus sequence suggested by Mount et al. (1982). No coding triplets are disrupted by splicing sites. This integrity of codons seems to be well conserved among *Bca* genes from different species. No exception from this organization has been reported.

Sequence analysis of human *Bca* cDNAs from various libraries (obtained from Clontech; and cDNA libraries prepared from mammary epithelial cells from lactating Swedish women) has indicated a variation with respect to the presence or absence of a triplet (CAG) coding for

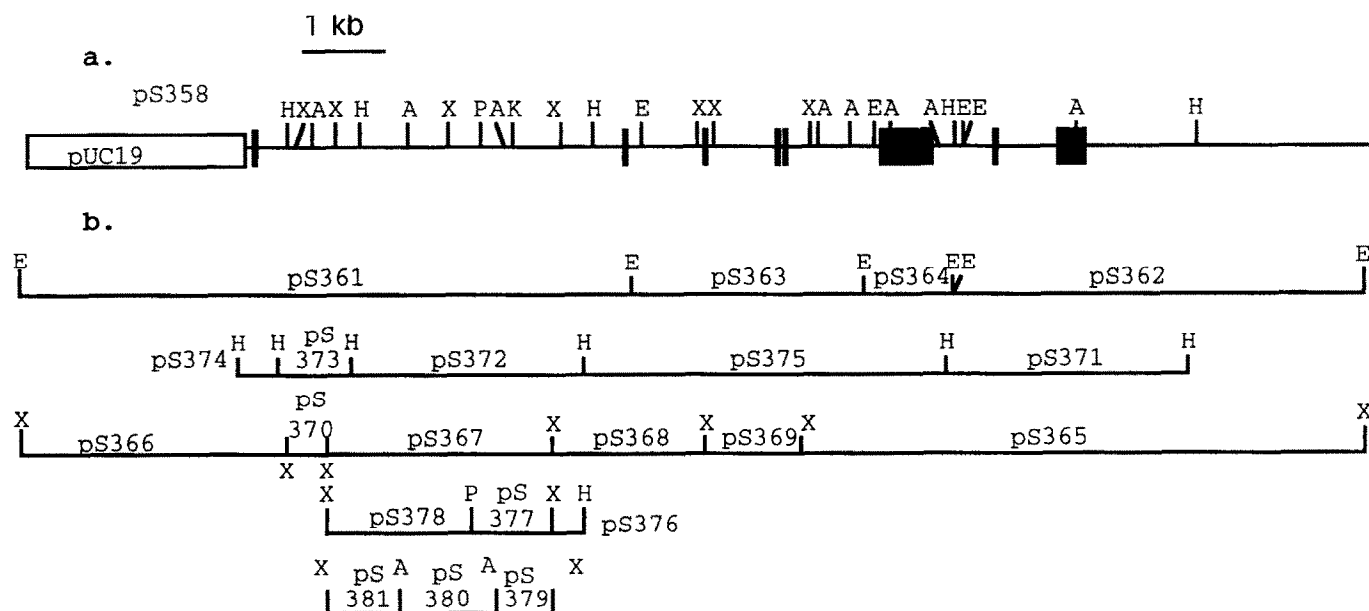


Fig. 1. Exon-intron structure and restriction map of the human *Bca* gene. (a) A map of the *Sal*I fragment isolated from the isolated λ EMBL3 phage and cloned into pUC19, resulting in plasmid pS358. The exons are shown as solid boxes. (b) Map of the 21 restriction fragments subcloned into pUC19 used for sequencing and restriction mapping of the gene. Two of the *Xba*I sites are Dam methylated and therefore not cleaved in these experiments. E, *Eco*RI; H, *Hind*III; X, *Xba*I; A, *Acc*I; P, *Pst*I and K, *Kpn*I.

TABLE I
Intron splice junctions of human *Bca* gene

Intron	Conserved nt ^a	Exon	Intron	Exon	Conserved nt ^a
1	7/9	A A C G	<u>G</u> T A A G A . . . T T T C T A C A T T C A C A <u>G</u>	G A C T	13/15
2	7/9	G G A G	<u>G</u> T A T G T . . . T T T T T T T T C C C A C A <u>G</u>	A C C A	14/15
3	8/9	T G A G	<u>G</u> T A A G T . . . T A T A T T T T T C T G T A <u>G</u>	G A A T	13/15
4	8/9	C A A G	<u>G</u> T A A A T . . . A T G A T A T A T C T T T A <u>G</u>	C A G A	9/15
5	7/9	A G A G	<u>G</u> T A A T T . . . T T T C C T T C T T T C C A <u>G</u>	G A T G	15/15
6	7/9	T A G T	<u>G</u> T A A G T . . . C T G T T T A T T T T T A A <u>G</u>	G T C T	12/15
7	7/9	T T T G	<u>G</u> T A A G T . . . C T C T C T A T T C C A C A <u>G</u>	A A T T	13/15
Consensus ^b		M A G	G T R A G T . . . Y Y Y Y Y Y Y Y Y Y N Y A G	G	

^aFraction of nucleotides conserved relative to Mount consensus sequence.

^bM, A or C; R, A or G; Y, C or T; N, A or C or G or T.

a Gln residue at position 19 in the mature protein (unpublished data). To explore whether this sequence variation reflects genetic polymorphism or stage-specific splice variants, *Bca* mRNA prepared from samples of human milk from different stages of lactation were sequenced. Milk samples, collected via a breast-pump, were obtained from healthy Swedish women. Milk from each donor (50 ml) was centrifuged to collect the free cells, which to a part are epithelial cells originating from the lactating mammary gland. Total RNA was extracted from the pelleted cells as previously described (Chomczynski and Sacchi, 1987; Puissant and Houdebine, 1990). The prepared total RNA was used for reverse transcriptase and PCR amplification. The amplified DNA fragments were then subjected to sequence analysis and the obtained results showed that these five females produced *Bca* mRNA containing the codon for the Gln¹⁹ residue. This aa residue was also reported by Greenberg et al. (1984), and the codon for this glutamine is included in the mRNA sequence submitted to GenBank (accession No. X17070) by R.S. Menon. The conclusion that this is the most abundant variant is further supported by the observation that *Bca* sequences obtained from mammary gland cDNA libraries prepared from Scandinavian females encodes this variant. At the gene level this difference can be explained by alternative splicing. The Gln codon, CAG, is located in the 5' end of exon-5. In the Gln¹⁹ variant exon-5 starts with the sequence 5'-CAG AAA GTT-3', whereas in the variant lacking Gln¹⁹ exon-5 starts with 5'-AAA GTT GAG-3'. In both variants the exon-5 splice sites conform to the Breatnach (GT/AG) rule and to the Mount consensus sequence (Mount et al., 1982). When the exon sequences obtained from the genomic clone were compared to the previously published cDNA derived sequence (Lönnerdal et al., 1990), only one additional difference was observed; while the nt 285 in exon-8 is C in the genomic sequence it is a T in the

cDNA (Fig. 2). However, this position is located downstream from the *Bca* ORF.

To map the *tsp* of the human *Bca* gene, primer extension analysis was performed using total RNA prepared from lactating human mammary gland. The result indicated that the *tsp* is 48-bp upstream from the first intron (Fig. 4). Based on sequence similarities, a TATA-box was found 30-bp upstream from the *tsp* (Fig. 2). Furthermore, a sequence showing a high degree of homology to the previously reported 'D' consensus sequence (Roberts et al., 1992) was found 33-bp upstream from the TATA-box (Fig. 2).

An analysis of the human *Bca* gene resulted in identification of several conserved intron sequences similar to those reported by Yoshimura and Oka (1989). In addition, long direct repeats of about 200 bp were identified located in intron-1 and 6 (Fig. 2).

(c) Comparison to other *Bca* genes

The *Bca* gene organization has been described for mouse (Yoshimura and Oka, 1989), cow (Bonsing et al., 1988), rat (Jones et al., 1986), rabbit (Thépot et al., 1991) and goat (Roberts et al., 1992). The *Bca* gene from all these species have nine exons and range in size from 6.8 kb to 9.9 kb. Despite the finding that the human gene consists of only eight exons, it is longer than other characterized *Bca* genes. In particular, the first intron is longer than what was found in other species (Fig. 3).

An alignment of the human β -casein deduced aa sequence with the corresponding β -casein aa sequences derived from other species, i.e., cow, sheep, rabbit, rat and mouse, shows extensive homologies over the entire proteins. Recently, it was shown that exon-3 is omitted in the human mature *Bca* cDNA (Martin and Leroux, 1992; Menon et al., 1992). In these studies it could not be excluded that exon-3 is under the control of alternative splicing and that only variants lacking exon-3 were se-

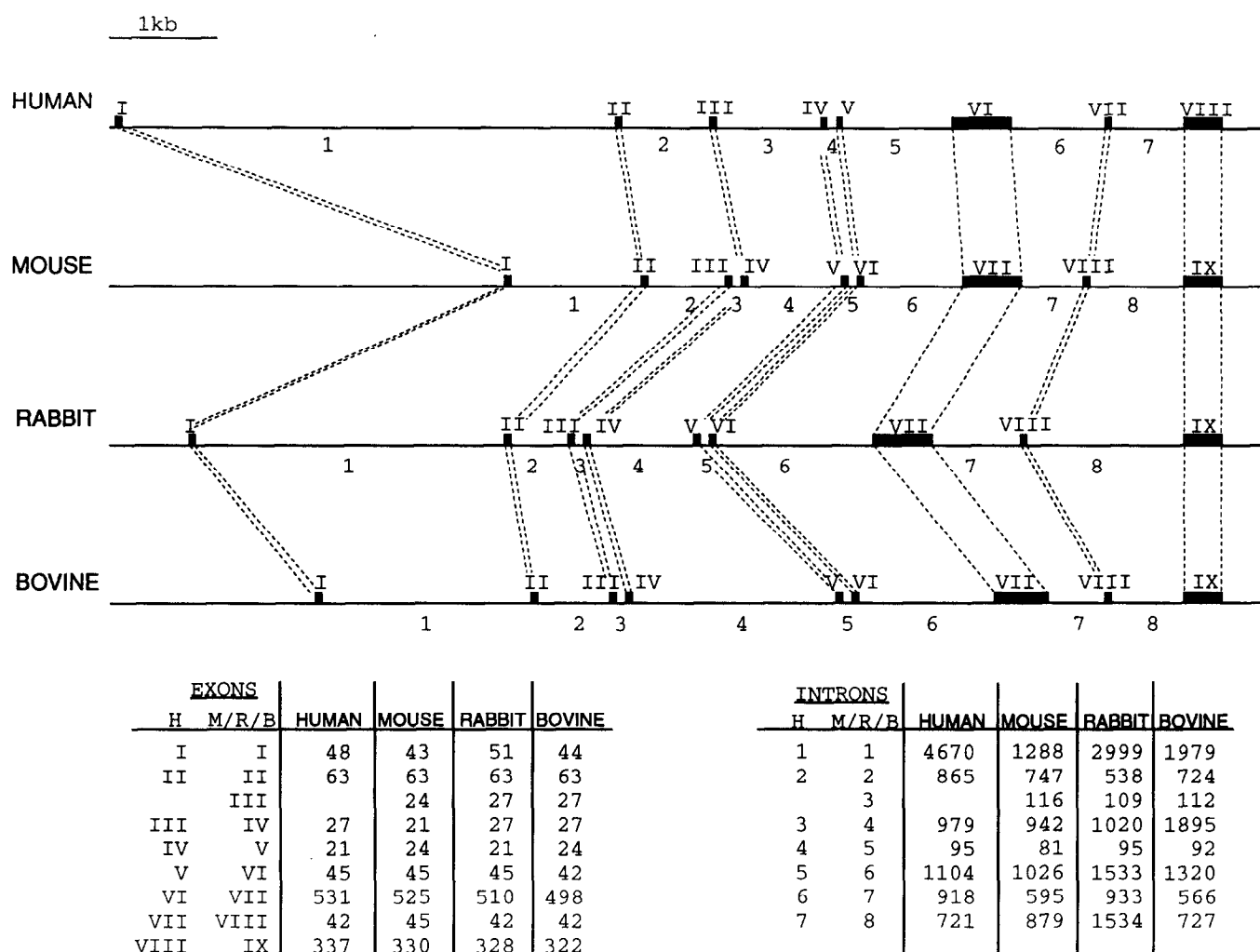


Fig. 3. Comparison of the structure of human, bovine, rabbit and mouse *Bca* genes. The genes are aligned from the 3' end of the last exon. Exons are numbered I–VIII and introns are numbered 1–7 for the human gene (H), for mouse, rabbit and bovine (M/R/B) genes the exons are numbered I–IX and the introns 1–8, respectively. Tables below the genes list exact sizes, in bp, of the exons and introns of the four genes.

sequence of this missing exon is homologous to exon-3 of other species.

(2) Despite the lower number of exons, the human gene is longer than other known *Bca* genes.

(3) The 226-aa ORF extends from exon-2 to exon-7.

(4) No coding triplets are disrupted by splice junctions.

(5) Alternative splicing of exon-5 may explain the presence of two *Bca* mRNA sequences which differ with respect to a Gln¹⁹ codon of the mature protein. Of these, the variant containing the Gln¹⁹ residue seems to be the most abundant.

ACKNOWLEDGEMENTS

We thank Ms. Brith Leidvik for excellent technical assistance and Drs. Anders Bergman and Gunnar Skogman for their generous support. This work was financially sup-

ported by Symbicom AB and the Swedish Medical Research Council (16x-07922; 19x-05708).

REFERENCES

- Alexander, L.J. and Beattie, C.W.: The sequence of porcine β -casein cDNA. *Anim. Genet.* 23 (1992) 369–371.
- Båga, M., Göransson, M., Normark, S. and Uhlin, B.E.: Transcriptional activation of a Pap pilus virulence operon from uropathogenic *Escherichia coli*. *EMBO J.* 4 (1985) 3887–3893.
- Blackburn, D.E., Hobbs, A.A. and Rosen, J.M.: Rat β -casein cDNA: sequence analysis and evolutionary comparisons. *Nucleic Acids Res.* 10 (1982) 2295–2307.
- Bonsing, J., Ring, J.M., Stewart, A.F. and Mackinlay, A.G.: Complete nucleotide sequence of the bovine β -casein gene. *Aust. J. Biol. Sci.* 41 (1988) 527–537.
- Breatnach, R., Benoist, C., O'Hare, K., Gannon, F. and Chambon, P.: Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. *Proc. Natl. Acad. Sci. USA* 75 (1978) 4853–4857.

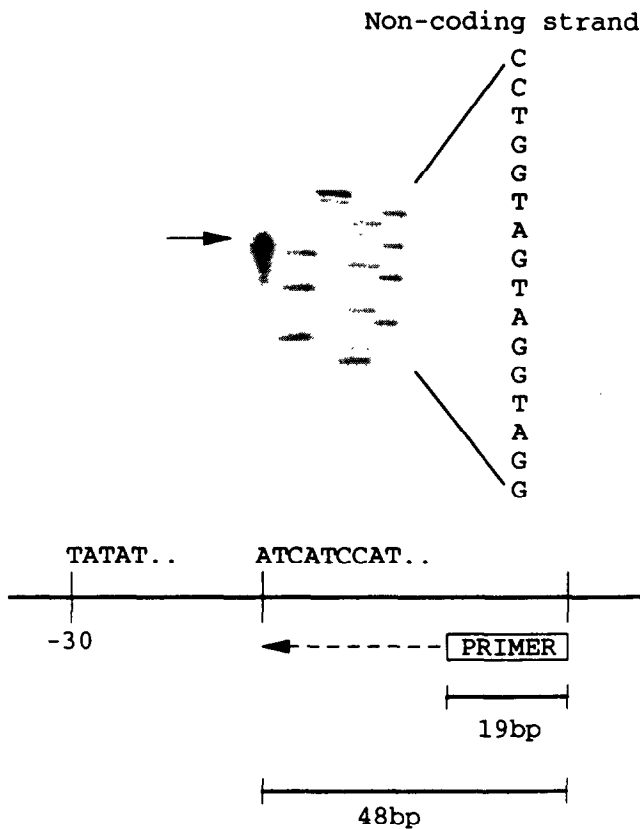


Fig. 4. Mapping *tsp*. To map the *tsp* of the human *Bca* gene, the 5' end of *Bca* mRNA was determined by primer extension. The sequence of the synthetic primer complementary to human *Bca*, positions 30–48, is 5'-CGTTTTTCCAAGATTGAA. The primer extension products were co-electrophoresed on a sequencing gel with a sequencing reaction with the same primer. The sequence of the non-coding strand surrounding the transcription initiation site is indicated alongside the autoradiogram. The diagram below indicates the location of the primer and the arrow shows the direction of extension reaction. The size of the extension product and the *tsp* mapped to the A nt located 30 bp downstream from the TATA-box is also shown.

- Chomczynski, P. and Sacchi, N.: Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem.* 162 (1987) 156–159.
- Collet, C., Joseph, R. and Nicholas, K.: Molecular characterization and *in-vitro* hormonal requirements for expression of two casein genes from a marsupial. *J. Mol. Endocrinol.* 8 (1992) 13–20.
- Devinoy, E., Hubert, C., Jolivet, G., Thépot, D., Clergue, N., Desaleux, M., Dion, M., Servely, J.L. and Houdebine, L.M.: Recent data on the structure of rabbit milk protein genes and on mechanism of the hormonal control of their expression. *Reprod. Nutr. Dev.* 28 (1988) 1145–1164.
- Göransson, M., Forsman, K., Nilsson, P. and Uhlin, B.E.: Upstream activating sequences that are shared by two divergently transcribed operons mediate cAMP-CRP regulation of pilus adhesin in *Escherichia coli*. *Mol. Microbiol.* 3 (1989) 1557–1565.

- Greenberg, R., Groves, M.L. and Dower, H.J.: Human β -casein. *J. Biol. Chem.* 259 (1984) 5132–5138.
- Hansson, L., Bergström, S., Hernell, O., Lönnerdal, B., Nilsson, A.-K. and Strömqvist, M.: Expression of human milk β -casein in *Escherichia coli*; comparison of recombinant protein with native isoforms. *Prot. Expr. Purif.* 4 (1993) 373–381.
- Jahn, G.A., Djiane, J. and Houdebine, L.M.: Inhibition of casein synthesis by progestagens *in vitro*: modulation in relation to concentration of hormones that synergize with prolactin. *J. Steroid Biochem.* 32 (1989) 373–379.
- Jones, W.K., Yu-Lee, L.-Y., Clift, S.M., Brown, T.L. and Rosen, J.M.: The rat casein multigene family. *J. Biol. Chem.* 260 (1985) 7042–7050.
- Lönnerdal, B., Bergström, S., Andersson, Y., Hjalmarsson, K., Sundqvist, A.-K. and Hernell, O.: Cloning and sequencing of a cDNA encoding human milk β -casein. *FEBS Lett.* 269 (1990) 153–156.
- Martin, P. and Leroux, C.: Exon-skipping is responsible for the 9 amino acid residue deletion occurring near the N-terminal of human β -casein. *Biochem. Biophys. Res. Commun.* 183 (1992) 750–757.
- Menon, R.S., Chang, Y.-F., Jeffers, K.F. and Ham, R.G.: Exon skipping in human β -casein. *Genomics* 12 (1992) 13–17.
- Miller, M.J.S., Witherly, S.A. and Clark, D.A.: Casein: a milk protein with diverse biologic consequences. *Proc. Soc. Exp. Biol. Med.* 195 (1990) 143–159.
- Mount, S.M.: A catalogue of splice junction sequences. *Nucleic Acids Res.* 10 (1982) 459–471.
- Provot, C., Persuy, M.-A. and Mercier, J.-C.: Complete nucleotide sequence of ovine β -casein cDNA: interspecies comparison. *Biochimie* 71 (1989) 827–832.
- Puissant, C. and Houdebine, L.-M.: An improvement of the single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *BioTechniques* 5 (1990) 148–149.
- Roberts, B., DiTullio, P., Vitale, J., Hehir, K. and Gordon, K.: Cloning of the goat β -casein-encoding gene and expression in transgenic mice. *Gene* 121 (1992) 255–262.
- Sambrook, J., Fritsch, E.F. and Maniatis, T.: *Molecular Cloning. A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1989.
- Schaerer, E., Devinoy, E., Krähenbühl, J.P. and Houdebine, L.M.: Sequence of the rabbit β -casein cDNA: comparison with other casein cDNA sequences. *Nucleic Acids Res.* 16 (1988) 11814.
- Stewart, A.F., Bonsing, J., Beattie, C.W., Shah, F., Willis, I.M. and Mackinlay, A.G.: Complete nucleotide sequences of bovine α s2- and β -casein cDNAs: comparisons with related sequences in other species. *Mol. Biol. Evol.* 4 (1987) 231–241.
- Thépot, D., Devinoy, E., Fontaine, M.L. and Houdebine, L.M.: Structure of the gene encoding rabbit β -casein. *Gene* 97 (1991) 301–306.
- Yoshimura, M., Banerjee, M.R. and Oka, T.: Nucleotide sequence of a cDNA encoding mouse β -casein. *Nucleic Acids Res.* 14 (1986) 8224.
- Yoshimura, M. and Oka, T.: Isolation and structural analysis of the mouse β -casein gene. *Gene* 78 (1989) 267–275.