Nucleic Acids Research, Vol. 19, No. 20 5591–5596

# Genomic organization of the bovine alpha-S1 casein gene

Dirk Koczan, G.Hobom and H.-M.Seyfert*
Institut für Mikrobiologie und Molekularbiolgie der Justus-Liebig-Universität Gießen, Frankfurterstraße 107, D-6300 Gießen, FRG

Received July 18, 1991; Revised and Accepted September 24, 1991 EMBL accession no. X59856

## ABSTRACT

**We report the sequence of the complete bovine α-s1 casein gene eludicating for the first time the genomic organization of an α-s type casein gene. Extending over 17508 bp the gene is split into 19 exons, ranging in size from 24 bp to 385 bp. Except for the translational stop codon not a single coding triplet of the α-s1 reading frame is disrupted by any of the splice junctions, which all confirm to known splice consensus sequences. Nine out of 16 coding exons begin with a 'GAX' codon, specific for glutamate. Splicing of this codon from exon 10 to the preceding exon creates a major phosphorylation site. An intron-exon-intron stretch of 154 bp comprising exons 10 and 13 is found precisely duplicated. Associated with the gene, copies of 8 atriodactyla retroposons are found, 6 of which are interspersed into the sequences of the three longest introns. We discuss the possibility that three functional parts of the gene have been recruited and evolutionary conserved at a time before gene diversification gave rise to the separate evolution of α- and β-type casein-genes.**

## INTRODUCTION

The caseins are the major milk proteins of mammals. Their dual function for the suckling infant is to serve as a major source of amino acids, as well as to transport phosphate and calcium in sufficient amounts to support growth of bones. In the cow these functions are carried out by three different proteins, two α-and one β-type casein. Therefore, these have the properties of phosphoproteins. They aggregate in solution as micelles, sequestering up to 5% of their dry weight as $Ca^{2+}$. A major role for the non-phosphorylated κ-casein is to stabilize those micelles.

Coordinate expression of the four casein genes is regulated by different hormones involved in lactation (1) and concerted regulation is borne out in a tight clustering of all four genes within less than 200 kb of DNA on bovine chromosome 6 (2, 3). Such an arangement might have been expected for the phosphorylated caseins, since according to our current understanding—based on cDNA sequencing of all bovine and rat caseins and some more from other species (4–10)—these caseins have evolved from an ancesteral gene by duplication and diversification, which itself originated from exon shuffling (11). However, tight linkage of the κ-casein gene to the three others is unexpected, since it is functionally distinct and conceivably may represent a different branch in evolution of casein gene structure (12).

While the genomic sequence of the bovine β- and κ-casein genes have been determined (12, 13) no genomic α-casein gene organization has been reported yet. Only partial segments concerning the promoter region and first exon of the bovine α-s1 casein gene as well as the same promotor region extented into exon five of the rat α-casein gene have been sequenced (14). These studies supported the view that duplication of short exons contributed to casein gene evolution and in addition pointed out conserved sequence elements in the promotor region as potential sites of gene regulation. However, a full understanding of the genomic organzation and knowledge of the sequence of the bovine α-s1 casein gene would be desirable, since its product is the most abundant bovine milk protein, exceeding a concentration of 13 mg/ml of milk (15) and hence is one of the most important quantitative traits in cattle breeding. Based on the complete sequence information it is possible to design specific primers to exploit the advantages of PCR application to monitor the occurrence of the reported (2) α-s1-casein linked RFLPs and follow electrophoretic variants of this protein (16) in breeding analysis. Furthermore, it will faciliate the use of this gene to specifically target and booster the expression of transgenes in the mammary gland, which in pilot experiments has already been demonstrated to be possible (17–21). For these purposes a detailed knowledge of α-s1 intron sequences may be extremely helpful, since it was proven for a variety of transgenes, that inclusion of intron sequences may improve transgene expression by an order of magnitude (22). Thus, we sequenced the entire gene.

## MATERIALS AND METHODS

A genomic gene bank was constructed by cloning partially MboI digested DNA from a 'Deutsche Schwarzbunte' cow into BamHI digested lambda EMBL 3 DNA.

Initially this library was screened using as probes two HaeIII fragments of the cDNA clone pBC184 (6; provided by Dr. MacKinlay, Australia), encompassing most of the α-s1 cDNA sequence. Lateron, appropriate genomic DNA probes were used for genomic walking after sequencing.

* To whom correspondence should be addressed


a
b
c
d
e
1
2
3
4
5
6
7
8
//Intron I//
//Intron III//
//Intron IV//
//Intron V//
//Intron VI//
//Intron VII//

**Fig. 1.** Sequence of the bovine $\alpha_{s1}$ casein gene. Bases are numbered on the right site of the sequence blocks; exon positions are indicated at the opposite margines and their sequence is printed with large, bold face letters. Functional elements (TATA-box, polyadenylation signal) are boxed, direct repeats underlined. Retroposon-elements are indicated by intermediate sized letters. Sequence blocks containing intron sequences only have been deleted, for conciseness.

Probe labelling with $^{32}$P-dCTP and hybridization of filter lifts and Southern blots followed standard protocols (23). Stringency for hybridizations was 5×SSC at 65°C and for washings 0.5% SSC at 55°C, 0.1% SDS.

Sanger's dideoxy chain termination method (24) using a $T_7$ DNA polymerase based sequencing kit (Pharmacia) with $^{35}$S-dATP as label was used to sequence single stranded DNA after subcloning the genomic DNA into M13-phages

## RESULTS

The sequence of the entire gene has been determined (Fig. 1) from five overlapping lambda clones (Fig.2). The inserted genomic DNA of both orientations had been subcloned into 115 different M13 phages.

Altogether, 22087 contiguous basepairs of bovine DNA have been sequenced, covering both strands throughout the entire

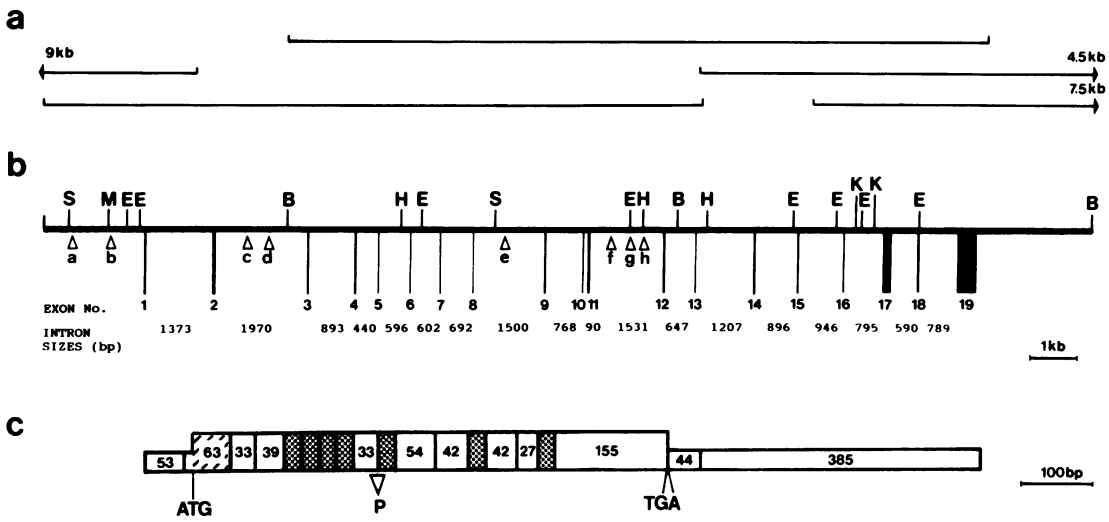


**Fig. 2.** Gene structure of the $\alpha_{s1}$ -casein gene. a) Overlapping lambda EMBL3 clones containing the $\alpha_{s1}$ genomic sequences are aligned annd drawn to scale relative the deduced gene structure. b) Restriction map of the αs1 casein gene (B= BamHI; E =EcoRI; H= HindIII; K= KpnI; M= SmaI; S= SphI). Exon number, position and relative size are indicated (black bars, numbers). Intron sizes are given below. Arrow heads specify the positions of artiodactyla retroposon elements (lower case letters a–h). c) cDNA structure: Vertical black lines define exon boundaries. 5'and 3'untranslated regions (half bar) are separated by the coding region (full bar). Exon sizes are given inside (bp); stippeling denotes the 7 small 24 bp exons, while the leader peptide, encoded by the second exon, is hatched. (ATG): initiation codon of translation; (P): major phosphorylation site, spliced together by exon 9 and 10; (TGA): translation stop codon.

5'-flanking region, for every exon and extending into the flanking intron segments, in total about 50% of the whole sequence. The other half of the intron sequences has been determined from one strand only, using lower parts of the sequencing gels. The sequences of all restriction sites used for cloning were crosschecked on corresponding overlapping clones.

## Gene Structure

The bovine α-s1 casein gene extends over 17508 bp, with 1138 bp of exon and 16370 bp of intron DNA. Thus, the size ratio of exon vs intron DNA is 1:14.4. The gene is split into 19 exons, ranging in size from 24 to 385 bp, and 18 introns from 90 bp to 1967 bp (Fig. 2). Several features are noteworthy:

1. While the first exon of 53 bp is not coding at all, the entire leader-peptide as well as the fist two amino acids of the mature protein are encoded by exon 2, spanning 63 bp, precisely as found in the mouse and bovine β-casein genes (13, 25) and similarily in all rat casein genes (14).

2. Not a single coding triplet of the α-s1 reading frame is disrupted by any of the splice junctions. Consequently, the coding exons 3 to 16 contain multiples of 3 bp. Only the translation stop codon UGA is created by splicing the final nucleotides UG of exon 17 onto the first nucleotide A of exon 18 (Fig 2). All splice junctions follow the 5' GT/3'AG splice rule (26, 27).

3. We found 7 exchanges compared to the published α-s1-cDNA sequence (6), all of which are base transitions. All three exchanges found in the coding region are confined to the third positions within coding triplets and do not cause amino acid substitutions.

4. Nine out of 16 coding exons begin with a 'GAX' sequence, confirming a prediction based on the cDNA analysis (6). A major phosphorylation site within the α-s1 sequence is created by splicing the first codon of exon 10 (GAA: glutamate) to the preceding exon. A similar phenomenon was found in the rat a casein gene (14), and the β-casein genes of rat, mouse and the bovine species (3, 14, 25).

5. An intron-exon-intron stretch of 154 bp is found precisely duplicated (Fig. 3; positions +9101 to +9254 and +11489 to +11642), revealing 97.4% homology with only 4 C/T base transitions and no gaps at all. This area encompasses exons 10 and 13, i.e. two of the short 24 bp exons, together with their flanking intron regions.

6. Two potentially functional 'TATA' boxes have been described in the α-s1 promoter region (14), with the sequence TTTAAAT at −29 being linked to the major transcription initiation site. We found that part of this sequence belongs to an 11 bp direct repeat motive AAATAGCTTGG which is also located at −572 (Fig. 1).

## Artiodactyla Retroposons

At 8 locations copies of artiodactyla retroposons are found (Figs. 1, 2). Element (a) resides in the distal promotor region between bp −1807 and −1531, in the vicinity embeded by a direct 13bp repeat sequence GGGATCTTCCCAA (positions −2130/−1478). It appears to be a split and rearranged art2 element (28). Its 5' end reveals a 79.6% homology extending over 186 bp to the 3' end of the published art2 element (positions 411−536 in ref 28) in the complementary strand. However, the 3' half of element (a) (positions −1602/−1531) confirms over 71 bp with 77.6% homology to the strand sequence of the 5' half of the art2 element (positions 88−146, ref 28). Overlapping with the latter region and extending up to −1482, i.e. just 4 bp 5' of the direct repeat at −1478, we detected an 80% sequence homology to element (d) in intron 2 over a stretch of 108 bp.

Element (b) (−753 to −644) relates to a monomer BMF element (29) and displays a homology of 78% to the published sequence over 111 bp. No direct repeat sequences can be detected in its vicinity.

Intron 2 harbors two dimer BDF elements (30), elements (c) (+1885 to +2151) and (d) (+2409 to 2675). They are organized in a very similar way with a central unit of 88.8% sequence homology within 268 bp, that is flanked by 11 bp direct repeats.

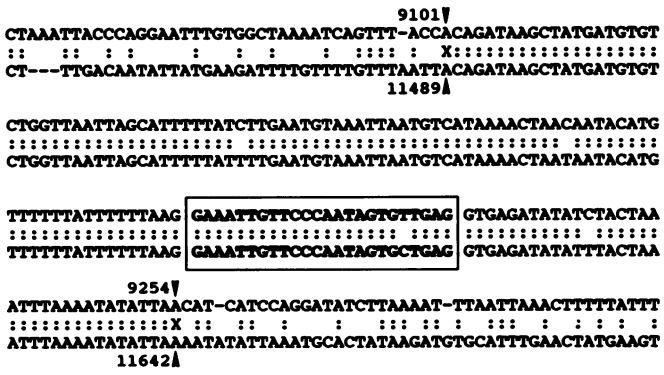


**Fig. 3.** Sequence dupliation of 154 bp comprising exons (boxed) 10 (upper) and 13 (lower). The dupliation (97,4% homology) is framed by crosses and runs from position +9101 in intron 9 to +9254 in intron 10 and is matched by the sequence between +11489 to +11642 in introns 12 and 13 respectively.

The repeat sequences flanking element (c) read CCAAATATA-CT, while TTGTGACCTCT is flanking element (d).

A segment of 9 bp (+2389/ +2397) preceding the left direct repeat of element (d) is identical with the 3′ left end sequence of the MLV retrovirus TCTGTTCCT (31), and a sequence TA-CTTTC, which is part of the 3′ right end of the MLV retrovirus, is found in the centre of element (d) in opposite orientation (+2637). Together, these sequence motives confirming to the ends of the MLV retroviral RNA are bracketing just one arm of the BDF dimer element in intron 2.

Element (e), which again relates to an art2 element, is located in intron 8 (+7285/ +7721). It is flanked by 11 bp direct repeats (TGGGCTATGCT).

Another three elements are all located in intron 11. None of their ends retain direct repeats. Element (f) (+9801/+9900) and (g) (+10170/+10285) are BMF elements arranged in opposite orientation, while element (h) (+10371/+10662) again confirms over a stretch of 295 bp to the published art2 element sequence with 74.6% homology.

### Potentially Regulatory Sequences

A segment of 700 bp flanking the bovine α-s1 casein gene promoter region and exon 1 have been sequenced and were searched for potential binding sites of glucocorticoid hormon and progesteron receptor molecules (14). This analysis has been extended over the entire sequence as presented here and several additional potential receptor binding sites can be pointed out. In addition to those known sites we find a glucocorticoid hormon receptor binding motive TAATGTG (32) at position +8184. At nearby position +8215 starts the chicken egg white steroid homone receptor binding motive AAAATTGAC (33). An identical sequence motive has also been found within the mouse β-casein gene (25). Furthermore, a potential binding site of the progesterone receptor TGTTCACT (34) is found at position +12173 in intron 13. The significance of potential regulatory DNA binding sites as presented here remains obscure though, unless their function has been examined in future experiments.

## DISCUSSION

The sequence of the bovine α-s1-casein gene eludicates for the first time the genetic structure of an α-type casein gene. Covering 17508 bp (without counting promotor sequences) it is considerably longer than all other known genes coding for phosphorylated caseins (rat β, g, mouse β, bovine β). Moreover, it is split into 19 exons, thus differing from the general 9 exon pattern of all β-casein genes and the bovine k-casein gene with five exons. However, by analogy we expect that the bovine α-s2 casein gene may also be distributed over more than 9 exons, and possibly more than 16 functional subunits as has been suggested (7). The verification of the predicted occurrence (6) of several short 24 bp exons—in particular exons 5 through 8 with 24 bp each—is in agreement with the hypothesis that exon duplication was one means of casein gene assembly during evolution. The only indication of a previous sequence duplication is the very high degree of homology (97.4%) of the two 154 bp stretches of DNA containing exons 10 and 13.

We observe a ratio of 1: 14.4 for exon vs. intron size in the α-s1-casein gene. Thus, all 19 exons contribute only 7% of the total α-s1 gene sequence. In contrast, nine exons make up more than 10% of the β-casein genes of rat, mouse and cattle . Such a comparison may indicate, that extensive deletion of intron sequences occurred within the β-casein-genes after diversification gave rise to the separate evolution of α- and β casein genes. This event must have taken place prior to the evolution of those species of mammals.

Furthermore, a comparison of the bovine α-s1 and β-casein-gene structure (for β, see EMBL Acc. no. X14711; MacKinlay '89) allows to define three parts of the casein genes, which may have been assembled prior to evolutionary diversification of both genes:

First, the 5′ flanking region together with exons 1 and 2, which not only show sequence similarities , as already noted (11, 14), but which also (i) contain a similar BMF retroposon (element b, Fig 1) in the same orientation and in similar distance from the promoter in the 5′ flanking region of both genes and (ii) the strong conservation of exon 2, including its boundary behind the first two amino acids of the mature proteins, in addition to coding for the entire leader peptide in both cases.

Second, the short 24 bp exons (α-s1: exon 10; β: exon 5) contributing to the major phosphorylation site of the proteins are separated by extremely short introns (α-s1: 90 bp; β: 92 bp) from the next exon in both in both genes, indicating an evolutionary conservation of this arangement.

Third, the 3′ ends of both genes are organized in a very similar way. The largest protein coding exon (exon 17 in the α-s1 gene, exon 7 in the β-casein gene) is followed by a 44 bp or 42 bp exon (α-s1 vs β), and finally by the last, non-coding exon, which is quite long (385 bp and 323 bp, α vs β respectively). Since even the introns separating these exons are quite similar in size (590 vs 601 bp and 789 vs 730 bp), it is likely that this part of the casein gene has been recruited before diversification into the α and β types of the casein genes occurred. Alteration of one splice junction between the two genes may explain the extension of the reading frame as well as the location of the stop codon on the last exon in the β-casein gene as opposed to the α-s1 gene, in which splicing of exon 17 onto 18 will create the stop codon.

Repetitive sequences are found associated with many genes of artiodactyla. These analogues to the human alu-type repetitive sequences (35) have been grouped into 3 families (24−26). We found eight complete or partial retroposon elements in the α-s1 casein gene. Considering sequential and structural similarity with the published prototypes along with the occurrence of flanking direct repeats as indicative for a recent insertion, it appears that elements c) and d) are rather young insertions, while elements

f), g), h) in intron 11 and element b) in the 5'flanking region are oldest by these criteria.

A surprising observation with respect to the artiodactyla BDF dimer retroposons is the detection of a sequence homology for the central part of these elements, i.e the connection between the two arms of these elements, with the functional part of the human β-interferon-gene (INF) box III enhancer element (36) on the one hand, which in turn is known share extensive sequence homology to the human hsp 70 enhancer box element necessary for serum stimulation (37). The central sequence TGAAAGTG-AAAAGTGAAAGTG of element c) is distinct from the β-INF enhancer box by only two conversions of the 21 bases, while the central sequence GAAGGGAAAG of element d) needs only the insertion of one 'A' nucleotide to match the functional sequence of the human hsp 70 enhancer element. The significance of this observation is unclear, but possibly permutations of sequence motives of the abundantly available alu-type repetitive elements in the vicinity of these genes resulted in selective advantage, and hence have been conserved.

With respect to the bovine α-s1 casein gene it appears, that the insertion of the artiodactyla retroposons increased the intron sizes, thus altering the otherwise fairly constant spacing of exons throughout large parts of the gene.

## ACKNOWLEDGEMENTS

We are grateful to Prof. Dr. A.MacKinlay for providing the α-s1 cDNA clone C184. This work contributes to the doctoral thesis of D.K. and was supported by grants from the 'Deutsche Forschungsgemeinschaft' and 'FAZIT-Stiftung'.

## REFERENCES

1. Rosen, J.M., Matusik, R.J., Richards, D.A., Gupta, P. and Rodgers, J.R. (1980) Rec. Prog. Hor. Res. 36, 157–193.
2. Threadgill, D.A. and Womack, J.E. (1990) Nuc. Acids Res. 18, 6935–6942.
3. Ferretti, C. and Scaramena, U. (1990) Nuc. Acids Res. 18, 6829–6833.
4. Hobbs, A.A. and Rosen, J.M. (1982) Nuc. Acids Res. 10, 8079–8089
5. Blackburn, J.E., Hobbs, A.A. and Rosen, J.M. (1982) Nuc. Acids Res. 10, 2295–2307.
6. Stewart, A.F., Willis, J.M. and MacKinlay, A.G. (1984) Nuc. Acids Res. 12, 3895–3907.
7. Stewart, A.F., Bonsing, J., Beattie, C.W., Shah, F., Willis, J.M. and MacKinlay, A.G,. (1987) Mol. Biol. Evol. 4, 231–241
8. Henninghausen, L.G., Steudle, A. and Sippel, A.E. (1982) Eur.J. Biochem. 126, 569–572.
9. Devinoy, E., Schaerer, E., Jolivet, G., Fontaine, M.L., Kraehenbuhl,J.-P. and Houdebine, L.-M. (1988) Nuc. Acids Res. 16, 11813
10. Schaerer, E., Devinoy, E., Kraehenbuhl, J.-P. and Houdebine, L.-M. (1988) Nuc. Acids 16, 11814
11. Jones, W.K., Yu-Lee, L.-Y., Cliffs, S.M., Brow, T.L. and Rosen, J.M. (1985) J. Biol. Chem. 260, 7042–7050.
12. Alexander, L.J., Stewart, A. F., MacKinlay, A.G., Kapelinskaya, T.V., Trach, T.M. and Gorodetsky, S. J. (1988) Eur. J. Biochchem. 395–401.
13. Gorodetsky, S.J., Tkach, T.M. and Kapelinskaya, T.V. (1988) Gene 6, 87–96.
14. Yu-Lee, L.Y., Richter-Mann, L., Couch, C.H., Stewart, A.F., MacKinlay, A.G. and Rosen, J.M. (1986) Nuc. Acids Res. 14, 1883–1902.
15. Jennes, R. (1970) Milk proteins, chemistry and molecular biology, p 25, Voll, H.A. McKenzie (ed), Acad. Press, New York.
16. Eigel, W.N., Butler, J. E., Ernstrom, C.A., Farrell, JR,H.M., Harwalker, V.R., Jenness, R. and Whitney, R.McL. (1984) J. Dairy Sci. 67, 1599–1631.
17. Simons, P., McClenaghan, M. and Clark, J. (1987) Nature 328, 530–532.
18. Simons, P., Wilmut, J., Clark, A.J., Archibald, A.L., Bishop, J.O. and Cathe, R. (1988) Biotechnology 6, 179–183.
19. Bühler, Th.A., Bruyère, Th., Stranzinger, G. and Burki, K. (1990) Biotechnology 8, 140–143.
20. Meade, H., Gates, L., Lacy, E. and Lonberg, N. (1990) Biotechnology, 8, 443–446.
21. Archibald, A. L., McClenaghan, M., Hornsey, V. and Simons, P. (1990) Proc. Natl. Acad. Sci. 87, 5178–5182.
22. Brinster, R.L., Allen, J.M., Behringer, R.R., Gelinas, R.E. and Palmiter, R.D. (1988) Proc. Natl. Acad. Sci. 85, 836–840.
23. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Cold Spring Harbor
24. Sanger, F., Nickler, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463–5467.
25. Yoshimura, M. and Oka, T, (1989) Gene 78, 267–275.
26. Breathnatch, R., Benoist, C., O'Hare, K., Cannon, F. and Chambon, P. (1978) Proc. Natl. Acad. Sci. USA 75, 4853–4857.
27. Shapiro, M. B. and Senapathy, P. (1987) Nuc. Acids Res. 15, 7155–7174.
28. Duncan, C. H. (1987) Nuc. Acids Res. 15, 1340–
29. Watanabe, Y., Tsukada, T., Notake, M., Nakanishi, S. and Numa, L. (1982) Nuc. Acids Res. 10, 1459–1469.
30. Sokowronski,J., Plucienniczak,A., Bedarek, A. and Jaworski, J. (1984) J. Mol. Biol. 177, 399–416.
31. Brown, P., Bowerman, B., Varnus, H.E. and Bishop, J.M. (1989) Proc. Natl. Acad. Sci. 86, 2525–2529.
32. Beato, M. (1987) Biochem. Biophys. Acata 910, 95–102.
33. Dean, D.C., Knoll, B. J., Riser, M.E. and O'Malley, B.W. (1983) Nature 305, 551–554.
34. Bailly, A, Lepage, Rauch, M and Milgrom, E. (1986) EMBO J. 5, 3235–3241.
35. Kariya, Y., Kato, K Hayashizaki, Y. Himeno S., Tarui S. and Matsubara, K. (1987) Gene 53, 1–10.
36. Zinn, K. and Maniatis, T. (1986) Cell 45, 611–618.
37. Wu, B. J., Kingston, R.E. and Morimoto, R.J. (1986) Proc. Natl. Acad. Sci. USA 83, 629–633.