Volume 279, number 1, 9–13 FEBS 09367 February 1991

ADONIS 001457939100044

Highly conserved sequences in the 3′-untranslated region of the COL1A1 gene bind cell-specific nuclear proteins

Arto Määttä[1], Paul Bornstein[2] and Risto P.K. Penttinen[1]

[1]*Departments of Medical Biochemistry, University of Turku, SF 20500, Finland* and [2]*Biochemistry, University of Washington, Seattle, WA 98195, USA*

Received 10 December 1990

Sequencing of the 3′ untranslated region (3′-UTR) of the human COL1A1 gene revealed numerous putative regulatory motifs and two highly conserved regions flanking the two polyadenylation sites. The conserved regions were separated by about 700 bp of less conserved sequences. The first region consists of almost all the 3′-UTR of the shorter (4.8 kbp) COL1A1 transcript. The second conserved domain includes a motif shared with several collagen genes. Both conserved domains bind cell-specific nuclear proteins suggesting that the 3′-UTR is important for cell specific expression of the COL1A1 gene.

Collagen gene; Untranslated region; Polyadenylation; Nuclear protein

1. INTRODUCTION

Studies of the transcriptional control of type I and II collagen genes have revealed functional promoters in their 5′ flanking regions and transcriptional activator and suppressor elements within the first intron (for references see [1–3]) but only little attention has been paid to the 3′ noncoding regions. The first report on this topic by Herget et al. [4] showed co-expression of the rat proα(I) collagen mRNA and a nuclear protein that binds the 3′-UTR. Regulatory elements have, however, been identified in the 3′-UTR of several other genes, e.g. β-actin [5], myosin light chain [6] and a *Drosophila* retroposon [7].

Collagen genes often yield several transcripts which differ only in their 3′-UTR. In involuting human placenta the shorter 4.8 kbp COL1A1 transcript predominates [8] whereas TGF-β treatment of fibroblasts increases both mRNAs, especially the longer one [9]. This increase is inhibited by cycloheximide suggesting that the effect is mediated by short-lived transcription factor(s). The 3′-UTR determines the half life of many mRNAs [10,11], a finding which may include collagens [9,12]. Treatment with γ-interferon, tumor necrosis factor α or with glucocorticoids reduces, and with interleukin 1 or TGF-β [9,12] increases, the levels of type I collagen mRNAs but details of these regulatory events are not known. In this paper we present the sequence of the 3′-UTR of the COL1A1 gene, identify putative regulatory motifs and demonstrate cell-specific binding of nuclear proteins to two highly conserved domains flanking the two polyadenylation sites.

Correspondence address: R. Penttinen, Department of Medical Biochemistry, University of Turku, Kiinamyllynkatu 10, 20520 Turku, Finland

Abbreviations: COL1A1 and COL1A2, type I procollagen α1 and α2 chain genes; COL2A1, type II procollagen α1 chain gene; COL4A1, type IV procollagen α1 chain gene; PMSF, phenylmethylsulphonyl fluoride; bp, base pair; kbp, kilobase pair

2. MATERIALS AND METHODS

2.1. *DNA sequencing*

The 5.2 kbp *Eco*RI–*Eco*RI fragment of the cosmid clone CG 103 contains the 3′-UTR of the COL1A1 gene [8]. The 5′ *Eco*RI site of this fragment (base 1 in Fig. 2) is located in the last exon, followed by two *Hin*dIII sites 0.3 and 2.2 kbp downstream which were used for cloning into pUC plasmids. Overlapping restriction fragments were subcloned into M13 mp18 and 19 vectors. Gaps in the sequence were covered by creating a set of deletions with *Bal*31 nuclease [13]. The restriction and *Bal*31-generated fragments were ligated into M13 vectors (Fig. 1) for dideoxy sequencing of single-stranded templates with modified T7 DNA polymerase (Sequenase, US Biochemicals) and M13 universal primer. The data were analyzed with Genepro and UWGCG programs and compared with EMBL and GenBank libraries.

2.2. *Nuclear extracts and gel retardation assays*

Nuclear proteins from human foetal Achilles tendon (FTF) and embryonic chick tendon fibroblasts (CTF) and from HeLa and NS-1 mouse myeloma cells [1,14], were dialysed against nuclear dialysis buffer (12 mM Hepes, pH 7.9, 100 mM KCl, 1 mM EDTA, 0.2 mM EGTA, 1 mM DTT, 20% glycerol and 1 mM PMSF) and stored at −70°C.

DNA probes were prepared from subcloned inserts (Fig. 1) labeled either at the 5′ end by T4 polynucleotide kinase and [γ-^{32}P]dATP (5000 Ci/mmol) or at the 3′ end by the Klenow enzyme and [α-^{32}P]dCTP, and purified by agarose gel electrophoresis. Gel retardation experiments [1] were carried out in the presence of competitor

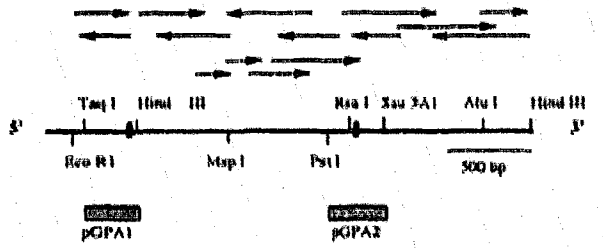

Fig. 1. Sequencing strategy and the restriction map of the 2.3 kbp *Eco*RI/*Hin*dIII fragment of the CG 103 cosmid clone [8]. The upper rows of arrows indicate the subcloned restriction fragments and the two lower rows the *Bal*31-generated deletion series. The hatched rectangles indicate the inserts for protein binding analysis. For the pA1 probe a 302 bp *Taq*I/*Hin*dIII fragment was subcloned into *Acc*I/*Hin*dIII-cleaved pGEM-1 plasmid. The PA2 probe was generated by inserting a 334 bp *Pst*I/*Sau*3A1 fragment into *Pst*I/*Bam*HI-cleaved pGEM-1. Only the relevant *Sau*3A1 site is shown. Black vertical bars: AATAAA polyadenylation signals.

poly(dI-dC) DNA. The probes (10000 cpm; about 1 ng DNA) were incubated with the reaction mixtures for 15–20 min at 20°C in the absence or presence of 1 mM $MgCl_2$ and $ZnSO_4$. The complexes were resolved by electrophoresis in 4% native polyacrylamide gels using 50 mM Tris, 400 mM glycine buffer, pH 8.5 and subjected to autoradiography.

3. RESULTS

3.1. *The 3' end of the COL1A1 gene*

The sequence of 2.3 kbp of human COL1A1 3'-UTR revealed two polyadenylation AATAAA signals (pA1 and pA2) separated by 1113 bp (Fig. 2). This corresponds to the size difference of 4.8 and 5.8 kbp mRNAs [9,15]. The first 540 bp of the human sequence are homologous with those of the mouse [16] (Fig. 2) and include a triplicated AATAAA signal. The less conserved region (about 700 bp) is followed by approx-

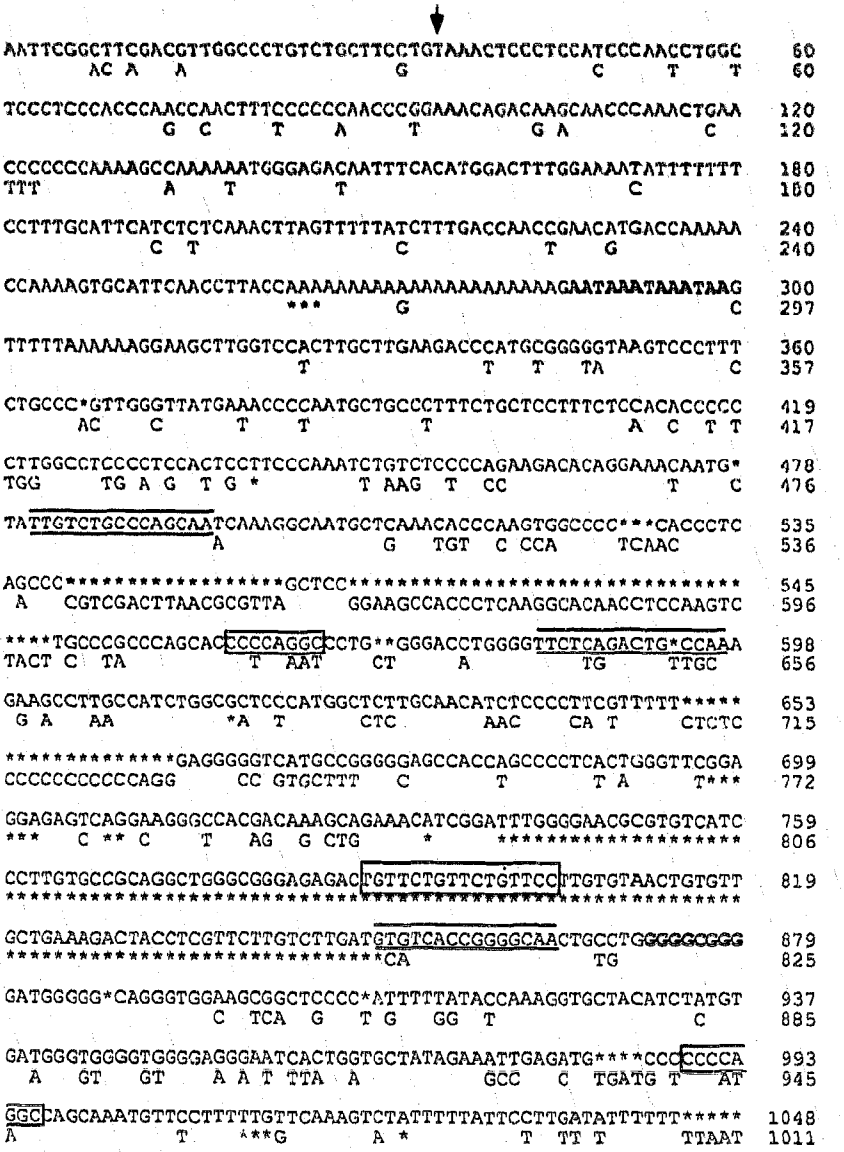

Fig. 2 (see following page for legend).

imately 200 bp of homologous sequence overlapping the pA2 site [16]. The murine gene has an additional AATAAA motif 38 bp upstream of the latter AATAAA signal. The poly-T stretch between the pA sites differs in size and base composition from the corresponding murine segment. The 3'-UTR contains putative gene regulatory motifs [17], e.g. two AP-2-sites, an Sp1-site, two viral core enhancers, one in each strand, an adenovirus E1A enhancer consensus sequence, an almost triplicated glucocorticoid responsive-element TGTTCT and several NF-1-like domains. Comparison of our sequence with previously published short human genomic and mRNA sequences [15,18] showed a few differences: residues 56 and 126 are Cs instead of Ts; beginning with nucleotide 96 our sequence has three As instead of two, and finally the poly-A stretch preceding the pA1 site consists of 22 nucleotides instead of 19.

The pA2 site or the equivalent pA5 site of the COL1A2 [19] is also extremely conserved. In human COL1A1 and COL1A2, mouse COL1A1 and chick COL1A2 genes the last AATAAA signal is followed by a GCATGT motif and a conserved element TGTACC-TATTTTGTAT (incomplete in chick) is found about 30 base pairs upstream from the pA2-site AATAAA motif. A/T versus G/C analysis (not shown) reveals three A/T rich domains located at the two pA sites and at the long poly-T-track. This poly-T-track is surrounded by an almost complete direct repeat of 11 nucleotides starting at residues 991 and 1133; a similar arrangement has been found in 3'-UTRs of many genes [20].

3.2. *DNA-binding proteins*

Herget et al. [4] demonstrated that a conserved 3'-UTR region ('tame sequence') of the rat COL1A1

```
**************************CTTTCTTTTTTTTTTTTTTTG********** 1069
GGATAGGGACTTGTGTGAATTGTTGGGG    T                GTTTTGTTTTT 1071

**TGGATGGGGACTTGTGAATTTTTCTAAAGGTGCTATTTAACATGGGAGGAGAGCGTGT 1127
TT T TGTTT TT    TTT    C G            G A    A   A   C    1131

GCG**CTCCAGCCCAGCCCGCTGCTCACTTTCCACCCTCTCTCCACCTGCCTCTGGCTTC 1185
 G GA  T  *   T    A CG  T  ****** A    T TAG T   GG C  AG   1184

TCAGGCCTCTGCTCTCCGACCTCTCTCCTCTGAAACCCTCCTCCACAGCTGCAG**CCCA 1243
  T AT    **********    T  T              C TCTT     T CT   T 1234

TCCTCCCGGCTCCCTCCTAGTCTGTCCTGCGTCCTCTGTCCCCGGGTTTCAGAG*ACAAC 1302
C  CT T C   T   T G          A  A*********          C  C     1285

TTCCCAAAGCACAAAGCAGTTTTT*CCCTAGGGGTGGGAGGAAGCAAAAGACTCTGTACC 1361
  T          A        A   CT   C         A   G              1345

TATTTTGTATGTGTATAA**TAATTTGAGATGTTTTTAATTATTTTGATTGCTGGAATAA 1419
                  ATA                                        1405

AGCATGTGGAAATGACCCAAACATAATCCGCAGTGGCCTCCTAATTTCCTTCTTTGGAGT 1479
                    TTGC TGTGCAT        TG   *** C  C G   CG 1462

TGGGGGAGGGGTAGACATGGGGAAGGGGCCTTGGGGTGATGGGCTTGCCTTCCATTCCTG 1539
     AG *********       TCC 1479

CCCTTTCCCTCCCCACTATTCTCTTCTAGATCCCTCCATAACCCCACTCCCCTTTCTCTC  1599

ACCCTTCTTATACCGCAAACCTTTCTACTTCCTCTTTCATTTTCTATTCTTGCAATTTCC  1659

TTGCACCTTTTCCAAATCCTCTTCTCCCCTGCAATACCATACAGGCAATCCACGTGCACA  1719

ACACACACACACACTCTTCACATCTGGGGTTGTCCAAACCTCATACCCACTCCCCTTCAA  1779

GCCCATCCACTCTCCACCCCCTGGATGCCCTGCACTTGGTGGCGGTGGGATGCTCATGGA  1839

TACTGGGAGGGTGAGGGGAGTGGAACCCGTGAGGAGGACCTGGGGGCCTCTCCTTGAACT  1899

GACATGAAGGGTCATCTGGCCTCTGCTCCCTTCTCACCCACGCTGACCTCCTGCCGAAGG  1959

AGCAACGCAACAGGAGAGGGGTCTGCTGAGCCTGGCGAGGGTCTGGGAGGGACCAGGAGG  2019

AAGGCGTGCTCCCTGCTCGCTGTCCTGGCCCTGGGGGAGTGAGGGAGACAGACACCTGGG  2079

AGAGCTGTGGGGAAGGCACTCGCACCGTGCTCTTGGGAAGGAAGGAGACCTGGCCCTGCT  2139

CACCACGGACTGGGTGCCTCGACCTCCTGAATCCCCAGAACACAACCCCCCTGGGCTGGG  2199

GTGGTCTGGGGAACCATCGTGCCCCCGCCTCCCGCCTACTCCTTTTTAAGCTT 2252
```

Fig. 2. Sequence of the COL1A1 3'-UTR downstream from the conserved *Eco*RI site in exon 51. The AATAAA polyadenylation signals and the consensus Sp1 sequence are presented in bold letters. Explanation of the symbols: *, bases missing in alignment; boxed, glucocorticoid responsive-element-like and AP-2 consensus sequences; double overlined, viral core enhancer-like sequences; underlined, Adenovirus EA1 enhancer-like sequence; under- and overlined, NF1-binding site-like sequences. The arrow shows the translation stop signal of the exon 51.

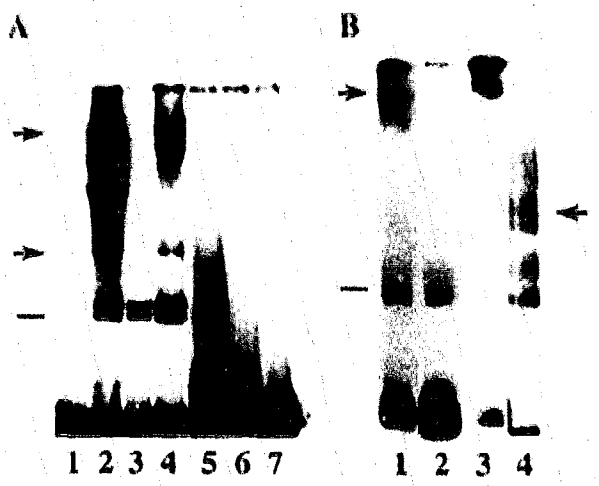

Fig. 3. Gel shift analysis of the pA1 (*TaqI*/*HindIII*) probe. (A) Probe only (lane 1), probe with 3 μg HeLa (lanes 2–4) and 12 μg NS-1 nuclear extracts (lanes 5–7). (B) Probe with 4 μg FTF (lanes 1–3) and 15 μg CTF (lane 4) extracts. For the competition in (A), 5 μg of a pGEM plasmid carrying the pA1 insert was added to reactions 3 and 6, and 5 μg of a pGEM plasmid with a nonspecific insert to reactions 4 and 7. In (B), 5 μg of the specific competing plasmid was added to reaction 2 and 5 μg of nonspecific plasmid to the reaction 3. Specific retarded complexes (—→) and nonspecific complexes (——) are indicated.

gene interacts with regulated nuclear proteins. We have extended these experiments to the pA2 site and, using three other cell lines, detected cell-specific binding of nuclear proteins to these regions (Figs 3, 4). A high molecular weight complex was formed between the pA1 probe and HeLa proteins (Fig. 3A, lane 2) and the specific binding was inhibited by an excess of non-labeled probe (lane 3). No specific binding was obtained with the NS-1 extract (Fig. 3A, lane 5); both specific and nonspecific competitors totally abolish the nonspecific binding (lanes 6 and 7). Nuclear proteins of human foetal (FTF) and embryonic chicken (CTF) tendon fibroblasts also recognize this region (Fig. 3B, lanes 1 and 4, respectively) but the FTF complex was more distinct. CTF extracts were tested because these cells are particularly suitable for transient transfections [1].

Similar results were obtained with the FTF, HeLa and NS-1 extracts and the pA2 probe (Fig. 4). In other experiments on RNA binding (manuscript in preparation) the same NS-1 and HeLa extracts bound to mRNA equally well indicating that the difference of the extracts in DNA binding reflected cell specific differences in the nuclear proteins.

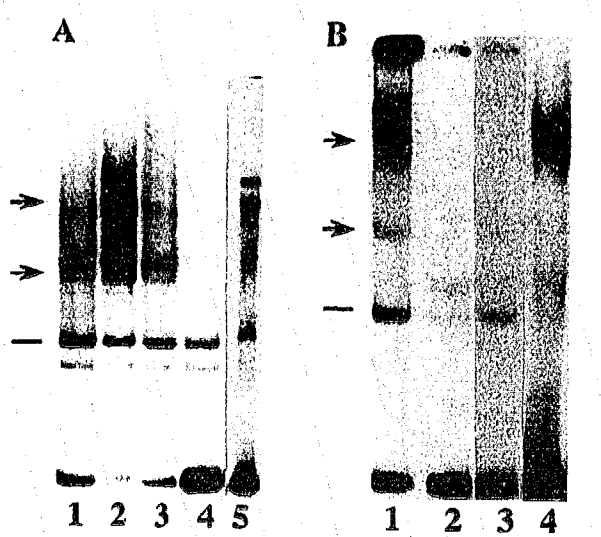

Fig. 4. Gel shift analysis of the pA2 (*PstI*/*Sau*3A1) probe. (A) FTF-extract in all reactions: (lane 1) 4 μg, (lane 2) 8 μg, (lane 3) 4 μg of extract plus 1 mM $ZnSO_4$, $MgCl_2$; (lane 4) 4 μg of extract plus 2 μg of specific competing plasmid; pGEM-1 with the pA2 insert; (lane 5) 4 μg of extract and 2 μg of nonspecific competitor. (B) (Lane 1) 1.5 μg of HeLa nuclear extract; (lane 2) 12 μg of NS-1 extract; (lane 3) 1.5 μg of HeLa extract and 2 μg of specific competitor; (lane 4) 1.5 μg of HeLa extract and 2 μg of nonspecific competitor. Specific (—→) and nonspecific complexes (——) are indicated.

4. DISCUSSION

4.1. *The structure of the COL1A1 3′-UTR*

Conservation of the polyadenylation sites has been observed in cDNA clones of collagens and other genes [21]. For example the human COL4A1 gene contains four AATAAA motifs, but only the most 3′ one, preceded by conserved nucleotides, is used [21]. However, in COL2A1 both the functional and the upstream nonfunctional polyadenylation signal regions are conserved [22] and the conserved region flanking the pA2 site in COL1A1 extends 5′ to the nonfunctional AATAA signal (Fig. 2). This might allow cell-specific binding of nuclear proteins to the nascent RNA. Flanking sequences 5′ to the pA1 and pA2 sites may be especially important for polyadenylation [23]. The genomic 21 bp poly-A stretch immediately before the first AATAAA signal might be functional in mRNA and associate with poly-A binding proteins which protect the mRNA from degradation [24].

Sequences which might affect the stability of the corresponding mRNA have also been identified in COL1A1. Examples are a poly-T stretch that can form a loop with the poly-A tail and the UAUUU and AUUUA motifs found in several copies in oncogene and other short-lived mRNAs [24,25]. In the plus strand of the 3′-UTR of COL1A1 there are 14 ATTT (AUUU) motifs, most of which are clustered around the poly-T region and around the pA2-site.

4.2. *DNA binding proteins recognizing 3′-UTR sequences*

In this paper we show cell-specific differences in binding of two conserved domains of the COL1A1 3′-UTR to nuclear proteins in fibroblasts, HeLa cells and NS-1 cells. Herget et al. [4] detected nuclear pro-

tein binding to the first of these sites (pA1) in rat myoblasts synthesizing type I collagen. The expression of both collagen and the binding protein were markedly reduced during myotube differentiation. It remains to be resolved whether the 3'-UTR-binding protein observed by us in fibroblasts and HeLa cells is a homolog of that in rat myoblasts and whether the same or different proteins bind the pA1 and pA2 sites in DNA. More studies are also needed to explain the role of the 3'-UTR in the expression of COL1A1 gene.

Acknowledgements: Supported by grants from the Turku University Foundation, City of Turku and The National Institutes of Health. We thank Dr Sirpa Jalkanen, Dept. of Microbiology, Univ. of Turku for the NS-1 myeloma line and Mrs Marita Potila for expert technical assistance. The sequence of the COL1A1 3' UTR is available in GenBank, Accession number M55998.

REFERENCES

[1] Liska, D.J., Slack, J.L. and Bornstein, P. (1990) Cell Regul. 1, 487–498.
[2] Karsenty, G. and De Crombrugghe, B. (1990) J. Biol. Chem. 265, 9934–9942.
[3] Boast, S., Su, M.-W., Ramirez, F., Sanchez, M. and Avvedimento, E.V. (1990) J. Biol. Chem. 265, 13351–13356.
[4] Herget, T., Burba, M., Schmoll, M., Zimmermann, K. and Starzinski-Powitz, A. (1989) Mol. Cell Biol. 9, 2828–2836.
[5] DePonti-Zilli, L., Seiler-Tuyns, A. and Paterson, B.M. (1988) Proc. Natl. Acad. Sci. USA 85, 1389–1393.
[6] Donoghue, M., Ernst, H., Wentworth, B., Nadal-Ginard, B. and Rosenthal, N. (1988) Genes Dev. 2, 1779–1790.
[7] Dorsett, D. (1990) Proc. Natl. Acad. Sci. USA 87, 4373–4377.
[8] Barth, G.S., Roush, C.L. and Gelinas, R.E. (1984) J. Biol. Chem. 259, 14906–14913.
[9] Penttinen, R., Kobayashi, S. and Bornstein, P. (1988) Proc. Natl. Acad. Sci. USA 85, 1105–1108.
[10] Ross, J. (1988) Mol. Biol. Med. 5, 1–14.
[11] Brawerman, G. (1989) Cell 57, 9–10.
[12] Raghow, R. and Thompson, J.P. (1989) Mol. Cell. Biochem. 86, 5–18.
[13] Nixon, B.T. (1989) in: Current Protocols in Molecular Biology (Ausbuel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A. and Struhl, K., eds) John Wiley and Sons, New York.
[14] Shapiro, D.J., Sharp, P.A., Wahli, W.W. and Keller, M.J. (1988) DNA 7, 47–55.
[15] Chu, M.-L., De Wet, W., Bernard, M. and Ramirez, F. (1985) J. Biol. Chem. 260, 2315–2320.
[16] Mooslehner, K. and Harbers, K. (1988) Nucleic Acids Res. 16, 773.
[17] Mitchell, P.J. and Tjian, R. (1989) Science 245, 371–378.
[18] Bernard, M.P., Chu, M.-L., Myers, J.C., Ramirez, F., Eikenberry, E.F. and Prockop, D.J. (1983) Biochemistry 22, 5213–5223.
[19] Myers, J.C., Dickson, L.A., De Wet, W.J., Bernard, M.P., Chu, M.-L., Di Liberto, M., Pepe, G., Sangiorgi, F.O. and Ramirez, F. (1983) J. Biol. Chem. 258, 10128–10135.
[20] Hennessy, S.W., Frazier, B.A., Kim, D.D., Deckwerth, T.L., Baumgarten, D.M., Rotwein, P. and Frazier, W.A. (1989) J. Cell Biol. 108, 729–736.
[21] Myers, J.C., Brinker, J.M., Kefalides, N.A., Rosenbloom, J., Wang, S.-Y. and Gudas, L.J. (1986) Nucleic Acids Res. 14, 4499–4517.
[22] Elima, K., Vuorio, T. and Vuorio, E. (1987) Nucleic Acids Res. 15, 9499–9504.
[23] DeZazzo, J.D. and Imperiale, M.J. (1989) Mol. Cell Biol. 9, 4951–4961.
[24] Bernstein, P. and Ross, J. (1989) Trends Biochem. 14, 373–377.
[25] Shaw, G. and Kamen, R. (1986) Cell 46, 659–667.