Volume 250, number 1, 99–105 FEB 07280 June 1989

Cloning and sequencing of the human nucleolin cDNA

Meera Srivastava[+], Patrick J. Fleming[+], Harvey B. Pollard and A. Lee Burns

Laboratory of Cell Biology and Genetics, National Institute of Diabetes, Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892 and [+]*Department of Biochemistry, Georgetown University School of Medicine, Washington, DC 20007, USA*

Received 5 May 1989

A cDNA containing the entire coding region for human nucleolin has been isolated from a λ gt10 human retinal library using a bovine cDNA probe. The cDNA hybridized to a transcript of 3000 bases from fast-dividing cells, as well as terminally differentiated tissues of several species. Translation of the nucleotide sequence revealed a long open reading frame which predicts a 707 amino acid protein with several distinct domains. These include repeating elements, four conserved RNA-binding regions, a glycine-rich carboxy-terminal domain and sites for phosphorylation, glycosylation and dibasic cleavage. Human and bovine nucleolin exhibited more additions and/or substitutions of aspartate, glutamate and serine residues in the chromatin-binding domains by comparison with the hamster and mouse nucleolins. These differences may be related to species-specific functions in transcription.

Nucleolin; Sequence; Retina; Homology

1. INTRODUCTION

Nucleolin, also known as C23, is an abundantly expressed acidic phosphoprotein of exponentially growing cells and has been shown to be mainly localized in dense fibrillar regions of the nucleolus [1,2]. This multifunctional protein [3] was shown to be involved in the control of transcription of ribosomal RNA genes by RNA polymerase I [4,5], in ribosome maturation and assembly [6,7], and in the nucleocytoplasmic transportation of ribosomal components [8]. The above activities are regulated not only by phosphorylation and proteolysis [9], but also by basic fibroblast growth factor [10] and by other thus far unidentified cytoplasmic regulators of nuclear activities [8].

While complete protein sequences for nucleolin have been isolated from hamster and mouse [11,12], the human nucleolin gene has thus far not been reported. This deficiency has precluded not only analysis of species-specific functions but also analysis of this protein in viral or chemically transformed human cells, in which one might anticipate this protein to have important regulatory functions. In the present paper, we describe experiments leading to our discovery of the human nucleolin gene.

Correspondence address: M. Srivastava, Department of Biochemistry, Georgetown University School of Medicine, Washington, DC 20007, USA

2. MATERIALS AND METHODS

2.1. *Construction of cDNA library from adrenal medulla*

Poly A^+ containing RNA from bovine adrenal medulla was purified by oligo(dT) column chromatography [13] of the guanidine thiocyanate isolated RNA [14]. The cDNA library was constructed using 2 μg poly A^+ according to a previously described technique [15]. The cDNA was treated with *Eco*RI methylase and the ends were blunted by the Klenow fragment of DNA polymerase I. *Eco*RI linkers were then added (0.5 μg linkers/μg cDNA) using T4 DNA ligase. After treatment with *Eco*RI, the free linkers were separated from cDNA by chromatography on a Bio-Gel A-50m column. 25 ng cDNA were ligated to 1 μg *Eco*RI-cleaved λ gt11. The phages were packaged in vitro and were amplified using *E. coli* strain Y1088.

2.2. *Isolation of bovine and human nucleolin cDNA*

Recombinant proteins expressed by the bovine cDNA library were probed with anti-cytochrome *b*-561 antibody as described [16,17], and a 610 bp partial cDNA clone coding for nucleolin was isolated. Other libraries (bovine brain library, Clonetech;

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/89/$3.50

bovine retinal library, Dr H.G. Khorana; and human retinal library, Dr Jeremy Nathans) were screened by plaque hybridization [18] with the bovine nucleolin cDNA, leading to the isolation of two human clones (CH2, 2.1 kb and CH4, 1.8 kb) and a 770 bp bovine clone. The human retinal library was rescreened using a 36 bp oligonucleotide (5'-AGCCACACCAGGCAAAGCATTGGTAGCAACTCCT-3') from the 5'-end of CH2 clone. Ten positive phages were isolated from 1×10^5 plaques and five contained the same full-length cDNA.

2.3. *Nucleotide sequence analysis*

*Eco*RI cDNA inserts or fragments generated by restriction with *Hae*III, *Rsa*I or *Alu*I were subcloned into M13mp18 for sequencing using the dideoxy chain termination method [19] with Sequenase. Sequence analysis was performed using the Microgenie Program of Beckman and secondary structure was predicted by the method of Garnier [20] using a computer program obtained from Dr Janet Finer-Moore [21].

2.4. *RNA blot hybridization*

Northern blot analysis was done by separating 5 μg poly A^+ RNA by electrophoresis on a 1.0% agarose gel in the presence of 2.2 M formaldehyde [22] and by transferring the RNA to a Nytran membrane (Schleicher & Schuell). The blots were then hybridized and washed according to the manufacturer's recommendations. Hybridization probes were prepared by the random-primer method [23].

2.5. *Other reagents*

Restriction endonucleases, T4 DNA ligase, T4 polynucleotide kinase, MLV reverse transcriptase, RNase H and DNA polymerase I were purchased from either BRL or New England Biolabs. Sequenase kits were obtained from United States Biochemicals.

3. RESULTS

3.1. *Determination of the human nucleolin sequence*

While screening a bovine adrenal medulla cDNA library using antiserum raised against cytochrome *b*-561, a major membrane protein of chromaffin cells [24], a clone was isolated which had 90% homology to hamster nucleolin [11]. Several other libraries were screened for full-length clones using bovine nucleolin cDNA. The largest insert (2570 bp) was obtained from a λ gt10 human retinal library, which when sequenced turned out to be full-length. The complete nucleotide sequence of the cDNA was obtained as illustrated in fig.1A and is shown in fig.1B. The clone contains a 114 bp 5'-untranslated region and a 2121 bp coding region. The initial ATG has a well conserved Kozak [25] consensus nucleotide sequence needed for translation initiation. The 332 bp untranslated region at the 3'-end includes the unusual polyadenylation signal TATAAA at nucleotide position 2424.

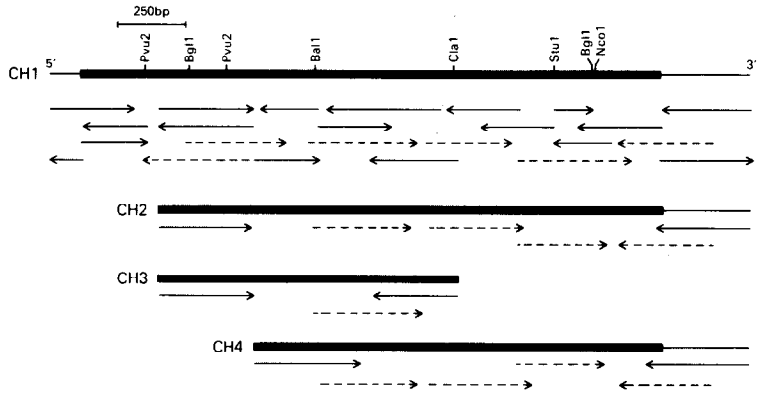
The open reading frame for human nucleolin cDNA codes for a 707 amino acid long protein of 77 kDa. The amino acid sequence in the coding region is 84% and 81% identical to that of hamster and mouse nucleolin cDNA, respectively. In addition, the sequence has distinct structural features such as: (i) highly charged acidic stretches at the amino-terminus with characteristic repeats; (ii) phosphorylation and glycosylation sites; (iii) ribonucleoprotein consensus sequences and (iv) a glycine-rich carboxy-terminal sequence. These data strongly support our identification of the clone as that of human nucleolin. Comparison of the amino acid sequences of human nucleolin with those of bovine, hamster, mouse and chicken is shown in fig.2 and is presented in more detail in section 4.

3.2. *Northern analysis of human nucleolin*

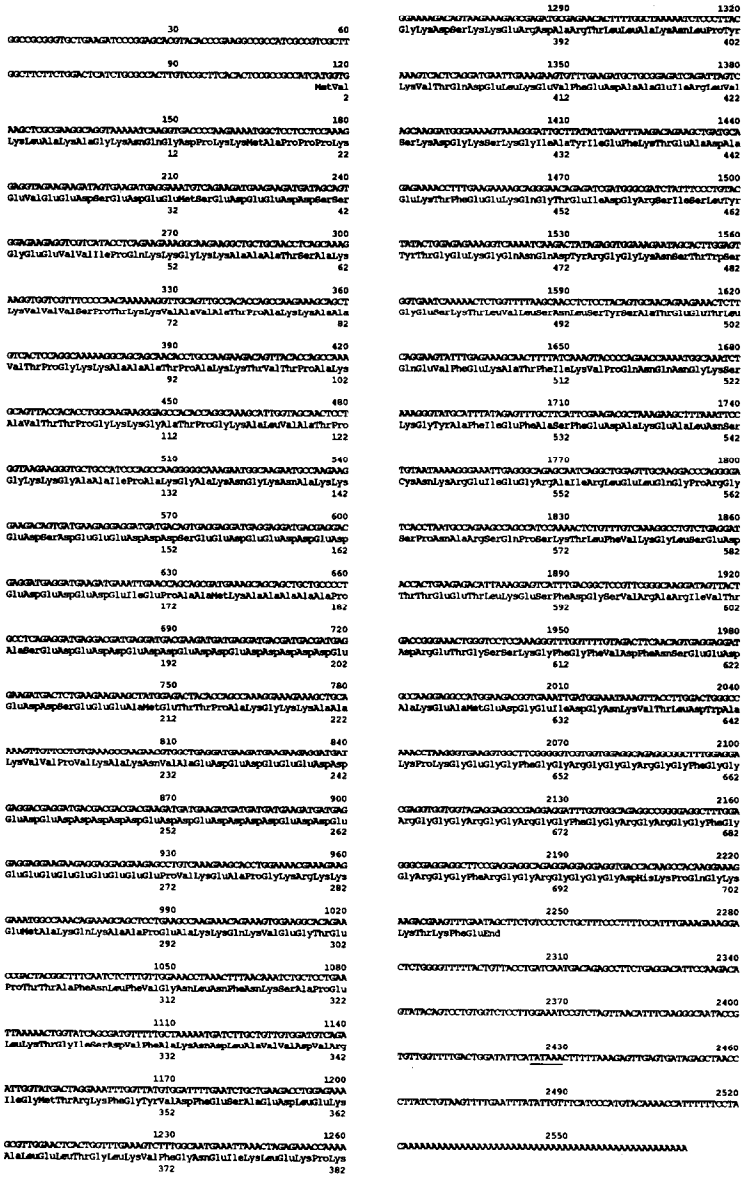
Poly A^+ RNA was prepared from various types of tissues and analyzed for expression of nucleolin mRNA by Northern blot. A major band of 3000 bp was detected in human retina, adrenal gland, liver, and a neuroblastoma cell line SK-N-SH. Similar results were obtained with bovine retina, adrenal medulla, liver, kidney and brain, plus rat brain and retina (fig.3). The two lower faint hybridizing bands seen in bovine liver and kidney could be alternative forms of nucleolin mRNA or degradation products. Expression of mRNA of the same size in different tissues is consistent with the notion that nucleolin is a ubiquitous protein involved with the critical processes of ribosomal RNA synthesis and subsequent ribosome maturation. As previously reported, nucleolin was found to be most abundant in ex-

Fig.1. (A) Restriction endonuclease map of nucleolin cDNAs and sequencing strategy. The thick line represents the coding region which is flanked by 5'- and 3'-noncoding regions indicated by thin lines. The direction of the sequence obtained from CH1, CH2, CH3, CH4 or restriction fragments using either specific primers (----→) or the universal primers (———→) are indicated by arrows. (B) Nucleotide and deduced amino acid sequence of the human nucleolin cDNA insert. Numbers for the nucleotides and amino acids are placed above or below their respective sequences. The presumed polyadenylation signal near the 3'-end is underlined.

A.


250bp
Pvu2
Bgl1
Pvu2
Bal1
Cla1
Stu1
Bgl1
Nco1
CH1
5'
3'
CH2
CH3
CH4

B.



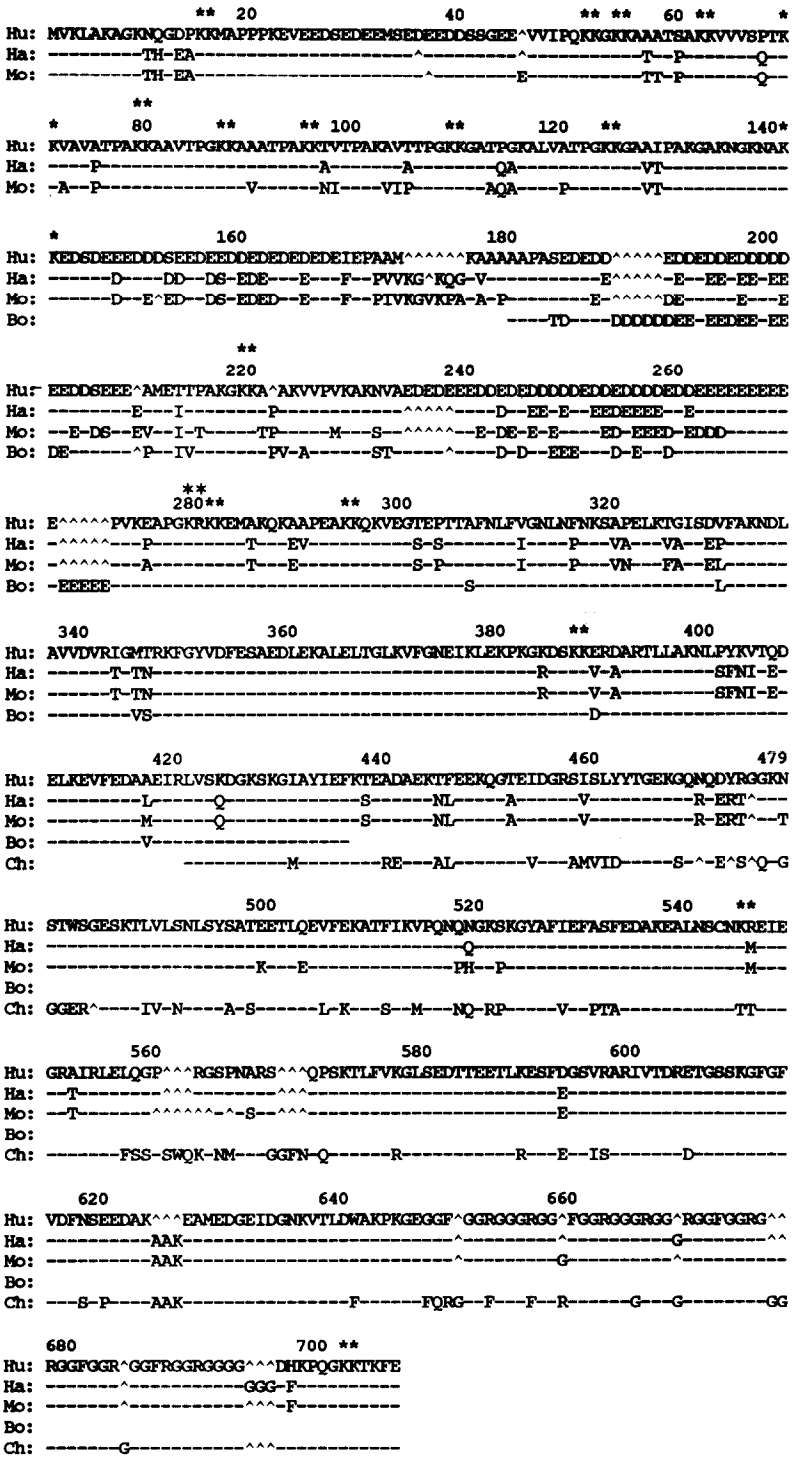


Fig.2. Comparison of the amino acid sequences of human and bovine nucleolin with hamster, mouse and chicken. Hu, human nucleolin; Ha, hamster nucleolin; Mo, mouse nucleolin; Bo, bovine nucleolin; Ch, chicken nucleolin. The numbering of the amino acids of human nucleolin is indicated above the single letter code (top sequence). Amino acids in the other protein sequences which are identical to human nucleolin are indicated by dashes and the amino acids which are different are denoted by single letters. Circumflexes indicate spaces introduced into the sequences for better alignment. Asterisks above the human sequence indicate potential dibasic cleavage sites.

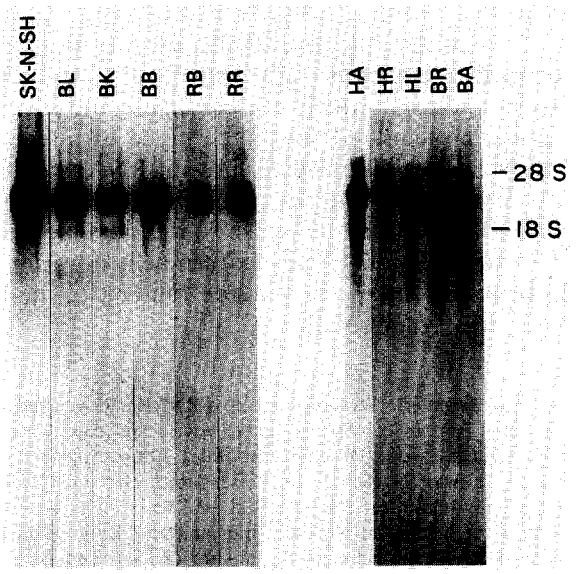


Fig.3. Northern blot analysis of nucleolin mRNA from different tissues. 5 μg from various tissues were analyzed. (SK-N-SH) Human neuroblastoma cell line; (BL) bovine liver; (BK) bovine kidney; (BB) bovine brain; (RB) rat brain; (RR) rat retina; (HA) human adrenal; (HR) human retina; (HL) human liver; (BR) bovine retina; (BA) bovine adrenal.

ponentially growing cells such as the neuroblastoma SK-N-SH. However, nucleolin synthesis was considerable in tissues containing nondividing cells such as retina, adrenal medulla and brain.

4. DISCUSSION

4.1. *Analysis of human nucleolin sequence*

A full-length human nucleolin cDNA from a retinal library was isolated using a bovine nucleolin partial cDNA. The 5′-untranslated region of 114 bp in human is approximately 50% different from hamster and mouse sequences. The first 105 nucleotides in the 3′-untranslated region of human nucleolin have a striking similarity (85%) to hamster and mouse, while the remainder of the nucleotides show considerable variation by 38% and 50%, respectively. As expected, the coding sequences for nucleolin in several species are highly conserved.

4.2. *Nucleolin structure and function*

We have observed that nucleolin has seventeen very conserved dibasic cleavage sites (15, 51, 54, 62, 70, 79, 87, 95, 109, 124, 141, 219, 279, 281, 294, 387, and 702). These could be the processing signals giving rise to different nucleolar proteins and possibly to other, as yet unidentified, biologically active peptides. Furthermore, human nucleolin has an additional site at position 545. The presence of potential dibasic cleavage sites for proteolytic processing has been observed in the case of chromogranin A [26] and some prohormones [27]. Nucleolin is highly susceptible to proteolytic degradation [28]. Polypeptides with masses of 95, 76, 70, 60 and 50 kDa have been identified with specific antisera to the 100 kDa protein, and it was suggested that these presumed nucleolin fragments resulted from thiol protease cleavage of the 100 kDa molecule [29]. It would be interesting to investigate which dibasic cleavage sites actually are utilized in vivo.

Recent reports suggest that karyophilic signals present in the proteins may be responsible for transporting cytoplasmic ribosomal components into the nucleolus [8,30]. The conserved, repeated sequence [TP(GAV)KK(GAV)2] at positions 59, 76, 84, 92, 106 and 121 could be putative nucleolar localization signals. These sequences contain a hydroxy amino acid, a proline, and paired dibasic residues, all of which are found in nuclear localization signals [31,32].

Other regions rich in aspartate and glutamate (positions 143–169, 184–209, and 234–272) are thought to be involved in binding to histones [33] and show variation between species. In the first segment, a potential phosphorylation site is displaced 5 residues toward the N-terminus of the human gene. In the second acidic segment, the phosphorylation site is abolished only in the bovine, which also has five additional glutamate residues relative to the rodent and human proteins. In the third acidic segment, the number of acidic amino acids is greatest in bovine (44), followed by human (39), hamster (34), mouse (34) and *Xenopus* (20). Most notably, in all three segments the substitutions and additions are only with aspartate or glutamate and occasionally with serine. In between the first two acidic segments, the human sequence exhibits both a deletion and high variability when compared to the hamster and mouse sequences. Conserved phosphorylation sites are located at position 28, 34, 145, 498, 580, 584 and 619. The variation in the number of acidic amino acids and in the location of phosphorylation

sites could be related to species specificity in the control of rRNA transcription [34], or to cation binding which could affect the specificity of phosphoprotein kinases [35]. Experiments involving deletion and mutation of the acidic domains might clarify the specific roles of these regions.

The four RNA-binding domains (positions 324–379, 406–461, 498–553 and 588–643) containing the central consensus sequences are present in human nucleolin. The fourth domain has the greatest conservation of sequence between species as recently reported in a comparison of the rodent and *Xenopus* sequences [36]. A glycine-rich carboxy-terminal domain (position 648–703) is also conserved and may be involved in protein-protein and/or protein-nucleic acid interactions [37]. The three additional glycine residues after position 695 seem specific for hamster nucleolin and 2 additional glycines after position 677 seem specific for chicken. Two conserved glycosylation sites are observed at position 317 and 492. The glycosylation sites at position 399 and 403, previously observed in hamster and mouse, are not detected in either human or bovine, and the glycosylation site at 478 is present in human and hamster, but not in mouse and chicken. We report an additional glycosylation site at position 541.

We also compared the predicted secondary structures for the human, hamster, and mouse nucleolin amino acid sequences. All three predicted structures were almost identical with more than 60% helix predicted for each protein. Even where the human sequence differed with the mouse or hamster, such as residues 401–405 and 470–474, the predicted secondary structures were essentially identical. Significant lengths of predicted α-helix were found in all regions of the protein except the glycine-rich carboxy-terminal region. Very large probabilities of α-helix formation were observed immediately on either side of the four consensus sequences (positions 344–354, 426–436, 520–530 and 607–617) of RNA binding. These predicted helix structures were approximately 10–15 residues in length. The consensus sequences themselves are predicted to have β-structures with low probability. Thus, it appears that a variation of the helix-turn-helix motif may be present in each of the four RNA-binding segments. A number of DNA-binding proteins with known structure exhibit a helix-turn-helix segment which interacts with the DNA [38]. In the case of nucleolin the turn segments (containing the consensus sequences) are between 10 and 20 residues in length which is larger than in the DNA-binding proteins. Although the mode of interaction in this region of nucleolin with RNA may be different than the interaction of DNA-binding proteins with double-stranded DNA, it is possible that the structural motif of a helix-turn-helix in various proteins has been adapted for binding to both types of polynucleotide.

Acknowledgements: We are grateful to Drs Karin Magendzo and Anat Shirvan for their valuable advice and cooperation throughout this work; Drs Frank Gonzales, Ettore Appella and Anil K. Jaiswal for their helpful discussions; Barbara Chung and Conny Cultraro for technical assistance; Dr Vicente Notario for helping in Northern blot analysis; Katie Sidman and Lois Hunt for helping in computer analysis and Delma Tyler for typing the manuscript. This work was supported by grants from USPHS (GM 27695) and the Cystic Fibrosis Foundation.

REFERENCES

[1] Escande, M.L., Gas, N. and Stevens, B.J. (1985) Biol. Cell 53, 99–110.
[2] Smetana, K., Ochs, R., Lischwe, M.A., Gyorkey, F., Freireich, E., Chudomel, V. and Busch, H. (1984) Exp. Cell Res. 1521, 195–203.
[3] Jordon, G. (1987) Nature 329, 489–490.
[4] Bousche, G., Caizergues-Ferrer, M., Bugler, B. and Amalric, F. (1984) Nucleic Acids Res. 12, 3025–3035.
[5] Olson, M.O.J., Rivers, Z.M., Thompson, B.A., Kao, W.Y. and Case, S.T. (1983) Biochemistry 22, 3345–3351.
[6] Herrara, A.H. and Olson, M.O.J. (1986) Biochemistry 25, 6258–6264.
[7] Bugler, B., Bourbon, H.M., Lapeyre, B., Wallace, M., Chang, J.H., Amalric, F. and Olson, M.O.J. (1987) J. Biol. Chem. 262, 10922–10925.
[8] Borer, R.A., Lehner, C.F., Eppenberger, H.M. and Nigg, E.A. (1989) Cell 56, 379–390.
[9] Bourbon, H.M., Bugler, B., Caizergues-Ferrer, M. and Amalric, F. (1983) FEBS Lett. 155, 218–222.
[10] Bousche, G., Gas, N., Prats, H., Baldin, V., Tauber, J.-P., Teissie, J. and Amalric, F. (1987) Proc. Natl. Acad. Sci. USA 84, 6770–6774.
[11] Lapeyre, B., Bourbon, H.M. and Amalric, F. (1987) Proc. Natl. Acad. Sci. USA 84, 1472–1476.
[12] Bourbon, H.M., Lapeyre, B. and Amalric, F. (1988) J. Mol. Biol. 200, 627–638.
[13] Aviv, M. and Leder, P. (1972) Proc. Natl. Acad. Sci. USA 69, 1408–1412.
[14] Chirgwin, J.M., Przybyla, A.E., MacDonald, R.J. and Rutter, W.J. (1979) Biochemistry 18, 5294–5299.
[15] Gubler, U. and Hoffman, B.J. (1983) Gene 25, 263–269.

[16] Young, R.A. and Davis, R.W. (1983) Proc. Natl. Acad. Sci. USA 86, 1194–1198.
[17] Young, R.A. and Davis, R.W. (1983) Science 222, 778–782.
[18] Benten, W.D. and Davis, R.W. (1977) Science 196, 1880–1882.
[19] Sanger, F., Nicklen, S. and Coulsen, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 560–564.
[20] Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) J. Mol. Biol. 120, 97–120.
[21] Finer-Moore, J. and Stroud, R.M. (1984) Proc. Natl. Acad. Sci. USA 81, 155–159.
[22] Hod, Y., Morris, S.M. and Hanson, R.W. (1984) J. Biol. Chem. 259, 25603–25608.
[23] Feinberg, A.P. and Vogelstein, B. (1983) Anal. Biochem. 137, 266–267.
[24] Winkler, H. and Westhead, E. (1980) Neuroscience 5, 1803–1823.
[25] Kozak, M. (1981) Nucleic Acids Res. 9, 5233–5252.
[26] Icangelo, A., Affolter, M.-U., Eiden, L.E., Herbert, E. and Grimes, M. (1986) Nature 324, 82–86.
[27] Docherty, K. and Steiner, D.F. (1982) Annu. Rev. Physiol. 44, 625–638.
[28] Lischwe, M.A., Richards, R.L., Busch, R.K. and Busch, M. (1981) Exp. Cell Res. 136, 101–109.
[29] Bugler, B., Caizergues-Ferrer, M., Bouche, G. and Amalric, F. (1982) Eur. J. Biochem. 128, 475–480.
[30] Peters, R. (1986) Biochim. Biophys. Acta 864, 305–359.
[31] Kalderon, D., Roberta, B.L., Richardson, W.D. and Smith, A.E. (1984) Cell 39, 499–509.
[32] Slomi, H., Shida, H., Nam, S.K., Nosaka, T., Maki, M. and Hatanaka, M. (1988) Cell 55, 197–209.
[33] Erard, M., Belenguer, P., Caizergues-Ferrer, M., Pantaloni, A. and Amalric, F. (1988) Eur. J. Biochem. 175, 525–530.
[34] Dhar, V.N., Miller, D.A., Kulkarni, A.B. and Miller, O.J. (1987) Mol. Cell. Biol. 7, 1289–1292.
[35] Mamrack, M.D., Olson, M.O.J. and Busch, H. (1979) Biochemistry 18, 3381–3386.
[36] Caizergues-Ferres, M., Mariottini, P., Curie, C., Lapeyre, B., Gas, N., Amalric, F. and Amaldi, F. (1989) Genes Dev. 3, 324–333.
[37] Chung, S.Y. and Wooley, J. (1986) Proteins: Struct. Func. Gen. 1, 195–210.
[38] Schleif, R. (1988) Science 241, 1182–1187.