Volume 216, number 2, 281-286 FEB 04741 June 1987

Nucleotide sequence coding for the human type IV collagen α_2 chain cDNA reveals extensive homology with the NC-1 domain of α_1(IV) but not with the collagenous domain or 3'-untranslated region

Sirkka Liisa Hostikka, Markku Kurkinen+ and Karl Tryggvason

Biocenter and Department of Biochemistry, University of Oulu, SF-90570 Oulu, Finland and +*Department of Medicine, University of Medicine and Dentistry of New Jersey, Robert Wood Johnson Medical School, Piscataway, NJ 08554, USA*

Received 2 March 1987

We have isolated two overlapping cDNA clones that provide the complete nucleotide sequence coding for the NC-1 domain and 3'-untranslated region of the α_2 chain of human type IV collagen as well as a sequence encoding 232 residues of the collagenous domain. An extensive homology was observed between the sequences of the NC-1 domain of the α_1(IV) and α_2(IV) chains, but considerably less between the sequences encoding collagenous and 3'-untranslated regions. There were four interruptions in the collagenous sequence studied whereas the comparable region of the α_1(IV) chain had only two. A potential oligosaccharide attachment site was found in a 6-residue long interruption of the collagenous domain but none in the NC-1 domain.

Basement membrane; Collagen; Sequence homology; (Human)

1. INTRODUCTION

Type IV collagen is the major structural component of basement membranes that is found only in these structures [1]. This collagen type is composed of two kinds of polypeptide chains, α_1 (IV) and α_2 (IV), that form a triple helic molecule. Type IV collagen differs from the major fibrillar collagens of type I, II and III, in that the collagenous Gly-X-Y repeat sequence is frequentyly interrupted, has a globular domain (NC-1) only at the C-terminus that is not removed extracellularly and forms a network-like structure instead of cross-striated fibrils [1,2]. The differences between type IV and fibrillar collagens are reflected at the gene level: the exons coding for the Gly-X-Y repeats of the α_1 (IV) and α_2 (IV) chains do not follow the 54 bp rule found in the genes of fibrillar collagens [3-5]. Thus, type IV collagen resembles more type VI, VII, VIII, IX and X collagens which do not form fibrils [6-10].

The amino acid sequence of the NC-1 domain of the human α_1 (IV) chain has been determined [11] as well as 907 residues from the C-terminal end of the triple-helical region [12] and 216 residues from the amino-terminal end [13], or about 70% of this 185 kDa chain. The amino acid sequence of the NC-1 domain and about 520 residues of the helical region of the mouse α_1 (IV) have also been reported [14-16]. An amino acid sequence of 511 residues from the collagenous domain of the mouse α_2 (IV) chain [17,18] and the predicted amino acid sequence of the NC-1 domain [19,20] were recently reported.

Here, we report the isolation and characterization of cDNA clones coding for 232 residues of the

Correspondence address: K. Tryggvason, Biocenter and Department of Biochemistry, University of Oulu, 90570 Oulu, Finland

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/87/$3.50 © 1987 Federation of European Biochemical Societies

C-terminal end of the helical region of the α_2 (IV) chain in man, the complete NC-1 domain and the 3′-nontranslating region.

2. MATERIALS AND METHODS

2.1. *Isolation of cDNA clones*

A human cDNA library in λgt11 (Clonetech, Palo Alto, CA) was screened with mouse α_2 (IV) coding cDNA clones PE18 and PE10A [3,20]. PE18 [3] is 1.1 kb and codes for a region of the helical part of the mouse α_2 (IV) chain whereas the 1.85 kb PE10A codes for the NC-1 domain and the untranslated region [20]. A mixture of the two ^{32}P-labelled cDNA inserts was used as probe to hybridize duplicate nitrocellulose filters at low stringency at 42°C in 5× Denhart's, 5×SSC, 0.1% SDS, 50 μg/ml ssDNA [21]. The filters were washed with 0.2×SSC, 0.1% SDS at 55°C.

2.2. *Other procedures*

The DNA sequence was determined by the use of M13 cloning [22] and the dideoxynucleotide sequencing method of Sanger [23] using the 'universal primer' as well as synthetic specific oligonucleotides for priming. Restriction endonuclease mapping was made on the cDNA inserts subcloned from λgt11 into the pBR322 plasmid *Eco*RI site and Southern analysis [24] was used to orientate the clones with respect to each other and known mouse cDNA sequences. Northern analysis of RNA was carried out as in [21].

3. RESULTS AND DISCUSSION

3.1. *Identification of cDNA clones*

Two overlapping human cDNA clones with insert sizes of 1.2 kb (HD-3) and 2.0 kb (HD-4) were isolated from the placenta cDNA library using the mouse α_2 (IV) coding cDNA clones as probes. Analysis of the clones demonstrated that they cover 2220 bp (fig.1), or about one-third of the mRNA (7.3 kb, not shown). Comparison of the encoded amino acid sequence with that reported for mouse [17–20] demonstrated that the HD-3 and HD-4 clones coded for the α_2 (IV) collagen chain. The sequence encodes 232 amino acids of the C-terminal end of the collagenous domain, the complete NC-1 domain and the untranslated 3′-region with a poly(A) tail (fig.2).

3.2. *Sequence of the α_2 (IV) collagenous domain*

The nucleotide sequence coding for the collagenous domain and the derived amino acid sequence, aligned with that of the human α_1 (IV) chain, is shown in fig.2. Out of 232 amino acid residues 105 or 45% are identical between the human α_2 (IV) and human α_1 (IV). However, the degree of homology was highly uneven between the residues of the Gly-X-Y repeats. Thus, 71 out of 75 glycine residues in the α_1 (IV) chain had a match in the α_2 (IV) chain, meaning that 95% of them were identical, whereas only 22% of the amino acids in the X-Y position of the Gly-X-Y repeats were identical (table 1). No cysteine residues were found in the collagenous sequence characterized in this study. The human α_2 (IV) sequence studied here contains four interruptions in the Gly-X-Y repeats (I–IV, fig.2) whereas the α_1 (IV) chain has only two in the same region. In two of the interruptions (II,IV, fig.2) the Gly-X-Y sequence was discontinued so that a glycine was replaced by an alanine. In both cases the codon for alanine was GCT. At comparable sites in the aligned human α_1 (IV) chain sequence there was a complete Gly-X-Y sequence with glycine codons GGG and GGT (see [11]).

In the third interruption of six residues (I, fig.2) two glycines had been substituted. This sequence is interesting since it contains an Asn-Ile-Ser sequence that could form a site for an asparagine-linked oligosaccharide reported to be present in type IV collagen [25]. This interruption is out of frame with the α_1 (IV) helical sequence because one amino acid has been deleted from the latter. This might result in the formation of one of the kinks in the type IV collagen molecule observed by rotary shadowing electron microscopy [1,2].

Nucleotide sequencing of the two cDNA clones studied here revealed one difference in the overlapping regions of the HD-3 and HD-4 clones indicating a polymorphic site in the coding region of the gene. This difference involved the codon for one of the amino acids in the 7-residue interruption (III, fig.2). As seen in the box in fig.2, the HD-4 cDNA clone had a codon ATC for isoleucine (−87) whereas the HD-3 cDNA clone had the codon GTC coding for valine at the same site. This polymorphic sequence was confirmed by restriction endonuclease digestion with *Fok*I whose recognition sequence is 5′...N_{13}CATCC...3′.

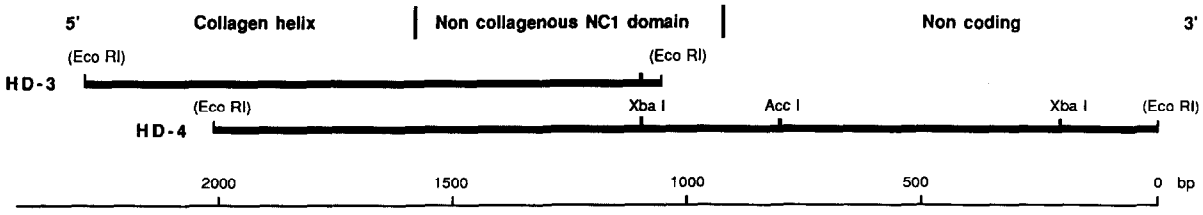

Fig.1. Map of two cDNA clones encoding human α_2 (IV) collagen. The position of the two cDNA clones (HD-3 and HD-4) with respect to the α_2 (IV) chain domains is illustrated at the top. The *Eco*RI sites in parentheses are contributed by linkers to the cDNA.

The enzyme cleaved the HD-4 but not the HD-3 cDNA clone (not shown). Interestingly, the mouse α_2 (IV) chain has a valine at this site [17,18] whereas the human α_1 (IV) has isoleucine ([11]; see fig.2).

Comparison of the sequence with that of mouse [17,18] showed that 80% of the nucleotides and 84% of the amino acids were identical. None of the nucleotide changes results in an exchange of glycine in the Gly-X-Y repeats.

3.3. *C-terminal noncollagenous (NC-1) domain*

The nucleotide sequence of the cDNA clones revealed that the NC-1 domain of the human α_2 (IV) chain contains 227 amino acid residues (fig.2). This is the same as in mouse [20] but two residues less than in the α_1 (IV) of man ([11]. fig.2) and mouse [14] (see fig.3). The amino acid sequence did not contain any Asn-X-Ser or Asn-X-Thr sequences indicating that the NC-1 domain does not contain any asparagine-linked carbohydrates. Comparison of the human α_2 (IV) chain amino acid sequence with that of the α_1 (IV) chain (fig.2) demonstrated an extensive homology between them as has recently been reported for mouse [19,20]. Out of the 227 amino acids in the α_2 (IV) chain NC-1 domain 147 amino acids or 65% were identical with the α_1 (IV) chain (table I). This means a homology of 65% which is considerably higher than what we observed between the sequences of the two chains in the collagenous domain (see above). In addition to the differences in the amino acid sequence, there are a few mismatches between the chains where an amino acid is missing from one chain. Thus Ser (+92), Ala (+94) and Glu (+200) in the α_1 (IV) chain do not have a match in the α_2 (IV) chain and Asn (+174) and Pro (+195) in the α_2 (IV) chain do not have a match in the sequence of the α_1 (IV) chain (see fig.2).

Comparison of the human α_2 (IV) sequence with that of mouse [19,20] demonstrated 77 differences in bases between the two species meaning that 89% are identical. Of those differences 66 involved the third base of the codon, and only one of them resulted in replacement of an amino acid. Out of the 227 amino acids residues of the NC-1 domain, only 8 differ between man and mouse yielding a homology of 97%. In seven of the eight cases an uncharged amino acid is replaced by an uncharged one and in one case an aspartate is substituted by a glutamate residue. None of the differences involve cysteine indicating that the intrachain bonds are essential for the structure and function of the protein.

The extensive homology of the amino acid sequence between the NC-1 domains of the α_1 (IV) and α_2 (IV) chain is further signified by the comparison of the sequences from both chains from man and mouse. The sequences were aligned for maximal homology (fig.3) and it can be seen that there is a conservation of all the cysteine residues and thus the three internal homologies (I–III) present in the two halves of the domain [11,14]. There is a conservation of about 64% of the amino acids between the two chains from both species.

3.3. *Nucleotide sequence of the 3'-untranslated region*

The 3'-untranslated region of the human α_2 (IV) cDNAs contains a TGA stop codon, 835 nucleotides and a poly(A) tail (fig.3). The base composition in the human α_2 (IV) chain was 24.8% A, 23.3% T, 22.2% G and 29.6% C when it was 31.3% A, 33.0% T, 16.6% G and 19.2% C in the human α_1 (IV) chain, respectively [12]. No ap-

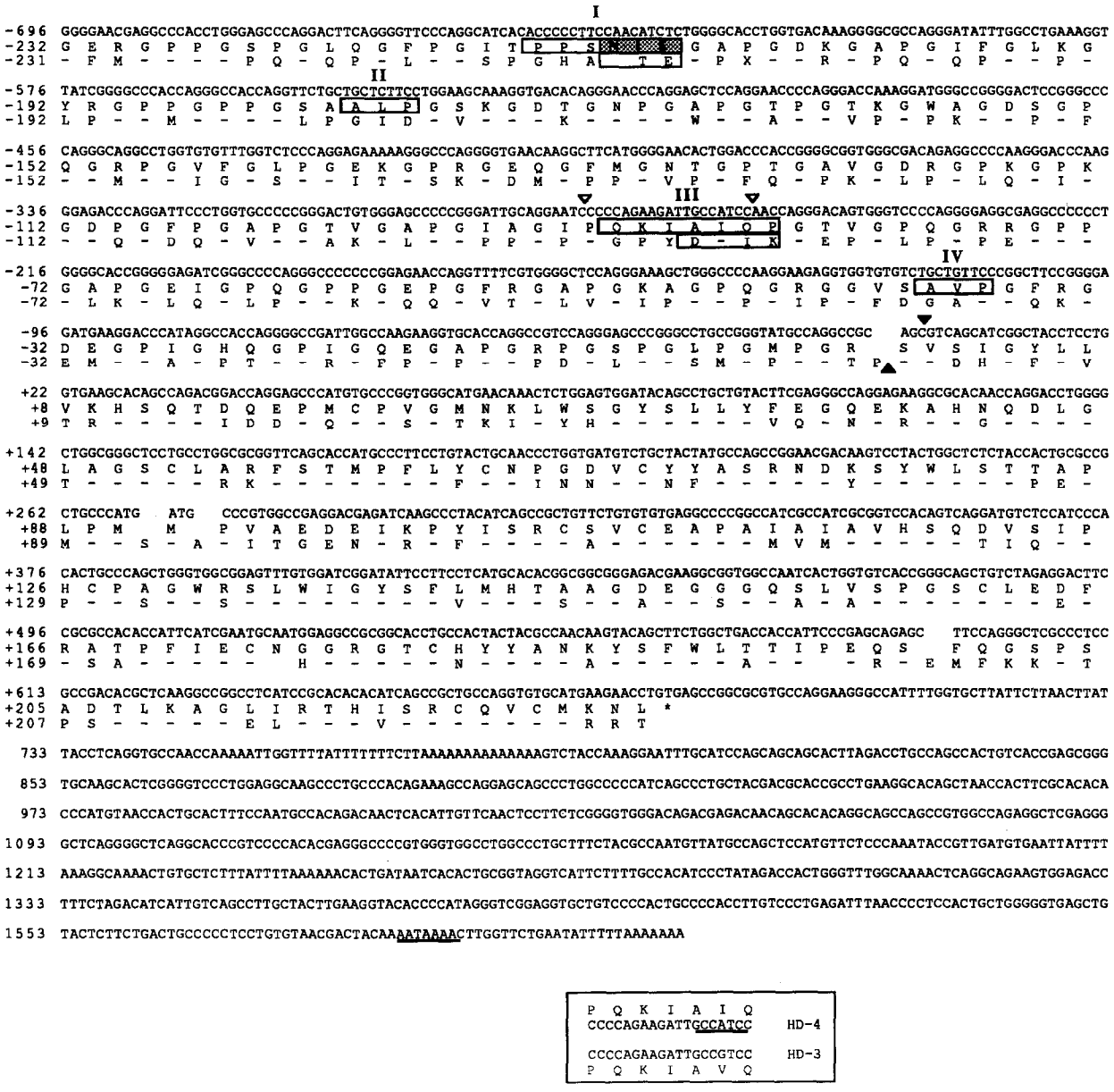

Fig.2. Nucleotide sequence and the deduced amino acid sequence for human α_2 (IV) chain coding cDNAs. The first row shows the nucleotide sequence of cDNA clones, the second row depicts the predicted amino acid sequence for the α_2 (IV) and the third row indicates the amino acid sequence of the human α_1 (IV) chain [11,12] where they differ from the α_2 (IV). Numbered boxes within the sequence surround interruptions in the collagenous Gly-X-Y repeats. The shaded part of box I illustrates a potential carbohydrate attachment site. Nucleotides (and amino acids) are numbered so that the last nucleotide of codon (or amino acid) for the Gly-X-Y domain is designated as '−1' and the first codon and amino acid of the NC-1 domain are designated as '+1' (see [11]). Filled arrowheads indicate the boundary of the collagenous and NC-1 domains. The open arrowheads designate a segment in the HD-4 clone with a recognition site for *Fok*I not present in HD-3 (see text). This segment from HD-4 is compared with the a segment from the same region of HD-3 clone in the separate box shown below. The single base substitution results in the replacement of isoleucine in HD-4 by valine in HD-3. The *Fok*I recognition site is underlined.

Table 1

Homology between amino acid sequence of human α_2 (IV) and α_1 (IV) chains

	Collagenous domain[a]		NC-1 domain
	Gly-	-X-Y-	
Identical residues (%)	95	22	65

[a] Involves 232 amino acid residues from the C-terminal end of the domain. Gly refers to the first and -X-Y- to the second and third amino acid in the collagenous Gly-X-Y repeats

parent homologies with cDNAs for pro-α_1 (I) and pro-α_2 (I) chains from man and chick, pro-α_1 (II) from chick and man, and pro-α_1 (III) from chick were observed.

3.4. *Conclusion*

The two cDNA clones described here provide the first complete sequence coding for the NC-1 domain and 3′-untranslated region of the human type IV collagen α_2 chain, as well as the sequence encoding 232 residues of the collagenous domain. A striking feature was the extensive homology between the sequence of the NC-1 domain of the α_1 (IV) and α_2 (IV) chains from both man and mouse [19,20] and considerably less homology between sequences coding for the collagenous domain or 3′-untranslated region. This high degree of conservation of sequences between two related proteins from different species indicates that the maintenance of structure of the NC-1 domain is crucial for the function of the protein, be it chain assembly or intermolecular crosslinking. It also suggests that the two genes have evolved from the same ancestor gene (see [19,20]). Another striking finding was the presence of interruptions in the collagenous Gly-X-Y repeat sequences that did not have complete matches in the α_1 (IV) chain. Such short discontinuities of the helix could explain the presence of kinks found in type IV collagen by rotary shadowing [1]. It was also interesting to note that only 22% of the amino acids in the X-Y position of the Gly-X-Y- repeats were conserved in both chains, whereas almost all the glycines, or 95%, were conserved. This indicates that there has been much higher tolerance for amino acid changes in the collagenous domain than in the NC-1 domain, given that the triple helical structure was maintained by conserving glycine in every third position. One triplet sequence, Asp-Ile-Ser, a potential attachment site for oligosaccharides, was found in one of the interruptions. Since mannose has been found in type IV collagen and there are no Asn-X-Ser or Asn-X-Thr sequences in the NC-1 domains or knowingly elsewhere in type IV collagen, this site could really be the attachment site for mannose.

ACKNOWLEDGEMENTS

This work was supported in part by grants from the National Research Council for Natural

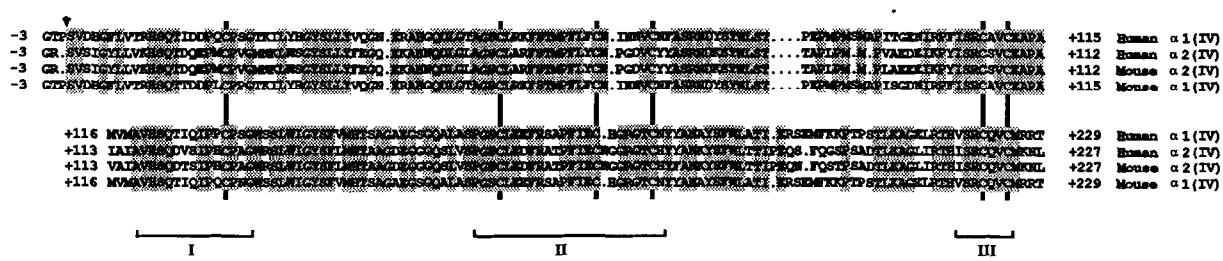

Fig.3. Alignment of the amino acid sequences of α_1 (IV) and α_2 (IV) NC1 domains fo human and mouse. The sequences of the two halves of the chains have been aligned for maximal homology. Identical residues in all four chains are indicated by shaded boxes. Cysteine-rich domains of internal homology [11,14] are indicated below.

Sciences, Academy of Finland and NIH, GM 34090.

REFERENCES

[1] Timpl, R. and Dziadek, M. (1986) Int. Rev. Exp. Pathol. 29, 1-112.
[2] Timpl, R., Wiedemann, H., Van Delden, V., Furthmayr, H. and Kühn, K. (1981) Eur. J. Biochem. 120, 203-211.
[3] Kurkinen, M., Bernard, M.P., Barlow, D.P. and Chow, L.T. (1985) Nature 317, 117-179.
[4] Soininen, R., Tikka, L., Chow, L., Pihlajaniemi, T., Kurkinen, M., Prockop, D.J., Boyd, C.D. and Tryggvason, K. (1986) Proc. Natl. Acad. Sci. USA 83, 1568-1572.
[5] Sakurai, Y., Sullivan, M. and Yamada, Y. (1986) J. Biol. Chem. 261, 6654-6657.
[6] Jander, R., Rauterberg, J., Voss, T. and Von Bassewitz, D.B. (1981) Eur. J. Biochem. 114, 17-25.
[7] Bentz, H., Morris, N.P., Murray, L.W., Sakai, L.Y., Hollister, D.W. and Burgeson, R.E. (1983) Proc. Natl. Acad. Sci. USA 80, 3168-3172.
[8] Sage, H., Trüeb, B. and Bornstein, P. (1983) J. Biol. Chem. 258, 13391-13401.
[9] Van der Rest, M., Mayne, R., Ninomiya, Y., Seidah, N.G., Greidien, M. and Olsen, B.R. (1985) J. Biol. Chem. 260, 220-225.
[10] Schmid, T.M. and Linsenmayer, T.F. (1985) J. Cell Biol. 100, 598-605.
[11] Pihlajaniemi, T., Tryggvason, K., Myers, K., Kurkinen, M., Lebo, R., Cheung, M.-C., Prockop, D.J. and Boyd, C.D. (1985) J. Biol. Chem. 260, 7681-7687.
[12] Babel, W. and Glanville, R.W. (1984) Eur. J. Biochem. 143, 545-556.
[13] Glanville, R.W., Qian, R., Siebold, B., Risteli, J. and Kühn, K. (1985) Eur. J. Biochem. 152, 213-219.
[14] Oberbäumer, I., Laurent, M., Schwarz, U., Sakurai, Y., Yamada, Y., Vogeli, G., Voss, T., Siebold, B., Glanville, R. and Kühn, K. (1985) Eur. J. Biochem. 147, 217-224.
[15] Nath, P., Laurent, M., Horn, E., Sobel, M.E., Zon, G. and Vogeli, G. (1986) Gene 43, 301-304.
[16] Schuppan, D., Glanville, R.W. and Timpl, R. (1982) Eur. J. Biochem. 123, 505-512.
[17] Schwarz, U., Schuppan, D., Oberbäumer, I., Glanville, R.W., Deutzmann, R., Timpl, R. and Kühn, K. (1986) Eur. J. Biochem. 157, 49-56.
[18] Vogeli, G., Horn, E., Carter, J. and Kaytes, P.S. (1986) FEBS Lett. 206, 29-32.
[19] Schwarz-Magdolen, U., Oberbäumer, I. and Kühn, K. (1986) FEBS Lett. 208, 203-207.
[20] Kurkinen, M., Condon, M.R., Blumberg, B., Barlow, D., Quinones, S., Saus, J. and Pihlajaniemi, T. (1987) J. Biol. Chem., in press.
[21] Maniatis, T., Fritsch, E. and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
[22] Messing, J. (1983) Methods Enzymol. 101, 20-78.
[23] Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
[24] Southern, E.M. (1975) J. Mol. Biol. 98, 503-517.
[25] Crouch, E. and Bornstein, P. (1979) J. Biol. Chem. 254, 4197-4204.