Volume 289, number 2, 260–264 FEBS 10150 September 1991

ADONIS 001457939100861H

Characterization of full-length cDNAs and the gene coding for the human GM2 activator protein

Horst Klima[1], Akemi Tanaka[2], Doris Schnabel[1], Takeshi Nakano[2], Maria Schröder[1], Kunihiko Suzuki[2] and Konrad Sandhoff[1]

[1]*Institut für Organische Chemie und Biochemie, Gerhard-Domagk-Straße 1, 5300 Bonn 1, Germany*, and [2]*Brain and Development Research Center, University of North Carolina School of Medicine, Chapel Hill, NC 27599-7250, USA*

Received 21 June 1991; revised version received 11 July 1991

Full-length cDNAs coding for the human GM2-activator protein has been isolated and characterized, and its genomic structure studied in two overlapping clones in λ-EMBL-4 isolated from a human brain genomic library. Two different cDNAs were found that were identical to the 5′-terminus to nt 1311 (counted from the A of the initiation codon, ATG) including the entire protein coding sequence. However, they were entirely dissimilar in the 3′-non-coding sequences. The genomic clones covered 94% of the full-length cDNA sequence. Three introns were found. The last exon spans contiguously the carboxyl terminus of the protein and the entire 3′-untranslated region of one of the two cDNAs with different 3′-ends. The origin of the 3′-portion of the other cDNA clone is not clear at this time.

GM2-activator protein; Genomic structure; Glycolipid-binding protein

1. INTRODUCTION

The GM2 activator protein is a lysosomal glycolipid-binding protein of a molecular weight of approximately 20 kDa. It is an essential component for in vivo degradation of the GM2 ganglioside by β-hexosaminidase A [1]. Genetic defects in the gene coding for the GM2 activator protein cause GM2-gangliosidosis AB variant, a severe, progressive and fatal neurological disorder [1]. Based on the information from purified GM2 activator and its amino acid sequence [2], a partial cDNA clone including the coding sequence of the entire mature GM2 activator protein was isolated [3]. In the present report we describe sequences of two full-length cDNA clones and their partial genomic organization. While the two cDNAs both encode an identical protein, they differ in the long 3′-untranslated sequences. The genomic sequence matches one of the two cDNA sequences.

2. MATERIALS AND METHODS

2.1. *Screening of cDNA Library*

A cDNA library prepared from cultured fibroblasts of a patient who had a juvenile form of Sandhoff disease [4] was screened with the partial cDNA reported earlier [3]. Several clones were isolated, their restriction maps evaluated and then sequenced. A few of them included the poly-A tract at the 3′-terminus but none included the initiation codon or 5′-untranslated sequence.

2.2. *Anchored PCR*

Several attempts at obtaining cDNA with the complete 5′-terminus by screening cDNA libraries were unsuccessful, and the 5′-most sequence was obtained by the 'anchored PCR methodology' (Fig. 1) [4,5]. Oligonucleotides synthesized were 5′-TGGATCCCCCCCCC-CCCCCC (No. 1,5′-end coding strand against poly dG tail with an additional *Bam*HI site), 5′-TCTGCAGGGGTCGCGAGAAGCAA (No. 2, non-coding strand of the 5′-terminus of the already known sequence with an additional *Pst*I site), 5′-GCTGCAGCTACTGA-GCTGGGATG (No. 3, non-coding strand at 20-bp downstream of No. 2 with an additional *Pst*I site), 5′-TCAGGCTTCTGATCACCG (No. 4, non-coding strand at 33-bp downstream of No. 3). The first reverse transcription was done with the primer No. 4 with AMV reverse transcriptase in the presence of 1 unit/μl of RNAsin at 42°C for 60 min. The DNA product was purified by phenol and chloroform extraction and precipitation from ethanol. A poly-dG tail was attached at the 3′-terminus with terminal deoxynucleotidyl transferase and dGTP at 37°C for 30 min. The segment was amplified using the primers No. 1 and No. 3. Finally the amplified products were electrophoresed, blotted to a Nylon membrane and screened with the oligonucleotide No. 2. The use of 3 different non-coding strand oligonucleotides for reverse transcription, PCR and the final screening was empirically necessary to reduce non-specific products. Finally, the region of agarose gel corresponding to the area of the positive signal was excised, DNA extracted and ligated into pUC18.

2.3. *Colony Hybridization*

Competent *E. coli* cells were transformed with the ligation products and grown on LB plates overnight at 37°C. Replicas were made with a Nylon membrane (Colony/Plaque Screen, NEN Research Products, Boston, MA). They were denatured in 0.5 N NaOH, neutralized in 1 M Tris-HCl, pH 7.5, washed in 3 × SSC in 0.1% SDS 3–5 times at room temperature and then once at 65°C for 5–6 h. Final screening was with the oligonucleotide No. 2, as above.

Abbreviations: β-hexosaminidase, β-*N*-acetyl-D-hexosaminidase (EC 3.2.1.52).

Correspondence address: K. Sandhoff, Institut für Organische Chemie und Biochemie, Gerhard-Domagk-Straße 1, 5300 Bonn 1, Germany. Fax: (49) (228) 73-56-83.

Published by Elsevier Science Publishers B.V.

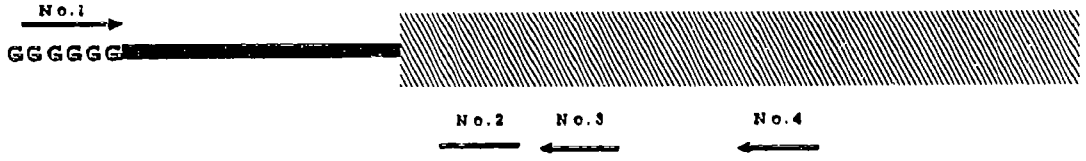

Fig. 1. Strategy for anchored PCR to obtain the 5'-terminus. The hatched bar represents the already existing cDNA, and the solid line the 5'-terminus being sought. Primer No. 4 was used to synthesize the first strand (non-coding). After attaching a G-tail, primers Nos. 1 and 3 were used for PCR amplification of the segment, and finally, the PCR products were screened using the oligonucleotide No. 4.

2.4. *Sequence Analysis*

The nucleotide sequence of the insert was determined by the dideoxy chain termination method [6] in the double-stranded form in both directions.

2.5. *Isolation of Genomic Clones*

A genomic library was prepared from human brain tissue by the method of Frischauf [7]. With the GM2 activator cDNA isolated from the plasmid by *Bam*HI digestion as the probe, 2.5×10^6 genomic clones were screened. The probe was labelled with ^{32}P by random priming (Random Labelling Kit from Pharmacia, Freiburg, Germany). Two positive genomic clones were identified and purified according to the standard protocol [8].

2.6. *Gene Mapping and DNA Sequencing*

The genomic clones were characterized by digesting with appropriate restriction enzymes (Pharmacia, Freiburg, Germany). The fragments were analyzed by agarose gel electrophoresis, Southern blotting

```
-58   AAGGCACC TCTGCCGCCA CAGACCTTGC AGTTAACTCC GCCCTGACCC ACCCTTCCCG
1     ATGCAGTCCC TGATGCAGGC TCCCCTCCTG ATCGCCCTGG GCTTGCTTCT CGCGACCCCT
61    GCGCAAGCCC ACCTGAAAAA GCCATCCCAG CTCAGTAGCT TTTCCTGGGA TAACTGTGAT
121   GAAGGGAAGG ACCCTGCGGT GATCAGAAGC CTGACTCTGG AGCCTGACCC CATCGTCGTT
181   CCTGGAAATG TGACCCTCAG TGTCGTGGGC AGCACCAGTG TCCCCCTGAG TTCTCCTCTG
241   AAGGTGGATT TAGTTTTGGA GAAGGAGGTG GCTGGCCTCT GGATCAAGAT CCCATGCACA
301   GACTACATTG GCAGCTGTAC CTTTGAACAC TTCTGTGATG TGCTTGACAT GTTAATTCCT
361   ACTGGGGAGC CCTGCCCAGA GCCCCTGCGT ACCTATGGGC TTCCTTGCCA CTGTCCCTTC
421   AAAGAAGGAA CCTACTCACT GCCCAAGAGC GAATTCGTTG TGCCTGACCT GGAGCTGCCC
481   AGTTGGCTCA CCACCGGGAA CTACCGCATA GAGAGCGTCC TGAGCAGCAG TGGGAAGCGT
541   CTGGGCTGCA TCAAGATCGC TGCCTCTCTA AAGGGCATAT AACATGGCAT CTGCCACAGC
601   AGAATGGAGC GGTGTGAGGA AGGTCCCTTT TCCTCTGTTT TGTGTTTGCC AAGGCCAAAC
661   TCCCACTCTC TGCCCCCCTT TAATCCCCTT TCTACAGTGA GTCCACTACC CTCACTGAAA
721   ATCATTTTGT ACCACTTACA TTTTAGGCTG GGGCAAGCAG CCCTGACCTA AGGGAGAATG
781   AGTTGGACAG TTCTTGATAG CCCAGGGCAT CTGCTGGGCT GACCACGTTA CTCATCCCCG
841   TTAACATTCT CTCTAAAGAG CCTCGTTCAT TTCCAAAGCA GTTAAGGAAT GGGAACCAGA
901   GTGTTTTAGG ACCTGAAGAA TCTTTATGAC TCTCTCTCTT TCACTCTTTT TTTTTTTGT
961   CACTAAGTTA AAAGCGAAGT GAGAGTATTA ACGTTTTTGT TCTCCTCCGG CCCCCTGTTA
1021  CAATGAAGGG GCAAAAGTAT TTGCTCTTAG TCTATTCCTC CCTTAACTTC TGTGACTAAT
1081  TTTTATTTCC TTTCTAGATT TGCCCAATTA ATACTAGGGT GCAGTGTATC CTGGAGAGGT
1141  AGGGTGTGTG GGGGAGGAAT CCCTTGGGGG AGATATTAGG AGTGCTCTGT TGTTTACAAA
1201  CTCACGGTAC CCGCAGGGCC TAGCAAGAGA CTTAAATGAC TGATAAGAAC CGTGAGAAAC
1261  ATGTTGCTTC CAGGCTTGAT TTCGATTTTT CGCTTTTTTT TTTTTGAGA C
1312  AGAATCTCAC TTTGTCACCA GGCTGGAGTG CAGTGGTGCA ATCTCACCTC ACTGCAACCT
1372  CCGCCTCCTG GGTTCAAGCA ATTCTCCTGC CTCAGCCTCC CAAGTAGCTT GGACTACAGG
1432  CCCTGCCACC ACGCCCGGCT AATTTGTGTA TTTTTAGTAG AGATGGGGTT TCACCATGTT
1492  GGCCAGGATG GTCTCGATCT CTTGACCTCG TGATCTGTCC ACCTTGGCCT TGCAAAGCGC
1552  TGGATTACAG GCATGAGCCA CTACACCCAG CCGATTTTTC CTTTTTGATT AAAGATGCTA
1612  TTACAATGTA AATATTTCTT ACACAGAAAG TCACAGCACA TGTGCCCATT GATACAAGGC
1672  TGCTGAGGCC TGGTCTCCAG TTGGAAATAT AATTAAGGGT GGCAAGGACT GGAGTCAGTT
1732  GGAGAGTGCA TAGCCAGTCT GTGAAGACAA CTGCCAGATA CTGGCAATAC TCCAGCCTGG
1792  TGACAGAGTG AGACTCTGTC TCAAAAAAAA AGTTTCAATG TTTACTCCTA GAGAAGCCAA
1852  AAATCCAGAT TTGTATATGA AATCTTACCA TTTTAAAAGA TTGGCAGCTA ATTATTTTTT
1912  TAAAAAGCTG TGCAGTGTGA TGTGTCCCAA ACGGACTGGC TCATGGGTGG CCACGTCACA
1972  ACCTCTGATC TCAGACCGTG CATGCCTTGT CCTCTTAAGA CAACTCCTGT GGCACCGTTT
2032  CTCCCTCCAC AGGGCCAAAG CCATAGTGTC CGGTCCCAAG GACAAGGCTC TTCCAGTGCT
2092  AGGAGAGGTA TGAGCAGCCT CTCACCTGTG AGCTGTGGGG ATCACAAGGC TGCCTGCCTC
2152  AGTCTTGGAG TCCTGTTGGG TGAATGAGGC AGATGGGAAA GAGCCTCACC AGCAGCTGCT
2212  TTTGGAGCAG GGGTCCAAGG AAGAGAGGGT GGCCTCGACA TCAAACTGCC TGGATTTTTC
2272  TACCACCCTG TTACATCATA ACAACTTCTG AAACACACAC CAGCCCTGAG TTCTGGGCTC
2332  ATTTGAAGCC TGGAATAGCA ATAAATCTTT TTAACTTGCG GACAGTT
```

Fig. 2. Nucleotide sequence of the full-length cDNA coding for the human GM2 activator protein (anchored PCR product plus #4–9). Two potential initiation codons and the termination codon are underlined. The stop codon in this sequence is TAA. In some instances, TAG was found. There is another polymorphism at nt 55 (bold-face) (A or G). The sequence of #6–10 differs entirely from nt 1312, and it also has one less T in the long stretch of Ts just upstream of the diversion point (bold-face). The downstream sequence of #6–10 after the diversion point is not shown here because the possibility of its being an artifact cannot be excluded. However, a hard copy of the sequence will be provided upon request.

Exon no.	Exon size (bp)	5'-Border	Intron (No.)	3'-Border	Intron (bp)
1	81	Lys^{27} 81 5' CTG AAA AAG	(1) (not available for analysis)	tgctggctta ttgtccccag	Pro^{28} 82 CCA TCC CAG	>1800
2	162	Lys^{81} 243 5' CCT CTG AAG gtgagcctgg gggtgggtgg	(2)	gttttatttt tttttaccag	Val^{82} 244 GTG GAT TTA	7300
3	182	Glu^{142} 426 5' TTC AAA GAA gtaagtactt agggaggaga gagcgttacc gtttggagag aagggtcttt gcattctcct tggatttgta agccagtgtg acctatcagg cctcagttat ccatctacga aatgggagac cagggccctt tatccatata atccatgctc tcatcttgtg cggctgtttt gagaatggga ctgcaatcct agcagtggct ctaggagaaa actgactggc ggtccactgg ctttcccgca	(3)	cctgtggcta aagagatggg tctgcagatc tgcatgtctc aatcacttat cttccgggag ttgaacttag atgtgatctt tacagtgcta tggccgtctc agaggggtgg tagttcatgg gaccccatca gtaggctccc g	Gly^{143} 427 GCA ACC TAC	381

Fig. 3. Exon/intron junctions of the GM2 activator gene. Exon nucleotides are written in upper case letters, and intron nucleotides in lower case. The numbers above the exon sequences refer to the nucleotide numbers in the GM2 cDNA (1 = A of the initiation codon, ATG). Above these numbers are the corresponding amino acids flanking each intron and their numbers. Intron 3 is relatively short and the complete nucleotide sequence is given. On the other hand, the last exon starting from nt 427 is 1951 bp long.

and hybridization. Fragments of interest were isolated with 'Glasmilk' (Biogen, La Jolla, CA, USA) and subcloned into pUC18. The recombinant plasmid DNA was isolated according to the method of Birnboim and Doly [9]. For sequencing, the plasmid DNA was further purified by columns (Qiagen-tip 20, Qiagen, Düsseldorf, Germany). The inserts were sequenced by the chain termination method [6], using the T7-Sequencing Kit (Pharmacia, Freiburg, Germany) and [^{35}S]dATP (Amersham, Braunschweig, Germany).

3. RESULTS

3.1. *Characterization of Two Different cDNAs*

Two cDNAs different at the 3'-untranslated sequence were obtained by screening the cDNA library. Both were incomplete at the 5'-terminus in that the initiation codon was not found. They were operationally num-

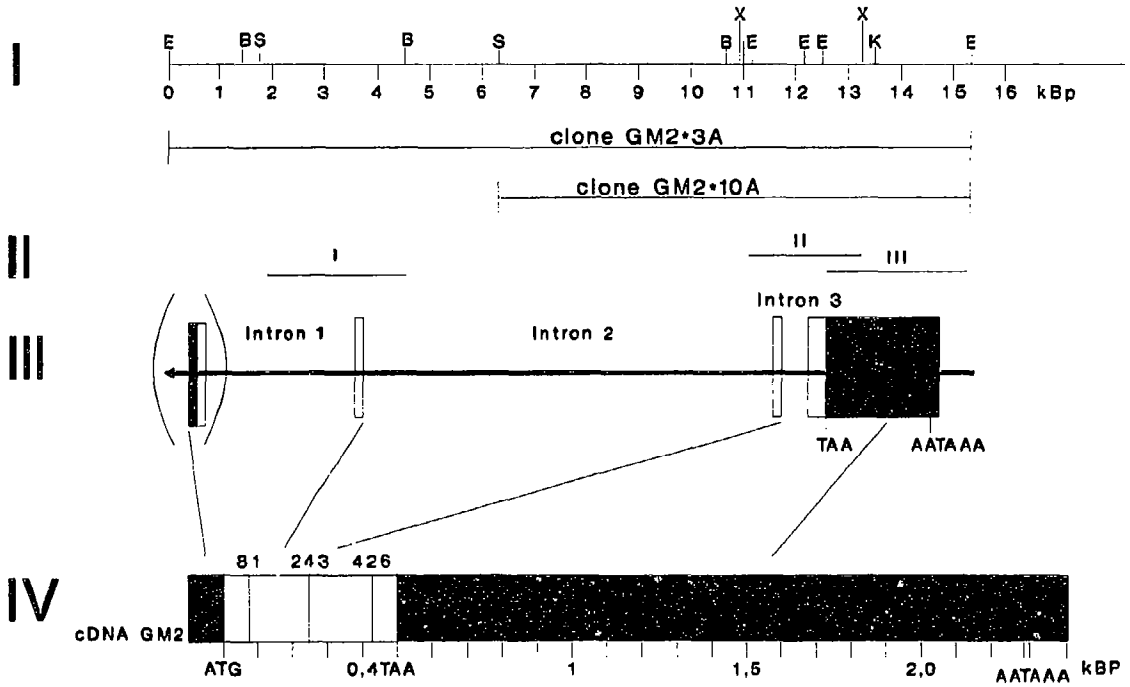

Fig. 4. Map of the GM2 activator gene. Two overlapping clones were isolated by screening a human genomic library with the cDNA for the GM2-activator. (I) Restriction map of the clones, GM2*3A and GM2*10A (B = *Bam*HI, E = *Eco*RI, K = *Kpn*I, S = *Sac*I, X = *Xba*I). (II) Sequenced regions. (III) The intron/exon organization. The open boxes refer to the exons. The black area belongs to the untranslated region of the cDNA. The putative exon 1 (in brackets) that covers the missing 5'-end of the gene is still missing. (IV) The full-length cDNA of the GM2 activator. Exons and introns are indicated as in (III). Position 1 is the A of the initiation codon, ATG.

bered #4–9 and #6–10, respectively. The protein coding sequences were identical for both cDNAs. Both had long 3'-untranslated sequences which remained identical until 729 bases downstream of the termination codon, except for an extra T in #4–9 at six bases upstream of the diversion point. From the diversion point, the sequences of the two cDNAs were entirely different without detectable similarities.

Seven genuine 5'-terminis of the cDNA clones were obtained out of 480 clones screened after the anchored PCR procedure. They added 100 nucleotides to the clone #4–9. Two potential in-frame initiation codons were found at 59- and 71-bases from the 5'-terminus. Since an in-frame termination codon is present at 32–34 bases, the cDNAs with the sequence added from the anchored PCR included the entire protein coding sequence. The combined length of the 5'-terminus and #4–9 is 2434 bp ending with a poly-A tract with a typical polyadenylation signal (Fig. 2), while that of the 5'-terminus plus #6–10 is 2034 bp without a typical polyadenylation signal or a poly-A tail. Both clones code for an identical protein of 194 amino acids and have very long 3'-untranslated regions – 1794 bp for #4–9 and 1394 bp for #6–10. No similarities could be detected between the two 3'-untranslated sequences.

Two polymorphisms were noted: A or G at nt 55 (Thr or Ala) and TAA or TAG for the termination codon.

3.2. *The Genomic Organization*

The gene coding for the GM2 activator was localized earlier on chromosome 5 [10]. Two overlapping clones for the gene of the GM2 activator were isolated from the λ-EMBL-4 human genomic library. They were appropriately digested, subcloned and sequenced. The sequence indicated that the 2 genomic fragments together covered 94% of the cDNA sequence, still missing the 5'-most portion of the gene. At least 58 of bp 5'-untranslated and 81 bp translated regions of the cDNA are still missing from the isolated genomic fragments. The gene coding for the human GM2 activator protein is relatively small, although the still missing 5'-terminus can possibly add a substantial size. The 3'-sequence corresponded to one of the two cDNAs described above (#4–9). Among the two polymorphisms noted in the cDNA, the stop codon in the genomic clones was TAA. The nt 55 of the cDNA was not within the isolated genomic sequence. While upstream exons are relatively short, the last exon is 1951 bp long from the 5-end to the polyadenylation site, spanning contiguously the last 51 amino acids of the protein and the entire 3'-untranslated region.

Three introns were identified. All introns are class 0 introns. They follow the ag/gt rule for the exon/intron boundaries. There is a good homology to the splice junctions [11]. Since the 5'-end of the genomic sequence is missing and since the missing portion includes the signal sequence which is often coded by its own exon, the number of introns may well be larger. From the available genomic sequence we can only conclude that the incomplete 'intron 1' is at least 1800 bp, intron 2 is 7.3 kb, and intron 3 is 381 bp long (Fig. 3). Intron 3 has been sequenced in its entirety, and intron 2 has been characterized by restriction digest and hybridization studies with oligonucleotides (Fig. 4). The 3'-region matches exactly that of the cDNA #4–9.

4. DISCUSSION

The partial cDNA we reported earlier included the entire coding region of the mature GM2 activator protein [3]. Both of the cDNAs described in this report included not only the entire protein-coding sequence. The 5'-portion obtained by anchored PCR is likely to represent the real 5'-terminus because multiple clones started from the same nucleotide. There are no experimental data to indicate which of the 2 initiation codons which are present only four codons apart, is used for translation. This is of no practical consequence because the nascent polypeptide is processed proteolytically at the N-terminal end to generate the mature activator protein starting at amino acid 32 (Ser) counted from the first methionine.

The origins of the two cDNAs with different 3'-untranslated sequences are of some interest. The 3'-untranslated sequence of #4–9 exactly matches the genomic sequence while that of #6–10 does not. Since the last portion of the coding region and the entire noncoding region of #4–9 are contiguous as a single exon, #6–10 cannot have been derived from hnRNA with an intron yet to be spliced out. The availabale data only allow us to conclude that #4–9 with the additional 5'-sequence represents the full-length cDNA coding for the GM2 activator protein but that the origin of the 3'-untranslated region of #6–10 is unclear. The possibility of this clone being an artifact cannot yet be excluded.

Acknowledgements: We thank Haeyoung Kwon for her excellent technical assistance throughout this investigation. This work was supported by grants from the Deutsche Forschungsgemeinschaft (SFB 284, C3), from the USA PHS (P30-HD03110, RO1-NS24289 and RO1-NS28997), and an Alexander von Humboldt Senior Scientist Award, IV-A-97 (to K. Suzuki).

REFERENCES

[1] Sandhoff, K., Conzelmann, E., Neufeld, E.F., Kaback, M.M. and Suzuki, K. (1989) in: The Metabolic Basis of Inherited Diseases, 6th ed (Scriver, C.R., Beaudet, A.L., Sly, W.S. and Valle, D. eds) pp. 1807–1839, McGraw-Hill, New York.

[2] Fürst, W., Schubert, J., Machleidt, W., Meyer, H.E. and Sandhoff, K. (1990) Eur. J. Biochem. 192, 709–714.

[3] Schröder, M., Klima, H., Nakano, T., Kwon, H., Quintern, L.E., Gärtner, S., Suzuki, K. and Sandhoff, K. (1989) FEBS Lett. 251, 197–200.

[4] Frohman, M.A., Dush, M.K., Agostino, S., Gomi, T. and Cantoni, G.L. (1988) Proc. Natl. Acad. Sci. USA 85, 8998–9002.

[5] Loh, E.Y., Elliott, J.F., Cwirla, S., Lanier, L. and Davies, M.M. (1989) Science 243, 217–220.
[6] Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463–5467.
[7] Frischauf, A.M., Lehrach, H., Poustka, A. and Murray, N. (1983) J. Mol. Biol. 170, 827–831.
[8] Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual, 2nd edition, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
[9] Birnboim, H.C. and Doly, J. (1979) Nucl. Acids Res. 7, 1513.
[10] Burg, J., Conzelmann, E. and Sandhoff, K. (1985) Ann. Human Genet. 49, 41–45.
[11] Breathnach, R. and Chambon, P. (1981) Ann. Rev. Biochem. 50, 349–383.