Vol. 178, No. 1, 1991 BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS
July 15, 1991 Pages 158-164

ORGANIZATION OF THE MOUSE ACID ß-GALACTOSIDASE GENE*

Eiji Nanba and Kunihiko Suzuki[1]

Brain and Development Research Center
Departments of Neurology and Psychiatry
University of North Carolina School of Medicine
Chapel Hill, NC 27599

Received May 29, 1991

ABSTRACT: Overlapping murine genomic DNA fragments containing the entire cDNA sequence were isolated from a cosmid library prepared from the DBA/2J strain of mice. The gene is more than 80-kb long, consisting of 16 exons. All of the exon-intron boundaries have been defined. The organization of the gene is highly conserved between the murine and human genes at least for the several exons and introns for which information for the human gene is available [Morreau et al., J. Biol. Chem. 264:20655, 1989]. Primer extension and RNase protection experiments indicated the presence of three potential transcription initiation sites, which are preceded by GC-rich SP1 binding sites but without typical TATA or CAT boxes, as is often the case for genes coding for housekeeping proteins. Compared to the cDNA sequence from C57BL/6J mouse, there were five nucleotide polymorphisms in the protein-coding region, two of which resulted in altered amino acids.

Genetic defects in the lysosomal acid ß-galactosidase gene cause GM1-gangliosidosis and Morquio B disease in humans [2]. Despite the ubiquitous presence of the Escherichia coli ß-galactosidase gene (lacZ) in all spheres of molecular biology, human acid ß-galactosidase cDNA has been isolated relatively recently [3 -5]. Nine mutations have so far been identified in the human gene that are underlying causes of GM1-gangliosidosis and Morquio B disease [6, 7]. Based on the expected homolgy with the human enzyme, we isolated and characterized a cDNA coding for the murine acid ß-galactosidase [8]. As anticipated, the murine enzyme was similar to the human enzyme in both the nucleic acid and amino acid sequences, but it was similar to the E. coli ß-galactosidase only in the nucleic acid sequence. In this report we describe isolation of the entire mouse acid ß-galactosidase gene in four segments from a genomic library in a cosmid. The organization of the gene has been characterized with respect to the transcription initiation site, approximate positions of exons, the nucleotide sequences of exon-intron junctions, and sites for several restriction enzymes.

* The content of this article was presented in part at the 22nd annual meeting of the American Society for Neurochemistry, Charleston, SC, March 1991 and published in an abstract form [1]. In order to maintain consistency among different laboratories, nucleotides of the mouse acid ß-galactosidase cDNA are numbered in this report with the adenine residue of the initiation codon, ATG, as nucleotide #1.

[1]To whom correspondence and reprint requests should be addressed.

0006-291X/91 $1.50

MATERIALS AND METHODS

Commercial Materials: Bethesda Research Laboratories (Gaithersburg, MD), Boerhinger-Mannheim (Indianapolis, IN), New England Biolabs (Beverly, MA), Pharmacia LKB Biotechnology, Inc. (Piscataway, NJ), Promega Corp., (Madison, WI), Fisher Scientific (Fair Lawn, NJ), Sigma Chemical Co, (St. Lousi, MO), and United States Biochemical (Cleveland, OH) were the main sources for standard enzymes, reagents, and other molecular biological supplies. Radioisotopes were purchased from ICN Biochemicals, Inc. (Costa Mesa, CA). Sources for nonstandard materials will be indicated below as appropriate.

Screening of Cosmid Library: The full-length mouse acid ß-galactosidase cDNA in pGEM 7Zf(+) [8] was purified by restriction enzyme digestion of the plasmid, preparative agarose gel electrophoresis, and electroelution. In addition to the full-length probe, three fragments were also prepared by restriction enzyme digestion and preparative electrophoresis from the 2353-bp cDNA; two fragments from the 5'-terminus, 178-bp (Eco RI - Taq I) and 444-bp (Eco RI - Apa I), and a 384-bp fragment (Apa I - Hind III, nt#370-753). These pieces of DNA were labelled with [^{32}P]dCTP by nick translation. A mouse genomic library in a cosmid [9] was screend with these labelled probes according to the standard procedures [10]. Clones positive by the high-stringency hybridization conditions described above [8] were isolated by the modified cleared-lysate method [11].

Purified cosmid clones were mapped for several restriction enzymes [11]. After a single or double digestions, fragments were screened with the probes described above and also for a series of oligonucleotide probes synthesized to match various segments of the exon sequences. The synthetic oligonucleotides used were nt# 405-426 (exon 4, sense), nt# 527-548 (exon 5, antisense), nt# 919-940 (exon 9, antisense), and nt# 2241-2261 (exon 16, antisense). The oligonucleotides were labelled at the 5'-terminus with T4 polynecleotide kinase and [γ-^{32}P]ATP. Positions of exons were further defined by additional restriction sites, Eco RI, Xba I, Bam HI, Hind III, and Sph I, within exons and exon-flanking regions of introns.

DNA Sequencing: Restriction fragments of the genomic clones that hybridized to the cDNA or oligonucleotide probes were subcloned into pGEM 7Zf(+). They were sequenced by the dideoxy chain termination method [12] with the T7 or SP6 promoter primer, or the above synthetic oligonucleotides corresponding to exon sequences and ^{35}S-labelled dATP [13] using the double-stranded plasmid as the template. The DNA polymerase used was a modified T7 polymerase (Sequenase version 2, US Biochem. Corp., Cleveland, OH). When it was necessary to sequence regions far from either of the promoter site, a series of nested deletional clones were prepared using exonuclease III and S1 nuclease [10]. With these procedures, nucleotide sequences of all exon-intron junctions were determined.

RNase Protection Experiment: RNase protection experiments were carried out according to the procedure of Zinn et al. [14]. A 710-bp Eco RI - Bam HI genomic DNA containing the 5'-terminal sequence of the ß-galactosidase cDNA was subcloned into pGEM 7Zf(+) using the corresponding poly-linker sites. After digestion with Eco RI, a ^{32}P-labelled RNA transcript was generated by SP6 RNA polymerase and [α-^{32}P]GTP (specific activity, 600 Ci/mmole). Total cellular RNA was prepared from cultured mouse fibroblasts using guanidine isothiocyanate [15]. Total RNA, 100 μg, was mixed with 5 x 10^5 cpm of the ^{32}P-labelled RNA in the hybridization solution [14]. After annealing, the mixture was digested with RNase A and RNase T1. The final products were electrophoresed on the sequencing gel consisting of 6% polyacrylamide/8M urea and radioactive bands detected in the same manner as for DNA sequencing.

Primer Extension Experiment

A 36-mer oligonucleotide complementary to the coding strand near the 5'-terminus of the mouse ß-galactosidase cDNA (nt# -29 to 7) was synthesized and labelled at the 5'-end with [γ-^{32}P]ATP and T4 polynucleotide kinase to a specific activity of 5 x 10^6 cpm/pmole and used as the primer for an primer extension experiment. The labelled primer (1 x 10^5 cpm) was mixed with 5 μg of poly A+ RNA prepared by oligo(dT)-cellulose chromatography in 30 μl of 22 mM Tris-HCl (pH 8.0), 660 mM NaCl, and 4.4 mM EDTA. After incubation at 85°C for 15 min, the mixture was allowed to cool slowly to 42°C and kept at this temperature for 2 hrs. The primer was extended according to Krug and Berger [16] with Moloney murine leukemia virus reverse transcriptase and the product was electrophoresed on the sequencing gel as for the RNase protection experiments. As a control, E. coli tRNA was used in place of the mouse poly A+ RNA fraction.

RESULTS

Out of twenty cosmid clones positive for presence of exon sequences, four were selected for further study. These four overlapping genomic segments covered an approximately 100-kb segment of the mouse genome and included the entire mouse acid ß-galactosidase gene, which spans approximately 80 kilobases and contains 16 exons (Fig. 1). The lengths of the exons generally range between 41-bp and 181-bp (Fig. 2). Only exons 15 (255-bp) and 16 (540-bp) are substantially longer. On the other hand, except for the very long intron 1 (22-kb) and the very short 80-bp intron 8, introns are generally 2- to 9-kb long. All exon-intron junctional sequences were consistent with the consensus sequence for the 5'-donor and 3'-acceptor sites [17, 18]. Comparison of the nucleotide sequence showed five disprepancies between the cDNA sequence from the C57BL/6J mouse we described earlier [8] and the exon sequences of the gene in this study; $C^{462}\rightarrow T$, $T^{804}\rightarrow C$, $T^{1320}\rightarrow C$, $A^{1549}\rightarrow G$, and $G^{1615}\rightarrow A$. Re-examination of the data indicated that these discrepancies are not due to sequencing errors but real differences between the two alleles. The discrepancies at nt #462, 804, and 1320 have no consequences because these do not affect the respective amino acids. However, $A^{1549}\rightarrow G$ changes Asn^{517} to Asp, and $G^{1615}\rightarrow A$ changes Gly^{539} to Arg.

The Eco RI - Bam HI fragment spanning from the 5'-terminus of the gene to approxiamtely 180 bases into intron 1 was sequenced in its entirety (Fig. 3). The primer extension and RNase protection experiments defined the 5'-terminus of the gene. The primer extension generated three products. Their sizes were 13, 14 and 55 bases after subtracitng the size of the primer (36-mer) (Fig. 4). These findings placed the probable transcription initiation sites at nt# -84, -43, and -42 (Fig. 3). The results of the RNase protection experiment were consistent with those of the primer extension experiments (Fig. 5). Two major products were generated. The larger of the two had an estimated size between 167 and 152 bases, while the smaller was about 40 bases shorter than the other. These values are consistent with the positions of the three transcription intiation sites estimated by the primer

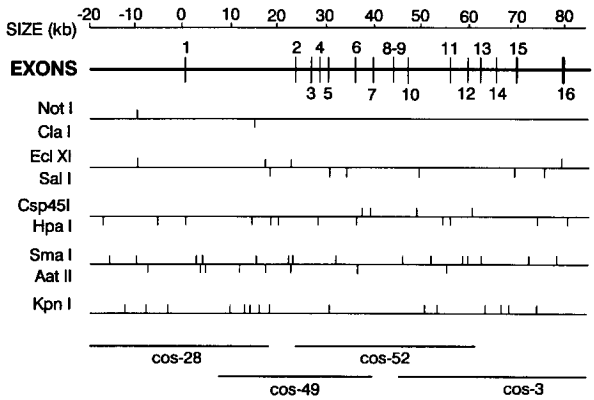

Fig. 1. Organization of the mouse acid ß-galactosidase gene. The positions of 16 exons, sites for nine restriction enzymes, and the sizes and regions of the four cosmid clones used for the study are given. The 5'-most clone included approximately 20-kb of the region above the gene (indicated by negative size designations). The gene starts at the 0-kb mark.

Exon #	Size (bp)	5'-Splice Site*	3'-Splice Site	intron size (kb)
1	≤162	CACGGCATCTAT(78)/gtaagtctccgg--------ggacttgccctgcag/AATGTCACCCAG		22
2	170	TGCTATCCAGAT(248)/gtaaggaaggat-------tgtgtgtcttcacag/GTACGTGCCCTG		3.4
3	151	GAGTGGGACATG(399)/gtgagcttccag-------ttttctctctcaaag/GGGGGCTTACCT		1.9
4	61	CTTCTGACCCAG(460)/gtgagttaccgt-------catcctgttttgcag/AATACCTTGTAG		2.3
5	95	ATAACCGTGCAG(555)/gtacccggggat-------tttctaaacacacag/GTTGAGAATGAG		6.5
6	181	ATTTTGGAACAG(736)/gtcggtgtgtgt-------tcttgaaatttccag/GCAACAATATCA		3.8
7	59	AAAGGACCTTTG(795)/gtaagaactggg-------ccttttccctttcag/ATCAATTCCGAG		3.2
8	122	CAACGTGAACTT(917)/gtgagcacccag-------cttatttctctgtag/GTACATGTTTAT		0.08
9	41	CCTATTGGAATG(958)/gtaagatccgtt-------ttttctgtcttccag/GTGCCAACACGC		5.1
10	113	GTCATTCAGATG(1071)/gtgggtccttcg------cagtgtgttttatag/TTTAAAGAAGTC		7.7
11	75	GCTCTGAGAAAG(1146)/gtaagacaggat------tctgcttctttgcag/TTCAAGACAGTG		3.6
12	90	CAGGTAAAACAG(1236)/gtaggtctgtct------ttcttttctcaccag/TATTTTGGGTAT		2.2
13	117	TCTGTGGACGGG(1353)/gtaagcattctc------ggctttaactcacag/GTCCCCCAAGGA		3.0
14	132	AATGACTTCAAG(1485)/gtaggacacttg------gtttccttttttccag/GGTTTGATTTCC		4.8
15	255	GGGTGGTCCAAG(1740)/gtatgcatcttc------tctccctctctctag/GGTCAAGTATGG		9.1
16	540	GTCATCTTTCTGCTGAATAAAATTGGTTTGCAAGT/tgtccccctc---		

*The numbers in parentheses indicate the cDNA nucleotide numbers of the last base of respective exons, counted from the adenine residue of the intiation codon, ATG.

Fig. 2. Locations, sizes and junctional sequences of exons and introns of mouse acid ß-galactosidase gene. The 80-bp intron 8 has been sequenced in its entirety (data not shown).

extension to the 3'-terminus of exon 1 (Fig. 3). Then, the size of exon 1 can be 162-, 121- or 120-bp. The typical GC-rich consensus sequence for the SP1 binding site, GGGCGG, is present at 14-, 26-, and 31-bases upstream of the 5'-most transcription initiation site. The gene lacks the typcial TATA or CAT boxes.

DISCUSSION

The mouse acid ß-galactosidase gene is known to be on chromosome 9 [19, 20]. It is relatively large for a lysosomal enzyme. The sixteen exons comprising the protein coding sequence were dispersed more or less evenly within the gene, and the 5'-terminus appeared typical for a housekeeping gene with GC-rich regions and without obvious TATA and CAT boxes just upstream of the transcription initiation site. Such an organization of the promotor region is often characteristic of housekeeping genes. Among the genes coding for lysosomal enzymes, that for arylsulfatase A has a smimilar promoter organization [21].

1 GAATTCCTCTGTAACCTACTACAGTGCTTCTTCACCTTTGCTTTCACTGAACCTCGGAGC 60

61 TGGGTCCGGGTCTAGGACTTCAGGGTTTCTGGGCTGGGGCCCAGAGTCTGCATTTCCATG 120

121 TTTAAGTCTCGTTGGCCTGTGGACCACACCTTGTGAATGTTAGACCACGTGCTAGGGAGA 180

181 GTTTTAGGGGCTAGGCTTCTAAGGCCAGAGCACTGAGGACTGACTGGCTGGTTTCCTCTT 240

241 CCCTGCTCAGTTCTTTTCTGGCAGCCGGGACAGGCTGAAGGCCCTGTGAATGGCCAGCCC 300

301 CTGAACTTCATCATTGCTCATTCAGGGGTGCCGCCATTCCGGAAGGAAACACAGTTGTGG 360

361 GAGATGCATAGACCACAGCTGCTTCTGAGTCACGTGGCCTACAGAGGCTGGACCTCCACT 420

421 CGCAGCACAGAGCCTCGGGGCGGGGCGGAGCTGGGGGCGGGTCCTGGCATAGGGTCCTCA 480

481 ACAACCCCCGGAGGTGCAGCGGCTGGCCAGAGCGCCCACTGCCTAACGGAGAGACCCCAT 540

541 CGTGGCGCGATCATGCTCCGGGTCCCCCTGTGTACGCCGCTCCCGCTCCTGGCACTGCTG 600

601 CAACTGCTGGGCGCTGCGCACGGCATCT*ATGTAAGTCTCCGGTGCCACCGCGCGGGGACG* 660

661 *CTCCGCCGCTGGGAAGCGGGCAGAGGCCGAGGAAAGCCGCGGCTCTGCAAAGTCGGGATCC* 721

Fig. 3. 5'-terminus of mouse acid ß-galactosidase gene. The Eco RI - Bam HI fragment covering this region was sequenced (sites underlined). The initiation codon, ATG, is double-underlined. Possible transcription initiation sites, as indicated by the primer extension and RNase protection experiments,are in bold face and double-underlined (A, G, and A). In the upstream of these transcription initiation sites are regions of the consensus sequence consistent with the SP1 binding sites (also underlined). This nucleotide sequence is in the GenBank database (accession #M63242).

Previously a high degree of similarity was noted between the nucleotide sequences of the human and murine acid ß-galactosidase cDNA [8]. The few exon-intron junctions in the human gene defined by Morreau et al. [4] matched exactly with the exon-intron junctions of the murine gene. Based on the determined positions of introns in the mouse gene and on homology of the two genes, we predict locations of introns in the human gene. The junctions in the human gene already defined by Morreau et al. [4] and known to match between the two species are marked with asterisks. Others are our predictions. Roman numerals indicate intron numbers and Arabic numerals in parentheses indicate the nucleotide numbers of the last nucleotides of the preceding exons counted from adenine of the initiation ATG (add 50 to arrive at the nucleotide numbers as given by Morreau et al. [4]); I (75), II* (245), III (396), IV* (457), V* (552), VI* (733), VII (792), VIII (914), IX (955), X (1068), XI (1143), XII (1233), XIII (1348), XIV (1479), XV (1734). These predictions have been used in our recent study on mutations in human GM1-gangliosidosis patients to generate exon-specific PCR primers [7]. The results of the genomic DNA amplification using these primers were consistent that our predictions of the intron locations in the human gene are probably correct.

As indicated in the Results section, we found differences in five locations within the protein coding sequences between the mouse cDNA sequence we reproted earlier [8] and that of the present study. We believe that these differences represent genuine polymorphisms

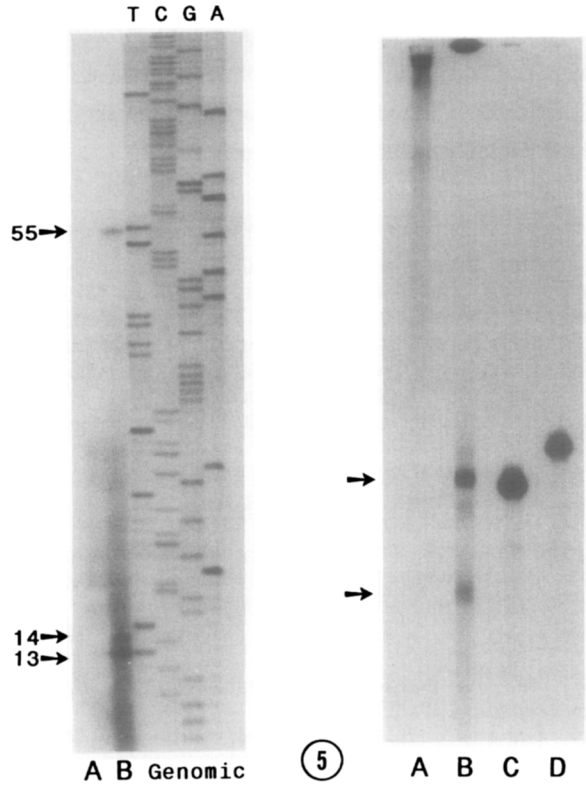

Fig. 4. Primer extension experiment. Technical details are described in the text. Lane A: E. coli transfer RNA was used as the control template (no distinct bands); Lane B: Five μg of mouse fibroblast poly A+ RNA fraction was used as the template for primer extension. Three distinct bands are visible, corresponding to 13-, 14-, and 55-bases, respectively, above the primer sequence. The genomic fragment encompassing this region was sequenced with the same primer ("genomic") for identification of the sizes and nucleotides corresponding to the three bands in the primer extension experiments. The bands in the primer extension experiments match to T, C, and T (A, G, and A in the sense strand; also see Fig. 3).

Fig. 5. RNase protection experiment. Technical details are give in the text. Lane A: synthesized RNA without RNase treatment; Lane B: after RNase treatment; Lane C and D: RNA size standards prepared from templates with known lengths. The sizes are 152-bp (C) and 167-bp (D). The two major bands in the experimental lane (B) suggest that transcription is initiated at two locations consistent with the results of the primer extension experiment (see Fig. 4).

between the two inbred strains used in the two studies. Our previous cDNA was isolated from a library from a C57BL/6J mouse, while the genomic library used for the present study was from a DBA/2J mouse. In a preliminary study, we observed restriction fragment length polymorphisms between these two strains in the acid ß-galactosidase gene that were detectable by Eco RI and Xba I digestion with the full-length cDNA as the probe (Nanba and Suzuki, unpublished observation). Then, we must conclude that the amino acid at position 517 can be either Asn or Asp, and that at position 539 can be either Gly or Arg without inactivating the catalytic activity of the resultant enzyme protein. Asn^{517} is not a glycosyltation site. In this regard, it is of interest that ß-galactosidase activity has been reported to be substantially lower and the concentration of GM1-ganglioside higher in DBA mice, compared to C57BL mice [22, 23].

GM1-gangliosidosis due to genetic deficiency of acid ß-galactosidase activity occurs in some larger animals, such as dogs, in addition to humans. However, no genetically authentic GM1-gangliosidosis is known among smaller laboratory animals. The availability of the murine genomic clones and the knowledge of its organization should help generating a mouse of the disease by the homologous recombination technology.

ACKNOWLEDGMENTS

The mouse genomic library in cosmid used to isolate the ß-galactosidase gene was kindly provided by Dr. Brian Popko at our Center. We thank Mr. Jose Langaman for his excellent assistance in the fibroblast cell culture. The oligonucleotide probes were synthesized in the Nucleotide Synthesis Laboratory of the Program in Molecular Biology and Biotechnology of the University of North Carolina under supervision of Dr. Dana Fowlkes. This investigation was supported in part by research grant, RO1 NS-24289, and Mental Retardation Research Center Core Grant, P30 HD-03110 from the United States Public Health Service.

REFERENCES

1. Nanba, E., and Suzuki, K. (1991) Abstract for the 22nd annual meeting of the American Society for Neurochemistry, Charleston, NC, March 10-15.
2. O'Brien, J. S. (1989) in *The Metabolic Basis of Inherited Disease* (Scriver, C. R., Beaudet, A. L., Sly, W. S. and Valle, D., eds.), pp. 1797-1806, McGraw-Hill, New York.
3. Oshima, A., Tsuji, A., Nagao, Y., Sakuraba, H., and Suzuki, Y. (1988) *Biochem Biophys Res Commun* **157**, 238-244.
4. Morreau, H., Galjart, N. J., Gillemans, N., Willemsen, R., van der Horst, G. T..J., and d'Azzo, A. (1989) *J Biol Chem* **264**, 20655-20663
5. Yamamoto, Y., Hake, C. A., Martin, B. M., Kretz, K. A., Ahernrindell, A. J., Naylor, S. L., Mudd, M., and O'Brien, J. S. (1990) *DNA & Cell Biol* **9**, 119-127
6. Yoshida, K., Oshima, A., Shimmoto, M., Fukuhara, Y., Sakuraba, H., Yanagisawa, N., and Suzuki, Y. (1991) *Am J Human Genet*, in press.
7. Nishimoto, J., Nanba E, Inui, K., Okada, S., and Suzuki, K. (1991) *Am J Human Genet*, in press.
8. Nanba, E., and Suzuki, K. (1990) *Biochem. Biophys. Res. Commun.* **173**, 141-148.
9. Popko, B., Puckett, C., Lai, E., Shine, H. D., Readhead, C., Takahashi, N., Hunt, S. W., III, Sidman, R., L., and Hood, L. (1987) *Cell*, **48**, 713-721.
10. Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989) *Molecular Cloning, A Laboratory Manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY 1989
11. Steinmetz, M., Winoto, A., Minard, K., and Hood, L. (1982) *Cell* **28**, 489-498.
12. Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Nat. Acad. Sci., USA* **74**, 5463-5467.
13. Biggin, M. D., Gibson, J. J. and Hong, G. F. (1983) *Proc. Natl. Acad. Sci., USA* **80**, 3963-3965.
14. Zinn, K., DiMaio, D., and Maniatis, T. (1983) *Cell* **34**, 865-879.
15. Okayama, H., Kawauchi, M., Brownstein, M., Lee, F., Yokota, T., and Arai, K. (1987) *Methods Enzymol.* **154**, 3-28.
16. Krug, M. S., and Berger, S. L. (1987) *Methods in Enzymol.* **152**, 316-325.
17. Padgett, R. A., Grabowski, P. J., Konarska, M. M., Seiler, S., and Sharp, P. A. (1986) *Ann. Rev. Biochem.* **55**, 1119-1150.
18. Shapiro, M. B., and Senapathy, P. (1987) *Nucleic Acid Res.* **15**, 7155-7174.
19. Shows, T. B., Scrafford-Wolff, L. R., Brown, J. A., and Meisler, M. H. (1979) *Somat. Cell Genet.* **5**, 147-158.
20. Naylor, S. L., Elliott, R. W., Brown, J. A., and Shows, T. B. (1982) *Am. J. Human Genet.* **34**, 235-244.
21. Kreysing, J., von Figura, K., and Gieselmann, V. (1990) *Europ. J. Biochem.* **191**, 627-631.
22. Felton, J., Meisler, M., and Paigen, K. (1974) *J. Biol. Chem.* **249**:3267-3272.
23. Bouvier, J. D., and Seyfried, T. N. (1990) *Develop. Neurosci.* **12**:126-132.