INFORMATION TO USERS

The most advanced technology has been used to photograph and reproduce this manuscript from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

U·M·I

University Microfilms International
A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
313/761-4700 800/521-0600

Order Number 9108187

Biochemical and molecular genetic studies of human alpha-*N*-acetylgalactosaminidase and Schindler disease

Wang, Anne May, Ph.D.

City University of New York, 1990

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

BIOCHEMICAL AND MOLECULAR GENETIC STUDIES OF HUMAN ALPHA-*N*-ACETYLGALACTOSAMINIDASE AND SCHINDLER DISEASE

by

ANNE MAY WANG

A dissertation submitted to the Graduate Faculty in Biomedical Sciences
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy, The City University of New York.

1990

This manuscript has been read and accepted for the Graduate Faculty in Biomedical Sciences in satisfaction of the dissertation requirement for the Degree of Doctor of Philosophy, The City University of New York.

5/25/90
Date

Robert J. Desnick, Ph.D., M.D.
Chair of Examining Committee

8/22/90
Date

Terry A. Krulwich, Ph.D.
Executive Officer

Kenneth H. Astrin, Ph.D.

Gregory A. Grabowski, M.D.

D. Schindler
Detlev Schindler, M.D.

James. G. Wetmur, Ph.D.

Supervisory Committee

The City University of New York

ABSTRACT

BIOCHEMICAL AND MOLECULAR GENETIC STUDIES OF

HUMAN ALPHA-*N*-ACETYLGALACTOSAMINIDASE AND SCHINDLER DISEASE

by

ANNE M. WANG

Advisor: Robert J. Desnick, Ph.D., M.D.

Human α-*N*-acetylgalactosaminidase (α-GalNAc; E.C. 3.2.1.49), the lysosomal glycohydrolase that cleaves α-*N*-acetylgalactosaminyl moieties from glycoconjugates, is encoded by a gene localized to chromosome 22q13→ter. The deficient activity of this enzyme results in Schindler disease, an autosomal recessive disorder characterized by the increased urinary excretion of glycopeptides and oligosaccharides containing α-*N*-acetylgalactosaminyl moieties. Two clinical phenotypes of Schindler disease have been described. The Type I disease is similar to the infantile form of neuroaxonal dystrophy, whereas the Type II disorder is characterized by disseminated angiokeratoma corporis diffusum with no neurological involvement. In these studies, a full-length cDNA encoding human α-GalNAc was isolated and characterized. The full-length 2158 bp α-GalNAc sequence encoded 411 amino acids and predicted a signal peptide sequence of 17 amino acids as well as six *N*-glycosylation sites. The functional integrity of the cDNA was demonstrated by transient expression in COS-1 cells. Northern hybridization analysis of mRNA revealed two transcripts of about 3.6 and 2.2 kb. The α-GalNAc cDNA had 46.9% amino acid homology with the human α-galactosidase A (α-Gal A) cDNA suggesting the evolutionary relatedness of these genes. The isolation of a full-length α-GalNAc cDNA facilitated the characterization of the molecular lesions in the first identified affected homozygotes with Schindler disease. The specific molecular lesions were determined by the

polymerase chain reaction (PCR) amplification and sequencing of reverse-transcribed α-GalNAc transcripts from affected homozygotes with Type I and II disease. In a sib with Type I disease, a single G to A transition at nucleotide 973 was detected and resulted in a glutamic acid to lysine substitution in residue 325 of the α-GalNAc polypeptide. In the type II homozygote, a single C to T transition was identified at nucleotide 985 which resulted in an arginine to tryptophan substitution at position 329. These base substitutions were confirmed by dot-blot analyses of PCR-amplified genomic DNA from members of each family using allele-specific oligonucleotides. Furthermore, transient expression of α-GalNAc constructs containing these nucleotide transitions resulted in the expression of an immunoreactive polypeptide which had no detectable α-GalNAc activity. The isolation of an α-GalNAc cDNA also facilitated the isolation and characterization of the α-GalNAc structural gene. The α-GalNAc gene was ~14 kb and contained nine exons. All exon/intron junctions conformed to the GT/AG rule. Analysis of 1432 bp of 5' flanking sequence revealed two Sp1 and one GC-box promotor elements in this lysosomal housekeeping gene. Six *Alu*-repetitive elements were identified and all were in the reverse orientation. Comparison of the α-GalNAc gene with the α-Gal A gene revealed homologous exonic placement of α-GalNAc introns 2 through 7 with α-Gal A introns 1 through 6. Predicted amino acid homology among α-GalNAc exons 2-7 with α-Gal A exons 1-6 ranged from 46.2% to 62.7% with a few short gaps. In contrast, there was little similarity between α-Gal A exon 7 and α-GalNAc exons 8 and 9 which had only 15.8% homology with numerous gaps. The high exonic homologies, in addition to homologous intron placement, suggest that these genes are evolutionarily related and arose through duplication and divergence from a common ancestral gene.

ACKNOWLEDGEMENTS

I wish to express my sincerest thanks and gratitude to Dr. Robert J. Desnick for providing the environment in which these studies were conducted. His dedication to science and academic excellence and rare insight have served as a model and motivating force during these years. It is with his faith and perseverance that this thesis was realized.

A significant portion of these studies, as well as identification of a new lysosomal disorder, is due to Dr. Detlev Schindler. His commitment to clinical care, diagnosis and scientific pursuit has contributed significantly to the area of the inborn errors of metabolism. His influence, in addition to that of Drs. David F. Bishop and Gregory A. Grabowski, have enriched my educational experience.

I am also thankful for the excellent assistance of Thomas Fitzmaurice, Raman Reddy, Yiannis Ioannou and Richard Gotlib. Their expertise has leavened the gravity on many occasions.

Lastly, I am indebted to my parents, Pin and Chen Hsiu Wang, whose unfaltering love and encouragement have supported me through not only these studies, but my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

α-Gal A	α-galactosidase A
α-Gal B	α-galactosidase B
α-GalNAc	α-*N*-acetylgalactosaminidase
A	adenosine
bp	base pair(s)
C	cytosine
cDNA	complementary DNA
DNA	deoxyribonucleic acid
G	guanosine
kCal	kilocalorie
kDa	kilodalton
mRNA	messenger RNA
4-MU-α-Gal	4-methylumbelliferyl-α-D-galactopyranoside
4-MU- or pNP-α-GalNAc	4-methylumbelliferyl- or p-nitrophenyl-α-*N*-acetylgalactosaminide
$NaDodSO_4$/PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
nt	nucleotide(s)
PCR	polymerase chain reaction
RNA	ribonucleic acid
T	thymidine

BACKGROUND

A. Biochemical Characterization of α-*N*-Acetylgalactosaminidase

It is now recognized that human lysosomal α-galactosidase B (α-Gal B) is an α-*N*-acetylgalactosaminidase (α-D-*N*-acetylgalactosamine-*N*-acetylgalactosaminohydrolase, α-GalNAc; E.C 3.2.1.49). The confusion regarding the natural substrate specificity of α-Gal B resulted from the use of artificial galactosides as substrates which detected two α-galactosidase isozymes in human tissues (1, 2). The major isozyme, designated α-Gal A, represented about 80-90% of total α-galactosidase activity and was heat labile, while the other isozyme, designated α-Gal B, represented the remaining activity and maintained 80% of enzymatic activity at 50°C for at least 4 hours (3-9). Myoinositol specifically inhibited α-Gal A activity (5, 7) whereas the B isozyme was competitively inhibited by α-*N*-acetylgalactosamine (3, 4, 6). The two "isozymes" were separable by starch and cellulose acetate gel electrophoresis (7, 10), isoelectric focusing (9, 11, 12), and ion exchange chromatography (5, 6, 9, 13, 14). These studies revealed the presence of multiple forms of α-Gal A with pI values ranging from 4.3 to 5.1 while α-Gal B had a single form with a pI of about 4.5. It was also clearly demonstrated that α-Gal A activity was deficient in Fabry disease (1, 7, 8, 10-12, 15) and that the residual activity detected in classical hemizygotes with Fabry disease was due to α-Gal B. Neuraminidase treatment of α-Gal A and B converted the multiple α-Gal A forms to a single activity band at pI 4.6, while the migration of α-Gal B was unaffected (2, 16, 17). In the early 1970s, long before it was recognized that α-Gal B was an α-*N*-acetylgalactosaminidase, these findings were erroneously interpreted; α-Gal A was thought to be a sialated form of α-Gal B.

However, the two isozymes were immunologically distinguishable (5, 7, 15): anti-α-Gal A did not crossreact with α-Gal B, and anti-α-Gal B did not crossreact with α-Gal A. The residual activity in Fabry disease could be immunoprecipitated with anti-α-Gal B, but not with

anti-α-Gal A. These findings suggested that the two isozymes were indeed distinct proteins. Subsequently, these genes were mapped to separate chromosomes; α-Gal B to 22q13→ter (18, 19) and α-Gal A to Xq21.33→22 (20), confirming their independent nature.

In 1977, Dean and Sweeley in the United States, and Schram and colleagues in the Netherlands independently demonstrated that α-Gal B functioned as an α-*N*-acetylgalactosaminidase which hydrolyzed artificial and natural substrates with terminal α-*N*-acetylgalactosaminyl moieties and was competitively inhibited by α-*N*-acetylgalactosamine (3, 4, 6, 21, 22). While studies suggested that α-Gal B was able to cleave terminal α-galactosyl residues from glycolipid substrates (albeit at a very slow rate) and water soluble oligosaccharides, α-Gal B was 2-5 fold more active as an α-*N*-acetylgalactosaminidase on glycolipid and water soluble oligosaccharide substrates with terminal α-*N*-acetylgalactosaminyl moieties (14). Furthermore, α-Gal B had a higher affinity for o- or p-nitrophenyl-α-*N*-acetylgalactosaminide (K_m ~ 1-2 mM) than for o- or p-nitrophenyl-α-D-galactoside (K_m ~ 7-20 mM) (3, 4, 14). α-Gal B also hydrolyzed the terminal α-*N*-acetylgalactosaminyl residues from Forssman hapten, a canine intestinal neutral glycosphingolipid, and blood group A trisaccharide (6). Thus, it was concluded that α-Gal B was actually an α-*N*-acetylgalactosaminidase which did not cleave the glycolipid substrates accumulated in Fabry disease.

Human α-GalNAc has been purified from a variety of sources, including placenta and liver (3-6, 9, 14). The enzyme is a homodimeric glycoprotein with a subunit molecular weight of about 48 kDa. Support for the homodimeric structure is based on biosynthetic studies of α-GalNAc in human fibroblasts which indicated that the enzyme was synthesized as a 65 kDa precursor which was processed to a 48 kDa mature lysosomal form (23). Additionally, analysis of radiolabelled oligosaccharides from immunoprecipitated enzyme revealed a high-mannose type oligosaccharide structure ($Man_{8-9}GlcNAc$) on both precursor and processed enzyme forms; only

the precursor had phosphorylated mannose residues. The absence of complex-type oligosaccharide moieties containing terminal sialic residues was supported by neuraminidase studies which did not alter the pI of the enzyme, and by qualitative oligosaccharide analysis of the mature, processed enzyme (24, 25). The range of substrates hydrolyzed by α-GalNAc have not yet been identified, but any compound containing an α-*N*-acetylgalactosaminyl residue is a potential substrate. These glycoconjugates include glycosphingolipids, mucins, *O*- and *N*-linked glycoproteins, and the mucopolysaccharide, keratin sulfate.

B. Lysosomal Biology

α-Gal B is representative of a class of enzymes present in lysosomes, virtually all of which are acid hydrolases (26). Genetic deficiencies of specific lysosomal hydrolases result in the different lysosomal storage diseases. To date, over 30 lysosomal storage diseases in humans have been identified and have generated wide interest in the cell biology of lysosomes and the biochemistry of lysosomal enzymes. Intense studies have lead to the delineation of the unique properties of lysosomal enzymes, especially related to their biosynthesis, processing, and trafficking to the lysosome. Briefly, lysosomal enzymes are typically synthesized as pre-proenzymes which are converted to proenzymes in the endoplasmic reticulum and cis-Golgi, and then processed into mature forms within the lysosome (27). These extensive co- and post-translational modifications include glycosylation and sequential carbohydrate modification and phosphorylation. Transport to the lysosome is mediated by the mannose-6-phosphate residues on most of the enzymes. These residues are added in the Golgi and are recognized by mannose-6-phosphate receptors in membranes which are responsible for transporting the enzymes to the endosomes and ultimately to the lysosomes.

The study of lysosomal enzymes and lysosomal storage diseases has proceeded to the molecular level. In general, each lysosomal enzyme constitutes approximately 0.01% of total cellular protein (28). Despite this low abundancy, several cDNA sequences have been cloned for lysosomal enzymes, including α-fucosidase (29), α-Gal A (30, 31), and β-hexosaminidase α- (32-34) and β-chains (35). Several strategies have been employed in cloning these sequences, including screening cDNA libraries with radiolabelled probes which may be oligonucleotides constructed from peptide sequence (34), cDNA sequences from other species (36, 37), or an exonic fragment of a human gene from the same gene family (38). Screening of cDNA libraries enhanced for message by hybrid selection and polysome immunoabsorption (32) and of expression libraries constructed in λgt11 with monospecific antisera (29, 30, 39-41) also have been successful. More recently, the polymerase chain reaction (PCR) and mixed oligonucleotide primed amplification of cDNA (MOPAC) has permitted the cloning of partial cDNA sequences using amino acid microsequence data. These partial cDNAs may then be used as specific probes for more efficient library screening (42). Analysis of these cDNA sequences has confirmed at the molecular level sequences corresponding to pre- and prepro-peptide sequences which were presumed cleaved during the transport of lysosomal enzymes. These pre- and prepro-peptide sequences were hydrophobic as expected (43).

To date, only five genomic lysosomal enzyme sequences have been reported: α-Gal A (44), the α- (45) and β-chains (46) of β-hexosaminidase, β-glucosidase (47) and acid phosphatase (48). α-Gal A is 12 kb with seven exons and has been sequenced in its entirety (49). The first exon contains the 5' untranslated region, the signal peptide, and the first 33 amino acids of the mature enzyme. The gene is unusual in that it lacks a 3' untranslated region. The β-hexosaminidase α- and β-chains are structurally similar, and extensive homology between the two genes suggests evolution of the two sequences from a common ancestral gene. The α-

chain gene spans about 40 kb with 14 exons, and the β-chain gene spans about 35 kb and also has 14 exons. Twelve of the 13 introns in the β-chain were shown to interrupt the gene at homólogous positions in the α-chain. β-glucosidase is 7.6 kb with 11 exons. The gene and its tightly linked 5.8 kb pseudogene which has 96% nucleotide homology have been sequenced. Acid phosphatase spans nine kb with 11 exons. Like the α-Gal A gene, exon 1 encodes the signal peptide and the first eight residues of the mature protein.

As a result of cloning cDNA and genomic sequences for many of the proteins associated with genetic diseases, it has been possible to determine the precise molecular pathology underlying these disorders and study their effects. Among the lysosomal disorders, a number of mutations in the α-subunit of β-hexosaminidase A have been described in Tay-Sachs disease (β-hexosaminidase A deficiency). The major α-subunit mutation in the Ashkenazi Jewish population is an insertion of four bases into exon 11 (50) which causes a frameshift and subsequent termination nine nucleotides downstream. A second mutation occurring within this population is a G to C transversion at the 5' border of intron 12 (51-53) which alters the consensus GT dinucleotide necessary for intron splicing and results in no detectable mRNA. A mutation responsible for the adult-form or late-onset G_{M2} gangliosidosis has been identified as a G to A transition in exon 7 which causes the substitution of serine for glycine at residue 269 (54, 55). This substitution results in poor association of the α- and β-subunits in the active tetrameric configuration (56). Among the French-Canadian population, the prevalent mutation has been identified as a 7.6 kb deletion of exon 1 and flanking sequences of the α-subunit gene which may have occurred by homologous recombination of misaligned *Alu* sequences (57, 58). Other mutations which cause β-hexosaminidase A deficiency have also been identified and include base substitutions and deletions which cause amino acid substitutions, frameshifts and premature termination (59-62). Analyses of these mutations have revealed the abnormal

processing and trafficking of the defective α-subunits as well as altered kinetics and substrate specificities. As illustrated by the example of Tay-Sachs disease, the molecular lesions resulting in this disorder, as well as in other lysosomal disorders (e.g., Fabry disease, Gaucher disease), are heterogeneous in nature and include partial gene duplications (63).

A variety of methods have been used to delineate the specific nature of these mutations, including the construction and screening of cDNA and genomic libraries from the DNA of affected individuals. S1 nuclease protection studies (64, 65) and RNase A cleavage analysis (66) have been successfully used to detect single base substitutions as well as small insertions and deletions. Most recently, primer directed amplification of specific DNA sequences by the PCR (67) has allowed direct, rapid identification and sequencing of mutations. The effects of these mutations have been studied using a variety of techniques including transient expression, pulse-chase labelling and immunoprecipitation (61, 68).

C. The Evolution of Human Gene Families

Comparative studies at the nucleotide and deduced amino acid levels have provided insight into the mechanisms of molecular evolution. Gene duplication from a single ancestral gene is the primary mechanism in the evolution of genes with new functions (69). This initial event may occur by several mechanisms: 1) nonhomologous chromosomal breakage and reunion, 2) homologous, but unequal recombination between repeated elements such as *Alu* I or *Kpn* I sequences, 3) RNA-mediated duplications, and 4) gene amplification. After this initial event, one copy of the gene may retain its original function while a duplicated copy may diverge by the accumulation of mutations which alter its function or regulation.

The most extensive work in the structure and molecular evolution of genes has involved the preproinsulin and the β-globin gene families. Perler and colleagues (70) established a

method for calculating evolutionary divergence time based on the nucleotide replacement substitution rate. Mutations have been estimated to occur in DNA sequences at a minimum rate of 7×10^{-9} per nucleotide site per year. When exonic regions of two homologous sequences are compared, two types of substitutions are evident: replacement substitutions which result in amino acid replacement, and silent substitutions which do not alter the amino acid. Nucleotide differences among homologous proteins have been used to estimate the replacement site divergence of the corresponding genes by determining the minimum number of base changes necessary to generate the observed amino acid replacements. Corrections have been devised for multiple base change events with individual codons. Within a protein family, replacement changes appear to accumulate linearly with divergence time; this is supported by the fossil record. In general, silent and intronic changes do not provide good evolutionary information as they accumulate more rapidly than replacement substitutions.

The Perler method for measuring evolutionary time was based on sequence analysis of the rat, human, and chicken preproinsulin genes. Most vertebrate species have one insulin gene, but rat, mice and three species of fish have two nonallelic insulin genes. The rat preproinsulin gene II contains a 499 bp intron in the C peptide that is not present in gene I. Thus, one gene either gained or lost the intron during duplication. To determine which gene was ancestral, the preproinsulin gene from chicken was cloned. The chicken gene contained a 3.5 kb intron in the C peptide, and therefore, the rat gene II was the ancestral gene.

Using the Perler method, the structure and evolution of the human β-globin gene family was delineated by Efstradiatis *et al.* (71). The complete sequences of embryonic ε- (72), fetal δ- and γ- (73), and adult α- (74) and β- (75) globins have been determined. From these sequences, pairwise calculations of replacement and silent substitutions according to the Perler method were performed. This analysis allowed the construction of an evolutionary tree for the

human β-like globin genes based entirely on nucleotide sequence comparisons. The tree agreed with data available from the fossil record.

Nucleotide sequence data may also provide structural information regarding the evolutionary conservation of homologous proteins. Among the lysosomal enzymes, extensive homology has been found between the β-hexosaminidase α- and β-chain cDNA and chromosomal genes (46). The α- and β-chains are the polypeptide subunits of the two major isozymes of β-hexosaminidase; hexosaminidase A is composed of two α- and two β-subunits, and hexosaminidase B, contains four β-subunits. A third minor isozyme, hexosaminidase S, is composed of α-subunits. Comparison of the cDNA sequences for the α- and β-chains revealed 55% nucleotide and 57% amino acid similarity, suggesting divergence of the two genes from a common ancestral gene (45). The finding of extensive similarity was not unexpected, as the subunits share several common features. First, homodimers of each chain have the ability to hydrolyze the same artificial substrates in vitro, and therefore must have similar substrate binding and active sites. Secondly, several independently prepared antisera against hexosaminidase S have shown crossreactivity with hexosaminidase B, suggesting that the two subunits share related epitopes. Characterization of the α- and β- chain genomic sequences (45, 46) has shown that the structural organization of the genes were highly conserved. Comparison of the two genes revealed that 12 of the 13 introns interrupted the coding regions at homologous positions, clearly indicating that they were duplicated from a single ancestral gene and that the single difference in intron/exon structure was probably due to subsequent divergence.

Study of the genomic structure of genes in relation to molecular evolution has also provided support for hypotheses concerning the conservation of active site regions and for exons encoding functional domains. X-ray crystallography studies of the same protein purified from different species have shown that secondary structure is highly conserved (76, 77). In a study

of the phosphoglycerate kinase gene from six species (76), the amino acids forming the substrate binding cleft were very highly conserved. The homology between yeast and mammalian sequences for the binding cleft was greater than 94%, whereas the overall homology was less than 65%. Trypanosome phosphoglycerate kinase had only 44% overall homology with the mammalian enzyme, but showed 74% homology at the binding cleft. Analysis of the sequences of genes with functionally conserved regions has shown that these regions undergo replacement substitutions at lower rates than for nonconserved regions (77). The lysosomal hydrolase, α-glucosidase, has also been shown to have significant homology with human isomaltase and the rabbit intestinal sucrase-isomaltase complex (78). Alignment of these sequences revealed 26% homology. Ten of 13 amino acids were identical in the region surrounding the known active site of sucrase and isomaltase, suggesting that the active site of α-glucosidase may be in this region. From these sequences, it was proposed that the isomaltase-sucrase single gene precursor arose by duplication of an ancestral isomaltase gene. Since α-glucosidase has the same degree of homology with isomaltase and sucrase, α-glucosidase must have duplicated from the same ancestral gene. A more recently emerging gene family with potential active site conservation is the sulfatases. Protein homologies have been identified among the human arylsulfatases A and B, glucosamine-6-sulfatase and steroid sulfatase (42, 79). While these proteins share an overall amino acid homology of ~20%, the highest degree of homology among the coding regions of these genes was concentrated at the *N*-terminal third of the four mature polypeptides; studies with arylsulfatase A indicate that these regions may be involved in the enzyme's active site (79).

Conservation of active sites or functional protein domains by genomic structural organization has also been proposed (80). Structural characterization of the phosphoglycerol kinase gene (81) has provided evidence supporting the concept of exon shuffling. Comparative

analysis of human phosphoglycerol kinase, maize alcohol dehydrogenase and chicken glyceraldehyde-3-phosphate dehydrogenase revealed that the sequence for the nucleotide binding domain showed homologous intron placement in each gene. This domain is specified by five exons; three exons coded for the first mononucleotide domain and two for the second domain. The homology of the exon organization in structurally similar regions suggests that the nucleotide binding domain evolved by gene duplication and was subsequently dispersed to different proteins through a process of intron-mediated recombination. Thus, this gene family supports the concept of exon shuffling in the conservation of functional domains.

D. Schindler Disease

The first patients with α-GalNAc deficiency were reported in the last year (82-85). Although these cases are due to markedly deficient α-GalNAc activity, the clinical manifestations in the unrelated patients differed remarkably. Type I disease is characterized by a clinical course identical to that of an infantile form of neuroaxonal dystrophy, whereas the patient with Type II disease has no neurological involvement, but does have angiokeratoma corporis diffusum.

The first patients with Type I Schindler disease were two German brothers, the consanguineous offspring of fourth cousins. Development of the brothers was normal until about 15 months of age, after which each experienced rapid regression. Clinical onset was signaled by grand mal seizures in the younger sib at 9 months, and by clumsiness and falling episodes at 12 months in the older sib. From 15 months of age, both developed strabismus, nystagmus, optic atrophy, muscular hypotonia and frequent myoclonic seizures. They also exhibited profound psychomotor retardation. By three to four years of age, they were immobile, had decorticate postures, cortical blindness, and had little, if any, contact with the environment.

The deficient activity of α-GalNAc was demonstrated using the artificial substrate p-nitrophenyl-α-*N*-acetylgalactosaminide (82, 83). This result was later confirmed using a newly synthesized, highly sensitive and specific fluorogenic substrate, 4-methylumbelliferyl-α-*N*-acetylgalactosaminide (4-MU-α-GalNAc) (84). This substrate allowed reliable detection of the enzyme defect in plasma, lymphoblasts, and fibroblasts from the affected brothers as well as the intermediate activity levels in their heterozygous parents, consistent with an autosomal recessive mode of inheritance.

An abnormal urinary oligosaccharide profile was also observed in the affected brothers (82-84). The excreted products were later shown to be *O*-linked glycopeptides (87) with the structures: NeuNAcα2→3Galβ1→3(NeuNAcα2→6)GalNAcα1→O-serine and -threonine, NeuNAcα2→3Galβ1→4GlcNAcβ1→6(NeuNAcα2→3Galβ1→3)GalNAcα1→O-serine and -threonine, and GalNAcα1→O-serine and -threonine. Only the older affected sib, who was blood group type A positive, excreted the blood group A trisaccharide: GalNAcα1→3(Fucα1→2)Gal.

Further clinical evaluation of the two brothers revealed several interesting findings. Ultrastructural evaluation of biopsied rectal tissue revealed abnormal tubulovesicular material free in the cytoplasm of a few preterminal and terminal axons in the myenteric plexus. Subsequent examination of the frontal lobe cortical biopsy showed the characteristic neuropathology of neuroaxonal dystrophy (88), which included the characteristic "spheroids" in terminal axons that contained tubulovesicular and other membranous arrays. These structures were essentially identical to those observed in the inherited neuroaxonal dystrophies, including Seitelberger disease and Hallervorden-Spatz disease. The morphologic hallmark in these disorders is generalized axonal dystrophy with peculiar swellings or "spheroids" in terminal autonomic axons. Each is inherited as an autosomal recessive trait (89).

Since the clinical and neuropathologic findings clearly classified this disease as a neuroaxonal dystrophy, efforts were directed to determine the α-GalNAc activity in unrelated patients with Seitelberger disease (84). In eight biopsy- or autopsy-proven cases, the α-GalNAc activities in plasma, cultured cells, or cortical tissue were within the normal range. Thus, α-GalNAc was not the enzymatic defect in these unrelated patients with infantile neuroaxonal dystrophy. Moreover, it could be concluded that the biochemical defects underlying the neuroaxonal dystrophy phenotype were heterogeneous.

In the second reported case of Schindler disease, and the first with the Type II phenotype, (86) the propositus was a 46-year old Japanese woman, a daughter of a first cousin marriage (85). Development of this patient was normal with the notable exception of the eruption of disseminated angiokeratoma corporis diffusum during the third decade of life. The development of angiokeratoma is characteristic of several lysosomal disorders including Fabry disease (α-Gal A deficiency), fucosidosis (α-fucosidosis deficiency) and galactosialidosis (β-galactosidase and sialidase deficiency). However, the patient had normal levels of activity for these and a number of other lysosomal hydrolases. An abnormal urinary oligosaccharide profile was observed in the affected individual (85, 90) and the products were purified and determined to be essentially identical to the ones identified in Type I disease. This finding suggested the diagnosis of α-GalNAc deficiency (86). The patient had less than 1% normal activity, while her two children had intermediate activity levels, consistent with the autosomal recessive mode of inheritance. Electron microscopy of a skin biopsy from this individual revealed vacuolation of the cytoplasm in endothelial cells. Similar vacuoles were also observed in sweat gland cells, dermal fibrocytes, vascular pericytes, smooth muscle cells, fat cells, Schwann cells and neuron axons. Such vacuoles were essentially absent in the brothers affected with Type I disease.

Thus far, two distinct subtypes of α-GalNAc deficiency have been described. These subtypes are phenotypically distinct, but result from the same deficient enzymatic activity.

REFERENCES

1. Beutler, E. and Kuhl, W. (1971) J Lab Clin Med 78:987.

2. Kint, J. A. (1971) Arch Int Physiol Biochem 78:633-634.

3. Dean, K. J., Sung, S. and Sweeley, C. C. (1977) Biochem Biophys Res Com 77:1411-1417.

4. Schram, A. W., Hamers, M. N. and Tager, J. M. (1977) Biochim Biophys Acta 482:138-144.

5. Beutler, E. and Kuhl, W. (1972) J Biol Chem 247: 7195-7200.

6. Callahan, J. W., Lassilla, E. L., DenTandt, W. and Philippart, M. (1973) Biochem Med 7:424-431.

7. Rietra, P. J. G. M., VanDenBergh, F. A. J. T. M. and Tager, J. M. (1975) Clin Chim Acta 62:401-413.

8. Johnson, D. L. and Desnick, R. J. (1978) Biochim Biophys Acta 538:195-204.

9. Kusiak, J. W., Quirk J. M. and Brady, R. O. (1978) J Biol Chem 253:184-190.

10. Beutler, E. and Kuhl, W. (1972) Amer J Hum Genet 24:237-249.

11. Salvayre, R., Maret, A., Negre, A. and Douste-Blazy, L. (1979) Eur J Biochem 100:377-383.

12. Salvayre, R., Negre, A., Maret, A., Lenoir, G. and Douste-Blazy, L. (1981) Biochim Biophys Acta 659:445-456.

13. Dean, K. J. and Sweeley, C. C. (1979) J Biol Chem 254:9994-10000.

14. Dean K. J. and Sweeley, C. C. (1979) J Biol Chem 254:10001-10005.

15. Schram, A. W., Hamers, M. N., Brouwer-Kelder, B., Donker-Koopman, W. E. and Tager, J. M. (1977) Biochim Biophys Acta 482:125-37.

16. Ho, M. W., Beutler, E., Tennant, L. and O'Brien, J. S. (1972) Am J Hum Genet 24:256-266.

17. Mapes, C. A. and Sweeley, C. C. (1973) Arch Biochim Biophys 158:297-304.

18. deGroot, P. G., Westerveld, A., Meera-Khan, P. and Tager, J. M. (1978) Hum Genet 44:305-312.

19. Mysczkiewicz, B. A., Wilkinson, F. E., Kung, H. J. and Sweeley, C. C. (1984) Fed Proc 43:1529.

20. Unpublished results.

21. Dean, K. J., Sung, S. and Sweeley, C. C. (1978) Adv Exp Med Biol 101:515-523.

22. Schram, A. W., Hamers, M. N. and Tager, J. M. (1978) Adv Exp Med Biol 101:525-529.

23. LeDonne, N. C., Fairley, J. L. and Sweeley, C. C. (1983) Arch Biochim Biophys 224:186-195.

24. Dean. K. J. and Sweeley, C. C. (1979) J Biol Chem 254:10006-10010.

25. Unpublished results.

26. deDuve, C. (1983) Eur J Biochem 137:391-397.

27. VonFigura, K. and Hasilik, A. (1986) Ann Rev Biochem 55:167-193.

28. Myerowitz, R., and Neufeld, E. F. (1981) J Biol Chem 256:3044-3048.

29. Fukushima, H., DeWet, J. R. and O'Brien, J. S. (1985) Proc Natl Acad Sci USA 82:7289-7293.

30. Calhoun, D. H., Bishop, D. F., Bernstein, H. S., Quinn, M., Hantzopoulos, P. and Desnick, R. J. (1986) Proc Natl Acad Sci USA 82:7364-7368.

31. Bishop, D. F., Calhoun, D. H., Bernstein, H. S., Hantzopoulos, P., Quinn, M. and Desnick, R. J. (1986) Proc Natl Acad Sci USA 83:4859-4863.

32. Myerowitz, R. and Proia, R. (1984) Proc Natl Acad Sci USA 81:5394-5398.

33. Myerowitz, R., Peikarz, R., Neufeld, E. F., Shows, T. B. and Suzuki, K. (1985) Proc Natl Acad Sci USA 82:7830-7834.

34. Korneluk, R. G., Mahuran, D. J., Neote, K., Klavins, M. H., O'Dowd, B. F., Tropak, M., Willard, H. F., Anderson, M., Lowden, J. A. and Gravel, R. A. (1986) J Biol Chem 261:8407-8413.

35. O'Dowd, B. F., Quan, F., Willard, H. F., Lanhonwah, A., Korneluk, R. G., Lowden, J. A., Gravel, R. A. and Mahuran, D. J. (1985) Proc Natl Acad Sci USA 82:1184-1188.

36. Guise, K. S., Korneluk, R. G., Waye, J., Lanhonwah, A. M., Quan, F., Palmer, R., Ganschow, R. E., Sly, W. S. and Gravel, R. A. (1985) Gene 34:105-110.

37. Chan, S. J., Segundo, B. S., McCormick, M. B. and Steiner, D. F. (1986) Proc Natl Acad Sci USA 83:7721-7725.

38. Faust, P. L., Kornfeld, S. and Chirgwin, J. M. (1985) Proc Natl Acad Sci USA 82:4910-4914.

39. Sorge, J., West, C., Westwood, B. and Beutler, E. (1985) Proc Natl Acad Sci USA 82:7289-7293.

40. Tsuji, S., Choudary, P. V., Martin, B. M., Winfield, S., Barranger, J. A. and Ginns, E. I. (1986) J Biol Chem 261:50-53.

41. Martiniuk, F. Mehler, M., Pellicer, A., Tzall, S., LaBadie, G., Hobart, C., Ellenbogen, A. and Hirschhorn, R. (1986) Proc Natl Acad Sci USA 83:9641-9644.

42. Schuchman, E. H., Jackson, C. E. and Desnick, R. J. (1990) Genomics 6:149-158.

43. vonHeijne, G. (1986) Nucleic Acids Res 14:4683-4690.

44. Bishop, D. F., Kornreich, R. and Desnick, R. J. (1988) Proc Natl Acad Sci USA 85:3903-3907.

45. Proia, R. L. and Soravia, E. (1987) J Biol Chem 262:5677-5681.

46. Proia, R. L. (1988) Proc Natl Acad Sci USA 85:1883-1887.

47. Horowitz, M., Wilders, S., Horowitz, A., Reiner, O., Gelbart, T. and Beutler, E. (1989) Genomics 4:87-96.

48. Geier, C., von Figura, D. and Pohlmann, R. (1989) Eur J Biochem 183:611-616.

49. Kornriech, R., Desnick, R. J. and Bishop, D. F. (1989) Nuc Acids Res 14:3301-3302.

50. Myerowitz, R. and Costigan, F. C. (1988) J Biol Chem 263:18587-18589.

51. Myerowitz, R. (1988) Proc Natl Acad Sci USA 85:3955-3949.

52. Arpaia, E., Dumbrille-Ross, A., Maler, T., Neote, K., Tropak, M., Troxel, C., Stirling, J. L., Pitts, J. S., Bapat, B., Lamhonwah, A. M., Mahuran, D. J., Schuster, S. M., Clarker, J. T. R., Lowden, J. A. and Gravel, R. A. (1988) Nature 333:85-86.

53. Ohno, K. and Suzuki, K. (1988) Biochem Biophys Res Comm 153:463-469.

54. Navon, R. and Proira, R. L. (1989) Science 243:1471-1474.

55. Paw, B. H., Kaback, M. M. and Neufeld, E. F. (1989) Proc Natl Acad Scie USA 86:2413-2417.

56. D-Azzo, A., Proira, R. L., Kolodny, E. H., Kaback, M. M. and Neufeld, E. F. (1984) J Biol Chem 259:11070-11074.

57. Myerowitz, R. and Hogikyan, N. D. (1985) Science 232:1646-1648.

58. Myerowitz, R. and Hogikyan, N. D. (1987) J Biol Chem 262:15396-15899.

59. Proira, R. L. and Neufeld, E. F. (1982) Proc Natle Acad Sci USA 79:6360-6364.

60. Nakano, T., Muscillo, M., Ohno, K., Hoffman, A. J. and Suzuki, K. (1988) J Neurochem 51:984-987.

61. Lau, M. M. and Neufeld, E. F. (1988) J Cell Biol 107:342A.

62. Ohno, K. and Suzuki, K. (1988) J Neurochem 50:316-318.

63. Kornreich, R., Bishop, D. F. and Desnick, R. J. (in press) J Biol Chem.

64. Prokop, D. J. (1984) Am J Hum Genet 36:499-505.

65. Oshima, A., Kyle, J. W., Miller, R. D., Hoffman, J. W., Powell, P. P., Grubb, J. H., Sly, W. S. Tropak, M., Guise, K. S. and Gravel, R. A. (1987) 84:685-589.

66. Chamberlain, M. and Ryan, T. (1982) In The Enzymes, Vol XV. Boyer, P. Editor. Academic Press, NY.

67. Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. and Erlich, H. A. (1987) Science 239:487-491.

68. Brown, C. A., Neote, K., Leung, A., Gravel, R. A. and Mahuran, D. J. (1989) J Biol Chem 264:21705-21710.

69. Maeda, N. and Smithies, O. (1986) Ann Rev Genet 20:81-108.

70. Perler, F., Efstradiatis, A., Lomeico, P., Gilbert, W., Kolodner, R. and Dodgson, J. (1980) Cell 20:555-566.

71. Efstradiatis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., DeRiel, J. K., Forget, B. G., Weissman, S. M., Slightom, J. L., Blechel, A. E., Smithies, O., Baralle, F. E., Shoulders, C. C. and Proudfoot, N. J. (1980) Cell 21:653-558.

72. Baralle, F. E., Shoulders, C. C. and Proudfoot, N. J. (1980) Cell 21:621-26.

73. Slightom, J. L., Blechel, A. E. and Smithies, O. (1980) Cell 21:627-38.

74. Spritz, R. A., DeRiel, J. K., Forget, B. G. and Weissman, S. (1980) Cell 21:639-46.

75. Lawn, R. M., Efstradiatis, A., O'Connell, C. and Maniatis, T. (1980) Cell 21:647-651.

76. Mori, N., Singer-Sam, J. and Riggs, A. D. (1986) FEBS Letters 204:313-317.

77. McDonald, J. D., Lin, F. and Goldwasser, E. (1986) Mol Cell Biol 6:842-848.

78. Hoefsloot, L. H., Hoogeveen-Westerveld, M., Kroos, M. A., Vanbeeumen, J., Reuser, A. J. J. and Oostra, B. A. (1988) EMBO Journal 7:1697-1704.

79. Peters, C., Schmidt, B., Rommerskrich, W., Rupp, K., Zuhlsdorf, M., Vingron, M., Meyer, H. E., Pohlmann, R. and von Figura, K. (1990) J Biol Chem 265:3374-3381.

80. Gilbert, W. (1978) Nature, 271:501.

81. Michelson, A. M., Blake, C. C., Evans, S. T. and Orkin, S. H. (1985) Proc Natl Acad Sci USA 82:6965-6969.

82. van Diggelen, O. P., Schindler, D., Kleijer, W. J., Huijmans, J. M. G., Galjaard, H., Linden, H. U., Peter-Katalinic, J., Egge, H., Dabrowski, U., Cantz, M. (1987) Lancet ii:804.

83. van Diggelen, O. P., Schindler, D., Willemson, R., Boer, M., Kleijer, W. J., Hiujman, J. G. M., Blom, W. and Galjaard, H. (1988) J Inher Met Dis 11:349-357.

84. Schindler, D., Bishop, D. F., Wolfe, D. E., Wang, A. M., Egge, H., Lemieux, R. U. and Desnick, R. J. (1989) N Engl J Med 320:1735-1740.

85. Kanzaki, T., Yokota, M., Mizuno, N., Matsumoto, Y., and Hirabayashi, Y. (1989) Lancet April 22, 1989:875-877.

86. Wang, A. M., Kanzaki, T., Schindler, D. and Desnick, R. J. (1989) Am J Hum Genet 43:4 A228.

87. Linden, H. U., Klein, R. A., Egge, H., Peter-Katalinic, J., Dabrowski, J. and Schindler, D. (1989) Biol Chem Hoppe-Seyer 370:661-672.

88. Wolfe, D. E., Schindler, D., Desnick, R. J. and Perl, D. (1989) J Neuropathol Exp Neurol 48:349.

89. Seitelberger, F. (1986) In: Handbook of Clinical Neurology Vol. 49. Elsevier, Amsterdam. p. 391-415.

90. Hirabayashi, Y., Matsumoto, Y., Matsumoto, M., Toida, T., Iida, N., Matsubara, T., Kanzaki, T., Yokota, M. and Ishizuka, I. (1990) J Bio Chem 265:1693-1701.

OBJECTIVES AND SIGNIFICANCE

The overall objectives of this research are to 1) determine the molecular genetics of α-*N*-acetylgalactosaminidase (α-GalNAc), and 2) elucidate the nature of the molecular lesions in Schindler disease (α-GalNAc deficiency). To achieve these objectives, each of the following were successfully accomplished: 1) Purification of homogeneous human α-GalNAc enzyme for amino acid microsequencing of *N*-terminal and tryptic peptides, and for quantitative tryptic and chymotryptic mapping; 2) Selection and synthesis of mixed oligonucleotide probes corresponding to amino acid sequences with minimal codon redundancy; 3) Isolation of a full-length cDNA encoding human α-GalNAc and characterization of the sequence, including 5' and 3' untranslated regions, pre- and prepro-peptide sequences, determination of the transcriptional initiation site(s), and analysis of the α-GalNAc transcripts; 4) Transient expression of the full-length α-GalNAc cDNA to demonstrate its functional integrity; 5) Isolation of genomic sequences for α-GalNAc and characterization of the genomic structure including intron/exon organization, 5' regulatory elements, and 3' flanking region. Both the cDNA and genomic sequences of α-GalNAc were compared to α-Gal A for identification of evolutionary relationships; and 6) Characterization of the molecular defects in Schindler disease Types I and II.

These studies relate directly to our understanding of lysosomal enzymes and provide insight into their function and possible evolution. The isolation and characterization of the cDNA and genomic sequences encoding α-GalNAc and their comparison with those of α-Gal A should provide insight into the potential structural and functional domains of these enzymes as well as into their evolutionary relationships. These finding along with those of other lysosomal enzymes known to share homology (e.g., β-hexosaminidase α- and β-chains, and

arylsulfatases A and B) may lead to the identification of potential "lysosomal domains" which may be conserved among lysosomal enzymes.

In addition, these studies are related to the most recently identified lysosomal disorder, Schindler disease. The Type I disorder is a new form of infantile neuroaxonal dystrophy and is the first neuroaxonal dystrophy in which the biochemical basis is known. The Type II disorder appears to be a new cause of angiokeratoma corporis diffusum. The isolation and characterization of the α-GalNAc cDNA and genomic sequences has allowed the determination of the molecular lesions underlying these two subtypes in the first individuals identified with the Schindler disease subtypes.

These studies have provided the molecular basis for the further study of human α-GalNAc and lysosomal enzymes in general. The characterization of the α-GalNAc cDNA and genomic sequences, as well as the identification of the specific base substitutions resulting in two distinct clinical phenotypes due to α-GalNAc deficiency, have provided the framework toward further delineating the function and specificity of α-GalNAc in man and the pathology of its deficiency.

CHAPTER ONE:

HUMAN α-*N*-ACETYLGALACTOSAMINIDASE: MOLECULAR CLONING, NUCLEOTIDE SEQUENCE, AND EXPRESSION OF A FULL-LENGTH cDNA

Homology with Human α-Galactosidase A Suggests Evolution From a Common Ancestral Gene

SUMMARY

Human α-*N*-acetylgalactosaminidase (α-GalNAc; E.C. 3.2.1.49), the lysosomal glycohydrolase that cleaves α-*N*-acetylgalactosaminyl moieties from glycoconjugates, is encoded by a gene localized to chromosome 22q13→qter. The deficient activity of α-GalNAc is the enzymatic defect in Schindler disease, an inherited neuroaxonal dystrophy. To isolate a full-length cDNA, the enzyme from human lung was purified to homogeneity, 129 non-overlapping amino acids were determined by microsequencing the *N*-terminus and seven tryptic peptides, and four synthetic oligonucleotide mixtures were used to screen a human fibroblast cDNA library. A full-length cDNA, pAGB-3, isolated from a placental λgt11 cDNA library, had a 2158 bp insert with an open reading frame which predicted an amino acid sequence that was colinear with all 129 microsequenced residues of the purified enzyme. The pAGB-3 insert had a 344 bp 5' untranslated region, a 1236 bp open reading frame encoding 411 amino acids, a 514 bp 3' untranslated region, and a 64 bp poly(A) tract. A signal peptide sequence of 17 amino acids as well as six *N*-glycosylation sites were predicted. The pAGB-3 cDNA was subcloned into the p91023(B) mammalian expression vector and human α-GalNAc activity was transiently expressed in COS-1 cells, demonstrating the functional integrity of the full-length cDNA. Northern hybridization analysis of mRNA revealed two transcripts of about 3.6 and 2.2 kb, and primer extension studies indicated a cap site at nt -347. A second full-length cDNA, pAGB-35, was isolated from a human retinal library and corresponded to the 3.6 kb transcript. The pAGB-35 cDNA insert was 3597 bp and had 468 bp and 1893 bp 5' and 3' untranslated sequences, respectively. Isolation of a genomic clone, gAGB-1, and sequencing the 2048 bp region including pAGB-3 revealed an 1754 bp intron between codons 319 and 320, which also was the site of a 70 bp insertion and a 45 bp deletion in other cDNA clones. Notably, the α-GalNAc cDNA had remarkable amino acid homology with the human α-galactosidase A (α-Gal A)

cDNA suggesting the evolutionary relatedness of these genes. The α-GalNAc cDNA had 46.9 to 64.7% amino acid identity in sequences (codons 1-319) corresponding to α-Gal A exons 1 through 6, while the comparable exon 7 sequence (pAGB-3 codons 320-411) had only 15.8% homology with numerous gaps. These findings implicate the genomic region at and surrounding codon 319 as a potential site for the abnormal processing of α-GalNAc transcripts as well as for a recombinational event in the evolution and divergence of α-Gal A and α-GalNAc. The availability of the full-length cDNA for human α-GalNAc will permit studies of the genomic organization and evolution of this lysosomal gene, as well as the characterization of the molecular lesions causing Schindler disease.

INTRODUCTION

In the early 1970's, several investigators demonstrated the existence of two α-galactosidase isozymes, designated A and B, which hydrolyzed the α-galactosidic linkages in 4-MU- and/or p-NP-α-D-galactopyranosides (1-7). In tissues, about 80-90% of total α-galactosidase (α-Gal) activity was due to a thermolabile, myoinositol-inhibitable α-Gal A isozyme, while a relatively thermostable isozyme, α-Gal B, accounted for the remainder. The two "isozymes" were separable by electrophoresis, isoelectric focusing, and ion exchange chromatography. After neuraminidase treatment, the electrophoretic migrations and pI values of α-Gal A and B were very similar (1), initially suggesting that the two enzymes were the differentially glycosylated products of the same gene. The finding that the purified glycoprotein enzymes had similar physical properties including subunit molecular weight (~46 kDa), homodimeric structures, and amino acid compositions also indicated their structural relatedness (8-14). However, the subsequent demonstration that polyclonal antibodies against α-Gal A or B did not cross-react with the other enzyme (8, 11), that only α-Gal A activity was deficient in hemizygotes with Fabry disease (1-8) and that the genes for α-Gal A and B mapped to different chromosomes (7, 15), clearly demonstrated that these enzymes were genetically distinct. Thus, it was not surprising when α-Gal B was shown in 1977 to be an α-*N*-acetylgalactosaminidase (α-GalNAc), a homodimeric glycoprotein which hydrolyzed artificial and natural substrates with terminal α-*N*-acetylgalactosaminyl moieties (10, 11, 14) including various *O*- and *N*-linked glycopeptides and glycoproteins, glycosphingolipids, and the proteoglycan, cartilage keratin sulfate II (7).

Purified α-GalNAc has reported native and subunit molecular weights of 90 to 117 kDa and 46 to 48 kDa, respectively (8-14). Kinetic studies demonstrated that the enzyme was inhibited by α-*N*-acetylgalactosamine (K_i ~ 2.1 mM) and hydrolyzed synthetic substrates with

either terminal α-*N*-acetylgalactosaminide (K_m ~ 1-2 mM) or α-D-galactoside moieties (K_m ~ 7-10 mM) (8-14). Biosynthetic studies performed with cultured fibroblasts indicated that the human enzyme was synthesized as a 65 kDa glycosylated precursor which was processed to a mature 48 kDa lysosomal form; both the precursor and mature forms had high-mannose type oligosaccharide chains, but only the precursor's mannose residues were phosphorylated (16).

The deficient activity of α-GalNAc was demonstrated in two brothers with Schindler disease (17, 18), a newly recognized form of infantile neuroaxonal dystrophy (18). The affected brothers excreted increased amounts of *O*-linked glycopeptides and oligosaccharides containing α-*N*-acetylgalactosaminyl moieties which were detectable in urinary screening profiles (17, 19). Biochemical and immunologic studies revealed that neither α-GalNAc activity or enzyme protein was present in fibroblast lysates from the affected sibs (18). Thus, efforts were undertaken to isolate and express a full-length α-GalNAc cDNA in order to determine the nature of the molecular lesions in patients with Schindler disease, and to characterize the genomic organization and expression of the human gene encoding this lysosomal hydrolase.

While expression studies of a hybrid α-GalNAc sequence were in progress, Tsuji *et al.* reported the isolation of a human α-GalNAc cDNA (20). Unlike the full-length pAGB-3 α-GalNAc cDNA sequence reported here, their clone, pcD-HS1204, contained a 70 bp insertion after pAGB-3 nt 957 which altered the reading frame for pAGB-3 residues 330 to 411 and resulted in a truncated polypeptide of only 358 residues. Although their predicted amino acid sequence did not include our tryptic peptide containing residues 335 to 344, the 70 bp insertion may have resulted from alternative splicing. Thus, efforts were directed to determine if such an alternatively spliced transcript occurs in man. In this communication, we report the isolation, nucleotide sequence and transient expression of a full-length cDNA encoding α-GalNAc. Genomic sequencing did not reveal the presence of the putative 70 bp insertion, thereby

affirming that the expressible pAGB-3 transcript is authentic. In addition, remarkable homology between the predicted α-GalNAc and α-Gal A amino acid sequences was identified, suggesting the evolutionary relatedness of the autosomal and X-linked genes encoding these lysosomal hydrolases.

EXPERIMENTAL PROCEDURES

Affinity Purification, Microsequencing and Antibody Production. Human lung α-GalNAc was purified to homogeneity, polyclonal rabbit anti-human α-GalNAc antibodies were produced and purified, and cell supernatants were immunoblotted as described previously (18, 21, 22). For the isolation of tryptic peptides, the concentrated post-hydroxylapatite fraction was subjected to preparative 10% $NaDodSO_4$/PAGE and the 48 and 117 kDa α-GalNAc species were electroeluted separately (23), digested with tosylphenylalaninechloromethylketone-treated trypsin, and the resulting tryptic peptides from each species were separated by C8 reversed-phase HPLC (24). The *N*-terminal amino acid sequences of both the 48 and 117 kDa species and the sequences of selected tryptic peptides from the 48 kDa species were determined by automated gas-phase microsequencing and HPLC identification of phenylthiohydantoin amino acid derivatives (25).

Construction of Synthetic Oligonucleotide Probes. Mixed and unique oligonucleotides were synthesized on an Applied Biosystems Model 380B oligonucleotide synthesizer. Four oligonucleotide mixtures were constructed to regions of minimal codon redundancy for the following *N*-terminal and tryptic peptide sequences: *N*-terminus [5'-CA(AG)ACNCCNCCNATGGG-3']; peptide T-106A [AA(TC)AT(TCA)GA(TC)GA(TC)-TG(TC)TGGAT(TCA)GGNGG-3']; peptide T-72 [5'-ACNTT(TC)GCNGA(AG)TGGAA-3']; and peptide T-133, [5'-TGGCCNGCNTA(TC)GA(AG)GG-3']. Oligonucleotide probes for library screening were 5' end-labelled with [γ-^{32}P]ATP using T4 polynucleotide kinase (26). Unique sequence oligonucleotides (17-mers) were synthesized and used as primers in sequencing reactions. To determine the cap site, two unique, overlapping 30-mers were synthesized for primer extension, 5'-TCGGGACTCCCAGCACTGCAGAGGGTGTGA-3' and 5'-CTG CAGAGGGTGTGAGGTCTGACATCCAGG-3'. To detect alternatively spliced transcripts,

PCR sense and antisense primers for the exonic region flanking the putative 70 bp insertion had the sequences 5'-AGTCGAATTCTGATGTCCACAGACCTGCGT-3', and 5'-AGTCGTCGAGCATATCGGTCCTGCAGCTGA-3', respectively. The four PCR primer sequences for the construction of the α-Gal A and α-GalNAc hybrid cDNA were α-Gal A sense, 5'-TGGGGAGTAGATCTGCTAAAA-3'; α-Gal A antisense, 5'-GATGAGAGATTTTT-CCTGTCTAAGCTGGTACCC-3'; α-GalNAc sense, 5'-TACCAGCTTAGACAGGAAAAA-TCTCTCATCGAA-3'; and α-GalNAc antisense, 5'-AAGAGGTCAGATCTCTCTACT-3'.

Isolation and Characterization of cDNA and Genomic Clones. The pcD human fibroblast cDNA library, kindly provided by Dr. Hiroto Okayama (NIH), was screened with the radiolabelled 26-mer oligonucleotide mixture corresponding to tryptic peptide T-106A by colony hybridization (24). Plasmid DNA isolation and Southern hybridization analyses of positive clones were performed as previously described (26). A λgt10 human retinal library, kindly provided by Dr. Jeremy Nathans (Johns Hopkins University) was screened with the radiolabelled pAGB-3 cDNA insert using the conditions previously described. For isolation of a full-length cDNA, a 0.9 kb *Bam*HI fragment corresponding to the 5' portion of the pAGB-1 insert was then isolated, nick-translated, and used to screen recombinants from a λgt11 human placental library (Clontech Laboratories, Palo Alto, CA) by plaque hybridization (26). To isolate genomic clones containing the entire α-GalNAc sequence, 1×10^6 recombinants from a human genomic cosmid library were screened with the radiolabelled pAGB-3 cDNA insert using the conditions described above for cDNA library screening. The genomic library was prepared from size-selected human lymphoblast DNA and kindly was provided by Dr. Henrik Vissing (Mount Sinai School of Medicine).

DNA Sequencing and Computer-Assisted Analyses. The *Bam*HI inserts from pAGB-1, a *Eco*RI-*Bam*HI restriction fragment of pAGB-3 and the *Eco*RI insert from pAGB-35 were

subcloned into M13 mp18 and mp19. All DNA sequencing reactions were carried out by primer extension using either M13 universal primers or α-GalNAc-specific synthetic oligonucleotide primers by the dideoxy method in both orientations (27). Searches for nucleotide and amino acid sequence similarity were carried out with the University of Wisconsin Genetics Computer Group Software (28). Computer-assisted RNA folding was performed with the PCFOLD program (29).

Transient Expression Assays. The human α-GalNAc full-length pAGB-3 cDNA insert was subcloned into the p91023(B) eukaryotic expression vector (30), kindly provided by Dr. R. J. Kaufmann (Genetics Institute, Boston, MA). Plasmid DNA from the construct (designated p91-AGB-3) was purified and COS-1 monkey kidney cells were transfected with 10 μg of the p91-AGB-3 plasmid DNA by calcium-phosphate precipitation (31). Cells were harvested at 24 hr intervals after transfection and assayed for α-GalNAc activity as previously described (18). One unit (U) of enzymatic activity is equal to that amount of enzyme required to hydrolyze 1 nmol of 4-MU-α-GalNAc per hour. Protein concentrations were determined by the fluorescamine method (21).

Northern Hybridization and Cap Site Analyses. Total RNA was isolated from human lymphoblasts, fibroblasts, and placentae and northern hybridization was performed using the nick-translated pAGB-3 insert as probe (26). Alternatively, the pAGB-3 insert was subcloned into pGEM-4Z (Promega, Madison, WI) and radiolabelled α-GalNAc riboprobe, rAGB-3, was generated using the Promega riboprobe system and used for northern hybridization. For identification of the α-GalNAc cap-site, two unique, overlapping 30-mer oligonucleotide primers were synthesized corresponding to regions 60 and 75 bp from the 5' end of the pAGB-3 cDNA and end-labelled (26). Each primer (100 ng) was used to extend 10 μg of total placental RNA with the BRL cDNA Synthesis Kit (BRL, Gaithersburg, MD). First-strand synthesis was

terminated by phenol extraction and ethanol precipitation. The pellet was washed three times with 70% ethanol, resuspended in 6 µl of H_2O, and then mixed with 6 µl of loading dye (0.3% xylene cyanol, 0.3% bromophenol blue, 0.37% EDTA, pH 7.0). The RNA/DNA heteroduplexes were denatured at 65 °C for 3 min and an aliquot was electrophoresed on a standard 8 M urea, 8% polyacrylamide sequencing gel.

Construction of p91α-GalA6/α-GalNAc7. A plasmid containing α-Gal A exons 1 through 6 from pcDAG-126 (32) was fused to the 3' region of pAGB-3 α-GalNAc insert which corresponded in position to α-Gal A exon 7. The hybrid cDNA, designated pα-GalA6/α-GalNAc7 was constructed with the sense and antisense primers indicated above using a PCR-based method (33) and sequenced. The pα-GalA6/α-GalNAc7 insert was subcloned into the expression vector, p91023(B), and the construct was transiently expressed in COS-1 cells as described above. The α-Gal A and α-GalNAc enzymatic activities and enzyme proteins were detected with 4-MU substrates and by immunoblotting with the respective polyclonal antibodies as described above.

Primer Extension and PCR Amplification of cDNA and Genomic Sequences. For PCR amplification of the putative alternatively spliced region, the 30-mer sense and antisense primers (described above) were used to amplify the 1) reverse-transcribed mRNA from various human sources, 2) cDNA inserts from clones pAGB-4 to 34, and 3) the gAGB-1 genomic sequence. DNAs from pAGB-4 to 34 cDNA clones and the gAGB-1 genomic clone were isolated as described (24, 26). cDNA was synthesized from 10 µg of lymphoblast, fibroblast, and placental total RNA or 2.5 µg of brain poly $(A)^+$ mRNA (Clontech, Palo Alto, CA) using the BRL cDNA Synthesis Kit. Bacteriophage DNA (~0.1 µg) and reverse-transcribed mRNA (~0.1 µg) or genomic cosmid DNA (~ 1 µg) was PCR-amplified using 20 µM of each primer and the GeneAmp DNA Amplification Reagent Kit (Perkin Elmer Cetus, Norwalk, CT). Each PCR

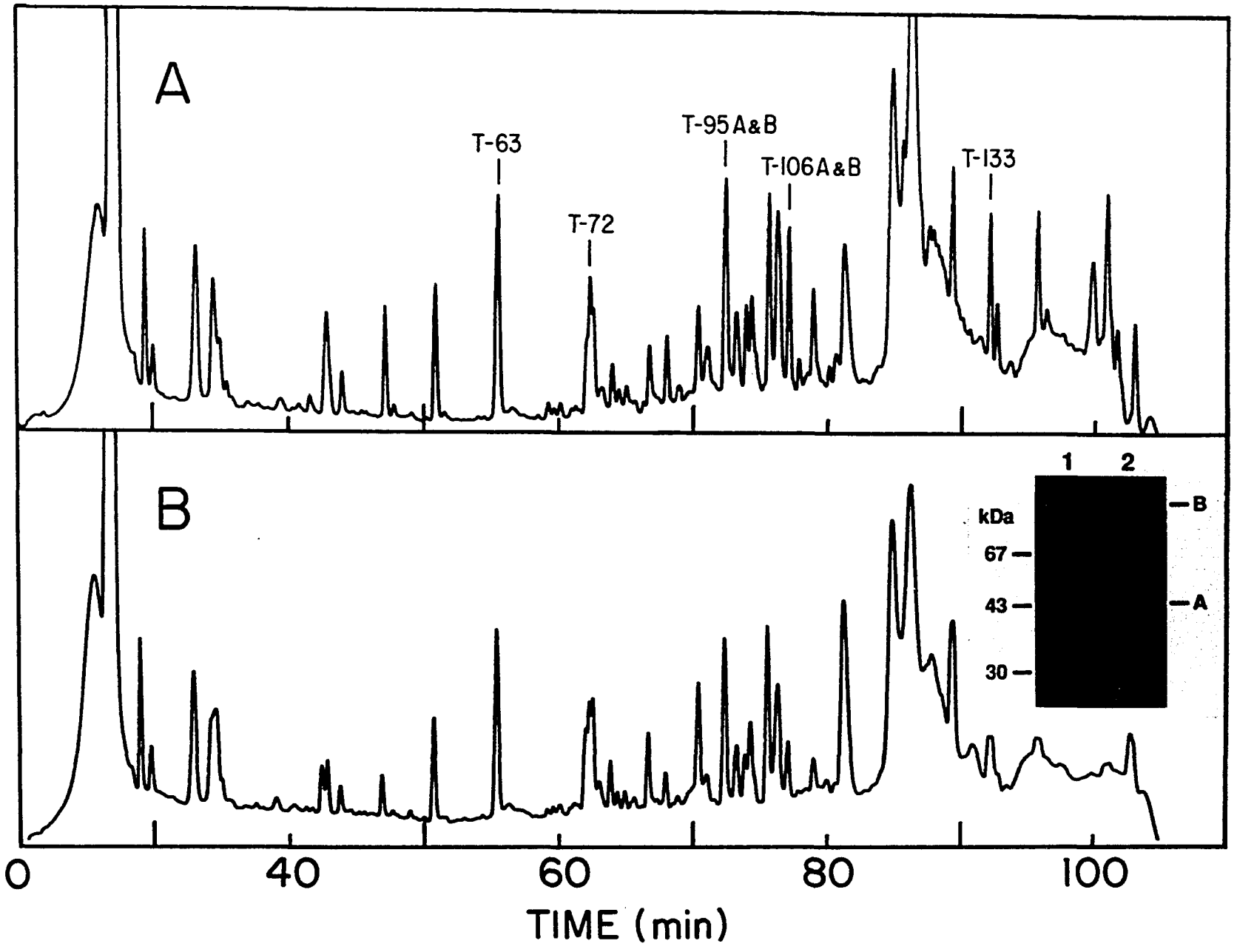
cycle consisted of 1 min denaturation at 94 °C; 2 min annealing at 37 °C; and a 7 min extension at 60 °C. The PCR products were phenol extracted, ethanol precipitated and resuspended in 20 µl of H_20. An aliquot (2 µl) of each PCR reaction was analyzed by electrophoresis on agarose gels using *Hind*III digested lambda and *Hae*III digested ϕX174 DNAs as size standards. For identification of potential stops during reverse transcription of the region surrounding the pcD-HS1204 insertion, a unique 32-mer, 5'-AGTAGTAAGCTTTCATATATCACAGACCCGGT-3', was used to extend 10 µg of total placental RNA or 1 µg of rAGB-3 generated *in vitro* by the Promega riboprobe system as described above.

RESULTS AND DISCUSSION

Purification and Characterization of Human α-GalNAc. Human α-GalNAc was purified to homogeneity (specific activity = ~370,000 U/mg protein) as assessed by the presence of only the 48 and 117 kDa species on $NaDodSO_4$/PAGE (Fig. 1, inset). The 117 kDa was not reduced by boiling or by dialysis against 8 M urea in the presence of β-mercaptoethanol. The 27 microsequenced *N*-terminal residues of the electroeluted 117 kDa species were identical to those of the 48 kDa species. Further evidence that the 117 kDa species was a homodimer of the 48 kDa glycoprotein subunit was the finding that the tryptic digests (and chymotryptic digests; not shown) of both species had essentially identical HPLC profiles (Fig. 1). Microsequencing of the *N*-terminus and seven tryptic peptides from the 48 kDa species identified a total of 129 non-overlapping α-GalNAc residues. For library screenings, synthetic oligonucleotide mixtures (17- to 26-mers) were constructed to contain all possible codons for selected amino acid sequences from the *N*-terminus and three internal tryptic peptides (Figs. 1 and 2).

Isolation, Characterization and Expression of a Full-Length cDNA. Screening of 2 x 10^6 recombinants from the pcD human fibroblast cDNA library with a 26-mer oligonucleotide mixture of 576 species corresponding to internal peptide T-106A detected two putative positive clones. pAGB-1, which hybridized with all four oligonucleotide mixtures, had a 1.8 kb insert with an open reading frame of 1242 bp, a 514 bp 3' untranslated region, and a poly(A) tract, but no apparent 5' untranslated sequence. Authenticity was established by colinearity of the pAGB-1 insert's predicted amino acid sequence with 129 microsequenced residues of the purified protein. In order to isolate a full-length cDNA, the 0.9 kb 5' *Bam*HI fragment from the pAGB-1 insert was radiolabelled and used to screen a human placental cDNA library. Of 32 putative positive clones (pAGB-3 to 34), pAGB-3 contained the longest insert and was sequenced in both

Figure 1. Reversed-phase HPLC separation of tryptic peptides from electroeluted 117 kDa (A) and 48 kDa (B) species of purified human α-GalNAc. The indicated peptides were microsequenced. (Inset) NaDodSO$_4$/PAGE of purified α-GalNAc. See text for details.

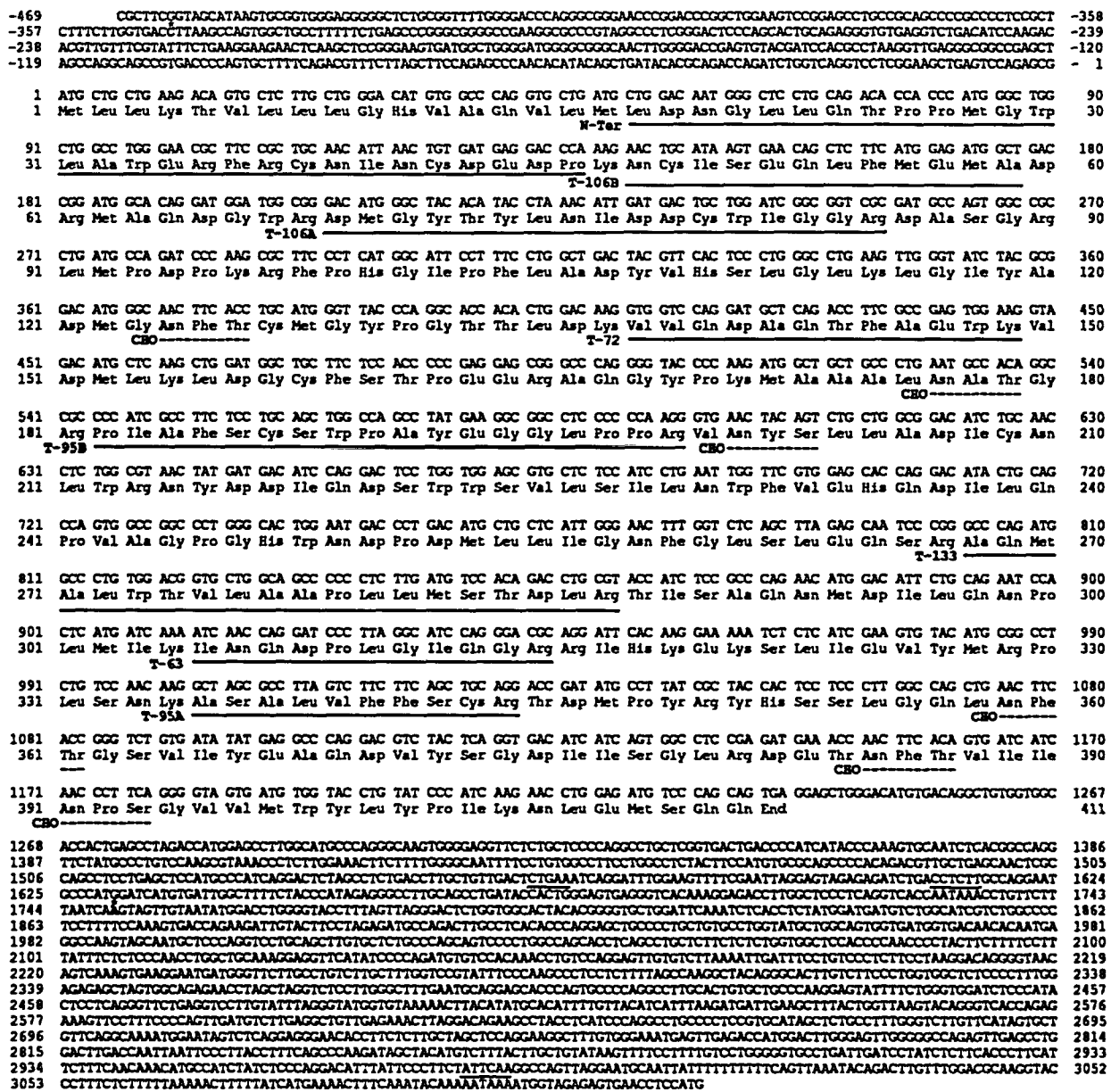

A
T-63
T-72
T-95A&B
T-106A&B
T-133
B
1
2
kDa
67
43
30
B
A
0
40
60
80
100
TIME (min)

orientations. As shown in Fig. 2, the 2158 bp pAGB-3 insert had a 344 bp 5' untranslated region, a 1236 bp open reading frame which encoded 411 amino acids, a 514 bp 3' untranslated region and a 64 bp poly(A) tract. An upstream, inframe ATG occurred at -192 nt, but there were inframe termination codons at -141, -135, and -120 nt, indicating that the -192 ATG was non-functional. A single consensus polyadenylation signal (AATAAA) and a consensus recognition sequence (CACTG) for the U4 small nuclear ribonucleoprotein (34) were located 16 and 65 bp from the poly(A) tract, respectively. In retrospect, the partial cDNA, pAGB-1 had the entire 1236 bp coding region as well as 6 bp of 5' untranslated sequence.

Analysis of the deduced amino acid sequence of pAGB-3 indicated a signal peptide sequence of 17 residues since Leu-18 was the *N*-terminal residue of the microsequenced mature enzyme. When the weight-matrix method of von Heijne (35) was used to predict the peptidase cleavage site, the preferred site, between Ala-13 and Gln-14, had a score of 4.34, whereas cleavage after Met-17 had a score of 2.38. The predicted molecular mass of the 394 residue mature, unglycosylated enzyme subunit (M_r=44,700) was consistent with that (48 kDa) estimated by NaDodSO$_4$/PAGE of the purified glycosylated enzyme. These findings suggest that the mature glycoprotein subunit had at least two *N*-linked oligosaccharide chains, although there were six putative *N*-glycosylation sites at Asn residues 124, 177, 201, 359, 385 and 391 (Fig. 2).

For transient expression, the pAGB-3 full-length cDNA insert was subcloned into the eukaryotic expression vector p91023(B) and the construct, p91-AGB-3, was transfected into COS-1 monkey kidney cells. Compared to the endogenous mean α-GalNAc activity in mock transfected COS-1 cells (35 U/mg; range: 23-50 U/mg; n=6), the transfected cells had a mean activity of 600 U/mg (range: 104-2,400 U/mg; n=6) 72 hr after transfection, or about 17 times the endogenous activity. The expressed human enzyme protein also was detected by

Figure 2. Nucleotide and predicted amino acid sequences of the pAGB-3 and pAGB-35 cDNA inserts containing the complete coding region for human α-*N*-GalNAc. The A of the initiation ATG is nt +1 and the *N*-terminal Met of the signal peptide is amino acid 1. Bold underlines indicate colinear amino acid sequence obtained by microsequencing the *N*-terminal (N-Ter) and tryptic peptides (T) of the purified enzyme. CHO indicates potential sites of *N*-glycosylation. Overlines indicate the polyadenylation signals (AATAAA) and the pentanucleotide sequence (CACTG) recognized by the U4 small nuclear ribonucleoprotein.


N-Ter
T-106B
T-106A
CBO
T-72
T-95B
T-133
T-63
T-95A

immunoblot analysis using rabbit-anti-human α-GalNAc antibodies, whereas the endogenous monkey enzyme was variably visible as a faint band at ~40 kDa or was not detectable (Fig. 3). The expressed human enzyme subunit had a molecular weight of ~48 kDa, indicating that it was glycosylated.

Northern Hybridization and Cap-Site Analyses. Northern hybridization analyses revealed two transcripts in total, cytoplasmic, or poly(A)$^+$ RNA of about 2.2 and 3.6 kb, which were present in similar amounts (not shown). The cap-site was determined to be at -347, or 3 nt beyond the 5' end of the pAGB-3 cDNA insert by primer extension of total placental RNA using two overlapping oligonucleotide probes. The 3.6 kb transcript was determined to be due to a second polyadenylation signal of AATAAA 1366 nt downstream of the first hexamer identified in the pAGB-3 cDNA. This 3.6 kb cDNA encoding α-GalNAc was isolated from a human retinal library and was identical to the pAGB-3 cDNA sequence with the exception of an additional 125 bp and 1379 bp of 5' and 3' untranslated sequences respectively (Fig. 2). The pAGB-35 cDNA confirmed the alternative use of polyadenylation hexamers in producing two mRNA transcripts, and also suggests that the α-GalNAc gene has multiple transcriptional start sites, and that the -347 cap-site is only one of at least two such sites.

Sequence Homology Between α-GalNAc with α-Gal A. Computer-assisted searches of nucleic acid and protein data bases revealed no significant amino acid sequence similarities between α-GalNAc and that of any other DNA or protein sequence except for human α-Gal A (32). Comparison of the nucleic acid and deduced amino acid sequences of the full-length α-GalNAc and α-Gal A cDNAs revealed 55.8% and 46.9% overall homology, respectively. Since the intron/exon junctions and the entire genomic sequence encoding human α-Gal A have been determined (32, 36), it was possible to compare the α-GalNAc amino acid sequence with those deduced from each of the seven α-Gal A exons (Fig. 4). Notably, there was remarkable identity

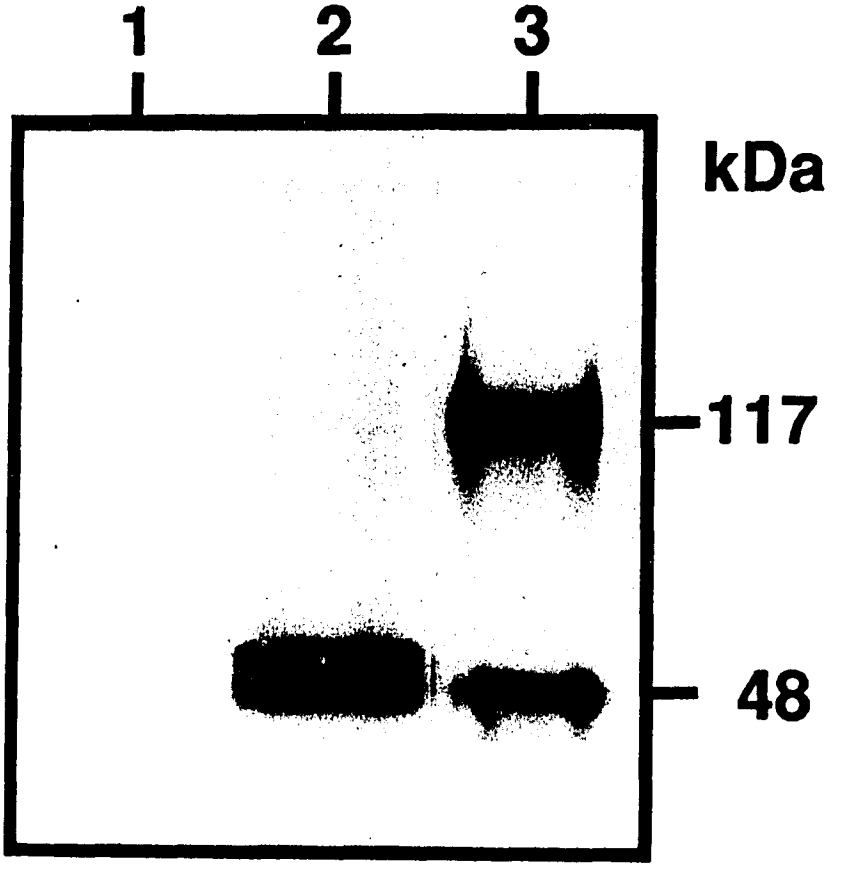


Figure 3. Immunoblot of human α-GalNAc expressed in COS-1 cells. Lanes: 1, mock-transfection; 2, p91-AGB-3 transfection; 3, purified human lung α-GalNAc.

Figure 4. Alignment of amino acid sequences deduced from the full-length cDNAs encoding human α-GalNAc (α-Gal B), α-Gal A, yeast *Mel 1*, and *E. coli Mel A*. Colons, identical residues; single dots, isofunctional amino acids; and boxes, identical or isofunctional residues in α-GalNAc, α-Gal A, *Mel 1* and/or *Mel A*. Gaps were introduced for optimal alignment. Numbered vertical lines indicate exon boundaries for α-Gal A (32).


Cal B: 1
Cal A: 1
Mal 1: 1
1¡2
2¡3
3¡4
4¡5
5¡6
6¡7

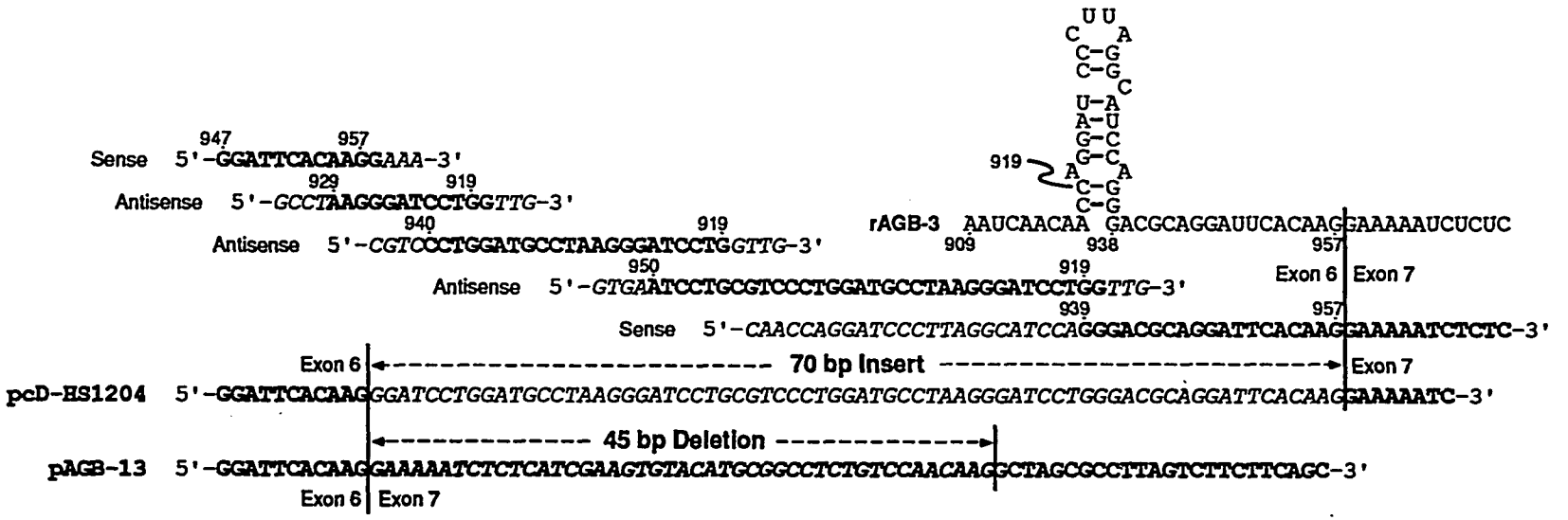
(56.4%) between the α-GalNAc sequences corresponding to those of α-Gal A exons 1 through 6. For example, all eight cysteine residues in α-GalNAc were present in the identical positions in α-Gal A. Of the 14 proline and 23 glycine residues in α-Gal A, 10 and 20 were conserved in identical positions in α-GalNAc, respectively. Furthermore, all four of the α-Gal A *N*-glycosylation sites were conserved in α-Gal B. Putative functional domains were suggested by shorter stretches of amino acid homology shared by α-GalNAc, α-Gal A, yeast α-galactosidase (*Mel 1*) (37) and/or *E. coli* α-galactosidase (*Mel A*) (38) in α-Gal A exons 1 through 6. In contrast, there was little, if any, similarity in the predicted α-GalNAc carboxy-terminal amino acid sequence after residue 319 which corresponded to α-Gal A exon 7 (15.8% homology with numerous gaps). In addition, there were no significant similarities for the cDNAs encoding other human lysosomal polypeptides, with the exception of a short α-GalNAc sequence (residues 365 to 371) in which six out of seven amino acids were identical to residues 194 to 200 in the β-hexosaminidase α-chain, a lysosomal polypeptide with *N*-acetylgalactosaminidase specificity (39). These findings suggested that a cDNA construct containing α-Gal A exons 1-6 joined to α-GalNAc exon 7 might express a hybrid protein with both α-Gal A and B activities. Therefore, a hybrid cDNA containing α-Gal A exons 1 through 6 (nt -60-1029) and α-GalNAc exon 7 (nt 958-1258) was constructed and expressed in COS-1 cells; however, neither immunoreactive protein nor enzymatic activity for either α-Gal A or α-GalNAc were detected.

The finding of extensive homology between α-GalNAc and α-Gal A suggested that they evolved by duplication and divergence of an ancestral sequence for α-Gal A exons 1 through 6. Although there is little, if any, homology among the other lysosomal amino acid sequences (i.e., no "lysosomal domains"), there are notable examples of lysosomal enzyme subunits, pseudogenes or gene families which presumably evolved by duplication and divergence (e.g., 39-41). Future comparison of the α-GalNAc and α-Gal A intron/exon boundaries should

provide further information on the evolution of these lysosomal genes which encode structurally related, but functionally specific glycohydrolases.

Primer Extension and PCR and Sequence Analyses of cDNA and Genomic Sequences. During the course of these studies, Tsuji *et al.* reported a similar human α-GalNAc cDNA sequence (20) which differed from pAGB-3 by a 70 bp insertion after nt 957 (Fig. 5A) and by several substitutions (nt 493, 494, 524, 614, and 667). The 70 bp insertion consisted of three inverted repeats (nt 919-926, 919-936, and 919-944) and a direct repeat (nt 940-957) from the pAGB-3 coding sequence nt 919 to 957. Analysis of the pAGB-3 cDNA sequence from nt 760-1053 using an RNA folding program (29) predicted a stem and loop structure from nt 918 to 937 (Fig. 5A) which could stall or stop reverse transcription of the α-GalNAc mRNA during cDNA synthesis. To determine if this secondary structure could cause cDNA synthesis errors in library construction, a 32-mer oligonucleotide primer was used to extend total placental RNA and α-GalNAc transcripts generated *in vitro* with the riboprobe construct, rAGB-3. Stops of varying intensity were observed from nt 903 to 1009, including two weak stops at the 3' base (nt 940) and 5' end (nt 921) of the stem and loop structure (Fig. 5A). However, there were no strong stops in this region. Although the actual mechanism is unknown, these findings were consistent with the 70 bp insertion resulting from a complex abnormality involving an RNA-DNA duplex in cDNA library construction (42). Another possibility would be an insertion due to a complex strand-switching event involving DNA polymerase I (43). Alternatively, this 70 bp insertion may have resulted from alternative splicing, although the insertion predicts a truncated α-GalNAc polypeptide of 358 residues. To investigate the possible occurrence of α-GalNAc transcripts with a 70 bp insertion after pAGB-3 nt 957, PCR was used to amplify this region in 1) reverse-transcribed mRNA from various sources, 2) the cDNA inserts from clones pAGB-4 to 34, and 3) the gAGB-1 genomic clone. If the cDNA inserts or reverse-transcribed RNAs

Figure 5. Partial genomic sequence of human α-GalNAc including an intron between coding nt 957 and 958. (A) rAGB-3: partial α-GalNAc RNA sequence (nt 909 to 969) corresponding to the 3' end of α-Gal A exon 6 (32). The indicated stem and loop structure between nt 918 and 937 had a ΔG of -11.6 (29). The overlapped antisense and sense sequences shown in bold are inverted and direct repeats derived from nt 919 to 957 of pAGB-3 that are in the 70 bp insertion in pcD-HS1204 (20). The 45 bp deletion in clone pAGB-13 is indicated in italics. (B) The genomic α-GalNAc sequence from coding nt 760 to 1053 (upper case) includes a 1754 nt intron between nt 957 and 958 which corresponds in position to the α-Gal A exon 6 and 7 boundaries. Dashed line, the 5' splice donor sequence; solid underlines, putative branch point sequences; dotted underlines, putative polypyrimidine tracts at the 3' acceptor sites for the normal gene and mutant pAGB-13; and asterisks, differences from the consensus sequence (44).



B

```
760  CTG CTC ATT GGG AAC TTT GGT CTC AGC TTA GAG CAA TCC CGG GCC CAG ATG GCC CTG TGG ACG GTG CTG GCA GCC CCC CTC TTG ATG TCC  848
     Leu Leu Ile Gly Asn Phe Gly Leu Ser Leu Glu Gln Ser Arg Ala Gln Met Ala Leu Trp Thr Val Leu Ala Ala Pro Leu Leu Met Ser

849  ACA GAC CTG CGT ACC ATC TCC GCC CAG AAC ATG GAC ATT CTG CAG AAT CCA CTC ATG ATC AAA ATC AAC CAG GAT CCC TTA GGC ATC CAG  938
     Thr Asp Leu Arg Thr Ile Ser Ala Gln Asn Met Asp Ile Leu Gln Asn Pro Leu Met Ile Lys Ile Asn Gln Asp Pro Leu Gly Ile Gln

939  GGA CGC AGG ATT CAC AAG | gtactagggtgtggagggaaggaaggggagggctgaggaactgggttctcctgagagaaaggctgccagctccctgggggcaacacctggcgagg  957
     Gly Arg Arg Ile His Lys |
                        Exon 6

     tacaggagtcgcccagtccccaaccagggctacccctttctggttgcttatggttgaggactctgatgggagctgctccaactgtcctcctcttgctgggtgagagcagggctgagcagg
     acagctcaagggagtcggggatgagaggtgtcagccacataagtgcacatagcaagggtgaggcacagagcttctatacacccgtgatggcctgcagagagcttggacttccctccaga
     gcaggaggagctggtttgtttgtttttgagacagggtctcactctgtcacccaggctggagtgcagtggcacaatttcgactcactgcaatctctacctgccaggttcaagcaattctc
     gtgcctcagcctcctgagtagctggcactacaggcgcctgcaccacacccagctaatttttgtatttttagtagagacaccatgttggccaggcttgtctcgaactcctggcctcaggt
     gatccacccgtatcagcctcccaaagtgctgggattacaggcatgagcaccgcactcggccaggagaagctgttatagccaaggaatactacgactactggtggctgctatttattgag
     tacctaccatgtgctgggagttttagataatttttctcagcaaggtagttatcttgccattttacaaatgagaaaaatgaaacttcgagagtctgagtaactttatcccaaggctacac
     agttggtacaaacaagactggacttcagtgtcacctcaaagccttttttttttttttttttttgagatggagtctcacgctgtagcccaggctggagtgcagtggcaccatctcagct
     cactgcaacctctggctcccaggttcaagcgattttcctgcctcagcctcccaggtagctgggattacaggtgtgcgccaccacacccggctaatttttttttgtatttttttcagtaga
     gacagggtttcaccatgttggccaggctactctcaaaactcctgacgtcagctgatccactgcctcggcctcacaaagtaatgggattacagcatgagccactgtgcctgtctgccttt
     gctctttaccaaatcctggattctggtaaaaagaaacctacagaactatggaaggcacctatagaactggtgatgcccagaggaagtaacaattccctgccagaggggctgatggtgga
     gctgggcctggaaaaccttctggaggatgggagttcacatccagctccactctccaccctcctggaacagagttcactgttcccactggacagcaccctccaggccagcactggcagct
     gtttggggccagcactcatacgctgtactgttgttgcgcttccctgtttctgcgtttatccctcccgttgtcctatgagcttctggggcagggctcatgcagcacttgtctcagtgtgc
     tagcatagggccgggctcagagtaggtgttgatgagtatctgctgagtcagggaaggtgggcagatagggttagataagctggggtgctggaggcccgtgcgatcctccctaaacctg
958  tgtgacatggagctgtgaactgggggacccagaactcagggagggccagggaggcaatggtaggtcctgtctgagcaagggacccagccagtagccaccttctgtgcccagGAA AAA  963
                                                                                                                  Glu Lys
                                                                                                                  Exon 7
964  TCT CTC ATC GAA GTG TAC ATG CGG CCT CTG TCC AAC AAG | GCT AGC GCC TTA GTC TTC TTC AGC TGC AGG ACC GAT ATG CCT TAT CGC TAC 1053
     Ser Leu Ile Glu Val Tyr Met Arg Pro Leu Ser Asn Lys | Ala Ser Ala Leu Val Phe Phe Ser Cys Arg Thr Asp Met Pro Tyr Arg Tyr
                                                        Exon 7 - pAGB-13
```

contained the 70 bp insert, a 290 bp PCR product would be observed, whereas the absence of the insert would result in a 220 bp PCR product. Only the 220 bp product was observed in PCR-amplified reverse-transcribed total RNA from human lymphoblasts, fibroblasts, and placenta, or in Poly(A)$^+$ mRNA from brain (not shown). Thus, these analyses did not detect longer or shorter transcripts. All of the pAGB-4 through 34 cDNA inserts had only the 220 bp PCR product with the exception of pAGB-13, which had an inframe 45 bp deletion after pAGB-3 nt 957 (*i.e.* deleted nt 958 to 993). A short direct repeat (ACAAG) was present at both breakpoint junctions. Notably, the deletion occurred at the identical 5' site of the 70 bp insertion in pcD-HS1204 (14) (Fig. 5A).

Subsequent sequencing of the region including pAGB-3 codons 254 to 351 in the genomic clone, gAGB-1, revealed a 2048 bp sequence containing a 1754 bp intron between pAGB-3 nt 957 and 958. The intronic sequence had no homology with α-Gal A intron 6, contained two *Alu*-repetitive sequences in reverse orientation and did not have the 70 bp insertion in either orientation (Fig. 5B). It was remarkable that both the pAGB-13 deletion and the pcD-HS1204 insertion occurred at the 5' donor splice site, nt 957, of this intron. Perhaps the location of the consensus lariat branch point sequences in the intron far upstream (94 and 199 bp) from the 3' splice site may impair splicing (44). This concept is supported by the pAGB-13 deletion in which the more closely positioned cryptic lariat branch point and 3' splice site were used. Thus, this intron or surrounding region may have a unique sequence and/or secondary structure that impairs the fidelity of hnRNA processing. Since the intron/exon junction after coding nt 957 also is the site of divergence between the α-Gal A and B sequences, this region also may be mechanistically important in the evolution of human α-GalNAc.

In conclusion, the availability of an authentic full-length cDNA encoding human α-GalNAc should permit the characterization of the structure/function and evolutionary

relationships of α-GalNAc and α-Gal A as well as the identification of the molecular lesions that cause Schindler disease.

REFERENCES

1. Kint, J. A. (1971) *Arch. Int. Physiol. Biochem.* **79,** 633-644

2. Beutler, E. and Kuhl. W. (1972) *Amer. J. Hum. Genet.* **24,** 237-249

3. Romeo, G., Childs, B. and Migeon B. R. (1972) *FEBS Lett.* **27,** 161-166

4. Wood, S. and Nadler, H. L. (1972) *Am. J. Hum. Genet.* **24,** 250-255

5. Ho, M. W., Beutler, S., Tennant, L. and O'Brien, J. (1972) *Am. J. Hum. Genet.* **24,** 256-266

6. Desnick, R. J., Allen, K. Y., Desnick, S. J., Raman, M. K., Bernhlohr, R. W. and Krivit, W. (1973) *J. Lab. Clin. Med.* **81,** 157-171

7. Desnick, R. J. and Bishop, D. F. (1989) in *The Metabolic Basis of Inherited Disease.* (Scriver, C. R., Beaudet, A. L. Sly, W. S. and Valle, D., eds), pp. 1751-1796, McGraw Hill, New York.

8. Beutler, E. and Kuhl, W. (1972) *J. Biol. Chem.* **247,** 7195-7200

9. Callahan, J. W., Lasilla, E. L., Den Tandt, W. and Philippart, M. (1973) *Biochem. Med.* **7,** 424-431

10. Dean, K. J., Sung, S. and Sweeley, C. C. (1977) *Biochem. Biophys. Res. Comm.* **77,** 1411-1417

11. Schram, A. W., Hamers, M. N. and Tager, J. M. (1977) *Biochim. Biophys. Acta.* **482,** 138-144

12. Kusiak, J. W., Quirk, J. M. and Brady, R. O. (1978) *J. Biol. Chem.* **253,** 184-190

13. Dean K. J., and Sweeley, C. C. (1979) *J. Biol. Chem.* **254,** 10001-10005

14. Bishop, D. F., Dean, K. J., Sweeley, C. C. and Desnick, R. J. (1980) in *Enzyme Therapy in Genetic Disease:2,* (Desnick, R.J., ed.), pp. 17-32, Alan R. Liss, Inc., New York

15. deGroot, P. G., Westerveld, A., Meera-Khan, P. and Tager, J. M. (1978) *Hum. Genet.* **44,** 305-312

16. Sweeley, C. C., LeDonne, N. C. and Robbins, P. W. (1983) *Arch. Biochim. Biophys.* **223,** 158-165

17. van Diggelen, O. P., Schindler, D., Willemson, R., Boer, M., Kleijer, W. J., Huijmans, J. G. M., Blom, W. and Galjaard, H. (1988) *J. Inher. Met. Dis.* **11,** 349-357

18. Schindler, D., Bishop, D. F., Wolfe, D. E., Wang, A. M., Egge, H., Lemieux, R. U. and Desnick, R. J. (1989) *N. Engl. J. Med.* **320,** 1735-1740

19. Schindler, D. Kanzaki, T. and Desnick, R. J. (in press) *Clin. Chim. Acta.*

20. Tsuji, S., Yamanchi, T., Hiraiwa, M., Isobe, T., Okuyama, T., Sakimura, K., Takahashi, Y., Mishizawa, M., Uda, Y. and Miyatake, T. (1989) *Biochem. Biophys. Res. Comm.* **163,** 1498-1504

21. Bishop, D. F. and Desnick, R. J. (1981) *J. Biol. Chem.* **256,** 1307-1316

22. Calhoun, D., Bishop, D. F., Berstein, H. S., Quinn, M., Hantzopoulos, P. and Desnick, R. J. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 7364-7368

23. Hunkapillar, M. W., Lujan, E., Ostrander, F. and Hood, L. E. (1973) *Methods Enzymol.* **91,** 227-236

24. Tsai, S. F., Bishop, D. F. and Desnick, R. J. (1988) *Proc. Natl. Acad. Sci. USA* **55,** 7049-7053

25. Hunkapillar, M. W. and Hood, L. E. (1983) *Science* **219,** 650-569

26. Maniatis, T., Fritsch, E. F. and Sambrook, J. (eds.) (1982) in *Molecular Cloning: A Laboratory Manual,* pp. 309-328, Cold Spring Harbor Laboratories, NY, NY

27. Sanger, F., Coulson, A. R., Barrell, B. G., Smith, A. J. H. and Roe, B. A. (1980) *J. Mol. Biol.* **143,** 161-178

28. Wolf, H., Modrow, S., Motz, M., Jameson, B. A., Hermann, G., and Fortsch, B. (1988) *CABIOS.* **4,** 187-191

29. Zuker, M. (1989) *Methods Enzymol.* **180,** 262-88

30. Wong, G. G., Witek, J. S., Temple, P. A., Wilkens, K. M., Leary, A. C., Luxenberg, D. P., Jones, S. S., Brown, E. L., Kay, R. M., Orr, E. C., Shoemaker, C. S., Golde, D. W., Kaufman, R. J., Hewick, R. M., Wang, E. A. and Clark, S. C. (1985) *Science* **228,** 810-813

31. Chen, C. and Okayama, H. (1987) *Mol. Cell. Biol.* **7,** 2745-2752

32. Bishop, D. F., Kornreich, R. and Desnick, R. J. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 3903-3907

33. Ho, S. N., Hunt, H. D., Horton, R. M., Pullen, J. K. and Pease, L. R. (1989) *Gene* **77,** 51-59

34. Berget, S. M. (1984) *Nature (London)* **309,** 179-182

35. von Heijne, G. (1986) *Nucleic Acids Res.* **14,** 4683-4960

36. Kornreich, R., Desnick, R. J. and Bishop, D. F. (1989) *Nucleic Acids Res.* **17,** 3301-3302

37. Liljeström, P. L. (1985) *Nucleic Acids Res.* **13,** 7257-7268

38. Liljeström, P. L. and Lijeström, P. *Nucleic Acids Res.* **15,** 2213-2220

39. Proira, R. L. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 1883-1887

40. Horowitz, M., Wilder, S., Horowitz, A., Reiner, O., Gelbart, T. and Beutler, E. (1989) *Genomics* **4,** 87-96

41. Schuchman, E. H., Jackson, C. E. and Desnick, R. J. (1990) *Genomics* **6,** 149-158

42. Roberts, J. D., Preston, B. D., Johnston, L. A., Soni, A., Loeb, L. A. and Kunkel, T. A. (1989) *Mol. Cell. Biol.* **9,** 468-476

43. Papanicolaou, C. and Ripley, L. S. (1989) *J. Mol. Biol.* **207,** 335-353

44. Reed, R. and Maniatis, T. (1988) *Genes Dev.* **2,** 1268-1276

CHAPTER TWO:

SCHINDLER DISEASE: THE MOLECULAR LESIONS IN THE α-*N*-ACETYLGALACTOSAMINIDASE GENE THAT CAUSE TYPE I AND II DISEASE

ABSTRACT

Schindler disease is a recently recognized disorder resulting from the deficient activity of the lysosomal hydrolase, α-*N*-acetylgalactosaminidase (α-GalNAc). Two clinical phenotypes of α-GalNAc deficiency have been described. The Type I disease is similar to the infantile form of neuroaxonal dystrophy, whereas the Type II disorder is characterized by disseminated angiokeratoma corporis diffusum with no neurological involvement. The recent isolation and expression of the full-length cDNA encoding α-GalNAc has facilitated the identification of the molecular lesions in the first affected homozygotes with each of the remarkably distinct forms of this autosomal recessive disease. Southern and northern hybridization analyses of DNA and total RNA from the Type I and II affected homozygotes revealed normal α-GalNAc gene structures and normal transcript sizes and amounts. Therefore, the reverse-transcribed α-GalNAc transcripts from affected homozygotes with Type I and Type II disease were amplified individually by the polymerase chain reaction (PCR) for DNA sequence analysis. In a sib from Family D with Type I disease, a single G to A transition at nucleotide 973 was detected in multiple subclones of the PCR-amplified α-GalNAc transcript. This point mutation results in a glutamic acid to lysine substitution in residue 325 of the α-GalNAc polypeptide. In the Type II homozygote from Family K, a single C to T transition was identified at nucleotide 985 which results in an arginine to tryptophan substitution at position 329. These base substitutions were confirmed by dot-blot analyses of PCR-amplified genomic DNA from members of each family using allele-specific oligonucleotides. Furthermore, transient expression of α-GalNAc constructs containing the 973 G to A transition and the 985 C to T transition each resulted in the expression of an immunoreactive polypeptide which had no detectable α-GalNAc activity.

INTRODUCTION

Schindler disease is a recently recognized disorder resulting from the deficient activity of the lysosomal glycohydrolase, α-*N*-acetylgalactosaminidase (E.C. 3.2.1.29; α-GalNAc) (1-5). The enzymatic defect, inherited as an autosomal recessive trait, leads to the tissue accumulation and increased urinary excretion of glycopeptides and oligosaccharides containing α-*N*-acetylgalactosaminyl moieties (6-8). Two subtypes, designated Type I and II, with variant phenotypes have been described (2, 5). The Type I disorder is an infantile form of neuroaxonal dystrophy first described in two brothers, the offspring of a consanguineous couple of German descent (Family D). The affected infants had normal births and early developments for the first 9 to 12 months of age. Then, they experienced a period of developmental delay which was followed by rapid regression in the second year of life. By three to four years of age, the progressive neurologic impairment led to cortical blindness, myoclonic seizures, spasticity, decorticate posturing and profound psychomotor retardation. A cortical biopsy revealed the presence of "spheroids" in terminal axons, the unique neuropathology which classified this disease as a neuroaxonal dystrophy (3, 9). The clinical and morphologic manifestations of the two affected brothers with this disease are essentially identical to those in patients with Seitelberger disease (9); however, the finding of normal α-GalNAc activity in eight unrelated patients with Seitelberger disease demonstrated that Schindler disease is biochemically distinct and that the infantile neuroaxonal dystrophies are genetically heterogeneous (3).

The Type I affected homozygotes had α-GalNAc activity levels that were less than 1% of normal in various sources as assayed with the newly synthesized substrate, 4-methylumbelliferyl-α-*N*-acetylgalactosaminide (4MU-α-GalNAc) (3, 10). Their parents had intermediate levels of activity consistent with being obligate heterozygotes for this autosomal recessive disorder. Using monospecific rabbit anti-human α-GalNAc antibodies, immunoblots

of fibroblast extracts from the affected brothers revealed no detectable immunoreactive enzyme protein, while in extracts from their parents and a brother, there were two immunoreactive peptides of 48 and 117 kDa, the monomeric and dimeric forms of α-GalNAc observed in cultured fibroblasts from normal individuals and the purified human enzyme (3). These findings suggested that the α-GalNAc mutation in this family markedly altered the enzyme's activity and stability.

Type II Schindler disease was described in a 46-year old Japanese woman, the daughter of a first-cousin marriage (Family K). This affected homozygote had no neurological involvement, but had disseminated angiokeratoma similar to that observed in males with Fabry disease (4). Urinary oligosaccharide analysis revealed that this individual excreted the same glycopeptides as those characterized in the urine from the two brothers with Type I Schindler disease (6, 7). These findings suggested that this patient may have α-GalNAc deficiency. In fact, the affected homozygote had less than 1% of normal α-GalNAc activity in various sources as assayed with 4-MU-α-GalNAc. Her children had intermediate levels of activity consistent with being obligate heterozygotes for this autosomal recessive disorder. Immunoblotting of fibroblast extracts from this individual also revealed no immunoreactive enzyme protein, while the two peptide species were observed in extracts from her son. These findings suggested that the α-GalNAc mutation in this family also altered the enzymes's activity and stability.

The recent isolation, sequencing and expression of the full-length cDNA encoding human α-GalNAc (11) has facilitated the investigation of the molecular lesions in the α-GalNAc gene that cause Schindler disease in these families. The α-GalNAc cDNA encodes a polypeptide of 411 amino acids including a signal peptide of 17 residues. The precursor polypeptide is co-translationally glycosylated and, following carbohydrate modifications in the Golgi, the mature polypeptide of 394 amino acids dimerizes to form the active enzyme of ~117

kDa. Studies of the biosynthesis of human α-GalNAc in fibroblasts indicated that both the 65 kDa precursor and 48 kDa mature lysosomal forms had only high mannose-type oligosaccharide structures (12). In this communication, the specific base substitutions in the α-GalNAc coding region of each family are identified by sequencing the PCR-amplified products of the reverse-transcribed mutant α-GalNAc mRNAs. Confirmation of this mutation was demonstrated by dot-blot analyses of PCR-amplified genomic DNA from family members, and by transient expression studies of the mutant cDNAs. Of interest was the fact that both of the expressed mutant enzymes had no activity, but were immunologically detectable, whereas no immunoreactive enzyme protein was present in cultured cells from the patients. These findings suggest that the enzyme polypeptide is synthesized *in vivo*, even though sensitive immunologic analyses may reveal the absence of crossreactive immunological material in cultured cells or tissue extracts.

METHODS

Cell lines. Primary cultures of fibroblasts and lymphoblasts were established from skin biopsies and peripheral blood samples obtained from the Family D and K members with Schindler disease and from normal controls with informed consent. The COS-1 cells lines were obtained from the American Type Tissue Collection (Rockville, MD). The fibroblasts, lymphoblasts and COS-1 cell lines were grown in RPMI with 10% fetal bovine serum and 1% penicillin-streptomycin and 2 mM glutamine (Gibco, Grand Island, NY) by standard procedures (13).

Assays of α-GalNAc activity and protein. The α-GalNAc activity in cultured cells and COS-1 cells was determined with the synthetic flurogenic substrate 4-methylumbelliferyl-α-*N*-acetylgalactosaminide (10) as previously described (3). One unit (U) of enzymatic activity is equal to that amount of enzyme required to hydrolyze 1 nmol of 4MU-α-GalNAc per hour. Protein concentration was determined by the fluorescamine assay (14). Immunoblot analyses of α-GalNAc in transfected COS-1 cells was performed using monospecific rabbit anti-human antibodies as described (3, 12).

Southern and northern hybridization analyses. For analysis of the α-GalNAc gene integrity, genomic DNA was isolated from at least 10^6 cultured fibroblasts or lymphoblasts (15). The DNA was digested with restriction endonucleases (e.g., *BamH* I, *Pst* I, *EcoR* I, *Taq* I) (New England Biolabs, Beverly, MA), electrophoresed in 1% agarose, transferred to BioTrace RP charge-modified nylon 66 binding matrix (Gelman Sciences, Inc., Ann Arbor, MI) (16) and analyzed with nick-translated α-GalNAc cDNA (17). For analysis of the relative sizes and amount of the α-GalNAc transcripts, total RNA was isolated from at least 10^8 cultured fibroblasts or lymphoblasts (18). RNA samples were electrophoresed in 1% agarose/formaldehyde denaturing gels (19), transferred to BioTrace (Gelman Sciences), and

analyzed with radiolabelled riboprobe synthesized with SP6 polymerase from the α-GalNAc cDNA cloned into the pGEM4Z vector (Promega, Madison, WI).

DNA amplification and sequencing of the mutant allele. Sense and anti-sense oligonucleotide primers designed to amplify the entire coding region of the α-GalNAc gene in two overlapping fragments were synthesized on a model 380B DNA synthesizer (Applied Biosystems, Foster City, CA). As shown in Fig. 1, the 5' region of the α-GalNAc transcript [cDNA nt -54 to 689 (11)] was amplified using the 32-mer sense primer, P1 (5'-AGTAGTGAATTCCTGATACACGCAGACCAGAT-3') corresponding to α-GalNAc nt -34- -53 with an additional 12 nt which included an *EcoR* I site and the 32-mer antisense primer, P2 (5'-AGTAGTAAGCTTTTCAGGATGGAGAGCTCGCT-3') corresponding to α-GalNAc cDNA nt 670-689 with a *Hind* III site for forced cloning. The 3' sequence of the coding (cDNA nt 595 to 1292) was amplified using the 32-mer sense primer, P3 (5'-AGTAGTGAATTCAGGGTGAACTACAGTCTGCT-3') corresponding to nt 595-614 with an *EcoR* I site and a 32-mer antisense primer, P4 (5'-AGTAGTAAGCTT-GCTCCATGGTCTAGGCTCAG-3') corresponding to nt 1273-1292 and contained a *Hind* III site. Total RNA (10 μg) was reverse-transcribed to cDNA using the BRL cDNA Synthesis Kit (Bethesda Research Laboratories, Gaithersburg, MD). One-fourth of the cDNA product was PCR-amplified (20) using the GeneAmp DNA Amplification Reagent Kit (Perkin Elmer Cetus, Norwalk, CT) and 1 μM of each primer. Each of the 30 PCR cycles consisted of denaturation at 94 °C for 1 min; annealing at 37 °C for 2 min; and extension at 60 °C for 7 min. The PCR products were phenol extracted, ethanol precipitated, resuspended in 20 μl H_20 and 2 μl were analyzed on a 2% agarose gel. The remaining 18 μl was digested with *EcoR* I and *Hind* III and subcloned into M13mp18 and 19 vectors (21). Clones containing the PCR products were identified by plaque hybridization using the nick-translated α-GalNAc cDNA (17). Single-

stranded template was isolated (21) from six separate M13 clones, and each was sequenced in both orientations by the dideoxy chain termination method (22) using universal and sequence-specific primers.

Oligonucleotide hybridization of genomic DNA. To confirm the mutation at the genomic level, 1 μg of genomic DNA from the Family D and K affected homozygotes, obligate heterozygotes, and normal individuals was PCR-amplified by the conditions described above using primers that flanked the region of the mutations. The 5' sense primer was a 20-mer (5'-TGATGTCCACAGACCTGCGT-3') corresponding to nt 842-861 of the full-length cDNA (11) and the 3' antisense primer was a 20-mer (TCAGCTGCAGGACCGATATG) corresponding to nt 1005-1024. Normal and mutation specific oligonucleotide probes (21-mers) were synthesized with the normal sequence 5'-TCTCTCATCGAAGTGTACATG-3' and the mutation-specific sequence 5'-TCTCTCATCAAAGTGTACATG-3' for detection of the Type I base substitution, and 5'-GTGTACATGCGGCCTCTGTCC-3' and 5'-GTGTACATGTGGCCTCTGTCC-3', for identification of the Type II lesion. The calculate T_m values for the Type I normal and mutation specific oligonucleotides were 58 °C and 55 °C, respectively, while the corresponding values for the Type II allele-specific oligonucleotides were 65 °C, and 63 °C, respectively. Dot blots were denatured, hybridized, and washed as described (23).

Construction of the mutant cDNAs and transient expression of the Type I and II mutations in COS-1 cells. A *BamH* I-*Kpn* I fragment containing the Type I G to A or Type II C to T transitions from the respective M13 clones were used as cassettes to replace the corresponding sequence in the pAGB-3 cDNA. The Type I and II constructs were individually subcloned into the eukaryotic expression vector p91023(B) (27) and the sequence of each construct was confirmed by dideoxy sequencing. COS-1 cells were transfected (28) with the

Type I or II full-length construct, harvested at 72 h post-transfection, assayed and immunoblotted as previously described (2).

Computer-assisted analyses. Local secondary structure for both the normal and mutant enzymes were predicted by the algorithms of Chou and Fasman (24) and Garnier *et al.* (25) using the University of Wisconsin Computer Group software (26). A region of 20 amino acids surrounding the site of the Type I and II amino acid substitutions were used for the prediction of the local secondary structure.

RESULTS

Molecular characterization of Families D and K. In order to characterize the nature of the mutations in Families D and K, Southern analysis was performed with several restriction endonucleases of genomic DNA from the affected homozygotes and obligate heterozygotes. No gross gene rearrangements were observed (data not shown). Interestingly, Family D was polymorphic for two restriction endonucleases, *BamH* I and *Taq* I. In a normal Caucasian population of 38 and 58 alleles, the polymorphisms were found at a frequency of about 44% and 2.6%, respectively. Northern analysis of total RNA from the Family D and K homozygotes and obligate heterozygotes revealed the presence of the two α-GalNAc transcripts of ~2.2 kb and ~ 3.0 kb. Each of the transcripts were present in normal amounts in all individuals studied (data not shown), thereby excluding the possibility of splice, promotor, or mRNA stability mutations.

These findings, and the fact that no α-GalNAc immunoreactive protein was detected in fibroblasts from Type I or II homozygotes (2, 5), suggested that the mutations causing Type I and II Schindler disease in these families were due to point mutations or small insertions or deletions within the coding region of the α-GalNAc gene. To identify the mutations in the α-GalNAc coding region of these Type I and II homozygotes, their α-GalNAc transcripts were reverse-transcribed, PCR-amplified and sequenced in both orientations (Fig. 1). In the sequence analyses of the Type I and II α-GalNAc cDNAs, a single unique substitution was found in all Type I or II subclones. The transcript from the Type I homozygote had a G to A transition at nucleotide 973 which resulted in a glutamic acid to lysine substitution at residue 325 (Fig. 2). In all the Type II cDNAs, a C to T transition occurred at nt 985 which would substitute a tryptophan for an arginine at residue 329 (Fig. 2). These substitutions were confirmed by dot-blot hybridization analyses with PCR-amplified genomic DNA from the respective Family D and K affected homozygotes, obligate heterozygotes with allele-specific oligonucleotides.

Figure 1. Strategy for the identification of the specific molecular lesions in Schindler disease Types I and II. Total RNA was isolated from lymphoblasts of affected homozygotes, reverse-transcribed to cDNA, PCR-amplified in two overlapping reactions, cloned into M13 sequencing vectors, and sequenced in both orientations.

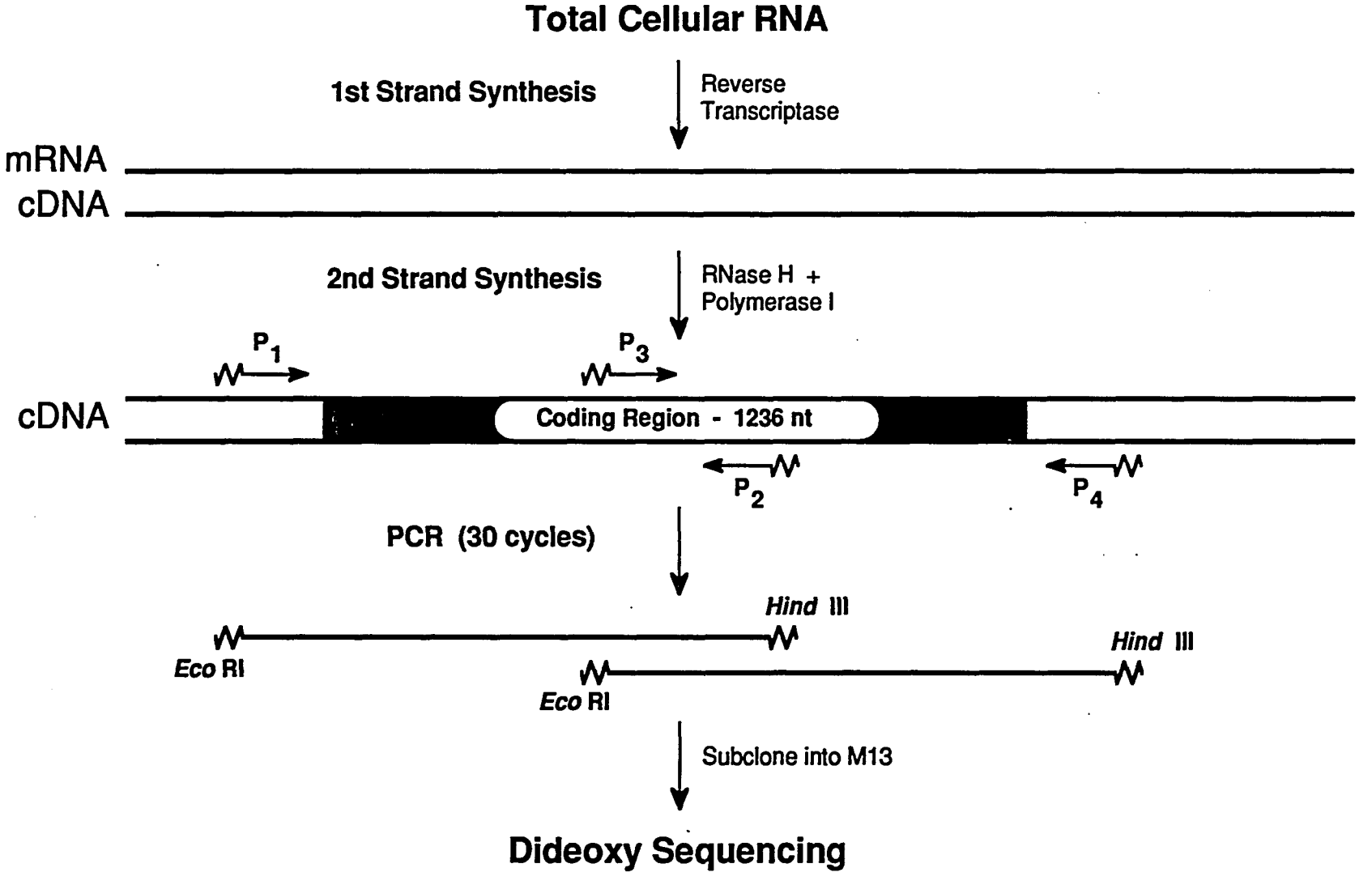
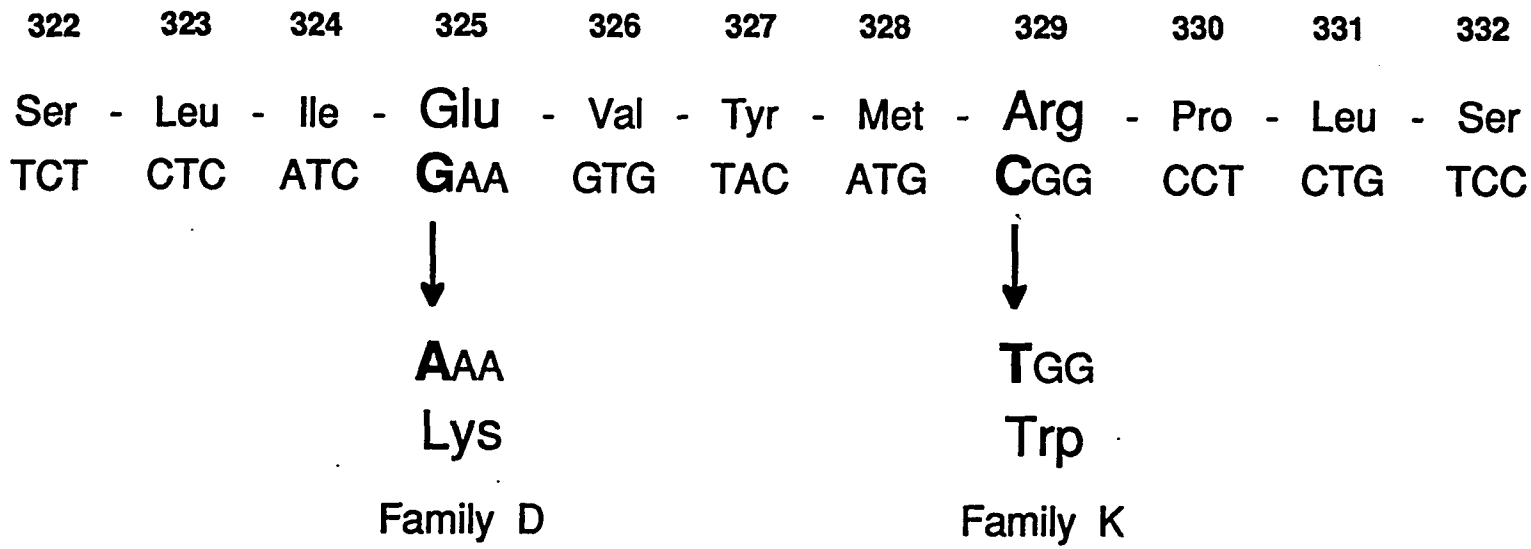

Total Cellular RNA
1st Strand Synthesis
Reverse Transcriptase
mRNA
cDNA
2nd Strand Synthesis
RNase H + Polymerase I
P1
P3
cDNA
Coding Region - 1236 nt
P2
P4
PCR (30 cycles)
Hind III
Eco RI
Hind III
Eco RI
Subclone into M13
Dideoxy Sequencing

Figure 2. The mutation in the Type I homozygote was identified by DNA sequencing to be a G to A transition at nucleotide 973 which caused a glutamic acid to lysine substitution at residue 325. The mutation in the Type II homozygote was identified by DNA sequencing to be a C to T transition at nucleotide 985 which caused an arginine to tryptophan substitution at residue 329.


322 323 324 325 326 327 328 329 330 331 332
Ser - Leu - Ile - Glu - Val - Tyr - Met - Arg - Pro - Leu - Ser
TCT CTC ATC GAA GTG TAC ATG CGG CCT CTG TCC
AAA
Lys
Family D
TGG
Trp
Family K

As shown in Figs 3A and 3B, the respective normal probes hybridized with to normal and heterozygote amplified DNA products, while the respective mutation-specific probes hybridized only with the amplified genomic sequences from the respective Type I or II homozygotes indicating that each homozygote was homoallelic for the respective Type I or II mutations. In support of these findings, PCR-amplified genomic DNA from both obligate heterozygous parents of the Type I homozygotes and both obligate heterozygous children of the Type II homozygote hybridized with the respective mutant probes.

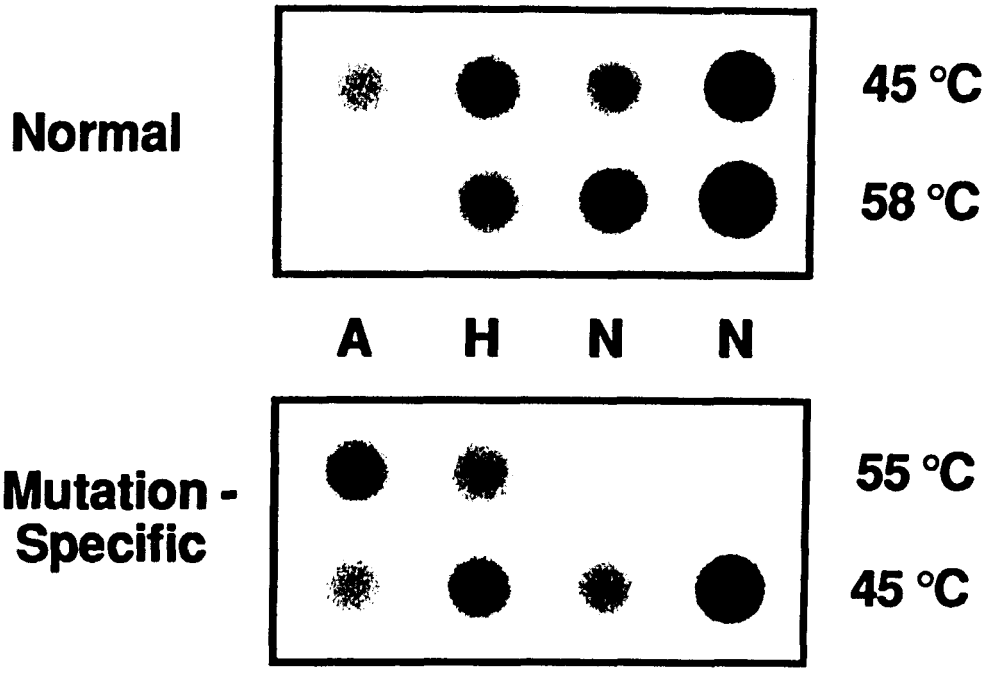
Expression of the α-GalNAc mutations in COS-1. To further characterize the Type I and II mutations, each was introduced as a *BamH* I-*Kpn* I cassette (nt 921-1193) into the full-length pAGB-3 cDNA. The mutated cDNAs, designated pAGB-973A and pAGB-985T, were then subcloned into the eukaryotic expression vector p91023(B) and sequenced. Each construct, p91-AGB-973A or p91-AGB-985T was transfected into COS-1 cells, harvested at 72 h post-transfection, and the extracts assayed for α-GalNAc activity and immunoreactive protein. Immunoblot analysis revealed the presence of immunologically detectable proteins with subunit molecular weights of ~48 kDa expressed form the p91-AGB-973A and p91-AGB-985T constructs (Fig. 4A and B). However, neither mutant protein expressed 4MU-α-GalNAc activity above endogenous COS-1 levels, whereas the normal construct p91-AGB-3 expressed immunologically detectable protein which had α-GalNAc activity that was 20-fold greater than in mock-transfected COS-1 cells (Table 1).

Secondary structure analyses. Computer-assisted regional secondary structural analysis of residues 310 to 330 was carried out using the algorithm of Garnier *et al.* (22). Compared to the predicted secondary structure of the normal sequence, the glutamic acid to lysine substitution in the Type I homozygote extended a β-pleated sheet region 5' to the mutation, a created a region of α-helicity around and including the mutation (Fig. 5). The arginine to tryptophan

substitution in the Type II homozygote predicted a shortening of β-pleated sheet 3' to the mutation (Fig. 5).

Figure 3. Allele specific oligonucleotide probes were constructed and hybridized to PCR-amplified genomic DNA by dot blot analysis and washed at the differential temperatures as indicated. A, affected homozygote; H, obligate heterozygote; and N, normal control. (A) The dot blots for Family D with Type I Schindler disease. (B) The dot blots for Family K with Type II Schindler disease.

A


Normal
45 °C
58 °C
A H N N
Mutation - Specific
55 °C
45 °C

B

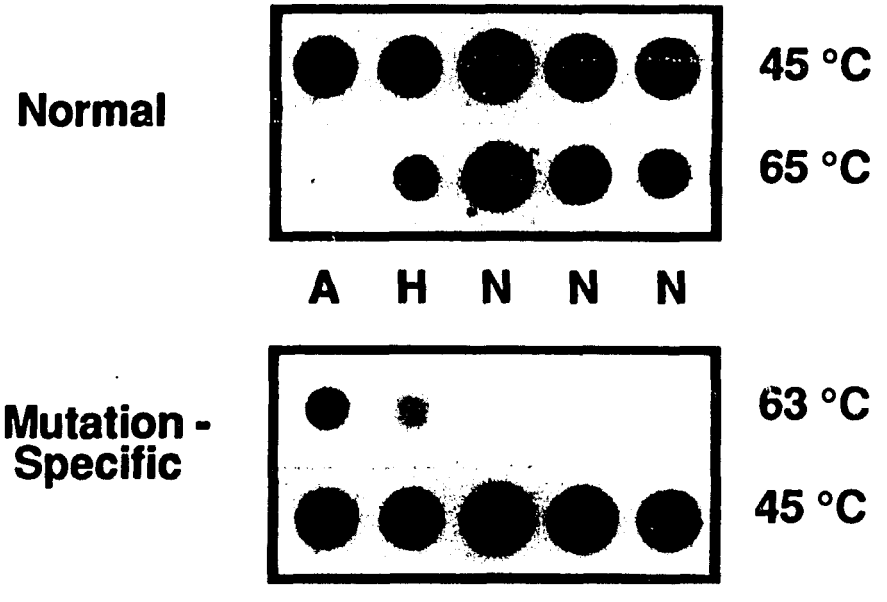
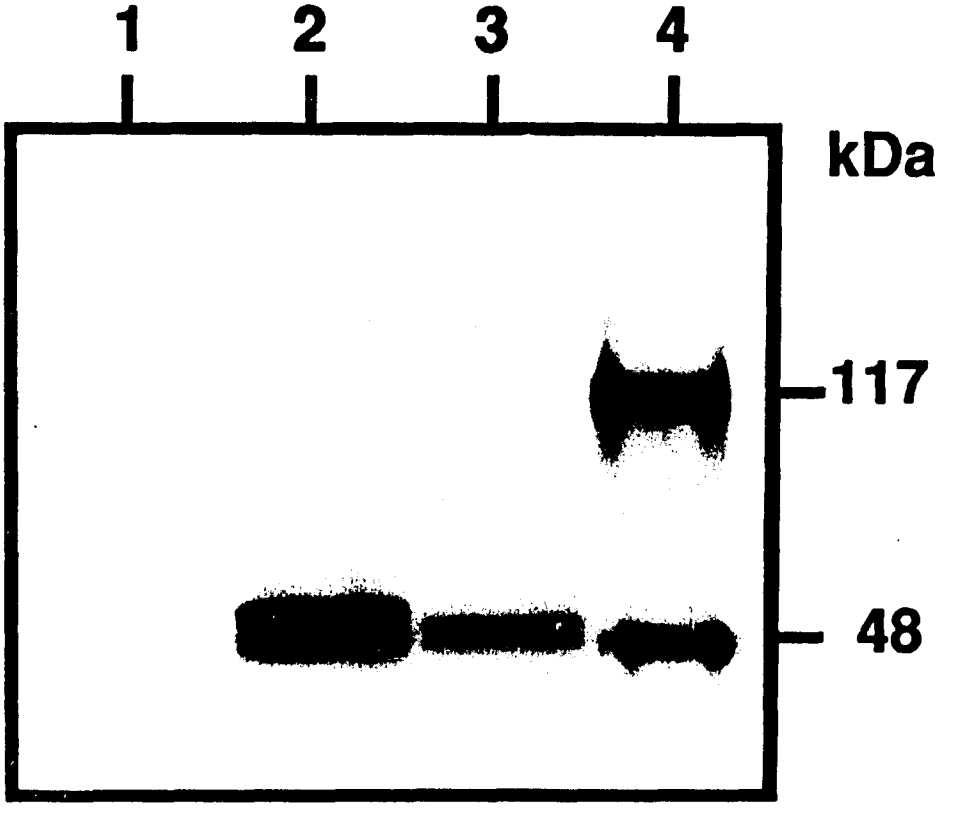

Normal
45 °C
65 °C
A H N N N
Mutation - Specific
63 °C
45 °C

Figure 4. Immunoblot of transiently expressed α-GalNAc in COS-1 cells. (A) Family D: lanes 1, mock transfection; 2, p91-AGB-3; 3, p91-AGB-973A; 4, human fibroblast control. (B) Family K: lanes 1, mock transfection; 2, p91-AGB-3; 3, p91-AGB-985T; 4, human fibroblast control.

A



B

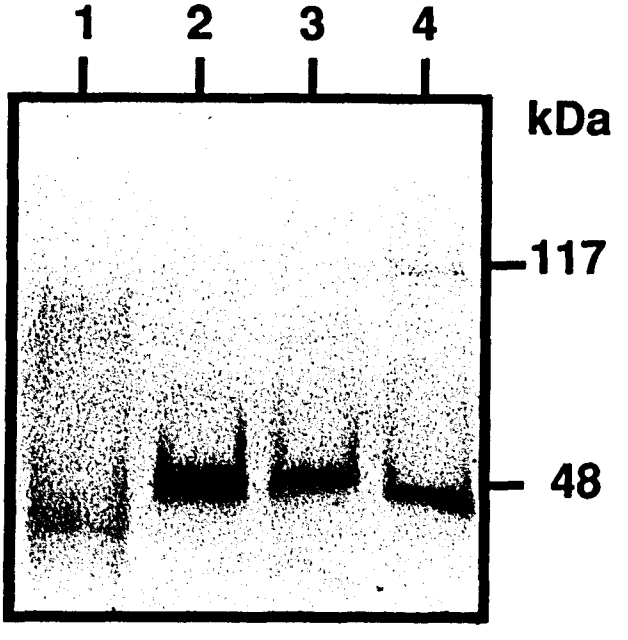
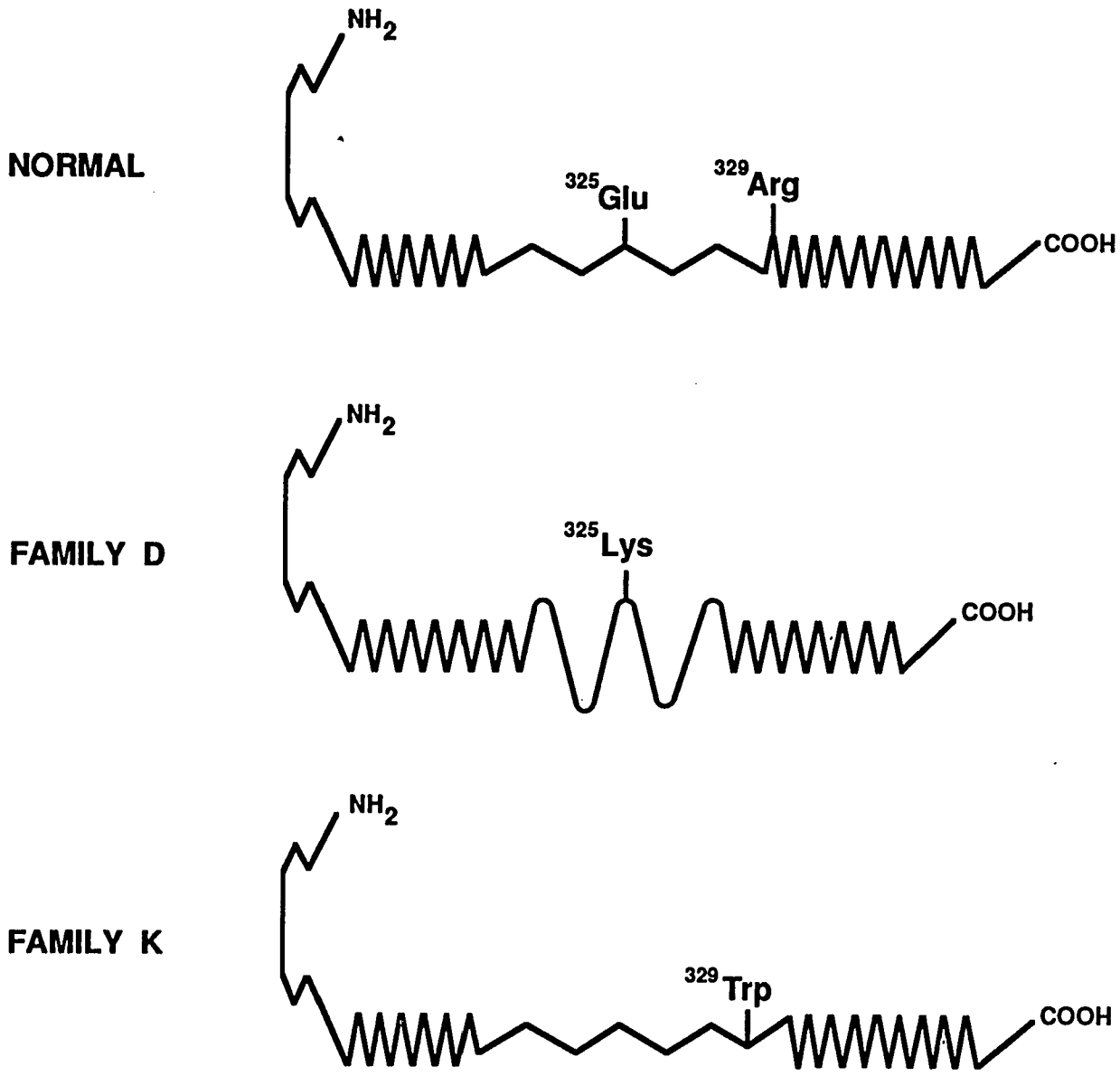


Table 1. Transient expression of normal and mutant α-GalNAc constructs in COS-1 cells.

Construct	α-GalNAc Activity*
	U/mg
p91-AGB-3	
Mean	661
Range (n=6)	105-2415
p91-AGB-973A	
Mean	44
Range (n=6)	27-65
p91-AGB-985T	
Mean	36
Range (n=5)	31-64
Mock Transfection (No DNA)	
Mean	35
Range (n=6)	23-50

*As defined using 0.8 mM 4MU-α–GalNAc as substrate.

Figure 5. Computer assisted secondary structure was determined for the residues flanking and including the mutations for the normal and mutant sequences. Top: normal sequence; middle: Type I substitution of glutamic acid to lysine at residue 325; bottom: Type II substitution of arginine to tryptophan at residue 329.



Amino Acids 310-330

DISCUSSION

The recent identification of two remarkably distinct phenotypes with severely deficient α-GalNAc activity stimulated studies to investigate the molecular nature of these enzymatic defects. Using the full-length α-GalNAc cDNA as a probe, Southern and northern analyses revealed no gross chromosomal rearrangements and transcripts of normal size and abundancy in the DNA and RNA from the affected homozygotes with Type I and II Schindler disease. Subsequent reverse-transcription, amplification, and sequencing of the α-GalNAc transcripts from these affected individuals revealed specific base substitutions. In the two brothers with Type I disease, a G to A transition at nt 973 resulted in a glutamic acid to lysine substitution at residue 325 of the α-GalNAc polypeptide. In the affected woman with Type II disease, a C to T transition at nt 985 was identified which would substitute a tryptophan for an arginine at residue 329. These nucleotide changes were confirmed in genomic DNA from the consanguineous parents of the Type I homozygotes and the children of the Type II homozygote using allele-specific oligonucleotide probes. The mutations causing Type I and II Schindler disease were homoallelic, consistent with the fact that the Type I and II homozygotes were the products of consanguineous marriages. Computer-assisted secondary analysis of this region revealed that these substitutions changed the local secondary structure; the Type I mutation causing an extended β-pleated sheet region 5' to the substitution and a region of α-helicity around and including the mutation, whereas the Type II mutation appeared to alter the secondary structure to a lesser degree by decreasing the β-pleated sheet regions 3' to the mutation.

The fact that fibroblast extracts from the affected homozygotes were CRIM-negative suggested that the two mutations caused the enzyme polypeptide to be unstable and rapidly degraded in cultured cells and, presumably, *in vivo*. Similar findings of CRIM-negative mutant proteins in fibroblasts which were immunologically detectable when transiently expressed in

COS-1 cells have been observed for mutations in the vitamin D receptor gene (29) and the cystathionine β-synthase gene (L. E. Rosenberg, personal communication). The detection of immunoreactive enzyme in the COS-1 cells most likely reflects the "pulsed" synthesis of the altered human polypeptide which then is degraded due to instability and/or the inability of the polypeptide subunits to associate into the more stable multimeric forms. This may be the case for the α-GalNAc polypeptide since the homodimeric form is extremely stable (11). Thus, the classification of a mutation as CRIM-negative based on studies of the patient's cells or tissues does not imply that the enzyme is not synthesized. Subsequent stable expression of normal and mutant constructs may provide sufficient quantities of these enzyme proteins for comparison of their physical and kinetic properties. For example such studies may determine if the mutations alter subunit association, the interaction with different substrates, and/or other factors in the catalytic complex.

REFERENCES

1. van Diggelen, O. P., D. Schindler, R. Willemson, M. Boer, W. J. Kleijer, J. G. M. Huijmans, W. Blom, and H. Galjaard. 1988. α-*N*-acetylgalactosaminidase deficiency, a new lysosomal storage disorder. *J. Inher. Met. Dis.* 11:349-357.

2. Schindler, D., D. F. Bishop, D. E. Wolfe, A. M. Wang, H. Egge, R. U. Lemieux, and R. J. Desnick. 1989. Neuroaxonal dystrophy due to lysosomal α-*N*-acetylgalactosaminidase deficiency. *N. Engl. J. Med.* 320:1735-1740.

3. Desnick, R. J., and D. F. Bishop. 1989. Schindler disease: α-*N*-acetylgalactosaminidase deficiency. *In* The Metabolic Basis of Inherited Disease. C.R. Scriver, A.L. Beaudet, W.S. Sly and D. Valle, editors. McGraw-Hill, New York. 1751-1796.

4. Kanzaki, T., M. Yokota, M. Mizuno, Y. Matsumoto, and Y. Kirabayashi. 1989. Novel lysosomal lycoamino acid storage disease with angiokeratoma corpois diffusum. *Lancet* April 22, 1989:875-877.

5. Wang, A. M., T. Kanzaki, D. Schindler, and R. J. Desnick. 1989. Schindler disease: Molecular defects in the infantile- and recently identified adult-onset forms. Am. J. Hum. Genet. 43:4A228.

6. Linden, H. U., R. A. Klein, H. Egge, J. Peter-Katalinic, J. Dabrowski, and D. Schindler. 1989. Isolation and characterization of sialic acid-containing glycopeptides of the *O*-glycosidic type from the urine of two patients with hereditary deficiency in α-GalNAc activity. *Biol Chem. Hoppe-Seyler.* 370:661-672.

7. Hirabayashi, Y., Y. Matsumoto, M. Matsumoto, T. Toida, N. Iida, T. Masubara, T. Kanzaki, M. Yokota, and I. Shizuka. 1990. Isolation and characterization of major urinary amino acid *O*-glycosides and a dipeptide *O*-glysocide from a new lysosomal storage disorder (Kanzaki disease). *J. Biol. Chem.* 265:1693-1701.

8. Schindler, D., T. Kanzaki, and R. J. Desnick. In press. *Clin Chim Acta.*

9. Seitelberger, F. 1986. Neuroaxonal dystrophy: its relation to aging and neurological disease. In Hankbook of Clinical Neurology, Vol 49. P. J. Vinken, G. W. Bruyn, and H. L. Klawans, editors. Elsevier, Amsterdam. 391-415.

10. Lemieux, R. U., and R. M. Ratcliffe. 1979. The azidonitration of tri-*O*-acetyl-D-galactal. *Can. J. Chem.* 57:1244-1251.

11. Wang, A. M., D. F. Bishop and R. J. Desnick. In press. Human α-*N*-acetylgalactosaminidase: molecular cloning, nucleotide sequence and expression of a full-length cDNA. *J. Biol. Chem.*

12. Sweeley, C. C., N. C. LeDonne, and P. W. Robbins. 1983. Post-translational processing reactions involved in the biosynthesis of lysosomal α-*N*-

acetylgalactosaminidase in cultured human fibroblasts. *Arch. Biochem. Biophys.* 223:158-165.

13. Bernstein, H. S., D. F. Bishop, K. H. Astrin, R. Kornreich, C. M. Eng, H. Sakuraba, and R. J. Desnick. 1989. Fabry disease: Six gene rearrangements and an exonic point mutation in the α-galactosidase gene. *J. Clin. Invest.* 83:1390-1399.

14. Bishop, D. F., and R. J. Desnick. 1981. Affinity purification of α-galactosidase A from human spleen, placental and plasma with elimination of pyrogen contamination. *J. Biol. Chem.* 256:1307-1316.

15. Aldridge, J., L. Kunkel, G. Bruns, U. Tantravahi, M. Lalande, T. Brewster, E. Moreau, M. Wilson, W. Bromley, T. Roderick, and S. A. Latt. 1984. A strategy to reveal high frequence RFLPs along the human X chromosome. *Am. J. Hum. Genet.* 36:546-564.

16. Southern, E. M. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* 98:503-517.

17. Maniatis, T., E. F. Fritsch, and J. Sambrook. 1982. Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratories, New York.

18. Chrigwin, J. M., A. E. Pryzbyla, R. J. MacDonald, and W. J. Rutter. 1979. Isolation of biologically active RNA from sources enriched in ribonuclease. *Biochemistry.* 18:5294-5298.

19. Lehrach, H., D. Diamond, J. M. Wozney, and H. Doedtker. 1977. RNA molecular weight determinations by gel electrophoresis under denaturing conditions, a critical look. *Biochemistry.* 16:4743-4751.

20. Saiki, R. K., S. Scharf, F. Faloona, K. B. Mullis, G. T. Horn, H. A. Erlich, and N. Arnheim. 1985. Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science (Wash. DC).* 230:1350-1359.

21. Messing, J., and J. Vieira. 1982. A new pair of M13 vectors for selecting either DNA strand of double digest restriction fragments. *Gene (Amst.).* 19:269-276.

22. Sanger, F., A. R. Coulson, B. G. Garrell, A. J. H. Smith and B. A. Roe. 1980. Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. J. Mol. Biol. 143:161-178.

23. Theophilus, B., T. Latham, G. A. Grabowski, and F. I. Smith. 1989. Gaucher disease: Molecular heterogeneity and phenotype-genotype correlations. *Am. J. Hum. Genet.* 45:212-225.

24. Chou, P.Y., and G. D. Fasman. 1978. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol.* 47:45-147.

25. Garnier, J., D. J. Osguthorpe, and B. Robson. 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* 120:97-120.

26. Wolf, H., S. Modrow, M. Motz, B. A. Jameson, G. Hermann, and B. Fortsch. 1988. An integrated family of amino-acid sequence-analysis programs. *CABIOS.* 4:187-191.

27. Wong, G. G., J. S. Witek, P. A. Temple, K. M. Wilkins, A. C. Leary, D. P. Luxenberg, S. S. Jones, E. L. Brown, R. M. Kay, E. C. Orr, C. S. Shoemaker, D. W. Golde, R. J. Kaufman, R. M. Hewick, E. A. Wang and S. C. Clarke. 1985. Human GM-CSF: Molecular cloning of the complementary DNA and purification of the natural and recombinant proteins. *Science (Wash, DC)* 228:810-813.

28. Chen, C., and H. Okayama. 1987. High efficiency transformation of mammalian cells by plasmind DNA. *Mol. Cell. Biol.* 7:2745-2752.

29. Ritchie, H. H., M. R. Hughes, E. T. Thompson, P. J. Malloy, Z. Hochberg, D. Feldman, J. W. Pike, and B. W. O'Malley. 1989. An ochre mutation in the vitamin D receptor gene causes hereditary 1,25-dihydroxyvitamin D_3-resistant rickets in three families. *Proc. Natl. Acad. Sci. USA.* 86:9783-9787.

CHAPTER THREE:

STRUCTURAL ORGANIZATION OF THE HUMAN α-*N*-ACETYLGALACTOSAMINIDASE GENE: CONSERVATION OF GENOMIC STRUCTURE WITH HUMAN α-GALACTOSIDASE A

ABSTRACT

Human α-*N*-acetylgalactosaminidase (α-GalNAc; E.C. 3:2.1.49), the lysosomal glycohydrolase that cleaves α-*N*-acetylgalactosaminyl moieties from glycoconjugates, is encoded by a gene localized to chromosome 22q13→ter. The deficient activity of this enzyme results in Schindler disease, an autosomal recessive disorder characterized by the increased urinary excretion of glycopeptides and oligosaccharides containing α-*N*-acetylgalactosaminyl moieties. The cDNA sequence encoding α-GalNAc has been isolated and its characterization revealed significant nucleotide and amino acid homology (55.8% and 46.9%, respectively) to α-galactosidase A (α-Gal A), a lysosomal hydrolase localized to Xq21.33→22. To determine the structural organization of the α-GalNAc gene and to compare its structural features to that of α-Gal A, the α-GalNAc chromosomal gene was isolated from a genomic cosmid library using the radiolabelled full-length α-GalNAc cDNA. One clone, gAGB-1, which contained the entire α-GalNAc gene, was characterized by restriction mapping and sequencing. The α-GalNAc gene was ~14 kb and contained nine exons. All exon/intron junctions conformed to the GT/AG rule. Analysis of 1432 bp of 5' flanking sequence revealed two Sp1 and three GC-box promotor elements in this lysosomal housekeeping gene. The region was GC-rich (56%), but no HTF-island (Hpa II tiny fragments) was identified. Six *Alu*-repetitive elements were identified and all were in the reverse orientation. Comparison of the α-GalNAc gene with the α-Gal A gene revealed homologous exonic placement of α-GalNAc introns 2 through 7 with α-Gal A introns 1 through 6. Two additional introns, 1 and 8, were identified in the α-GalNAc gene. Predicted amino acid homology among α-GalNAc exons 2-7 with α-Gal A exons 1-6 ranged from 46.2 % to 62.7% with a few short gaps. In contrast, there was little, if any, similarity between the remaining coding sequence (α-Gal A exon 7 and α-GalNAc exons 8 and 9) which had only 15.8% homology with numerous gaps. In addition, no similarities were found between α-

GalNAc and α-Gal A intronic or flanking sequences. The high exonic homologies, in addition to homologous intron placement, suggest that these genes are evolutionarily related and arose through duplication and divergence from a common ancestral gene.

INTRODUCTION

Human α-*N*-acetylgalactosaminidase (E.C. 3.2.1.29; α-GalNAc) is a lysosomal glycohydrolase that cleaves glycoconjugates with α-*N*-acetylgalactosaminyl moieties (1-3). The human enzyme, encoded by a gene mapped to chromosome 22q13→ter, has been purified to homogeneity from various sources, and its physical and kinetic properties have been determined (1-6). The subunits of this homodimeric enzyme are synthesized as 65 kDa precursor glycopeptides, and following modifications in the Golgi, the mature 48 kDa subunits dimerize to form the active enzyme of 117 kDa (7).

The deficient activity of α-GalNAc results in Schindler disease, an autosomal recessive disorder characterized by the increased urinary excretion of glycopeptides and oligosaccharides containing α-*N*-acetylgalactosaminyl moieties (8). Two subtypes of this enzymatic deficiency have been identified: Type I disease has been described in two children who have the phenotype of an infantile neuroaxonal dystrophy (9), whereas Type II disease occurred in an adult who is neurologically asymptomatic, but has angiokeratoma corporis diffusum (10, 11). The patients with Type I or II disease have less than 1% of normal enzymatic activity and no immunoreactive enzyme protein in fibroblast extracts. Subsequent delineation and identification of the molecular lesions in these two subtypes revealed that each was due to a point mutation that resulted in an amino acid substitution (11, 12). However, the physiologic basis for the remarkable clinical heterogeneity of the two distinct subtypes in unknown.

Recently, we reported the isolation and nucleotide sequence of a cDNA encoding human α-GalNAc (13). The predicted amino acid sequence of the cDNA had significant homology with that of human α-galactosidase A (α-Gal A) (13), suggesting the evolutionary relatedness of these two genes. In this communication, we report the isolation, characterization and sequence of the human α-GalNAc chromosomal gene. The structure, 5' regulatory elements,

3' flanking sequence, repetitive elements, and structural similarities with the human α-Gal A gene are described.

EXPERIMENTAL PROCEDURES

Construction of Synthetic Oligonucleotide Probes. Unique oligonucleotides for sequencing the human α-GalNAc gene and for use as polymerase chain reaction (PCR) primers were synthesized on an Applied Biosystems Model 380B.

Isolation of Genomic Clones. To isolate genomic clones containing the entire α-GalNAc sequence, $1x10^6$ recombinants from a human genomic cosmid library were screened by colony hybridization with the full-length cDNA, pAGB-3, as radiolabelled probe (13). The genomic library was prepared from size-selected human lymphoblast DNA and kindly provided by Dr. Henrik Vissing, Mount Sinai School of Medicine.

Characterization of Genomic Clones. Cosmid DNA was isolated from purified positive clones, digested with various restriction endonucleases, separated by agarose gel electrophoresis, transferred to nylon membranes and then probed with end-labelled oligonucleotides corresponding to the pAGB-3 cDNA sequence to identify exonic sequences (14). Fragments which contained the entire genomic sequence were subcloned into the pGEM4Z vector (Promega, Madison, WI). Approximate intronic sizes were determined from the PCR products generated by amplification (15) of selected regions of the cosmid DNA with the appropriate flanking exonic primers. The GeneAmp DNA Amplification Reagent Kit (Perkin Elmer Cetus, Norwalk, CT) was used to amplify 1 μg of cosmid DNA using 1μM of each primer. Each of 30 PCR cycles consisted of 1 min denaturation at 94 °C, 2 min annealing at 37 °C; and 7 min extension at 60 °C. An aliquot (10-20 μl) was analyzed by agarose gel electrophoresis using *Hind* III digested lambda and *Hae* III digested ϕX174 DNAs as size standards.

DNA Sequencing and Computer-Assisted Analyses. All DNA sequencing reactions were carried out by the dideoxy method (16) using universal or α-GalNAc-specific synthetic oligonucleotide primers in both orientations using Sequenase (U.S. Biochemical Corp, Cleveland,

OH). Searches for nucleotide and amino acid sequence similarities were carried out with the Microgenie DNA analysis program (Beckman, Fullerton, CA).

RESULTS AND DISCUSSION

Isolation and Restriction Mapping of Genomic Clones. A total human genomic DNA cosmid library was screened with the nick-translated pAGB-3 cDNA insert. Of $1x10^6$ independent recombinants screened, six putative positive clones were isolated, purified, and subjected to Southern hybridization with end-labelled oligonucleotides corresponding to α-GalNAc cDNA sequences to identify exonic fragments. One clone, gAGB-1, which had an ~35 kb insert and appeared to contain the entire genomic sequence, was selected for further analysis. The entire α-GalNAc genomic sequence occurred in a ~14 kb *EcoR* I fragment which was subcloned into the pGEM4Z vector and designated gAGB-1E.

Intron-Exon Boundaries. Since the α-GalNAc and α-Gal A cDNA sequences were highly homologous and suggested that the two genes may have evolved from a common ancestral sequence (13), it was expected that the genomic structure of the two genes would be conserved. Thus, initial efforts were directed to determine the approximate size and location of α-GalNAc introns. Exonic oligonucleotides corresponding in position to the α-Gal A intron/exon junctions (14, 17) were used to PCR-amplify the homologously located α-GalNAc intron/exon regions. Sequencing of the PCR products revealed that all six corresponding α-Gal A intron/exon junctions were conserved in the α-GalNAc gene. This was analogous to the previously demonstrated conservation of the gene structures for the β-hexosaminidase α- and β-chain genes, where 12 of 13 introns in the β-chain interrupted the gene at homologous positions in the α-chain (18). Additional sequencing of exonic regions revealed the presence of two additional introns interrupting the α-GalNAc gene, one 5' and one 3' to the six previously identified introns. To further characterize the α-GalNAc gene, the entire 14 kb gAGB-1E fragment was sequenced in both orientations (Fig 1). The nine α-GalNAc exons

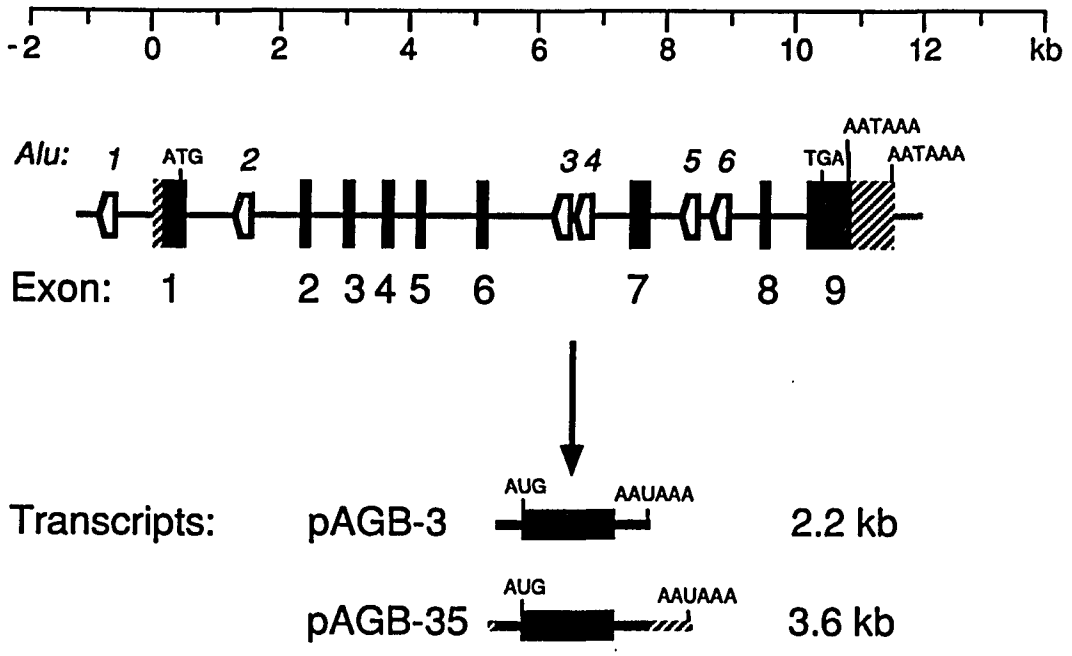


Figure 1. Organization of the α-GalNAc gene.

ranged in length from 95 to 649 bp, with exon 1 containing the 347 nt 5' untranslated region and the sequence encoding the first six amino acids of the 17-residue signal peptide. The eight introns ranged in length from 0.3 to 2.7 kb. All splice junctions followed the GT/AG rule (19) and were consistent with the 5' and 3' consensus sequences for splice junctions of RNA polymerase II (20). Putative lariat branch points were identified between -40 and -18 nt from the 3' splice junction for all eight introns by similarity to the less well conserved consensus sequence (C/T)N(C/T)T(A/G)A(A/C/T) (21). All three codon phases were observed at the exon junctions in the α-GalNAc gene (Table 1). The α-GalNAc genomic sequences also confirmed at the genomic level the presence of two polyadenylation hexamers of AATAAA whose alternative use results in the two α-GalNAc transcripts of 2.2 and 3.6 kb (13) (Fig. 1).

Regulatory Elements. Computer-assisted analysis of the α-GalNAc 1432 bp 5' flanking region identified two Sp1 binding sites at -26 and -258 in sense and antisense orientations, respectively, and three GC boxes (at -23, -68, and -257) from the transcription start site (22) (Fig. 2). No CAAT or TATA-like elements were found. The 5' flanking region was GC-rich (~56%) but there was not an HTF-island. The CpG/GpC ratio was 0.625, whereas an HTF-island has a ratio greater than 1.0 (23, 24). The absence of a TATA box and the high GC content are known features in the promotor region of other housekeeping genes (25). Imperfect inverted repeats were identified at nt -108 to -127 and at -27 to -17 which could pair with a predicted ΔG of -34.8 kCal/mol (Fig. 2). The structure predicted by this pairing may be important to promotor function. Further evaluation of these putative promotor elements will require footprinting and functional assays with deletion/mutation mapping.

Alu-Repetitive Elements. Six *Alu*-repetitive elements (26) were identified in the 14 kb sequence: one each in the 5' flanking region and intron 1, and two in introns 6 and 7 (Fig. 3).

Table 1. Nucleotide sequence of the intron/exon boundaries in the human α-GalNAc gene.

Exon number and (size)	cDNA position of exon	5' Splice donor	Intron number and (size)	3' Splice acceptor	Codon phase
(nt)			(kb)	5' ttggtGACCTTAAG	
1 (363)	-347- 16	GAAGACAG gtacct	1 (1.7)	gggggcaactgatagtcaccatcccctccgtccgttgcccggtag TGCTCTTG	?
2 (136)	17-152	TGCATAAG gtgagt	2 (0.5)	cgggctgggtcccctgagctaccccatccctctgcccctcactag TGAACAGC	II
3 (172)	153-324	CTGACTAC gtgagc	3 (0.5)	ccttgagggtcctgcttgacccactgtgtcctcaaaccccacag GTTCACTC	0
4 (178)	325-502	GGCCCAGG gtgagt	4 (0.3)	cctcctgcacttcctcaggactgaaacaggttgggctttctctag GGTACCCC	I
5 (95)	503-597	CCCCCAAG gtaagc	5 (0.8)	gcgtgtgattctggtgctggactctgcctttccctcgacatccag GTGAACTA	0
6 (162)	598-759	CTGACATG gtacca	6 (2.7)	tccttgaggagatgagagctgactactcctcttctctgcccccag CTGCTCAT	0
7 (198)	760-957	TTCACAAG gtacta	7 (1.7)	tgtctgagcaagggaccccagccagtagccaccttctgtgcccag GAAAAATC	0
8 (144)	958-1101	TATATGAG gtagag	8 (0.5)	atggtggcattggtgatgtccctccccactgcag GCCCAGGA	0
9 (649)	1102-1750	TTAATCAAgtagtt 3'			

Consensus sequences: donor: $^{A}_{C}AG\ gt^{a}_{g}agt$ lariat: $^{c}_{t}n^{c}_{t}t^{a}_{g}a^{c}_{t}$ acceptor: $^{cccccccccc}_{ttttttttttt}ncag$

Exon sequences are in upper case letters; intron sequences in lower case. Donor and acceptor consensus sequences as well as those for the lariat branch point are indicated. The putative lariat branch points are underlined in the 3' splice acceptor sequences. Codon phase O intron/exon junctions occur between codons, whereas phases I and II interrupt codons after the first and second nucleotides, respectively.

Figure 2. The nucleotide sequence of the α-GalNAc gene 5' flanking region. The 1532 nt sequence was determined in both orientations from genomic clone, gAGB-1. The sequence includes 1432 nt of 5' flanking sequence and 100 nt of 5' untranslated sequence from exon 1. The G at nt +1 indicates an transcriptional initiation site as determined by primer extension (13). Sp1 binding sites (Sp1) are underlined with wavy arrows. A sequence corresponding to a GC box is boxed. An imperfect inverted repeat (IR) is underlined. An *Alu*-repetitive element from nt -958- -1234 is italicized.

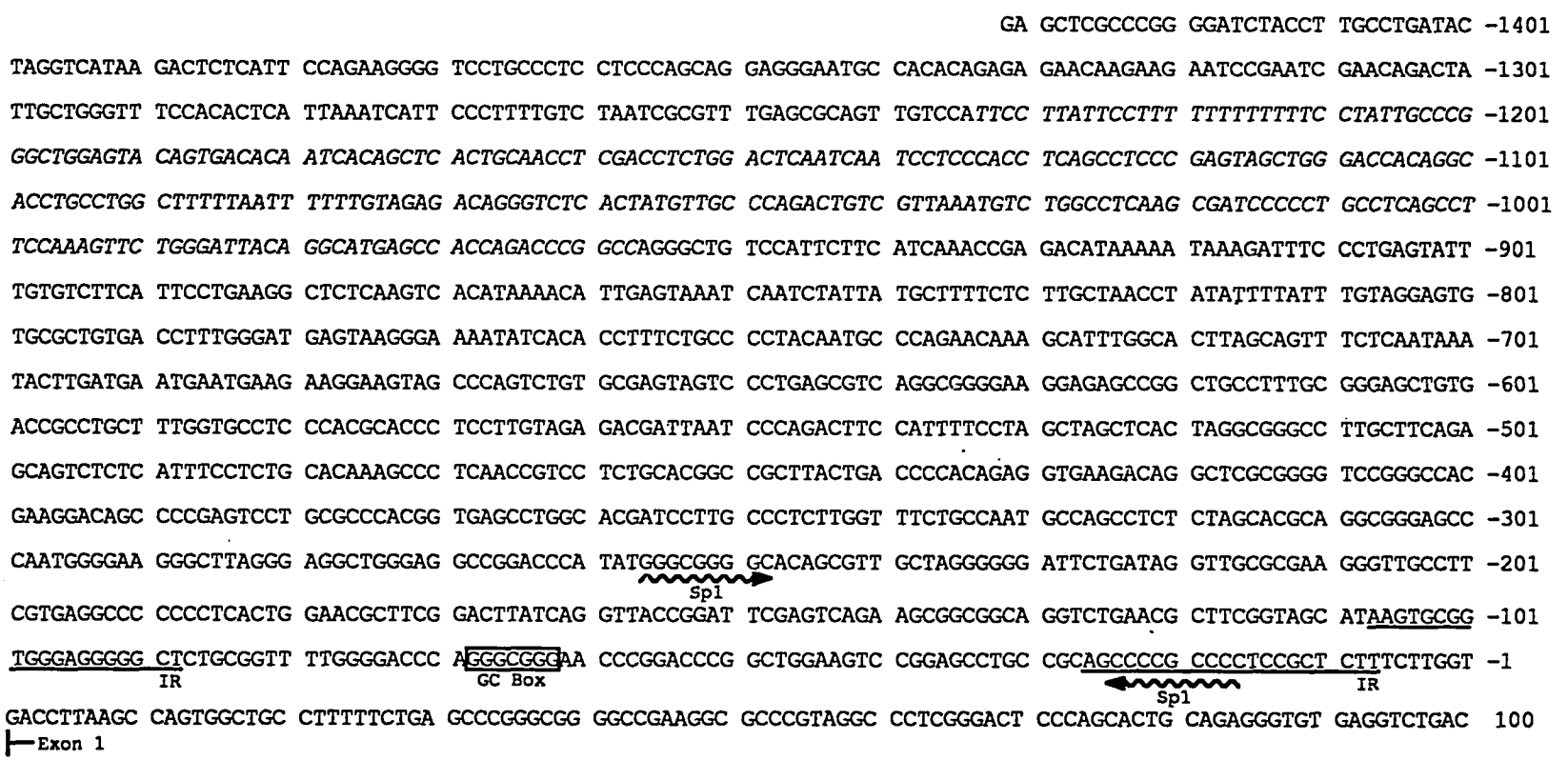
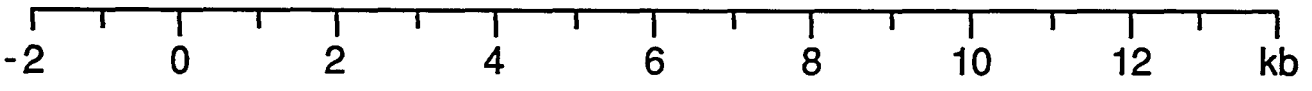
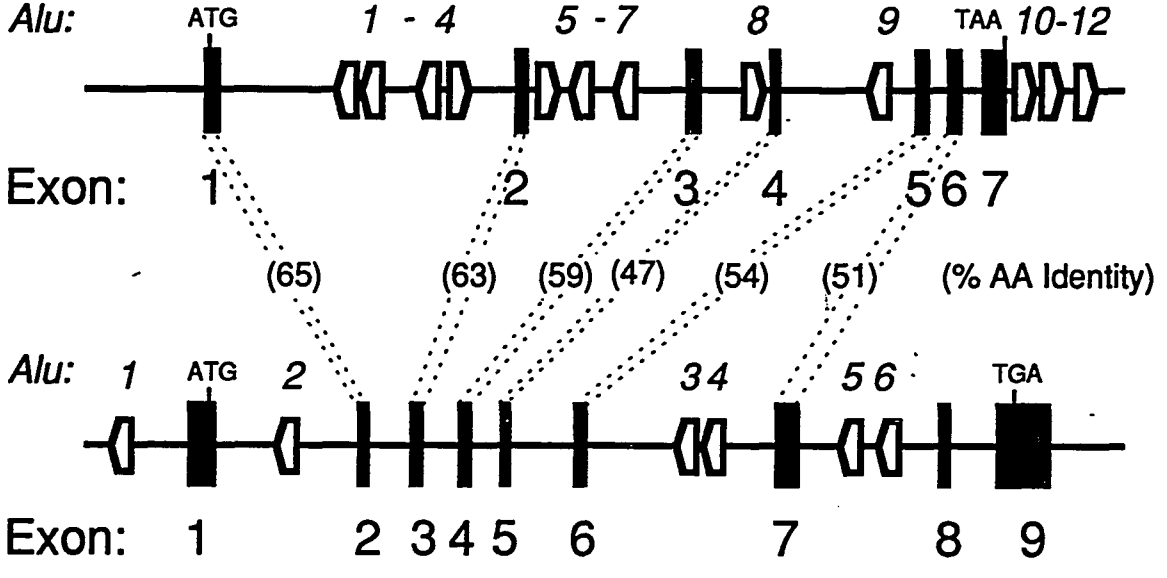

-1432 GA GCTCGCCCGG GGATCTACCT TGCCTGATAC -1401
-1400 TAGGTCATAA GACTCTCATT CCAGAAGGGG TCCTGCCCTC CTCCCAGCAG GAGGGAATGC CACACAGAGA GAACAAGAAG AATCCGAATC GAACAGACTA -1301
-1300 TTGCTGGGTT TCCACACTCA TTAAATCATT CCCTTTTGTC TAATCGCGTT TGAGCGCAGT TGTCCATTCC TTATTCCTTT TTTTTTTTTC CTATTGCCCG -1201
-1200 GGCTGGAGTA CAGTGACACA ATCACAGCTC ACTGCAACCT CGACCTCTGG ACTCAATCAA TCCTCCCACC TCAGCCTCCC GAGTAGCTGG GACCACAGGC -1101
-1100 ACCTGCCTGG CTTTTTAATT TTTTGTAGAG ACAGGGTCTC ACTATGTTGC CCAGACTGTC GTTAAATGTC TGGCCTCAAG CGATCCCCCT GCCTCAGCCT -1001
-1000 TCCAAAGTTC TGGGATTACA GGCATGAGCC ACCAGACCCG GCCAGGGCTG TCCATTCTTC ATCAAACCGA GACATAAAAA TAAAGATTTC CCTGAGTATT -901
-900 TGTGTCTTCA TTCCTGAAGG CTCTCAAGTC ACATAAAACA TTGAGTAAAT CAATCTATTA TGCTTTTCTC TTGCTAACCT ATATTTTATT TGTAGGAGTG -801
-800 TGCGCTGTGA CCTTTGGGAT GAGTAAGGGA AAATATCACA CCTTTCTGCC CCTACAATGC CCAGAACAAA GCATTTGGCA CTTAGCAGTT TCTCAATAAA -701
-700 TACTTGATGA ATGAATGAAG AAGGAAGTAG CCCAGTCTGT GCGAGTAGTC CCTGAGCGTC AGGCGGGGAA GGAGAGCCGG CTGCCTTTGC GGGAGCTGTG -601
-600 ACCGCCTGCT TTGGTGCCTC CCACGCACCC TCCTTGTAGA GACGATTAAT CCCAGACTTC CATTTTCCTA GCTAGCTCAC TAGGCGGGCC TTGCTTCAGA -501
-500 GCAGTCTCTC ATTTCCTCTG CACAAAGCCC TCAACCGTCC TCTGCACGGC CGCTTACTGA CCCCACAGAG GTGAAGACAG GCTCGCGGGG TCCGGGCCAC -401
-400 GAAGGACAGC CCCGAGTCCT GCGCCCACGG TGAGCCTGGC ACGATCCTTG CCCTCTTGGT TTCTGCCAAT GCCAGCCTCT CTAGCACGCA GGCGGGAGCC -301
-300 CAATGGGGAA GGGCTTAGGG AGGCTGGGAG GCCGGACCCA TATGGGCGGG GCACAGCGTT GCTAGGGGGG ATTCTGATAG GTTGCGCGAA GGGTTGCCTT -201
Sp1
-200 CGTGAGGCCC CCCCTCACTG GAACGCTTCG GACTTATCAG GTTACCGGAT TCGAGTCAGA AGCGGCGGCA GGTCTGAACG CTTCGGTAGC ATAAGTGCGG -101
-100 TGGGAGGGGG CTCTGCGGTT TTGGGGACCC AGGGCGGGAA CCCGGACCCG GCTGGAAGTC CGGAGCCTGC CGCAGCCCCG CCCCTCCGCT CTTTCTTGGT -1
IR GC Box Sp1 IR
1 GACCTTAAGC CAGTGGCTGC CTTTTTCTGA GCCCGGGCGG GGCCGAAGGC GCCCGTAGGC CCTCGGGACT CCCAGCACTG CAGAGGGTGT GAGGTCTGAC 100
Exon 1

Figure 3. Comparison of the α-Gal A and α-GalNAc structural genes. Exons are denoted by solid rectangles. The positions and orientations of the *Alu* repeats are indicated. The homology between α-Gal A exons 1-6 and α-GalNAc exons 2-7 are denoted by dashed lines and the amino acid identity indicated in parentheses.



α-GALACTOSIDASE A GENE



α-N-ACETYLGALACTOSAMINIDASE GENE

All six were in the reverse orientation. The percent homology of these repetitive elements with the consensus 300 nt sequence ranged from 54.1 to 87.4%.

Structural Homology of the α-GalNAc and α-Gal A Genes; Potential Evolutionary Relationships. Sequencing of the α-GalNAc structural gene revealed that six of the eight α-GalNAc introns (2 through 7) interrupted the gene at positions that were identical to those interrupting the α-Gal A coding sequence. Characterization of the exonic sequences encoding these two functionally related genes revealed 55.8% and 46.9% overall homology at the nucleotide and deduced amino acid levels. Analysis of the amino acid identity between homologous exons demonstrated 46.2% to 62.7% identity with only a few gaps between α-GalNAc exons 2 through 7 and α-Gal A exons 1 through 6 (Table 2). In contrast, the homology between α-GalNAc exons 8 and 9 with α-Gal A exon 7 dropped significantly to 15.8% with multiple gaps. The homologous intron placement and exonic amino acid identity suggested that the α-GalNAc and α-Gal A genes had a common origin, presumably they were duplicated from the same ancestral gene and underwent subsequent divergence to acquire their specific functions (18, 27).

A possible mechanism for the duplication of these genes is a recombinational event, perhaps involving *Alu*-repetitive elements. Although the α-GalNAc and α-Gal A genes have five and 12 *Alu*-repetitive elements, respectively, *Alu-Alu* mediated recombination probably was not the mechanism for duplication since the *Alu* sequences were not homologously placed in these genes, nor was there any homology in intron size or sequence or flanking sequences (Table 2) of α-GalNAc and α-Gal A. Comparison of the deduced amino acid sequences in α-galactosidase cDNAs from human, mouse, yeast, and E. coli, and from the human α-GalNAc cDNA (13, 28), suggested that α-GalNAc exons 2 though 7 were conserved due to functional

Table 2. Comparison of the exonic and intronic homologies between α-GalNAc and α-Gal A.

α-GalNAc exon number and (size)	α-Gal A exon number and (size)	% Amino acid identity	α-GalNAc intron number and (size)	α-Gal A intron number and (size)
(nt)	(nt)		(kb)	(kb)
1 (363)	–	–	1 (1.7)	–
2 (136)	1 (254)	46.2	2 (0.5)	1 (3.8)
3 (172)	2 (175)	62.7	3 (0.5)	2 (2.0)
4 (178)	3 (178)	59.0	4 (0.3)	3 (1.5)
5 (95)	4 (92)	46.9	5 (0.8)	4 (1.2)
6 (162)	5 (162)	53.7	6 (2.7)	5 (0.2)
7 (198)	6 (198)	51.3	7 (1.7)	6 (0.4)
8 (144)	7 (291)	15.8	8 (0.8)	–
9 (649)				

domains common to these enzymes. In support of this concept are regions of up to 15 consecutive identical residues and stretches of 21 of amino acids with 20 identical residues in the human α-GalNAc and α-Gal A polypeptides. These domains may have been under strict functional constraints, while other domains (α-GalNAc exons 8 and 9, and α-Gal A exon 7) may have more freely diverged to generate unique functional specificities such as *N*-acetylgalactosamine recognition. Because the homology between α-GalNAc and α-Gal A falls sharply after α-GalNAc exon 7, it is intriguing to speculate on the origin of α-GalNAc exons 8 and 9. When the predicted amino acid sequence encoded by α-GalNAc exons 8 and 9 were used to search protein sequences (13), the only homology identified was a region of seven α-GalNAc residues (365-371) which was homologous to residues 194-200 in the β-hexosaminidase α-chain, a subunit of the human lysosomal β-*N*-acetylgalactosaminidase (18). Thus, the functional specificity for a possible *N*-acetylgalactosaminyl recognition, binding, or catalytic site may have arisen by "exon shuffling" (2, 30). This concept would require that both α-GalNac exons 8 and 9 were shuffled together since α-GalNAc intron 8 interrupted the gene between the nucleotides encoding residues 367 and 368.

Speculations regarding the divergence time and conserved structural domains between α-GalNAc and α-Gal A will require α-Gal A and α-GalNAc sequences from other species. The isolation and characterization of these genes will allow the calculation of their divergence times. X-ray crystallographic studies of purified recombinant α-Gal A and α-GalNAc should lend insight into the conserved structural domains between the two lysosomal glycohydrolases.

REFERENCES

1. Dean. K. J., Sung, S. and Sweeley, C. C. (1977) *Biochem. Biophys. Res. Comm.* **77**, 1411-1417

2. Schram, A. W., Hamers, M. N. and Tager, J. M. (1977) *Biochim. Biophys. Acta.* **482**, 138-144

3. Bishop, D. F., Dean, K. J., Sweeley, C. C. and Desnick, R. J. (1980) in *Enzyme Therapy in Genetic Disease:2*, (Desnick, R. J., ed.), pp. 17-32. Alan R. Liss, Inc. New York

4. Beutler, E. and Kuhl, W. (1972) *J. Biol. Chem.* **247**, 7195-7200

5. Callahan, J. W., Lasilla, E. L., DenTandt, W. and Philippart, M. (1973) *Biochem. Med.* **7**, 424-431

6. Kusiak, J. W., Quirk, J. M. and Brady, R. O. (1978) *J. Biol. Chem.* **253**, 184-190

7. Sweeley, C. C., LeDonne, N. C. and Robbins, P. W. (1983) *Arch. Biochim. Biophys.* **223**, 158-165

8. van Diggelen, O. P., Schindler, D., Willemson, R., Boer, M., Kleijer, W. J., Huijmans, J. G. H., Blom, W. and Galjaard, H. (1988) *J. Inher. Met. Dis.* **11**, 349-357

9. Schindler, D., Bishop, D. F., Wolfe, D. E., Wang, A. M., Egge, H., Lemieux, R. U. and Desnick, R. J. (1989) *N. Engl. J. Med.* **320**, 1735-1740

10. Kanzaki, T., Yokota, M., Mizuno, N., Matsumoto, Y. and Hirabayashi, Y. (1989) *Lancet* April 22, 1989, 875-877

11. Wang, A. M., Kanzaki, T., Schindler, D. and Desnick, R. J. (1989) *Am. J. Hum. Genet.* **43**, 4 A228

12. Wang, A. M. and Desnick, R. J. (in review) *J. Clin. Invest.*

13. Wang, A. M., Bishop, D. F. and Desnick, R. J. (in review) *J. Biol. Chem.*

14. Bishop, D. F., Kornreich, R. and Desnick, R. J. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 3903-3907

15. Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. and Erlich, H. A. (1987) *Science* **239**, 487-491

16. Sanger, F., Nicklen, S. and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463-5467

17. Kornreich, R., Desnick, R. J. and Bishop, D. F. (1989) *Nucleic Acids Res.* **14**, 3301-3302

18. Proia, R. L. (1988) *Proc. Natl. Acad Sci. USA* **85**, 1883-1887

19. Breathnach, R. and Chambon, P. (1981) *Ann. Rev. Biochem.* **50**, 349-383

20. Mount, S. M. (1982) *Nucleic Acids Res.* **10**, 459-472

21. Reed, R. and Maniatis, T. (1988) *Genes Dev.* **2**, 1268-1276

22. Dynan, W. S. and Tijan, R. (1985) *Nature* **316**, 774-778

23. Bird, A. P. (1986) *Nature* **321**, 209-213

24. Lindsay, S. and Bird, A. P. (1987) *Nature* **327**, 336-338

25. Dynan, W. S. (1986) *Trends Genet.* **2**, 196-197

26. Deininger, P. L., Jolly, D. J., Rubin, C. M., Friedmann, T. and Schmid, C. W. (1981) *J. Mol. Biol.* **151**, 17-33

27. Maeda, N. and Smithies, O. (1986) *Ann. Rev. Genet.* **20**, 81-108

28. Unpublished results

29. Gilbert, W. (1978) *Nature* **271**, 501

30. Michelson, A. M., Blake, C. C., Evans, S. T. and Orkin, S. H. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 6965-6969

CONCLUDING REMARKS

α-*N*-acetylgalactosaminidase (α-GalNAc) is representative of a class of enzymes present in lysosomes. Genetic deficiencies of specific lysosomal hydrolases result in different lysosomal storage diseases. To date, over 30 lysosomal storage disorders have been described in man, and have sparked wide interest in lysosomal cell biology and the biochemistry of lysosomal enzymes. Biochemical and processing studies of α-GalNAc and other lysosomal hydrolases have now well defined the proteolytic, carbohydrate, and phosphorylation modifications necessary for the appropriate trafficking of these enzymes to the lysosomes. The deficient activity of α-GalNAc results in Schindler disease. Two distinct clinical subtypes of this enzymatic deficiency have been described: Type I is characterized by a clinical course identical to infantile neuroaxonal dystrophy, whereas Type II is characterized by disseminated angiokeratoma corporis diffusum with no neurological involvement.

The revolution in molecular biology has permitted the study of lysosomal enzymes and their deficiencies at the molecular level. Approximately 20 cDNA sequences encoding lysosomal proteins have been isolated. As a result, a myriad of mutations causing specific lysosomal disorders have been identified and characterized and determined to be heterogeneous. The studies described here provide the molecular basis for the further study of human α-GalNAc. The similarities of α-GalNAc with α-galactosidase A (α-Gal A) should provide insight into the potential structural and functional domains as well as into their evolutionary relationships. As the sequences encoding α-GalNAc and α-Gal A from other species are isolated, it should be possible to determine the divergence time of these enzymes. Furthermore, X-ray crystallographic studies on recombinant α-GalNAc and α-Gal A should provide information on the conserved structural and functional domains of these proteins, as well as potential differences in domains responsible for functional specificity. Schindler disease has

provided an interesting study of α-GalNAc deficiency. While the mutations and their resulting amino acid substitutions in each subtype have been identified and expressed, the physiological basis for the remarkable clinical heterogeneity between the two subtypes is not understood. However, the characterization of the α-GalNAc cDNA sequence and knowledge of the specific base substitution in each Schindler disease subtype should permit further delineation of the enzyme's function, stability, and potential specificity and protein interactions.

In conclusion, these studies have provided the framework for the further study and understanding of the properties of human α-GalNAc and the pathophysiology of its deficiency.