J. Biochem. **134**, 675–682 (2003)
DOI: 10.1093/jb/mvg192

Molecular Cloning of the Human β1,4 *N*-Acetylgalactosaminyltransferase Responsible for the Biosynthesis of the Sd^a Histo-Blood Group Antigen: The Sequence Predicts a Very Long Cytoplasmic Domain

Libera Lo Presti, Erik Cabuy*, Mariella Chiricolo and Fabio Dall'Olio†
Dipartimento di Patologia Sperimentale, Università di Bologna, Via S. Giacomo 14, 40126, Bologna, Italy

Received July 10, 2003; accepted August 25, 2003

The Sd^a antigen is a carbohydrate determinant expressed on erythrocytes, the colonic mucosa and other tissues. This epitope, whose structure is Siaα2,3[GalNAcβ1,4]Gal β1,4GlcNAc, is synthesized by a β1,4 *N*-acetylgalactosaminyltransferase (β4GalNAc-T) that transfers a β1,4-linked GalNAc to the galactose residue of an α2,3-sialylated chain. We have cloned from human colon carcinoma Caco2 cells a cDNA whose transfection in COS cells induces a GalNAc-T active on sialylated but not on asialylated fetuin and putatively represents the human Sd^a β4GalNAc-T. The cDNA predicts a 566 aa protein showing 66.6% and 39% identity with mouse CT β4GalNAc-T and human GM2/GD2 synthase, respectively, with a typical type II glycosyltransferase organization, no potential *N*-glycosylation sites and a 67 aa cytoplasmic tail, which is probably the longest among the glycosyltransferases cloned to date. The gene maps in chromosome 17q23, and is composed of at least 11 exons. Exons 2–11 are homologous to exons 2–11 of the previously cloned CT β4GalNAc-T from murine cytotoxic T lymphocytes while exons 1 of the two enzymes are totally different. The mRNA is expressed at a high level in differentiated Caco2 cells and in colonic mucosa and at a much lower level in lymphocytes and other colon cancer cell lines.

Key words: β1,4 *N*-acetylgalactosaminyltransferase, Caco2 cells, colon cancer, CT antigen, glycosyltransferases, Sd^a antigen.

Abbreviations: FTA, phosphotungstic acid; Gal, galactose; GalNAc, *N*-acetylgalactosamine; GlcNAc, *N*-acetylglucosamine; PBL, peripheral blood lymphocytes; ORF, open reading frame; RACE, rapid amplification of cDNA ends; RT-PCR, reverse transcriptase polymerase chain reaction; Sia, sialic acid; UTR, untranslated region.

The surface of animal cells is decorated with complex carbohydrate structures that, in some cases, behave as antigens. More than 90% of individuals of Caucasian origin express the carbohydrate antigen Sd^a on their erythrocytes and in a few other tissues, including colon and kidney (*1*). The structure of the Sd^a epitope expressed by the *N*-linked chains of urinary Tamm-Horsfall glycoprotein was elucidated in 1983 (*2*), and turned out to comprise an *N*-acetylgalactosamine, a sugar rarely present in the *N*-linked chains of glycoproteins, β1,4-linked to the galactose residue of an α2,3-sialylated type 2 chain (Siaα2,3Galβ1,4GlcNAc-R), giving rise to the following tetrasaccharide: Siaα2,3(GalNAcβ1,4)Galβ1,4GlcNAc-R (where Sia is sialic acid, GalNAc is *N*-acetylgalactosamine, Gal is galactose and GlcNAc is *N*-acetylglucosamine). The α2,3-sialylated type 2 chains are very frequently expressed by glycoproteins and glycolipids, while the presence of the β1,4-linked GalNAc is peculiar. Thus, the enzyme responsible for the biosynthesis of the Sd^a determinant is a β1,4 *N*-acetylgalactosaminyltransferase (β4GalNAc-T), which was first identified in guinea pig kidney (*3*). The oligosaccharide chains on which β4GalNAc-T act are virtually identical to the oligosaccharide portion of ganglioside GM3 (Siaα2,3Galβ1,4Glc). However, β4GalNAc-T is distinct from GM2/GD2 synthase in that it is completely inactive towards GM3 ganglioside (*4*, *5*). β4GalNAc-T shows a narrow tissue distribution (*6*), being expressed mainly by guinea pig (*5*) and human kidney (*4*), subpopulations of mouse cytotoxic T lymphocytes (*7*, *8*), and colonic mucosa cells of different species (*9–11*). The expression of β4GalNAc-T in colon shows a clear relationship with tissue differentiation. In fact, β4GalNAc-T activity dramatically decreases in human colon carcinoma (*12*), whereas in rat colonic epithelium, it develops only after weaning (*9*). According with the notion of a strict dependence of β4GalNAc-T on colon tissue differentiation, β4GalNAc-T activity could not be detected in a panel of poorly differentiated human colon carcinoma cell lines (*12*). However, in the colon cancer cell line Caco2, which shows a more differentiated phenotype and which can further differentiate *in vitro*, low β4GalNAc-T activity that increases upon differentiation can be detected (*13*).

Murine cytotoxic lymphocytes express carbohydrate antigens (CT antigens) involved in lytic function, whose structures are identical to the Sd^a determinant (*14*). In previous years, the sequence of a cDNA clone from murine cytotoxic lymphocytes encoding a β4GalNAc-T able to induce the biosynthesis of CT antigens (*15*) and a sequence encoding part of the putative catalytic domain

*Present address: Genetic Cancer Susceptibility Unit, IARC, 150 Cours Albert Thomas, 69372 Lyon cedex 08, France
†To whom correspondence should be addressed. Fax: +39-05-1209-4746, E-mail: dallolio@alma.unibo.it

Table 1. **PCR primers.**

Primer name	Position	Sequence (5′-3′)	Annealing temperature used (°C)
GeneRacer*		CGACTGGAGCACGAGGACACTGA	69
L.1**	Exon 7	ATCCGCCATCCTGTCATA	60
R.1	Exon 9	CACCAGCACCTCAATCTTG	60
L.5	5′-UTR	ACGAACTCTGCACCCCCAGGAAT	63
L.13	3′-UTR	GGTCATATCCAATTAATGTCCCCTGG	62
L.14	3′-UTR	GACATTGTACAGGGGTGAGGGAGTG	62
L.16	Exon 1	CACCATGGGGAGCGCTGGCTTTTCC	65
R.7	Exon 6	CCCTGCACCACACTGTCT	53
R.8	Exon 6	CCCCAGAGAAGCTGTCAGGGTGA	69
R.10	3′-UTR	CCAGTAACTGAGCCATTTCCCTTTTCC	63
R.19	3′-UTR	CCAACATTTTCCCCTTGGAACTCTC	62
R.21	3′-UTR	TGTGAACTGCAACCTTACAAGAAGGA	62
ACTL.3		GGCATCGTGATGGACTCCG	60
ACTR.3		GCTGGAAGGTGGACAGCGA	60

*GeneRacer primer is a Trade Mark of Invitrogen. **The letters "L" and "R" denote forward and reverse primers, respectively. Primers ACTL.3 and ACTR.3 are for the amplification of the β-actin transcript.

of the human β4GalNAc-T (*16*) were reported. In this paper, we report the molecular cloning from differentiated Caco2 cells of the full coding sequence of human Sd^a β4GalNAc-T. Although highly homologous to the enzyme from mouse lymphocytes, the sequence of the human β4GalNAc-T from Caco2 cells predicts a putative cytoplasmic tail that is probably the longest among the glycosyltransferases cloned to date and is totally different from that of the mouse enzyme.

MATERIALS AND METHODS

Cell Culture and mRNA Preparation—Caco2 and COS-7 cells were grown in DMEM (GIBCO, Paisley, UK) containing 100 U ml^{-1} penicillin, 100 mg ml^{-1} streptomycin and 15% or 10% FCS, respectively. Caco2 cells were harvested either just after reaching confluency (non differentiated state) or after three weeks of postconfluent culture (differentiated state); their differentiation state was evaluated as previously described (*13*, *17*). Cell line COLO205 was grown in RPMI 1640, cell lines SW480, SW620, SW1417, SW48, SW948, SW948FL were grown in Leibovitz's L-15 medium, and LoVo cells were grown in Ham F10 medium. Cells were harvested during log phase of growth and frozen at –80°C. Total RNA was prepared by the RNAzol (Biotecx Laboratories, Houston, TX) method and mRNA was prepared by oligo-dT chromatography.

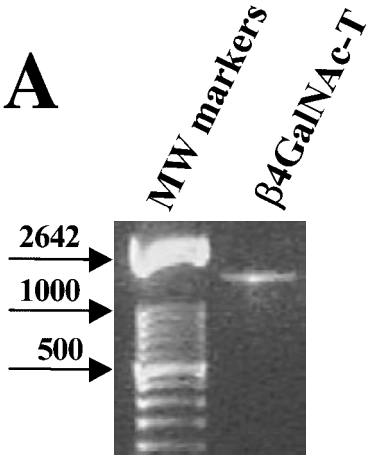
RT-PCR Analysis—Five micrograms of total RNA were reverse-transcribed using a TaKaRa RT-PCR kit, version 2.1 (TaKaRa Shouzo) according to the manufacturer's instructions using random 9-mers as primers in a final volume of 20 μl. Two microliters of the RT reaction mixture were subjected to PCR amplification. The sequences of the different primers used and the corresponding annealing temperatures are indicated in Table 1, while their approximate positions are shown in Fig. 3A. Each reaction, in a final volume of 50 μl, contained: 2 μl of cDNA, 1× Taq polymerase buffer, 1.7 mM $MgCl_2$, 0.2 mM dNTPs, 250 nM each primer and 0.5 U of *PfuTurbo* DNA polymerase (Stratagene, La Jolla, CA). After a preliminary denaturation step (1 min at 94°C), cycling conditions were: denaturation (94°C, 1 min), annealing (*1* min at the temperatures reported in Table 1) and extension (1 min/kb product, 72°C), usually for 30–35 cycles. PCR products were analyzed on 2% agarose gels, stained with ethidium bromide. A semi-quantitative estimation of the β4GalNAc-T transcript was obtained by the simultaneous amplification of β4GalNAc-T and β-actin transcripts using primer pair L.1/R.1 at the usual concentration of 250 nM and the β-actin–specific primers ACTL.3/ACTR.3 at a concentration of 25 nM. For this purpose we used InViTaq, DNA polymerase (Eppendorf, Milan, Italy). The intensity of the bands was quantified by Kodak Digital Science 1D software.

Rapid Amplification of the 5′-cDNA End (5′-RACE) and Sequencing—Three microliters of mRNA (equivalent to 10 μg of total RNA) were subjected to reverse transcription using oligonucleotide R7 as a primer with the Invitrogen (Paisley, UK) 5′-RACE kit according to the manufacturer's instructions. The resulting cDNA was 5′-ligated with GeneRacer oligonucleotide (Invitrogen) and subjected to PCR amplification with GeneRacer primer (Invitrogen) and oligonucleotide R.8 as follows: preliminary denaturing step: 94°C, 2 min; 5 cycles (denaturing 94°C, 30 s; annealing-extension 72°C 1 min), followed by 30 cycles (denaturing 94°C, 30 s; annealing 69°C 30 s; extension 72°C 1 min). The resulting PCR product was gel isolated, cloned using a TOPO TA cloning kit (Invitrogen) and sequenced. Sequence analyses were performed automatically using a Beckman-Coulter CEQ2000XL DNA analysis system.

Construction of the β4GalNAc-T Expression Vector and Transfection—The β4GalNAc-T cDNA was amplified using primer pair L16/R.10 with the following program: preliminary denaturing step: 94°C, 2 min; then 35 cycles of the following program: denaturing 94°C, 1 min; annealing 60 1 min; extension 72°C 2 min, followed by a final extension of 5 min at 72°C. The purified PCR product was cloned directly in the pcDNA3.1 Directional TOPO Expression vector (Invitrogen). The presence of the insert in proper orientation in four randomly chosen colonies was confirmed by sequencing.

COS-7 cells were transiently transfected with the pcDNA3.1/β4GalNAc-T expression vector as follows: 75 cm^2 flasks containing cells at 70% confluency were incu-

bated in 3 ml of serum-free DMEM containing 250 μg/ml DEAE-dextran, 50 μg/ml cloroquine and 10 μg/ml expression vector. After 3 h at 37°C, the cells were incubated for 2 min with 10% dimethylsulfoxide in PBS. After two rinses with serum-free DMEM, the cells were incubated with complete medium and harvested 3 days later.



B

```
                                                                  at   -58
aggtggctgcagaggcgaggtgacggcgcgtgcggaacgaactctgcacccccaggaatg        3
                                                                   M     1
gggagcgctggctttccgtgggaaaattccacgtggaggtggcctctcgcggccgggaa       63
 G  S  A  G  F  S  V  G  K  F  H  V  E  V  A  S  R  G  R  E      21
tgtgtctcggggacgcccgagtgtgggaatcggctcgggagtgcgggcttcggggatctc      123
 C  V  S  G  T  P  E  C  G  N  R  L  G  S  A  G  F  G  D  L      41
tgcttggaactcagaggcgctgacccagcctggggcccgtttgctgcccacgggaggagc      183
 C  L  E  L  R  G  A  D  P  A  W  G  P  F  A  A  H  G  R  S      61
cgccgtcagggctcgagatttctgtggctcctcaagatattggtcataatcctggtactt      243
 R  R  Q  G  S  R  F  L  W  L  L  K  I  L  V  I  I  L  V  L      81
ggcattgttggatttatgttccggaagcatgttccttcaagcagtgttcagcagccccaag      303
 G  I  V  G  F  M  F  G  S  M  F  L  Q  A  V  F  S  S  P  K     101
ccagaactcccaagtcctgccccgggtgtccagaagctgaagcttctgcctgaggaacgt      363
 P  E  L  P  S  P  A  P  G  V  Q  K  L  K  L  L  P  E  E  R     121
ctcaggaacctcttttcctacgatggaatctggctgttcccgaaaaatcagtgcaaatgt      423
 L  R  N  L  F  S  Y  D  G  I  W  L  F  P  K  N  Q  C  K  C     141
gaagccaacaaagagcagggaggttacaactttcaggatgcctatggccagagcgacctc      483
 E  A  N  K  E  Q  G  G  Y  N  F  Q  D  A  Y  G  Q  S  D  L     161
ccagcggtgaaagcgaggagacaggctgaatttgaacactttcagaggagagaagggctg      543
 P  A  V  K  A  R  R  Q  A  E  F  E  H  F  Q  R  R  E  G  L     181
ccccgcccactgcccctgctggtccagcccaacctccccttgggtacccagtccacgga       603
 P  R  P  L  P  L  L  V  Q  P  N  L  P  F  G  Y  P  V  H  G     201
gtggaggtgatgcccctgcacacggttcccatcccaggcctccagtttgaaggacccgat      663
 V  E  V  M  P  L  H  T  V  P  I  P  G  L  Q  F  E  G  P  D     221
gcccccgtctatgaggtcaccctgacagcttctctggggacactgaacacccttgctgat      723
 A  P  V  Y  E  V  T  L  T  A  S  L  G  T  L  N  T  L  A  D     241
gtcccagacagtgtggtgcagggcagaggccagaagcagctgatcatttctaccagtgac      783
 V  P  D  S  V  V  Q  G  R  G  Q  K  Q  L  I  I  S  T  S  D     261
cggaagctgttgaagttcattcttcagcacgtgacatacaccagcacggggtaccagcac      843
 R  K  L  L  K  F  I  L  Q  H  V  T  Y  T  S  T  G  Y  Q  H     281
cagaaggtagacatagtgagtctggagtccaggtcctcagtggccaagtttccagtgacc      903
 Q  K  V  D  I  V  S  L  E  S  R  S  S  V  A  K  F  P  V  T     301
atccgccatcctgtcatacccaagctatacgaccctggaccagagaggaagctcagaaac      963
 I  R  H  P  V  I  P  K  L  Y  D  P  G  P  E  R  K  L  R  N     321
ctggttaccattgctaccaagactttcctccgcccccacaagctcatgatcatgctccgg     1023
 L  V  T  I  A  T  K  T  F  L  R  P  H  K  L  M  I  M  L  R     341
agtattcgagagtattacccagacttgaccgtaatagtggctgatgacagccagaagccc     1083
 S  I  R  E  Y  Y  P  D  L  T  V  I  V  A  D  D  S  Q  K  P     361
ctggaaattaaagacaaccacgtggagtattacactatgcccttgggaagggttggttt     1143
 L  E  I  K  D  N  H  V  E  Y  Y  T  M  P  F  G  K  G  W  F     381
gctggtaggaacctggccatatctcaggtcaccaccaaatacgttctctgggtggacgat     1203
 A  G  R  N  L  A  I  S  Q  V  T  T  K  Y  V  L  W  V  D  D     401
gattttctcttcaacgaggagaccaagattgaggtgctggtggatgtcctggagaaaaca     1263
 D  F  L  F  N  E  E  T  K  I  E  V  L  V  D  V  L  E  K  T     421
gaactggacgtggtaggcggcagtgtgctgggaaatgtgttccagtttaagttgttgctg     1323
 E  L  D  V  V  G  G  S  V  L  G  N  V  F  Q  F  K  L  L  L     441
gaacagagtgagaatggggcctgccttcacaagaggatgggatttttccaaccccctggat     1383
 E  Q  S  E  N  G  A  C  L  H  K  R  M  G  F  F  Q  P  L  D     461
ggcttccccagctgcgtggtgaccagtggcgtggtcaacttcttcctggcccacacggag     1443
 G  F  P  S  C  V  V  T  S  G  V  V  N  F  F  L  A  H  T  E     481
cgactccaaagagttggctttgatccccgcctgcaacgagtggctcactcagaattcttc     1503
 R  L  Q  R  V  G  F  D  P  R  L  Q  R  V  A  H  S  E  F  F     501
attgatgggctagggaccctactcgtggggtcatgcccagaagtgattataggtcaccag     1563
 I  D  G  L  G  T  L  L  V  G  S  C  P  E  V  I  I  G  H  Q     521
tctcggtctccagtggtggactcagaactggctgccctagagaagacctacaatacatac     1623
 S  R  S  P  V  V  D  S  E  L  A  A  L  E  K  T  Y  N  T  Y     541
cggtccaacaccctcacccgggtccagttcaagctggccctccactacttcaagaaccat     1683
 R  S  N  T  L  T  R  V  Q  F  K  L  A  L  H  Y  F  K  N  H     561
ctccaatgtgccgcataaaggtgtgagggcataggagaaacactaggctggctggttatg     1733
 L  Q  C  A  A                                                   566
gtatctatagcaggccaccaaaaactggactcctgataggtgaacgttgtaccaaaccag     1793
ctggtgggtagggaaaagggaaatggctcagttactgg
```

RESULTS

Identification of Human β4GalNAc-T Exon Sequences in Genomic Databases—A search in public mouse genomic databases revealed that the reported cDNA sequence of mouse CT β4GalNAc-T (GenBank accession No. L30104) (*15*) is composed of 11 exons mapping in chromosome 11. Based on the published partial cDNA sequence of human β4GalNAc-T (*16*) (GenBank accession No. S83275) and the homology with mouse exons, we identified in human public genomic databases (chr. 17q23, contig NT_10783) sequences that are highly homologous to mouse exons 2–11. The predicted sequences were confirmed in cDNA from Caco2 cells by PCR amplification and sequencing. However, the search for a human genomic sequence showing convincing homology with mouse exon 1 was unfruitful.

Cloning of the Full Coding Sequence—To identify the 5′-regions of human β4GalNAc-T non homologous to the mouse cDNA we subjected the RNA from differentiated Caco2 cells to 5′-RACE. An 800 bp RACE product was found to contain a novel region nearly identical to a human genomic sequence located in contig NT_10783, which is about 7,000 bp upstream of exon 2. This region turned out to be completely different from mouse exon 1. Using a 5′-primer designed on the basis of this new sequence information, a 1,862 bp product containing the whole coding sequence, as well as the 5′UTR and part of the 3′-UTR untranslated sequences, was PCR amplified (Fig. 1A) and successively cloned. The nucleotide sequence and the predicted amino acid sequence of the β4GalNAc-T clone are reported in Fig. 1B.

Analysis of the Open Reading Frame (ORF) and Genomic Organization of Human and Mouse β4GalNAc-T—The β4GalNAc-T cDNA sequence expressed by differenti-

Fig. 1. **A: PCR amplification of the full coding sequence of β4GalNAc-T.** Primer pair L5/R.10 yields a 1862 bp product containing the full coding sequence, the 5′-UTR and part of the 3′-UTR. **B: Nucleotide- and predicted amino acid (single letter code) sequence of the cloned β4GalNAc-T from differentiated Caco2 cells.** The predicted transmembrane domain is underlined. This sequence has been deposited in the GenBank under the accession number AF510036.

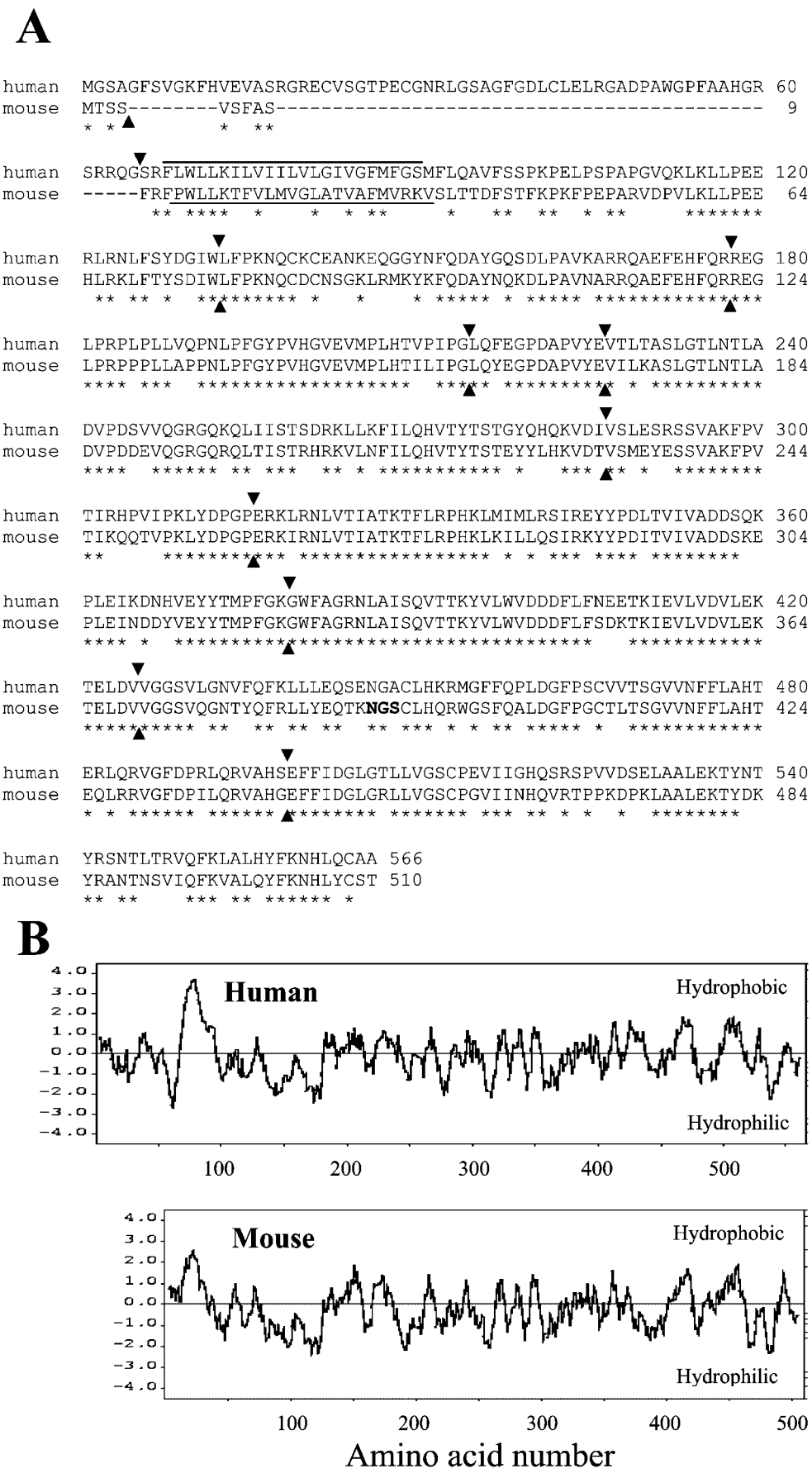


Fig. 2. **A: Multiple sequence alignment (ClustalW) of human and mouse β4GalNAc-T amino acid sequences.** The two putative transmembrane domains are underlined. The vertical arrowheads, mark the exon borders in the human and mouse sequences. The single potential N-glycosylation site in the mouse lymphocyte enzyme, not conserved in the Caco2 sequence, is in bold. "*" denotes identity. **B: Comparison of the Kyte-Doolitle plots of the β4GalNAc-T from human Caco2 and mouse lymphocytes.** The Caco2 enzyme shows a stronger hydrophobicity in the transmembrane domain and a much longer cytoplasmic tail.

Table 2. **Exon-intron characterization of human and mouse β4GalNAc-T genes.**

Species	Exon No.	Exon size	Human/mouse homology (%)	Exon junctions	5′ donor	3′ acceptor	Intron size*
human	1	253					
mouse	1	52					
human	2	201		1–2	CCGTCAGGGgtatgtgag	tcccaacagCTCGAGATT	>7,000
mouse	2	216	70.6**	1–2	GACTTCCAGgtagtgagt	ctttttcagCGTGTCTTT	22,846
human	3	138		2–3	TGGAATCTGgtgagagac	ttgtttcagGCTGTTCCC	587
mouse	3	138	73.0	2–3	TGACATCTGgtgagactc	ttgttctagGCTCTTCCC	647
human	4	107		3–4	TCAGAGGAGgtaatgcgg	ttcctgcagAGAAGGGCT	10,627
mouse	4	107	84.1	3–4	TCAGAGGAGgtgatgcag	ctcctgaagAGAAGGGCT	2,059
human	5	38		4–5	CCATCCCAGgtaagtaca	gtcctccagGCCTCCAGT	3,659
mouse	5	38	86.8	4–5	TCATCCCAGgtaggtaca	ctcttctagGCCTCCAGT	4,826
human	6	181		5–6	GTCTATGAGgtgagtcct	ccaccctagGTCACCCTG	2,433
mouse	6	181	84.0	5–6	GTCTATGAGgtaagagtc	catgtctagGTCATCCTG	7,662
human	7	87		6–7	TAGACATAGgtgagagcc	tccctccagTGAGTCTGG	1,337
mouse	7	87	74.7	6–7	TGGACACAGgtctgtctt	tccttccagTAAGTATGG	2,149
human	8	188		7–8	CTGGACCAGgtaaggccc	tctgcccagAGAGGAAGC	3,446
mouse	8	188	86.2	7–8	CTGGACCTGgtaagactc	cttctctagAGAGGAAGA	4,440
human	9	141		8–9	TTTGGGAAGgtatgtccc	catctacagGGTTGGTTT	1,838
mouse	9	141	88.7	8–9	TTTGGGAAGgtatgtccc	tgcctgcagGGCTGGTTT	743
human	10	220		9–10	CTGGACGTGgtaagggac	gctggctagGTAGGCGGC	2,426
mouse	10	220	77.3	9–10	CTGGATGTGgtaaggaac	gctgactagGTGGGTGGC	1,380
human	11	202***		10–11	CTCACTCAGgtgggaagg	tctttgcagAATTCTTCA	622
mouse	11	202	70.3	10–11	CCCACGGAGgtgagaggc	tctttgcagAGTTCTTTA	473

Exon and intron sequences are indicated in uppercase and lowercase letters, respectively. *The intron size was obtained from public genomic databases. For some human introns the size cannot be determined exactly because of the presence of undetermined nucleotides (N) in the databases. **The value refers to the homology between the whole human exon 2 and nucleotides 16–216 of mouse exon 2. ***The length of exon 11 refers to the coding region only.

ated Caco2 cells, reported in Fig. 1B, predicts an ORF of 566 aa with a theoretical molecular mass of 63.3 kDa, showing 66.6% identity with mouse CT GalNAc-T (*15*) and 39% identity with human GM2/GD2 synthase (*18*). The first ATG codon, preceded by a 59 bp 5′-UTR, is flanked by a purine in position –3 and by a G in position +4. According to Kozak's rules for translation initiation (*19*), this can be considered a strong initiator. The cytoplasmic tail (aa 1–67, software TMHMM V. 2.0) is extremely long and contains two serine residues (positions 35 and 61) that can potentially undergo phosphorylation (software NetPhos 2.0, score 0.976 and 0.891, respectively). To our knowledge, this is the longest cytoplasmic tail among the glycosyltransferases cloned to date. No other in frame ATG codons exist upstream of the putative transmembrane domain (aa 68–90). In Fig. 2, A and B is shown a direct comparison of the amino acid sequences and the hydropathy analysis (*20*) of human Sd^a- and mouse CT β4GalNAc-T (see also Table 2). The two enzymes display an extensive homology throughout the sequence, except in the cytoplasmic tail, which is much shorter in the mouse lymphocyte enzyme (12 *vs.* 67 residues), and in the putative transmembrane and stem domains, where the homology is limited. The single potential *N*-glycosylation site present in mouse β4GalNAc-T (Asn 390, in bold in Fig. 2A) is not conserved in the human Caco2 enzyme because of the substitution of serine with alanine in the *N*-glycosylation consensus sequence. The lack of potential *N*-glycosylation signals is an uncommon feature among glycosyltransferases. The Kyte-Doolittle plots of the two β4GalNAc-Ts reveal the presence of a prominent hydrophobic sequence in the NH_2-terminal region of both molecules, indicative of the type II transmembrane topology characteristic of many other glycosyltransferases. However, the transmembrane domain of the Caco2 enzyme displays a stronger hydrophobic nature and is preceded by a strongly hydrophilic sequence, that is not present in the mouse lymphocyte enzyme. Both the human and mouse enzymes contain the DxD motif (aa 400–402 and aa 343–345 in the human and mouse enzymes, respectively) shared by many glycosyltransferases (*21*); the sequence is embedded in the longest (37 aa) region of homology between the two enzymes. Consistent with the notion that this region is actively involved in catalysis, it also shows the highest homology with GM2/GD2 synthase.

In both enzymes, the coding region is contained in 11 exons (Table 2). Exons 3–10 and the coding part of exon 11 are of identical size in the two forms and display high degrees of homology and conservation of exon/intron boundaries. The exonsborders in the amino acid sequence of the two enzymes are indicated by vertical arrowheads in Fig. 2A. As shown, in the mouse enzyme the first exon encodes only the first four amino acids of the cytoplasmic tail, while the rest of this domain is encoded by the first 24 nucleotides of exon 2. On the contrary, in the human Caco2 enzyme, exon 1 encodes 64 out of the 67 residues of the cytoplasmic tail. It is not currently known whether the difference between the predicted NH_2-terminal domains of the mouse and human enzymes is due to a species difference or, more likely, to the different tissue origins of the two forms. The human gene spans at least 35000 bp of genomic sequence, while the mouse gene spans at least 47,000 bp.

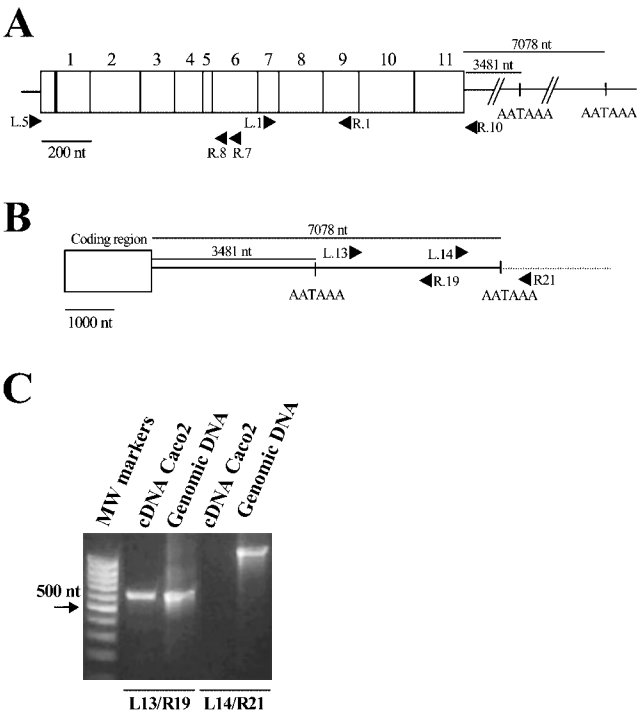


Fig. 3. **A: Schematic representation of the β4GalNAc-T transcript with the positions of the PCR primers used in this study and PCR analysis of the 3′ UTR.** A: The coding region (boxed) is preceded by a 59 bp 5′ UT sequence and is dispersed over 11 exons. The vertical line inside exon 1 represents the transmembrane domain. In genomic DNA, there are two classical AATAAA polyadenylation signals located 3481 and 7078 bp downstream of the translational stop codon. **B: Schematic representation of the putative β4GalNAc-T 3′ UTR.** PCR primer pair L.13/R19 amplifies a product (C) of the expected size (651 bp) from both the cDNA and genomic DNA, while primer pair L.14/R21 amplifies a product of the expected size (1285 bp) only from genomic DNA, suggesting that the termination of transcription may occur at the level of the second polyadenylation signal.

A schematic representation of the coding region and of the putative 3′-UTR with the position of the PCR primers used in this study is provided in Fig. 3, A and B. A search of the genomic sequences reveals the presence of two classical AATAAA polyadenylation signals located 3,481 and 7,078 bp downstream from the translational stop codon. The length of the 3′-UTR was preliminarily investigated by RT-PCR using primer pairs complementary to different regions of the 3′-UTR. As shown in Fig. 3B, the primer pair L13/R19 is complementary to a distal region of the putative 3′-UTR between the first and the second AATAAA polyadenylation signals, while the primer pair L14/R21 contains the second AATAAA sequence in between. Fig. 3C shows that the first primer pair amplifies a product of the expected size in both the cDNA from differentiated Caco2 cells and in genomic DNA. On the contrary, the second primer pair (L14/R21) yields a product only with genomic DNA. These data strongly suggest that in Caco2 cells the transcription of the β4GalNAc-T gene may be terminated at the level of the second polyadenylation signal which is consistent with the expression of a very long transcript.

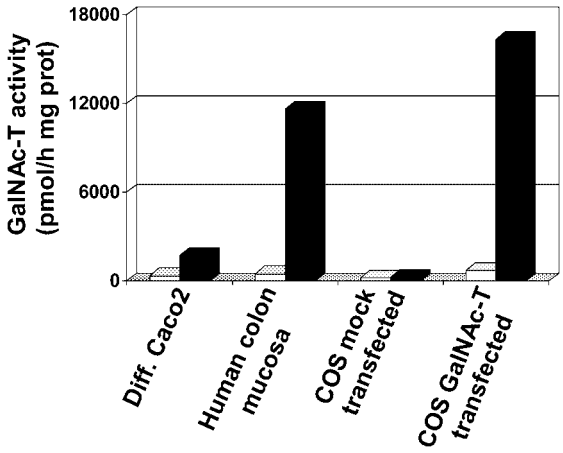


Fig. 4. **Transient expression of β4GalNAc-T.** COS-7 cells were transfected with an expression vector containing the β4GalNAc-T cDNA driven by the cytomegalovirus promoter or with an insertless pcDNA3.1 vector (mock transfection). Cells were harvested three days later and the GalNAc transferase activity was measured using either fetuin (black bars) or asialofetuin (white bars) as acceptors. The difference between the incorporation on the two acceptors is a measure of the β4GalNAc-T activity. A comparison with the level of enzyme activity reached by differentiated Caco2 cells and by human colon mucosa, which is probably the tissue displaying the highest β4GalNAc-T activity in the body, reveals that the β4GalNAc-T activity induced by transfection with the cloned cDNA is extremely high.

Transient Expression of the β4GalNAc-T cDNA in COS-7 Cells—As previously demonstrated, the β4GalNAc-T enzyme activity can be measured as the differential incorporation of radioactive GalNAc on an α2,3-sialylated glycoprotein (such as fetuin) and its corresponding asialylated counterpart (*12*). As shown in Fig. 4, transfection in COS-7 cells of the β4GalNAc-T cDNA under the control of the cytomegalovirus promoter results in the induction of an extremely high GalNAc-T activity that is active on fetuin but not on asialofetuin (Fig. 4). On the contrary, in mock-transfected COS-7 cells the GalNAc transferase activities for both asialofetuin and fetuin are expressed at negligible levels. The comparison with the activity expressed by differentiated Caco2 cells and by a human colon mucosa sample, the tissue probably expressing the highest level of β4GalNAc-T, reveals that the activity reached by transfected COS-7 cells is extremely high. This demonstrates conclusively that the cloned cDNA encodes a GalNAc transferase specific for sialylated acceptors whose features are identical to those of the Sd^a GalNAc transferase.

Expression of β4GalNAc-T—The expression of β4GalNAc-T was investigated by a semi-quantitative approach based on the simultaneous amplification of the β-actin transcript in several colon cancer cell lines as well as in normal human colonic mucosa and peripheral blood lymphocytes (Fig. 5, upper panel). As shown in the lower panel, which reports the β4GalNAc-T/β-actin ratios, high levels of the β4GalNAc-T transcript are expressed only by differentiated Caco2 cells and normal colonic mucosa, while all other cell lines and PBL express a low level of transcript. Interestingly, the differentiation of Caco2 cells is accompanied by the accumulation of the β4GalNAc-T transcript. Collectively, these data are con-

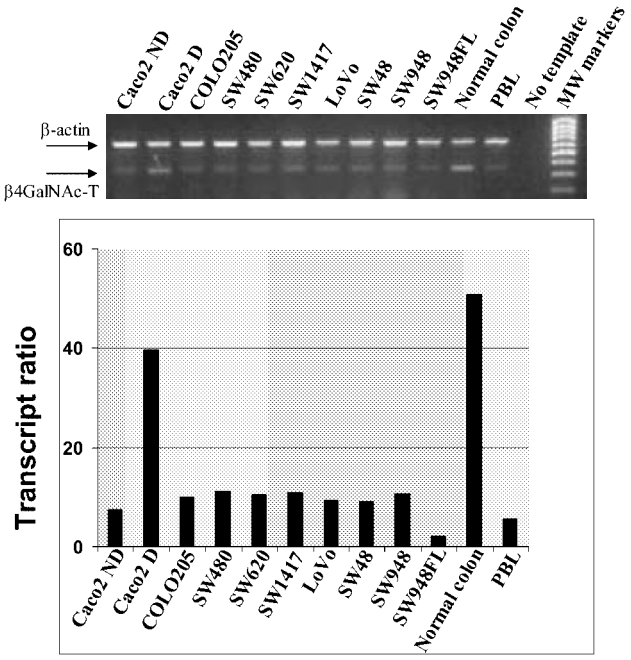


Fig. 5. **cDNAs from colon cancer cell lines, normal colonic mucosa and peripheral blood lymphocytes (PBL) were subjected to simultaneous amplification of the β4GalNAc-T and of β-actin transcripts.** The ratio between the intensity of the β4GalNAc-T/β-actin bands, reported in the lower panel, indicates high levels of the β4GalNAc-T transcript only in differentiated Caco2 cells and in normal colonic mucosa.

sistent with current and previous enzyme activity data reporting a very high level of activity in normal colonic mucosa (*12*) and a marked increase in activity upon differentiation of Caco2 cells (*13*). The detection of the β4GalNAc-T transcript, even though at a low level, in cell lines where the activity was not previously detected is not surprising considering the extreme sensitivity of the RT-PCR technique.

DISCUSSION

In this paper we report the molecular cloning of a cDNA from a human intestinal source encoding a GalNAc transferase strictly requiring sialic acid in the acceptor[1]. This feature, together with the high homology with the previously cloned mouse CT β4GalNAc-T and the tissue distribution of the transcript indicates that this enzyme is the human Sd^a β4GalNAc-T. The ORF predicts a protein with the type II orientation typical of many glycosyltransferases, with an extremely long cytoplasmic tail and the absence of any potential N-glycosylation site. The very long cytoplasmic tail, probably the longest among the glycosyltransferases cloned to date, contains two serine residues that are potential sites of phosphorylation. At present, it is not known whether these serine residues may actually be phosphorylated, but the possible involvement of phosphorylated serine residues in regulatory roles is intriguing. Another possibility that should be considered is the subcellular localization of the enzyme. In fact, besides the well known Golgi localization, glycosyltransferases have sometimes been found localized in post-Golgi compartments or on the plasma membrane (*22*). The fact that this GalNAc-T acts after sialyltransferases is consistent with a very distal localization of the enzyme, and the unusually long cytoplasmic tail may conceivably be involved in an unconventional subcellular localization of the enzyme. In the human β4GalNAc-T, the DxD signature is found at the end of a 30-aa-long sequence that shows 100% identity with the mouse enzyme and 96.6% identity with the human GM2/GD2 synthase. This high degree of homology is consistent with the notion that this part of the molecule is involved in substrate recognition by the two enzymes. Mutation of either aspartate residue in the DxD motif of GM2/GD2 synthase results in a complete loss of enzyme activity but leaves the UDP-binding activity unaffected (*23*), suggesting that this motif is involved in catalysis, rather than in nucleotide binding.

N-Acetylgalactosamine is present at the terminal reducing end of *O*-linked chains and is also frequently present in glycolipids, but is rarely found in *N*-linked chains of glycoproteins. The presence of a β1,4GalNAc on α2,3-sialylated termini typically found in *N*-linked chains, constitutes an unconventional oligosaccharide epitope that could conceivably be involved in specific recognition phenomena. Several studies published in past years demonstrate the specific role played by the Sd^a oligosaccharide epitope in different systems. For example, antibodies directed against the CT antigens are able to block *in vitro* the cytotoxic function of cytotoxic T lymphocytes (*24*, *25*). Moreover, β4GalNAc-T expression and β4GalNAc-containing oligosaccharides are confined at the neuromuscular junction where they mediate adhesion with laminin β2, a component of the synaptic cleft (*26*, *27*). In transgenic mice where β4GalNAc-T is overexpressed in extrasynaptic regions in skeletal myofibers, the development of the neuromuscular synapse is profoundly altered, suggesting that the Sd^a antigen is involved in controlling the expression of synaptic molecules (*28*).

In normal human colon, the Sd^a epitope is a major structure terminating core 3 mucin oligosaccharides (GlcNAcβ1,3GalNAc), while sialyl Lewis antigens are found as minor components (*29*). On the contrary, in colon cancer, sialyl Lewis antigens are frequently expressed but, curiously, their level of expression correlates poorly with the level of expression of the sialyl- and fucosyltransferases involved in their biosynthesis (*30*). Owing to the fact that both β4GalNAc-T and the fucosyltransferases involved in the biosynthesis of sialyl Lewis antigens act on the same acceptor substrate, it is likely that the level of expression of β4GalNAc-T plays a key role in determining the expression of sialyl Lewis antigens. Thus, the lack of sialyl Lewis antigen expression in normal colon may be related to the very high level of β4GalNAc-T expression previously reported in normal colon (*12*) and confirmed in this study, while the appearance of sialyl Lewis antigens in colon cancer would be allowed by the more or less pronounced downregulation of β4GalNAc-T in cancer tissues (*12*, *16*). If it is considered that sialyl Lewis antigens play a fundamental role in the adhesion of cancer cells to the endothelium during metastasis, it is reasonable to expect that β4GalNAc-T may act as a metastasis-suppression gene in colon cancer

and, possibly, in other malignancies. The elucidation of the full coding sequence of human Sd[a] β4GalNAc-T will provide the basis for future studies on the biology of the Sd[a] carbohydrate determinant in colon cancer and other biological systems.

GenBank accession No. AF510036. During the review process of this manuscript, other cDNA sequences of human β4GalNAc-T have been deposited in the GenBank with the accession numbers AJ517770 and AJ51777. This work was supported by grants from MIUR (ex60% and ex40%) and from the University of Bologna (funds for selected research topics). We thank Dr. Marco Trinchera, University of Insubria, Italy, for the gift of COS-7 cells.

REFERENCES

1. Morton, J.A., Pickles, M.M., and Terry, A.M. (1970) The Sda blood group antigen in tissues and body fluids. *Vox Sang.* **19**, 472–482
2. Donald, A.S., Yates, A.D., Soh, C.P., Morgan, W.T., and Watkins, W.M. (1983) A blood group Sda-active pentasaccharide isolated from Tamm-Horsfall urinary glycoprotein. *Biochem. Biophys. Res. Commun.* **115**, 625–631
3. Serafini-Cessi, F. and Dall'Olio, F. (1983) Guinea-pig kidney β-N-acetylgalactosaminyltransferase towards Tamm- Horsfall glycoprotein. Requirement of sialic acid in the acceptor for transferase activity. *Biochem. J.* **215**, 483–489
4. Piller, F., Blanchard, D., Huet, M., and Cartron, J.P. (1986) Identification of a α-NeuAc-(2—-3)-β-D-galactopyranosyl N-acetyl-β-D-galactosaminyltransferase in human kidney. *Carbohydr. Res.* **149**, 171–184
5. Serafini-Cessi, F., Dall'Olio, F., and Malagolini, N. (1986) Characterization of N-acetyl-β-D-galactosaminyltransferase from guinea-pig kidney involved in the biosynthesis of Sda antigen associated with Tamm-Horsfall glycoprotein. *Carbohydr. Res.* **151**, 65–76
6. Dall'Olio, F., Malagolini, N., and Serafini-Cessi, F. (1987) Tissue distribution and age-dependent expression of β-4-N-acetylgalactosaminyl-transferase in guinea-pig. *Biosci. Rep.* **7**, 925–932
7. Conzelmann, A. and Kornfeld, S. (1984) A murine cytotoxic T lymphocyte cell line resistant to Vicia villosa lectin is deficient in UDP-GalNAc: β-galactose β 1, 4-N-acetylgalactosaminyl-transferase. *J. Biol. Chem.* **259**, 12536–12542
8. Conzelmann, A. and Bron, C. (1987) Expression of UDP-N-acetylgalactosamine: β-galactose β 1, 4-N- acetylgalactosaminyltransferase in functionally defined T-cell clones. *Biochem. J.* **242**, 817–824
9. Dall'Olio, F., Malagolini, N., Di Stefano, G., Ciambella, M., and Serafini-Cessi, F. (1990) Postnatal development of rat colon epithelial cells is associated with changes in the expression of the β 1, 4-N-acetylgalactosaminyltransferase involved in the synthesis of Sda antigen and of α 2, 6-sialyltransferase activity towards N-acetyl-lactosamine. *Biochem. J.* **270**, 519–524
10. Malagolini, N., Dall'Olio, F., Di Stefano, G., Minni, F., Marrano, D., and Serafini-Cessi, F. (1989) Expression of UDP-GalNAc: NeuAc α 2, 3Gal β-R β 1, 4(GalNAc to Gal) N-acetylgalactosaminyltransferase involved in the synthesis of Sda antigen in human large intestine and colorectal carcinomas. *Cancer Res.* **49**, 6466–6470
11. Malagolini, N., Dall'Olio, F., Guerrini, S., and Serafini-Cessi, F. (1994) Identification and characterization of the Sda beta 1, 4, N-acetylgalactosaminyltransferase from pig large intestine. *Glycoconj. J.* **11**, 89–95
12. Malagolini, N., Dall'Olio, F., Di Stefano, G., Minni, F., Marrano, D., and Serafini-Cessi, F. (1989) Expression of UDP-GalNAc: NeuAc α 2, 3Gal β-R β 1, 4(GalNAc to Gal) N-acetylgalactosaminyltransferase involved in the synthesis of Sda antigen in human large intestine and colorectal carcinomas. *Cancer Res.* **49**, 6466–6470
13. Malagolini, N., Dall'Olio, F., and Serafini-Cessi, F. (1991) UDP-GalNAc: NeuAc α 2, 3Gal β-R (GalNAc to Gal) β 1, 4-N-acetylgalactosaminyltransferase responsible for the Sda specificity in human colon carcinoma CaCo-2 cell line. *Biochem. Biophys. Res. Commun.* **180**, 681–686
14. Conzelmann, A. and Lefrancois, L. (1988) Monoclonal antibodies specific for T cell-associated carbohydrate determinants react with human blood group antigens CAD and SDA. *J. Exp. Med.* **167**, 119–131
15. Smith, P.L. and Lowe, J.B. (1994) Molecular cloning of a murine N-acetylgalactosamine transferase cDNA that determines expression of the T lymphocyte-specific CT oligosaccharide differentiation antigen. *J. Biol. Chem.* **269**, 15162–15171
16. Dohi, T., Yuyama, Y., Natori, Y., Smith, P.L., Lowe, J.B., and Oshima, M. (1996) Detection of N-acetylgalactosaminyltransferase mRNA which determines expression of Sda blood group carbohydrate structure in human gastrointestinal mucosa and cancer. *Int. J. Cancer* **67**, 626–631
17. Dall'Olio, F., Malagolini, N., Guerrini, S., Lau, J.T., and Serafini-Cessi, F. (1996) Differentiation -dependent expression of human β-galactoside α 2, 6-sialyltransferase mRNA in colon carcinoma CaCo-2 cells. *Glycoconj. J.* **13**, 115–121
18. Nagata, Y., Yamashiro, S., Yodoi, J., Lloyd, K.O., Shiku, H., and Furukawa, K. (1992) Expression cloning of β 1, 4 N-acetylgalactosaminyltransferase cDNAs that determine the expression of GM2 and GD2 gangliosides. *J. Biol. Chem.* **267**, 12082–12089
19. Kozak, M. (1989) The scanning model for translation: an update. *J. Cell Biol.* **108**, 229–241
20. Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132
21. Wiggins, C.A. and Munro, S. (1998) Activity of the yeast MNN1 alpha-1, 3-mannosyltransferase requires a motif conserved in many other families of glycosyltransferases. *Proc. Natl Acad. Sci. USA* **95**, 7945–7950
22. Berger, E.G. (2002) Ectopic localizations of Golgi glycosyltransferases. *Glycobiology*, **12**, 29R–36R
23. Li, J., Rancour, D.M., Allende, M.L., Worth, C.A., Darling, D.S., Gilbert, J.B., Menon, A.K., and Young, W.W., Jr. (2001) The DXD motif is required for GM2 synthase activity but is not critical for nucleotide binding. *Glycobiology*, **11**, 217–229
24. Lefrancois, L. and Bevan, M.J. (1985) Functional modifications of cytotoxic T-lymphocyte T200 glycoprotein recognized by monoclonal antibodies. *Nature*, **314**, 449–452
25. Lefrancois, L. and Bevan, M.J. (1985) Novel antigenic determinants of the T200 glycoprotein expressed preferentially by activated cytotoxic T lymphocytes. *J. Immunol.* **135**, 374–383
26. Martin, P.T., Scott, L.J., Porter, B.E., and Sanes, J.R. (1999) Distinct structures and functions of related pre- and postsynaptic carbohydrates at the mammalian neuromuscular junction. *Mol. Cell. Neurosci.* **13**, 105–118
27. Parkhomovskiy, N., Kammesheidt, A., and Martin, P.T. (2000) N-acetyllactosamine and the CT carbohydrate antigen mediate agrin-dependent activation of MuSK and acetylcholine receptor clustering in skeletal muscle. *Mol. Cell. Neurosci.* **15**, 380–397
28. Xia, B., Hoyte, K., Kammesheidt, A., Deerinck, T., Ellisman, M., and Martin, P.T. (2002) Overexpression of the CT GalNAc transferase in skeletal muscle alters myofiber growth, neuromuscular structure, and laminin expression. *Dev. Biol.* **242**, 58–73
29. Capon, C., Maes, E., Michalski, J.C., Leffler, H., and Kim, Y.S. (2001) Sd(a)-antigen-like structures carried on core 3 are prominent features of glycans from the mucin of normal human descending colon. *Biochem. J.* **358**, 657–664
30. Kudo, T., Ikehara, Y., Togayachi, A., Morozumi, K., Watanabe, M., Nakamura, M., Nishihara, S., and Narimatsu, H. (1998) Up-regulation of a set of glycosyltransferase genes in human colorectal cancer. *Lab. Invest.* **78**, 797–811